Some Applications of Divergences to Robust Inference with Mixed Data

ARIJIT PYNE

Interdisciplinary Statistical Research Unit

Applied Statistics Division, Kolkata

Indian Statistical Institute

DOCTORAL THESIS

Some Applications of Divergences to Robust Inference with Mixed Data

---

Author:

Arijit Pyne

Thesis Advisors:

Prof. Ayanendranath Basu,

Prof. Abhik Ghosh

Interdisciplinary Statistical Research Unit

Applied Statistics Division, Kolkata

Indian Statistical Institute

A thesis submitted to the Indian Statistical Institute

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Statistics

Declaration of Authorship

It is hereby certified that the doctoral thesis titled "*Some Applications of Divergences to Robust Inference with Mixed Data*" by Mr Arijit Pyne has been carried out under our supervision. We also declare that the work it contains is original, and has not been submitted elsewhere for a degree.

Signed: Date: June 2, 2025

Prof. Ayanendranath Basu, ISRU, Kolkata

Prof. Abhik Ghosh, ISRU, Kolkata

"Discovery is the privilege of the child: the child who has no fear of being once again wrong, of looking like an idiot, of not being serious, of not doing things like everyone else. "

Alexander Grothendieck

Acknowledgements

It was a red-letter day, in my life, when I first entered the hallowed premises of the Indian Statistical Institute (ISI), Kolkata. It still gives me goosebumps. I take this opportunity to express my deepest gratitude and sincere regards to my thesis advisors, Prof. Ayanendranath Basu and Prof. Abhik Ghosh, for everything that they have done for me. Their continuous enthusiasm is felt right from the very beginning, such as finalizing a thesis topic and reviewing all the works with a critical eye to help me cross the final hurdle. More importantly, it perpetually shaped my research career over this important phase of my life. Words can not adequately express my appreciation for them. I feel extremely obliged to acknowledge their relentless and painstaking attention to the minutest details in all my work. This story would remain incomplete without appreciating them for reposing their faith and patience in me and providing continuous support and encouragement even when the chips were down. I hope this continues even further, beyond the stipulated time of my PhD tenure. It will not be an exaggeration, if I acknowledge them as my *knights in shining armour!*

I convey my gratitude to the faculty of the Applied Statistics Division (ASD), Interdisciplinary Statistical Research Unit (ISRU) and Statistics and Mathematics Unit (SMU) of ISI, Kolkata. The courses offered by Prof. Ayanendranath Basu, Prof. Tapas Samanta, Prof. Arijit Chakrabarti, Prof. Anil Kumar Ghosh, Prof. Diganta Mukherjee, Prof. Debapriyo Sengupta, Prof. Krishanu Maulik, Prof. Amita Pal have had a definitive influence on my overall understanding of the subject– Statistics, as a whole, which in turn helps my research. The first-hand experiences I gained through assisting Prof. Amita Pal, Prof. Antar Bandopadhyay, and Prof. Arnab Chakraborty have been immensely helpful to me in becoming a researcher.

At this point, I must single out and acknowledge the ever-indebted friendship of Subhrajyoty (Roy), my fellow researcher, whose perseverance and comprehension skills ranging from helping me with efficient R programming to insightful comments deserve special mention. Special thanks go to my seniors and fellow research scholars– Adhidev (Biswas) da, Kaushik (Jana) da, Debasis (Chatterjee) da, Anurag (Dey), Soumya (Chakraborty), Rahul (Roy), Sujay (Das) da, Subhankar (Chattopadhyay), Amarnath (Nandy), Monitirtha (Dey), Anik (Roy), Biswadeep (Ghosh), Nabaneet (Das), Javed da and Soutik (Haldar) for helping me out through different phases of uncertainties in my research career.

Also, I would like to thank the administration of the ISI for providing an exquisite infrastructure– be it the congenial institutional environment and the library that boasts a rich collection of books and journals, or the extensive computing resources and all the other cutting-edge facilities that a researcher could have wished for. All the support staff especially Somnath da of the Interdisciplinary Statistical Research Unit (ISRU) has been immensely helpful throughout my stay.

I am blessed to have a soul mate like Souti in my life. Her continuous support, from helping me understand difficult mathematical concepts to lending her ears and precious time to all my emotional gibberish to get me through the trying phases of my life, always keeps me motivated. Her never-ending encouragement to pursue a research career is my constant source of inspiration to work hard. Also, the support that her parents Mr Sanjit Kumar and Mrs Sanjulika Kumar lend, irrespective of anything whatsoever, has been rock solid that I could ever cherish for my life.

This journey would not have started, had I not been introduced to the beautiful world of Statistics. I am deeply indebted to all my professors at my alma mater, the University of Kalyani. Also, my special thanks go to all my friends, out there, who stuck through

thick and thin.

A special corner in my mind will always be reserved for my great high school Mathematics teacher– Ashok (Ghosh) sir who instilled a love for Mathematics into me that flows underneath.

Most importantly, none of these would have been possible without the love, patience, and care of my parents. I wish my father late Mr Aloke Kumar Pyne were with us, which could not happen as he departed for his heavenly abode so early in his life. Had he been around, he would have been the happiest person along with my mother Mrs Anuradha Pyne to see me complete this great journey. Their love, spirit, and encouragement have been a constant source of strength and support in all of my life.

Arijit Pyne

List of Symbols

θ :	unknown parameter
Θ :	parameter space
$\mathbb{R}$:	set of real numbers
$\mathbb{R}^p$:	p-dimensional Euclidean space
$\mathcal{D}$:	set of all distribution functions
f_θ :	model density indexed by the unknown parameter θ
F_θ :	distribution function of the model density f_θ
$\mathcal{F}$:	family of parametric model densities
G :	true distribution function
g :	true density
G_n :	empirical distribution function based on a sample of size n
g_n :	empirical density based on a sample of size n
χ :	common support of true and model densities
I_p :	Identity matrix of order $p \times p$
A^T :	Transpose of a matrix A
$tr(A)$:	Trace of a matrix A
$\mathcal{N}(\mu, \Sigma)$:	normal distribution with mean vector μ and covariance matrix Σ
$\Phi_1(x), \phi_1(1)$:	distribution and density function of $\mathcal{N}(0,1)$ at x
$\xrightarrow{\mathbb{P}}$:	convergence in probability
$\xrightarrow{a.s.}$:	convergence with probability one/ almost sure convergence

$\xrightarrow{\mathcal{L}}$:	convergence in distribution
$x \mapsto f(x)$:	maps x to $f(x)$
$o_{\mathbb{P}}(1), \mathcal{O}_{\mathbb{P}}(1)$:	stochastic orders
$\|\cdot\|$:	Euclidean norm
∇ :	partial derivative under θ
∇^2 :	second-order partial derivative under θ
$\nabla_{ik\cdots}$:	partial derivatives under the indicated components of θ
$\mathbb{E}_f(\cdot), Var_f(\cdot), Cov_f(\cdot)$:	mathematical expectation, variance and covariance under the density f
$u_\theta(\cdot)$:	likelihood-score function when the model density is f_θ
$I(\theta)$:	Fisher information based on X which is modelled by f_θ
$\otimes$:	Kronecker product
$I_d(X_1, \ldots, X_n)$:	mutual information among $\{X_1, \ldots, X_n\}$ under a divergence 'd'
$\sigma(T_1, \ldots, T_m)$:	sigma-field generated by the set of random variables $\{T_1, \ldots, T_m\}$

List of Abbreviations

ARE	Asymptotic Relative Efficiency
CAN	Consistent and Asymptotically Normal
CUAN	Consistent and Uniformly Asymptotically Normal
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
DCT	Dominated Convergence Theorem
DPD	Density Power Divergence
EP	Exponential-Polynomial Divergence
GSB	Generalized S-Bregman
HD	Hellinger Distance
IF	Influence Function
iid	Independent and Identically Distributed
KLD	Kullback-Leibler Divergence
LD	Likelihood Disparity
LIF	Level Influence Function
MCT	Monotone Convergence Theorem
MDPDE	Minimum Density Power Divergence Estimator
MHDE	Minimum Hellinger Distance Estimator
MLE	Maximum Likelihood Estimator
MI	Mutual Information

NCS	Neyman's chi-square
PD	Power Divergence
PCS	Pearson's chi-square
PIF	Power Influence Function
RAF	Residual Adjustment Function
S-HD	S-Hellinger Distance

List of Figures

List of Tables

Contents

Dedicated to my loving family– Mother Mrs Anuradha Pyne, Father Late Mr Aloke Kumar Pyne and Souti.

This page is intentionally left blank.

Chapter 1

Prologue

1.1 Introduction

Statistical inference partly depends upon observations. An equally important issue is the choice of a probabilistic model and the underlying assumptions that go along with it, such that these random observations are supposed to be generated by the (probabilistic) model. A *statistical distance* between the data and a model naturally quantifies a *discrepancy* between them in a certain way. Using such a distance to draw statistical inference, therefore, aids a statistician to a great extent in elucidating the inferential results.

However, the ways to construct statistical distances between the data and a model are generally varied in the literature. The choice of any such distance function may greatly influence the "behaviour" of a statistical decision rule. This, in a way, grabs much attention from the researchers in the paradigm of *Minimum Distance Estimation* for the last few decades which, in turn, broadens the scope and further drives research in this direction.

Significant methodological works were done in this direction by Wolfowitz and his

associates; see, e.g., Wolfowitz (1952; 1953; 1954; 1957) and Kac et al. (1955). These works include the studies of those estimators that minimize a general class of statistical distances. With that, they also study their large sample properties. These distances are further useful to develop goodness-of-fit tests. A useful review of the minimum distance methods is presented by Parr (1981).

As mentioned before, an estimate of an unknown parameter may be obtained as a minimizer of a chosen statistical distance between the data and a suitable model which is known except for the unknown parameters. Broadly, two types of distances are predominantly used in the statistical literature. These comprise–

(i) the distances between the distribution functions of a model and the data; examples of such distances include the well-known Kolmogorov-Smirnov metric (Kolmogorov, 1933), Cramér-von Mises criterion (von Mises, 1947), Anderson-Darling distance (Anderson and Darling, 1952), and

(ii) the distances between the density functions (when they exist!) of the data and a model. Some such familiar distances in this class are the Pearson's chi-square distance (Pearson, 1900b), the Kullback-Leibler divergence (Kullback and Leibler, 1951), the Hellinger distance, the family of ϕ-divergences (Csiszár, 1964; Ali and Silvey, 1966), the Bregman divergence (Bregman, 1967), the entropy differential metric (Burbea and Rao, 1982) and many more.

The reason why these distances are crucial in statistical inference is the following. To extract valuable information from real-life data we first formulate a probabilistic model that "fairly" explains these data. However, in all probability, these data might contain a small fragment of observations that cannot go along with that model, or rather one would say that they are at odds with the model; this happens even when the model

has been chosen with a lot of care. These observations, that defy the model specification, may be viewed as *outliers* with respect to that model. In reality, one hardly possesses complete knowledge about the probability distribution underlying any data-generating process; hence the question of outliers remains pertinent all along, even if we try out different other models. Yet, the model, by itself, maybe a fair representation of the structure that will govern the generation of future data, and we may want to suitably estimate the model parameters that are not unduly affected by the outliers. It is well known that most classical inferential techniques are quite sensitive to outliers because they fail miserably in producing results that should remain stable in such circumstances. Discarding these unruly observations is not often suggested, primarily because this might have a negative impact on the *asymptotic efficiency* of an estimator. On the contrary, valuable insights about the underlying process often come out from these anomalous data points. The minimum distance inferential methods are particularly useful in this context; when a small proportion (maybe 5% – 20%) of outliers in the data is allowed without changing the model specifications.

The stability study of any inferential procedures in the presence of outliers and model misspecifications belongs to the general area of *Robust Statistics*. The formal theory of robustness started developing around the middle of the last century. Early works, which significantly influenced future research in this area, include Box (1953), Tukey (1960), Huber (1968; 1992) and Hampel (1968; 1971; 1974). Hampel et al. (1986), Huber and Ronchetti (2009), and Maronna et al. (2019) represent some popular textbooks in the area of robust statistics. Minimum distance methods form an important part of the literature in robust statistical inference.

In the context of minimum distance estimation based on distribution functions, the works of Anderson and Darling (1952; 1954), Keifer (1959), Shapiro (1965), Lilliefors (1967), Stephens (1974), Parr (1980; 1981; 1982), Boos (1981; 1982), Scholz and Stephens

(1987), Collins and Weins (1989), Özturk and Hettmansperger (1996), Thas and Ottoy (2003), Mansuy (2005) deserve special mention over here.

Minimum distance estimators defined through density-based divergences are also noted for their strong robustness features, in addition, some of them are seen to be fully *asymptotically efficient*. Breakthroughs towards this direction may be primarily attributed to the works of Beran (1977a; 1977b), Tamura and Boos (1986), Simpson (1987; 1989), Lindsay (1994), and Basu and Lindsay (1994) among many others. Unlike the distribution-based approach, the density-based approach partially settles the time-honoured conflict between two desirable criteria of an estimator– robustness and efficiency which are generally at variance. Also, the properties analogous to the minimum distance estimators continue to hold for the test statistic in the parametric hypothesis testing. Karl Pearson's breakthrough invention of the chi-square statistic (Pearson, 1900b) is an early example of the use of a density-based divergence. Though it has been used ever since in numerous practical applications, much attention was not paid to developing a fully rigorous formal theory out of it at least in the first half of the last century. This landscape changed dramatically from 70s onwards with the influential papers of Beran (1977). This important landmark also heralds a new era, and consequently the entire panorama changes forever and has never been the same as before. Ever since the literature has witnessed a substantial growth in depth and breadth. The books written by Basu et al. (2011), Pardo (2006), and Vajda (1989) serve as useful resources to the researchers for the description of the major research and advancements in this field.

1.2 Mathematical Background and Useful Definitions

1.2.1 General Notations

A certain amount of underlying mathematical rigour is necessary to carry on with the above discussion in the context of specific problems as dealt with in this thesis. This section is devoted to introducing some general notations and a few other useful concepts. This will help the readers in the ensuing chapters.

(i) The term "density function" will be used interchangeably to represent both the probability mass function (pmf) and probability density function (pdf) respectively in the context of discrete and continuous random variables. The lower-case letters (suitably scripted, if necessary) such as g, f, f_θ, δ are used for the density functions. In contrast, its upper-case counterparts, e.g., G, F, F_θ, Λ will be reserved for the probability distribution functions.

(ii) The $0-1$ indicator of the set A is denoted by $\mathbb{1}(A)$. The notation $\delta_y(x)$ stands for a degenerate density function in $\mathbb{R}$ whose entire mass is concentrated at a point y. In other words, $\delta_y(x) = \mathbb{1}\{x = y\}$ and its distribution function is given by $\Lambda_y(x) = \mathbb{1}\{x \geq y\}$.

(iii) The symbol "G" generally represents the true distribution function, while its empirical analogue, based on a data set $\{X_1, \ldots, X_n\}$, is defined as

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \Lambda_{X_i}(x) \text{ for all } x \in \mathbb{R}. \tag{1.1}$$

Similarly, the letter "g" usually refers to the true data-generating density.

(iv) The word "distance" loosely refers to any statistical divergence or disparity function which will be formally introduced in a moment.

(v) Unless specified otherwise, we assume that the unknown parameter θ belongs to the Euclidean space of dimension p. The j-th component of θ is denoted by θ_j $j = 1, \ldots, p$. The associated family of parametric models is referred by either $\mathscr{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ or $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ whenever the latter is appropriate.

(vi) The likelihood-score function is defined as

$$u_\theta(x) = \nabla \log f_\theta(x), \tag{1.2}$$

where ∇ denotes the gradient with respect to θ. Similarly, ∇^2 denotes the Hessian matrix with respect to θ and so on for the higher-order derivatives. The symbol $\nabla_{jk\ldots}$ will denote the partial derivatives with respect to the indicated components of θ. Unless mentioned otherwise, both "*log*" and "ln" mutually represent the natural logarithm.

(vii) The notations $\xrightarrow{\mathbb{P}}, \xrightarrow{a.s.}, \xrightarrow{\mathcal{L}}$ are conventionally used to refer *convergence in probability*, *almost sure convergence* and *convergence in distribution* respectively.

(viii) Sometimes it is mathematically convenient to use $o_{\mathbb{P}}(1)$ (read "small oh-P-1") for convergence in probability to zero; similarly $\mathcal{O}_{\mathbb{P}}(1)$ (read "big oh-P-1") for bounded in probability. More generally, for any sequence of random variables $\{X_n\}$ and a positive real sequence $\{r_n\}$:

$$X_n = o_{\mathbb{P}}(r_n) \text{ means } \frac{X_n}{r_n} \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty, \tag{1.3}$$

$$X_n = \mathcal{O}_{\mathbb{P}}(r_n) \text{ means } \sup_n \mathbb{P}\Big\{\|r_n^{-1} X_n\| > M\Big\} \to 0 \text{ as } n \to \infty \tag{1.4}$$

for some finite constant M. All these limits are defined with respect to the norm $\|\cdot\|$ appropriately defined in the space of random variables. Expressions in (1.3) and (1.4) are understood as: the sequence $\{\|X_n\|\}$ converges to 0 in probability at

the rate r_n, and $\{\|X_n\|\}$ stays bounded at the rate r_n with probability one. Alternatively, when (1.4) is satisfied, we call that X_n is uniformly tight with respect to $\|\cdot\|$ at a rate r_n. When $\{X_n\}$ is not a sequence of random variables, the stochastic order symbols become the usual order symbols "o" and "$\mathcal{O}$" as in Calculus.

Let us recall some common definitions over here.

Definition 1.1. *The parametric family $\mathscr{F}$ is said to be identifiable if the mapping $\theta \mapsto F_\theta$ is one-to-one. When a probability density function f_θ exits, identifiability means that the set $\{f_{\theta_1} = f_{\theta_2}\}$ has a probability measure zero with respect to their common dominating measure for all $\theta_1, \theta_2 \in \Theta$ satisfying $\theta_1 \neq \theta_2$.*

Definition 1.2. *The true density g is said to be compatible with the model family $\mathcal{F}$, if both g and f_θ are supported on a common set uniformly for all $f_\theta \in \mathcal{F}$.*

Definition 1.3. *A statistical divergence between two densities g and f is defined as a map $d : \mathcal{A} \times \mathcal{A} \to \overline{\mathbb{R}}_{\geq 0}$ satisfying*

$$d(g, f) = 0 \iff g = f \text{ a.s.}, \tag{1.5}$$

where $\mathcal{A}$ is the class of densities with respect to a common dominating measure and $\overline{\mathbb{R}}_{\geq 0} := [0, \infty]$ is the non-negative half of the extended real line.

Note that a statistical divergence is not necessarily a mathematical metric, because, generally it is not symmetric in its arguments, and neither does it satisfy the triangle inequality. Hellinger distance and the L_2 distance are prominent exceptions.

1.2.2 Fisher Information

The Fisher information is one of the most fundamental concepts in statistical inference. In a way, it measures the amount of information that a random observation X carries about the unknown parameter θ which comes through a statistical model F_θ for X. The precise mathematical definition is the following. Let X be a random variable which is assumed to follow a density f_θ defined with respect to a σ-finite measure μ. Appropriate regularity conditions (Loève, 1977) are assumed such that

$$\nabla \int_B f_\theta d\mu = \int_B \nabla f_\theta d\mu \text{ for any measurable set } B \tag{1.6}$$

in a suitably defined probability space. The exponential family satisfies these regularity conditions; see Lehmann and Casella (2006) for further illustrations. Often, such a family of parametric distributions are called the *Regular Family*. A simple application of (1.6) implies that the likelihood-score function $u_\theta(x)$ as in (1.2) satisfies $\mathbb{E}_{f_\theta}[u_\theta(X)] = 0$. Then, the Fisher information based on the observable X is defined as

$$I(\theta) = \mathbb{E}_{f_\theta}\Big[u_\theta(X)u_\theta(X)^T\Big]. \tag{1.7}$$

Note that $I(\theta)$ is a non-negative definite matrix. Next, we shall see that $I(\theta)$ is central to the first-order efficiency of an estimator.

1.2.3 First-Order Efficiency

Let $\{X_1, \ldots, X_n\}$ be an independent random sample which is supposed to be modelled by the parametric family $\mathcal{F}$. If $T_n = T_n(X_1, \ldots, X_n)$ is an estimator of an unknown parametric function $m(\theta)$ in $\mathbb{R}$, then it is of substantial interest to restrict our attention

to the class of estimators satisfying

$$\sqrt{n}v^{-\frac{1}{2}}(\theta)\Big(T_n - m(\theta)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ as } n \to \infty, \tag{1.8}$$

for all $\theta \in \Theta$ and some finite $v(\theta) > 0$. See that T_n is a $\sqrt{n}$-consistent estimator for $m(\theta)$. Estimators of this kind are called consistent and asymptotically normal (CAN) estimators. It is important to know if there is a lower bound for the asymptotic variance $v(\theta)$. The Cramér-Rao lower bound for any unbiased estimator of $m(\theta)$ based on an independent sample of size n is given by

$$CRLB(\theta) = \frac{h^T I^{-1}(\theta) h}{n} \tag{1.9}$$

whenever $h = \nabla m(\theta)$ exists, and the Fisher information $I(\theta)$ of a single observation is assumed to be positive definite. For long, it was thought that $v(\theta) \geq CRLB(\theta)$ for all θ. However, there exists a class of estimators, called the super-efficient estimators (Le Cam, 1953), such that they are asymptotically normal with their asymptotic variance $v(\theta)$ never exceeding and sometimes lower than the $CRLB(\theta)$ for some θ. This caused a stir in the Statistics community. Later on Cox and Hinkley (1974) showed that super-efficient estimators are generally of little practical importance to practitioners.

To rule out these super-efficient estimators and base the notion of first-order efficiency on a solid mathematical foundation, it becomes necessary to impose further restrictions on the class of CAN estimators. The consistent and uniformly asymptotically normal (CUAN) estimators are defined to be the CAN estimators for which the above asymptotic normality holds uniformly in θ over compact subsets of Θ. This additional restriction of uniform convergence in distribution eliminates any point of super-efficiency that an estimator may have. In the class of CUAN estimators an estimator T_n is said to

be first-order efficient if

$$v(\theta) = \frac{h^T I^{-1}(\theta) h}{n} \text{ for all } \theta. \tag{1.10}$$

The multivariate generalizations can be analogously obtained.

1.3 Statistical Functional and the Influence Function

A statistical functional T is a map (von Mises, 1939) from the space of distribution functions $\mathcal{D}$ to the parameter space Θ, i.e., $T : \mathcal{D} \to \Theta \subseteq \mathbb{R}^p$. An estimator or a test statistic, when viewed as a statistical functional, is denoted by $T(G_n)$ which depends on a random sample $\{X_1, \ldots, X_n\}$ through the empirical distribution function G_n. Its population version may be similarly given by $T(G)$. Setting them in the context of statistical functionals helps us in deriving the asymptotic distribution of $T(G_n)$, and further study the infinitesimal stability of $T(G)$ at a point. The latter measure primarily stems from the work of Hampel (1974) which will be discussed in this section further down the line. A statistical functional T is called linear if

$$T(\epsilon F + (1-\epsilon)G) = \epsilon T(F) + (1-\epsilon)T(G) \text{ for all } F, G \in \mathcal{D} \text{ and } \epsilon \in [0,1]. \tag{1.11}$$

It is not difficult to verify that a statistical functional T is weakly continuous and linear if and only if it can be expressed as

$$T(G) = \int \tau dG \tag{1.12}$$

for some real-valued function τ (independent of G). When the von Mises derivative T'_G in the direction of G exists, the following expansion holds

$$T(G_n) = T(G) + T'_G(G_n - G) + R_n, \tag{1.13}$$

where R_n is the remainder in this expansion. It is called the von Mises expansion (Fernholz, 2012), and the associated derivative T'_G is called the von Mises derivative which is a linear functional. The first-order (similarly, the higher-order) von Mises derivative of $T(G_n)$ at the direction of G is defined as

$$T'_G(G_n - G) = \frac{\partial}{\partial \epsilon} T(G + \epsilon(G_n - G))\Big|_{\epsilon=0} \tag{1.14}$$

if there exists a real-valued function ψ such that

$$T'_G(G_n - G) = \int \psi d(G_n - G). \tag{1.15}$$

In most applications, ψ maps to the parameter space. Since ψ is uniquely defined up to an additive constant, we can choose ψ such that it satisfies the normalizing condition $\int \psi dG = 0$. This gives

$$T(G_n) - T(G) = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i) + R_n. \tag{1.16}$$

A simple application of the CLT along with Slutsky's theorem gives

$$\sqrt{n}(T(G_n) - T(G)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma) \text{ as } n \to \infty, \tag{1.17}$$

provided $\Sigma = \int \psi \psi^T dG$ is positive definite and $\sqrt{n} R_n = o_{\mathbb{P}}(1)$. However a first-order von Mises derivative is not sufficient to prove $\sqrt{n} R_n = o_{\mathbb{P}}(1)$, but the existence of a

higher-order von Mises derivative is required. The latter condition is rather too strong for many statistical functionals, and so is its first-order Frećhet derivative. Also, a first-order Hadamard differentiability can do the job. Though it is a stronger concept than the von Mises derivative, most statistical functionals possess a first-order Hadamard derivative. We will not discuss these in any further detail. From (1.17) it is clear that $T(G_n)$ is also $\sqrt{n}$-consistent. When both the Hadamard and von Mises derivatives exist they are equal.

Notice that the asymptotic distribution of $T(G_n)$ is driven by ψ when the von Mises derivative exists. This is called the first-order influence function (IF) of the functional T at a distribution degenerate at a point y towards the direction of the distribution G. It is entirely a local concept. Formally, it is obtained as

$$\mathcal{IF}(T, G, y) = T'_G(\Lambda_y - G) = \psi(y) \tag{1.18}$$

using (1.14). However, the influence function can still exist even when the von Mises derivative does not. In simple terms, the value of the influence function $\mathcal{IF}(T, G, y)$ at a point y may be interpreted as an approximation to the relative error in the value of T when the true distribution G is perturbed with infinitesimal proportion by the degenerate distribution Λ_y. It may be seen that the stability of $T(G_n)$ is closely tied to the boundedness of ψ. In the view of this, it is desirable to have a bounded influence function for better robustness of T. The influence function is a local measure of robustness. Thus we see that IF plays a dual role as it is connected to both the robustness and the asymptotic variance of $T(G_n)$. To estimate the standard error of the estimator $T(G_n)$ one often uses

$$\widehat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \psi(X_i)\psi^T(X_i). \tag{1.19}$$

This is called the sandwich estimator of variance. Next, we present the influence functions for the two most commonly used statistical functionals. First, consider the mean functional

$$T_{mean}(G) = \int x dG(x). \tag{1.20}$$

A simple calculation shows that

$$\mathcal{IF}(T_{mean}, G, y) = y - T_{mean}(G). \tag{1.21}$$

Clearly, (1.21) is unbounded when viewed as a function of y. It gives a piece of strong evidence that the mean functional is not robust. Next, we consider the median functional $T_{Me}(G)$ which is defined as

$$T_{Me}(G) = G^{-1}\Big(\frac{1}{2}\Big). \tag{1.22}$$

Here the quantile function is defined as: $G^{-1}(p) = \inf\Big\{x : G(x) \geq p\Big\}$ for $0 \leq p \leq 1$. A standard calculation yields

$$\mathcal{IF}(T_{Me}, G, y) = \begin{cases} \frac{1}{g(T_{Me}(G))} & \text{for } y > T_{Me}(G), \\ -\frac{1}{g(T_{Me}(G))} & \text{for } y < T_{Me}(G) \end{cases} \tag{1.23}$$

which is bounded; this validates the fact that the median is far more robust than the mean as an estimator of location. Later we shall interchangeably use the notations $\mathcal{IF}_1$ and $\mathcal{IF}$ for first-order influence functions. Next, we give the following definition.

Definition 1.4. *Let $T(G_n)$ be an estimator of θ under the parametric family $\mathcal{F}$. Then the functional T is called Fisher consistent if $T(F_{\theta_0}) = \theta_0$ whenever $G = F_{\theta_0}$ for some θ_0.*

A Fisher consistent functional captures the true value of the parameter whenever the true distribution G belongs to the parametric family $\mathcal{F}$. Fisher consistency is a useful concept that often characterizes a statistical functional.

1.4 Maximum Likelihood Estimation

The technique of maximum likelihood estimation is the cornerstone of classical parametric inference. It is one of the oldest concepts for parameter estimation, and its first use can be traced back to as early as the beginning of the nineteenth century in the works of Gauss (1821), Laplace (Stigler, 1986a). However, its systematic development is more recent and may be attributed to Sir R. A. Fisher (1922; 1925; 1934; 1935). To know more about the historical developments of the maximum likelihood method, see Savage (1976), Stigler (1986b), Aldrich (1997), and Hand (1998).

Let $\{X_1, X_2, \ldots, X_n\}$ be an iid random sample from a distribution modelled by the parametric family $\mathcal{F}$ defined in Subsection 1.2.1. Assume that the model distribution F_θ possesses a density function f_θ. The likelihood function of θ based on that random sample is given by

$$L(\theta) \equiv L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i). \tag{1.24}$$

Notice that (1.24) also represents the joint density of the random sample. Any value $\hat{\theta}_{ML}$ in the parameter space Θ, which maximizes the likelihood function $L(\theta)$, is called a maximum likelihood estimator (MLE) of θ. For mathematical tractability, it is often useful to work with the logarithm of the likelihood function which is usually denoted by $\ell(\theta) = \log L(\theta)$. When $\ell(\theta)$ is differentiable (as a function of θ), $\hat{\theta}_{ML}$ is obtained as a solution of the likelihood score equation which sets the partial derivative of the

log-likelihood to zero. This means $\hat{\theta}_{ML}$ solves equation

$$\nabla \log L(\theta) = \sum_{i=1}^{n} u_\theta(X_i) = 0 \tag{1.25}$$

for θ, where u_θ is the likelihood-score function. As discussed before, it is often useful to write $\hat{\theta}_{ML} = T_{ML}(G_n)$ such that

$$\int u_{T_{ML}(G_n)} dG_n = 0. \tag{1.26}$$

Similarly the maximum likelihood functional $T_{ML}(G)$ satisfies

$$\int u_{T_{ML}(G)} dG = 0. \tag{1.27}$$

It is not hard to see that T_{ML} is Fisher consistent. Under suitable regularity conditions, we have

$$\sqrt{n}(T_{ML}(G_n) - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta_0)) \text{ as } n \to \infty \tag{1.28}$$

when $G = F_{\theta_0}$ for some $\theta_0 \in \Theta$ and $I(\theta)$ is the Fisher information. Hence $T_{ML}(G_n)$ is first-order efficient among the class of CUAN estimators (Huzurbazar, 1947; Cramer, 1946; Wald, 1949).

Let $G = F_{\theta_0}$ and contaminate it as $G_\epsilon = (1-\epsilon)G + \epsilon\Lambda_y$ for $\epsilon \in [0,1]$. Then (1.27) gives $\int u_{T_{ML}(G_\epsilon)} dG_\epsilon = 0$. An implicit differentiation of that with respect to ϵ and further evaluation at $\epsilon = 0$ gives the influence function as

$$\mathcal{IF}(T_{ML}, F_{\theta_0}, y) = I^{-1}(\theta_0) u_{\theta_0}(y) \text{ for } y \in \mathbb{R}. \tag{1.29}$$

As the likelihood-score function u_θ is unbounded in most common parametric models,

so is the influence function of T_{ML}. This highlights the fact that observations inconsistent with respect to a parametric model can make $T_{ML}(G_n)$ unstable. Though MLE is a first-order efficient estimator, it is easily affected by outliers and generally has weak robustness properties.

1.5 Why Robust Statistics?

In parametric inference the underlying assumptions are never guaranteed to be exactly true; sometimes they are just mathematically convenient rationalizations of our fuzzy knowledge about the data-generating processes. Often such rationalizations or simplifications may be difficult to verify for real data. This incompleteness of knowledge on our part mandates a statistician to bring forth such statistical methods so that a minor error in the mathematical model should cause only a small deviation in the conclusions. Unfortunately, this criterion is not always satisfied for most common statistical procedures.

Since the middle of the last century, it has become increasingly clear that some well-known methods are overly sensitive to seemingly minor deviations from the assumptions. As an alternative, a plethora of robust procedures have been proposed.

Though the word *robust* is filled with many connotations that are sometimes at variance with one another, we primarily use it to indicate those methods that are insensitive to the observations defying the underlying assumptions. Since the concepts of outliers are varied in the literature, it is often useful to provide a brief account of the type of outliers that we will be dealing with along the way all the time, and further try to give rationales for accommodating them through robust statistical procedures instead of just eliminating them from analyses.

An outlier, as generally understood, is an observation that sits far away from the bulk of data points. This is a geometrical concept. It differs from our notion of outliers as described earlier. Ours is a probabilistic one in nature. However, either of these concepts often leads to the same set of outliers, they are not necessarily the same.

The problems of discarding outliers have been spurring debates among mathematicians and scientists over the last two centuries. (Boscovich, 1757; Bernoulli, 1777; Peirce, 1852; Chauvenet, 1863; Student, 1927; Pearson and Sekar, 1936). However modern research in different biological and other scientific fields indicates that outliers may often contain valuable information about the processes under investigation. There could be several other important factors why such observations should not be just thrown away from the study. One can do better by downweighting dubious observations through suitable robust procedures. But, the rejection or downweighting of outliers constitutes only a tiny fragment of the theory of robust statistics, it also applies to a much wider class of problems dealing with all sorts of deviations from various underlying assumptions in parametric models.

Though the need to come up with a coherent statistical theory was felt for quite some time and many statisticians tried to point out the dramatic lack of robustness in many classical procedures by producing one-off counter-examples, it was realized only through the works of Huber (1964; 1965; 1992), Hampel (1968; 1974; 2011), Bickel (1965) who carried out its systemic development in many directions. Ever since the success stories of the robust statistical procedures to applied problems have been nothing short of overwhelming.

There are several approaches en route to this common goal. Some procedures deal with abstract and more general notions of stability through different topological and

geometric considerations. Others attempt to broaden the scope to some kind of non-parametric statistics (including adaptive estimations) or replace the given parametric model with an enlarged supermodel.

In minimum distance estimation, often a model is used to downweight the effect of inconsistent data points. We see in minimum disparity estimation (Lindsay, 1994; Markatou et al., 1998) that the model densities are employed to construct a *residual adjustment function* (RAF) which provides the necessary downweighting. Observations corresponding to large positive values of these residuals will be the candidates for downweighting. However, we choose a different scheme in Section 1.7. In this, the model density raised to an appropriate power in terms of a related tuning parameter is the central theme of downweighting.

1.5.1 M-Estimation: General Concepts

Let $\mathscr{X}$ be a sample space and $\mathscr{B}_X$ being the smallest σ-field generated by $\mathscr{X}$. The pair $(\mathscr{X}, \mathscr{B}_X)$ is called a measurable space. Further recall that $\mathbb{R}_{\geq 0} = [0, \infty]$. Suppose we draw an iid random sample taking values in $(\mathscr{X}, \mathscr{B}_X)$, and $\rho_\theta : \Theta \times \mathscr{X} \to \mathbb{R}_{\geq 0}$ be a general loss function. Then an M-estimator is defined as

$$T_\rho(G_n) = \arg\min_{\theta \in \Theta} \int \rho_\theta dG_n. \tag{1.30}$$

The corresponding M-functional is given by

$$T_\rho(G) = \arg\min_{\theta \in \Theta} \int \rho_\theta dG, \tag{1.31}$$

provided the integration exists. If ρ_θ admits partial derivatives with respect to each component of θ, then $T_\rho(G)$ and $T_\rho(G_n)$ can be obtained, respectively, as solutions to

the following sets of equations

$$\int \psi_\theta dG = 0 \text{ and } \int \psi_\theta dG_n = 0 \text{ for } \theta, \tag{1.32}$$

where $\psi_\theta = \nabla\rho_\theta$. An implicit differentiation as before gives

$$\mathcal{IF}(T_\rho, G, y) = -J^{-1}\psi_{T_\rho(G)}(y), \tag{1.33}$$

where $J = \int \nabla\psi_{T_\rho(G)}dG$. Under appropriate regularity conditions (Huber, 2011; Hampel et al., 2011; Maronna et al., 2019) it follows that

$$\sqrt{n}(T_\rho(G_n) - T_\rho(G)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_G) \text{ as } n \to \infty \tag{1.34}$$

where $\Sigma_G = J^{-1}KJ^{-1}$ and $K = \int \psi_{T_\rho(G)}\psi^T_{T_\rho(G)}dG$. The class of M-estimators constitutes a very general class of estimators. Asymptotic properties of many well-known M-estimators follow from this result. The parallel development of the general theory of M-estimation and the robust statistics have a considerable overlap. One plays a pivotal role in others' development. For instance, Huber's bounded ψ-function (Huber, 2011) produces one of the early examples of an M-estimator which is very robust. On the other hand, many robust estimators are M-estimators. However, it should be noted that all M-estimators are not robust. For instance, the mean, maximum likelihood estimator, and median are among the most commonly used examples of M-estimators. The first two are non-robust, while the latter is very robust as an estimator of location.

1.6 Minimum Distance Estimation: Discrete Setup

Estimators minimizing statistical distances based on density functions form an important class of robust estimators. As mentioned earlier, many members of this class are

fully asymptotically efficient without compromising its robustness (Lindsay, 1994) per se. In this case we shall primarily discuss the chi-square type distances which belong to a more general family of ϕ-divergence or f-divergence or g-divergence (Csiszár, 1964; Ali and Silvey, 1966; Lindsay, 1994). Pardo (2006) and Basu et al. (2011) are useful resources to learn more about important properties of minimum distance methods. For the rest of this section, we restrict our discussions to the discrete models only. Without loss of generality, we will assume the common support to be $\{0, 1, 2, \ldots, \}$.

1.6.1 Disparities

Suppose an iid sample $\{X_1, \ldots, X_n\}$ drawn from G having a probability density function g is modelled by $f_\theta \in \mathcal{F}$. Let $g_n(x)$ be the relative frequency of data at a point x. Our objective is to find an estimator which is (first-order) efficient, but robust for θ. In the next definition, we define a "disparity" (Lindsay, 1994) between $g_n = (g_n(0), g_n(1), \ldots)^T$ and $f_\theta = (f_\theta(0), f_\theta(1), \ldots)^T$.

Definition 1.5. *(Basu et al., 2011, Definition 2.1) Let the function C defined on $[-1, \infty)$ be strictly convex and thrice differentiable. Suppose it also satisfies $C(0) = 0$. Then a disparity or divergence between g_n and f_θ induced by the C-function is defined as*

$$\rho_C(g_n, f_\theta) = \mathbb{E}_{f_\theta}\Big[C(\delta_n(X))\Big], \tag{1.35}$$

where $\delta_n(x) = \frac{g_n(x)}{f_\theta(x)} - 1$ is the Pearson residual at x.

The conditions, imposed on the C-function, are called the *disparity conditions*. Using strict convexity of C one can easily show that $\rho_C(g_n, f_\theta) \geq 0$, and the equality holds if and only if $g_n(x) = f_\theta(x)$ for all x. Next, we mention some well-known disparities. The

likelihood disparity, given by

$$LD(g_n, f_\theta) = \sum \left\{ g_n \log\left(\frac{g_n}{f_\theta}\right) + (f_\theta - g_n) \right\} = \sum g_n \log\left(\frac{g_n}{f_\theta}\right), \tag{1.36}$$

is generated by $C(\delta_n) = (\delta_n + 1)\log(\delta_n + 1) - \delta_n$. The symmetric opposite of LD is the Kullback-Leibler divergence (Kullback and Leibler, 1951) that is given by

$$KLD(g_n, f_\theta) = \sum \left\{ f_\theta \log\left(\frac{f_\theta}{g_n}\right) + (g_n - f_\theta) \right\} = \sum f_\theta \log\left(\frac{f_\theta}{g_n}\right). \tag{1.37}$$

This is generated by $C(\delta_n) = \delta_n - \log(1 + \delta_n)$. The (twice, squared) Hellinger distance having the form

$$HD(g_n, f_\theta) = 2\sum (g_n^{1/2} - f_\theta^{1/2})^2 \tag{1.38}$$

is generated by $C(\delta_n) = 2((\delta_n + 1)^{1/2} - 1)$. Similarly, Pearson's chi-square (divided by 2) is defined as

$$PCS(g_n, f_\theta) = \sum \frac{(g_n - f_\theta)^2}{2f_\theta} \tag{1.39}$$

for $C(\delta_n) = \frac{\delta_n^2}{2}$. When $C(\delta_n) = \frac{\delta_n^2}{2(\delta_n+1)}$ we get the Neyman's chi-square (divided by 2) as

$$NCS(g_n, f_\theta) = \sum \frac{(g_n - f_\theta)^2}{2g_n}. \tag{1.40}$$

There are several important classes of disparities each generating several commonly used divergences. Among them, the Cressie-Read family (Cressie and Read, 1984) of power divergences (PD) is arguably the most important one which has the following

form

$$PD_\lambda(g_n, f_\theta) = \frac{1}{\lambda(1+\lambda)} \sum g_n \left[\left(\frac{g_n}{f_\theta} \right)^\lambda - 1 \right] \text{ for } \lambda \in \mathbb{R} \setminus \{-1, 0\}. \tag{1.41}$$

The corresponding C-function is given by

$$C(\delta_n) = \frac{(\delta_n + 1)^{1+\lambda} - (\delta_n + 1)}{\lambda(1+\lambda)} - \frac{1}{\lambda + 1}. \tag{1.42}$$

The power divergence is defined as continuous limits in λ when $\lambda = -1, 0$. Notice that (1.41) generates the NCS, KLD, HD, LD, and the PCS, respectively, for $\lambda = -2, -1, -\frac{1}{2}, 0, 1$. It is easy to see that HD is the only metric in the power divergence family. The summations in (1.36) - (1.41) are taken over the common support $\{0, 1, \dots, \}$ which is suppressed as it is clear from the context.

From (1.35) it is clear that it is possible to modify the C-function in such a way that it becomes non-negative without changing the values of the disparity $\rho_C(g_n, f_\theta)$. In such a case, $C(0) = 0$ is the minimum value of the non-negative C-function, and therefore we should have $C'(0) = 0$ whenever the latter exits. Furthermore, if we assume that $C''(0) \neq 0$, then the C-function may be standardized as $C(x) \mapsto C^*(x) = \frac{C(x) - xC'(0)}{C''(0)}$. Notice that C^* is a convex function, and not only that, it also satisfies $C^*(0) = 0$, $C^{*'}(0) = 0$, $C^{*''}(0) = 1$. This standardization yields $\rho_{C^*}(g_n, f_\theta) = \frac{1}{C''(0)} \rho_C(g_n, f_\theta)$, but the property of the minimum distance estimator remains unchanged all along. This is often helpful for mathematical convenience. Thus, without loss of generality, we assume that the C-function also satisfies $C'(0) = 0, C''(0) = 1$. Vajda (1972) produces some results related to the upper bound of such divergences for any two arbitrary densities.

1.6.2 Asymptotics and Influence Function

Let $\rho_C(g_n, f_\theta)$ be as before, and the minimum disparity estimator (MDE) $\hat{\theta}$ is defined as

$$\rho_C(g_n, f_{\hat{\theta}}) = \min_{\theta \in \Theta} \rho_C(g_n, f_\theta), \tag{1.43}$$

provided it exists. Suppose f_θ viewed as a function of θ is differentiable. Then the MDE $\hat{\theta}$ is obtained as a zero of the estimating equation

$$-\nabla \rho_C(g_n, f_\theta) = \sum A(\delta_n) \nabla f_\theta = 0 \text{ for } \theta, \tag{1.44}$$

where $A(\delta_n) = C'(\delta_n)(\delta_n + 1) - C(\delta_n)$ is called the *Residual Adjustment Function* (RAF) of the disparity. See that $A(x) \uparrow x$, and further, it satisfies $A(0) = 0$ and $A'(0) = 1$. We call $A(\delta_n)$ regular if both $A'(\delta_n)$ and $A''(\delta_n)(1 + \delta_n)$ are bounded on $[-1, \infty)$.

For the Cressie-Read family of power divergences, it turns out that

$$A_\lambda(\delta_n) = \begin{cases} \log(1 + \delta_n) & \text{for } \lambda = -1, \\ \frac{(\delta_n + 1)^{\lambda+1} - 1}{\lambda + 1} & \text{otherwise .} \end{cases} \tag{1.45}$$

From (1.45) it is clear that $A_0(\delta_n) = \delta_n$ which is linear and unbounded. It is because $\delta_n(x)$ can become unusually large at the outlying values of x. In that case, the outlier x has an undue effect on the estimation process when $\lambda = 0$. This is another way of showing that MLE (as $\lambda = 0$) is non-robust. However, $A_\lambda(\delta_n)$ stays bounded for some λ, for instance, when $\lambda = -0.5$ even if x is an outlier. If RAF is unbounded, this yields stable estimators. This is how RAF plays a crucial role in determining the *behaviour* of the minimum distance estimator. Next, we describe its role in the influence function.

Let $T(G) = \theta_g$ be a best-fitting parameter which is obtained via minimization of $\rho_C(g, f_\theta)$

over the parameter space Θ, and $\delta = \frac{g}{f_\theta} - 1$ be the population version of δ_n. Then θ_g solves $\sum A(\delta)\nabla f_\theta = 0$. From the standard theory, it follows that the influence function is obtained as

$$\mathcal{IF}(T, G, y) = T'(y) = J_g^{-1}\left\{A'(\delta(y))u_{\theta_g}(y) - \mathbb{E}_g\Big(A'(\delta)u_{\theta_g}\Big)\right\}, \tag{1.46}$$

when $J_g = \mathbb{E}_g\Big(u_{\theta_g}u_{\theta_g}^T A'(\delta)\Big) - \sum_x A(\delta)\nabla_2 f_{\theta_g}$ is non-singular. In particular, the influence function becomes $T'(y) = I^{-1}(\theta_0)u_{\theta_0}(y)$ when the true distribution belongs to the model family, i.e., $G = F_{\theta_{\theta_0}}$ for some θ_0. In this case, the influence function of any minimum disparity functional coincides with that of the maximum likelihood functional at the true model, which, in turn, implies that $T(G)$ is also first-order efficient. When a RAF is bounded the resulting MDE is robust, but this is not revealed by its influence function at the true model. This highlights the limitations of the influence function.

Next, we state the asymptotic normality result of the MDE.

Theorem 1.1. *(Lindsay, 1994) Let the Assumptions (A1) - (A7) (Basu et al., 2011, pp. 60-61) be true. Then there exists a consistent sequence of roots $\hat{\theta}$ of (1.44) such that $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_g$ and*

$$n^{1/2}(\hat{\theta} - \theta_g) \xrightarrow{\mathcal{L}} \mathcal{N}(0, J_g^{-1}V_g J_g^{-1}) \text{ as } n \longrightarrow \infty, \tag{1.47}$$

where $V_g = Var_g\Big[A'(\delta)u_{\theta_g}\Big]$.

Suppose $G = F_{\theta_0}$ for some θ_0. Then θ_g is Fisher consistent, and the asymptotic variance becomes $I^{-1}(\theta_0)$. Thus, under appropriate conditions, all the minimum disparity estimators attain the minimum possible variance– the inverse of Fisher information, and hence they are first-order efficient at the model. While the discussion on disparities helps to set up the perspective, we will not use them directly in this thesis.

1.7 Minimum Density Power Divergence Estimator: Independent and Identically Distributed Data

A continuous model poses an immediate challenge to construct a distance (divergence) between the empirical density g_n and the model f_θ, as the former is discrete it becomes incompatible with the model. One possible approach to make things work in this setup is to construct a continuous density estimate from the data using a non-parametric kernel density function (Beran, 1977a; 1977b). This induces an element of subjective bias into the conclusion because the problem of kernel selection is still open and yet to be fully settled. To overcome this, Basu and Lindsay (1994) propose that both f_θ and g_n be smoothed with the same kernel along the way. This partly solves this issue as it significantly reduces an estimator's dependency on a kernel function, but still, the question regarding the choice of kernels persists.

Being motivated by this, Basu et al. (1998) propose the density power divergence that allows us to avoid the use of kernel density estimation even for a continuous model. Not only that, it also produces highly robust estimators, albeit at the expense of marginal loss in asymptotic efficiency.

Firstly, we give a short account of the minimum L_2 distance estimator which prepares the ground as a precursor to developing the minimum density power divergence estimator. The squared L_2 distance between g and f_θ is defined as

$$L_2(g, f_\theta) = \int (g - f_\theta)^2. \tag{1.48}$$

The integration over common support is understood to be taken with respect to an appropriate dominating measure. The minimum L_2 distance functional $T_1(G)$ minimizes

$L_2(g, f_\theta)$ over $\theta \in \Theta$; this is equivalent to minimizing

$$\int f_\theta^2 - 2\int f_\theta dG. \tag{1.49}$$

The minimum L_2 distance estimator similarly minimizes

$$\int f^2 - 2\int f_\theta dG_n = \int f_\theta^2 - 2n^{-1}\sum_{i=1}^{n} f_\theta(X_i). \tag{1.50}$$

Note that the construction of the empirical divergence in (1.50) does not require any kernel smoothing even for continuous models. Under differentiability, the minimum L_2 distance estimator may be obtained as a solution to the estimating equation

$$n^{-1}\sum_{i=1}^{n} u_\theta(X_i) f_\theta(X_i) - \int u_\theta f_\theta^2 = 0, \tag{1.51}$$

where u_θ is the likelihood-score function. The estimating equation in (1.51) turns out unbiased under the model. When the model belongs to the location family, i.e., $f_\theta(x) = f(x-\theta)$ for all x satisfying $\int f = 1$, the expression in (1.51) further simplifies to

$$\sum_{i=1}^{n} u_\theta(X_i) f_\theta(X_i) = 0. \tag{1.52}$$

Contrasting it with the estimating equation in (1.26) of MLE, we immediately recognize that (1.52) is a weighted likelihood estimating equation with the weight being chosen as the model density itself. Such a weight function automatically downweights the effect of those observations which are unlikely to occur under the model f_θ. It is not hard to see that $\sup_x ||u_\theta(x) f_\theta(x)||$ stays bounded unlike $\sup_x ||u_\theta(x)||$ for most common parametric families (e.g., the exponential family). This means that the effect of outlying observations is limited in the estimating equation as in (1.52). This, in turn, yields the highly robust minimum L_2 distance estimator. However, this comes at the expense of a

significant loss in its asymptotic efficiency.

Taking a cue from the earlier example, Basu et al. (1998) propose the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i) f_\theta^\alpha(X_i) - \int u_\theta(x) f_\theta^{1+\alpha}(x) dx = 0, \alpha \in [0,1]. \tag{1.53}$$

Note that the tuning parameter α smoothly connects the MLE (first-order efficient but non-robust) with the minimum L_2 distance estimator (very robust but relatively inefficient). Suppose $\hat{\theta}_\alpha$ solves (1.53). Then it is not hard to see that the robustness of $\hat{\theta}_\alpha$ increases with α, and at the same time its asymptotic efficiency decreases with α. Therefore a nice trade-off between the asymptotic efficiency and robustness is achieved through appropriately choosing the tuning parameter α. The associated divergence called the density power divergence (Basu et al., 1998) is given by

$$d_\alpha(g, f_\theta) = \begin{cases} \int \left\{ f_\theta^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) f_\theta^\alpha g + \frac{1}{\alpha} g^{1+\alpha} \right\} & \text{for } \alpha > 0, \\ \int g \log(g/f_\theta) & \text{for } \alpha = 0. \end{cases} \tag{1.54}$$

Thus $\hat{\theta}_\alpha$ is called the minimum density power divergence estimator (MDPDE). Similarly, the best-fitting parameter is usually denoted by $\theta_\alpha = T_\alpha(G)$. Since $\hat{\theta}_\alpha$ is also an M-estimator its asymptotic properties can be easily derived from the general theory of M-estimation. Though not directly related to the power-divergence family, Patra et al. (2013) find an indirect connection between the power divergence and density power divergence families. Note that the density power divergence is a special family in the class of Bregman divergences (Bregman, 1967) which will be introduced in Chapter 6.

1.7.1 Asymptotic Properties and Influence Function

Next, we briefly describe the asymptotic properties of $\hat{\theta}_\alpha$. Define

$$J_\alpha(\theta) = \int u_\theta u_\theta^T f_\theta^{1+\alpha} + \int \{i_\theta - \alpha u_\theta u_\theta^T\}(g - f_\theta) f_\theta^\alpha, \tag{1.55}$$

$$K_\alpha(\theta) = \int u_\theta u_\theta^T f^{2\alpha} g - \xi_\alpha(\theta)\xi_\alpha(\theta)^T, \text{ and } \xi_\alpha(\theta) = \int u_\theta f_\theta^\alpha g \tag{1.56}$$

where $i_\theta = -\nabla u_\theta$. Basu et al. (1998) prove the following important result.

Theorem 1.2. *Suppose the Assumptions (D1) - (D5) (Basu et al., 1998) are true. Then there exists a consistent sequence of roots $\hat{\theta}_\alpha$ to the estimating equation as in (1.53) such that $\hat{\theta}_\alpha \xrightarrow{\mathbb{P}} \theta_\alpha$ and $n^{1/2}(\hat{\theta}_\alpha - \theta_\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}(0, J^{-1}KJ^{-1})$ as $n \to \infty$, where $J = J_\alpha(\theta_\alpha)$ and $K = K_\alpha(\theta_\alpha)$.*

A direct application of the M-estimation theory or an implicit differentiation of the estimating equation– $\nabla d_\alpha(g, f_\theta) = 0$, gives the influence function as

$$\mathcal{IF}(T_\alpha, G, y) = T'_\alpha(y) = J^{-1}\Big\{u_{\theta_\alpha}(y) f_{\theta_\alpha}^\alpha(y) - \xi_\alpha(\theta_\alpha)\Big\}. \tag{1.57}$$

Suppose $||\xi_\alpha(\theta_\alpha)||$ and the elements of J are bounded. Then the influence function stays bounded as long as $\sup_y ||u_{\theta_\alpha}(y) f_{\theta_\alpha}^\alpha(y)||$ is bounded. This happens for most commonly used parametric models when $\alpha > 0$. In particular, when $g = f_{\theta_0}$ the quantities J, K and ξ simplify to

$$J = \int u_{\theta_0} u_{\theta_0}^T f_{\theta_0}^{1+\alpha}, K = \int u_{\theta_0} u_{\theta_0}^T f_{\theta_0}^{1+2\alpha} - \xi\xi^T \text{ and } \xi = \int u_{\theta_0} f_{\theta_0}^{1+\alpha}. \tag{1.58}$$

In this context, the influence function of the MDPDE differs from that of the minimum disparity estimator (whose IF is the same as MLE at the true model $g = f_{\theta_0}$). In addition,

when $\alpha = 0$ one gets $J = K = I(\theta_0)$, and consequently

$$\mathcal{IF}(T_\alpha, G, y) = I^{-1}(\theta_0) u_{\theta_0}(y). \tag{1.59}$$

Basu et al. (2011) plot the IF of the mean functional when $g \equiv \mathcal{N}(\theta, 1)$. They observe that all the curves are bounded except when $\alpha = 0$. From the theory of M-estimation, we know that the asymptotic variance $J^{-1}KJ^{-1}$ is obtained by the variance of the corresponding influence function. Reflecting on that, a consistent estimate of $J^{-1}KJ^{-1}$ may be given by

$$\hat{J}_\alpha^{-1}(\hat{\theta}_\alpha)\left((n-1)^{-1}\sum_{i=1}^{n} R_i R_i^T\right)\hat{J}_\alpha^{-1}(\hat{\theta}_\alpha) \text{ where } R_i = u_{\hat{\theta}_\alpha}(X_i) f_{\hat{\theta}_\alpha}^{\alpha}(X_i) - \hat{\xi}_\alpha(\hat{\theta}_\alpha). \tag{1.60}$$

$\hat{\xi}_\alpha(\hat{\theta}_\alpha)$ and $\hat{J}_\alpha(\hat{\theta}_\alpha)$ are the empirical versions of $J_\alpha(\theta_\alpha)$ and $\xi_\alpha(\theta_\alpha)$.

1.8 Minimum Density Power Divergence Estimator: Independent but Non-homogeneous Data

Suppose that the random observations are independent but not necessarily identically distributed. Then the density power divergence may be generalized in a variety of different ways. In this case, we follow the approach of Ghosh and Basu (2013). Let X_i be generated by the density g_i which is modelled by $f_{\theta,i}$: $i = 1, \ldots, n$. All the densities are assumed to be defined with respect to a common σ-finite measure. In this setup, Ghosh and Basu (2013) define the DPD as

$$\frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i, f_{\theta,i}), \tag{1.61}$$

where $d_\alpha(g_i, f_{\theta,i})$ is the usual DPD as in (1.54). As before, notice that (1.61) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\int g_i \ln\frac{g_i}{f_{\theta,i}} \text{ at } \alpha = 0, \tag{1.62}$$

which yields the maximum likelihood functional in the independent but not-homogeneous setup. To find the minimum density power divergence estimator, we substitute the true density g_i by its empirical analogue $\hat{g}_i = \delta_{X_i}$ (which is a degenerate density defined earlier) and further minimize (1.61) over $\theta \in \Theta$. Under the differentiability of the model with respect to θ, the minimum density power divergence estimator $\hat{\theta}_\alpha$ solves the estimating equation

$$\sum_{i=1}^{n}\Big\{\int f_{\theta,i}^{1+\alpha}u_{\theta,i} - f_{\theta,i}^{\alpha}(X_i)u_{\theta,i}(X_i)\Big\} = 0 \text{ where } u_{\theta,i} = \frac{\nabla f_{\theta,i}}{f_{\theta,i}}. \tag{1.63}$$

Next, we briefly discuss its asymptotic properties.

1.8.1 Asymptotic Properties

The best-fitting parameter θ_α minimizes the objective function

$$H(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int\Big\{f_{\theta,i}^{1+\alpha} - \Big(1+\frac{1}{\alpha}\Big)f_{\theta,i}^{\alpha}g_i\Big\}. \tag{1.64}$$

Now define

$$\Psi_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left[\int u_{\theta_\alpha,i}u_{\theta_\alpha,i}^{T}f_{\theta_\alpha,i}^{1+\alpha} - \int\Big\{\nabla u_{\theta_\alpha,i} + \alpha\cdot u_{\theta_\alpha,i}u_{\theta_\alpha,i}^{T}\Big\}\Big\{g_i - f_{\theta_\alpha,i}\Big\}f_{\theta_\alpha,i}^{\alpha}\right], \tag{1.65}$$

$$\Omega_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int u_{\theta_\alpha,i}u_{\theta_\alpha,i}^{T}f_{\theta_\alpha,i}^{2\alpha}g_i - \xi_i(\alpha)\xi_i(\alpha)^{T}\right\} \text{ and } \xi_i(\alpha) = \int u_{\theta_\alpha,i}f_{\theta_\alpha,i}^{\alpha}g_i. \tag{1.66}$$

Next, we present the consistency and asymptotic normality results.

Theorem 1.3. *(Ghosh and Basu, 2013) Suppose the Assumptions (A1) - (A7) are true. Then*

there exists a consistent sequence of roots $\hat{\theta}_\alpha$ *such that* $\hat{\theta}_\alpha \xrightarrow{\mathbb{P}} \theta_\alpha$ *and* $\Omega_n^{-1/2}\Psi_n\left\{n^{1/2}(\hat{\theta}_\alpha - \theta_\alpha)\right\} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p)$ *as* $n \to \infty$, *where* $\Omega_n = \Omega_n(\alpha), \Psi_n = \Psi_n(\alpha)$.

Remark 1.1. *Suppose* X_i*s are iid, i.e.,* $g_i \equiv g$ *for* $i = 1, 2, \ldots, n$. *Then we get to choose only one density such as* f_θ *to model the true density* g. *In this case, the Assumptions (A1) - (A7) (Ghosh and Basu, 2013) reduce to Assumptions (D1) - (D5) (Basu et al., 1998). In this case, also it turns out that* $\Psi_n = J$ *and* $\Omega_n = K$ *as in Theorem 1.2.*

1.8.2 Influence Function and Asymptotic Breakdown Point

Let the i-th true density g_i be contaminated with ϵ-proportion at a point t_i as $g_{i,\epsilon} = (1-\epsilon)g_i + \epsilon\delta_{t_i}$, but the other true densities remain unchanged. Under this setup, the influence function of the MDPD functional is obtained as

$$\mathcal{IF}_i(\theta_\alpha, G_1, \cdots, G_n, t_i) = \Psi_n^{-1}(\alpha)\frac{1}{n}\left\{f_{\theta_\alpha,i}(t_i)^\alpha u_{\theta_\alpha,i}(t_i) - \xi_i(\alpha)\right\}. \tag{1.67}$$

Here contamination happens only at the i-th direction. When all the true densities are similarly contaminated one gets the influence function as

$$\mathcal{IF}(\theta_\alpha, G_1, \cdots, G_n, t_1, \cdots, t_n) = \Psi_n^{-1}(\alpha)\frac{1}{n}\sum_{i=1}^{n}\left\{f_{\theta_\alpha,i}(t_i)^\alpha u_{\theta_\alpha,i}(t_i) - \xi_i(\alpha)\right\}. \tag{1.68}$$

Notice that (1.68) is a simple arithmetic mean of (1.67) over $i = 1, \ldots, n$. In both these cases we see that the effect of the observations, which are inconsistent with the models, is shrunk.

Following Hampel et al. (2011), and Ghosh and Basu (2013), the unstandardized and self-standardized gross error sensitivities (GES) corresponding to the influence function

(1.67) is given by

$$GES_i^u(\theta_\alpha, G_1, \cdots, G_n) = \sup_t \|\mathcal{IF}_i(\theta_\alpha, G_1, \cdots, G_n, t)\|, \tag{1.69}$$

$$GES_i^s(\theta_\alpha, G_1, \cdots, G_n) = \sup_t \|(\Psi_n^{-1}\Omega_n\Psi_n^{-1})^{-1/2}\mathcal{IF}_i(\theta_\alpha, G_1, \cdots, G_n, t)^T\|. \tag{1.70}$$

Similarly, the GES may be defined using (1.68). In the context of normal linear regression, Ghosh and Basu (2013) show that influence functions are bounded for all values of the tuning parameter except at $\alpha = 0$. This demonstrates, as in the iid setup, that higher values of α ensure better stability whereas lower values of α produce higher asymptotic efficiencies.

Although the influence function is one of the most commonly used measures of robustness it is a local concept. Therefore it must be complemented by a global measure of robustness which should quantify the maximum proportion of contamination in a data set that a functional may endure before giving erratic values when the sample size is held fixed, but the elements of contaminating sequence are allowed to go as far as we please. The critical contamination proportion, as mentioned before, is called the asymptotic breakdown point of the functional.

The definition of the breakdown point is varied across the literature. Often, the general definitions such as Hampel (1968; 1971) and Huber (2011) may be quite hard to work with in many practical situations. This may require tweaking the definition depending upon a particular setup keeping the underlying philosophy intact. With that, we lay down our working definition. This will be primarily followed with suitable modification as the case may be, throughout this thesis unless we reach to the point when the theory of two-sample non-parametric tests based on the divergence measures is developed in Chapter 6.

Let g_i be contaminated by a sequence of densities $\{k_{i,M}\}$ as $h_{i,\epsilon,M} = (1-\epsilon)g_i + \epsilon k_{i,M}$; $i = 1, \ldots, n$. Then the best-fitting parameter is given by

$$\theta_\alpha^{h_{\epsilon,M}} := \arg\min_{\Theta} = \frac{1}{n}\sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,M}, f_{\theta,i}). \tag{1.71}$$

Following Simpson (1987) and Park and Basu (2004) we know that the minimum density power divergence functional θ_α breaks down at ϵ if $\|\theta_\alpha^{h_{\epsilon,M}} - \theta_\alpha\| \to \infty$ as $M \to \infty$.

Definition 1.6. *The asymptotic breakdown point of θ_α at the model is defined as the least proportion of contamination $\epsilon \in [0,1]$ such that $\|\theta_\alpha - \theta_\alpha^{h_{\epsilon,M}}\| \to \infty$ when $M \to \infty$ at fixed sample size n.*

Remark 1.2. *The working definition of the breakdown point used throughout this thesis in the parametric setup is an asymptotic one. Given that the true density becomes contaminated, a minimum density power divergence (MDPD) functional breaks down when the Euclidean norm represented in Definition 1.6 diverges as M increases to infinity. Though the breakdown point is not computed for the MDPD estimates, the following results have important implications because the MDPD estimates become very close to the MDPD functionals for large sample sizes with very high probability under standard regulatory conditions. In an iid setup, the MDPD functional depends on the true density only. However, in an independent but non-homogeneous setup, the MDPD functionals additionally depend on the sample size, not the actual sample points. The following results that use this definition lead to breakdown points that do not depend on the sample size in the sense the breakdown point considered in this thesis may be qualified as an asymptotic measure.*

Now consider the following location-scale family

$$f_{\theta,i}(y) = \frac{1}{\sigma} f\left(\frac{y - l_i(\mu)}{\sigma}\right) \text{ for } (\mu,\sigma) \in \Theta, \tag{1.72}$$

where $\mu \mapsto l_i(\mu)$ is one-to-one; $i = 1, \ldots, n$. The next result gives the asymptotic breakdown point of the location functional at the above location-scale family.

Theorem 1.4. *(Ghosh and Basu, 2013) Suppose the Assumptions (BP1)–(BP3) are true, and $\alpha > 0$. Then the asymptotic breakdown point ϵ^* of the MDPD functional μ_α is at least $\frac{1}{2}$ at the true model belonging to the location-scale family (1.72) with a fixed scale parameter.*

Theorem 1.4 establishes that the minimum density power divergence yields location estimators with very high breakdown points for all $\alpha > 0$ when the true density belongs to the location-scale family with a fixed scale parameter (1.72). See Ghosh and Basu (2013) for more related discussions.

1.8.3 Tuning Parameter Selection

From earlier discussions, it is clear that the choice of tuning parameter plays an important role in the performance of minimum density power divergence estimators. This problem is difficult because an experimenter does not know a priori the amount of contamination in a data set. However, there have been several attempts to find optimum tuning parameters based on different data-driven strategies. The key ideas revolve around minimizing some empirical estimate of the summed asymptotic variance or summed asymptotic mean squared error over the tuning parameter α. Such a tuning parameter selection strategy helps practitioners to apply this method meaningfully to real-life data. Hong and Kim (2001) suggest using α_{opt}^{HK} that minimizes the sum of asymptotic variances across all components of $\hat{\theta}_\alpha$, i.e.,

$$\alpha_{opt}^{HK} = \arg\min_{\alpha} \mathbb{E}(\hat{\theta}_\alpha - \theta_\alpha)^T(\hat{\theta}_\alpha - \theta_\alpha) \tag{1.73}$$

$$\approx \arg\min_{\alpha} \frac{1}{n} tr(J_\alpha^{-1}(\hat{\theta}_\alpha) K_\alpha(\hat{\theta}_\alpha) J_\alpha^{-1}(\hat{\theta}_\alpha)) \tag{1.74}$$

where $tr(M)$ denotes the trace of a matrix M. Warwick and Jones (2005), on the other hand, suggest the minimization of an asymptotic approximation of the mean squared error of $\hat{\theta}_\alpha$. Their method gives α_{opt}^{WJ} which is obtained as

$$\alpha_{opt}^{WJ} = \arg\min_{\alpha} \mathbb{E}(\hat{\theta}_\alpha - \theta_P)^T(\hat{\theta}_\alpha - \theta_P) \tag{1.75}$$

$$\approx \arg\min_{\alpha} \left\{ (\hat{\theta}_\alpha - \theta_P)^T(\hat{\theta}_\alpha - \theta_P) + \frac{1}{n} tr(J_\alpha^{-1}(\hat{\theta}_\alpha) K_\alpha(\hat{\theta}_\alpha) J_\alpha^{-1}(\hat{\theta}_\alpha)) \right\}, \tag{1.76}$$

using some appropriate pilot θ_P. The pilot θ_P, deemed as a proxy for the true value, needs to be estimated from the data. If the bias component is zero, both these methods lead to the same optimal tuning parameter. Often it is easier, at least in principle, to estimate the variance component, if not numerically. To carry out the second approach we need an estimator for the pilot θ_P. Warwick and Jones (2005) try out several MDPDEs as pilot values and recommend using $\hat{\theta}_P = \hat{\theta}_1$. Ghosh and Basu (2015) have advocated the use of $\hat{\theta}_{0.5}$. Although both approaches provide reasonable solutions, the solutions are dependent on the pilot estimate in each case. Basak et al. (2021) have suggested an iterative algorithm to modify the method of Warwick and Jones to remove, or at least reduce, its dependency on the pilot estimator.

1.9 Aim and Layout of this Thesis

The main objective of this thesis is to present some applications of density-based divergences to the problems related to ordinal and mixed data. Roughly speaking, the first half of this thesis uses the density power divergence. In the remaining part, we use a more general family of distance measures, viz. the extended Bregman divergence. This section provides a brief sketch of the thesis as a whole. The rest of this thesis is organized as follows.

The present chapter (Chapter 1) sets the tone for the rest of this thesis. Since no parametric model perfectly matches the data-generating scheme, the occurrence of outlying observations inconsistent with the model is never completely unexpected. With this background, we discuss the necessity and use of robust methodologies in statistics with particular emphasis on minimum disparity estimators and minimum density power divergence estimators. Statistical inference based on the minimum disparity approach is very useful to accomplish this objective. In this context, we briefly review the existing approaches and discuss some of their important properties. All the minimum disparity estimators are first-order efficient at the true model, and some of them are highly robust. In this chapter, we also introduce the density power divergence measure (Basu et al., 1998). Also its extension to the case of independent but not identically distributed data (Ghosh and Basu, 2013) is discussed.

In Chapter 2 we use the density power divergence to estimate the parameters of an ordinal response model in a robust and efficient way. The roles of different link functions in ordinal response models are analyzed through the lens of the density power divergence. Asymptotic properties of the minimum density power divergence estimator are discussed theoretically. Its robustness is investigated through the influence function analysis. Analytically, we have shown that the proposed estimate has a very high asymptotic breakdown point against data contamination. Numerically it is further demonstrated that the proposed method yields slope estimates whose implosive breakdown point is also very high, unlike the MLE. The performance of these minimum density power divergence estimators in finite samples across different links is investigated through extensive numerical experiments either at the model or when data contamination occurs. It outperforms the maximum likelihood estimators, producing more stable results when robustness is a concern. Moreover, our estimators are very competitive with the other robust alternatives.

In Chapter 3 we use the density power divergence to estimate the polychoric correlation. A Wald-type statistic to test a simple null hypothesis is also proposed. The theoretical properties of the estimators and the test statistics are derived and substantiated with extensive simulation studies. A couple of real-life data examples are supplemented at the end of this chapter.

In continuation to the preceding chapter, a two-step approach of the DPD is adopted in Chapter 4. The performance of the estimators and test statistics remain almost equivalent for pure models, but their performances deteriorate slightly at data contamination in comparison to the usual DPD. However, this two-step adaptation significantly reduces the computational burden.

A simple modification is proposed over the two-step DPD in Chapter 5. At a higher level of data contamination, it works better than the one- and two-step MDPDE. Necessary theoretical details are provided along the way.

In Chapter 6 we give a generalized definition of mutual information (MI) using the class of extended Bregman divergences (Basak and Basu, 2022). Using that, we develop a class of two-sample non-parametric test for unstructured comparison between two absolutely continuous distributions. Several theoretical properties of these tests are investigated from the point of robust inference. This chapter includes only the theoretical details. In-depth numerical studies are presented separately for two of the special families, namely the generalized S-Bregman divergence and exponential-polynomial divergence, respectively, in Chapter 7 and Chapter 8. In comparison to the power divergence, we find that there exist tests that are both consistent and robust outside the power divergence family.

Finally, we end this thesis with some discussion about the future directions of our research in Chapter 9.

This page is intentionally left blank.

Chapter 2

Robust Estimation in Ordinal Response Models

2.1 Introduction

In recent years, the analysis of ordinal response data has become a popular topic in mainstream research. Such data arise naturally in many areas of scientific studies, for instance in psychology, sociology, economics, medicine, political science, and in several other disciplines where the final response of a subject belongs to a finite number of ordered categories based on the values of several explanatory variables in a way described later in this chapter. One such example may be the qualitative customer review of a particular vehicle where its price, mileage, carbon emission properties, etc., are to be considered to arrive at a qualitative response on an ordinal scale. While these ratings summarize many important explanatory variables and are primarily useful to a new customer, these customers' feedbacks often turn out to be equally important to the manufacturer itself, as the latter might want to dig deep into the statistical relationship between the ordinal response and its covariates to improve their product, or for post-manufacturing surveys to fix things.

A pioneering work in this field is done by McCullagh (1980). He advocates using an underlying continuous latent variable that drives ordinal responses based on some unknown cut-offs. This method has become popular as it enables us to view the ordinal response model within the purview of a unified framework of the generalized linear model (GLM); see, e.g., McCullagh and Nelder (2019) and Nelder et al. (1972). Moustaki (2000) uses the maximum likelihood (ML) method to fit a multi-dimensional latent variable model to a set of observed ordinal variables and also discusses the related goodness-of-fit problems. See Moustaki (2003) for further discussions. Piccolo (2003) and Iannario et al. (2016) suggest a different approach which uses the response variable as a combination of a discrete mixture of uniform and a shifted binomial (CUB) random variables.

Although the area of robust statistics has a rich and developed body of literature, applications in the direction of ordinal response data are rather limited. An early reference is Hampel (1968), where, in addition to developing a classical infinitesimal approach to robustness, some pointers about robustness in the case of binomial model fitting are discussed. Victoria-Feser and Ronchetti (1997) develop robust estimators for grouped data. Ruckstuhl and Welsh (2001) obtain estimators while fitting a robust binomial model to a data set. Moustaki and Victoria-Feser (2004; 2006) develop bounded-bias and bounded-influence robust estimators for the generalized linear latent (GLL) variable. Croux et al. (2002), and Müller and Neykov (2003) point out the robustness issues in the maximum likelihood estimators for the logistic regression model. Croux et al. (2013) propose a weighted maximum likelihood (WML) estimation method for the logit link function. Iannario et al. (2016) deal with the robustness issues for the class of CUB models. More recently, Iannario et al. (2017) suggest using a weighted likelihood score function as an estimating equation to obtain robust estimates. These where weights vary depending on the choice of link functions. Unlike the approaches

of Croux et al. (2013), who propose to use weights as functions of robust Mahalanobis type distances, Iannario et al. (2017) consider the Huber's weight functions that combine both the generalized residual and robust Mahalanobis distance or the normalized MAD as appropriate for different link functions. The primary objective is to control the influential observations with respect to the parametric model. Recently, Scalera et al. (2021) analyze the role of different link functions in robustness in this setup.

In this chapter, we propose to implement the density power divergence (DPD) measure, originally proposed by Basu et al. (1998), to obtain robust and (asymptotically highly) efficient estimates in ordinal response data under the same setup as Iannario et al. (2017). The independent but non-homogeneous version of DPD-based inference (Ghosh and Basu, 2013) is best suited for this application.

The rest of this chapter is organized as follows.

(a) The parametric model is discussed in Section 2.2.

(b) Estimating equations are presented in Section 2.3.

(c) To study the role of different link functions in estimating the slope parameter, we plot a DPD-version of the generalized residuals in Section 2.4. As it turns out, the DPD-version of the generalized residuals stays bounded even for the commonly used links which produce unbounded generalized residuals for the MLE. This gives a clear insight as to why the DPD should produce robust slope estimates.

(d) Asymptotic properties of the proposed minimum density power divergence estimator (MDPDE) for the parameters related to ordinal response models, are discussed in Section 2.5.

(e) The robustness properties of the minimum density power divergence functional are investigated through the influence function analysis in Subsection 2.6.1. As expected, we find the effect of tiny contamination in data to be limited, as compared to the MLE, whenever the tuning parameter α is strictly positive.

(f) In Subsection 2.6.2, we have shown that under suitable assumptions, the asymptotic breakdown point of the MDPD functional is at least $\frac{1}{2}$ at the model. This is way above those related to the maximum likelihood functional.

(g) In Subsection 2.6.3, we have empirically shown that the implosive breakdown point of the MDPDE of the regression parameter is very high. This means that, unlike the MLE, we can still obtain a stable MDPDE of the slope parameter even when the sample contains a high proportion of outlying observations.

(h) In continuation of the earlier point, these estimates are also used to find the prediction misclassification rate. When $\alpha > 0$. The misclassification rate incurred by the MDPDE becomes much lower than the MLE.

(i) Extensive simulation studies are presented in Section 2.7. Particularly in Subsection 2.7.1, it is demonstrated that the performances of the proposed estimators are almost as good as the MLE at relatively smaller values of α when data are truly generated by the model. Also, we show in Subsection 2.7.2 that the proposed estimator outperforms the MLE in the presence of outlying observations in data at higher values of the tuning parameter α. Our estimator performs better than the weighted likelihood estimator proposed by Croux et al. (2013), and it is very competitive to the M-estimator proposed by Iannario et al. (2017). In Subsection 2.7.3 we have empirically shown that our method is computationally less expensive than Iannario et al. (2017).

(j) To apply this method in real-life data examples, we make use of a tuning parameter selection algorithm, as proposed by Warwick and Jones (2005). The performance of this algorithm is validated through simulation studies in Section 2.8.

(k) Finally, we analyze a real data example in Section 2.9. In that, we choose the optimal tuning parameter using the above algorithm. The prediction from the resulting estimator achieves no lesser (higher in some cases) accuracy than the MLE.

2.2 Parametric Model and the Maximum Likelihood Estimation

Consider a random sample $\{(x_i, Y_i) : i = 1, 2, \ldots, n\}$ of size n. The i-th explanatory vector, denoted by $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, is assumed to be non-stochastic in $\mathbb{R}^p$. Here Y_i is the realization of the response variable Y when conditioned on x_i. Further, Y is supported on a finite set $\chi = \{1, 2, \ldots, m\}$. Following McCullagh (1980) we presume that there exists a continuous latent random variable Y^* such that it is related to Y as

$$Y = j \iff \gamma_{j-1} < Y^* \leq \gamma_j \text{ for } j \in \chi, \tag{2.1}$$

where $-\infty = \gamma_0 < \gamma_1 < \gamma_2 < \cdots < \gamma_{m-1} < \gamma_m = \infty$ are the unknown cut-off points (thresholds) in the continuous support of the latent variable. Moreover, Y_i depends on the explanatory variables x_i because

$$Y_i^* = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i = x_i^T\beta + e_i \text{ for all } i = 1, 2, \ldots, n. \tag{2.2}$$

Here $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector of regression coefficients in the latent linear (LL) regression model with e_i being a random error term. The e_is are assumed to be identically and independently distributed according to a known probability distribution function F. The inverse of F is called the link function. Commonly used links are– probit, logit, complementary log-log (or, simply the log-log) and Cauchy links. We assume that F admits a probability density function f, and further denote the thresholds by $\gamma = (\gamma_1, \ldots, \gamma_{m-1})$. Using (2.2) we find that

$$p_{\theta,i}(j) = Pr(Y = j|x_i) = F(\gamma_j - x_i^T\beta) - F(\gamma_{j-1} - x_i^T\beta) \text{ where } \theta = (\gamma, \beta), \tag{2.3}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. Notice that $p_{\theta,i}(1) = F(\gamma_1 - x_i^T\beta)$ and $p_{\theta,i}(m) = 1 - F(\gamma_{m-1} - x_i^T\beta)$. Here the parameter space is denoted by $\Theta \subseteq \mathbb{R}^{m+p-1}$. Later on, we shall interchangeably use the term "slope parameters" for the "regression parameters".

Now we wish to find an estimate of θ. A traditional way of doing that is to find its maximum likelihood estimate $\hat{\theta}_{ML}$ which maximizes the log-likelihood function

$$\sum_{i=1}^{n} \ell(\theta; x_i, Y_i) \text{ where } \ell(\theta; x_i, Y_i) = \sum_{j=1}^{m} \delta_i(j) \ln p_{\theta,i}(j). \tag{2.4}$$

Note that $\delta_i(j) = \mathbb{1}(Y = j|x_i)$ where $\mathbb{1}(\cdot|x_i)$ symbolizes the indicator of the set $\{Y = j\}$ given at x_i, and $\ln(\cdot)$ represents the natural logarithm. The log-likelihood function for the entire sample turns out to be the sum of individual log-likelihood functions $\ell(\theta; x_i, Y_i)$ evaluated at each data point (x_i, Y_i), $i = 1, 2, \ldots, n$. Under appropriate regularity assumptions $\hat{\theta}_{ML}$ consistently estimates true θ. Moreover, when the model truly generates data, the asymptotic variance of $\hat{\theta}_{ML}$ attains the Rao-Cramér lower bound. Though the latter is not a finite sample property, it is what makes MLE "optimum" (or asymptotically most efficient) in the class of consistent and uniformly asymptotically normal (CUAN) estimators. However, in practice, we hardly come across a data set

that truly follows an assumed model. Often a model is deemed to be a good fit to a data set if the majority of the data points follow that model, leaving out only a small proportion (maybe 5% - 10%) of observations that are inconsistent with the model. Observations, that defy the assumed probability distribution, are deemed outliers with respect to that model. The presence of even a single outlying observation can strongly affect the reliability of the MLE. This haunts the MLE whenever robustness is a concern. Often robustness comes at the cost of significant loss in asymptotic efficiency. To overcome this, it is therefore required to resort to alternative methods that would make use of the same model but yield estimators trading off between these extreme situations– asymptotic efficiency and robustness.

In the vast literature of robust statistics, methods often focus on two primary aspects to eliminate or limit the influence of outlying observations in the estimation process. The most intuitive way is to multiply the log-likelihood by a suitable weight function. Many different types of weights may be suggested along the way. Croux et al. (2013) propose one such weight function that uses the robust Mahalanobis distance in the space of explanatory variables. In the second approach, a weighted version of the likelihood score is set to zero, and solved for the unknown parameter. These are expected to lead to robust estimates. Iannario et al. (2017) take this second approach. Methods that deal with parameter estimation in ordinal response models are primarily limited to these two methods. In this chapter, we use the density power divergence (Basu et al., 1998) to find robust parameter estimates with high asymptotic efficiency.

2.3 Estimating Equations

Notice that the model $p_{\theta,i}$ for the i-th ordinal response Y_i depends very much on the i-th data point itself, yet the common set of unknown parameters θ remains the same across

all these different models. Assume that G_i be the true probability distribution function that generates Y_i; $i = 1, 2, \ldots, n$. These G_is should be independent but possibly different. It is further assumed that G_is admit probability density functions g_is with respect to a common dominating measure for $i = 1, 2, \ldots, n$. At each i, the true density g_i is modelled by $p_{\theta,i}$. Construct $d_\alpha(g_i, p_{\theta,i})$ as the DPD between the g_i and $p_{\theta,i}$; $i = 1, 2, \ldots, n$. Then the overall divergence (Ghosh and Basu, 2013) is given by

$$\frac{1}{n}\sum_{i=1}^{n} d_\alpha(g_i, p_{\theta,i}). \tag{2.5}$$

The minimum density power divergence functional θ_α, that minimizes (2.5), also depends on the true distributions $G_1, \ldots, G_n$. Essentially, θ_α minimizes the following objective function

$$H(\theta) = \frac{1}{n}\sum_{i=1}^{n} H^{(i)}(\theta), \tag{2.6}$$

where $H^{(i)}(\theta)$ is obtained from $d_\alpha(g_i, p_{\theta,i})$ excluding its last term which is independent of θ. To find out the minimum density power divergence estimate $\hat{\theta}_\alpha$, we substitute $\delta_i(j)$ for $g_i(j)$ in the above expressions as the former is an empirical estimate of the i-th true density. The empirical version of $H^{(i)}$ is given by

$$V_i(x_i, Y_i, \theta) = \begin{cases} \sum_{j=1}^{m} p_{\theta,i}(j)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) p_{\theta,i}(Y_i)^\alpha & \text{when } \alpha > 0, \\ -\ln p_{\theta,i}(Y_i) & \text{when } \alpha = 0. \end{cases} \tag{2.7}$$

Consequently, $\hat{\theta}_\alpha$ minimizes

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} V_i(x_i, Y_i, \theta) \tag{2.8}$$

over the parameter space Θ. Suppose the error-distribution F as in (2.3) is differentiable, so that the minimum density power divergence estimator $\hat{\theta}_\alpha$ can be obtained by solving the estimating equation

$$\nabla H_n(\theta) = \frac{(1+\alpha)}{n} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} p_{\theta,i}(j)^{1+\alpha} u_{\theta,i}(j) - p_{\theta,i}(Y_i)^{\alpha} u_{\theta,i}(Y_i) \right\} = 0, \tag{2.9}$$

where $u_{\theta,i}(j) = \frac{\nabla p_{\theta,i}(j)}{p_{\theta,i}(j)}$ is the likelihood-score function at a point $j \in \chi$. Notice that $\nabla p_{\theta,i}(j)$ is a vector-valued function which is given by

$$\nabla p_{\theta,i}(j) = \left(\frac{\partial}{\partial \gamma^T} p_{\theta,i}(j), \frac{\partial}{\partial \beta^T} p_{\theta,i}(j) \right)^T \in \mathbb{R}^{m+p-1}. \tag{2.10}$$

A simple calculation shows that

$$\frac{\partial}{\partial \gamma_s} p_{\theta,i}(j) = \begin{cases} f(\gamma_s - x_i^T \beta) & \text{when } j = s, \\ -f(\gamma_s - x_i^T \beta) & \text{when } j = s+1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.11}$$

and

$$\frac{\partial}{\partial \beta_k} p_{\theta,i}(j) = \left\{ f(\gamma_{j-1} - x_i^T \beta) - f(\gamma_j - x_i^T \beta) \right\} x_{ik} \tag{2.12}$$

for $s = 1, 2, \ldots, (m-1)$ and $k = 1, \ldots, p$.

Both (2.11) and (2.12) together imply that

$$\frac{1}{1+\alpha}\cdot\frac{\partial}{\partial\gamma_s}V_i(x_i,Y_i,\theta)=\left\{p_{\theta,i}(s)^{\alpha}f(\gamma_s-x_i^T\beta)-p_{\theta,i}(Y_i)^{\alpha-1}f(\gamma_{Y_i}-x_i^T\beta)\mathbb{1}(Y_i=s|x_i)\right\}$$
$$-\left\{p_{\theta,i}(s+1)^{\alpha}f(\gamma_s-x_i^T\beta)-p_{\theta,i}(Y_i)^{\alpha-1}f(\gamma_{Y_i-1}-x_i^T\beta)\mathbb{1}(Y_i=s+1|x_i)\right\}, \tag{2.13}$$

$$\frac{1}{1+\alpha}\cdot\frac{\partial}{\partial\beta_k}V_i(x_i,Y_i,\theta)=x_{ik}\Bigg[\sum_{j=1}^{m}p_{\theta,i}(j)^{\alpha}\left\{f(\gamma_{j-1}-x_i^T\beta)-f(\gamma_j-x_i^T\beta)\right\}$$
$$-p_{\theta,i}(Y_i)^{\alpha-1}\left\{f(\gamma_{Y_i-1}-x_i^T\beta)-f(\gamma_{Y_i}-x_i^T\beta)\right\}\Bigg] \tag{2.14}$$

for $s=1,2,\ldots,(m-1)$ and $k=1,2,\ldots,p$. The estimating equation in (2.9) may be further simplified using (2.13) and (2.14). Observe that the left-hand side of (2.9) is unbiased when the true densities belong to the model families, i.e., $g_i(j)=p_{\theta_*,i}(j)$ for all $i=1,\ldots,n$ and $j=1,\ldots,m$. In that case, the minimum density power divergence functional is Fisher consistent, i.e., $\theta_\alpha=\theta_*$ for true θ_*.

Another justification for adapting the general theory of the non-homogeneous DPD over the usual one in this context is the following. Assume that (X,Y) is jointly distributed according to a probability distribution that involves all the parameters of our interest. Then a single DPD can still be constructed between the data and a model using the original formulation of Basu et al. (1998). Consequently, all the parameters may be estimated by minimizing the divergence albeit with computational complexity that may arise due to modelling both the covariates and response variables in a higher dimension. This may be completely avoided or at least reduced to a great extent if we take the conditioning approach on the explanatory variables keeping the usual flavour of regression analysis as widely used in most of its applications. This, in a way, gives a reason for using the non-homogeneous version of the DPD in the context of the present

situation. Not only that, the loss of efficiency incurred by the ordinary MDPDE which usually occurs in higher dimensions at a fixed α, may be completely circumvented in this case.

2.4 DPD-version of the Generalized Residual

Next, we discuss the role of generalized residual (Iannario et al., 2017). We know that $-\frac{1}{1+\alpha}\sum_{i=1}^{n} V_i(x_i, Y_i, \theta)$ is akin to the log-likelihood function. It is called the β-likelihood function (cf. Fujisawa and Eguchi, 2006). To find the DPD-version of the generalized residual, we express

$$-\frac{1}{1+\alpha}\cdot\sum_{i=1}^{n}\frac{\partial}{\partial\beta_k}V_i(Y_i, x_i, \theta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathcal{E}_{ij}(\theta,\alpha)x_{ik} \text{ for all } k, \tag{2.15}$$

where $\mathcal{E}_{ij}(\theta,\alpha) = \Big[p_{\theta,i}(j) - \delta_i(j)\Big]e_{ij}(\theta)p_{\theta,i}(j)^{\alpha}$ and $e_{ij}(\theta)$ being the generalized residual as in (6) of Iannario et al. (2017). Here $\mathcal{E}_{ij}(\theta,\alpha)$ plays the same role in the estimation of $\hat{\beta}_\alpha$ as $\Big[\delta_i(j)e_{ij}(\theta)\Big]$ does for $\hat{\beta}_{ML}$. Excluding the term $[p_{\theta,i}(j) - \delta_i(j)]$, which is anyway bounded by 2, we may consider $\Big[e_{ij}(\theta)p_{\theta,i}(j)^{\alpha}\Big]$ as a simple analogue for *generalized residual* in the context of using the DPD. To study its behaviour, denote

$$B_j(t) = A_j(j)\Big[F(\gamma_j - t) - F(\gamma_{j-1} - t)\Big]^{\alpha} \text{ for } j = 1, 2, \ldots, m \tag{2.16}$$

where $t = x_i^T\beta$, and $A_j(t)$ is defined in (9) in Iannario et al. (2017). As in Figure 1-Figure 4 of Iannario et al. (2017), we plot $B_j(t)$ in Figure 2.1- Figure 2.4 in a panel for different values of the tuning parameter across different link functions. In all these graphs, we find that when α increases from 0 to 1, the magnitude of the DPD-version of the generalized residual is significantly dampened. This is, in fact, true for most of the distributions belonging to the exponential families.

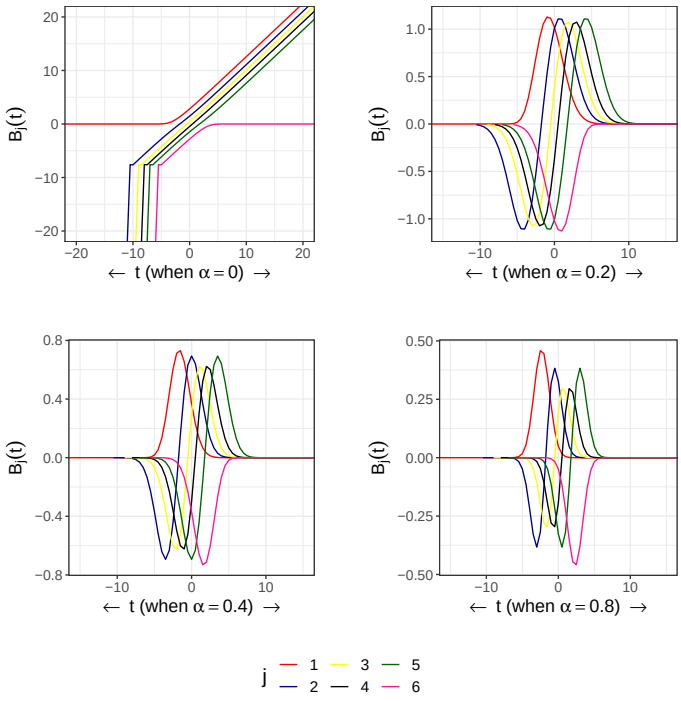

FIGURE 2.1: Generalized residuals for the probit link with $\gamma = (-2.5, -1, 0, 1, 2.5)^T$.

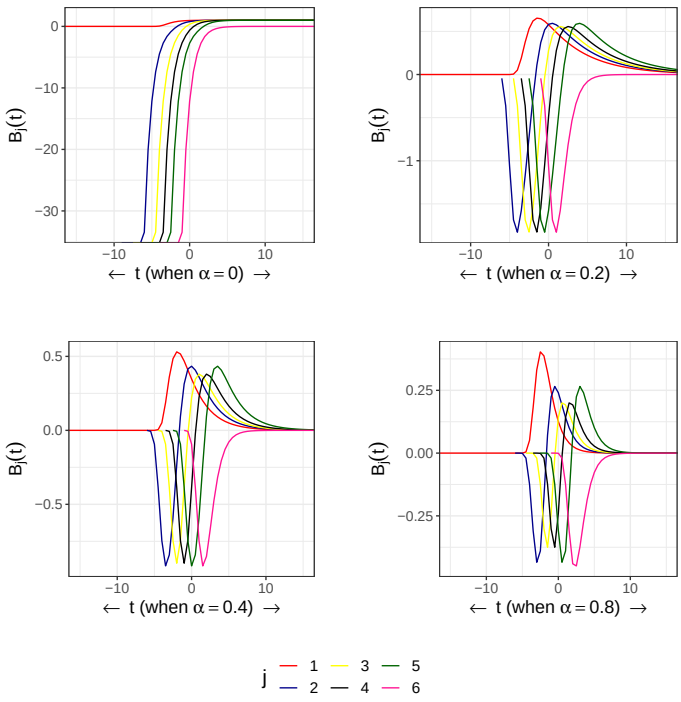

FIGURE 2.2: Generalized residuals for the log-log link with $\gamma = (-2.5, -1, 0, 1, 2.5)^T$.

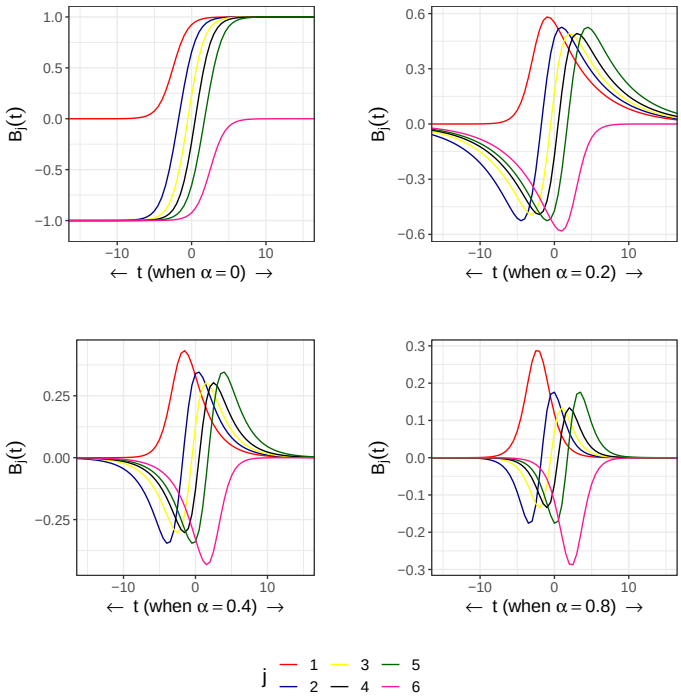

FIGURE 2.3: Generalized residuals for the logit link with $\gamma = (-2.5, -1, 0, 1, 2.5)^T$.

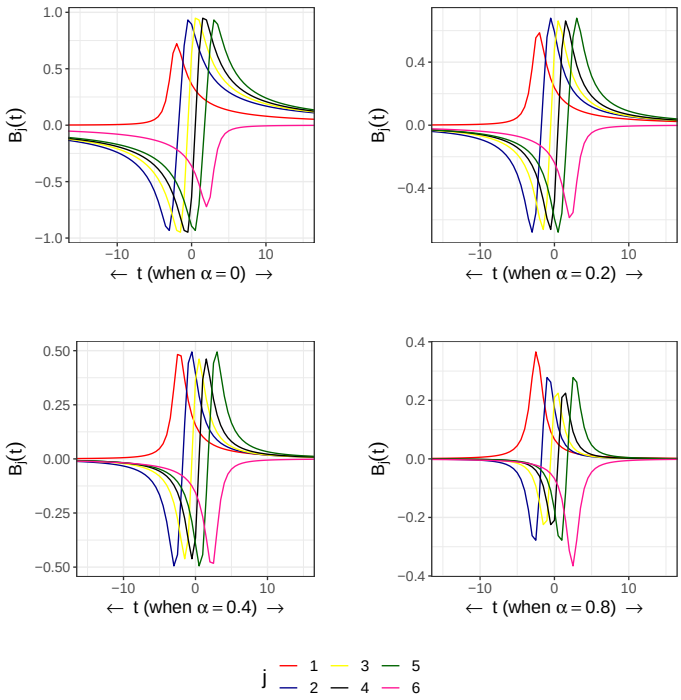

FIGURE 2.4: Generalized residuals for the Cauchy link with $\gamma = (-2.5, -1, 0, 1, 2.5)^T$.

An outlying observation can make the generalized residual unbounded when the probit and the complementary log-log link are used. However, the impact of these outliers becomes limited for the DPD. This also explains why the minimum density power divergence estimate of the slope parameter in the ordinal response models is more stable as compared to the MLE when robustness is a concern.

2.5 Asymptotic Properties

Here We shall present the weak consistency and asymptotic normality results of $\hat{\theta}_\alpha$. These follow from Ghosh and Basu (2013). Let us introduce

$$J^{(i)}(\alpha) = \frac{1}{1+\alpha}\mathbb{E}_{g_i}\Big[\nabla^2 V_i(x_i, Y_i, \theta_\alpha)\Big] \text{ and } \xi_i(\alpha) = \sum_{j=1}^{m} u_{\theta,i}(j)p^{\alpha}_{\theta,i}(j)g_i(j). \tag{2.17}$$

The matrix $J^{(i)}(\alpha)$ is assumed to be positive definite for $i = 1, 2, \ldots, n$. Also, define

$$\Psi_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} J^{(i)}(\alpha), \tag{2.18}$$

$$\Omega_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}ar_{g_i}\Big[\nabla V_i(x_i, Y_i, \theta_\alpha)\Big]. \tag{2.19}$$

Explicit forms of $\Psi_n(\alpha)$ and $\Omega_n(\alpha)$ are already given in (1.65) and (1.66). Denote $J = (1, 1, \ldots, 1)^T \in \mathbb{R}^{m-1}$, also recognize $e_i \in \mathbb{R}^{m-1}$ as the i-th unit vector in $\mathbb{R}^{m-1}$. Using the formulae as in (2.11) and (2.12) further simplifies the score vector as

$$u_{\theta,i}(j) = \begin{pmatrix} I_{m-1} \\ \\ -x_i J^T \end{pmatrix} D_{\theta,i}(j)(e_j - e_{j-1}), \tag{2.20}$$

where

$$D_{\theta,i}(j) = \frac{1}{p_{\theta,i}(j)} \begin{pmatrix} f(\gamma_1 - x_i^T\beta), & 0, & \dots & 0 \\ 0, & f(\gamma_2 - x_i^T\beta), & \dots & 0 \\ 0, & \dots & \ddots & 0 \\ 0, & 0, & \dots & f(\gamma_{m-1} - x_i^T\beta) \end{pmatrix}_{(m-1)\times(m-1)} \tag{2.21}$$

for $j = 1, 2, \dots, m$. We denote $e_0 = e_m = 0_{(m-1)\times 1}$ for notational convenience. It may be easily checked that

$$\frac{\partial}{\partial\gamma_s} D_{\theta,i}(j) = \begin{pmatrix} \frac{\partial}{\partial\gamma_s}\frac{f(\gamma_1 - x_i^T\beta)}{p_{\theta,i}(j)}, & 0, & \dots & 0 \\ 0, & \frac{\partial}{\partial\gamma_s}\frac{f(\gamma_2 - x_i^T\beta)}{p_{\theta,i}(j)}, & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0, & \dots, & 0, & \frac{\partial}{\partial\gamma_s}\frac{f(\gamma_{m-1} - x_i^T\beta)}{p_{\theta,i}(j)} \end{pmatrix}, \tag{2.22}$$

where

$$\frac{\partial}{\partial\gamma_s}\frac{f(\gamma_{s'} - x_i^T\beta)}{p_{\theta,i}(j)} = \begin{cases} \frac{p_{\theta,i}(j)f'(\gamma_s - x_i^T\beta) + (-1)^{j+s+1}f^2(\gamma_s - x_i^T\beta)}{p_{\theta,i}^2(j)} & \text{for } s' = s \text{ and } j = s, s+1, \\ \frac{(-1)^{j+s+1}f(\gamma_{s'} - x_i^T\beta)f(\gamma_s - x_i^T\beta)}{p_{\theta,i}^2(j)} & \text{for } s' \neq s \text{ and } j = s, s+1 \end{cases} \tag{2.23}$$

for all $s, s' = 1, 2, \dots, (m-1)$ and $j = 1, 2, \dots, m$. Similarly

$$\frac{\partial}{\partial\beta_k} D_{\theta,i}(j) = \begin{pmatrix} \frac{\partial}{\partial\beta_k}\frac{f(\gamma_1 - x_i^T\beta)}{p_{\theta,i}(j)}, & 0, & \dots & 0 \\ 0, & \frac{\partial}{\partial\beta_k}\frac{f(\gamma_2 - x_i^T\beta)}{p_{\theta,i}(j)}, & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0, & \dots, & 0, & \frac{\partial}{\partial\beta_k}\frac{f(\gamma_{m-1} - x_i^T\beta)}{p_{\theta,i}(j)} \end{pmatrix}, \tag{2.24}$$

where

$$\frac{\partial}{\partial \beta_k}\frac{f(\gamma_s - x_i^T\beta)}{p_{\theta,i}(j)} = -\frac{x_{ik}}{p_{\theta,i}^2(j)}\Bigg[p_{\theta,i}(j)f'(\gamma_s - x_i^T\beta) + f(\gamma_s - x_i^T\beta)\Big\{f(\gamma_{j-1} - x_i^T\beta) - f(\gamma_j - x_i^T\beta)\Big\}\Bigg] \tag{2.25}$$

for all s,k and j. Here f' denotes the first derivative of f with respect to its argument. Simple matrix algebra yields

$$\nabla u_{\theta,i}(j) = \begin{pmatrix} \frac{\partial}{\partial\gamma}u_{\theta,i}(j)^T \\ \\ \frac{\partial}{\partial\beta}u_{\theta,i}(j)^T \end{pmatrix}$$

$$= \begin{pmatrix} ((e_j - e_{j-1})^T \otimes I_{m-1})\frac{\partial}{\partial\gamma}D_{\theta,i}(j), & ((e_{j-1} - e_j)^T \otimes I_{m-1})\frac{\partial}{\partial\gamma}D_{\theta,i}(j)Jx_i^T \\ \\ ((e_j - e_{j-1})^T \otimes I_k)\frac{\partial}{\partial\beta}D_{\theta,i}(j), & ((e_{j-1} - e_j)^T \otimes I_k)\frac{\partial}{\partial\beta}D_{\theta,i}(j)Jx_i^T \end{pmatrix} \tag{2.26}$$

where $\otimes$ denotes the Kronecker product between two matrices. The blocks $[1,1]$ and $[1,2]$ of the partitioned matrix (2.26) can be further simplified using the $(m-1)^2 \times (m-1)$ matrices

$$\frac{\partial}{\partial\gamma}D_{\theta,i}(j) = \begin{pmatrix} \frac{\partial}{\partial\gamma_1}D_{\theta,i}(j) \\ \\ \frac{\partial}{\partial\gamma_2}D_{\theta,i}(j) \\ \vdots \\ \frac{\partial}{\partial\gamma_{m-1}}D_{\theta,i}(j) \end{pmatrix} \text{ and } \frac{\partial}{\partial\gamma}D_{\theta,i}(j)Jx_i^T = \begin{pmatrix} \frac{\partial}{\partial\gamma_1}D_{\theta,i}(j) \\ \\ \frac{\partial}{\partial\gamma_2}D_{\theta,i}(j) \\ \vdots \\ \frac{\partial}{\partial\gamma_{m-1}}D_{\theta,i}(j) \end{pmatrix} Jx_i^T. \tag{2.27}$$

Similar calculations also hold for the blocks $[2,1]$ and $[2,2]$ of (2.26). Using (2.20) and (2.26), one can explicitly calculate (2.18) and (2.19).

Next, we make the following assumptions to derive the consistency and asymptotic normality of $\hat{\theta}_\alpha$.

(A1) The best-fitting parameter θ_α is an interior point of Θ.

(A2) The error-distribution F is thrice continuously differentiable with bounded derivatives.

(A3) The matrices $J^{(i)}(\alpha)$s are positive definite for all i, and

$$\lambda_0 := \inf_n \Big[\min \text{ eigenvalue of } \Psi_n(\alpha) \Big] > 0. \tag{2.28}$$

(A4) The vector $x_i = (x_{i1}, \ldots, x_{ip})^T$ is such that the following conditions are true:

$$\frac{1}{n}\sum_{i=1}^{n} |x_{ij}x_{ij'}x_{ij^*}| = \mathcal{O}(1)\ ,\ \sup_n \max_{1\le i\le n} |x_{ij}| = \mathcal{O}(1) \text{ and } \sup_n \max_{1\le i\le n} |x_{ij}x_{ij'}| = \mathcal{O}(1) \tag{2.29}$$

for all $j, j', j^* = 1, 2, \ldots, p$.

Remark 2.1. *It may be easily checked that*

$$f'(x) = \begin{cases} \frac{e^{-x}(e^{-x}-1)}{(e^{-x}+1)^3} & \text{when } X \sim Logistic(0,1), \\ -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} x & \text{when } X \sim \mathcal{N}(0,1), \\ -\frac{1}{\pi} \cdot \frac{2x}{(1+x^2)^2} & \text{when } X \sim \mathcal{C}(0,1), \\ e^{x-e^x}(1-e^x) & \text{when } F(x) = 1 - e^{-e^x}. \end{cases} \tag{2.30}$$

See that f's are bounded for all these links as long as we assume $0 \times \pm\infty = 0$. In all these cases, density functions f and f'' are also bounded. This boundedness ensures that Assumption (A2) is satisfied for these link functions. Assumption (A3) refers to the condition that the smallest

eigen root of $\Psi_n(\alpha)$ should stay positive in the limit. Also, Assumption (A4) requires that the cross-products between the components of the non-stochastic covariates x_is be bounded.

Theorem 2.1. *Suppose the Assumptions (A1) to (A4) are true. Then the following holds:*

(a) $\hat{\theta}_\alpha$ *is weakly consistent for* θ_α *as* $n \to \infty$,

(b) $\sqrt{n}\Omega_n^{-\frac{1}{2}}(\alpha)\Psi_n(\alpha)\Big(\hat{\theta}_\alpha - \theta_\alpha\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, I_{m+p-1}\Big)$ *as* $n \to \infty$.

We omit the proof because it simply follows from Ghosh and Basu (2013) with an adaptation of (A4) that implies the boundedness of the derivatives.

Remark 2.2. *Let the true distributions belong to the model families, i.e.,* $g_i(j) = p_{\theta_*,i}(j)$ *for all* $j = 1, 2, \ldots, m$ *and* $i = 1, 2, \ldots, n$. *In this case, we get* $\xi_i(0) = 0$ *and* $\Psi_n(0) = \Omega_n(0) = I(\theta_*)$ *where* $I(\theta_*)$ *is the Fisher information matrix. So the asymptotic covariance matrix becomes* $I^{-1}(\theta_*)$.

Remark 2.3. *In simulation studies, one may compute the summed MSE to compare the performance of* $\hat{\theta}_\alpha$ *with the MLE. In such cases, the observed efficiency (Eff) of* $\hat{\theta}_\alpha$ *may be defined as* $Eff = \frac{MSE_{ML}}{MSE_{\hat{\theta}_\alpha}}$. *Because the estimators are consistent when the true distributions belong to the model family, it becomes approximately*

$$Eff \approx \frac{tr\Big(\Psi_n^{-1}(0)\Omega_n(0)\Psi_n^{-1}(0)\Big)}{tr\Big(\Psi_n^{-1}(\alpha)\Omega_n(\alpha)\Psi_n^{-1}(\alpha)\Big)} \text{ for sufficiently large } n, \tag{2.31}$$

where $tr(A)$ *denotes the trace of a matrix* A. *When true distributions belong to the model family, smaller values of* α *should lead to the estimators which are almost as good as the MLE. If there are some outliers in a data set, the performance of the MLE may become very unstable depending on the amount of anomaly in the data set. As the MDPDE naturally downweighs*

those outlying observations with respect to the model, one should expect that the MDPDE will have superior performance over the MLE, in that case, for relatively large values of α.

Remark 2.4. *The asymptotic covariance matrix of $\sqrt{n}\hat{\theta}_\alpha$ is given by $(\Psi_n^{-1}(\alpha)\Omega_n(\alpha)\Psi_n^{-1}(\alpha))$. This needs to be estimated in real data analysis. A consistent estimator of $\Psi_n(\alpha)$ is obtained by plugging in $\hat{\theta}_\alpha$ and δ_i respectively for θ_α and g_i in (2.18). This gives*

$$\begin{aligned}\hat{\Psi}_n(\alpha) =& \frac{1}{n}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{m}\left\{\nabla u_{\hat{\theta}_\alpha,i}(j) + (1+\alpha)u_{\hat{\theta}_\alpha,i}(j)u_{\hat{\theta}_\alpha,i}(j)^T\right\}p_{\hat{\theta}_\alpha,i}(j)^{1+\alpha} \\ &- \left\{\nabla u_{\hat{\theta}_\alpha,i}(Y_i) + \alpha u_{\hat{\theta}_\alpha,i}(Y_i)u_{\hat{\theta}_\alpha,i}(Y_i)^T\right\}p_{\hat{\theta}_\alpha,i}(Y_i)^{\alpha}\Bigg]. \qquad (2.32)\end{aligned}$$

However $\Omega_n(\alpha)$ cannot be estimated using only a single observation Y_i that comes from $g_i, i = 1, 2\ldots, n$. To avoid that we make use of the model densities as proxies for true densities in (2.19) and $\hat{\theta}_\alpha$ for θ_α. Thus we obtain

$$\hat{\Omega}_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{m}u_{\hat{\theta}_\alpha,i}(j)u_{\hat{\theta}_\alpha,i}(j)^T p_{\hat{\theta}_\alpha,i}(j)^{2\alpha+1} - \hat{\xi}_i(\alpha)\hat{\xi}_i(\alpha)^T\right\}, \qquad (2.33)$$

$$\hat{\xi}_i(\alpha) = \sum_{j=1}^{m}u_{\hat{\theta}_\alpha,i}(j)p_{\hat{\theta}_\alpha,i}(j)^{1+\alpha}. \qquad (2.34)$$

These estimates are required to apply the tuning parameter selection strategy to a real data example.

2.6 Robustness Studies

In this section, we will study the robustness of the minimum distance functional. This is, in fact, the main theme of this chapter. This section contains the following three parts– influence function analysis, asymptotic breakdown point analysis, and a discussion of the implosion resistance property of the slope estimates.

2.6.1 Influence Function Analysis

The influence function (IF) is one of the most popular measures of robustness in studying the impact of an infinitesimal data contamination on a statistical functional. Essentially, an estimate having a bounded influence function exhibits stable behaviour in the presence of outlying observations. Here, we shall present the influence function of the MDPD functional θ_α that minimizes $H(\theta)$. Given our setup, where x_is are fixed carriers, an outlier may only occur in the vertical direction (i.e., the Y-space). Therefore, having contamination at the i-direction in the vertical space only perturbs the distribution of total mass over the set of ordinal responses given at fixed x_i. This may be characterized when the true distribution G_i is contaminated at a point t_i as $G_{i,\epsilon} = (1-\epsilon)G_i + \epsilon\Lambda_{t_i}$, where Λ_{t_i} is the distribution function degenerate at $t_i = 1, 2, \ldots, m$.

Now we explain why all the true distributions may be simultaneously contaminated. Suppose we know the data-generating distributions a priori (as in the simulation studies) but the data points that arise all have ordinal responses which do not go along with the fixed regressors. This may happen due to temporary technical glitches in the process which are fixed after some time when that particular data set is already generated. These data should not be discarded altogether, as they might contain some valuable insights that cause technical issues. On the other hand, we should use our knowledge about the models that are expected to be true from the time when the experiment begins. Therefore, we must find a way to work with the erroneous data with the limited impact of such values on the overall inferential results. Though it is a one-off situation, statistically it may happen in any controlled experiment. In the context of multiple linear regressions, such observations are deemed outliers to a reference model because of diagnostic plots.

Let each true distribution G_i be contaminated as $G_{i,\epsilon}$ for $i = 1, 2, \ldots, n$. Through a straight forward differentiation, the influence function of θ_α is obtained as

$$\mathcal{IF}(\theta_\alpha, G_1, \cdots, G_n, t_1, \cdots, t_n) = \sum_{i=1}^{n} \frac{1}{n} \Psi_n^{-1}(\alpha) \Big\{ p_{\theta_\alpha,i}(t_i)^\alpha u_{\theta_\alpha,i}(t_i) - \xi_i(\alpha) \Big\}. \tag{2.35}$$

Observe that the i-th summand in (2.35) is exactly the influence function when contamination is present only at the i-th distribution G_i.

From (2.35) it is evident that the MDPD functional θ_α downweighs the influence of the data points that are inconsistent with the model with weights being chosen as model densities raised to the power of $\alpha \in [0, 1]$. This is a vector-valued function depending on $(t_1, t_2, \ldots, t_n)$ where $t_i \in \{1, 2, \ldots, m\}$ for all i. The influence function also depends on the fixed carriers $(x_1, x_2, \ldots, x_n)$ through the models. Since the number of levels is finite, it may be appropriate to plot the gross error sensitivity (GES) rather than the influence function itself. The GES based on (2.35) is given by

$$GES(\theta_\alpha) = \max_{t_1,\ldots,t_n} \Big\| \mathcal{IF}(\theta_\alpha, G_1, \ldots, G_n, t_1, t_2, \ldots, t_n) \Big\|, \tag{2.36}$$

where $\|\cdot\|$ is the Euclidean norm. The gross error sensitivity in (2.36) may be standardized using the asymptotic variance of θ_α. See Ghosh and Basu (2013) for more details. Let the components of the best-fitting parameter θ_α be denoted as $\gamma_\alpha = (\gamma_{1,\alpha}, \ldots, \gamma_{(m-1),\alpha})$ and $\beta_\alpha = (\beta_{1,\alpha}, \ldots, \beta_{p,\alpha})$. The gross error sensitivity of each component may be similarly obtained using the specified component of the IF given in (2.35). These are respectively denoted by $GES(\gamma_{1,\alpha}), \ldots, GES(\gamma_{(m-1),\alpha}), GES(\beta_{1,\alpha}), \ldots, GES(\beta_{p,\alpha})$. Dependency on the true distributions is kept implicit in these notations.

Next, we plot the GES of different components of θ_α related to Model 1 and Model 2 (described in Section 2.7) respectively in Figure 2.5 and Figure 2.6.

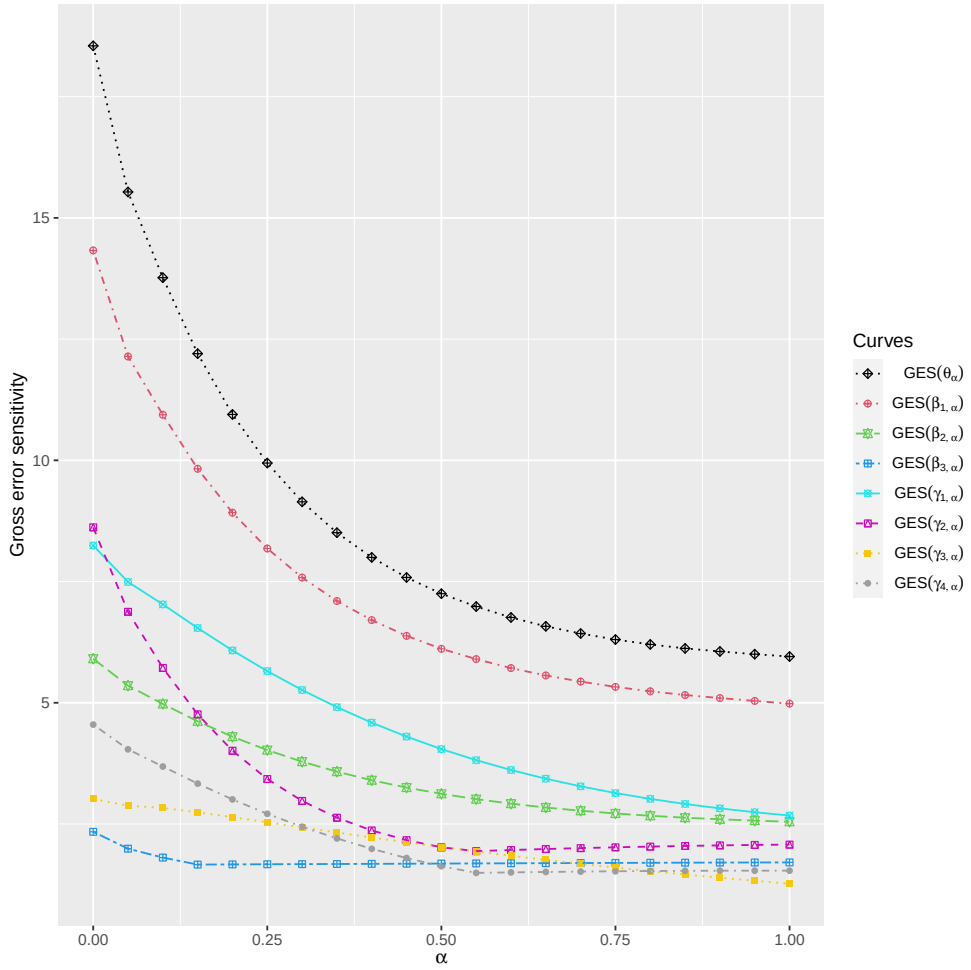

FIGURE 2.5: GES of the best-fitting parameters related to Model 1 with the probit link.

It is clear in Figure 2.5 and Figure 2.6 that $\alpha = 0$ represents the case for which GES attains its highest value. It decreases steadily as α increases from 0 towards 1. This gives a piece of strong evidence that when α is chosen close to zero the MDPD functional may tend to produce higher absolute bias at model misspecification; this indeed includes the case of maximum likelihood functional even for logit link (which is claimed to be robust in Scalera et al., 2021). As α increases, the MDPD functionals achieve better stability at model misspecification.

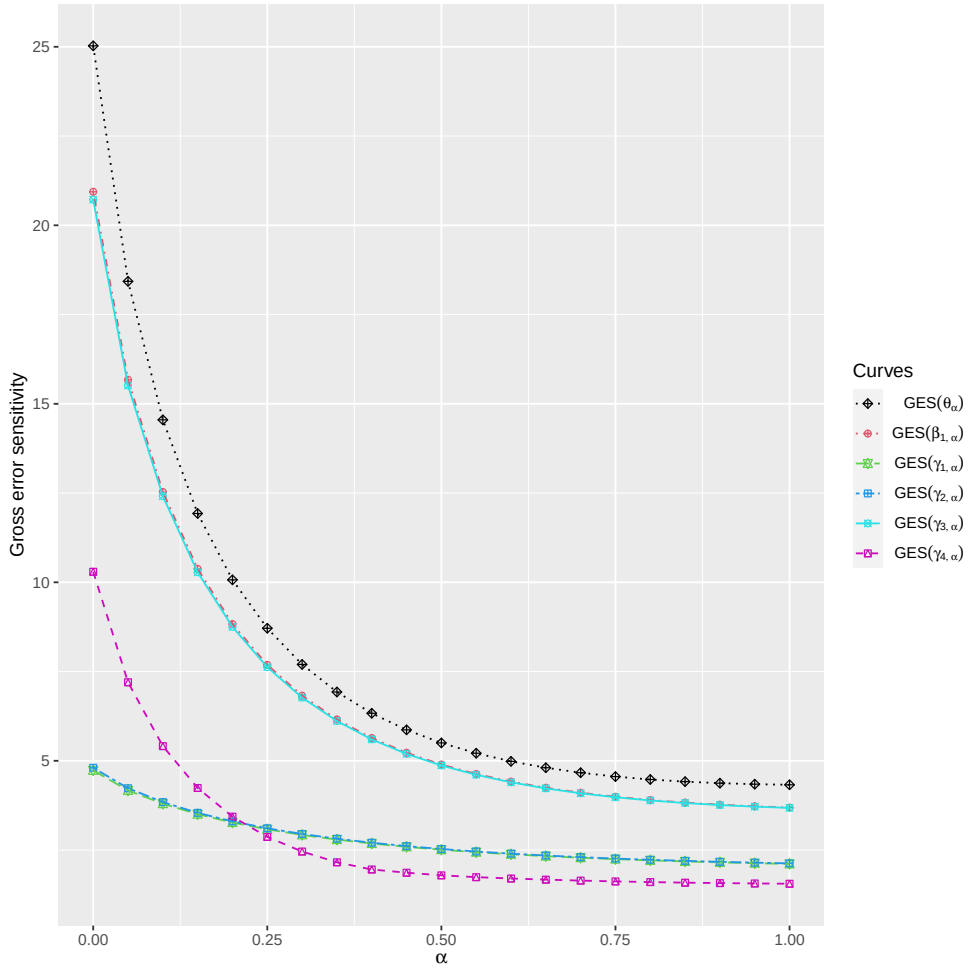

FIGURE 2.6: GES of the best-fitting parameters of Model 2 with the logit link.

2.6.2 Asymptotic Breakdown Point

We know that the influence function is a local measure of robustness that must be complemented with a global measure of stability. Now, we shall use Definition 1.6 to compute the (asymptotic) breakdown point of θ_α for $\alpha > 0$ with minor modifications suited to the present setup. Additionally, this requires that the number of ordinal responses "m" are fixed constants.

Suppose that the i-th true distribution G_i is contaminated at ϵ-proportion with a sequence of contaminating distributions $\{K_{i,M}\}_{M=1}^{\infty}$ as $H_{i,\epsilon,M} \equiv (1-\epsilon)G_i + \epsilon K_{i,M}$ at fixed

n, m. The corresponding probability density functions are denoted by the lower-case letters– $k_{i,M}$ and $h_{i,\epsilon,M}$, g_i. We assume that the true densities $\{g_i\}$, the models $\{p_{\theta,i} : \theta \in \Theta\}$ and the contaminating sequence of densities $\{k_{i,M}\}$ are all supported on χ. Let $\theta_\alpha^{h_{\epsilon,M}}$ be the MDPD functional when all G_is are ϵ-contaminated as above, i.e.,

$$\theta_\alpha^{h_{\epsilon,M}} := \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,M}, p_{\theta,i}). \tag{2.37}$$

We declare that θ_α breaks down at ϵ-contamination if $\|\theta_\alpha - \theta_\alpha^{h_{\epsilon,M}}\| \to \infty$ when $M \to \infty$ at fixed n, m (Simpson, 1989; Park and Basu, 2004). Define

$$D_\alpha\Big(g_i(j), p_{\theta,i}(j)\Big) = \left\{ p_{\theta,i}^{1+\alpha}(j) - \Big(1 + \frac{1}{\alpha}\Big) p_{\theta,i}^{\alpha}(j) g_i(j) + \frac{1}{\alpha} g_i(j)^{1+\alpha} \right\} \text{ for } j \in \chi, \tag{2.38}$$

and $L_i^\alpha = \sum_j p_{\theta_\alpha,i}(j)^{1+\alpha}$ for all i. Let us make the following assumptions to find the asymptotic breakdown point for θ_α. Throughout this section, m is assumed to be fixed.

(B1) g_i and $p_{\theta,i}$ belong to a common family for all $i = 1, 2, \ldots, n$.

(B2) There exists a set $B \subset \chi$ and a non-negative $\delta_1^*(j)$ depending on $j \in B$, such that

$$\big\{k_{i,M}(j) - p_{\theta,i}(j)\big\} \longrightarrow \delta_1^*(j) \geq 0 \text{ and } \sum_{j \in B} k_{i,M}(j) \longrightarrow 1 \text{ as } M \to \infty \text{ for all } i \tag{2.39}$$

uniformly for $\|\theta\| < \infty$. This condition means that on a set $B \subset \chi$, the contaminating sequence of densities $k_{i,M}$ asymptotically dominate $p_{\theta,i}$ when the parameter θ is uniformly bounded; moreover, B^c asymptotically becomes a $K_{i,M}$-null set as M increases to infinity for each i.

(B3) There exists a set $C \subset \chi$ and a non-negative $\delta_2^*(j)$ depending on $j \in C$, such that

$$\left\{p_{\theta_M,i}(j) - p_{\theta_\alpha,i}(j)\right\} \longrightarrow \delta_2^*(j) \geq 0 \text{ and } \sum_{j\in C} p_{\theta_M,i}(j) \longrightarrow 1 \text{ as } M \to \infty \text{ for all } i, \quad (2.40)$$

for any sequence $\{\theta_M\}$ such that $\|\theta_M\| \to \infty$ as $M \to \infty$. This means that as $\|\theta_M\|$ diverges, the associated sequence of models $p_{\theta_M,i}$ tend to dominate the true density $p_{\theta_\alpha,i}$ on a set C, and the sequence $p_{\theta_M,i}$ remain concentrated on C for $i = 1, 2, \ldots, n$.

(B4) (a) Assume that for any density q_i supported on χ, we have

$$d_\alpha(\epsilon q_i, p_{\theta,i}) \geq d_\alpha(\epsilon p_{\theta_\alpha,i}, p_{\theta_\alpha,i}) \text{ for all } \theta, i \text{ and } 0 < \epsilon < 1. \quad (2.41)$$

This means that $d_\alpha(\epsilon q_i, p_{\theta,i})$ will be minimized at $\theta = \theta_\alpha$ and $q_i = p_{\theta_\alpha,i}$ for all i.

(b) Further assume that

$$\limsup_{M\to\infty}\left(k_{i,M}(j)\right) \leq p_{\theta_\alpha,i}(j) \text{ for all } i, j, \text{ and } L^\alpha = \frac{1}{n}\sum_{i=1}^{n} L_i^\alpha < \infty \quad (2.42)$$

at fixed n, m.

Remark 2.5. *Since we know that* $p_{\theta,i}(j) = F(\gamma_j - x_i^T\beta) - F(\gamma_{j-1} - x_i^T\beta)$, *Assumption (B3) can be verified in the following situations.*

(S1) *Any particular* γ_j *decreases to* $-\infty$ *or increases to* ∞, *but* β*s remain bounded.*

(S2) *Let* $\gamma_{j_1} \to -\infty$ *and* $\gamma_{j_2} \to \infty$ *for any* $1 \leq j_1 < j_2 \leq m-1$, *but* β*s remain bounded.*

(S3) γ_j*s remain bounded but* β*s diverge to* $-\infty$ *or* ∞.

Consider the scenario (S1). Suppose $\gamma_j \to -\infty$. *Then* $\gamma_r \to -\infty$ *for* $r = 1, 2, \ldots, j$ *because* $\gamma_1 < \gamma_2 < \cdots \gamma_{j-1} < \gamma_j$. *Therefore* $p_{\theta,i}(r) \to 0$ *for* $r = 1, 2, \ldots, j$, *and* $\sum_{r=j+1}^{m} p_{\theta,i}(r) \to 1$. *So*

the probability under $p_{\theta,i}$ will become concentrated on the set $C^j(-\infty) := \{j+1,\ldots,m\}$. If we assume that $\gamma_j \to \infty$. Then $\gamma_r \to \infty$ for $r = j,\ldots,m$. This gives $p_{\theta,i}(r) \to 0$ for $r = j+1,\ldots,m$. In this case, the probability under $p_{\theta,i}$ gets concentrated on $C^j(\infty) := \{1,2,\ldots,j\}$.

Next we consider (S2). Now $\gamma_{j_1} \to -\infty$ and $\gamma_{j_2} \to \infty$ such that $\gamma_{j_1} < \gamma_{j_2}$. Using the above argument we see that the probability will get concentrated on the set $C^{j_1}(-\infty) \cap C^{j_2}(\infty) = \{j_1+1,\ldots,j_2\}$. As $\chi = C^0(-\infty) \cap C^m(\infty)$, both the sets $C^j(-\infty), C^j(\infty)$ are proper subsets of χ for $j = 1,\ldots,m-1$.

In (S3), it will depend on the sign of $x_i^T\beta$. If $x_i^T\beta \to \infty$, then all the terms $(\gamma_j - x_i^T\beta)$ goes to $-\infty$. In that case, the last term $p_{\theta,i}(m) = 1 - F(\gamma_{m-1} - x_i^T\beta) \to 1$. Thus the mass is concentrated on the singleton set $\{m\}$. On the other hand, if $x_i^T\beta \to -\infty$, then the probability mass gets concentrated on the set $\{1\}$. In all these above cases $\|\theta\| \to \infty$. These proper subsets can be taken as B and C as mentioned in Assumptions (B2) and (B3).

Assumption (B4) (a) ensures that the divergence in the left-hand side of (2.41) attains its minimum value at the models when θ being chosen as the best-fitting parameter. In (B4) (b) we state the extremity of contamination up to which the true distribution may be contaminated, but still produce reasonable MDPD estimates.

Theorem 2.2. *Suppose $g_i, p_{\theta,i}$ and the contaminating sequence of densities $\{k_{i,M}\}_{M=1}^{\infty}$ are supported on χ, $i = 1,\ldots,n$. Then under the Assumptions (B1) - (B4), the asymptotic breakdown point ϵ^* of the MDPD functional θ_α is atleast $\frac{1}{2}$ at the models for $\alpha > 0$.*

Proof. When breakdown occurs, $\|\theta_\alpha^{h_\epsilon,M}\|$ diverges to infinity. Therefore, we may choose θ_M to be $\theta_\alpha^{h_\epsilon,M}$ that satisfies the assumptions. First, we define the following sets

$$A_{i,M} = \left\{j : p_{\theta_\alpha,i}(j) > \max\left\{k_{i,M}(j), p_{\theta_M,i}(j)\right\}\right\} \text{ and } S_{i,M} = \left\{j : p_{\theta_\alpha,i}(j) > k_{i,M}(j)\right\}. \quad (2.43)$$

The i-th divergence between $h_{i,\epsilon,M}$ and $p_{\theta_M,i}$ can be decomposed as

$$d_\alpha(h_{i,\epsilon,M}, p_{\theta_M,i}) = \sum_{j:A_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) + \sum_{j:A_{i,M}^c} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big). \tag{2.44}$$

Clearly, $A_{i,M} \subset S_{i,M}$. It follows from Assumption (B2) that for sufficiently large $M \equiv M(j)$, $k_{i,M}(j) - p_{\theta_\alpha,i}(j) \geq 0$ for all $j \in B$. Since B is a finite set, taking the maximum $M_* = \max\{M(j) : j \in B\}$ ensures that $k_{i,M'}(j) - p_{\theta_\alpha,i}(j) \geq 0$ for all $j \in B$ and $M' > M_*$. Thus, it follows that $S_{i,M} \to S \subseteq B^c$. See that

$$\sum_{j\in A_{i,M}} k_{i,M}(j) \leq \sum_{j\in S_{i,M}} k_{i,M}(j) \leq \sum_{j\in B^c} k_{i,M}(j) \longrightarrow 0 \text{ as } M \to \infty. \tag{2.45}$$

by Assumption (B2). From Assumption (B3) it similarly follows that $\sum_{j:A_{i,M}} p_{\theta_M,i}(j) \to 0$ as $M \to \infty$. Therefore, the sets $A_{i,M}$ converge to a set of probability measures zero under both $p_{\theta_M,i}$ and $k_{i,M}$ for each i. Using Assumptions (B2) and (B3) together, we also get $\max\{k_{i,M}(j), p_{\theta_M,i}(j)\} \to 0$ for all $j \in A_{i,M}$ as $M \to \infty$. Therefore

$$D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) = \left\{p_{\theta_M,i}(j)^{1+\alpha} - \Big(1+\frac{1}{\alpha}\Big)p_{\theta_M,i}^{\alpha}(j)h_{i,\epsilon,M}(j) + \frac{1}{\alpha}h_{i,\epsilon,M}^{1+\alpha}(j)\right\} \tag{2.46}$$

$$\longrightarrow \frac{1}{\alpha}(1-\epsilon)^{1+\alpha}p_{\theta_\alpha,i}(j)^{1+\alpha} = D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big) \tag{2.47}$$

for all $j \in A_{i,M}$ as $M \to \infty$. Now applying the DCT gives the following

$$\left|\sum_{j:A_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \sum_{j:A_{i,M}} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big)\right| \longrightarrow 0 \text{ as } M \to \infty. \tag{2.48}$$

Assumptions (B2) and (B3) together gives

$$\left|\sum_{j:A_{i,M}} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big) - \sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big)\right| \longrightarrow 0 \text{ when } M \to \infty. \tag{2.49}$$

A simple application of the triangle inequality gives the following result

$$\left| \sum_{j:A_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big) \right| \longrightarrow 0 \text{ for } M \to \infty. \tag{2.50}$$

We also have

$$\sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), 0\Big) = \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} \sum_{j\in\chi} p_{\theta_\alpha,i}^{1+\alpha}(j) = \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} L_i^\alpha. \tag{2.51}$$

See that, when $M \to \infty$,

$$\sum_{j\in A_{i,M}} p_{\theta_\alpha,i}(j) = \sum_{j\in A_{i,M}\cap\{k_{i,M}\downarrow+0\}\cap\{p_{\theta_M,i}\downarrow+0\}} p_{\theta_\alpha,i}(j) \longrightarrow \sum_{j\in\chi} p_{\theta_\alpha,i}(j) = 1. \tag{2.52}$$

This gives $\sum_{j:A_{i,M}^c} p_{\theta_\alpha,i}(j) \to 0$ as $M \to \infty$. We also know that $\sum_{j:A_{i,M}^c} k_{i,M}(j) \to 1$ and $\sum_{j:A_{i,M}^c} p_{\theta_M,i}(j) \to 1$ when $M \to \infty$. Therefore

$$\left| \sum_{j:A_{i,M}^c} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \sum_{j\in\chi} D_\alpha\Big(\epsilon k_{i,M}(j), p_{\theta_M,i}(j)\Big) \right| \to 0 \text{ as } M \to \infty. \tag{2.53}$$

From Assumption (B4), we see that

$$\sum_{j\in\chi} D_\alpha\Big(\epsilon k_{i,M}(j), p_{\theta_M,i}(j)\Big) = d_\alpha(\epsilon k_{i,M}, p_{\theta_M,i}) \geq d_\alpha(\epsilon p_{\theta_\alpha,i}, p_{\theta_\alpha,i}) = a(\epsilon) L_i^\alpha, \tag{2.54}$$

where $a(\epsilon) = \Big\{1 - \Big(1 + \frac{1}{\alpha}\Big)\epsilon + \frac{1}{\alpha}\epsilon^{1+\alpha}\Big\}$. When n is fixed,

$$\liminf_{M\to\infty} \Big\{ d_\alpha(h_{i,\epsilon,M}, p_{\theta_M,i}) \Big\} \geq \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} L_i^\alpha + a(\epsilon) L_i^\alpha \tag{2.55}$$

for all i. Averaging over all i, we get

$$\liminf_{M\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,M}, p_{\theta_M,i})\right\} \geq \frac{1}{n}\sum_{i=1}^{n}\liminf_{M\to\infty}\left\{d_\alpha(h_{i,\epsilon,M}, p_{\theta_M,i})\right\} \tag{2.56}$$

$$\geq \frac{1}{\alpha}(1-\epsilon)^{1+\alpha}L^\alpha + a(\epsilon)L^\alpha = a_1(\epsilon) \tag{2.57}$$

at fixed n, where $L^\alpha = \frac{1}{n}\sum_i L_i^\alpha$. We shall have a contradiction to our assumption that $\{k_{i,M}\}_{M=1}^{\infty}$ is the sequence for which breakdown occurs if we can show that there exists a constant value θ_{**} in the parameter space such that $\theta_M \to \theta_{**}$ but

$$\limsup_{M\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n} d_\alpha(h_{i,\epsilon,M}, p_{\theta_{**},i})\right\} < a_1(\epsilon) \tag{2.58}$$

for the same sequences $\{k_{i,M}\}_{M=1}^{\infty}$. It means that the sequence $\{\theta_M\}$ could not minimize the density power divergence: $\frac{1}{n}\sum_i d_\alpha(h_{i,\epsilon,M}, p_{\theta_M,i})$ at each M.

Define $B_{i,M} = \left\{j : k_{i,M}(j) > \max\{p_{\theta_\alpha,i}(j), p_{\theta_M,i}(j)\}\right\}$ for some sequence $\{\theta_M\}$. From Assumptions (B2) and (B3), it follows that $\sum_{j:B_{i,M}} p_{\theta_\alpha,i}(j) \to 0$, $\sum_{j:B_{i,M}} p_{\theta_M,i}(j) \to 0$ and $\sum_{j:B_{i,M}^c} k_{i,M}(j) \to 0$ as $M \to \infty$. Therefore under $k_{i,M}$ the set $B_{i,M}^c$ converges to a probability null set, also under both $p_{\theta_\alpha,i}$ and $p_{\theta_M,i}$ the set $B_{i,M}$ also converge to the sets of zero probability measures. Thus for all $j \in B_{i,M}$,

$$\left|D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - D_\alpha\Big(\epsilon k_{i,M}(j), 0\Big)\right| \longrightarrow 0 \tag{2.59}$$

$$\implies \Big|\sum_{j:B_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \sum_{k_{i,M}>0} D_\alpha\Big(\epsilon k_{i,M}(j), 0\Big)\Big| \longrightarrow 0 \Big[\text{ DCT }\Big]. \tag{2.60}$$

Observe that $D_\alpha\Big(\epsilon k_{i,M}(j), 0\Big) = \frac{\epsilon^{1+\alpha}}{\alpha}k_{i,M}^{1+\alpha}(j)$ when $k_{i,M} > 0$ and $\alpha > 0$. It follows that

$$\left|\sum_{j:B_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \frac{\epsilon^{1+\alpha}}{\alpha}\sum_{j\in\chi} k_{i,M}^{1+\alpha}(j)\right| \longrightarrow 0 \text{ as } M\to\infty. \tag{2.61}$$

Similarly, we have

$$\left| \sum_{j:B^c_{i,M}} D_\alpha\Big(h_{i,\epsilon,M}(j), p_{\theta_M,i}(j)\Big) - \sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), p_{\theta_{**},i}(j)\Big)\right| \longrightarrow 0 \text{ for } M\to\infty. \quad (2.62)$$

Now see that

$$\begin{aligned}\limsup_{M\to\infty}\Big\{d_\alpha(h_{i,\epsilon,M}, p_{\theta_{**},i})\Big\} &= \limsup_{M\to\infty}\left\{\sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), p_{\theta_{**},i}(j)\Big) + \frac{\epsilon^{1+\alpha}}{\alpha}\sum_{j\in\chi} k^{1+\alpha}_{i,M}(j)\right\}\\ &\le \sum_{j\in\chi} D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), p_{\theta_{**},i}(j)\Big) + \frac{\epsilon^{1+\alpha}}{\alpha}L^\alpha_i \text{ for all } i. \quad (2.63)\end{aligned}$$

Averaging (2.63) overall $i = 1, 2, \ldots, n$ we get

$$\begin{aligned}\limsup_{M\to\infty}\left\{\frac{1}{n}\sum_{i=1}^n d_\alpha(h_{i,\epsilon,M}, p_{\theta_{**},i})\right\} &\le \frac{1}{n}\sum_{i=1}^n \limsup_{M\to\infty}\Big\{d_\alpha(h_{i,\epsilon,M}, p_{\theta_{**},i})\Big\}\\ &\le \frac{1}{n}\left\{\sum_{i=1}^n\sum_{j=1}^m D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), p_{\theta_{**},i}(j)\Big) + \frac{\epsilon^{1+\alpha}}{\alpha}\sum_{i=1}^n L^\alpha_i\right\}\end{aligned} \quad (2.64)$$

at fixed n. Let us choose $\theta_m = \hat{\theta}_\alpha$. Then Theorem 2.1 (a) implies that $\theta_{**} = \theta_\alpha$. Let us substitute θ_α for θ_{**} in the first part of (2.64), we get

$$\sum_{j=1}^m D_\alpha\Big((1-\epsilon)p_{\theta_\alpha,i}(j), p_{\theta_\alpha,i}(j)\Big) = a(1-\epsilon)\sum_{j=1}^m p_{\theta_\alpha,i}(j)^{1+\alpha} = a(1-\epsilon)L^\alpha_i \text{ for all } i. \quad (2.65)$$

Hence

$$\limsup_{M\to\infty}\left\{\frac{1}{n}\sum_{i=1}^n d_\alpha(h_{i,\epsilon,M}, p_{\theta_\alpha,i})\right\} \le a(1-\epsilon)L^\alpha + \frac{\epsilon^{1+\alpha}}{\alpha}L^\alpha = a_2(\epsilon). \quad (2.66)$$

Asymptotically there will be no breakdown at ϵ-contamination when $a_2(\epsilon) < a_1(\epsilon)$. Note that $a_1(\epsilon)$ and $a_2(\epsilon)$ are strictly decreasing and increasing functions respectively in ϵ and $a_1(\frac{1}{2}) = a_2(\frac{1}{2})$. So, the breakdown occurs asymptotically at the model when $\epsilon \ge \frac{1}{2}$. □

In Theorem 2.2, we establish that minimizing the DPD generates estimators with very high breakdown points for all $\alpha > 0$.

2.6.3 Resistant to Implosion Breakdown of the Slope Estimates

The notion of breakdown discussed earlier is called the explosive breakdown because it makes an estimator explode towards infinity. Following Croux et al. (2013) we know that the lack of robustness of the MLE is due to the implosion of the slope estimator $\hat{\beta}$ towards zero. Let the initial sample be $Z_n = \{z_1, \ldots, z_n\}$ where $z_i = (x_i, Y_i); i = 1, \ldots, n$. Upon adding m_o outliers to the initial sample, it looks like $Z'_{n+m_o} = \{z_1, \ldots, z_n, z_{n+1}, \ldots, z_{n+m_o}\}$. Then the implosion breakdown point of $\hat{\beta}$ is defined as $\epsilon^- = \frac{m_o^-}{m_o^- + n}$ where

$$m_o^- = \left\{ m_o \in \mathbb{N}_0 : \inf_{z_{n+1}, \ldots, z_{n+m_o}} \|\hat{\beta}(Z'_{n+m_o})\| = 0 \right\}. \tag{2.67}$$

It is difficult to calculate the implosion breakdown point theoretically. We shall empirically demonstrate that the minimum density power divergence estimate of β has a very high implosion breakdown point for $\alpha > 0$.

For illustration, let us draw a random sample of size 50 as in Figure 2.7. This data set contains non-stochastic covariates $x = (x_1, x_2)^T$ whose components are fixed at the values generated by two independent standard normal distributions. The associated random errors are distributed according to the standard logistic distribution. Ordinal responses, that classify the observations into 3 categories, are generated through (2.2) using $\beta = (-1, 1.5)^T, \gamma = (-1, 1)^T$. Norm of true β is $\|\beta\| = 1.803$. At the initial sample, the MLE of slope yields $\|\hat{\beta}_{ML}\| = 2.20$ with the misclassification rate equal to 36%.

However $\hat{\beta}_{ML}$ may be completely perturbed when an outlier in the form of $((s, -s), 3)$ is added to the initial sample along the diagonal line as in Figure 2.7. Here $x = (s, -s)$ is the

projection of the outlying observation in the x_1x_2-space. Notice that as soon as s goes outside $(-2,2)$, it becomes an outlying. However, the outlying region in the x_1x_2-space looks more deserted in the presence of fewer Y-values whenever s is positive. For each value of s, the parameters of the ordinal regression model are estimated by the MDPD methods and the corresponding misclassification rate is computed. The norms of these estimates, i.e., $||\hat{\beta}||$s and the prediction misclassification rates are reported, respectively, in Figure 2.8 and Figure 2.9 as functions of s. As expected, negative values of s do not incur much bias in the slope estimates. On the other hand, as soon as s gets positive, the impact of an added observation becomes visibly more severe in the ML estimation and gets even more extreme as s increases. Not only does the norm of the slope estimator go to zero, but the misclassification rate also reaches its maximum value which is about 57%. We notice that the MDPDE of β becomes more resistant to implosion whenever the tuning parameter increases; and it becomes the most resistant for the minimum L_2 distance estimate. Consequently, the misclassification rate decreases steadily along the way as the estimates become more stable which happens with the increment of the tuning parameter α from 0 towards 1.

In Table 2.1 and Table 2.2 we report the minimized norm of the slope estimates when the set of outliers– $I_s = \{z_{50+i} = (s,-s,3) : i = 1,2,\ldots,50\}$ are added to the initial data with $s = 8,10$. Also, the proportion of outliers corresponding to the lowest $||\hat{\beta}||$ is reported. The third column in these tables sort of gives an idea about which value of the tuning parameter α adds more resistance to the slope estimates such that it safeguards against implosion. It may be thought of as a finite sample analogue of the implosion breakdown point. Notice that $||\hat{\beta}||$ is close towards zero when $\alpha \downarrow 0+$, resulting into the lowest tolerance of outliers, i.e., 5.7% for $s = 8$, and 3.8% for $s = 10$. Moreover, as s increases more in the positive direction, this tolerance level of MLE decreases. In this case, the MLE of β can accommodate a smaller proportion of outliers before it starts

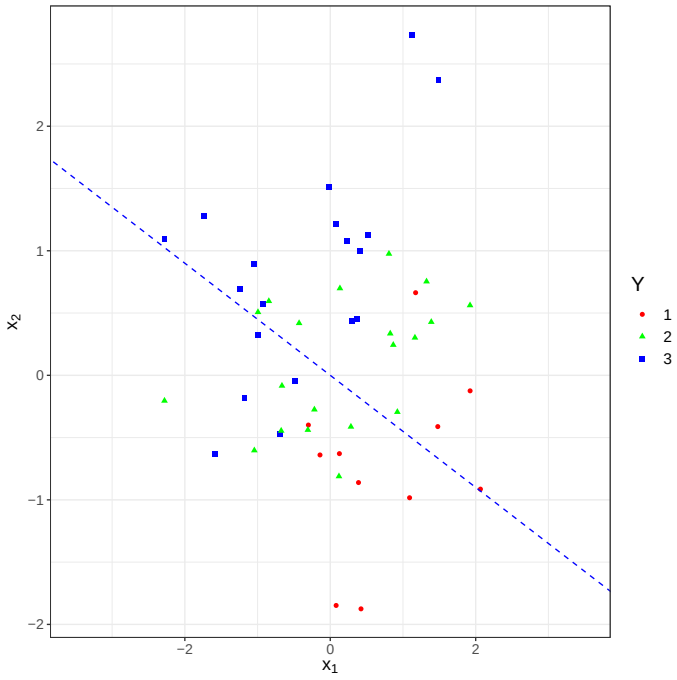

FIGURE 2.7: Outliers $((s,-s),3)$ will be added along the diagonal line to this scatter plot of a simulated data set.

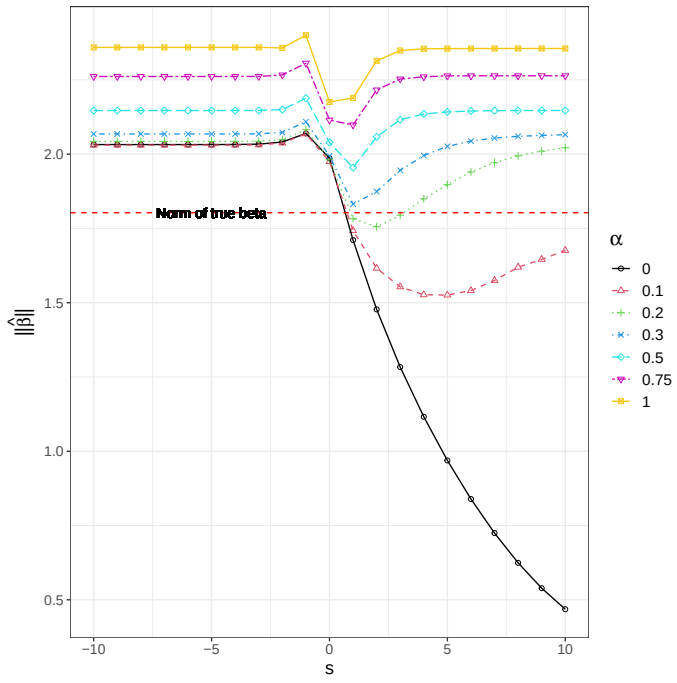

FIGURE 2.8: Norm of $\hat{\beta}$ when a single outlier is added along the diagonal line in Figure 2.7.

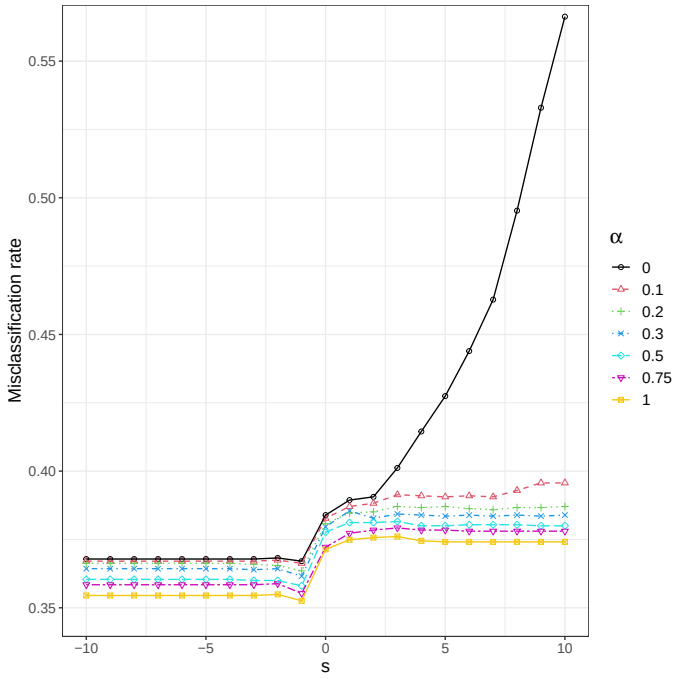

FIGURE 2.9: Misclassification rate of the slope estimates.

TABLE 2.1: Minimum of $||\hat{\beta}||$ over I_8

α	$\min ||\hat{\beta}||$	Prop. of outliers
0	0.323	0.0566
0.1	0.374	0.0566
0.2	0.468	0.438
0.3	1.070	0.479
0.5	2.080	0.495
0.75	2.250	0.390
1	2.340	0.306

TABLE 2.2: Minimum of $||\hat{\beta}||$ over I_{10}

α	$\min ||\hat{\beta}||$	Prop. of outliers
0	0.319	0.0385
0.1	0.379	0.180
0.2	0.613	0.390
0.3	1.720	0.474
0.5	2.120	0.500
0.75	2.250	0.390
1	2.340	0.306

imploding. However, as the tuning parameter α increases, the tolerance level improves significantly. Also, the decreasing trend of tolerance with the increment of s is not observed for the MDPDE. In this numerical study, we see that the implosion breakdown point of the MDPDE of β becomes very high for $\alpha > 0$ when compared to the MLE. In some cases, this may become as high as 50% for some positive values of α.

2.7 Numerical Studies

Samples of sizes $n = 150$ and $n = 200$ are drawn from each of the following models described in Subsection 2.7.1, and this experiment is repeated over 1000 (say, B) replications. For any given method, let $\hat{\theta}^{(b)} = (\hat{\gamma}^{(b)}, \hat{\beta}^{(b)})$ be the estimate of θ obtained in the b-th replication, $b = 1, 2, \ldots, B$. The simulated mean of the estimate is obtained as $\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$. The squared biases of $\hat{\gamma}$ and $\hat{\beta}$ are respectively given by $\|\hat{\gamma} - \gamma\|^2$ and $\|\hat{\beta} - \beta\|^2$. The mean squared error (MSE) of $\hat{\gamma}$ is defined as $MSE(\hat{\gamma}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\gamma}^{(b)} - \gamma)^T(\hat{\gamma}^{(b)} - \gamma)$, and similarly for $\hat{\beta}$. We obtain the MSE of $\hat{\theta}$ simply by adding together all the MSEs incurred in each of its components. When $\hat{\theta}$ is consistent its MSE consistently estimates the trace of the asymptotic covariance matrix.

In this above setup, we will numerically compare the performances of the minimum density power divergence estimates to the MLE and the robust alternatives proposed by Croux et al. (2013) and Iannario et al. (2017). For the sake of completeness, we briefly mention the particular choices of weight functions and tuning constants that are used in simulation studies for these last two methods.

In Croux et al. (2013), a weighted log-likelihood function is constructed by multiplying the usual log-likelihood function and weight function $w_i = \frac{p+3}{d_i+3}$ at the i-th data point, $i = 1, \ldots, n$. Here d_i denotes the robust Mahalanobis distance of the i-th covariate $x_i \in \mathbb{R}^p$ computed in the space of explanatory variables. The weighted maximum likelihood (WML) estimator is then obtained by maximizing the weighted log-likelihood function. On the other hand, Iannario et al. (2017) propose multiplying the usual likelihood-score

function by one of the following weight functions

$$w_1(Y_i, x_i, \theta) = \min\left\{1, \frac{c}{\sum_{j=1}^{m} \delta_i(j)|e_{ij}(\theta)|}\right\},\ w_2(Y_i, x_i, \theta) = \min\left\{1, \frac{c}{\sum_{j=1}^{m} \delta_i(j)|e_{ij}(\theta)| \cdot \|x_i\|}\right\}, \tag{2.68}$$

$$w_3(Y_i, x_i, \theta) = \min\left\{1, \frac{c}{\|x_i\|}\right\} \tag{2.69}$$

where c (> 0), $\|x_i\|$ and $e_{ij}(\theta)$ respectively denote the tuning constant, norm of x_i and the generalized residuals. For more details about the computations of $\|x_i\|$ and the definition of the generalized residuals, the readers are referred to Iannario et al. (2017). Given a particular data example, the choice of a weight function depends on the link function under consideration. Based on the numerical studies, Iannario et al. (2017) suggest using $c \in [1.1, 1.7]$ for the probit link, and $c \in [0.6, 1]$ (see Table 2 for Trace criterion in Iannario et al. (2017) for the logit link to keep the loss of efficiency below 5%. When the complementary log-log (or, simply the log-log) and the Cauchy link function are added to the simulation studies, we choose the same values of c as the probit and logit link, respectively, since no further suggestion is made in Iannario et al. (2017).

2.7.1 Simulation Studies: Pure Models

We consider the following models in the simulation studies.

Model 1 : The response variable Y, generated by (2.1), assumes the values $1, \ldots, 5$. Y depends on three dichotomous $0-1$ variables X_1, X_2, X_3 such that at most one of them can take the value 1. The cut-off points and the regression coefficients are respectively given by $\gamma = (-0.7, 0, 1.5, 2.9)^T$ and $\beta = (2.5, 1.2, 0.5)^T$ for both the probit and the complementary log-log link functions.

Model 2 : The response variable Y is generated through the latent variable $Y^* = 1.5X + e$ where the regressor X is assumed to have come from $\mathcal{N}(0,1)$. The categories of Y are determined by the cut-off points $\gamma = (-1.7, -0.5, 0.5, 1.7)^T$ and $\gamma = (-2.1, -0.6, 0.6, 2.1)^T$, respectively, when the error component follows $\mathcal{N}(0,1)$ and the logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$. Furthermore, the Cauchy link is used with the same cut-offs as the logit link.

Model 3 : The response variable Y assumes 4 categories. It depends on two regressors $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,4)$ with $Cov(X_1, X_2) = 1.2$. The regression coefficients are given by $\beta = (1.5, 0.7)^T$. We use the cut-off points $\gamma = (-2.3, 0, 2.3)^T$ for the probit link and $\gamma = (-2.6, 0, 2.6)^T$ for the logit link.

Model 4 : The response variable Y, that takes 3 categories, depends upon three regressors $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \mathcal{N}(0,4)$ and $X_3 \sim \mathcal{N}(0,9)$ such that $Cov(X_1, X_2) = 1.5$, $Cov(X_1, X_3) = 0.8$ and $Cov(X_1, X_3) = 2.5$. The regression parameters are given by $\beta = (2.5, 1.2, 0.7)^T$. The cut-off points are chosen as $\gamma = (-3.8, 3.8)^T$ and $\gamma = (-4, 4)^T$ respectively for the probit and the logit links.

Model 5 : The response variable Y is generated through $Y^* = 2.5D + 1.2X + 0.7XD + e$ where $D \sim Bernoulli(\frac{1}{2})$ and $X \sim \mathcal{N}(0,1)$. Here $\gamma = (-1, 1, 3)^T$ for the probit and the complementary log-log links; and $\gamma = (-1.4, 1.1, 3.4)^T$ for the logit link.

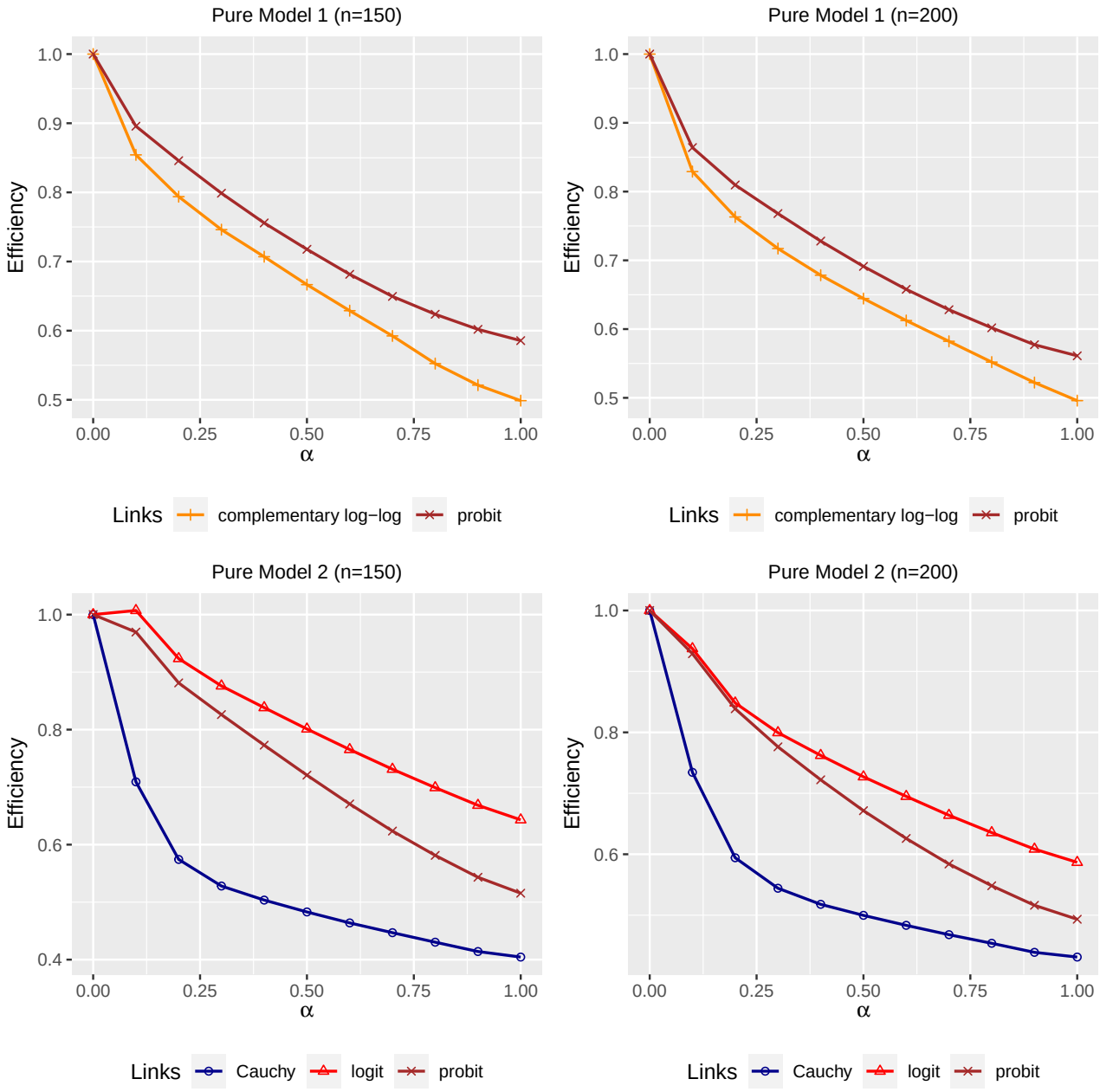

FIGURE 2.10: Graphs of efficiency at pure Model 1 and Model 2 for different sample sizes.

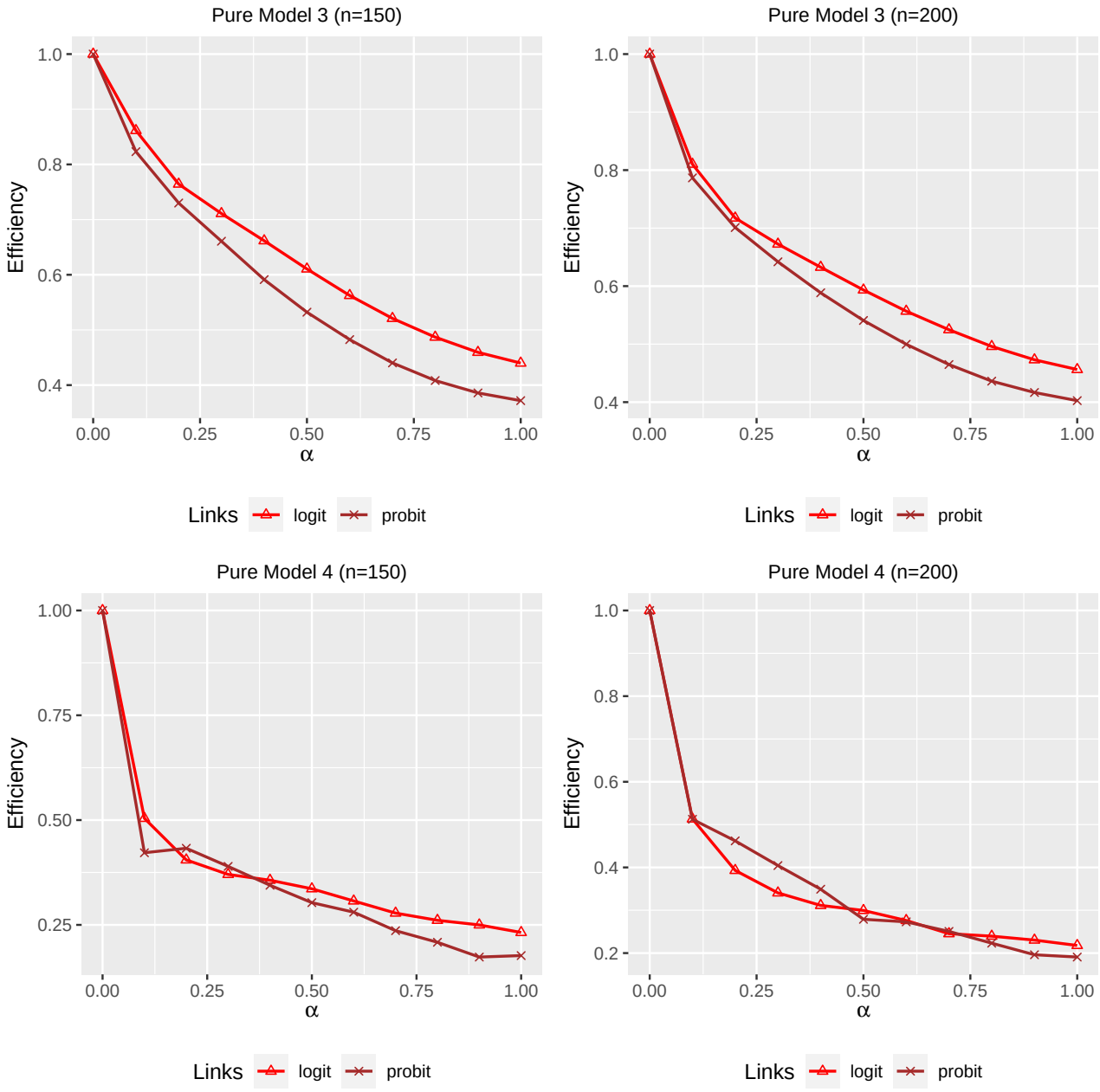

FIGURE 2.11: Graphs of efficiency at pure Model 3 and Model 4 for different sample sizes.

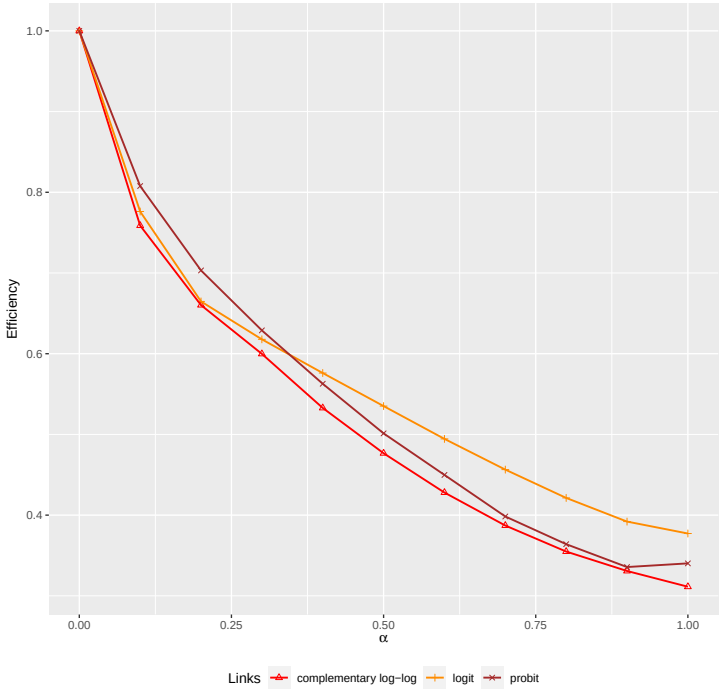

FIGURE 2.12: Graphs of efficiency at pure Model 5 with sample size n = 150.

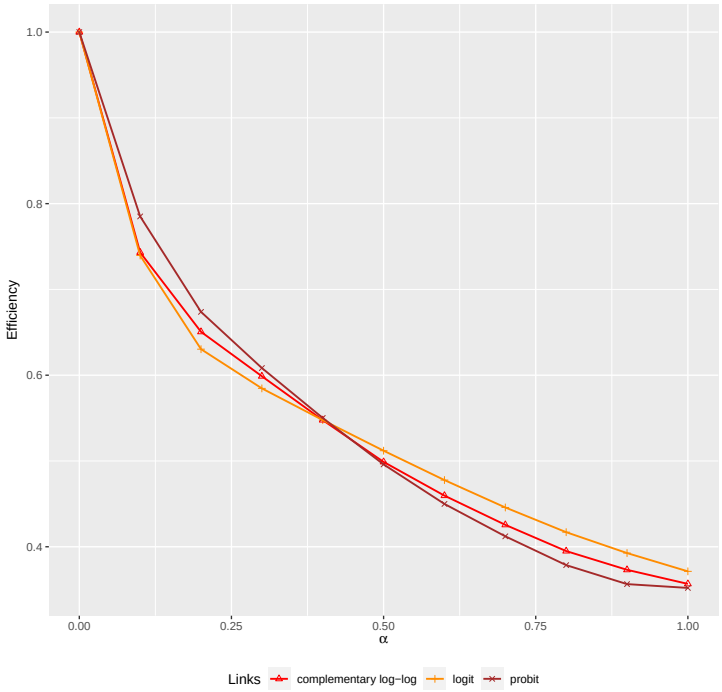

FIGURE 2.13: Graphs of efficiency at pure Model 5 with sample size n = 200.

TABLE 2.3: Squared bias and MSE of the estimates at Model 1 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.00059	0.00007	0.01810	0.02113
			(0.00027)	(0.00008)	(0.01232)	(0.01464)
		DPD (α)				
		0.1	0.00050	0.00020	0.02044	0.02335
			(0.00034)	(0.00009)	(0.01437)	(0.01683)
		0.2	0.00060	0.00025	0.02182	0.02456
			(0.00044)	(0.00012)	(0.01555)	(0.01775)
		0.3	0.00069	0.00030	0.02316	0.02595
			(0.00050)	(0.00015)	(0.01650)	(0.01861)
		0.5	0.00094	0.00042	0.02572	0.02894
			(0.00067)	(0.00023)	(0.01837)	(0.02064)
		0.8	0.00137	0.00064	0.02920	0.03369
			(0.00098)	(0.00040)	(0.02091)	(0.02389)
		1.0	0.00161	0.00078	0.03057	0.03642
			(0.00116)	(0.00053)	(0.02210)	(0.02594)
		Iannario (c)				
		1.1	0.91683	0.39137	1.11269	0.50400
			(0.88153)	(0.38396)	(0.94061)	(0.44379)
		1.4	0.50530	0.31753	0.53996	0.35084
			(0.49168)	(0.32074)	(0.51512)	(0.34237)
		1.5	0.35871	0.16981	0.41747	0.26115
			(0.36809)	(0.18241)	(0.43070)	(0.30426)
		1.7	0.11496	0.04636	0.18342	0.13085
			(0.10749)	(0.04001)	(0.16445)	(0.10902)
		Croux	0.00083	0.00030	0.02712	0.02527
			(0.00048)	(0.00013)	(0.01582)	(0.01791)

TABLE 2.4: Squared bias and MSE of the estimates at Model 1 with the complementary log-log link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	log-log	MLE	0.00061	0.00039	0.02038	0.02548
			(0.00029)	(0.00013)	(0.01460)	(0.01775)
		DPD (α)				
		0.1	0.00105	0.00072	0.02429	0.02939
			(0.00055)	(0.00029)	(0.01793)	(0.02110)
		0.2	0.00120	0.00084	0.02622	0.03155
			(0.00065)	(0.00037)	(0.01958)	(0.02282)
		0.3	0.00136	0.00099	0.02779	0.03367
			(0.00073)	(0.00044)	(0.02082)	(0.02429)
		0.5	0.00177	0.00144	0.03052	0.03827
			(0.00094)	(0.00064)	(0.02292)	(0.02729)
		0.8	0.00283	0.00277	0.03548	0.04754
			(0.00142)	(0.00121)	(0.02600)	(0.03262)
		1.0	0.00369	0.00395	0.03871	0.05322
			(0.00193)	(0.00189)	(0.02814)	(0.03711)
		Iannario (c)				
		1.1	1.26019	0.41864	1.35396	0.52829
			(1.18928)	(0.39308)	(1.23907)	(0.44351)
		1.4	1.07112	0.32326	1.38608	0.70113
			(1.10160)	(0.32870)	(1.37134)	(0.64300)
		1.5	0.46727	0.19361	0.83238	0.60076
			(0.48630)	(0.21106)	(0.81470)	(0.60872)
		1.7	0.16943	0.08002	0.37477	0.25788
			(0.15303)	(0.07182)	(0.34927)	(0.23157)
		Croux	0.00151	0.00106	0.02694	0.03201
			(0.00081)	(0.00048)	(0.02017)	(0.02320)

TABLE 2.5: Squared bias and MSE of the estimates at Model 2 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.00042	0.00045	0.02476	0.02020
			(0.00015)	(0.00039)	(0.01704)	(0.01360)
		DPD (α)				
		0.1	0.00025	0.00056	0.02563	0.02075
			(0.00007)	(0.00022)	(0.01865)	(0.01433)
		0.2	0.00031	0.00069	0.02786	0.02317
			(0.00011)	(0.00030)	(0.02037)	(0.01615)
		0.3	0.00038	0.00084	0.02942	0.02502
			(0.00016)	(0.00041)	(0.02167)	(0.01779)
		0.5	0.00061	0.00132	0.03284	0.02955
			(0.00032)	(0.00073)	(0.02432)	(0.02130)
		0.8	0.00115	0.00245	0.03912	0.03824
			(0.00066)	(0.00142)	(0.02875)	(0.02713)
		1.0	0.00159	0.00331	0.04331	0.04392
			(0.00092)	(0.00191)	(0.03154)	(0.03056)
		Iannario (c)				
		1.1	0.03232	0.38755	0.08256	0.43027
			(0.02757)	(0.38295)	(0.06176)	(0.41408)
		1.4	0.01624	0.17879	0.05374	0.21573
			(0.01395)	(0.17036)	(0.04071)	(0.19659)
		1.5	0.01333	0.13928	0.04923	0.17448
			(0.01136)	(0.13169)	(0.03712)	(0.15653)
		1.7	0.00889	0.08394	0.04209	0.11527
			(0.00751)	(0.07992)	(0.03196)	(0.10343)
		Croux	0.00039	0.00088	0.02829	0.02356
			(0.00012)	(0.00035)	(0.02071)	(0.01596)

TABLE 2.6: Squared bias and MSE of the estimates at Model 2 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.00023	0.00034	0.04561	0.03962
			(0.00019)	(0.00056)	(0.03116)	(0.02662)
		DPD (α)				
		0.1	0.00026	0.00030	0.04841	0.03622
			(0.00008)	(0.00027)	(0.03660)	(0.02504)
		0.2	0.00033	0.00042	0.05294	0.03936
			(0.00012)	(0.00033)	(0.04090)	(0.02722)
		0.3	0.00043	0.00057	0.05561	0.04169
			(0.00018)	(0.00044)	(0.04312)	(0.02915)
		0.5	0.00066	0.00096	0.06008	0.04632
			(0.00031)	(0.00072)	(0.04670)	(0.03277)
		0.8	0.00122	0.00185	0.06773	0.05420
			(0.00062)	(0.00129)	(0.05227)	(0.03865)
		1.0	0.00169	0.00251	0.07328	0.05925
			(0.00088)	(0.00169)	(0.05609)	(0.04239)
		Iannario (c)				
		0.6	0.00054	0.00046	0.05951	0.04641
			(0.00021)	(0.00042)	(0.04603)	(0.03200)
		0.8	0.00050	0.00046	0.05724	0.04394
			(0.00019)	(0.00038)	(0.04412)	(0.03038)
		0.9	0.00050	0.00048	0.05654	0.04309
			(0.00019)	(0.00037)	(0.04349)	(0.02985)
		1.0	0.00049	0.00049	0.05601	0.04240
			(0.00019)	(0.00036)	(0.04307)	(0.02948)
		Croux	0.00045	0.00057	0.05665	0.04031
			(0.00018)	(0.00044)	(0.04358)	(0.02811)

TABLE 2.7: Squared bias and MSE of the estimates at Model 2 with the Cauchy link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	Cauchy	MLE	0.00025	0.00044	0.04725	0.04025
			(0.00010)	(0.00006)	(0.03468)	(0.02549)
		DPD (α)				
		0.1	0.00118	0.00118	0.06698	0.05642
			(0.00071)	(0.00052)	(0.04670)	(0.03524)
		0.2	0.00236	0.00252	0.08108	0.07130
			(0.00138)	(0.00111)	(0.05670)	(0.04457)
		0.3	0.00312	0.00343	0.08731	0.07842
			(0.00180)	(0.00155)	(0.06129)	(0.04929)
		0.5	0.00450	0.00517	0.09456	0.08664
			(0.00257)	(0.00235)	(0.06628)	(0.05414)
		0.8	0.00676	0.00793	0.10515	0.09820
			(0.00374)	(0.00356)	(0.07282)	(0.05975)
		1.0	0.00824	0.00962	0.11165	0.10468
			(0.00451)	(0.00431)	(0.07680)	(0.06272)
		Iannario (c)				
		0.6	0.00316	0.00350	0.09661	0.09002
			(0.00187)	(0.00175)	(0.06960)	(0.06131)
		0.8	0.00311	0.00339	0.09288	0.08628
			(0.00183)	(0.00165)	(0.06642)	(0.05806)
		0.9	0.00308	0.00331	0.09188	0.08490
			(0.00183)	(0.00163)	(0.06548)	(0.05688)
		1.0	0.00305	0.00325	0.09114	0.08385
			(0.00184)	(0.00161)	(0.06495)	(0.05604)
		Croux	0.00289	0.00306	0.09302	0.08199
			(0.00185)	(0.00161)	(0.06625)	(0.05269)

TABLE 2.8: Squared bias and MSE of the estimates at Model 3 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.00199	0.00080	0.05187	0.02185
			(0.00078)	(0.00036)	(0.03272)	(0.01450)
		DPD (α)				
		0.1	0.00346	0.00155	0.06276	0.02679
			(0.00167)	(0.00077)	(0.04140)	(0.01863)
		0.2	0.00471	0.00206	0.07101	0.02997
			(0.00233)	(0.00100)	(0.04664)	(0.02072)
		0.3	0.00591	0.00260	0.07847	0.03309
			(0.00280)	(0.00117)	(0.05091)	(0.02265)
		0.5	0.00941	0.00416	0.09728	0.04124
			(0.00403)	(0.00164)	(0.06038)	(0.02695)
		0.8	0.01579	0.00710	0.12596	0.05459
			(0.00629)	(0.00254)	(0.07452)	(0.03374)
		1.0	0.01859	0.00850	0.13750	0.06075
			(0.00728)	(0.00296)	(0.08038)	(0.03692)
		Iannario (c)				
		1.1	0.00805	0.01269	0.48283	0.09817
			(0.01225)	(0.01653)	(1.20809)	(0.14244)
		1.4	0.00432	0.00739	0.12354	0.04234
			(0.00289)	(0.00593)	(0.04192)	(0.03300)
		1.5	0.00208	0.00579	0.03898	0.03495
			(0.00263)	(0.00522)	(0.41057)	(0.05300)
		1.7	0.00091	0.00338	0.02661	0.02187
			(0.00239)	(0.00372)	(0.30853)	(0.02478)
		Croux	0.00572	0.00246	0.07450	0.03049
			(0.00323)	(0.00144)	(0.05025)	(0.02161)

TABLE 2.9: Squared bias and MSE of the estimates at Model 3 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.00169	0.00041	0.07006	0.03587
			(0.00039)	(0.00020)	(0.04925)	(0.02389)
		DPD (α)				
		0.1	0.00247	0.00085	0.08368	0.03928
			(0.00124)	(0.00068)	(0.06127)	(0.02901)
		0.2	0.00380	0.00132	0.09504	0.04355
			(0.00195)	(0.00096)	(0.06969)	(0.03227)
		0.3	0.00496	0.00176	0.10218	0.04681
			(0.00247)	(0.00118)	(0.07434)	(0.03443)
		0.5	0.00774	0.00281	0.11874	0.05477
			(0.00359)	(0.00165)	(0.08417)	(0.03909)
		0.8	0.01324	0.00493	0.14840	0.06914
			(0.00571)	(0.00252)	(0.10053)	(0.04693)
		1.0	0.01644	0.00615	0.16388	0.07690
			(0.00692)	(0.00300)	(0.10902)	(0.05129)
		Iannario (c)				
		0.6	0.00003	0.00001	0.02027	0.01395
			(0.00002)	(0.00002)	(0.01596)	(0.01156)
		0.8	0.00001	0.00000	0.01770	0.01381
			(0.00002)	(0.00003)	(0.01294)	(0.00932)
		0.9	0.00002	0.00004	0.01857	0.01221
			(0.00000)	(0.00002)	(0.01248)	(0.00925)
		1.0	0.00003	0.00001	0.01763	0.01311
			(0.00001)	(0.00000)	(0.01210)	(0.00864)
		Croux	0.00406	0.00138	0.10083	0.04337
			(0.00233)	(0.00113)	(0.07433)	(0.03232)

TABLE 2.10: Squared bias and MSE of the estimates at Model 4 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.01279	0.00280	0.19448	0.06218
			(0.00373)	(0.00105)	(0.13881)	(0.04185)
		DPD (α)				
		0.1	0.08153	0.01777	0.47914	0.12899
			(0.03230)	(0.00709)	(0.27740)	(0.07516)
		0.2	0.09886	0.02146	0.46427	0.12885
			(0.04529)	(0.00973)	(0.30960)	(0.08144)
		0.3	0.11786	0.02552	0.51792	0.14103
			(0.05616)	(0.01215)	(0.35480)	(0.09195)
		0.5	0.17194	0.03662	0.67212	0.17492
			(0.09193)	(0.01939)	(0.52375)	(0.12404)
		0.8	0.29509	0.06185	0.98384	0.24715
			(0.15115)	(0.03113)	(0.65670)	(0.15294)
		1.0	0.38689	0.08023	1.16475	0.28748
			(0.19555)	(0.04004)	(0.76913)	(0.17819)
		Iannario (c)				
		1.1	0.00216	0.00190	0.06028	0.03010
			(0.00253)	(0.00167)	(0.05822)	(0.02531)
		1.4	0.00178	0.00116	0.05012	0.02313
			(0.00112)	(0.00101)	(0.03588)	(0.01929)
		1.5	0.00096	0.00104	0.04392	0.02456
			(0.00191)	(0.00111)	(0.03881)	(0.02191)
		1.7	0.00167	0.00070	0.51388	0.03080
			(0.00180)	(0.00100)	(0.07450)	(0.02285)
		Croux	0.48349	0.09860	2.24010	0.50087
			(0.18440)	(0.03642)	(0.85851)	(0.18960)

TABLE 2.11: Squared bias and MSE of the estimates at Model 4 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.01096	0.00198	0.21668	0.08238
			(0.00388)	(0.00058)	(0.12579)	(0.05433)
		DPD (α)				
		0.1	0.04325	0.00772	0.46626	0.12697
			(0.01336)	(0.00249)	(0.27120)	(0.07975)
		0.2	0.06967	0.01231	0.58611	0.15233
			(0.02838)	(0.00522)	(0.36083)	(0.09750)
		0.3	0.08707	0.01540	0.64092	0.16660
			(0.03970)	(0.00722)	(0.41894)	(0.11005)
		0.5	0.11184	0.01971	0.70599	0.18387
			(0.05778)	(0.01049)	(0.47639)	(0.12491)
		0.8	0.17972	0.03152	0.91572	0.23050
			(0.09571)	(0.01728)	(0.59872)	(0.15333)
		1.0	0.21797	0.03833	1.03225	0.25747
			(0.11784)	(0.02115)	(0.66003)	(0.16696)
		Iannario (c)				
		0.6	0.00003	0.00006	0.02826	0.01663
			(0.00008)	(0.00003)	(0.02277)	(0.01316)
		0.8	0.00000	0.00001	0.02799	0.01714
			(0.00000)	(0.00003)	(0.01618)	(0.01186)
		0.9	0.00001	0.00004	0.02604	0.01593
			(0.00001)	(0.00001)	(0.02043)	(0.01224)
		1.0	0.00004	0.00004	0.03388	0.01457
			(0.00005)	(0.00001)	(0.02539)	(0.01252)
		Croux	0.11632	0.02018	0.80170	0.18716
			(0.04683)	(0.00801)	(0.48515)	(0.11748)

TABLE 2.12: Squared bias and MSE of the estimates at Model 5 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.00161	0.00121	0.04809	0.05365
			(0.00105)	(0.00084)	(0.03407)	(0.03683)
		DPD (α)				
		0.1	0.00322	0.00265	0.06178	0.06419
			(0.00178)	(0.00140)	(0.04327)	(0.04704)
		0.2	0.00462	0.00378	0.07189	0.07281
			(0.00279)	(0.00222)	(0.05093)	(0.05429)
		0.3	0.00588	0.00483	0.08084	0.08095
			(0.00355)	(0.00283)	(0.05665)	(0.05989)
		0.5	0.00947	0.00787	0.10178	0.10119
			(0.00573)	(0.00463)	(0.06959)	(0.07334)
		0.8	0.01763	0.01486	0.13967	0.13986
			(0.01019)	(0.00837)	(0.09075)	(0.09651)
		1.0	0.02097	0.01575	0.14413	0.15493
			(0.01231)	(0.01037)	(0.09679)	(0.10461)
		Iannario (c)				
		1.1	0.19446	0.18837	0.31018	0.32085
			(0.17076)	(0.17003)	(0.25177)	(0.25863)
		1.4	0.12921	0.11254	0.22900	0.22789
			(0.11128)	(0.09657)	(0.18175)	(0.17693)
		1.5	0.11084	0.09426	0.20694	0.20676
			(0.09681)	(0.08167)	(0.16588)	(0.16155)
		1.7	0.08487	0.06983	0.17836	0.17859
			(0.07212)	(0.05832)	(0.13899)	(0.13638)
		Croux	0.00661	0.00500	0.08345	0.07720
			(0.00360)	(0.00271)	(0.06036)	(0.05544)

TABLE 2.13: Squared bias and MSE of the estimates at Model 5 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.00153	0.00072	0.07735	0.08172
			(0.00086)	(0.00055)	(0.05689)	(0.05612)
		DPD (α)				
		0.1	0.00187	0.00136	0.09672	0.10823
			(0.00159)	(0.00124)	(0.07070)	(0.08210)
		0.2	0.00313	0.00232	0.11330	0.12600
			(0.00267)	(0.00202)	(0.08381)	(0.09547)
		0.3	0.00426	0.00318	0.12227	0.13521
			(0.00368)	(0.00279)	(0.09136)	(0.10193)
		0.5	0.00709	0.00530	0.14168	0.15560
			(0.00587)	(0.00441)	(0.10587)	(0.11485)
		0.8	0.01389	0.01027	0.18022	0.19735
			(0.01050)	(0.00776)	(0.13219)	(0.13879)
		1.0	00.02191	0.01414	0.20490	0.21680
			(0.01413)	(0.00989)	(0.14936)	(0.15496)
		Iannario (c)				
		0.6	0.00302	0.00200	0.13200	0.14737
			(0.00264)	(0.00199)	(0.09823)	(0.11466)
		0.8	0.00301	0.00206	0.12474	0.13998
			(0.00264)	(0.00203)	(0.09317)	(0.10890)
		0.9	0.00305	0.00212	0.12273	0.13765
			(0.00265)	(0.00205)	(0.09162)	(0.10701)
		1.0	0.00309	0.00219	0.12132	0.13592
			(0.00269)	(0.00207)	(0.09054)	(0.10544)
		Croux	0.00540	0.00293	0.15006	0.13236
			(0.00326)	(0.00231)	(0.11015)	(0.10165)

TABLE 2.14: Squared bias and MSE of the estimates at Model 5 with the complementary log-log link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	log-log	MLE	0.00220	0.00136	0.05270	0.05933
			(0.00104)	(0.00054)	(0.03814)	(0.04114)
		DPD (α)				
		0.1	0.00430	0.00287	0.06963	0.07802
			(0.00237)	(0.00142)	(0.05123)	(0.05546)
		0.2	0.00611	0.00428	0.08108	0.08861
			(0.00333)	(0.00217)	(0.05868)	(0.06319)
		0.3	0.00749	0.00539	0.08978	0.09701
			(0.00397)	(0.00269)	(0.06382)	(0.06859)
		0.5	0.01250	0.00930	0.11349	0.12154
			(0.00599)	(0.00422)	(0.07664)	(0.08228)
		0.8	0.02330	0.01753	0.15353	0.16217
			(0.01023)	(0.00741)	(0.09707)	(0.10366)
		1.0	0.03010	0.02260	0.17581	0.18408
			(0.01275)	(0.00929)	(0.10739)	(0.11486)
		Iannario (c)				
		1.1	0.21656	0.18125	0.44069	0.42720
			(0.18602)	(0.15414)	(0.36320)	(0.33341)
		1.4	0.12067	0.10590	0.26816	0.27908
			(0.09209)	(0.08140)	(0.18498)	(0.19263)
		1.5	0.10448	0.09187	0.23472	0.25058
			(0.08140)	(0.07163)	(0.16667)	(0.17489)
		1.7	0.08478	0.07408	0.20196	0.21348
			(0.06634)	(0.05797)	(0.14491)	(0.15192)
		Croux	0.00776	0.00472	0.09849	0.09576
			(0.00525)	(0.00334)	(0.07474)	(0.07096)

Unless stated otherwise, all the links refer to the standard ones. To be able to apply the theory developed earlier, X must be non-stochastic. A justification is therefore required for these above models, which involve random covariates, to bring them into the realm of our proposed theory. In all such models, we assume that the values of X are fixed at the values generated by the aforesaid distributions, and the values of Y depend on the generated (fixed) values of X.

We find that squared biases (with one exception at $\alpha = 0$; see Table 2.3 for $||\hat{\beta} - \beta||^2$) and MSEs decrease as the sample size increases. Now, we make the following remarks based on the simulation studies of the pure models.

- Highest efficiency is achieved at pure models when $\alpha = 0$ (MLE) for all these link functions. As α increases, there is a drop in the efficiencies of the MDPD estimators. However, when $\alpha = 0.1$, the loss of efficiency for using the MDPDE is roughly within 10% for Model 1 - Model 3 and Model 5. However, it drops around 50% in Model 4. These observations are clear in Figures 2.10 to 2.13.

- Generally the logit link gives better efficiency than the probit link which in turn produces more efficient estimates than the complementary log-log link at a fixed value of α. Also, the probit link yields better efficiency than the Cauchy link. However in Model 4, we notice that the probit link dominates (in the sense of better efficiency) the logit link up to the point around $\alpha = 0.37$ ($n = 150$) and $\alpha = 0.45$ ($n = 200$). After that, the usual pattern follows. This is observed in Figure 2.10 and Figure 2.11. Also Figure 2.13 shows that the probit link dominates the complementary log-log link that in turn dominates the logit link up to a point close to $\alpha = 0.45$, and thereafter the order changes when the sample size is 200. A similar thing is also noticed in Figure 2.12 when the sample size is 150 with different order of domination.

- MDPDE performs better (in terms of lower MSE) than both Croux et al. (2013) and Iannario et al. (2017) estimators in Model 1, Model 2, Model 3 (only with probit link) and Model 5. However, Iannario et al. (2017) perform slightly better than the MDPDE in Model 3 (only with logit link) and Model 4 (only with probit, logit links). For that, see Tables 2.3 - 2.14.

2.7.2 Simulation Studies: Contaminated Models

Now the above models are contaminated in the following way.

Vertical outliers: In this chapter, a data point is said to be a vertical outlier when Y takes the highest categorical value in a manner such that it is inconsistent with the covariates. We add 5% and 10% vertical outliers to the data sets generated by Model 1 to Model 5 across all the link functions under consideration.

Horizontal outliers: We have developed the theory where the covariates are assumed to be non-stochastic. However, we might still come across a situation where a small proportion of x-values with high magnitudes may destabilise the MLE. We refer to these x-values as horizontal outliers because they do not naturally correspond to the ordinal responses. Let the covariate in the data sets, simulated through Model 2 with the probit link, be contaminated with 5% – 10% horizontal outliers, where the outlying values are chosen to be 5.

Misspecification of links: Now data sets are generated through Model 1 using the probit link. However, the complementary log-log is used in estimation. The results are reported in Table 2.24.

The numerical results for contaminated models are summarized in the following remarks.

- In contaminated models MDPDEs with $\alpha > 0$ perform better than the MLE. As α increases its efficiency increases up to a point, then sometimes it decreases slowly. But all the way, it performs much better than the MLE. In Figures 2.14 - 2.17 and Tables 2.15 - 2.22, we notice that the complimentary log-log link performs better than the probit link in Model 1. Also in Model 2, we find that the probit link yields better efficiency than the logit which in turn holds an edge over the Cauchy link; but, at lower values of the tuning parameter, the Cauchy link dominates the logit link. In both Model 3 and Model 4 the probit link performs better than the logit link. In Model 5 the complementary log-log link performs better than the probit link which performs better than the logit link.

- Let the data sets generated through Model 2 be contaminated by horizontal outliers as described before. The graphs of the efficiency are plotted for both the 5% and 10% level of contamination in Figure 2.18 and Figure 2.19. Also, we report the squared bias and squared MSE in Table 2.23 only for the 5% data contamination. We find that the estimated efficiency improves with the increment of α; after some point, it sharply increases and reaches its maximum at some point between $\alpha = 0.25$ and $\alpha = 0.4$. After that, the estimated efficiency slowly drops, but still stays higher than the MLE. Also, the MDPDE performs better than the M-estimates of Iannario et al. (2017).

- When comparing the MSE, we see that MDPDE performs better than both the Croux et al. (2013) and the Iannario et al. (2017) in– Model 1 with the probit link, Model 2 with the Cauchy and Model 5 with the complementary log-log link. Also, MDPDE performs better in the case when the probit link is misspecified in Model 1. Although the Iannario et al. (2017) beats the MDPDE in Model 3 with the logit link, our method performs better than the Croux et al. (2013) method.

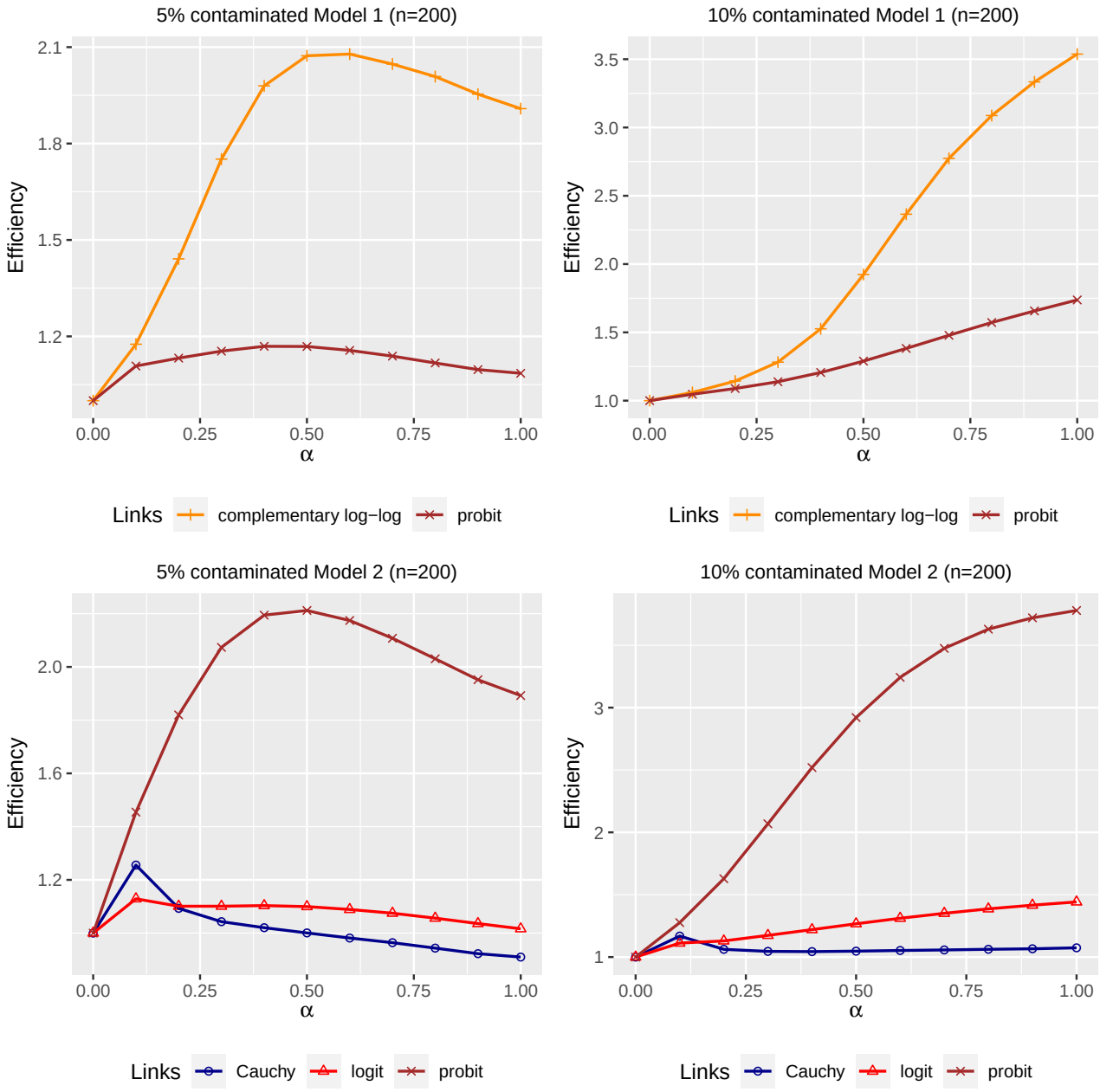

FIGURE 2.14: Graphs of efficiency when Model 1 and Model 2 are contaminated at 5% and 10% levels of contamination.

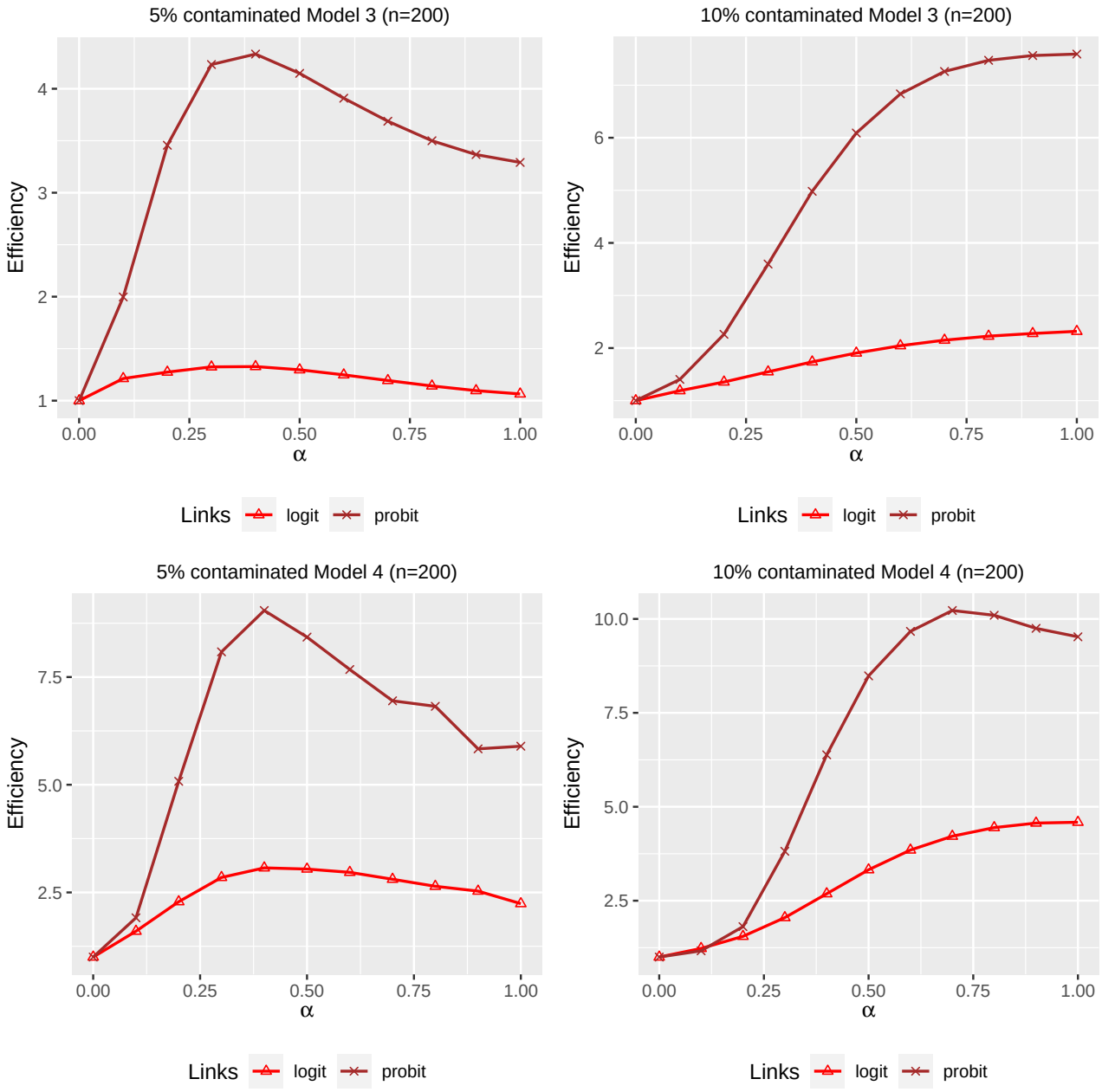

FIGURE 2.15: Graphs of efficiency when Model 3 and Model 4 are contaminated at 5% and 10% levels of contamination.

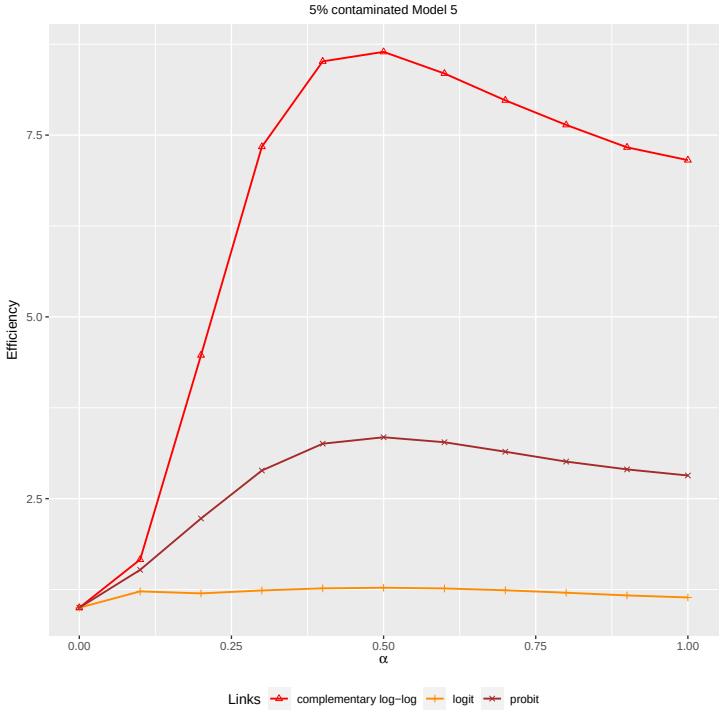

FIGURE 2.16: Graphs of efficiency when Model 5 is contaminated at 5% level of contamination.

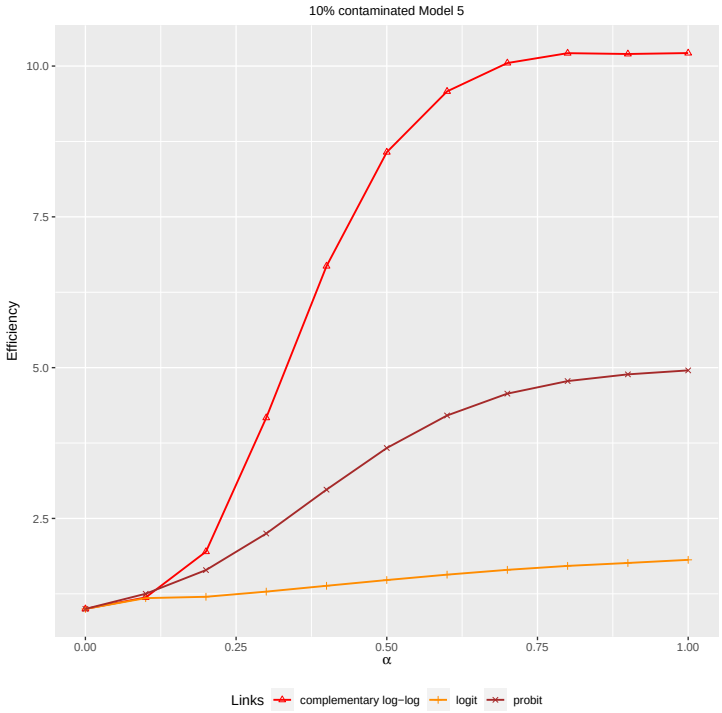

FIGURE 2.17: Graphs of efficiency when Model 5 is contaminated at 10% level of contamination.

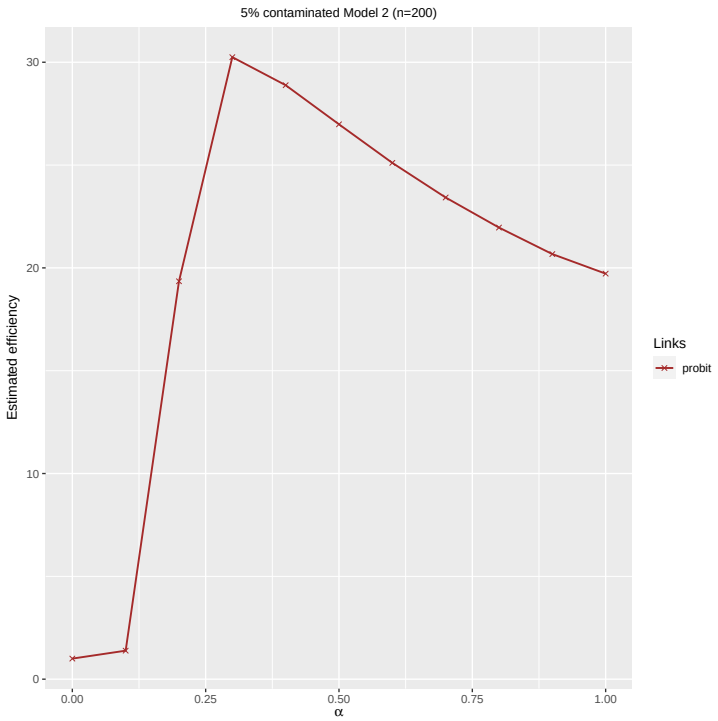

FIGURE 2.18: Graphs of efficiency when data generated by Model 2 is horizontally contaminated at 5% level of contamination.

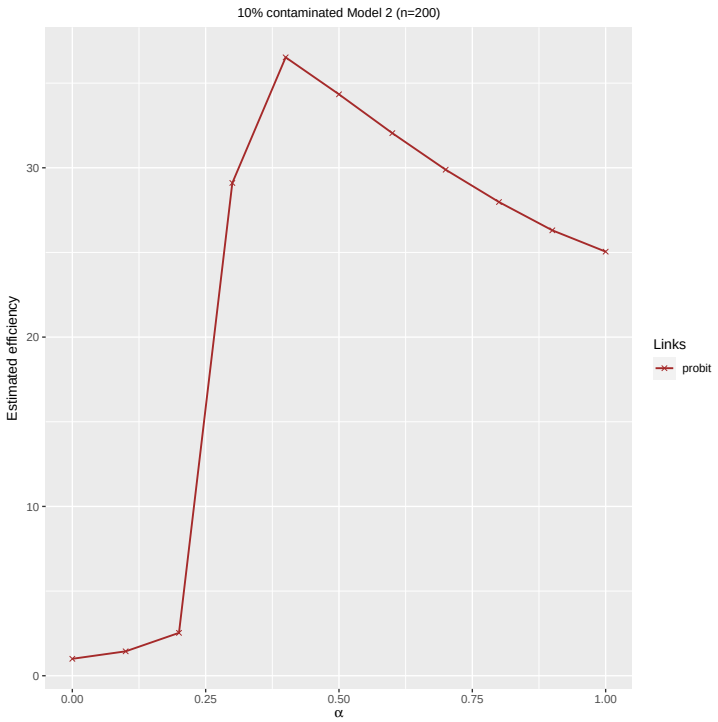

FIGURE 2.19: Graphs of efficiency when data generated by Model 2 is horizontally contaminated at 10% level of contamination.

TABLE 2.15: Squared bias and MSE when 5% vertical outliers are added to data generated by Model 1 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.01836	0.01357	0.03743	0.03508
			(0.02104)	(0.01564)	(0.03319)	(0.03140)
		DPD (α)				
		0.1	0.01550	0.01123	0.03274	0.03271
			(0.01749)	(0.01306)	(0.02982)	(0.02909)
		0.2	0.01353	0.00971	0.03193	0.03210
			(0.01550)	(0.01150)	(0.02854)	(0.02798)
		0.3	0.01118	0.00792	0.03095	0.03188
			(0.01301)	(0.00961)	(0.02706)	(0.02708)
		0.5	0.00691	0.00454	0.02969	0.03237
			(0.00833)	(0.00595)	(0.02474)	(0.02590)
		0.8	0.00312	0.00159	0.02941	0.03549
			(0.00411)	(0.00252)	(0.02334)	(0.02650)
		1.0	0.00201	0.00076	0.02912	0.03770
			(0.00283)	(0.00140)	(0.02283)	(0.02749)
		Iannario (c)				
		1.1	0.65004	0.33193	0.70072	0.39141
			(0.62722)	(0.32305)	(0.66563)	(0.36653)
		1.4	0.35808	0.22975	0.38934	0.26487
			(0.34148)	(0.22876)	(0.36280)	(0.25362)
		1.5	0.23591	0.11497	0.29638	0.25526
			(0.22664)	(0.11298)	(0.29271)	(0.26179)
		1.7	0.06634	0.02597	0.17265	0.23807
			(0.06072)	(0.02308)	(0.10544)	(0.07125)
		Croux	0.01827	0.01364	0.03867	0.03527
			(0.02106)	(0.01566)	(0.03327)	(0.03142)

TABLE 2.16: Squared bias and MSE when 10% vertical outliers are added to data generated by Model 1 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.06373	0.04571	0.08052	0.06505
			(0.06409)	(0.04689)	(0.07534)	(0.06159)
		DPD (α)				
		0.1	0.05969	0.04291	0.07531	0.06284
			(0.05964)	(0.04406)	(0.07144)	(0.05937)
		0.2	0.05647	0.04049	0.07276	0.06100
			(0.05631)	(0.04159)	(0.06845)	(0.05731)
		0.3	0.05235	0.03734	0.06950	0.05920
			(0.05224)	(0.03850)	(0.06508)	(0.05520)
		0.5	0.04207	0.02859	0.06174	0.05434
			(0.04206)	(0.02989)	(0.05699)	(0.04921)
		0.8	0.02725	0.01523	0.05101	0.04892
			(0.02772)	(0.01682)	(0.04594)	(0.04116)
		1.0	0.02073	0.00892	0.04599	0.04733
			(0.02155)	(0.01064)	(0.04087)	(0.03795)
		Iannario (c)				
		1.1	0.50295	0.28278	0.54448	0.32853
			(0.50931)	(0.27929)	(0.67603)	(0.36968)
		1.4	0.25635	0.17064	0.28689	0.21103
			(0.25041)	(0.16952)	(0.27327)	(0.20127)
		1.5	0.15558	0.08586	0.19710	0.14595
			(0.15703)	(0.08411)	(0.23498)	(0.14650)
		1.7	0.04451	0.02526	0.09104	0.07660
			(0.04461)	(0.02489)	(0.08272)	(0.06596)
		Croux	0.06351	0.04584	0.08197	0.06534
			(0.06414)	(0.04693)	(0.07545)	(0.06162)

TABLE 2.17: Squared bias and MSE when 5% vertical outliers are added to data generated by Model 2 with the Cauchy link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	Cauchy	MLE	0.01566	0.00444	0.08885	0.07220
			(0.01983)	(0.00782)	(0.07032)	(0.05127)
		DPD (α)				
		0.1	0.01319	0.00327	0.07317	0.05511
			(0.01572)	(0.00511)	(0.05779)	(0.03834)
		0.2	0.01347	0.00375	0.08207	0.06529
			(0.01695)	(0.00665)	(0.06449)	(0.04600)
		0.3	0.01262	0.00349	0.08502	0.06942
			(0.01638)	(0.00675)	(0.06638)	(0.04893)
		0.5	0.01056	0.00251	0.08754	0.07341
			(0.01445)	(0.00603)	(0.06761)	(0.05142)
		0.8	0.00788	0.00125	0.09167	0.07902
			(0.01166)	(0.00469)	(0.06930)	(0.05382)
		1.0	0.00652	0.00070	0.09462	0.08237
			(0.01007)	(0.00388)	(0.07058)	(0.05505)
		Iannario (c)				
		0.6	0.01963	0.00362	0.09957	0.08104
			(0.02450)	(0.00670)	(0.08074)	(0.06024)
		0.8	0.01777	0.00373	0.09455	0.07745
			(0.02232)	(0.00697)	(0.07586)	(0.05735)
		0.9	0.01715	0.00386	0.09305	0.07618
			(0.02156)	(0.00710)	(0.07425)	(0.05621)
		1.0	0.01668	0.00397	0.09193	0.07512
			(0.02100)	(0.00724)	(0.07318)	(0.05538)
		Croux	0.01604	0.00448	0.09273	0.07349
			(0.02022)	(0.00790)	(0.07326)	(0.05207)

TABLE 2.18: Squared bias and MSE when 10% vertical outliers are added to data generated by Model 2 with the Cauchy link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	Cauchy	MLE	0.08868	0.03983	0.14961	0.09796
			(0.09109)	(0.09109)	(0.13357)	(0.08222)
		DPD (α)				
		0.1	0.07571	0.02919	0.13245	0.08109
			(0.07598)	(0.03074)	(0.11735)	(0.06736)
		0.2	0.08032	0.03608	0.14006	0.09224
			(0.08214)	(0.03922)	(0.12509)	(0.07813)
		0.3	0.07940	0.03724	0.14091	0.09572
			(0.08195)	(0.04126)	(0.12531)	(0.08110)
		0.5	0.07461	0.03584	0.13873	0.09752
			(0.07816)	(0.04119)	(0.12315)	(0.08288)
		0.8	0.06658	0.03199	0.13573	0.09878
			(0.07139)	(0.03877)	(0.11955)	(0.08359)
		1.0	0.06146	0.02922	0.13415	0.09925
			(0.06694)	(0.03665)	(0.11743)	(0.08349)
		Iannario (c)				
		0.6	0.09957	0.03679	0.16811	0.10414
			(0.10203)	(0.04039)	(0.15051)	(0.08813)
		0.8	0.09401	0.03748	0.15930	0.10102
			(0.09652)	(0.04144)	(0.14224)	(0.08605)
		0.9	0.09227	0.03803	0.15655	0.10015
			(0.09469)	(0.04199)	(0.13954)	(0.08532)
		1.0	0.09099	0.03853	0.15444	0.09946
			(0.09343)	(0.04252)	(0.13771)	(0.08487)
		Croux	0.08982	0.04025	0.15346	0.09912
			(0.09211)	(0.04424)	(0.13672)	(0.08328)

TABLE 2.19: Squared bias and MSE when 5% vertical outliers are added to data generated by Model 3 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.09349	0.01687	0.15705	0.05238
			(0.11243)	(0.02058)	(0.16154)	(0.04721)
		DPD (α)				
		0.1	0.06391	0.00938	0.12838	0.04431
			(0.07868)	(0.01209)	(0.12957)	(0.03878)
		0.2	0.04892	0.00627	0.12044	0.04389
			(0.06251)	(0.00869)	(0.11766)	(0.03702)
		0.3	0.03575	0.00368	0.11397	0.04413
			(0.04794)	(0.00578)	(0.10756)	(0.03570)
		0.5	0.01934	0.00095	0.11330	0.04819
			(0.02909)	(0.00245)	(0.09932)	(0.03669)
		0.8	0.00908	0.00001	0.12563	0.05788
			(0.01642)	(0.00074)	(0.10112)	(0.04144)
		1.0	0.00656	0.00005	0.13355	0.06321
			(0.01278)	(0.00041)	(0.10397)	(0.04440)
		Iannario (c)				
		0.6	0.00097	0.00014	0.02866	0.01818
			(0.00136)	(0.00009)	(0.02427)	(0.01593)
		0.8	0.00079	0.00009	0.02593	0.01637
			(0.00094)	(0.00017)	(0.02294)	(0.01450)
		0.9	0.00086	0.00012	0.02390	0.01663
			(0.00129)	(0.00013)	(0.02367)	(0.01551)
		1.0	0.00078	0.00013	0.02415	0.01577
			(0.00090)	(0.00019)	(0.02246)	(0.01476)
		Croux	0.08967	0.01627	0.15754	0.05202
			(0.11268)	(0.02064)	(0.16526)	(0.04740)

TABLE 2.20: Squared bias and MSE when 10% vertical outliers are added to data generated by Model 3 with the logit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	logit	MLE	0.36860	0.07404	0.41935	0.10829
			(0.37264)	(0.07511)	(0.41266)	(0.10100)
		DPD (α)				
		0.1	0.30176	0.05413	0.35734	0.08980
			(0.30591)	(0.05532)	(0.34978)	(0.08219)
		0.2	0.25484	0.04331	0.31373	0.08001
			(0.26081)	(0.04507)	(0.30684)	(0.07238)
		0.3	0.21064	0.03276	0.27549	0.07158
			(0.21814)	(0.03495)	(0.26821)	(0.06351)
		0.5	0.14538	0.01830	0.22511	0.06313
			(0.15577)	(0.02121)	(0.21604)	(0.05351)
		0.8	0.09268	0.00860	0.19391	0.06302
			(0.10601)	(0.01207)	(0.18052)	(0.05021)
		1.0	0.07560	0.00618	0.18670	0.06528
			(0.08997)	(0.00980)	(0.17075)	(0.05084)
		Iannario (c)				
		0.6	0.00551	0.00069	0.05755	0.03125
			(0.00700)	(0.00063)	(0.05788)	(0.02746)
		0.8	0.00568	0.00059	0.05680	0.02608
			(0.00731)	(0.00078)	(0.05857)	(0.03094)
		0.9	0.00719	0.00123	0.06402	0.03791
			(0.00639)	(0.00077)	(0.05060)	(0.02869)
		1.0	0.00673	0.00073	0.05699	0.02974
			(0.00645)	(0.00059)	(0.05413)	(0.02740)
		Croux	0.35907	0.07206	0.41423	0.10671
			(0.37283)	(0.07520)	(0.41637)	(0.10117)

TABLE 2.21: Squared bias and MSE when 5% vertical outliers are added to data generated by Model 5 with the complementary log-log link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	log-log	MLE	0.67107	0.34065	0.75620	0.46745
			(0.79469)	(0.41913)	(0.84674)	(0.50749)
		DPD (α)				
		0.1	0.35965	0.16504	0.45120	0.27767
			(0.44723)	(0.21468)	(0.51548)	(0.29913)
		0.2	0.08589	0.03262	0.17626	0.12863
			(0.11621)	(0.04806)	(0.18499)	(0.11780)
		0.3	0.02325	0.00645	0.11282	0.10273
			(0.03735)	(0.01254)	(0.10295)	(0.08154)
		0.5	0.00233	0.00113	0.10605	0.11394
			(0.00784)	(0.00131)	(0.07781)	(0.07884)
		0.8	0.00112	0.00454	0.12674	0.14493
			(0.00167)	(0.00039)	(0.08337)	(0.09388)
		1.0	0.00202	0.00619	0.13590	0.15810
			(0.00091)	(0.00059)	(0.08772)	(0.10153)
		Iannario (c)				
		1.1	0.03569	0.07658	0.28077	0.31508
			(0.02280)	(0.05765)	(0.21503)	(0.23128)
		1.4	0.00845	0.03774	0.20197	0.22277
			(0.00388)	(0.02508)	(0.15595)	(0.16133)
		1.5	0.00433	0.02996	0.18551	0.20015
			(0.00220)	(0.01870)	(0.14522)	(0.14665)
		1.7	00.00096	0.01992	0.17543	0.18171
			(0.00175)	(0.01341)	(0.15017)	(0.14232)
		Croux	0.74334	0.39614	0.81978	0.50701
			(0.85310)	(0.46098)	(0.89923)	(0.53467)

TABLE 2.22: Squared bias and MSE when 10% vertical outliers are added to data generated by Model 5 with the complementary log-log link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	log-log	MLE	1.35901	0.67394	1.39573	0.75649
			(1.39844)	(0.70918)	(1.42347)	(0.76895)
		DPD (α)				
		0.1	1.12607	0.53678	1.17625	0.63136
			(1.17012)	(0.56997)	(1.20455)	(0.63803)
		0.2	0.64232	0.27279	0.73806	0.39434
			(0.67148)	(0.29143)	(0.74338)	(0.38151)
		0.3	0.24846	0.08975	0.35056	0.20063
			(0.26717)	(0.10096)	(0.34647)	(0.17944)
		0.5	0.06191	0.01273	0.16822	0.12938
			(0.07849)	(0.01925)	(0.15566)	(0.10010)
		0.8	0.02076	0.00066	0.13952	0.13650
			(0.03506)	(0.00358)	(0.11800)	(0.09665)
		1.0	0.01543	0.00005	0.13761	0.14375
			(0.02927)	(0.00230)	(0.11410)	(0.10052)
		Iannario (c)				
		1.1	0.00817	0.03981	0.28002	0.27322
			(0.00939)	(0.03204)	(0.23622)	(0.21272)
		1.4	0.02240	0.02028	0.26673	0.22526
			(0.02133)	(0.01424)	(0.22134)	(0.16968)
		1.5	0.02654	0.01753	0.28295	0.23761
			(0.02542)	(0.01157)	(0.23144)	(0.16900)
		1.7	0.04521	0.01310	0.30673	0.22814
			(0.04570)	(0.01024)	(0.25892)	(0.16987)
		Croux	1.35182	0.67836	1.39762	0.75624
			(1.40714)	(0.71510)	(1.44104)	(0.77266)

TABLE 2.23: Squared bias and MSE when 5% horizontal outliers are added to data generated by Model 2 with the probit link

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	probit	MLE	0.17980	1.08089	0.19515	1.08790
			(0.18975)	(1.12790)	(0.20091)	(1.13275)
		DPD (α)				
		0.1	0.10358	0.62587	0.13361	0.71104
			(0.12947)	(0.76588)	(0.14917)	(0.81316)
		0.2	0.00011	0.00175	0.03179	0.05376
			(0.00036)	(0.00296)	(0.02494)	(0.04399)
		0.3	0.00028	0.00049	0.03098	0.02797
			(0.00010)	(0.00015)	(0.02335)	(0.02075)
		0.5	0.00063	0.00141	0.03461	0.03072
			(0.00033)	(0.00074)	(0.02626)	(0.02317)
		0.8	0.00122	0.00266	0.04123	0.03939
			(0.00072)	(0.00152)	(0.03122)	(0.02951)
		1.0	0.00168	0.00358	0.04569	0.04522
			(0.00101)	(0.00206)	(0.03434)	(0.03330)
		Iannario (c)				
		1.1	0.01437	0.17591	0.05819	0.21230
			(0.01111)	(0.16005)	(0.04167)	(0.18579)
		1.4	0.00419	0.05228	0.03682	0.08047
			(0.00290)	(0.04204)	(0.02671)	(0.06332)
		1.5	0.00267	0.03273	0.03350	0.05880
			(0.00179)	(0.02507)	(0.02463)	(0.04524)
		1.7	0.00091	0.01047	0.02978	0.03424
			(0.00059)	(0.00654)	(0.02201)	(0.02488)
		Croux	0.17578	1.07969	0.19163	1.08658
			(0.18979)	(1.12763)	(0.20143)	(1.13247)

TABLE 2.24: Squared bias and MSE when the probit link is misspecified with the complementary log-log link in Model 1

Sample size	Link	Method	$\|\hat{\gamma}-\gamma\|^2$	$\|\hat{\beta}-\beta\|^2$	$MSE(\hat{\gamma})$	$MSE(\hat{\beta})$
150 (200)	log-log	MLE	0.22796	0.16405	0.27993	0.20407
			(0.21333)	(0.16156)	(0.24882)	(0.18965)
		DPD (α)				
		0.1	0.19688	0.14474	0.24136	0.18171
			(0.18525)	(0.14289)	(0.21630)	(0.16870)
		0.2	0.20842	0.14284	0.25887	0.18398
			(0.19578)	(0.14061)	(0.22985)	(0.16835)
		0.3	0.21571	0.13956	0.27373	0.18954
			(0.20093)	(0.13634)	(0.23818)	(0.16785)
		0.5	0.23037	0.14024	0.29848	0.20461
			(0.21397)	(0.13506)	(0.26013)	(0.17922)
		0.8	0.23812	0.14576	0.30862	0.21919
			(0.22256)	(0.14158)	(0.26900)	(0.19087)
		1.0	0.23509	0.14644	0.30301	0.22220
			(0.22104)	(0.14322)	(0.26707)	(0.19585)
		Iannario (c)				
		1.1	4.07929	1.13771	4.27116	1.30679
			(3.85829)	(1.05397)	(4.00773)	(1.19122)
		1.4	2.22450	0.25767	2.45428	0.46093
			(2.09108)	(0.25334)	(2.24494)	(0.40160)
		1.5	0.86343	0.56844	1.02307	0.69084
			(0.79026)	(0.54682)	(0.89390)	(0.62439)
		1.7	0.80507	0.57208	0.92810	0.66357
			(0.74604)	(0.55233)	(0.81903)	(0.60822)
		Croux	0.22830	0.16435	0.28051	0.20447
			(0.21362)	(0.16199)	(0.24918)	(0.19011)

2.7.3 Comparison of the Computational Time

Next, we compare the elapsed time incurred in the computation of MDPDE and the M-estimates of Iannario et al. (2017) for Model 1 with the probit link. Computational time is noted, respectively, for the MDPDE with $\alpha = 0.5$ and the M-estimates with $c = 1.5$ at different replication (B) sizes. To minimize the sampling fluctuations, we repeat this experiment 50 times at each value of "B", and take the time average across 50 repetition. We find that the computation time depends linearly on the B for both these methods. However, as B increases, the difference in computational time between the MDPDE and the M-estimates widens; consequently, the M-estimates proposed by Iannario et al. (2017) become computationally much more expensive.

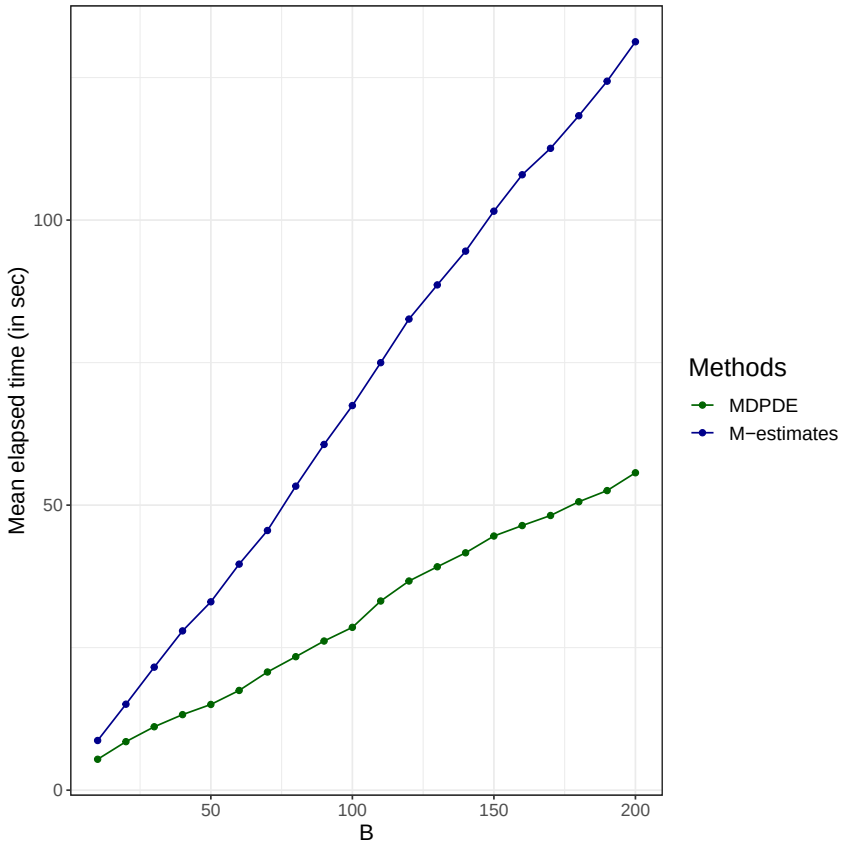

FIGURE 2.20: Comparison of time incurred in the computation of MDPDE and the M-estimates

2.8 Validation of Tuning Parameter Selection Strategy

In Table 2.25, we compare the replication-MSE and the squared bias of the MLEs with those of the MDPDEs, fitted through (1.75), for data sets generated through Model 3. In doing so, we have considered both $\hat{\theta}_{0.5}$ and $\hat{\theta}_1$ as pilot values to be used in (1.75). We see that the replication-MSE of the fitted MDPDEs corresponding to the pilot $\hat{\theta}_{0.5}$ gives a much better approximation of the replication-MSE of the MLE (smallest in this case) in pure data sets across two different links. Also, note that the squared biases due to the pilot $\hat{\theta}_1$ become almost twice the values generated by the MLE in pure data for both the probit and the logit links.

Now, we add 10% and 15% vertical outliers to these data sets. As we find, the fitted MDPDEs perform better than the MLE, and the pilot $\hat{\theta}_1$ gives the lowest replication-MSE and the squared bias. This gives a clear indication that the more robust we choose a pilot value, the fitted MDPDEs become more resistant to the outliers, but also they lose their efficiency along the way. Since $\hat{\theta}_{0.5}$ is highly robust (though less robust than $\hat{\theta}_1$) and they yield better-fitted MDPDEs in pure data, $\hat{\theta}_{0.5}$ chosen as a pilot would provide a nice trade-off between the efficiency and robustness. This conclusion also validates the empirical evidence of Ghosh and Basu (2015).

TABLE 2.25: Comparison of the replication-MSE and squared-bias (in bracket) between the MLE (first row for each link) and the fitted MDPDEs with $\hat{\theta}_{0.5}$ and $\hat{\theta}_1$ as pilots under different levels (ϵ) of data contamination in vertical direction

Model	Link	Pilot	$\epsilon = 0$	$\epsilon = 0.10$	$\epsilon = 0.15$
Model 3	probit		0.18278 (0.01067)	2.87686 (2.79551)	4.00026 (3.94187)
		$\hat{\theta}_{0.5}$	0.19797 (0.01283)	0.54000 (0.35334)	1.07291 (0.89671)
		$\hat{\theta}_1$	0.24429 (0.02210)	0.46164 (0.25809)	0.82525 (0.63395)
	logit		0.2782 (0.00514)	1.47196 (1.30487)	2.57257 (2.42587)
		$\hat{\theta}_{0.5}$	0.29862 (0.00528)	0.91008 (0.67627)	1.70114 (1.50296)
		$\hat{\theta}_1$	0.34358 (0.01027)	0.77661 (0.50389)	1.39069 (1.14921)

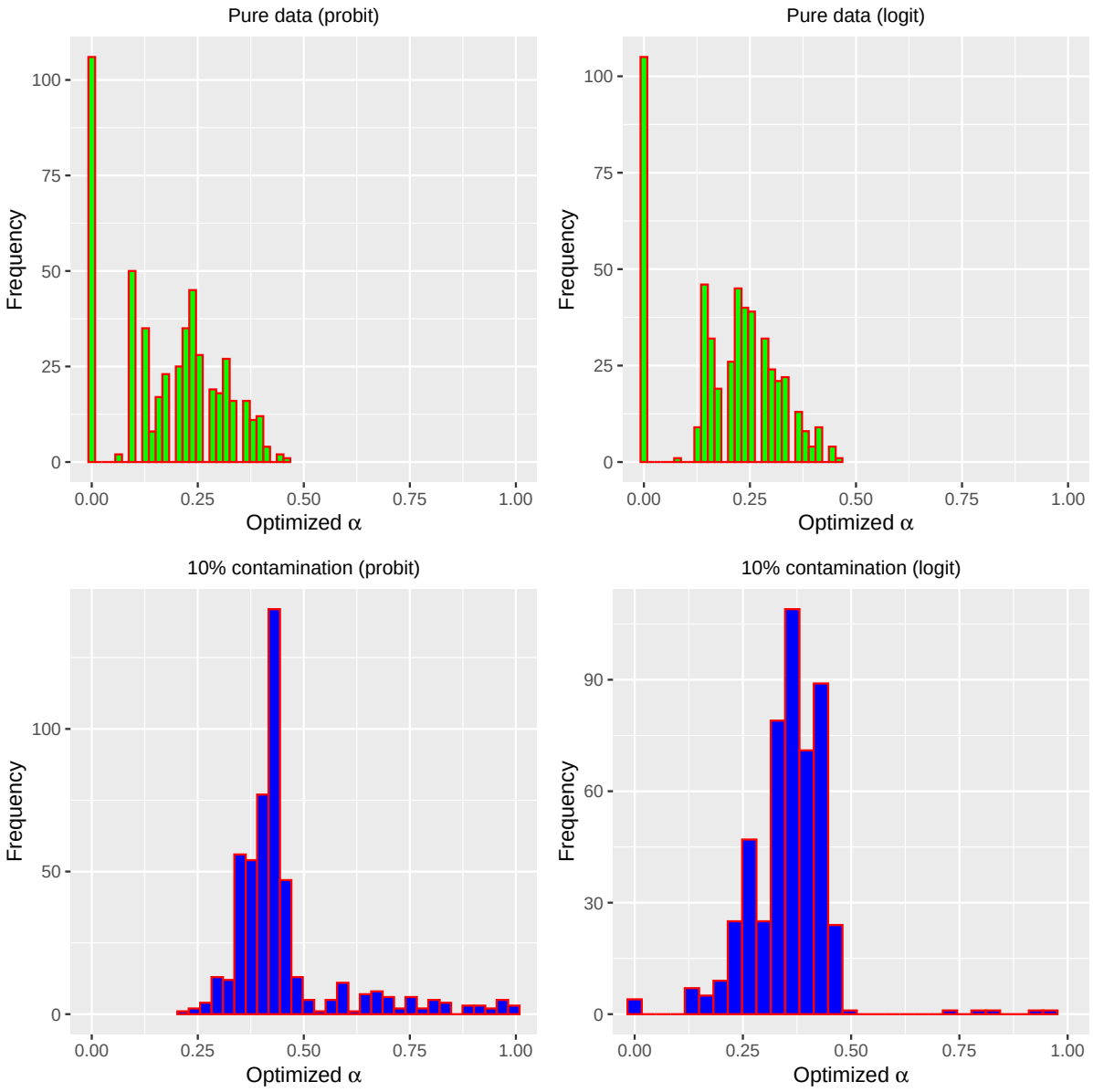

FIGURE 2.21: Histograms of the optimized α values based on the data sets simulated through Model 3 over 500 replications.

This demonstrates the effectiveness of the algorithm in picking out a suitable tuning parameter based on a data set. We apply this strategy in the next section to analyze a wine quality data set. Going forward, we will use the $\hat{\theta}_{0.5}$ as our pilot in tuning parameter selection criterion.

2.9 Real Data Analysis

In this setup, the link function, assumed to be known, is essential for determining the probabilities induced by the model. The relative merit of a link function is closely related to the optimal selection of the link function within a finite set of available options. In simulation studies, a link function may be favoured over another if it results in a lower MSE for the MDPDE at a particular value of the tuning parameter. However, this point-wise comparison alone is insufficient. To make a uniform comparison, the following criterion is proposed:

One link function, G_1, should be preferred over another, G_2, if and only if the minimized (over the tuning parameter) MSE associated with G_1 is lower than that of G_2.

This criterion can be applied to a finite set of candidate link functions. We aim to implement this algorithm to analyze a real-life data set. In practice, the true underlying link function and the true MSE are unknown to a statistician. Therefore, instead of a single link function, the analysis must begin with a finite set of potential link functions. The empirical MSE can serve as an effective proxy for the true MSE. Using the method of Warwick and Jones (2005), the optimal tuning parameter and the minimized empirical MSE can be determined for each link function. The link function yielding the lowest optimized empirical MSE within this class should then be selected for the given data set.

Here we analyze a white wine quality data set that is available in the UCI Machine Learning Repository. This data set contains 11 independent variables and an ordinal response variable. These continuous variables x_is determine the wine's quality rated on a scale of 3 – 9. As we see, the values of the covariates vary a lot. They are standardized using the median (*med*) and the mean absolute deviation (MAD) as

$$x_{ij}^{*} = \frac{x_{ij} - med(x_i)}{1.4828 \times MAD(x_i)} \text{ where } x_i = (x_{i1}, \ldots, x_{ip})^T \text{ and } j = 1, 2, \ldots, p = 11 \tag{2.70}$$

for each i-th data point, to aid better convergence for the optimization algorithms. Having done that, we find the parameter estimates using the probit and the logit links. In Table 2.26 and Table 2.27 the minimum density power divergence estimates are reported along with the 95%-trimmed MLE (trim. MLE). The 95%-trimmed MLE refers to the computation of the MLE after deletion of the statistical units whose covariates satisfy

$$(x_i^{*} - \hat{\mu})^T S^{-1} (x_i^{*} - \hat{\mu}) \geq \chi^2_{0.95,4897} \text{ for all } i. \tag{2.71}$$

Here $\hat{\mu}$ and S respectively denote the robust location and scale estimates of the data cloud $\{x_i^{*}; i = 1, \ldots, 4898\}$, and $\chi^2_{0.95,4897}$ being the upper 5% point of central χ^2-distribution with 4897 degrees of freedom. More precisely, $\hat{\mu}$ is obtained by taking component-wise medians, whereas S is constructed from the median absolute deviation for each row of the data matrix. As we find in Table 2.26 that when $\alpha = 0, 0.1, 0.3$, the MDPD estimates of $\beta_1, \beta_2, \beta_3$ differ with the trimmed MLE. This might suggest that the optimum tuning parameter might occur somewhere above 0.3 because α close to zero gives unstable estimates. This speculation is fairly supported in Table 2.28 which gives the optimum tuning parameter as $\alpha = 0.39$ for the lowest value of the MSE with the probit link. When the logit link is used, the optimum tuning parameter turns out $\alpha = 0.68$. Here MSE is

computed using the formulation of (1.75) with the pilot being chosen as $\hat{\theta}_{0.5}$. Also, we notice that, for the logit link, the matrix $\widehat{\Psi}_n(\alpha)$ is singular at $\alpha = 0$. Therefore, we report $\alpha = 0.01$ in Table 2.27 in place of $\alpha = 0$. We also report the total *standard error* (SE) in the second-last row of these two tables. The SE is computed as the sum of asymptotic standard deviations divided by the squared root of the sample size. For these links, SE generally increases with α, therefore it is the squared bias term that drives the MSE to have a parabolic shape along with α.

Next, to measure the performance of these estimates we split the entire data set into two parts, namely the training and test data sets. The training data set consisting of 75% data points is used to estimate the parameters that are further used to predict the wine quality levels in the test data set. The proportion of these cases, where the true levels match the predicted values in the test data set, measures the prediction accuracy of a particular method. In Table 2.26 we notice that the MDPDE with $\alpha > 0$ produces better accuracy than both the MLE (53.6%) and the trimmed MLE (54%) for the probit link. However in Table 2.27 we find that the trimmed MLE produces slightly better accuracy (55.4% respectively) than the MDPDEs with $\alpha > 0$.

Now we proceed to find the optimum value of the tuning parameter using the strategy of Warwick and Jones (2005). In Figure 2.24 we see that MSE decreases with α roughly up to the point $\alpha = 0.39$, after that the curve moves slightly upwards. This implies that MDPDEs with $\alpha > 0$ perform better than the MLE. This fact is fairly corroborated by the trend of the accuracy values as well. From the discussions of the simulation results, we can fairly conclude that this data set contains outlying observations with respect to the probit link function. Similarly, for the logit link, we observe a similar trend of the MSE values in Figure 2.25. We notice that MSE is minimized at the tuning parameter $\alpha = 0.68$ for the logit link. These values are reported in Table 2.28. Given this data set, we obtain different optimized MSE values for different link functions. We may take up

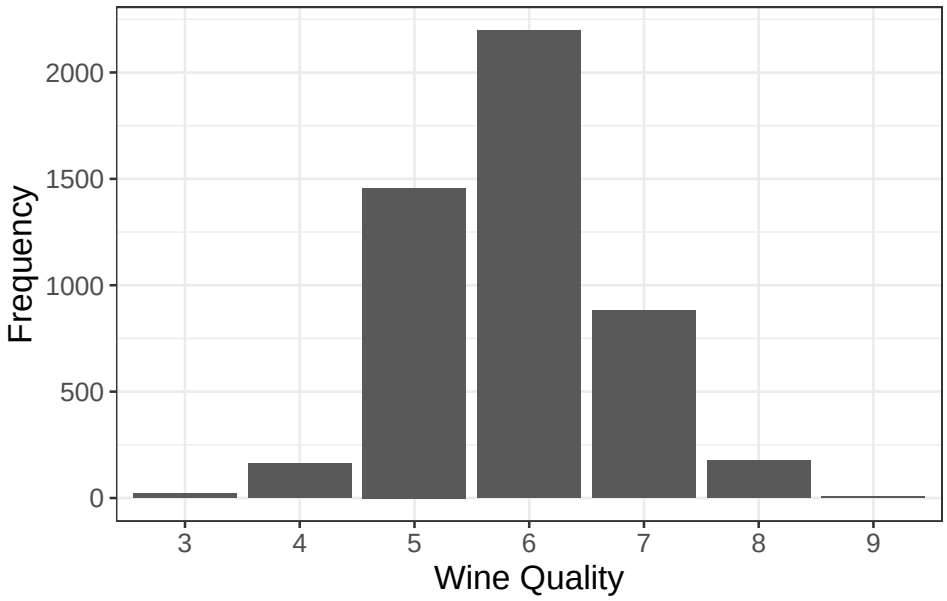

FIGURE 2.22: Histogram of the response variable in the wine quality data set.

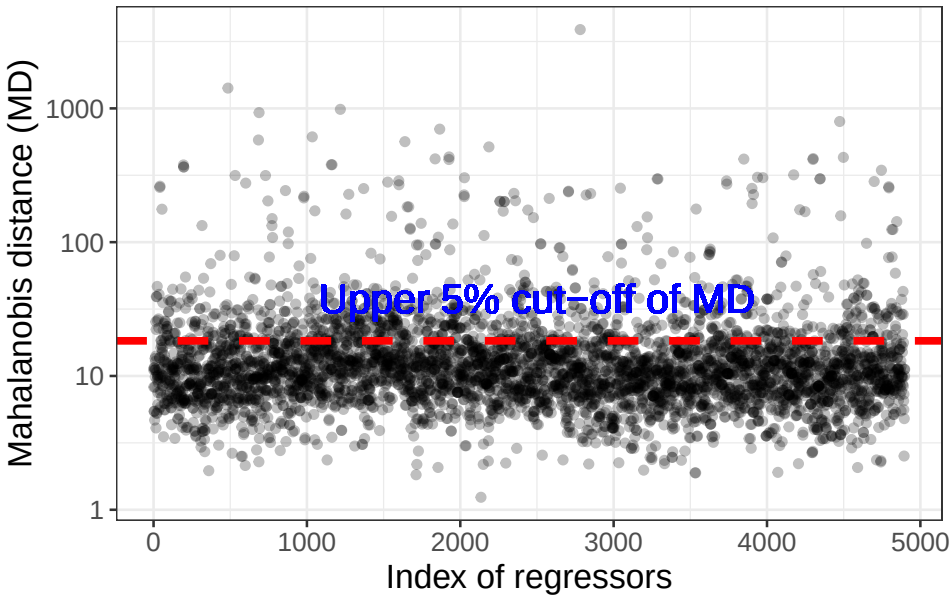

FIGURE 2.23: Mahalanobis distances corresponding to regressors.

TABLE 2.26: Parameters estimates in the wine quality data set with the probit link

Estimates	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$	95%-trim. MLE
$\hat{\beta}_1$	0.06268	0.07435	0.11204	0.11696	0.11992	0.11688	0.13085
$\hat{\beta}_2$	-0.2509	-0.25329	-0.25605	-0.25542	-0.25602	-0.25796	-0.2695
$\hat{\beta}_3$	0.00053	-0.00015	0.00791	0.01135	0.01241	0.01230	-0.01282
$\hat{\beta}_4$	0.61871	0.63508	0.72930	0.73082	0.72955	0.70955	0.6382
$\hat{\beta}_5$	-0.00358	-0.00514	-0.00374	-0.0037	-0.00344	-0.00282	-0.07292
$\hat{\beta}_6$	0.08815	0.11128	0.12612	0.1368	0.14413	0.14893	0.12094
$\hat{\beta}_7$	-0.01651	-0.02224	-0.03029	-0.04818	-0.06126	-0.07476	-0.01060
$\hat{\beta}_8$	-0.66650	-0.71737	-0.90893	-0.92605	-0.93696	-0.91311	-0.68881
$\hat{\beta}_9$	0.13937	0.15234	0.18989	0.20338	0.21168	0.21468	0.15736
$\hat{\beta}_{10}$	0.09548	0.1016	0.11539	0.12403	0.13168	0.13694	0.09799
$\hat{\beta}_{11}$	0.42875	0.42235	0.34789	0.35226	0.34956	0.36067	0.41748
$\hat{\gamma}_1$	-2.99276	-3.13133	-3.36179	-3.43993	-3.49516	-3.47491	-3.10773
$\hat{\gamma}_2$	-2.05813	-2.10963	-2.18638	-2.23159	-2.26262	-2.27492	-2.19455
$\hat{\gamma}_3$	-0.43326	-0.43169	-0.43081	-0.42982	-0.42819	-0.42475	-0.46651
$\hat{\gamma}_4$	1.06414	1.07661	1.10986	1.13198	1.14808	1.15783	1.02281
$\hat{\gamma}_5$	2.25888	2.29456	2.39137	2.45673	2.50497	2.53885	2.26259
$\hat{\gamma}_6$	24.6605	24.6605	24.6605	24.6605	24.6605	24.6605	24.6605
SE	0.56917	0.76282	0.75769	0.83709	0.9544	1.12446	NA
Accuracy	0.53551	0.54367	0.54449	0.54367	0.54286	0.54041	0.54017

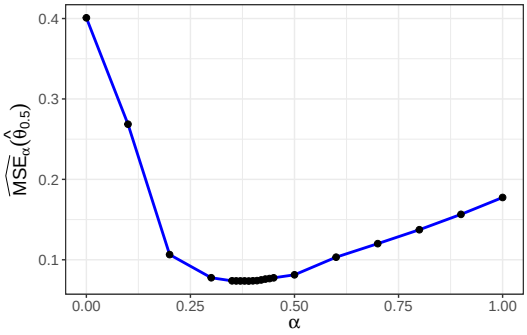

FIGURE 2.24: Graphs of MSE for the probit link

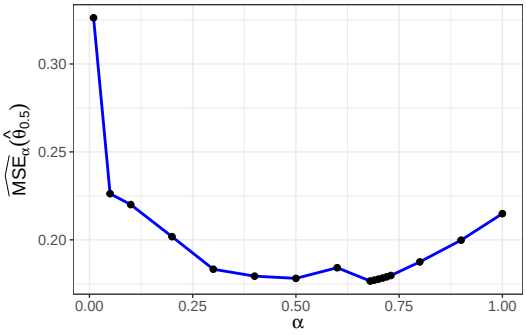

FIGURE 2.25: Graphs of MSE for the logit link

this as a future research problem in choosing an appropriate link function in this setup.

When the DPD-based method shows slightly lower accuracy compared to the 95%

TABLE 2.27: Parameters estimates in the wine quality data set with the logit link

Estimates	α = 0.01	α = 0.1	α = 0.3	α = 0.5	α = 0.7	α = 1	95%-trim. MLE
$\hat{\beta}_1$	0.29494	0.21962	0.23269	0.22743	0.21692	0.20681	0.29497
$\hat{\beta}_2$	-0.45306	-0.47338	-0.44141	-0.43249	-0.42688	-0.42918	-0.45306
$\hat{\beta}_3$	0.00639	0.02054	0.01721	0.01900	0.01979	0.01745	0.00638
$\hat{\beta}_4$	1.40803	1.39715	1.35650	1.32555	1.27118	1.22729	1.40802
$\hat{\beta}_5$	-0.13639	-0.00420	-0.00147	-0.00008	-0.00151	0.00141	-0.13660
$\hat{\beta}_6$	0.15613	0.20760	0.21563	0.22599	0.23991	0.23934	0.15611
$\hat{\beta}_7$	0.01588	-0.01629	-0.05074	-0.07677	-0.10496	-0.11556	0.01587
$\hat{\beta}_8$	-1.76746	-1.79154	-1.76090	-1.73325	-1.66398	-1.61093	-1.76747
$\hat{\beta}_9$	0.38776	0.35698	0.36654	0.37110	0.36943	0.36313	0.38776
$\hat{\beta}_{10}$	0.19478	0.20459	0.21147	0.22260	0.23045	0.23322	0.19478
$\hat{\beta}_{11}$	0.4817	0.49613	0.48572	0.48776	0.51220	0.52506	0.48172
$\hat{\gamma}_1$	-19.18072	-19.17419	-19.17860	-19.18054	-19.18059	-19.18051	-19.18072
$\hat{\gamma}_2$	-4.07523	-4.04373	-3.93328	-3.91250	-3.92988	-3.92934	-4.07523
$\hat{\gamma}_3$	-0.74738	-0.74259	-0.72432	-0.70962	-0.71042	-0.70239	-0.74737
$\hat{\gamma}_4$	1.84985	1.8959	1.89781	1.91055	1.91435	1.91503	1.84983
$\hat{\gamma}_5$	4.29262	4.26090	4.24203	4.26516	4.28086	4.30877	4.29263
$\hat{\gamma}_6$	20.87745	20.87584	20.87700	20.87743	20.87745	20.87745	20.87745
SE	1.33471	1.08470	1.09194	1.1541	1.24054	1.3822	NA
Accuracy	0.53959	0.54122	0.53796	0.53959	0.53878	0.53551	0.55426

TABLE 2.28: Optimum tuning parameter along with estimated MSE, Accuracy and SE.

Method	Link functions	α	MSE	SE	Accuracy
MDPDE	probit	0.39	0.07358	0.78485	0.54286
	logit	0.68	0.17656	1.23131	0.53959

trimmed MLE, alternative link functions should preferably be explored to assess potential improvements in accuracy. In this example, while the logit link yields slightly inferior results, the probit link demonstrates better accuracy. Additionally, the standard errors (SE) should be considered alongside accuracy for a more comprehensive comparison. The analytic expression of the standard error (SE), based on the asymptotic variance, should be used in this context. However, since the analytic expression of the asymptotic variance is not available for the 95% trimmed MLE in this setup, "NA" is reported.

2.10 Conclusions

The lack of robustness in the likelihood-based inferential procedures poses a major challenge in modelling ordinal response data. Here we explore an alternative robust methodology to estimate the parameters through minimization of the density power divergence while using those statistical models. It is based on the theory of an independent but non-homogeneous version of the DPD. We see from the estimating equations how the choice of tuning parameters enables the MDPDE to achieve a higher degree of stability against different types of outliers that are inconsistent with a reference model. The robustness of these estimators is discussed through the influence function and breakdown point analysis. Numerically, we have also shown that MDPD estimates also have a high implosive breakdown point at model misspecification. Robustness and asymptotic optimality are generally two competing concepts. The balance between these two is hard to achieve. We have demonstrated through the simulation studies how it is possible to find a suitable trade-off between these two extremities through the proper choice of a tuning parameter. Moreover, our proposed estimates perform better than Croux et al. (2013). Also, they are very competitive with Iannario et al. (2017). Factoring in all such possible challenges, we believe that the MDPDE for the ordinal response models provides a useful tool in the armoury of applied scientists.

Data Availability Statement

The data set that supports the findings of this study is openly available in the UCI Machine Learning Repository at `https://archive.ics.uci.edu/ml/datasets/wine+quality`.

This page is intentionally left blank.

Chapter 3

One-Step Inference about the Polychoric Correlation

3.1 Introduction

In behavioural sciences and many other social studies, subjects under a particular study are often categorized depending on their responses in different ordinal scales. The most convenient way to summarize these responses for a whole study is to consolidate them in the form of a contingency table. For example, consider a socio-economic study in which a group of people over the age of 18 from different economic backgrounds is selected, and subsequently, their levels of education are recorded on a scale of 0 – 2. The following is a typical example of such a contingency table.

TABLE 3.1: Levels of education under different economic backgrounds

	Educational Level			
Economic Status	Primary (0)	Secondary (1)	Advanced (2)	Total
Low (0)	10	10	13	33
Middle (1)	21	22	23	66
High (2)	31	32	33	96
Total	62	64	69	195

Here, the observed ordinal variables are *Educational Level* (levels: $0-2$) and *Economic Status* (levels: $0-2$). In this example, a person's *Education Level* is considered to be a continuous (but latent) score so that the higher values of this score correspond to higher education levels. Depending on the score and its placement within intervals defined by unknown cutoffs, the individual is categorized into one of three classes: Primary, Secondary, or Advanced. Similarly, the average monthly earnings in the current year should be considered to define an individual's *Economic Status*.

Having summarized the data in a contingency table, studying the association between these two ordinal variables will be the next job of the researcher. A natural way to accomplish this task is to find the *polychoric correlation* (Pearson, 1900a) which we will define in a moment.

Let (X, Y) be the vector of our interest that can quantify an individual response in two different ordinal scales, and together, they produce an entry in a $(r \times s)$ contingency table. Both r and s are considered finite constants unless specified otherwise. Suppose such an experiment is conducted on a sample of n $(\gg r \times s)$ subjects. The frequency corresponding to the (i, j)−th response is denoted by n_{ij}. An underlying parametric model must be assumed to generate the data to find the association between these ordinal variables X and Y using the polychoric correlation. A traditional approach would be to assume that there exists a pair of underlying unobserved latent variables (U, V) that has a probability density function $\phi_2(\cdot, \cdot; \rho)$ (i.e., the standard bivariate normal density with the correlation coefficient ρ) such that they are related to the former through the following equivalence

$$\Big\{X = i, Y = j\Big\} \overset{\mathcal{L}}{\equiv} \Big\{\eta_{i-1} < U \leq \eta_i, \beta_{j-1} < V \leq \beta_j\Big\} \text{ for } i = 1, \ldots, r \text{ and } j = 1, 2, \ldots, s. \tag{3.1}$$

Here "$\mathcal{L}$" denotes the probability law. Let γ be the vector of unknown cut-off points

$(\eta_1, \ldots, \eta_{r-1}, \beta_1, \ldots, \beta_{s-1})^T$ with the following restrictions $-\infty = \eta_0 < \eta_1 < \cdots < \eta_{r-1} < \eta_r = \infty$ and $-\infty = \beta_0 < \beta_1 < \cdots < \beta_{s-1} < \beta_s = \infty$. Under this model (3.1), the probability of having an individual response $\{X = i, Y = j\}$ is given by

$$\begin{aligned}\pi_{ij}(\rho, \gamma) &= \mathbb{P}\Big\{\eta_{i-1} < U \leq \eta_i, \beta_{j-1} < V \leq \beta_j\Big\} = \iint\limits_{R_{ij}} \phi_2(u, v; \rho) d(u \times v) \\ &= \sum_{i_1, j_1 = 0}^{1} (-1)^{i_1 + j_1} \Phi_2(\eta_{i-i_1}, \beta_{j-j_1}; \rho), \end{aligned} \tag{3.2}$$

where $R_{ij} = (\eta_{i-1}, \eta_i] \times (\beta_{j-1}, \beta_j]$ and $\Phi_2(x, y, \rho)$ being the CDF of a standard bivariate normal distribution with the correlation coefficient ρ evaluated at a point $(x, y) \in \mathbb{R}^2$. In this setup, ρ is called the polychoric correlation between X and Y. When both the observed variables are dichotomous, it is called the *tetrachoric correlation* (Pearson, 1900a; 1920).

One would like to solve ρ equating the model probabilities to the relative frequencies as

$$\pi_{ij}(\rho, \gamma) = p_{ij} \text{ where } p_{ij} = \frac{n_{ij}}{\sum_{i,j} n_{ij}} \text{ for } i = 1, 2, \ldots, r \text{ and } j = 1, 2, \ldots, s. \tag{3.3}$$

Notice that solving the sets of integral equations as in (3.3) for the polychoric correlation also requires them to be solved for a whole lot of other parameters clubbed into $\theta = (\rho, \gamma) \in \Theta = I \times \mathbb{R}^{r+s-2}$ where $I = (-1, 1)$. We use the following notations for the subsequent discussions:

$$\pi(\theta) = ((\pi_{ij}(\theta)))_{r \times s} \text{ and } p = ((p_{ij}))_{r \times s}. \tag{3.4}$$

The observed values of the marginal cumulative probabilities are given by

$$P_{is} = \sum_{l=1}^{i} \sum_{k=1}^{s} p_{lk} \text{ and } P_{rj} = \sum_{l=1}^{r} \sum_{k=1}^{j} p_{lk} \text{ for all } i, j. \tag{3.5}$$

Note that $P_{rs} = 1$. We assume that $g_{ij} = \mathbb{P}\{X = i, Y = j\}$ is the true probability that generates the observations for the (i, j)-th cell, and g is similarly represented as in (3.4). Throughout this chapter, we assume that none of the observed cells have probability zero under the model, and they should also be non-empty.

In general, one cannot find a ρ that solves (3.3), unless the specific case of a 2×2 contingency table where the tetrachoric series may be employed to approximate the model probability. However, this technique does not seem to extend to higher-dimensional contingency tables. Instead, we would turn this into an optimization problem. For the case of a general contingency table, some statistical distance $L(p, \pi(\theta))$ between the observed and model cell probabilities may be minimized to find a solution for ρ. Ritchie-Scott (1918) suggests the estimation of ρ by averaging over all possible tetrachoric correlations. Tallis (1962) proposes the maximum likelihood estimation of ρ along with the points of polytomy (or, the cut-off points) simultaneously in a 3×3 contingency table. Lancaster (1964) and Hamdan (1968; 1971) use the polychoric series to estimate the polychoric correlation, but their results differ from those of Tallis (1962). Hamdan (1970) also studies the structure of the tetrachoric correlation, and further, shows that the tetrachoric correlation is equivalent to the MLE of ρ for a 2×2 table. Martinson and Hamdan (1972; 1975) study the maximum likelihood method and a few other asymptotically efficient methods to estimate the polychoric correlation in a 2×2 contingency table. In the context of general $r \times s$ contingency tables, Olsson (1979) discusses the maximum likelihood estimation, and subsequently, a two-step approach in the maximum likelihood estimation is adapted. Lee (1985) extends the ML estimation method

to $r \times s \times t$ contingency tables. Joreskog (1994) outlines a general theory of the maximum likelihood estimation method in the context of a contingency table, and derives their asymptotic covariance matrices. Poon & Lee (1987) and Lee & Chiu (1990) extend the ML method to estimate the multivariate polychoric and polyserial correlation coefficients for complete data. Leung (1990) further extends it to incomplete data. The central theme of estimation in each of the above methods revolves around the problem of maximizing the likelihood function. The loss function associated with the likelihood function is the likelihood disparity (LD) which essentially takes the following form

$$LD(p, \pi(\theta)) = \sum_{i,j} p_{ij} \ln \frac{p_{ij}}{\pi_{ij}(\theta)}. \tag{3.6}$$

In all of the above situations, latent vectors are assumed to be normally distributed. Lee and Lam (1988) study the estimation of the polychoric correlation assuming that the latent variables follow some elliptic distribution. Quiroga (1994), Rascino and Pollice (2006) suggest using a bivariate skew-normal distribution instead, but, this theory is neither fully developed nor does it agree with Pearson's original definition that depends on the bivariate normal distributions. Tomofeeva and Khailenko (2016) propose an extension to the definition of the polychoric correlation assuming that the latent vector has a distribution with shape parameters. They have considered only symmetric distributions and assume that the true distribution may be leptokurtic as well. For this purpose, bivariate Student and generalized lambda distributions are used. Based on these distributional assumptions, estimation algorithms are developed.

In practice, all the observations in a contingency table are not 'usually' generated through a bivariate normal distribution, neither, could all such outlying observations be accommodated into an analysis without adding further complexity to a model. On the other hand, doing away with those unruly observations could entail an information loss. In

this paper, we shall not tinker with the model assumption. Also, remember that the maximum likelihood estimators are highly sensitive to model misspecifications. This necessitates using some robust methods alternatives to the MLE. Keeping all these in mind, we adopt the density power divergence replacing the LD without changing the bivariate normality assumption per se. This would allow a reasonable modelling of the majority of the data barring a small portion that is probabilistically discordant with the assumed model; however, they should not adversely impact the estimators and tests. The density power divergence will be implemented to this problem in one of two ways– namely the one-step method and the two-step method. In this chapter, we shall discuss the first type, while the second one will be discussed in the next chapter.

The rest of this chapter is organized as follows.

(a) In Section 3.2 the estimating equations for the minimum density power divergence estimators are presented.

(b) Their asymptotic properties are discussed in Section 3.3. More specifically, Subsection 3.3.1 includes two different proofs of consistency of the MDPDE, and Subsection 3.3.2 contains the asymptotic normality result. Asymptotic properties of the test statistic are discussed in Subsection 3.3.3.

(c) In Section 3.4 we study the stability of the estimators and the test statistics primarily using the influence function.

(d) We also study the asymptotic breakdown point in Section 3.5. More specifically, Subsection 3.5.1 contains the asymptotic breakdown point result in this setup. Also, we discuss how the asymptotic breakdown point changes under different kinds of reparametrization in Subsection 3.5.2.

(e) Simulation studies are presented in Section 3.6. Applications of this method to some real-life data examples are presented in Section 3.7. Finally, some concluding remarks are made in Section 3.8.

Remark 3.1. *Polychoric correlation assumes that observed categorical variables originate from an underlying bivariate normal distribution, enabling it to exploit the structural relationship between ordinal variables. Unlike the nonparametric measures such as Yule's coefficient, the coefficient of colligation, the chi-squared measure, or Pearson's coefficient of contingency, which disregard latent variable information, polychoric correlation can better reflect the true association. The polychoric correlation, taking values in* $[-1,1]$*, also provides insight into the direction and strength for the strongest possible relationship, something not achievable with some nonparametric measures. While measures like Kendall's* τ *focus on monotonic relationships and are robust to underlying distributional assumptions, they may fail to capture associations when ordinal categories are coarse or unevenly spaced. This occurs because, in such cases, more observations are grouped into a single category than is appropriate, leading to an incorrect count of concordant pairs. In such cases, polychoric correlation is preferred, highlighting the need for its robust version in statistical inference.*

3.2 Estimating Equations

We assume that $\pi_{ij}(\theta), g_{ij} > 0$ for all $i = 1, 2 \ldots, r$ and $j = 1, 2, \ldots, s$. Recalling the notation from (3.4) the density power divergence between g and $\pi(\theta)$ is given by

$$d_\alpha(g, \pi(\theta)) = \sum_{i,j} \left\{ \pi_{ij}^{1+\alpha}(\theta) - \left(1 + \frac{1}{\alpha}\right) \pi_{ij}^{\alpha}(\theta) g_{ij} + \frac{1}{\alpha} g_{ij}^{1+\alpha} \right\} \text{ for } \alpha > 0. \tag{3.7}$$

$d_0(g, \pi(\theta))$ is defined as a continuous limit of (3.7) when $\alpha \downarrow 0+$. In particular, one recovers the LD (which produces the MLE) at $\alpha = 0$, and it turns out to be the L_2 distance (which produces very robust but somewhat inefficient estimators) at $\alpha = 1$. So, a

proper balance between these two extreme situations can be achieved by adjusting the parameter α according to a specific situation.

In the one-step method, the divergence is minimized over all the parameters simultaneously. A best-fitting parameter is given by

$$\theta_\alpha = \arg\min_{\theta\in\Theta} d_\alpha(g, \pi(\theta)) \text{ at fixed } \alpha. \tag{3.8}$$

As the true density g is unknown, it may be replaced by its empirical version p in (3.8) to find the one-step minimum density power divergence estimator (MDPDE) as

$$\hat{\theta}_\alpha = \arg\min_{\theta\in\Theta} d_\alpha(p, \pi(\theta)) \text{ at fixed } \alpha. \tag{3.9}$$

Since $\pi_{ij}(\theta)$s are differentiable with respect to each component of θ, the set of estimating equations can be symbolically written as:

$$\nabla d_\alpha(p, \pi(\hat{\theta}_\alpha)) := (1+\alpha)\sum_{i,j}\left\{\pi_{ij}^{\alpha}(\hat{\theta}_\alpha) - \pi_{ij}^{\alpha-1}(\hat{\theta}_\alpha)p_{ij}\right\}\nabla\pi_{ij}(\hat{\theta}_\alpha) = 0 \text{ at fixed } \alpha, \tag{3.10}$$

where $\nabla\pi_{ij}(\theta)^T = \left(\frac{\partial\pi_{ij}(\theta)}{\partial\rho}, \frac{\partial\pi_{ij}(\theta)}{\partial\eta_1}, \ldots, \frac{\partial\pi_{ij}(\theta)}{\partial\eta_{r-1}}, \frac{\partial\pi_{ij}(\theta)}{\partial\beta_1}, \ldots, \frac{\partial\pi_{ij}(\theta)}{\partial\beta_{s-1}}\right) \in \mathbb{R}^{r+s-1}$. We further know (cf. p.510, Balakrishnan and Lai, 2009; Sungur, 1990) that

$$\frac{\partial}{\partial\rho}\Phi_2(x_1, x_2; \rho) = \phi_2(x_1, x_2; \rho) \text{ and } \frac{\partial}{\partial x_1}\Phi_2(x_1, x_2; \rho) = \phi_1(x_1)\Phi_1\left(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}}\right). \tag{3.11}$$

Recalling the expression of $\pi_{ij}(\theta)$ from (3.2), its partial derivatives can be calculated as

$$\frac{\partial\pi_{ij}(\theta)}{\partial\rho} = \sum_{i_1, j_1=0}^{1}(-1)^{i_1+j_1}\phi_2(\eta_{i-i_1}, \beta_{j-j_1}; \rho) \tag{3.12}$$

and

$$\frac{\partial \pi_{t_1 j}(\theta)}{\partial \eta_{t_1}} = -\frac{\partial \pi_{(t_1+1)j}(\theta)}{\partial \eta_{t_1}} = \phi_1(\eta_{t_1})\left\{\Phi_1\left(\frac{\beta_j - \rho\eta_{t_1}}{\sqrt{1-\rho^2}}\right) - \Phi_1\left(\frac{\beta_{j-1} - \rho\eta_{t_1}}{\sqrt{1-\rho^2}}\right)\right\}, \tag{3.13}$$

$$\frac{\partial \pi_{it_2}(\theta)}{\partial \beta_{t_2}} = -\frac{\partial \pi_{i(t_2+1)}(\theta)}{\partial \beta_{t_2}} = \phi_1(\beta_{t_2})\left\{\Phi_1\left(\frac{\eta_i - \rho\beta_{t_2}}{\sqrt{1-\rho^2}}\right) - \Phi_1\left(\frac{\eta_{i-1} - \rho\beta_{t_2}}{\sqrt{1-\rho^2}}\right)\right\} \tag{3.14}$$

for $t_1 = 1, \ldots, r-1$ and $t_2 = 1, \ldots, s-1$. Using (3.12) to (3.14), it can be readily seen that $\hat{\theta}_\alpha$ solves the following system of estimating equations:

$$\frac{1}{(1+\alpha)} \cdot \frac{\partial}{\partial \rho} d_\alpha(p, \pi(\theta)) = \sum_{i,j}\left\{\pi_{ij}^\alpha(\theta) - \pi_{ij}^{\alpha-1}(\theta)p_{ij}\right\} \sum_{i_1,j_1=0}^{1} (-1)^{i_1+j_1}\phi_2(\eta_{i-i_1}, \beta_{j-j_1}; \rho) = 0, \tag{3.15}$$

$$\frac{1}{(1+\alpha)} \cdot \frac{\partial}{\partial \eta_{t_1}} d_\alpha(p, \pi(\theta)) = \sum_j \left[\left\{\pi_{t_1 j}^\alpha(\theta) - \pi_{t_1 j}^{\alpha-1}(\theta)p_{t_1 j}\right\} - \left\{\pi_{(t_1+1)j}^\alpha(\theta) - \pi_{(t_1+1)j}^{\alpha-1}(\theta)p_{(t_1+1)j}\right\}\right] \times \phi_1(\eta_{t_1})\left\{\Phi_1\left(\frac{\beta_j - \rho\eta_{t_1}}{\sqrt{1-\rho^2}}\right) - \Phi_1\left(\frac{\beta_{j-1} - \rho\eta_{t_1}}{\sqrt{1-\rho^2}}\right)\right\} = 0 \tag{3.16}$$

and

$$\frac{1}{(1+\alpha)} \cdot \frac{\partial}{\partial \beta_{t_2}} d_\alpha(p, \pi(\theta)) = \sum_i \left[\left\{\pi_{it_2}^\alpha(\theta) - \pi_{it_2}^{\alpha-1}(\theta)p_{it_2}\right\} - \left\{\pi_{i(t_2+1)}^\alpha(\theta) - \pi_{i(t_2+1)}^{\alpha-1}(\theta)p_{i(t_2+1)}\right\}\right] \times \phi_1(\beta_{t_2})\left\{\Phi_1\left(\frac{\eta_i - \rho\beta_{t_2}}{\sqrt{1-\rho^2}}\right) - \Phi_1\left(\frac{\eta_{i-1} - \rho\beta_{t_2}}{\sqrt{1-\rho^2}}\right)\right\} = 0 \tag{3.17}$$

for all t_1, t_2 and fixed $\alpha > 0$. The essential objective function that comes from (3.7) excluding the constant term may be rewritten as

$$H_n(\theta) = \frac{1}{n}\sum_{l=1}^{n} V(\theta, Z_l) \text{ at fixed } \alpha, \tag{3.18}$$

where $V(\theta, Z_l) = \sum_{i,j}\left\{\pi_{ij}^{1+\alpha}(\theta) - \left(1 + \frac{1}{\alpha}\right)\pi_{ij}^\alpha(\theta)\delta_{ij}(Z_l)\right\}$ and $\delta_{ij}(Z_l) = \mathbb{1}\left\{Z_l = (i,j)\right\}$. Suppose $\mathbb{P}_n$ is the probability measure that puts an equal mass $\frac{1}{n}$ at each point of $\{Z_1, Z_2, \ldots, Z_n\}$

and G being the true distribution, then $H_n(\theta) = \mathbb{E}_{\mathbb{P}_n} V(\theta, Z_l)$. Note that $\mathbb{E}_G \delta_{ij}(Z_l) = \mathbb{P}_G\Big\{Z_l = (i,j)\Big\} = g_{ij}$; this gives the population version of the objective function (3.18) as

$$H(\theta) = \mathbb{E}_G V(\theta, Z_l) \text{ at fixed } \alpha. \tag{3.19}$$

Consequently, $\hat{\theta}_\alpha$ and θ_α respectively satisfy the equations

$$\mathbb{E}_{\mathbb{P}_n} \nabla V(\hat{\theta}_\alpha, Z_l) = 0 \text{ and } \mathbb{E}_G \nabla V(\theta_\alpha, Z_l) = 0 \text{ at fixed } \alpha. \tag{3.20}$$

The dependency on α is kept implicit in these symbols– $H(\theta)$ and $H_n(\theta)$. When the data are truly generated by the model, both the one-step and two-step estimators of the polychoric correlation converge to the same limit.

Remark 3.2. *(**Invariance under variable transformation**) In practice, often the ordinal responses come in the form of some transformation of variables. For instance, consider the transformation: $(X,Y) \mapsto \psi(X,Y) = \Big(\psi_1(X), \psi_2(Y)\Big)$ such that their inverses ψ_1^{-1} and ψ_2^{-1} exist. For this transformation, we get $\{X = i, Y = j\} \equiv \Big\{\psi_1(X) = \psi_1(i), \psi_2(Y) = \psi_2(j)\Big\}$ under every probability law. Hence, the objective function of the DPD in terms of θ does not change. Thus, $\hat{\theta}_\alpha$ remains invariant under such a transformation ψ. This includes common affine transformations such as: $(X,Y) \mapsto (aX + b, cY + d)$ with $a \neq 0$ and $c \neq 0$.*

Remark 3.3. *(**Equivariance under reparametrization**) Let us assume that the objective function $H_n(\theta)$ in (3.18) has a unique minimizer. Next we consider a one-to-one function $\tau = \psi(\theta) : \Theta \to \mathcal{T}$. Under such a transformation, the objective function $H_n(\theta)$ can be equivalently written as $H_n(\psi^{-1}(\tau))$. Then $\hat{\tau}_\alpha$ that minimizes the latter must satisfy the relation: $\psi^{-1}(\hat{\tau}_\alpha) = \hat{\theta}_\alpha$, where $\hat{\theta}_\alpha$ is the MDPDE already defined earlier. Therefore, we must have $\hat{\tau}_\alpha = \psi(\hat{\theta}_\alpha)$. Since the case $\alpha = 0$ is already covered through the properties of the MLE, we have this equivariance property for all $\alpha \geq 0$.*

Remark 3.4. ***(Effect of the affine transformation of the bivariate normal model)*** *Let us assume that the pair of underlying latent variables are not* $W = (U, V)^T$ *(the usual latent variables), but* $W^* = (U^*, V^*)^T$ *such that* $W^* \sim \mathcal{N}(\mu, \Sigma)$*. The non-zero mean vector and covariance matrix are respectively given by* $\mu = (\mu_1, \mu_2)^T$ *and*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \sigma_2\sigma_1\rho & \sigma_2^2 \end{bmatrix} \text{ with } \sigma_1\sigma_2 \neq 0. \tag{3.21}$$

The new pair of latent variables $(U^*, V^*)^T$ *may be obtained from the original variables as*

$$U^* = \sigma_1 U + \mu_1 \text{ and } V^* = \sigma_2 V + \mu_2. \tag{3.22}$$

Accordingly, the cut-offs are transformed as

$$\eta_i^* = \sigma_1\eta_i + \mu_1 \text{ and } \beta_j^* = \sigma_2\beta_j + \mu_2 \tag{3.23}$$

for $i = 1, \ldots, r$ *and* $j = 1, \ldots, s$*. Under this transformation, the model probabilities remain invariant because*

$$\begin{aligned} \pi_{ij}(\theta) &= \mathbb{P}\left\{\eta_{i-1} < U \leq \eta_i, \beta_{j-1} < V \leq \beta_j\right\} \\ &= \mathbb{P}\left\{\sigma_1\eta_{i-1} + \mu_1 < \sigma_1 U + \mu_1 \leq \sigma_1\eta_i + \mu_1, \sigma_2\beta_{j-1} + \mu_2 < \sigma_2 V + \mu_2 \leq \sigma_2\beta_j + \mu_2\right\} \\ &= \mathbb{P}\left\{\eta_{i-1}^* < U^* \leq \eta_i^*, \beta_{j-1}^* < V^* \leq \beta_j^*\right\}. \end{aligned} \tag{3.24}$$

Consequently, both the polychoric correlation functional and its MDPD estimate remain invariant under such an affine transformation of the bivariate normal model. Since the new cut-off parameters are affinely transformed from the original ones, one may apply Remark 3.3 to find its MDPD estimates under appropriate conditions.

3.3 Asymptotic Properties

In the rest of this chapter, the parameter space Θ is assumed to be equipped with the Euclidean metric. It means that the convergence in the parameter space will be understood in the sense of the Euclidean metric. In the two-step scenario, the same metric is assumed to be restricted to $I = (-1, 1)$. Observe that in either case, the parameter space is not a compact set. This, in a way, does not ensure the existence of best-fitting parameters. All the results in this chapter will be discussed for $\alpha > 0$ unless mentioned otherwise.

3.3.1 Consistency

In Basu et al. (1998), the existence of a best-fitting parameter is implicitly assumed. In our first result in this subsection, the existence of a best-fitting parameter is derived under a mild condition. An assumption of uniqueness in this case makes the proof of consistency simpler. This result closely follows the approach of Beran (1977a).

Theorem 3.1. *Suppose* $\pi_{ij}(\theta_1) \neq \pi_{ij}(\theta_2)$ *for all* $i \neq j$ *and* $\theta_1 \neq \theta_2$*; also* $\inf_{\theta \in \Theta \setminus H} d_\alpha(g, \pi(\theta)) > d_\alpha(g, \pi(\theta^*))$ *for some compact set* $H \subset \Theta$ *where* $\theta^* \in H$*. Then a best-fitting parameter* θ_α *exists. If* θ_α *is unique, we further have* $\hat{\theta}_\alpha \xrightarrow{a.s.} \theta_\alpha$ *as* $n \to \infty$*.*

Proof. The first condition implies that the divergence when viewed as a function of θ, differs at different values of the parameter θ. This "identifiability" assumption is used in the last part of this proof to prove the consistency of the estimators when a best-fitting parameter exists uniquely.

Now consider a sequence of parameters $\{\theta_n\}$ such that $\theta_n \longrightarrow \theta_0$ for some fixed θ_0. it is easy to see that

$$\left|d_\alpha(g,\pi(\theta_n)) - d_\alpha(g,\pi(\theta_0))\right| \leq \sum_{i,j}\left|\pi_{ij}^{1+\alpha}(\theta_n) - \pi_{ij}^{1+\alpha}(\theta_0)\right| + \left(1+\frac{1}{\alpha}\right)\sum_{i,j}\left|\pi_{ij}^{\alpha}(\theta_n) - \pi_{ij}^{\alpha}(\theta_0)\right|$$

at any fixed $\alpha > 0$. Since $\theta \mapsto \pi(\theta)$ is continuous, $|d_\alpha(g,\pi(\theta_n)) - d_\alpha(g,\pi(\theta_0))| \longrightarrow 0$ as $\theta_n \longrightarrow \theta_0$. So the map $\theta \mapsto d_\alpha(g,\pi(\theta))$ is continuous for fixed g and $\alpha > 0$. As $\inf_{\Theta\setminus H} d_\alpha(g,\pi(\theta)) > d_\alpha(g,\pi(\theta^*))$ for some compact set $H \subset \Theta$ and $\theta^* \in H$, the continuity of $\theta \mapsto d_\alpha(g,\pi(\theta))$ implies that $d_\alpha(g,\pi(\theta))$ achieves a minimizer θ_α in $H \subset \Theta$.

We shall prove the consistency of $\hat{\theta}_\alpha$. At fixed g, we define $h(\theta) = d_\alpha(g,\pi(\theta))$ and $h_n(\theta) = d_\alpha(p,\pi(\theta))$. We shall prove that $h(\hat{\theta}_\alpha) \xrightarrow{a.s.} h(\theta_\alpha)$ as $n \to \infty$. See that

$$\sup_\theta |h_n(\theta) - h(\theta)| \leq (1+\alpha^{-1})\sum_{i,j}|p_{ij} - g_{ij}| + \frac{1}{\alpha}\sum_{i,j}|p_{ij}^{1+\alpha} - g_{ij}^{1+\alpha}|. \tag{3.25}$$

We know that p_{ij} almost surely converges to g_{ij} for each i,j. Thus

$$\sup_\theta \left|h_n(\theta) - h(\theta)\right| \xrightarrow{a.s.} 0 \text{ as } n \to \infty. \tag{3.26}$$

Now, if $h(\theta_\alpha) \geq h_n(\hat{\theta}_\alpha)$, then

$$0 \leq h(\theta_\alpha) - h_n(\hat{\theta}_\alpha) \leq h(\hat{\theta}_\alpha) - h_n(\hat{\theta}_\alpha). \tag{3.27}$$

Similarly, if $h(\theta_\alpha) \leq h_n(\hat{\theta}_\alpha)$, then

$$0 \leq h_n(\hat{\theta}_\alpha) - h(\theta_\alpha) \leq h_n(\theta_\alpha) - h(\theta_\alpha). \tag{3.28}$$

This yields

$$\left|h_n(\hat{\theta}_\alpha) - h(\theta_\alpha)\right| \leq 2 \sup_\theta \left|h(\theta) - h_n(\theta)\right| \xrightarrow{a.s.} 0 \text{ as } n \to \infty. \tag{3.29}$$

Combining (3.26) and (3.29), we arrive at

$$\left|h(\hat{\theta}_\alpha) - h(\theta_\alpha)\right| \leq \left|h(\hat{\theta}_\alpha) - h_n(\hat{\theta}_\alpha)\right| + \left|h_n(\hat{\theta}_\alpha) - h(\theta_\alpha)\right| \xrightarrow{a.s.} 0 \text{ as } n \to \infty. \tag{3.30}$$

See that the map $\theta \mapsto h(\theta)$ is continuous. We know the best-fitting parameter $\theta_\alpha \in H$. If possible, there exists a subsequence of $\hat{\theta}_\alpha$ such that it does not converge to θ_α, but to a different limit. Then, the continuity of $h(\cdot)$ leads us to a different subsequential limit which contradicts the uniqueness of θ_α. Hence, $h(\hat{\theta}_\alpha) \to h(\theta_\alpha)$ implies that $\hat{\theta}_\alpha \to \theta_\alpha$ as $n \to \infty$. □

Often there exist multiple best-fitting parameters. In that case, the above proof of consistency does not work. In the next Theorem, we prove the consistency when multiple best-fitting parameters exist. This proof follows the approach of Wald (1949). Recall that for any open set U, its diameter is defined as $diam(U) = \max\left\{||x - y|| : x, y \in U\right\}$.

Theorem 3.2. *Suppose* $\inf_{\theta \in \Theta \setminus H} d_\alpha(g, \pi(\theta)) > d_\alpha(g, \pi(\theta^*))$ *for some compact set* $H \subset \Theta$ *and* $\theta^* \in H$. *Also assume that for every open ball* U *around* θ *such that* $diam(U) \downarrow 0+$, *the map* $Z_l \mapsto \inf_{\theta \in U} V(\theta, Z_l)$ *is measurable and* $\mathbb{E}_G\left[\inf_{\theta \in U} V(\theta, Z_l)\right] > -\infty$. *Then for any estimator* $\hat{\theta}_\alpha$ *satisfying* $d_\alpha(p, \pi(\hat{\theta}_\alpha)) \leq d_\alpha(p, \pi(\theta_\alpha)) + o_{\mathbb{P}}(1)$, *it holds that*

$$\mathbb{P}\left\{\hat{\theta}_\alpha \in H : \min_{\theta_\alpha} ||\hat{\theta}_\alpha - \theta_\alpha|| \geq \epsilon\right\} \longrightarrow 0 \text{ as } n \to \infty \tag{3.31}$$

for every $\epsilon > 0$.

Proof. As in Theorem 3.1, the first condition ensures the existence of at least one best-fitting parameter inside a compact set H which is a proper subset of the parameter space. The assumption that the expectation of that particular quantity is bounded away from $-\infty$ prevents the trivial case when all the points in the parameter space are best-fitting parameters. At fixed α and $\epsilon > 0$, consider the following set

$$A = \Big\{\theta \in H : \min_{\theta_\alpha} ||\theta - \theta_\alpha|| \geq \epsilon\Big\}. \tag{3.32}$$

Note that A is compact, as it is a closed subset of the compact set H. Since $\{U_\theta \text{ open } : \theta \in A, diam(U_\theta) < 1/n\}$ forms an open cover of A, we have $A \subseteq \cup_{t=1}^{k} U_{\theta_t}$ for some finite integer k. By construction, the open set U_{θ_t} shrinks to the point θ_t, and consequently, the infimum over U_{θ_t} increases when $n \to \infty$. Thus, $\inf_{\theta \in U_{\theta_t}} V(\theta, Z_l) \uparrow V(\theta_t, Z_l)$ as $n \to \infty$ almost surely.

Using the monotone convergence theorem, we get

$$\begin{aligned}\Big|\mathbb{E}_{\mathbb{P}_n} \inf_{\theta \in U_{\theta_t}} V(\theta, Z_l) - \mathbb{E}_G V(\theta_t, Z_l)\Big| &\leq \mathbb{E}_{\mathbb{P}_n}\Big|\inf_{\theta \in U_{\theta_t}} V(\theta, Z_l) - V(\theta_t, Z_l)\Big| \\ &\quad + \Big|\mathbb{E}_{\mathbb{P}_n} V(\theta_t, Z_l) - \mathbb{E}_G V(\theta_t, Z_l)\Big| \longrightarrow 0 \text{ as } n \to \infty \end{aligned} \tag{3.33}$$

for all $t = 1, 2, \ldots, k$. See also

$$\begin{aligned}\inf_{\theta \in U_{\theta_t}} \mathbb{E}_{\mathbb{P}_n}\Big[V(\theta, Z_l)\Big] &= \inf_{\theta \in U_{\theta_t}}\left[\frac{1}{n}\sum_{l=1}^{n} V(\theta, Z_l)\right] \\ &\geq \frac{1}{n}\sum_{l=1}^{n} \inf_{\theta \in U_{\theta_t}} V(\theta, Z_l) = \mathbb{E}_{\mathbb{P}_n}\Big[\inf_{\theta \in U_{\theta_t}} V(\theta, Z_l)\Big]. \end{aligned} \tag{3.34}$$

Now observe that

$$\begin{aligned}\inf_{\theta\in A}\mathbb{E}_{\mathbb{P}_n}V(\theta,Z_l) &\geq \min\Big\{\inf_{\theta\in U_{\theta_t}}\mathbb{E}_{\mathbb{P}_n}V(\theta,Z_l); t=1,2,\ldots,k\Big\}\\ &\geq \min_{t=1,\ldots,k}\mathbb{E}_{\mathbb{P}_n}\inf_{\theta\in U_{\theta_t}}\Big[V(\theta,Z_l)\Big]\\ &= \min_{t=1,\ldots,k}\mathbb{E}_G V(\theta_t,Z_l)+o_{\mathbb{P}}(1)\\ &> \mathbb{E}_G V(\theta_\alpha,Z_l)+o_{\mathbb{P}}(1) \text{ as } \theta_\alpha\notin A \text{ and } \theta_1,\ldots,\theta_k\in A. \qquad (3.35)\end{aligned}$$

Suppose $\hat{\theta}_\alpha \in A$, then $\inf_{\theta\in A} d_\alpha(p,\pi(\theta)) \leq d_\alpha(p,\pi(\hat{\theta}_\alpha))$. Also, we assume that $d_\alpha(p,\pi(\hat{\theta}_\alpha)) \leq d_\alpha(p,\pi(\theta_\alpha))+o_{\mathbb{P}}(1) = d_\alpha(g,\pi(\theta_\alpha))+o_{\mathbb{P}}(1)$. So we get

$$\inf_{\theta\in A}\mathbb{E}_{\mathbb{P}_n}V(\theta,Z_l) \leq \mathbb{E}_G V(\theta_\alpha,Z_l)+o_{\mathbb{P}}(1). \qquad (3.36)$$

This gives

$$\Big\{\hat{\theta}_\alpha\in A\Big\} \subseteq \Big\{\inf_{\theta\in A}\mathbb{E}_{\mathbb{P}_n}V(\theta,Z_l) \leq \mathbb{E}_G V(\theta_\alpha,Z_l)+o_{\mathbb{P}}(1)\Big\}. \qquad (3.37)$$

In the view of (3.35), we get

$$\mathbb{P}\Big\{\inf_{\theta\in A}\mathbb{E}_{\mathbb{P}_n}V(\theta,Z_l) \leq \mathbb{E}_G V(\theta_\alpha,Z_l)+o_{\mathbb{P}}(1)\Big\} \longrightarrow 0 \text{ as } n\to\infty. \qquad (3.38)$$

Therefore we have $\mathbb{P}\{\hat{\theta}_\alpha\in A\}\longrightarrow 0$ when $n\to\infty$. This completes the proof. □

Remark 3.5. *In Theorem 3.2, the convergence in probability is proved since we have assumed that $\hat{\theta}_\alpha$ approximately minimizes the objective function $d_\alpha(p,\pi(\theta))$ in a weaker sense. A stronger version of convergence may also be established if we analogously replace $o_{\mathbb{P}}(1)$ by $o_{a.s.}(1)$ in the assumption $d_\alpha(p,\pi(\hat{\theta}_\alpha)) \leq d_\alpha(p,\pi(\theta_\alpha))+o_{\mathbb{P}}(1)$. However, this modification*

may not be quite tenable to practical applications. Also, note that we can do away with the measurability restrictions, and carry out the proof in terms of appropriate outer expectations. These two results widen the applicability of the theory of the DPD when consistency is a concern.

3.3.2 Asymptotic Normality

Let the score vector be given by $u_{ij}^T(\theta) = \frac{\partial}{\partial \theta^T} \ln \pi_{ij}(\theta)$, also recall that $I_{ij}(\theta) = -\frac{\partial}{\partial \theta} u_{ij}(\theta)$. Following Basu et al. (1998) we know that

$$\begin{aligned} J_\alpha(\theta) &= \frac{1}{(1+\alpha)} \begin{pmatrix} \mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho^2} V(\theta, Z_l)\Big] & \mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho \partial \gamma^T} V(\theta, Z_l)\Big] \\ \\ \mathbb{E}_G\Big[\frac{\partial^2}{\partial \gamma \partial \rho} V(\theta, Z_l)\Big] & \mathbb{E}_G\Big[\frac{\partial^2}{\partial \gamma \partial \gamma^T} V(\theta, Z_l)\Big] \end{pmatrix} \\ &= \sum_{i,j} u_{ij}(\theta) u_{ij}^T(\theta) \pi_{ij}^{1+\alpha}(\theta) + \sum_{i,j} \Big\{ I_{ij}(\theta) - \alpha u_{ij}(\theta) u_{ij}^T(\theta) \Big\} \Big\{ g_{ij} - \pi_{ij}(\theta) \Big\} \pi_{ij}^{\alpha}(\theta), \end{aligned} \tag{3.39}$$

$$\begin{aligned} K_\alpha(\theta) &= Var_G\Big[\pi_{ij}^{\alpha}(\theta) u_{ij}(\theta) \Big] \\ &= \sum_{i,j} u_{ij}(\theta) u_{ij}^T(\theta) \pi_{ij}^{2\alpha}(\theta) g_{ij} - \xi(\theta) \xi^T(\theta) \text{ where } \xi(\theta) = \sum_{i,j} u_{ij}(\theta) \pi_{ij}^{\alpha}(\theta) g_{ij}. \end{aligned} \tag{3.40}$$

Also define $V^2_{\rho_\alpha, \gamma_\alpha} = (1+\alpha)^2 m^T K_\alpha(\theta_\alpha) m$ where

$$m = \left(\frac{\mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho^2} V(\theta_\alpha, Z_l)\Big] + \mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho \partial \gamma^T} V(\theta_\alpha, Z_l)\Big] F^{-1} \mathbb{E}_G\Big[\frac{\partial^2}{\partial \gamma \partial \rho} V(\theta_\alpha, Z_l)\Big]}{\Big(\mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho^2} V(\theta_\alpha, Z_l)\Big]\Big)^2} \quad , \frac{-\mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho \partial \gamma^T} V(\theta_\alpha, Z_l)\Big] F^{-1}}{\mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho^2} V(\theta_\alpha, Z_l)\Big]} \right)^T, \tag{3.41}$$

$$F = \mathbb{E}_G\Big[\frac{\partial^2}{\partial \gamma \partial \gamma^T} V(\theta_\alpha, Z_l)\Big] - \frac{\mathbb{E}_G\Big[\frac{\partial^2}{\partial \gamma \partial \rho} V(\theta_\alpha, Z_l)\Big] \mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho \partial \gamma^T} V(\theta_\alpha, Z_l)\Big]}{\mathbb{E}_G\Big[\frac{\partial^2}{\partial \rho^2} V(\theta_\alpha, Z_l)\Big]}. \tag{3.42}$$

Under appropriate regularity conditions Basu et al. (1998) prove that

$$\hat{\theta}_\alpha \xrightarrow{\mathbb{P}} \theta_\alpha \text{ and } \sqrt{n}(\hat{\theta}_\alpha - \theta_\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, J_\alpha^{-1}(\theta_\alpha) K_\alpha(\theta_\alpha) J_\alpha^{-1}(\theta_\alpha)\Big) \text{ as } n \to \infty. \tag{3.43}$$

Next, we shall particularly derive the marginal asymptotic distribution of $\sqrt{n}\hat{\rho}_\alpha$. Let us make the following assumptions.

(A1) A best-fitting parameter θ_α belongs to an open subset of Θ.

(A2) Assume that the matrix $J_\alpha(\theta_\alpha)$ defined in (3.39) is positive definite.

Theorem 3.3. *Under the Assumptions (A1) and (A2) a consistent sequence of roots $\hat{\rho}_\alpha$ exists such that*

$$\sqrt{n}(\hat{\rho}_\alpha - \rho_\alpha) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V^2_{\rho_\alpha,\gamma_\alpha}) \text{ as } n \to \infty. \tag{3.44}$$

Proof. Observe that the support of the model $\left\{(i,j) : \pi_{ij}(\theta) > 0\right\}$ does not depend on θ. Also, $\pi_{ij}(\theta)$s are differentiable with respect to each component of θ. So, $\sum_{i,j}\pi_{ij}^{1+\alpha}(\theta)$ and $\sum_{i,j}\pi_{ij}^{\alpha}(\theta)g_{ij}$ are differentiable with respect to θ, and the order of differentiation and summation can be interchanged. Thus, the Assumptions (D1) and (D3) of Basu et al. (1998) are trivially true in this case. Next, we shall show that each element of $|\nabla_{ll'k}V(\theta, Z_l)|$ is bounded by an integrable function in an open neighbourhood of θ_α. Here $\nabla_{ll'k}$ denotes the partial derivative with respect to the indicated components of the parameter θ. Note that $\phi_1'(x) = -x\phi_1(x)$ and $\phi_1''(x) = \phi_1(x)(x^2-1)$ where $\phi_1(x)$ is the density of $\mathcal{N}(0,1)$ at a point x. We obtain the third derivatives of $\Phi_2(x_1,x_2;\rho)$ as

$$\begin{aligned}\frac{\partial^3}{\partial\rho^3}\Phi_2(x_1,x_2;\rho) = \phi_2(x_1,x_2;\rho)^2&\left\{\frac{\rho}{1-\rho^2} - \frac{-x_1x_2(1-\rho^2)+\rho(x_1^2-2\rho x_1x_2+x_2^2)}{(1-\rho^2)^2}\right\}^2\\ &+\phi_2(x_1,x_2;\rho)\left\{\frac{1+\rho^2}{(1-\rho^2)^2} - \frac{(1+2\rho^2)(x_1^2+x_2^2)+(2\rho^3+4\rho^2-4\rho)x_1x_2}{(1-\rho^2)^3}\right\},\end{aligned} \tag{3.45}$$

$$\frac{\partial^3}{\partial x_1^3}\Phi_2(x_1,x_2;\rho) = \phi_1(x_1)\left[\frac{\rho}{\sqrt{1-\rho^2}}\phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\left\{\frac{2x_1-\rho^2x_1-\rho x_2}{1-\rho^2}\right\} + (x_1^2-1)\Phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\right], \tag{3.46}$$

$$\frac{\partial^3}{\partial\rho^2\partial x_1}\Phi_2(x_1,x_2;\rho) = -x_1\phi_1(x_1)\phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\left\{\frac{x_1x_2-\rho x_1^2}{(1-\rho^2)^{7/2}} - \frac{3}{2(1-\rho^2)^{5/2}}\right\}, \tag{3.47}$$

$$\frac{\partial^3}{\partial\rho\partial x_1^2}\Phi_2(x_1,x_2;\rho) = \phi_1(x_1)\phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\left\{\frac{x_1^2+1}{(1-\rho^2)^{3/2}} - \frac{x_1\rho(x_2-\rho x_1)}{(1-\rho^2)^{5/2}}\right\}, \tag{3.48}$$

$$\frac{\partial^3}{\partial x_2\partial x_1^2}\Phi_2(x_1,x_2;\rho) = \phi_1(x_1)\phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\left\{\frac{-x_1}{\sqrt{1-\rho^2}} + \frac{\rho(x_2-\rho x_1)}{(1-\rho^2)^{3/2}}\right\}, \tag{3.49}$$

$$\frac{\partial^3}{\partial\rho\partial x_2\partial x_1}\Phi_2(x_1,x_2;\rho) = x_1\phi_1(x_1)\phi_1\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)\left(\frac{x_2-\rho x_1}{(1-\rho^2)^{5/2}}\right). \tag{3.50}$$

In these above expressions x_1, x_2 varies over $(\eta_1,\ldots,\eta_{r-1},\beta_1,\ldots,\beta_{s-1})$. All these terms will be bounded at θ_α when $|\rho_\alpha| < 1$. Assumption (A1) ensures that $|\rho_\alpha| < 1$. Since these terms constitute $\nabla_{ll'k}V(\theta_\alpha, Z_l)$ which will be bounded as l, l', k varies over the indices of the parameter, Assumption (D5) of Basu et al. (1998) holds. Consequently consistency and asymptotic normality of $\hat{\theta}_\alpha$ follow. To derive the marginal distribution of $\hat{\rho}_\alpha$ we write

$$(1+\alpha)J_\alpha(\theta_\alpha) = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{3.51}$$

where

$$A = \mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho^2}V(\theta_\alpha,Z_l)\Big],\ B = \mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho\partial\gamma^T}V(\theta_\alpha,Z_l)\Big] \text{ and } D = \mathbb{E}_G\Big[\frac{\partial^2}{\partial\gamma\partial\gamma^T}V(\theta_\alpha,Z_l)\Big] \tag{3.52}$$

and $B = C^T$. A is positive since $J_\alpha(\theta_\alpha)$ is assumed to be positive definite. Using the formula of inverse for the block diagonal matrix, we obtain

$$J_\alpha^{-1}(\theta_\alpha) = (1+\alpha)\begin{bmatrix} A^{-1}+A^{-1}BF^{-1}CA^{-1} & -A^{-1}BF^{-1} \\ -F^{-1}CA^{-1} & F^{-1} \end{bmatrix} \tag{3.53}$$

where $F = D - CA^{-1}B$. The first row of $J_\alpha^{-1}(\theta_\alpha)$ is therefore given by

$$(1+\alpha)\underbrace{\left(A^{-1} + A^{-1}BF^{-1}CA^{-1}, -A^{-1}BF^{-1}\right)}_{m}. \tag{3.54}$$

Simple calculations give

$$A^{-1} + A^{-1}BF^{-1}CA^{-1} = \frac{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho^2}V(\theta_\alpha, Z_l)\right] + \mathbb{E}_G\left[\frac{\partial^2}{\partial\rho\partial\gamma^T}V(\theta_\alpha, Z_l)\right]F^{-1}\mathbb{E}_G\left[\frac{\partial^2}{\partial\gamma\partial\rho}V(\theta_\alpha, Z_l)\right]}{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho^2}V(\theta_\alpha, Z_l)\right]^2}, \tag{3.55}$$

$$-A^{-1}BF^{-1} = -\frac{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho\partial\gamma^T}V(\theta_\alpha, Z_l)\right]F^{-1}}{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho^2}V(\theta_\alpha, Z_l)\right]}, \tag{3.56}$$

$$F = \mathbb{E}_G\left[\frac{\partial^2}{\partial\gamma\partial\gamma^T}V(\theta_\alpha, Z_l)\right] - \frac{\mathbb{E}_G\left[\frac{\partial^2}{\partial\gamma\partial\rho}V(\theta_\alpha, Z_l)\right]\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho\partial\gamma^T}V(\theta_\alpha, Z_l)\right]}{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho^2}V(\theta_\alpha, Z_l)\right]}. \tag{3.57}$$

This completes the proof. □

In Figure 3.1 we plot the asymptotic variance of $\hat{\rho}_\alpha$ and see that it increases with the tuning parameter. This explains why the confidence intervals in the simulation studies become wider with increasing α under the true model.

3.3.3 Test Statistic

Now consider the problem of testing a simple null hypothesis against two-sided alternatives. The polychoric correlation functional is denoted by $T_\alpha(G) = \rho_\alpha$. We wish to test the statistical hypothesis

$$\mathbb{H} : \rho = r_h \text{ against } \mathbb{K} : \rho \neq r_h \text{ for some fixed } r_h. \tag{3.58}$$

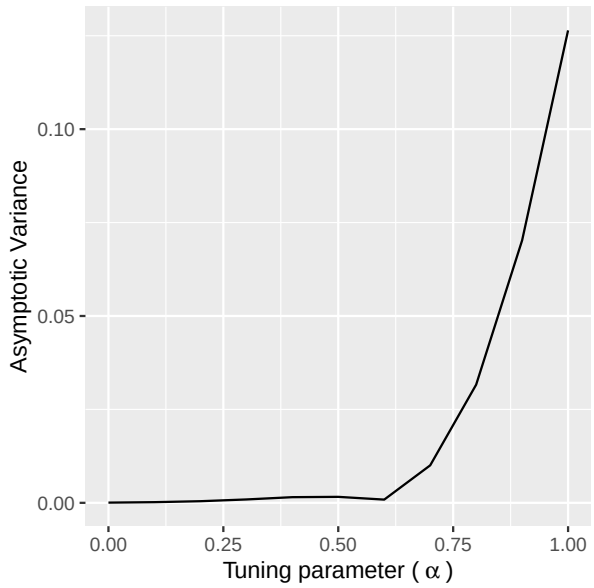

FIGURE 3.1: Asymptotic variance of the one-step estimator $\hat{\rho}_\alpha$ when the latent vector (U, V) is generated through standard bivariate normal distribution with $\rho = 0.75$, and the cut-offs are considered as $\eta = (-\infty, 0.7, 1.25, \infty)$, $\beta = (-\infty, -0.67, 0.67, \infty)$.

Let G_h be the true CDF under the null hypothesis $\mathbb{H}$ such that $T_\alpha(G_h) = r_h$. A family of Wald-type test statistics is given by

$$\widehat{W}_\alpha = \frac{n(\hat{\rho}_\alpha - r_h)^2}{V^2_{\hat{\rho}_\alpha, \hat{\gamma}_\alpha}}. \tag{3.59}$$

We reject the null hypothesis $\mathbb{H}$ for large values of $\widehat{W}_\alpha$ at fixed α. The exact null distribution and power function are very hard to obtain. However, their asymptotic distributions can be obtained for sufficiently large n. Under the assumptions of Theorem 3.3, it is easy to see that the asymptotic null distribution of $\widehat{W}_\alpha$ is central χ^2_1. Under two-sided alternatives the power function of the test statistic $\widehat{W}_\alpha$ at level-c is given by

$$\widehat{\xi}_\alpha(G) = \mathbb{P}\Big\{\widehat{W}_\alpha > \chi^2_{1,c} \text{ under true } G\Big\}, \tag{3.60}$$

where $\chi^2_{1,c}$ is the upper100c% point of central χ^2_1 distribution. Following Theorem 3.3, it can be easily shown that

$$\left|\widehat{\xi}_\alpha(G) - \left(1 - F_{\chi^2_1(\delta^2_n)}(\chi^2_{1,c})\right)\right| \longrightarrow 0 \text{ for } n \to \infty \text{ under } \mathbb{K}. \tag{3.61}$$

Here $F_{\chi^2(\delta^2_n)}$ is the CDF of the noncentral χ^2_1 distribution with the (n.c.p) noncentrality parameter $\delta^2_n = n\left[\frac{(T_\alpha(G)-r_h)}{V_{\rho_\alpha,\gamma_\alpha}}\right]^2$. When $\mathbb{H}$ is true, we get $\delta^2_n = 0$; so $\lim_{n\to\infty}\widehat{\xi}_\alpha(G_h) = c$. If the alternative hypothesis is true, the *n.c.p* depends on both n and the polychoric correlation functional. Therefore, it is impossible to calculate them explicitly. Suppose $\mathbb{K}$ is a simple alternative. Then there exists some $r_k \neq r_h$ such that $T_\alpha(P) \xrightarrow{\mathbb{P}} T_\alpha(G_k) = r_k$ for some G_k. The following theorem approximates the power of $\widehat{W}_\alpha$ under a simple alternative for large n. A more simplified expression is obtained under a sequence of contiguous alternatives. Before going into that, we define $q(\rho,\gamma) = \frac{(\rho-r_h)^2}{V^2_{\rho,\gamma}}$.

Theorem 3.4. *Suppose the assumptions of Theorem 3.3 hold, and G_k is defined as before. Assume $\frac{\partial q(\rho,\gamma)}{\partial\rho} \neq 0$ and $\frac{\partial^2 q(\rho,\gamma)}{\partial\rho^2}$ is bounded in a small neighbourhood of r_k and $\hat{\gamma}_\alpha$.*

(i) Then the power of $\widehat{W}_\alpha$ under the alternative $\mathbb{K}: \rho = r_k$ has the following limit:

$$\left|\widehat{\xi}_\alpha(G_k) - \left(1 - \Phi_1\Big(\frac{1}{\sigma(r_k)}\Big(\frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n}q(r_k,\hat{\gamma}_\alpha)\Big)\Big)\right)\right| \longrightarrow 0 \text{ as } n \to \infty \tag{3.62}$$

where $\sigma^2(r_k) = q'^2_*(r_k)V^2_{r_k,\gamma_\alpha(G_k)}$, and $q'_*(r_k) = \frac{\partial q(r_k,\gamma_\alpha(G_k))}{\partial\rho}$.

(ii) Consider the sequence of contiguous alternatives $\mathbb{K}_n: \rho = r_h + n^{-1/2}d$, where $d \neq 0$. If $\mathbb{K}_n \longrightarrow \mathbb{H}$ implies that $G_{k_n} \longrightarrow G_h$, then

$$\widehat{\xi}_\alpha(G_{k_n}) \longrightarrow 1 - F_{\chi^2_1(\delta^2)}(\chi^2_{1,c}), \tag{3.63}$$

where $\delta^2_n = \frac{d^2}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} \xrightarrow{\mathbb{P}} \frac{d^2}{V^2_{r_h,\gamma_\alpha(G_h)}} = \delta^2$ and G_{k_n} is the CDF associated with $\mathbb{K}_n$.

Proof. (i) See that $\widehat{W}_\alpha = nq(\hat{\rho}_\alpha, \hat{\gamma}_\alpha)$. The power of the test statistic $\widehat{W}_\alpha$ at $\mathbb{K} : \rho = r_k$ is given by

$$\begin{aligned}\widehat{\xi}_\alpha(G_k) &= \mathbb{P}_{G_k}\Big(\widehat{W}_\alpha > \chi^2_{1,c}\Big) \\ &= \mathbb{P}_{G_k}\Big[\sqrt{n}\Big(q(\hat{\rho}_\alpha, \hat{\gamma}_\alpha) - q(r_k, \hat{\gamma}_\alpha)\Big) > \frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n}q(r_k, \hat{\gamma}_\alpha)\Big]. \end{aligned} \tag{3.64}$$

A first-order Taylor series expansion of $\sqrt{n}q(\hat{\rho}_\alpha, \hat{\gamma}_\alpha)$ around r_k gives

$$\sqrt{n}\Big(q(\hat{\rho}_\alpha, \hat{\gamma}_\alpha) - q(r_k, \hat{\gamma}_\alpha)\Big) = \sqrt{n}(\hat{\rho}_\alpha - r_k)\Big[\frac{\partial q(r_k, \hat{\gamma}_\alpha)}{\partial \rho}\Big] + R_n, \tag{3.65}$$

where the remainder term in the expansion is given by

$$R_n = \sqrt{n}\frac{(\rho_\alpha - r_k)^2}{2}\left[\frac{\partial^2}{\partial \rho^2} q(\rho, \hat{\gamma}_\alpha)\right]_{\rho=\rho^*} \text{ as } \rho^* \text{ lies in } \hat{\rho}_\alpha, r_k. \tag{3.66}$$

Note that $\hat{\rho}_\alpha \xrightarrow{\mathbb{P}} r_k$ under $\mathbb{K}$, also $\hat{\gamma}_\alpha \xrightarrow{\mathbb{P}} \gamma_\alpha$. Using the boundedness condition of the second derivative, we can write the remainder term as

$$R_n = (\hat{\rho}_\alpha - r_k) \cdot \underbrace{\sqrt{n}(\hat{\rho}_\alpha - r_k)}_{\mathcal{O}_{\mathbb{P}}(1)} \cdot \mathcal{O}_{\mathbb{P}}(1) = o_{\mathbb{P}}(1). \tag{3.67}$$

Also, see that

$$\frac{\partial}{\partial \rho} q(r_k, \hat{\gamma}_\alpha) \xrightarrow{\mathbb{P}} \frac{\partial}{\partial \rho} q(r_k, \gamma_\alpha(G_k)) = q'_*(r_k). \tag{3.68}$$

Therefore

$$\sqrt{n}\left(q(\rho_\alpha, \hat{\gamma}_\alpha) - q(r_k, \hat{\gamma}_\alpha)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \underbrace{q'_*(r_k)^2 V^2_{r_k, \gamma_\alpha(G_k)}}_{\sigma^2(r_k)}\Big). \tag{3.69}$$

Thus we obtain

$$\left|\widehat{\xi}_\alpha(G_k) - \left\{1 - \Phi_1\Big(\frac{1}{\sigma(r_k)}\Big(\frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n}q(r_k, \hat{\gamma}_\alpha)\Big)\Big)\right\}\right| \longrightarrow 0 \text{ as } n \to \infty. \tag{3.70}$$

(ii) Denote $r_{k_n} = r_h + n^{-1/2}d$. So $r_{k_n} = T_\alpha(G_{k_n})$. Now see that

$$\begin{aligned}\widehat{W}_\alpha = \frac{n(\hat{\rho}_\alpha - r_h)^2}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} &= \frac{n(\hat{\rho}_\alpha - r_{k_n})^2}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} + \frac{n(r_{k_n} - r_h)^2}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} + 2\frac{n(\hat{\rho}_\alpha - r_{k_n})(r_{k_n} - r_h)}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} \\ &= \left(Z_n + \frac{d}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right)^2 \text{ where } Z_n = \frac{\sqrt{n}(\hat{\rho}_\alpha - r_{k_n})}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}.\end{aligned} \tag{3.71}$$

We assume that $\mathbb{K}_n \to \mathbb{H}$ implies that $G_{k_n} \to G_h$. Also $(\hat{\rho}_\alpha, \hat{\gamma}_\alpha) \xrightarrow{\mathbb{P}} (r_h, \gamma_\alpha(G_h))$ and

$$Z_n = \Big(\frac{V_{r_{k_n},\gamma_\alpha(G_{k_n})}}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\Big) \times U_n, \text{ where } U_n = \frac{\sqrt{n}(\hat{\rho}_\alpha - r_{k_n})}{V_{r_{k_n},\gamma_\alpha(G_{k_n})}}. \tag{3.72}$$

By Theorem 3.3 we know that

$$U_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ under } \mathbb{K}_n \text{ as } n \to \infty. \tag{3.73}$$

Also, the coefficient of U_n converges to 1 in probability. So applying the Slutsky's theorem we obtain $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Thus we obtain

$$\widehat{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^2) \text{ as } n \to \infty, \tag{3.74}$$

where $\delta^2 = \frac{d^2}{V^2_{r_h,\gamma_\alpha(G_h)}}$. This completes the proof.

□

However, it is still inevitable that a heavy computational burden will be pulled off to compute the power of the test at contiguous alternatives.

3.4 Influence Function Analysis

In this section, we shall study the stability behaviour of the proposed estimator and the test statistic when the true data-generating distribution becomes contaminated. This is mainly done by computing the influence functions of polychoric correlation functional and Wald-type test functional.

3.4.1 Influence Function of the Polychoric Correlation Functional

Under the true distribution $G = ((G_{ij}))$, the MDPD functional for the one-step scenario is given by

$$\theta_\alpha(G) = \arg\min_{\theta\in\Theta} d_\alpha(g, \pi(\theta)) \text{ where } \rho_\alpha = T_\alpha(G). \tag{3.75}$$

To find the influence function (IF) we consider the ϵ-contaminated version of the true density g as below:

$$g_{\epsilon,ij} = (1-\epsilon)g_{ij} + \epsilon\delta_{ij}(Z_l^o) \text{ for all } i, j \text{ and fixed } \epsilon \in [0,1]. \tag{3.76}$$

See that g is contaminated at a fixed point Z_l^o by a degenerate probability mass function with ϵ proportion. The contaminated CDF is similarly denoted by $G_{\epsilon,ij}$ for all i, j, and $G_\epsilon = ((G_{\epsilon,ij}))$. See that $\theta_\alpha(G_\epsilon)$ solves estimating equations

$$\mathbb{E}_{G_\epsilon} \nabla V(\theta_{\epsilon,\alpha}, Z_l) = 0 \text{ where } \theta_{\epsilon,\alpha} = \theta_\alpha(G_\epsilon). \tag{3.77}$$

Differentiating (3.77) with respect to ϵ, and evaluating at $\epsilon = 0$ gives the first-order influence function of $\theta_\alpha(G)$. From the standard result, we already know that

$$\mathcal{IF}_1(\theta_\alpha, G, Z_l^o) = \Big[\frac{\partial}{\partial \epsilon} T_\alpha(G_\epsilon)\Big]_{\epsilon=0} = -J_\alpha^{-1}(\theta_\alpha)\nabla V(\theta_\alpha, Z_l^o), \tag{3.78}$$

where Z_l^o varies over all the (i, j) cells. The first component of (3.78) gives the first-order influence function of $T_\alpha(G)$ at a point Z_l^o and true G as

$$\mathcal{IF}_1(T_\alpha, G, Z_l^o) = -(1+\alpha)m^T\nabla V(\theta_\alpha, Z_l^o) \text{ for all } Z_l^o, \tag{3.79}$$

where "m" is already defined in (3.41). The GES of the polychoric correlation functional $T_\alpha(G)$ is given by $GES(T_\alpha, G) = \max_{Z_l^o} |\mathcal{IF}_1(T_\alpha, G, Z_l^o)|$. Statistically, the GES quantifies the asymptotic bias of a functional due to data contamination. In Figure 3.2 we plot the gross error sensitivity (GES) of the polychoric correlation functional. In this graph, we find that the polychoric correlation functional becomes more stable under data contamination when the tuning parameter α moves further away from 0, because, in that case, the GES decreases.

Remark 3.6. ***(Influence function under reparametrization)*** *Consider a one-to-one transformation $\tau = \psi(\theta) : \Theta \to \mathbb{R}$ such that the MDPD functional under this transformation becomes $\tau_\alpha(G) = \psi(\theta_\alpha(G))$ under appropriate condition. A simple calculation shows that*

$$\mathcal{IF}_1(\tau_\alpha, G, Z_l^o) = \Big[\psi'(\theta_\alpha(G))\Big]^T \mathcal{IF}_1(\theta_\alpha, G, Z_l^o) \text{ for all } Z_l^o. \tag{3.80}$$

The boundedness of the influence function under reparametrization additionally depends on the boundedness of each component of the derivative ψ'. In particular, consider $\psi(\theta) = a^T\theta + b$ with $a \in \mathbb{R}^{r+s-1}$ and scalar b. Then the influence function is transformed into $\mathcal{IF}_1(\tau_\alpha, G, Z_l^o) = a^T\mathcal{IF}_1(\theta_\alpha, G, Z_l^o)$ which is bounded when both $||a|| < \infty$ and $||\mathcal{IF}_1(\theta_\alpha, G, Z_l^o)|| < \infty$.

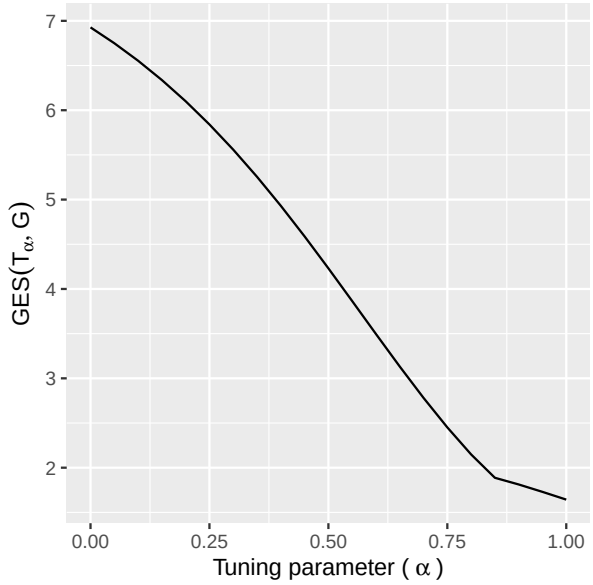

FIGURE 3.2: GES of the polychoric correlation functional when the latent vector (U, V) is generated through standard bivariate normal distribution with $\rho = 0.75$, and the cut-offs are $\eta = (-\infty, 0.7, 1.25, \infty)$ and $\beta = (-\infty, -0.67, 0.67, \infty)$.

3.4.2 Influence Function of the Wald-type Test Functional

Next we will study the stability behaviour of the Wald-type test functional corresponding to the test statistic defined in (3.59). The test functional ignoring the multiplier n is expressed as

$$W_\alpha(G) = \left[\frac{T_\alpha(G) - r_h}{V_{\hat{\rho}_\alpha, \hat{\gamma}_\alpha}}\right]^2 \text{ where } \alpha \in [0, 1]. \tag{3.81}$$

The Wald-type test functional at a contaminated true density is denoted as $W_\alpha(G_\epsilon)$. Differentiating $W_\alpha(G_\epsilon)$ with respect to ϵ, and evaluating at $\epsilon = 0$ gives the first-order

influence functions as

$$\mathcal{IF}_1(W_\alpha, G, Z_l^o) = \frac{\partial W_\alpha(G_\epsilon)}{\partial \epsilon}\bigg|_{\epsilon=0} = 2\left[\frac{T_\alpha(G) - r_h}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]\mathcal{IF}_1(T_\alpha, G, Z_l^o) \tag{3.82}$$

$$= -2\left[\frac{T_\alpha(G) - r_h}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right](1+\alpha)m^T\nabla V(\theta_\alpha, Z_l^o), \text{ for all } Z_l^o. \tag{3.83}$$

When the null hypothesis $\mathbb{H}$ is true, there exists a distribution function G_h such that $T_\alpha(G_h) = r_h$. Therefore $\mathcal{IF}_1(W_\alpha, G_h, Z_l^o) = 0$ for all Z_l^o and $\alpha \geq 0$. This phenomenon also holds for the test statistic based on the non-robust MLE which is in significant contrast to the strong stability behaviour exhibited by the observed levels derived from the proposed tests based on the MDPDE ($\alpha > 0$) (see Figure 3.10) as compared to the former. To exert further information regarding the robustness features under the null hypotheses $\mathbb{H}$, a second-order analysis may be more insightful (Lindsay, 1994; Basu et al., 2017). Differentiating a functional twice with respect to ϵ and evaluating at $\epsilon = 0$, gives its second-order influence function. Elaborate calculations give

$$\mathcal{IF}_2(W_\alpha, G, Z_l^o) = 2\left[\frac{T_\alpha(G) - r_h}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]\mathcal{IF}_2(T_\alpha, G, Z_l^o) + 2\left[\frac{\mathcal{IF}_1(T_\alpha, G, Z_l^o)}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]^2 \text{ for all } Z_l^o. \tag{3.84}$$

Under the null hypothesis $\mathbb{H}$, the expression of the second-order influence function becomes simplified.

Theorem 3.5. *The second-order influence functions of the Wald-type test functionals W_α under $\mathbb{H}$ is obtained as*

$$\mathcal{IF}_2(W_\alpha, G_h, Z_l^o) = 2\left[\frac{\mathcal{IF}_1(T_\alpha, G_h, Z_l^o)}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]^2 \textit{for all } Z_l^o. \tag{3.85}$$

The proof of Theorem 3.5 is omitted as it simply follows by substituting $T_\alpha(G_h)$ for $T_\alpha(G)$ in (3.84).

In Figure 3.3 we plot the GES of Wald-type test function defined as $GES_2(W_\alpha, G_h) = \max_{Z_l^o} |\mathcal{IF}_2(W_\alpha, G_h, Z_l^o)|$. It may be called the second-order GES. As before, in Figure 3.3, we see that the test functional W_α becomes much more stable as the tuning parameter α increases. This observation, in a way, validates the stability behaviour of the MDPD functional (for higher α) and the Wald-type test functional based on it.

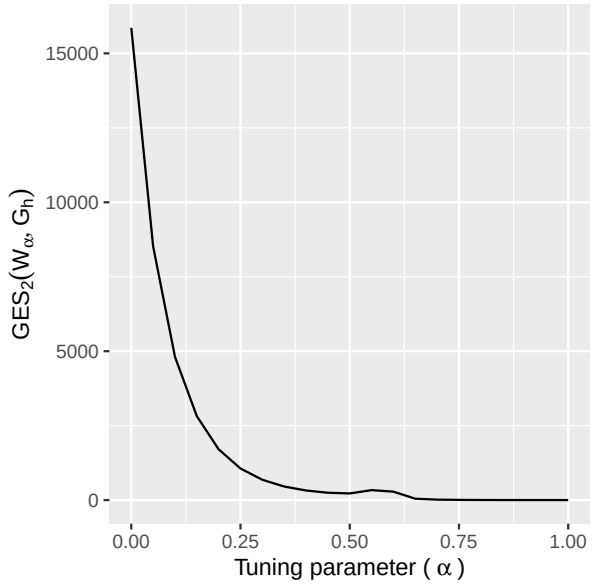

FIGURE 3.3: GES of the test functional W_α when the latent vector (U, V) is generated through standard bivariate normal distribution with $\rho = 0.75$, and the cut-offs are taken as $\eta = (-\infty, 0.7, 1.25, \infty)$, $\beta = (-\infty, -0.67, 0.67, \infty)$.

Things get harder if $G \neq G_h$ because in those cases we do not have simplified expressions for the second-order influence functions as the alternative hypothesis $\mathbb{K}$ is a composite one. An alternative way is to study the stability of the tests at the contiguous alternatives using the level and power influence functions presented in the next subsection.

3.4.3 Level and Power Influence Functions

We shall study the local stability of the type-I and type-II errors of the Wald-type test statistic. In particular, we shall calculate the first-order influence functions of the asymptotic level and power functions when the contaminated null and contaminated contiguous alternatives specify the true distributions. As before, the sequence of contiguous alternatives is given by $\mathbb{K}_n : \rho = r_h + \frac{d}{\sqrt{n}}$ for some constant $d \neq 0$.

Let $g_h = ((g_{h,ij}))$ and $g_{k_n} = ((g_{k_n,ij}))$ respectively denote the probability densities associated with $\mathbb{H}$ and $\mathbb{K}_n$. The null and alternative probability densities are contaminated at a point Z_l^o in the following way:

$$g^L_{h,\epsilon_n,ij} = \Big(1-\epsilon_n\Big)g_{h,ij} + \epsilon_n \delta_{ij}(Z_l^o) \text{ and } g^P_{k_n,\epsilon_n,ij} = \Big(1-\epsilon_n\Big)g_{k_n,ij} + \epsilon_n \delta_{ij}(Z_l^o), \tag{3.86}$$

where $\epsilon_n = n^{-1/2}\epsilon$ with $\epsilon > 0$. The CDFs corresponding to these densities are respectively denoted by G^L_{h,ϵ_n} and $G^P_{k_n,\epsilon_n}$. At the contaminated null and alternative distributions, the level and power of $\widehat{W}_\alpha$ at $100c\%$ nominal level of significance are given by

$$\alpha(G^L_{h,\epsilon_n}, Z_l^o) = \mathbb{P}_{G^L_{h,\epsilon_n}}\Big\{\widehat{W}_\alpha > \chi^2_{1,c}\Big\} \text{ and } \pi(G^P_{k_n,\epsilon_n}, Z_l^o) = \mathbb{P}_{G^P_{k_n,\epsilon_n}}\Big\{\widehat{W}_\alpha > \chi^2_{1,c}\Big\}. \tag{3.87}$$

Influence functions of these functionals (Rousseeuw et al., 2011) are defined as

$$\mathcal{LIF}(\alpha, G_h, Z_l^o) = \lim_{n\to\infty}\left[\frac{\partial \alpha(G^L_{h,\epsilon_n}, Z_l^o)}{\partial \epsilon}\right]_{\epsilon=0} \text{ and } \mathcal{PIF}(\pi, G_h, Z_l^o) = \lim_{n\to\infty}\left[\frac{\partial \pi(G^P_{k_n,\epsilon_n}, Z_l^o)}{\partial \epsilon}\right]_{\epsilon=0}. \tag{3.88}$$

To calculate $\mathcal{LIF}$ and $\mathcal{PIF}$, it is therefore required to find the large sample distributions of $\widehat{W}_\alpha$ under both G^L_{h,ϵ_n} and $G^P_{k_n,\epsilon_n}$. The first one is already derived in Theorem 3.4 (ii).

In the next result, we obtain the latter one. Let us define

$$\delta_{d,\epsilon} = \frac{d + \epsilon \mathcal{IF}_1(T_\alpha, G_h, Z_l^o)}{V^2_{\rho_\alpha,\gamma_\alpha(G_h)}}. \tag{3.89}$$

Recall that under G_h, the best-fitting parameter is $\rho_\alpha = r_h$.

Theorem 3.6. *Suppose T_α has a non-zero Hadamard derivative at G_{k_n}. Moreover, the assumptions of Theorem 3.4 hold. Then*

$$\widehat{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^2_{d,\epsilon}) \text{ under } G^P_{k_n,\epsilon_n} \tag{3.90}$$

as $n \to \infty$ with $\epsilon > 0$. The limiting distribution is noncentral chi-squared with the n.c.p $\delta^2_{d,\epsilon}$.

Proof. We express $\widehat{W}_\alpha$ as

$$\widehat{W}_\alpha = \frac{n(\hat{\rho}_\alpha - r_h)^2}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}} = S_{1n} + S_{2n} + S_{3n}, \tag{3.91}$$

where

$$S_{1n} = \left[\frac{\sqrt{n}(\hat{\rho}_\alpha - \rho^P_{n,\alpha})}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]^2, S_{2n} = \left[\frac{\sqrt{n}(\rho^P_{n,\alpha} - r_h)}{V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}\right]^2, S_{3n} = \frac{2n(\hat{\rho}_\alpha - \rho^P_{n,\alpha})(\rho^P_{n,\alpha} - r_h)}{V^2_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha}}, \tag{3.92}$$

and $(\rho^P_{n,\alpha}, \gamma^P_{n,\alpha}) = \arg\min_\Theta d_\alpha(g^P_{k_n,\epsilon_n}, \pi(\theta))$. When the true distribution is $G^P_{k_n,\epsilon_n}$, the term inside the squared bracket of S_{1n} is asymptotically $\mathcal{N}(0,1)$ by Theorem 3.4. So,

$S_{1n} \xrightarrow{\mathcal{L}} Z^2$ under $G^P_{k_n,\epsilon_n}$ where $Z \sim \mathcal{N}(0,1)$. The numerator of S_{2n} is expressed as

$$\begin{aligned}\Big[\sqrt{n}(\rho^P_{n,\alpha} - r_h)\Big]^2 &= \Bigg[\sqrt{n}\Big(T_\alpha(G^P_{k_n,\epsilon_n}) - T_\alpha(G_h)\Big)\Bigg]^2 \\ &= \Bigg[\underbrace{\sqrt{n}\Big(T_\alpha(G^P_{k_n,\epsilon_n}) - T_\alpha(G_{k_n})\Big)}_{S_{2na}} + \underbrace{\sqrt{n}\Big(T_\alpha(G_{k_n}) - T_\alpha(G_h)\Big)}_{S_{2nb}}\Bigg]^2. \end{aligned} \tag{3.93}$$

Let T_α be Hadamard differentiable (van der Vaart, 2000) at F_1. Then there exists a continuous, linear functional T'_{α,F_1} defined on the domain of T_α such that

$$\left|\frac{T_\alpha(F_1 + t_n H_n) - T_\alpha(F_1)}{t_n} - T'_{\alpha,F_1}(H)\right| \longrightarrow 0, \text{ when } t_n \downarrow 0+ \text{ for every } H_n \to H. \tag{3.94}$$

Substituting $F_1 = G_{k_n}$, $t_n = \frac{\epsilon}{\sqrt{n}}$ and $H_n = \Lambda(Z^o_l) - G_{k_n}$ in (3.94) gives

$$\left|\frac{T_\alpha\Big(G_{k_n} + \epsilon n^{-1/2}(\Lambda(Z^o_l) - G_{k_n})\Big) - T_\alpha(G_{k_n})}{\epsilon n^{-1/2}} - T'_{\alpha,G_{k_n}}\Big(\Lambda(Z^o_l) - G_h\Big)\right| \longrightarrow 0, \tag{3.95}$$

as $H_n \to H \equiv (\Lambda(Z^o_l) - G_h)$ and $\Lambda(Z^o_l)$ is the CDF of $\delta_{ij}(Z^o_l)$. Then it follows

$$\left|\sqrt{n}\Big(T_\alpha(G^P_{k_n,\epsilon_n}) - T_\alpha(G_{k_n})\Big) - \epsilon T'_{\alpha,G_{k_n}}\Big(\Lambda(Z^o_l) - G_h\Big)\right| \longrightarrow 0. \tag{3.96}$$

By continuity, $T'_{\alpha,G_{k_n}}(\Lambda(Z^o_l) - G_h) \to T'_{\alpha,G_h}(\Lambda(Z^o_l) - G_h)$. Since Hadamard differentiability implies the von Mises derivative, we also have

$$T'_{\alpha,G_h}\Big(\Lambda(Z^o_l) - G_h\Big) = \left[\frac{\partial}{\partial t} T_\alpha\Big(G_h + t(\Lambda(Z^o_l) - G_h)\Big)\right]_{t=0} = \mathcal{IF}_1(T_\alpha, G_h, Z^o_l). \tag{3.97}$$

Thus $S_{2na} = \epsilon\mathcal{IF}_1(T_\alpha, G_h, Z_l^o) + o_{\mathbb{P}}(1)$. We also know that $S_{2nb} = \sqrt{n}\Big(T_\alpha(G_{k_n}) - T_\alpha(G_h)\Big) = d$ as $\mathbb{K}_n : \rho = r_h + \frac{d}{\sqrt{n}}$. Further, see that

$$(G^P_{k_n,\epsilon_n} - G_{k_n}) \xrightarrow{\mathbb{P}} 0 \text{ and } (G_{k_n} - G_h) \longrightarrow 0 \Big[\text{ By assumption }\Big] \Longrightarrow (G^P_{k_n,\epsilon_n} - G_h) \xrightarrow{\mathbb{P}} 0 \tag{3.98}$$

as $n \to \infty$. Also note that $\Big(P - G^P_{k_n,\epsilon_n}\Big) \xrightarrow{\mathbb{P}} 0$ under $G^P_{k_n,\epsilon_n}$, which implies $P \xrightarrow{\mathbb{P}} G_h$ for $n \to \infty$. Hence $V_{\hat{\rho}_\alpha,\hat{\gamma}_\alpha} \xrightarrow{\mathbb{P}} V_{r_h,\gamma_\alpha(G_h)}$, and finally $S_{2n} \xrightarrow{\mathbb{P}} \delta^2_{d,\epsilon}$ for $n \to \infty$. Similarly, $S_{3n} \xrightarrow{\mathcal{L}} 2Z\delta_{d,\epsilon}$. Using Slutsky's theorem we finally get

$$\widehat{W}_\alpha \xrightarrow{\mathcal{L}} (Z + \delta_{d,\epsilon})^2 \text{ as } n \to \infty. \tag{3.99}$$

However, we know that $(Z + \delta_{d,\epsilon})^2$ have noncentral chi-squared distribution with 1 *degrees of freedom* (df) and n.c.p $\delta^2_{d,\epsilon}$. This completes the proof. □

Corollary 3.1. *Substituting $\epsilon = 0$ in Theorem 3.6, we get $\delta^2_{d,0} = \delta^2$ defined in Section 3.3.3. Thus, the power of the test under contiguous alternatives has the following limit*

$$\Big|\pi(G^P_{k_n,\epsilon_n}, Z_l^o) - \mathbb{P}\Big\{\chi^2_1(\delta^2) > \chi^2_{1,c}\Big\}\Big| \longrightarrow 0 \text{ as } n \to \infty. \tag{3.100}$$

Corollary 3.2. *When $d = 0$, we get $\sqrt{n}\Big(T_\alpha(G_{k_n}) - T_\alpha(G_h)\Big) = 0$. If the map $G \mapsto T_\alpha(G)$ is one-to-one, then $G_{k_n} \equiv G_h$. This implies that $G^P_{k_n,\epsilon_n} \equiv G^L_{h,\epsilon_n}$. Thus it easily follows form Theorem 3.6 that $\widehat{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^2_{0,\epsilon})$ under G^L_{h,ϵ_n}. Moreover, if we make $\epsilon \downarrow 0+$ along with $d = 0$, we get $\widehat{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1$ under G^L_{h,ϵ_n}. In that case, the following result holds*

$$\lim_{\epsilon \downarrow 0+} \Big|\alpha(G^L_{h,\epsilon_n}, Z_l^o) - c\Big| \longrightarrow 0 \text{ as } n \to \infty. \tag{3.101}$$

The exact values of the level and power under ϵ_n-contaminated versions of true null

and contiguous distributions are hard to compute. Only their limiting values can be obtained under appropriate assumptions. This makes it quite difficult for us to find the exact values of the functions $\mathcal{PIF}$ and $\mathcal{LIF}$ where n needs to be pushed towards infinity only after differentiation of level and power with respect to ϵ. Suppose the order of limit and differentiation can be interchanged. In that case, the exact expressions of $\mathcal{PIF}$ and $\mathcal{LIF}$ can be obtained with the aid of Theorem 3.6. In the next result, we explicitly mention these conditions when these simplifications hold.

Theorem 3.7. *Assume that both the sequences* $\left\{\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ *and* $\left\{\frac{\partial}{\partial\epsilon}\alpha(G^L_{h,\epsilon_n}, Z^o_l)\right\}$ *converge uniformly as a function of* ϵ *in* $[0,1]$ *at fixed* Z^o_l*. Further assume that the conditions of Theorem 3.6 are true.*

(i) Then the power influence function is obtained as

$$\mathcal{PIF}(\pi, G_h, Z^o_l) = \frac{d\cdot\mathcal{IF}_1(T_\alpha, G_h, Z^o_l)}{V^2_{r_h,\gamma_\alpha(G_h)}} e^{-\frac{\delta^2}{2}} \sum_{v=0}^{\infty} \frac{(\delta^2/2)^{v-1}}{v!}\Big(v-\frac{\delta^2}{2}\Big)\mathbb{P}\Big\{\chi^2_{1+2v} > \chi^2_{1,c}\Big\}. \tag{3.102}$$

(ii) The level influence function becomes $\mathcal{LIF}(\alpha, G_h, Z^o_l) \equiv 0$*.*

Proof. (i) When Z^o_l is fixed, $\pi(G^P_{k_n,\epsilon_n}, Z^o_l)$ is a sequence of functions on $[0,1]$. Using the result of Theorem 3.6, we know that the sequence $\pi(G^P_{k_n,\epsilon_n}, Z^o_l)$ is convergent to the following limit

$$\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n}, Z^o_l) = \sum_{v=0}^{\infty}\frac{e^{-\frac{\delta^2_{d,\epsilon}}{2}}}{v!}\Big(\frac{\delta^2_{d,\epsilon}}{2}\Big)^v\mathbb{P}\Big\{\chi^2_{1+2v} > \chi^2_{1,c}\Big\} \text{ for each } \epsilon\in[0,1]. \tag{3.103}$$

This along with the assumption of uniform convergence of $\left\{\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ imply the uniform convergence of the sequence of functions $\left\{\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ itself.

Hence the interchange of the limit and differentiation is permissible. So

$$\lim_{n\to\infty}\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n},Z^o_l)=\frac{\partial}{\partial\epsilon}\Big[\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n},Z^o_l)\Big]$$
$$=\frac{\partial}{\partial\epsilon}\Bigg[\sum_{v=0}^{\infty}\frac{e^{-\frac{\delta^2_{d,\epsilon}}{2}}}{v!}\Big(\frac{\delta^2_{d,\epsilon}}{2}\Big)^v\mathbb{P}\Big\{\chi^2_{1+2v}>\chi^2_{1,c}\Big\}\Bigg]. \tag{3.104}$$

Evaluating the right-hand side of (3.104) at $\epsilon=0$, we obtain

$$\mathcal{PIF}(\pi,G_h,Z^o_l)=\Bigg\{\frac{\partial}{\partial\epsilon}\Big[\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n},Z^o_l)\Big]\Bigg\}_{\epsilon=0}$$
$$=\frac{d\cdot\mathcal{IF}_1(T_\alpha,G_h,Z^o_l)}{V^2_{r_h,\gamma_\alpha(G_h)}}e^{-\frac{\delta^2}{2}}\sum_{v=0}^{\infty}\frac{(\delta^2/2)^{v-1}}{v!}\Big(v-\frac{\delta^2}{2}\Big)\mathbb{P}\Big\{\chi^2_{1+2v}>\chi^2_{1,c}\Big\}. \tag{3.105}$$

(ii) Putting $d=0$ in (3.105) gives $\mathcal{LIF}(\alpha,G_h,Z^o_l)\equiv 0$.

□

Remark 3.7. *When $\mathcal{IF}_1(T_\alpha,G_h,Z^o_l)$ is bounded, so is the $\mathcal{PIF}$. Hence the power of the test statistic based on MDPDE is stable at contaminated contiguous alternatives at the higher values of α. Also, note that $\mathcal{LIF}$ up to any order is always zero. So the influence function fails to reveal the robustness feature of the level at contaminated null distribution. However, we shall see that empirical levels in simulation studies are quite stable. But, this is not revealed by the influence function.*

3.5 Asymptotic Breakdown Point Analysis

In this section, we shall compute the asymptotic breakdown point of θ_α. Then, we will see how this changes under different kinds of reparametrization.

3.5.1 Asymptotic Breakdown Point

Let the true density g be contaminated at ϵ-proportion with a sequence of contaminating densities $\{k_M\}_{M=1}^{\infty}$ such that

$$h_{ij,\epsilon,M} = (1-\epsilon)g_{ij} + \epsilon k_{ij,M} \tag{3.106}$$

for $i = 1,\ldots,r$ and $j = 1,\ldots,s$. All these densities are assumed to be supported on the common set $\mathcal{S} = \{1,\ldots,r\} \times \{1,\ldots,s\}$. Under such a contamination, the MDPD functional becomes

$$\theta_{\alpha}^{h_{\epsilon,M}} := \arg\min_{\Theta} d_{\alpha}(h_{\epsilon,M}, \pi(\theta)) \text{ where } h_{\epsilon,M} = ((h_{ij,\epsilon,M})). \tag{3.107}$$

As before, define

$$D_{\alpha}\Big(g_{ij}, \pi_{ij}(\theta)\Big) = \left\{ \pi_{ij}(\theta)^{1+\alpha} - \Big(1+\frac{1}{\alpha}\Big)\pi_{ij}(\theta)^{\alpha} g_{ij} + \frac{1}{\alpha} g_{ij}^{1+\alpha} \right\} \text{ for } i,j. \tag{3.108}$$

Also, we introduce the following notations: $L_g = \sum_{i,j} g_{ij}^{1+\alpha}$, $L_{K_M} = \sum_{i,j} k_{ij,M}^{1+\alpha}$ and $L_{\pi(\theta)} = \sum_{i,j} \pi_{ij}^{1+\alpha}(\theta)$. Now, we are in the position to state our main result in this subsection regarding the asymptotic breakdown point with the help of the following assumptions. The number of cells of the $r \times s$ contingency table is fixed in this result.

(FB1) There exists a proper subset A (depending on M) and a non-negative number $\delta_A^*(i,j)$ such that

$$\Big(k_{ij,M} - \max\{g_{ij}, \pi_{ij}(\theta)\}\Big) \longrightarrow \delta_A^*(i,j) \geq 0 \text{ for all } (i,j) \in A \tag{3.109}$$

uniformly for $\|\theta\| < \infty$, satisfying $\sum_{(i,j)\in A} k_{ij,M} \to 1$ for $M \to \infty$. This means the contaminating density dominates the true and model densities on a proper subset

A of the support. The contaminating density is asymptotically concentrated on the set A itself.

(FB2) There exists a proper subset B (depending on M) and a non-negative number $\delta_B^*(i,j)$ such that

$$\Big(\pi_{ij}(\theta_M) - g_{ij}\Big) \longrightarrow \delta_B^*(i,j) \geq 0 \text{ for all } (i,j) \in B \tag{3.110}$$

satisfying $\sum_{(i,j)\in B} \pi_{ij}(\theta_M) \to 1$ when $\|\theta_M\| \to \infty$ for $M \to \infty$. This means the model density, evaluated at a divergent sequence θ_M, dominates the true density on the proper subset B. Asymptotically, the former density also concentrates on the set B itself.

(FB3) (Extremity of Contamination) $\liminf_{M\to\infty} L_{\pi(\theta_M)} \geq \limsup_{M\to\infty} L_{k_M}$ for any sequence $\{\theta_M\}$ with $\|\theta_M\| \to \infty$ as $M \to \infty$.

(FB4) $L_{\pi(\theta_\alpha)} \leq L_g \leq \sum_{i,j} g_{ij}\pi_{ij}^\alpha(\theta_\alpha)$ for all $\alpha > 0$.

Assumptions (FB1) and (FB2) describe those situations when different finitely-supported densities involved could have been singular to one another in this setup. Assumption (FB3) allows a certain kind of contaminating densities that should be involved to come up with the following result. Also, Assumption (FB4) is a technical that is required to compute the asymptotic breakdown point.

Theorem 3.8. *Under the Assumptions (FB1) - (FB4), the asymptotic breakdown point of the MDPD functional θ_α at the model is at least $\tilde{\epsilon}$ where*

$$\tilde{\epsilon} = \min\left\{\frac{\alpha}{1+\alpha}, \frac{1}{1+\alpha}\right\} \text{ at fixed } \alpha > 0. \tag{3.111}$$

Proof. Set $\theta_M = \theta_\alpha^{h_\epsilon,M}$. Notice that $\|\theta_\alpha^{h_\epsilon,M}\| \to \infty$ when $M \to \infty$. Let us define the following set

$$A_M = \Big\{(i,j) : g_{ij} > \max\big\{k_{ij,M}, \pi_{ij}(\theta_M)\big\}\Big\}. \tag{3.112}$$

The divergence between $h_{\epsilon,M}$ and $\pi(\theta_M)$ can be decomposed as

$$d_\alpha(h_{\epsilon,M}, \pi(\theta_M)) = \sum_{(i,j)\in A_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) + \sum_{(i,j)\in A_M^c} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big). \tag{3.113}$$

It is easy to see that $A_M \subset A^c$ when the Assumption (FB1) is true. Therefore, we get

$$\sum_{(i,j)\in A_M} k_{ij,M} \le \sum_{(i,j)\in A^c} k_{ij,M} \to 0 \text{ as } M \to \infty. \tag{3.114}$$

Similarly, it follows from Assumption (FB2) that $A_M \subset B^c$ and

$$\sum_{(i,j)\in A_M} \pi_{ij}(\theta_M) \le \sum_{(i,j)\in B^c} \pi_{ij}(\theta_M) \to 0 \text{ as } M \to \infty. \tag{3.115}$$

Therefore, we get

$$D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) = \left\{\pi_{ij}(\theta_M)^{1+\alpha} - \Big(1+\frac{1}{\alpha}\Big)\pi_{ij}(\theta_M)^\alpha h_{ij,\epsilon,M} + \frac{1}{\alpha}h_{ij,\epsilon,M}^{1+\alpha}\right\} \tag{3.116}$$

$$\longrightarrow \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} g_{ij}^{1+\alpha} = D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big) \tag{3.117}$$

for all $(i,j) \in A_M$ when $M \to \infty$. Now, applying the DCT gives the following

$$\left|\sum_{(i,j)\in A_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) - \sum_{(i,j)\in A_M} D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big)\right| \longrightarrow 0 \text{ as } M \to \infty. \tag{3.118}$$

The Assumptions (FB1) and (FB2) together implies that

$$\left|\sum_{(i,j)\in A_M} D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big) - \sum_{(i,j)\in \mathcal{S}} D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big)\right| \longrightarrow 0 \text{ when } M \to \infty. \tag{3.119}$$

A simple application of the triangle inequality gives the following result

$$\left| \sum_{(i,j)\in A_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) - \sum_{(i,j)\in\mathcal{S}} D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big) \right| \longrightarrow 0 \text{ for } M \to \infty. \tag{3.120}$$

Now, see that

$$\sum_{(i,j)\in\mathcal{S}} D_\alpha\Big((1-\epsilon)g_{ij}, 0\Big) = \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} \sum_{(i,j)\in\mathcal{S}} g_{ij}^{1+\alpha} = \frac{1}{\alpha}(1-\epsilon)^{1+\alpha} L_g. \tag{3.121}$$

Therefore, we arrive at

$$\sum_{(i,j)\in A_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) \longrightarrow \frac{(1-\epsilon)^{1+\alpha}}{\alpha} L_g \text{ when } M \to \infty. \tag{3.122}$$

Also, see that

$$\sum_{(i,j)\in A_M} g_{ij} \longrightarrow \sum_{(i,j)\in\mathcal{S}} g_{ij} = 1 \text{ when } M \to \infty, \tag{3.123}$$

by Assumptions (FB1) and (FB2). Since both $\sum_{(i,j)\in A_M^c} k_{ij,M} \to 1$ and $\sum_{(i,j)\in A_M^c} \pi_{ij}(\theta_M) \to 1$, we also have

$$\left| \sum_{(i,j)\in A_M^c} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) - \sum_{(i,j)\in\mathcal{S}} D_\alpha\Big(\epsilon k_{ij,M}, \pi_{ij}(\theta_M)\Big) \right| \tag{3.124}$$

$$= \left| \sum_{(i,j)\in A_M^c} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_M)\Big) - d_\alpha\Big(\epsilon k_M, \pi(\theta_M)\Big) \right| \to 0 \text{ as } M \to \infty. \tag{3.125}$$

Therefore, we get

$$\liminf_{M\to\infty}\Big[d_\alpha(h_{\epsilon,M}, \pi(\theta_M))\Big] \geq \liminf_{M\to\infty}\Big[d_\alpha(\epsilon k_M, \pi(\theta_M))\Big] + \frac{(1-\epsilon)^{1+\alpha}}{\alpha} L_g = a_1(\epsilon). \tag{3.126}$$

Define $B_M = \left\{(i,j) : k_{ij,M} \geq \max\{g_{ij}, \pi_{ij}(\theta_\alpha)\}\right\}$ where $\|\theta_\alpha\| < \infty$. See that the contaminating density dominates both the true and its model (evaluated at the best-fitting parameter θ_α) densities on the set B_M, hence $A = B_M$ as in Assumption (FB1). Thus we get

$$\sum_{(i,j)\in A} \max\{g_{ij}, \pi_{ij}(\theta_\alpha)\} \to 0 \text{ as } M \to \infty, \tag{3.127}$$

by Assumption (FB1). Notice that the set A, chosen as above, depends on M. See that $\sum_{(i,j)\in A^c} k_{ij,M} \to 0$ as $M \to \infty$. Thus, for all $(i,j) \in B_M$,

$$\left| D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_\alpha)\Big) - D_\alpha\Big(\epsilon k_{ij,M}, 0\Big)\right| \longrightarrow 0 \tag{3.128}$$

$$\implies \left| \sum_{(i,j)\in B_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_\alpha)\Big) - \sum_{k_{ij,M}>0} D_\alpha\Big(\epsilon k_{ij,M}, 0\Big)\right| \longrightarrow 0 \left[\text{ DCT }\right]. \tag{3.129}$$

Observe that $D_\alpha\Big(\epsilon k_{ij,M}, 0\Big) = \frac{\epsilon^{1+\alpha}}{\alpha} k_{ij,M}^{1+\alpha}$ when $\alpha > 0$. Therefore it follows that

$$\left| \sum_{(i,j)\in B_M} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_\alpha)\Big) - \frac{\epsilon^{1+\alpha}}{\alpha} \sum_{i,j} k_{ij,M}^{1+\alpha}\right| \longrightarrow 0 \text{ as } M \to \infty. \tag{3.130}$$

Similarly, we have

$$\left| \sum_{(i,j)\in B_M^c} D_\alpha\Big(h_{ij,\epsilon,M}, \pi_{ij}(\theta_\alpha))\Big) - \underbrace{\sum_{i,j} D_\alpha\Big((1-\epsilon)g_{ij}, \pi_{ij}(\theta_\alpha)\Big)}_{d_\alpha((1-\epsilon)g, \pi(\theta_\alpha))}\right| \longrightarrow 0 \text{ when } M \to \infty. \tag{3.131}$$

In Assumption (FB4) we have

$$L_{\pi(\theta_\alpha)} \leq L_g \leq \sum_{i,j} g_{ij} \pi_{ij}^\alpha(\theta_\alpha) \tag{3.132}$$

for all $\alpha > 0$. This gives

$$\limsup_{M\to\infty}\Big[d_\alpha\Big(h_{\epsilon,M}, \pi(\theta_\alpha)\Big)\Big] \leq a_2(\epsilon), \tag{3.133}$$

where

$$a_2(\epsilon) = \frac{\epsilon^{1+\alpha}}{\alpha}\limsup_{M\to\infty} L_{k_M} + \left[1 - \left(1 + \frac{1}{\alpha}\right)(1-\epsilon) + \frac{(1-\epsilon)^{1+\alpha}}{\alpha}\right] L_g. \tag{3.134}$$

Asymptotically there will be no breakdown at ϵ-contamination when $a_2(\epsilon) < a_1(\epsilon)$. This means

$$\liminf_{M\to\infty}\Big[d_\alpha(\epsilon k_M, \pi(\theta_M))\Big] > \frac{\epsilon^{1+\alpha}}{\alpha}\limsup_{M\to\infty} L_{k_M} + \left[1 - \left(1 + \frac{1}{\alpha}\right)(1-\epsilon)\right] L_g. \tag{3.135}$$

See that (3.135) is equivalent to the following:

$$d_\alpha(\epsilon k_M, \pi(\theta_M)) > \frac{\epsilon^{1+\alpha}}{\alpha} L_{k_M} + \left[1 - \left(1 + \frac{1}{\alpha}\right)(1-\epsilon)\right] L_g \tag{3.136}$$

for sufficiently large M along any sequence of contaminating densities $\{k_M\}$. See that we have $\liminf_{M\to\infty} L_{\pi(\theta_M)} \geq \limsup_{M\to\infty} L_{k_M}$ in Assumption (FB3). This, in turn, implies that $L_{\pi(\theta_M)} \geq L_{k_M}$ for all M. This yields

$$\begin{aligned}
d_\alpha(\epsilon k_M, \pi(\theta_M)) &= \sum_{i,j}\left[\pi_{ij}^{1+\alpha}(\theta_M) - \epsilon\left(1 + \frac{1}{\alpha}\right)\pi_{ij}^{\alpha}(\theta_M) k_{ij,M} + \frac{\epsilon^{1+\alpha}}{\alpha} k_{ij,M}^{1+\alpha}\right] \\
&= L_{\pi(\theta_M)} - \epsilon\left(1 + \frac{1}{\alpha}\right)\sum_{i,j}\pi_{ij}^{\alpha}(\theta_M) k_{ij,M} + \frac{\epsilon^{1+\alpha}}{\alpha} L_{k_M} \\
&\geq L_{\pi(\theta_M)} - \epsilon\left(1 + \frac{1}{\alpha}\right)\Big(\sum_{i,j}\pi_{ij}^{1+\alpha}(\theta_M)\Big)^{\frac{\alpha}{1+\alpha}}\left(\sum_{i,j} k_{ij,M}^{1+\alpha}\right)^{\frac{1}{1+\alpha}} + \frac{\epsilon^{1+\alpha}}{\alpha} L_{k_M} \text{ (Hölder's inequality)} \\
&= L_{\pi(\theta_M)} - \epsilon\left(1 + \frac{1}{\alpha}\right) L_{\pi(\theta_M)}^{\frac{\alpha}{1+\alpha}} L_{k_M}^{\frac{1}{1+\alpha}} + \frac{\epsilon^{1+\alpha}}{\alpha} L_{k_M}.
\end{aligned} \tag{3.137}$$

Since we have $L_{\pi(\theta_M)} \geq L_{k_M}$, thus

$$L_{\pi(\theta_M)} - \epsilon\left(1+\frac{1}{\alpha}\right)L_{\pi(\theta_M)}^{\frac{\alpha}{1+\alpha}}L_{k_M}^{\frac{1}{1+\alpha}} + \frac{\epsilon^{1+\alpha}}{\alpha}L_{k_M} \geq \left\{1-\epsilon\left(1+\frac{1}{\alpha}\right)+\frac{\epsilon^{1+\alpha}}{\alpha}\right\}L_{k_M}$$
$$= d_\alpha(\epsilon k_M, k_M). \tag{3.138}$$

This gives $d_\alpha(\epsilon k_M, \pi(\theta_M)) \geq d_\alpha(\epsilon k_M, k_M)$ for all M. The condition as in (3.135) that there will be no asymptotic breakdown will be satisfied when the following holds for sufficiently large M along any sequence such that

$$\begin{aligned} & d_\alpha(\epsilon k_M, k_M) \geq \frac{\epsilon^{1+\alpha}}{\alpha}L_{k_M} + \left[1-\left(1+\frac{1}{\alpha}\right)(1-\epsilon)\right]L_g \\ \iff & \left\{1-\epsilon\left(1+\frac{1}{\alpha}\right)+\frac{\epsilon^{1+\alpha}}{\alpha}\right\}L_{k_M} \geq \frac{\epsilon^{1+\alpha}}{\alpha}L_{k_M} + \left[1-\left(1+\frac{1}{\alpha}\right)(1-\epsilon)\right]L_g \\ \iff & \left\{1-\epsilon\left(1+\frac{1}{\alpha}\right)\right\}L_{k_M} \geq \left[1-\left(1+\frac{1}{\alpha}\right)(1-\epsilon)\right]L_g \\ \iff & \frac{L_{k_M}}{L_g}\left(\epsilon-\frac{\alpha}{1+\alpha}\right)+\left(\epsilon-\frac{1}{1+\alpha}\right) \leq 0. \end{aligned} \tag{3.139}$$

Clearly, (3.139) holds when $\epsilon < \tilde{\epsilon}$ where

$$\tilde{\epsilon} = \min\left\{\frac{\alpha}{1+\alpha}, \frac{1}{1+\alpha}\right\}. \tag{3.140}$$

This completes the proof. □

This result shows that the robustness of θ_α increases with α, which is clear from Figure 3.4.

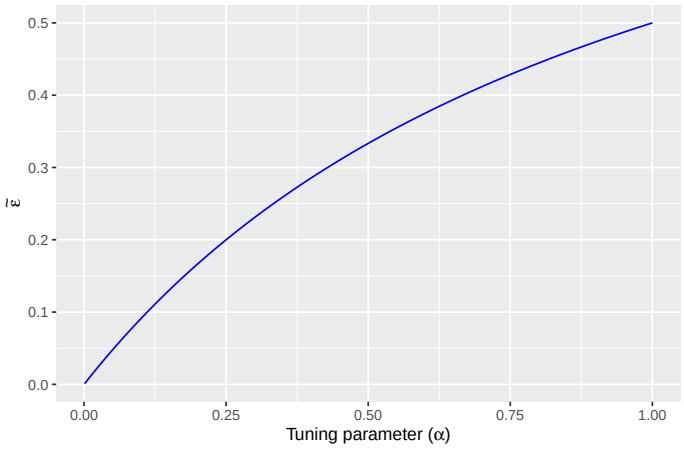

FIGURE 3.4: The plots of $\tilde{\epsilon}$ with respect to α

3.5.2 Asymptotic Breakdown Point Under Reparametrization

As earlier, the asymptotic breakdown point of $\psi(\theta_\alpha)$ is defined to be

$$\tilde{\epsilon}^{\psi} = \inf\left\{\epsilon \in [0,1] : ||\psi(\theta_\alpha) - \psi(\theta_\alpha^{h_\epsilon, M})|| \to \infty \text{ as } M \to \infty\right\}. \tag{3.141}$$

First, we present a sufficient condition that establishes the invariance of the asymptotic breakdown point under an affine transformation. Subsequently, we study how the asymptotic breakdown point gets impacted under reparametrization in different directions. Assume that $\tilde{\epsilon}$ be the asymptotic breakdown point of θ_α.

Theorem 3.9. *Consider the affine transformation $\psi(\theta) = B\theta + C$. Let $\lambda_{\min}$ and λ_{max} be the minimum and maximum eigenvalues of (B^TB). Then, we have the following results.*

(a) The asymptotic breakdown point remains invariant when $|B^TB| \neq 0$, or $\lambda_{\min} > 0$.

(b) The asymptotic breakdown point increases, that is $\tilde{\epsilon} \leq \tilde{\epsilon}^{\psi}$ when $|B^TB| = 0$ but $\lambda_{max} > 0$.

Proof. Define the following sets:

$$S_1 = \Big\{\epsilon \in [0,1] : ||\theta_\alpha - \theta_\alpha^{h_\epsilon,M}|| \to \infty \text{ as } M \to \infty\Big\}, \tag{3.142}$$

$$S_2 = \Big\{\epsilon \in [0,1] : ||\psi(\theta_\alpha) - \psi(\theta_\alpha^{h_\epsilon,M})|| \to \infty \text{ as } M \to \infty\Big\}. \tag{3.143}$$

Clearly, $\theta_\alpha \neq \theta_\alpha^{h_\epsilon,M}$. See that

$$\lambda_{\min} \leq \frac{||\psi(\theta_\alpha^{h_\epsilon,M}) - \psi(\theta_\alpha)||^2}{||\theta_\alpha^{h_\epsilon,M} - \theta_\alpha||^2} = \frac{(\theta_\alpha^{h_\epsilon,M} - \theta_\alpha)^T B^T B(\theta_\alpha^{h_\epsilon,M} - \theta_\alpha)}{(\theta_\alpha^{h_\epsilon,M} - \theta_\alpha)^T(\theta_\alpha^{h_\epsilon,M} - \theta_\alpha)} \leq \lambda_{\max}. \tag{3.144}$$

(a) Clearly, we have

$$\sqrt{\lambda_{\min}}||\theta_\alpha^{h_\epsilon,M} - \theta_\alpha|| \leq ||\psi(\theta_\alpha^{h_\epsilon,M}) - \psi(\theta_\alpha)|| \leq \sqrt{\lambda_{\max}}||\theta_\alpha^{h_\epsilon,M} - \theta_\alpha||. \tag{3.145}$$

If the determinant $|B^T B|$ is non-zero, we have $0 < \lambda_{\min} < \lambda_{\max}$ and $S_1 = S_2$. This gives $\tilde{\epsilon} = \tilde{\epsilon}^\psi$.

(b) In the second case, we have $||\psi(\theta_\alpha^{h_\epsilon,M}) - \psi(\theta_\alpha)|| \leq \sqrt{\lambda_{\max}}||\theta_\alpha^{h_\epsilon,M} - \theta_\alpha||$. This gives $S_2 \subset S_1$, consequently $\tilde{\epsilon} \leq \tilde{\epsilon}^\psi$.

□

Example 3.1. *Suppose we have*

$$\frac{||\psi(\theta_\alpha) - \psi(\theta_\alpha^{h_\epsilon,M})||}{||\theta_\alpha - \theta_\alpha^{h_\epsilon,M}||} = \mathcal{O}(M^k). \tag{3.146}$$

See that $S_1 \subset S_2$ according to $k > 0$. This gives $\tilde{\epsilon}^\psi < \tilde{\epsilon}$. Similarly, $\tilde{\epsilon} < \tilde{\epsilon}^\psi$ or $\tilde{\epsilon} = \tilde{\epsilon}^\psi$ according to $k < 0$ or $k = 0$. Next, we prove similar results for the Lipschitz functions.

Theorem 3.10. *Suppose $||\psi_k(\theta_1) - \psi_k(\theta_2)|| \leq C_k||\theta_1 - \theta_2||$ for all θ_1, θ_2 with some constant $C_k > 0$ for $k = 1, 2$. Then the following results are true:*

(i) $\tilde{\epsilon} \leq \min\{\tilde{\epsilon}^{\psi_1}, \tilde{\epsilon}^{\psi_2}\}$,

(ii) $\tilde{\epsilon} \leq \tilde{\epsilon}^{\psi}$ *where* $\psi(\theta) = \psi_1(\psi_2(\theta))$ *given such a composition is valid,*

(iii) $\tilde{\epsilon} < \tilde{\epsilon}^{a\psi_1 + b\psi_2}$ *for any finite a and b not zero simultaneously,*

Proof. (i) Using the same argument as before, we get $\tilde{\epsilon} \leq \tilde{\epsilon}^{\psi_k}$ for $k = 1, 2$. Hence, we have $\tilde{\epsilon} \leq \min\{\tilde{\epsilon}^{\psi_1}, \tilde{\epsilon}^{\psi_2}\}$.

(ii) Notice that

$$||\psi_1(\psi_2(\theta_1)) - \psi_1(\psi_2(\theta_2))|| \leq C_1||\psi_2(\theta_1) - \psi_2(\theta_2)|| \leq C_1C_2||\theta_1 - \theta_2|| \tag{3.147}$$

for all θ_1, θ_2. Using the same argument as before.

(iii) See that

$$||\big(a\psi_1(\theta_1) + b\psi_2(\theta_1)\big) - \big(a\psi_1(\theta_2) + b\psi_2(\theta_2)\big)|| \leq \big(|a|C_1 + |b|C_2\big) \cdot ||\theta_1 - \theta_2|| \tag{3.148}$$

for all θ_1, θ_2. Hence the result follows.

□

We can use the induction to prove the same result for any finite value of k in the last result.

3.6 Simulation Studies

Here we present numerical results based on some simulation studies. Random samples of size $N = 100$ are drawn from $\mathcal{N}_2(0, \Sigma)$ where the matrix of correlation coefficients is given by

$$\Sigma = \begin{pmatrix} 1, & 0.75 \\ 0.75, & 1 \end{pmatrix}. \tag{3.149}$$

Given a set of observations, a contingency table may be created using the following cut-offs: $\eta = (-\infty, -0.70, 1.25, \infty)$ and $\beta = (-\infty, -0.67, 0.67, \infty)$. We use the following naming conventions for different types of data sets.

Type 0 - A data set is called Type 0 when it truly follows the model. In other words, we may call it a pure data set.

Type 1 - In this case, data points are moved from the most to the least probable cell.

Type 2 - Here data points are randomly moved from the most probable to any other cell.

A data set such as $100\epsilon\%$ Type 1 cont. (or, contamination) means– $100\epsilon\%$ observations of the contingency table (that follows the model) are moved from the most probable cell to the least probable cell; $\epsilon = 0.05, 0.10$. These experiments are repeated over 200 replications. We also test the following statistical hypothesis

$$\mathbb{H} : \rho = 0.75 \text{ against } \mathbb{K} : \rho = 0.50 \tag{3.150}$$

at 5% nominal level of significance. In Figure 3.5 we see that small values of the tuning parameters produce estimates that are almost as good as the MLE in terms of having

TABLE 3.2: One-step estimates of the polychoric correlation

Methods	Pure data	5%Type 1 cont.	10%Type 1 cont.	5%Type 2 cont.	10%Type 2 cont.
$\hat{\rho}_0$	0.74802	0.48797	0.32604	0.66981	0.61063
$\hat{\rho}_{0.2}$	0.74498	0.54672	0.35519	0.68353	0.62629
$\hat{\rho}_{0.4}$	0.74476	0.62921	0.4254	0.69566	0.64237
$\hat{\rho}_{0.6}$	0.74504	0.68132	0.54007	0.70263	0.65397
$\hat{\rho}_{0.9}$	0.74477	0.70141	0.62682	0.70643	0.66218
HD	0.7664	0.65302	0.52291	0.70148	0.63983
NCS	0.7383	0.67981	0.56052	0.69895	0.65475
SCS	0.75742	0.71402	0.65564	0.70036	0.64336
NED	0.7500	0.70291	0.54874	0.69138	0.63380

very low bias and MSE. We also find that the observed levels and powers (albeit a bit lower for higher α) are quite stable across all the values of the tuning parameters. This trend is also observed in the confidence intervals of $\hat{\rho}_\alpha$ as in Figure 3.6. Under different types of data contamination, a higher value of α adds much stability to the performance of both the estimates (Figure 3.7, 3.8, 3.9) and the tests (Figure 3.10, 3.11).

To compare with the other robust methods such as– SCS (symmetric chi-square), NCS (Neyman chi-square), HD (Hellinger distance), and NED (Negative exponential disparity), we present the parameter estimates in Table 3.2. We see that the methods that perform better than MDPDE in contaminated data sets may work badly for pure data, and, the other way around. Thus, MDPDE may win over the different methods for performing reasonably well in pure and contaminated data sets.

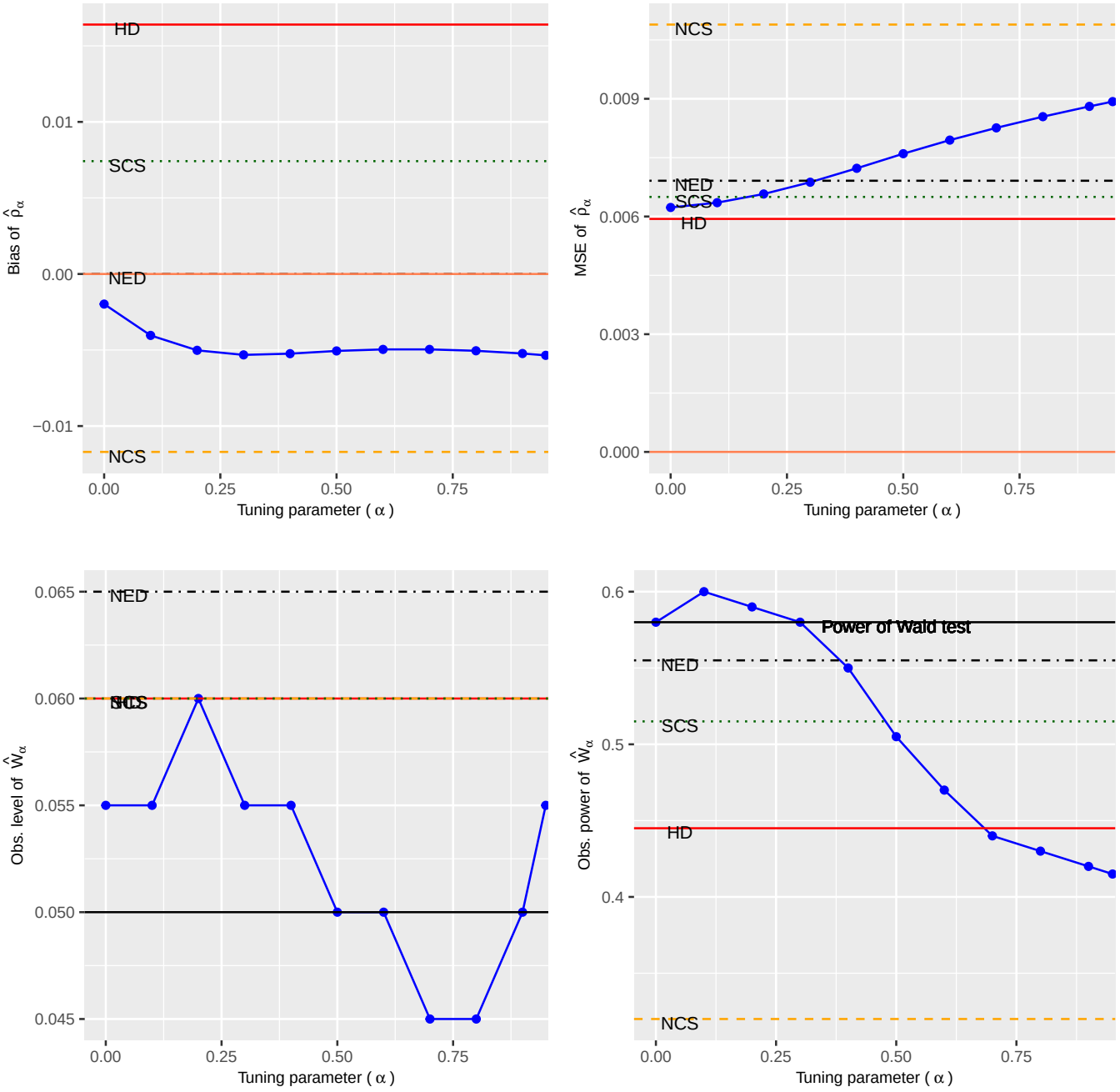

FIGURE 3.5: Plots under pure data.

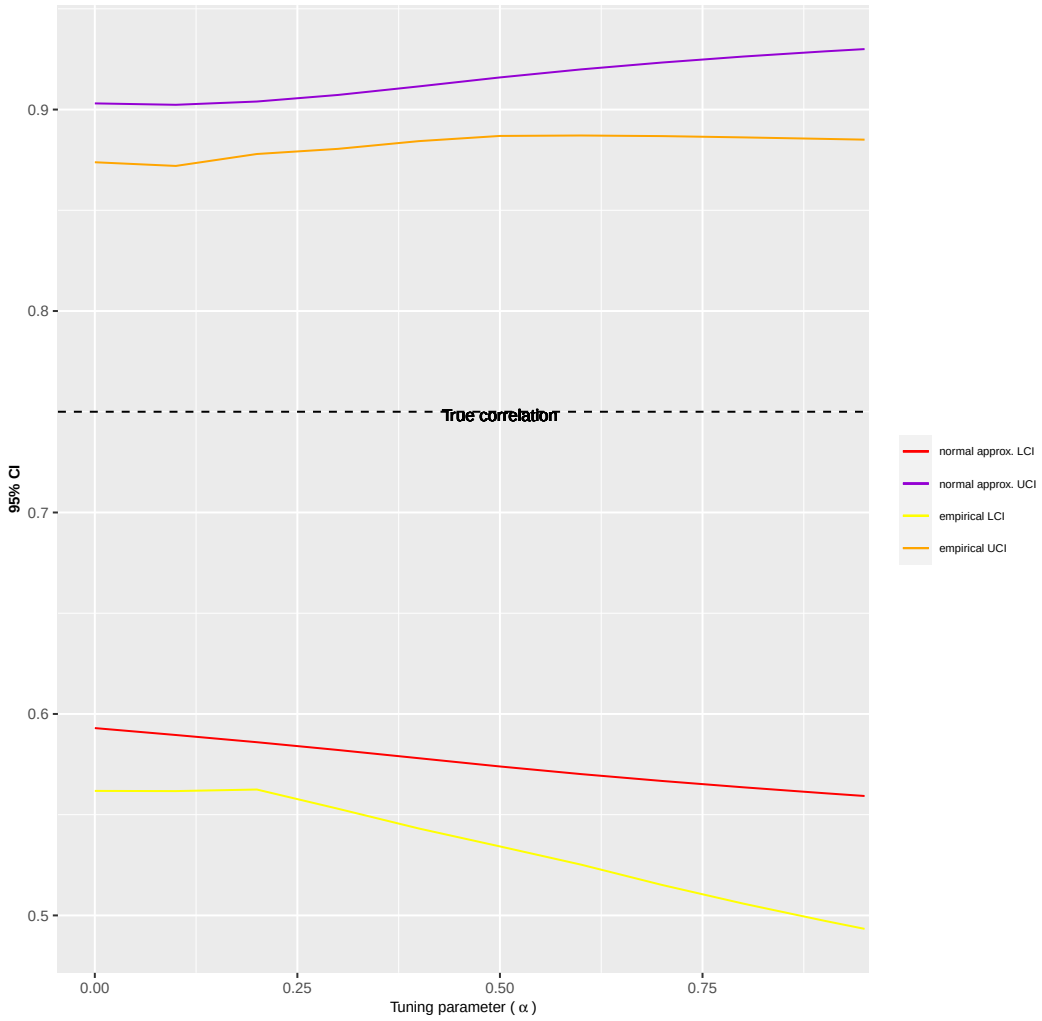

FIGURE 3.6: Plots of the confidence intervals of $\hat{\rho}_\alpha$ under pure data.

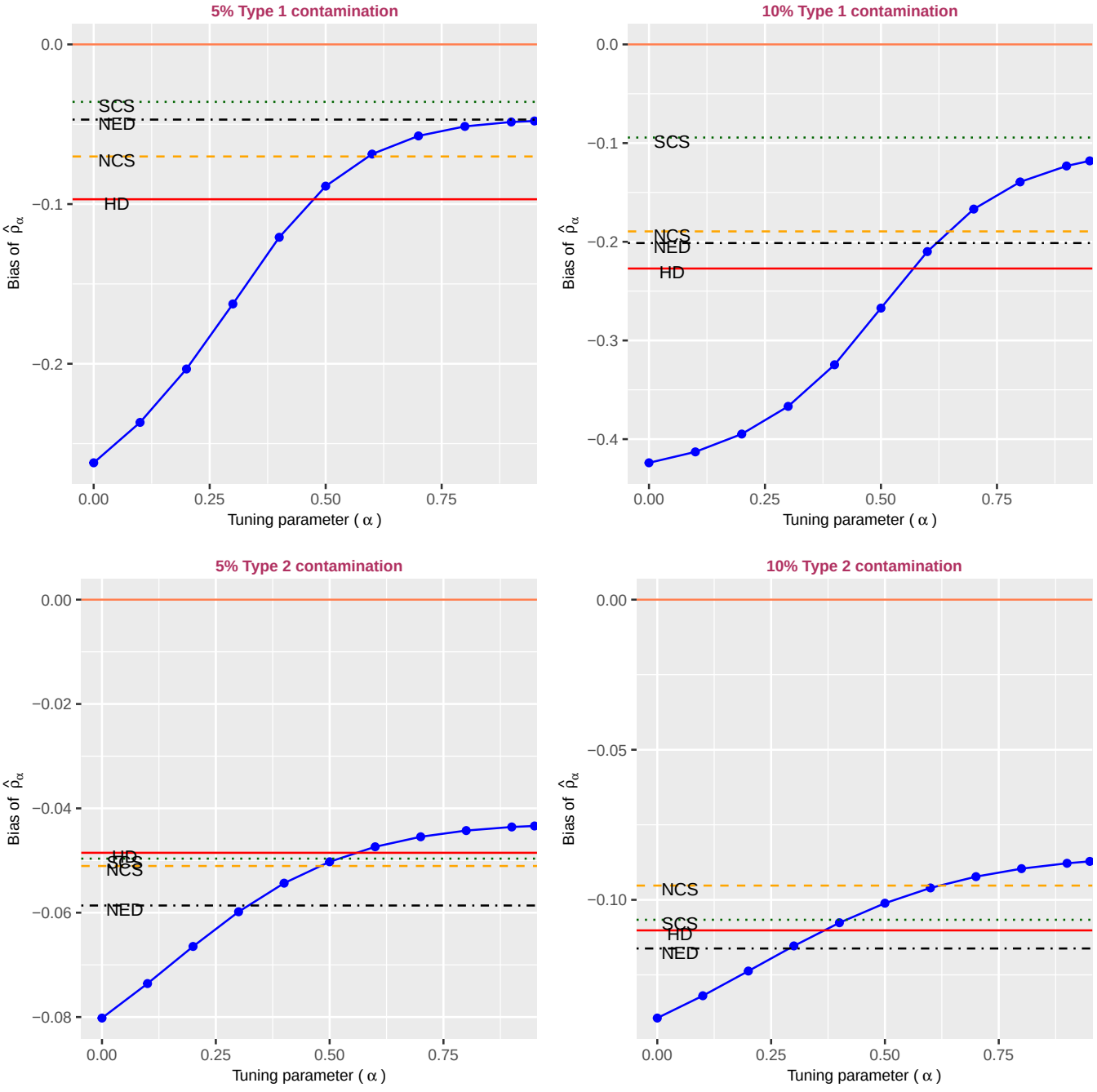

FIGURE 3.7: Bias in one-step estimates of the polychoric correlation under data contamination.

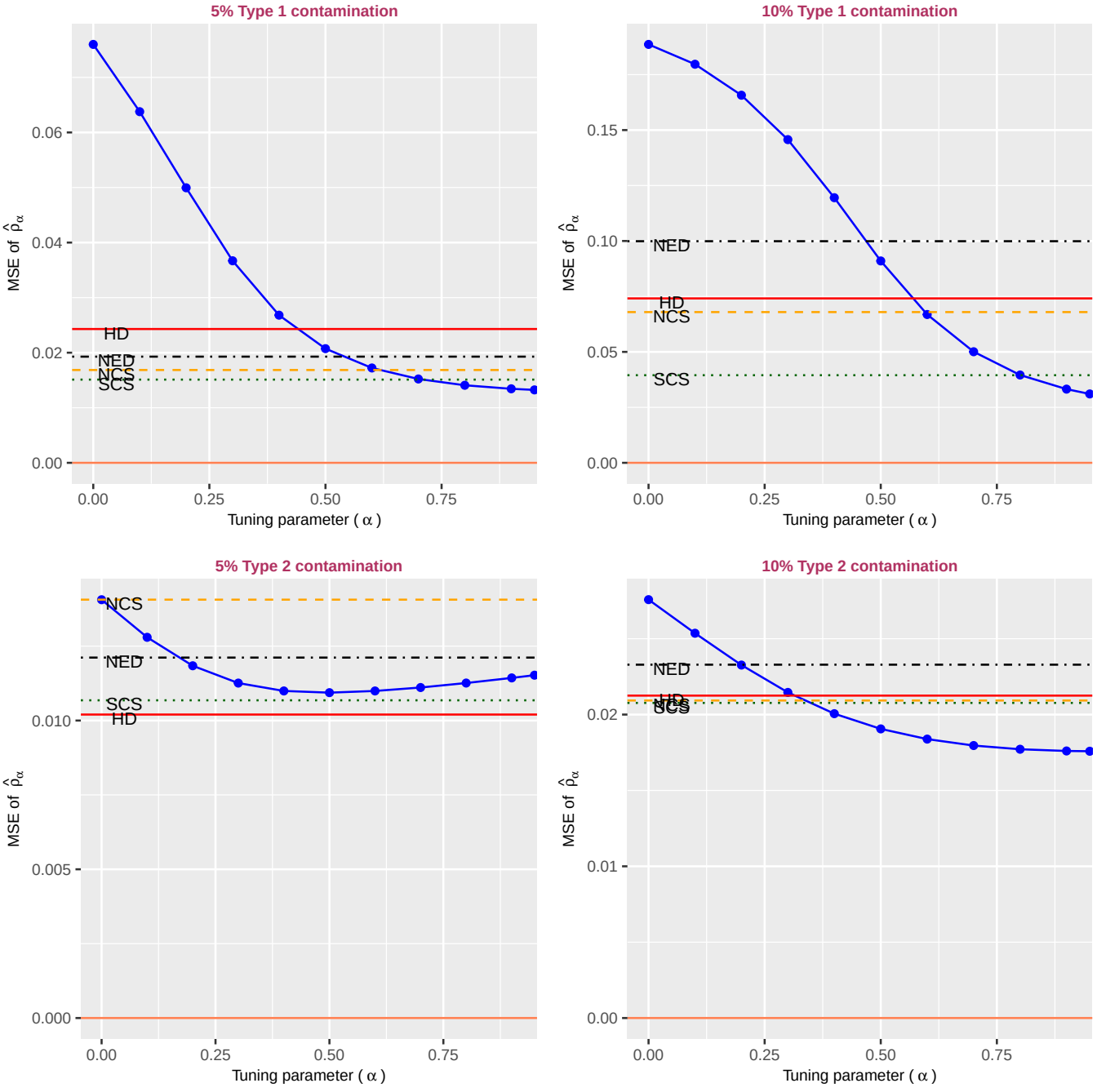

FIGURE 3.8: MSE in one-step estimates of the polychoric correlation under data contamination.

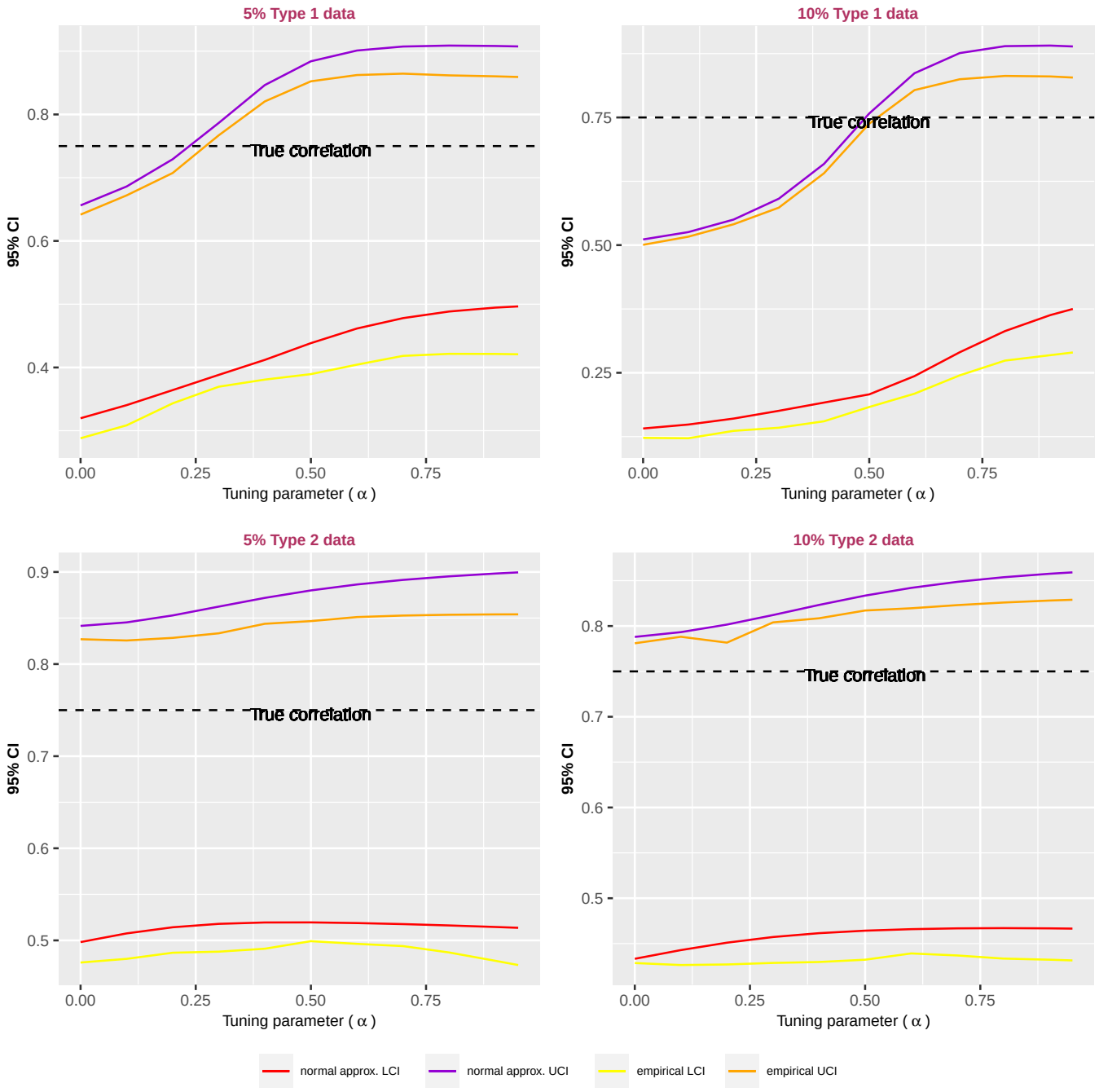

FIGURE 3.9: Confidence intervals (CI) of $\hat{\rho}_\alpha$ under data contamination.

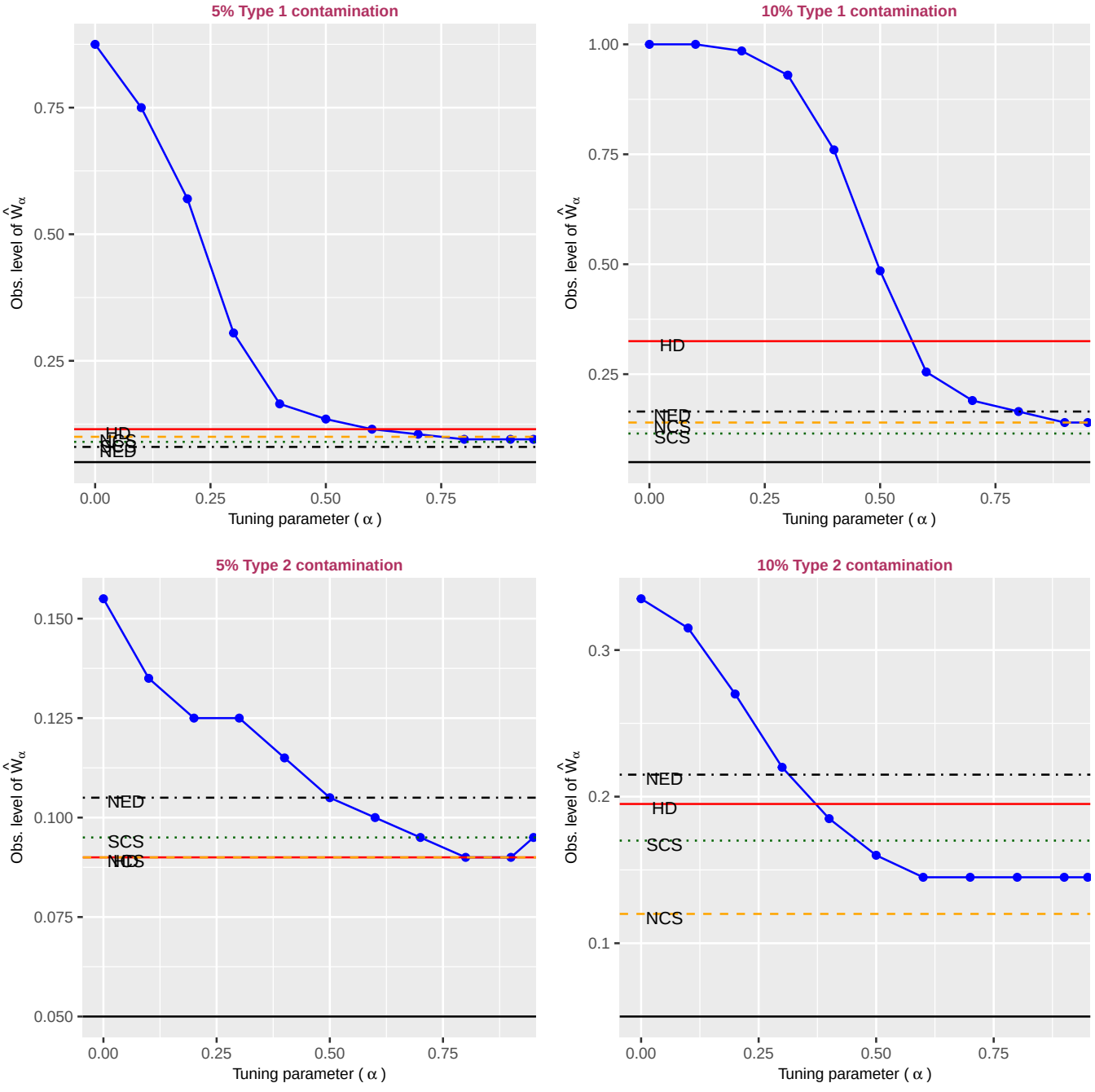

FIGURE 3.10: Observed levels of the Wald-type test statistic under data contamination.

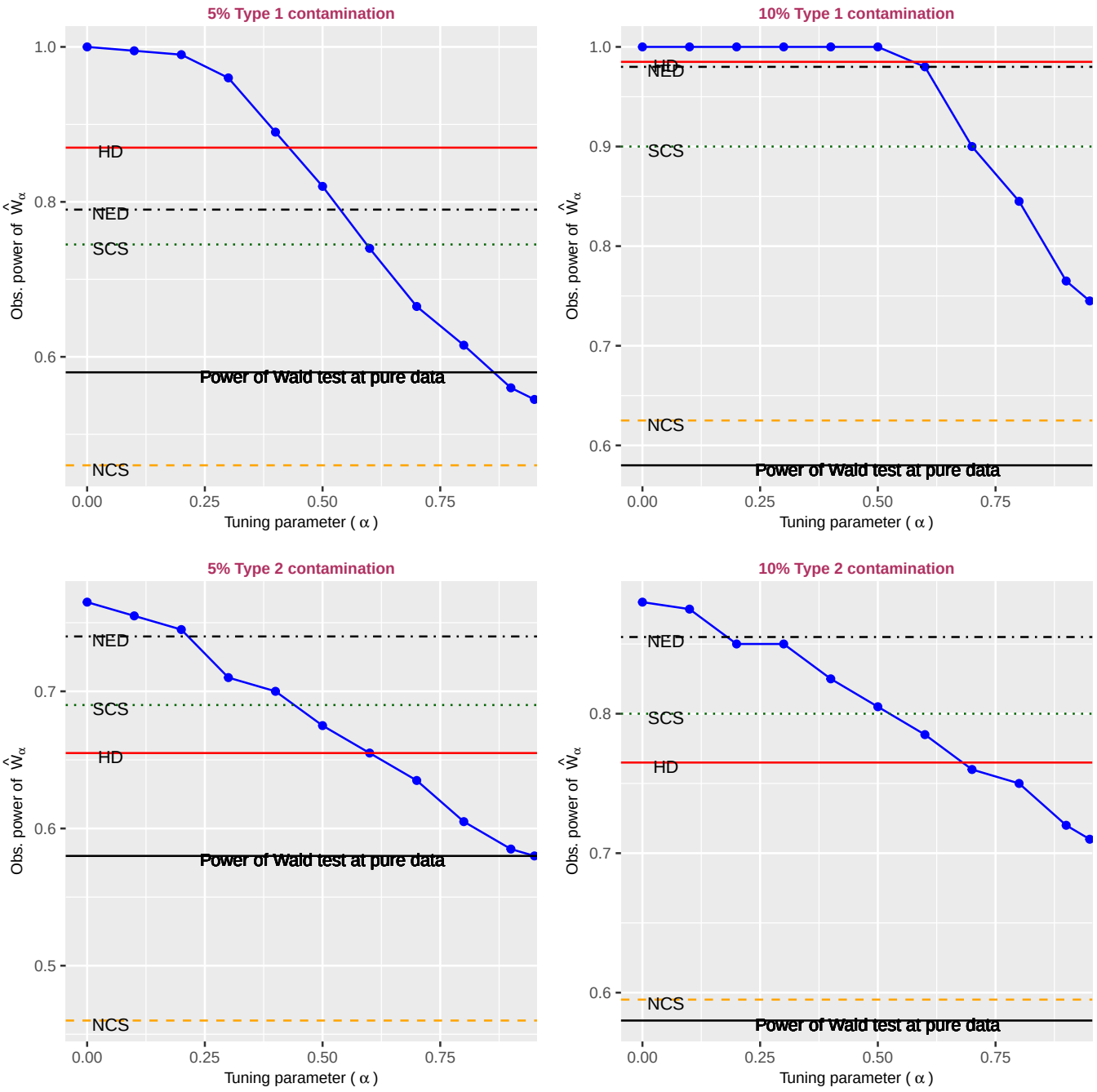

FIGURE 3.11: Observed powers of the Wald-type test statistic.

3.7 Real Data Examples

Here we analyze two real-life data examples collected from the Kaggle repository. For each data set, we will report the optimum tuning parameter, and the corresponding MDPD estimate and compare it with the minimum Hellinger Distance estimator (MHDE). Here we follow the approach of Warwick and Jones (2005) for tuning parameter selection.

Example 3.2. *(Ordinal Likert survey on HS math) This data set contains the results of a survey on HS math education. It includes several ordinal variables. Among them, we are interested in finding the relationship between "class-less-conf" (the change in style of my math class makes me feel less confident in my ability to succeed in math) and "teacher-quality" (my math teacher has done an excellent job in developing an online curriculum for us). Both these variables are measured on a scale of* 1 – 5 *which, respectively, represent the following categories– "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree".*

TABLE 3.3: Optimum α and MSE in Example 3.2 with different pilots ($\hat{\theta}_\alpha$) for one-step estimates

Tuning parameter α for the pilot estimator	Optimal α	Optimal $\widehat{MSE}$
0.0	1	0.2022772
0.1	1	0.1384393
0.2	1	0.08238846
0.3	0.83	0.04845193
0.4	0.83	0.0238762
0.5	0.83	0.008622607
0.6	0.83	0.003316258
0.7	0.83	0.002201
0.8	0.83	0.00165042
0.9	0.83	0.001933935
1.0	0.83	0.003818053

In Table 3.3 we report the optimal tuning parameters corresponding to different pilot values and the optimum MSE for each case. We see that optimum MSE attains a

minimum value 0.00165042 for $\alpha = 0.83$ corresponding to the pilot $\alpha = 0.8$. The pilot estimator $\hat{\theta}_{0.8}$ works best for having the lowest optimized empirical MSE for this data example. Since it requires a very high value of α to minimize the empirical MSE, these data may have a high amount of anomaly in relation to the parametric model. This may be corroborated by the fact that the histograms of these two variables, presented in the first row of Figure 3.12, heavily deviate from the normality assumption of the latent variables. This, in a way, implies that outliers exist in the values of the categorical variables such that they do not follow the model assumption. All parameters' estimates of all the parameters are reported in Table 3.4. This table also includes the estimates corresponding to the optimal tuning parameter $\alpha = 0.83$.

TABLE 3.4: One-step estimates in Example 3.2 for different methods

Method	ρ	η_1	η_2	η_3	η_4	β_1	β_2	β_3	β_4
MLE	0.588	−1.512	−0.760	0.175	1.246	−1.679	−0.880	−0.100	0.802
MDPDE									
$\alpha = 0.1$	0.648	-1.583	-0.773	0.169	1.239	-1.664	-0.886	-0.104	0.839
$\alpha = 0.2$	0.719	−1.678	−0.780	0.166	1.236	−1.656	−0.8976	−0.113	0.872
$\alpha = 0.3$	0.780	-1.763	-0.776	0.166	1.236	-1.673	-0.917	-0.124	0.886
$\alpha = 0.4$	0.820	−1.806	−0.762	0.166	1.236	−1.716	−0.938	−0.133	0.886
$\alpha = 0.5$	0.848	-1.826	-0.747	0.166	1.235	-1.774	-0.955	-0.137	0.881
$\alpha = 0.6$	0.866	-1.836	-0.734	0.164	1.233	-1.820	-0.966	-0.137	0.877
$\alpha = 0.7$	0.876	-1.832	-0.724	0.162	1.232	-1.840	-0.969	-0.135	0.875
$\alpha = 0.8$	0.881	-1.818	-0.715	0.161	1.233	-1.841	-0.968	-0.131	0.875
$\alpha = 0.83$	0.881	−1.812	−0.712	0.161	1.233	−1.838	−0.966	−0.129	0.875
$\alpha = 0.9$	0.883	−1.798	−0.706	0.161	1.234	−1.830	−0.963	−0.125	0.876
$\alpha = 1$	0.884	−1.777	−0.698	0.162	1.235	−1.816	−0.956	−0.119	0.878
MHDE	0.848	−1.987	−0.917	0.205	1.207	−1.887	−1.051	−0.229	0.836

Example 3.3. *(Network analysis of hobbies and interests) This data set contains survey responses from young persons aged between* 15 *and* 30*. This includes variables about their preferences in hobbies and interests. Two such variables– "reliability" and "keeping promises" are measured on an ordinal scale of* $1-5$*. As in the previous example, it is clear from Table 3.5 that* $\hat{\alpha}_1$ *works best as a pilot tuning parameter for this data set. Consequently, the optimum tuning*

parameter obtained from this table is $\alpha = 1$. *Histograms of these two variables are presented in the second row of Figure 3.12, and they slightly deviate from the normality assumption of the latent variables. All the parameters' estimates are presented in Table 3.6.*

TABLE 3.5: Optimum α and MSE in Example 3.3 with different pilots ($\hat{\theta}_\alpha$) for one-step estimates

Tuning parameter α for the pilot estimator	Optimal α	Optimal $\widehat{MSE}$
0.0	0	0.00625855
0.1	0.11	0.006101563
0.2	0.22	0.005790331
0.3	0.32	0.005423814
0.4	0.43	0.005067958
0.5	0.52	0.004766919
0.6	0.61	0.004510525
0.7	0.71	0.004287135
0.8	0.8	0.004089602
0.9	0.9	0.003913351
1.0	1.0	0.003758593

TABLE 3.6: One-step estimates in Example 3.3 for different methods

Method	ρ	η_1	η_2	η_3	η_4	β_1	β_2	β_3	β_4
MLE	0.453	−2.304	−1.396	−0.457	0.594	−2.255	−1.601	−0.635	0.473
MDPDE									
$\alpha = 0.1$	0.469	-2.321	-1.403	-0.461	0.594	-2.279	-1.607	-0.630	0.481
$\alpha = 0.2$	0.484	-2.334	-1.410	-0.466	0.593	-2.303	-1.613	-0.625	0.489
$\alpha = 0.3$	0.498	-2.342	-1.418	-0.471	0.592	-2.325	-1.618	-0.620	0.496
$\alpha = 0.4$	0.511	-2.341	-1.425	-0.476	0.592	-2.344	-1.621	-0.615	0.502
$\alpha = 0.5$	0.525	-2.331	-1.431	-0.481	0.591	-2.356	-1.622	-0.609	0.509
$\alpha = 0.6$	0.538	-2.310	-1.435	-0.485	0.591	-2.357	-1.618	-0.603	0.516
$\alpha = 0.7$	0.551	-2.280	-1.436	-0.487	0.592	-2.348	-1.611	-0.596	0.523
$\alpha = 0.8$	0.563	-2.243	-1.435	-0.489	0.593	-2.327	-1.600	-0.588	0.530
$\alpha = 0.9$	0.576	-2.201	-1.431	-0.489	0.595	-2.299	-1.587	-0.579	0.537
$\alpha = 1$	0.587	-2.164	-1.426	-0.489	0.597	-2.264	-1.571	-0.571	0.544
MHDE	0.460	−2.420	−1.429	−0.473	0.587	−2.359	−1.624	−0.647	0.470

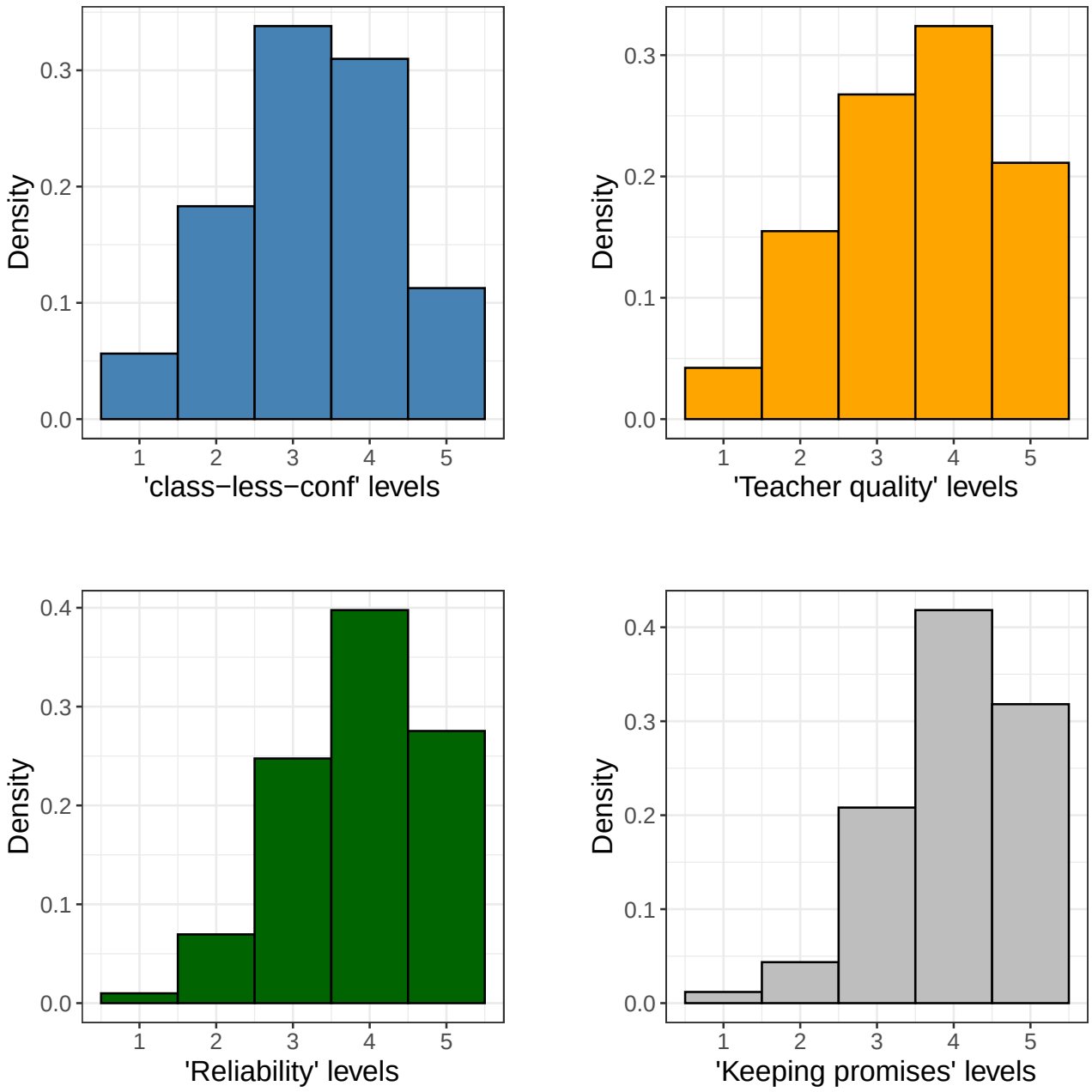

FIGURE 3.12: Histograms of the categorical variables of the real data sets in Example 3.2 and Example 3.3.

3.8 Conclusions

In these discussions, we see that one-step DPD performs almost as good as the ML method in estimating and testing the polychoric correlation when the data truly follow the model. Moreover, when the underlying data set fails to meet the model assumption, the MDPDE works much better, when compared to the MLE and some other robust methods. This fact is quite helpful to applied scientists who might want to use polychoric correlation to measure the association between ordinal variables in a way such that they should be robust and efficient.

Data Availability Statement

The data set that supports the findings of this study is openly available in the Kaggle repository at https://www.kaggle.com/datasets/matt40/ordinal-likert-survey-on-hs-math and https://www.kaggle.com/code/ankur310794/network-analysis-of-hobbies-interests/data.

This page is intentionally left blank.

Chapter 4

Two-Step Inference about the Polychoric Correlation

4.1 Introduction

In Chapter 3 we have discussed the one-step inference regarding the polychoric correlation using the density power divergence. In this chapter, we develop a two-step approach using this divergence measure. As before we assume a model as in (3.1) for a pair of ordinal variables constituting a contingency table. We use the same notations as in the previous chapter unless mentioned otherwise. The historical developments of the inferential methods regarding the polychoric correlation have already been discussed earlier. This present chapter aims to build a theory based on the density power divergence along the lines of the previous chapter but using a two-step approach.

The rest of this chapter is organized as follows.

(a) In Section 4.2, we introduce the two-step approach and explicitly derive the estimating equations.

(b) Asymptotic properties of the two-step estimator of the polychoric correlation are discussed in Section 4.3 which includes the consistency results in Subsection 4.3.1 and the asymptotic normality result in Subsection 4.3.2.

(c) As done previously, we introduce the Wald-type test statistic and discuss their asymptotic properties in Subsection 4.3.3.

(d) In Section 4.4, we study the stability of the estimators and the test statistic through the influence function analysis.

(e) Simulation studies are presented in Section 4.5.

(f) The tuning parameter selection is discussed in Section 4.6.

(g) Applications of this method to some real-life data examples are provided in Section 4.7. Finally, some concluding remarks are made in Section 4.8.

4.2 Estimating Equations

We assume that $\pi_{ij}(\theta), g_{ij} > 0$ for all $i = 1, 2 \ldots, r$ and $j = 1, 2, \ldots, s$. Also, the meaning of other notations carries forward from Chapter 3. See that

$$G_{is} = \mathbb{P}\Big(X \leq i\Big) = \sum_{l=1}^{i} \mathbb{P}\Big(\eta_{l-1} < U \leq \eta_l\Big) = \Phi_1(\eta_i) \tag{4.1}$$

at the model (3.1) when the underlying latent variables are assumed to follow a bivariate normal distribution. This implies that $\eta_i = \Phi_1^{-1}(G_{is})$, and similarly $\beta_j = \Phi_1^{-1}(G_{rj})$ for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. In this approach, we take the form of divergence to be $d_\alpha(g, \pi(\rho, \gamma^*))$ where

$$\gamma^* := \gamma(G) = (\Phi_1^{-1}(G_{1s}), \ldots, \Phi_1^{-1}(G_{(r-1)s}), \Phi_1^{-1}(G_{r1}), \ldots, \Phi_1^{-1}(G_{r(s-1)})). \tag{4.2}$$

In this setup, a best-fitting parameter is given by

$$\rho_\alpha^* = \arg\min_{\rho \in I} d_\alpha(g, \pi(\rho, \gamma^*)) \text{ at fixed } \alpha, \tag{4.3}$$

where $I = (-1, 1)$. Likewise, the minimum density power divergence estimator is defined as

$$\tilde{\rho}_\alpha := \arg\min_{\rho \in I} d_\alpha(p, \pi(\rho, \tilde{\gamma})) \text{ where } \tilde{\gamma} = \gamma(P), \tag{4.4}$$

and "P" being the distribution associated with the empirical density given by "p". We call this a *two-step* approach for an obvious reason. It is because some estimates of cut-off points are plugged into the objective function before the same is used in the minimization process for the unknown parameter ρ. Further, see that

$$H_n(\rho, \tilde{\gamma}) = \frac{1}{n}\sum_{l=1}^{n} V(\rho, \tilde{\gamma}, Z_l) = \mathbb{E}_{\mathbb{P}_n}\Big[V(\rho, \tilde{\gamma}, Z_l)\Big] \text{ at fixed } \alpha. \tag{4.5}$$

Its population version is similarly given by $H(\rho, \gamma^*) = \mathbb{E}_G\Big[V(\rho, \gamma^*, Z_l)\Big]$. The estimator $\tilde{\rho}_\alpha$ satisfies the estimating equation

$$\sum_{i,j}\Big\{\pi_{ij}^{\alpha}(\rho, \tilde{\gamma}) - \pi_{ij}^{\alpha-1}(\rho, \tilde{\gamma})p_{ij}\Big\}\sum_{i_1, j_1=0}^{1}(-1)^{i_1+j_1}\phi_2(\tilde{\eta}_{i-i_1}, \tilde{\beta}_{j-j_1}; \rho) = 0 \tag{4.6}$$

at fixed α. Similarly, it is true that

$$\mathbb{E}_{\mathbb{P}_n}\Big[\frac{\partial}{\partial\rho}V(\tilde{\rho}_\alpha, \tilde{\gamma}, Z_l)\Big] = 0 \text{ and } \mathbb{E}_G\Big[\frac{\partial}{\partial\rho}V(\rho_\alpha^*, \gamma^*, Z_l)\Big] = 0 \text{ at fixed } \alpha. \tag{4.7}$$

See that the two-step estimate of polychoric correlation differs from the usual MDPDE as their respective objective functions differ. However, we shall show later that under appropriate conditions both $\hat{\rho}_\alpha$ and $\tilde{\rho}_\alpha$ converge to the limiting value.

4.3 Asymptotic Properties

Notice that the parameter space $I = (-1, 1)$, where ρ belongs, is not a compact set by itself. Thus, in general, a best-fitting parameter does not exist, but under a mild condition, it does. Here we discuss the asymptotic properties of $\tilde{\rho}_\alpha$. Firstly, we present the consistency results and the asymptotic normality of $\tilde{\rho}_\alpha$. Subsequently, we discuss the asymptotic properties for the class of Wald-type test statistics. In the following results, it is implicitly assumed that $\alpha > 0$, unless stated otherwise.

4.3.1 Consistency

In the first result, we present an elementary proof of consistency when the best-fitting parameter is assumed to be unique. This proof follows the approach of Beran (1977a).

Theorem 4.1. *Suppose the models are conditionally identifiable in the sense that* $\pi_{ij}(\rho_1, \gamma^*) \neq \pi_{ij}(\rho_2, \gamma^*)$ *for all* i, j *and* $\rho_1 \neq \rho_2$*; also* $\inf_{\rho \in I \setminus H} d_\alpha(g, \pi(\rho, \gamma^*)) > d_\alpha(g, \pi(\rho^*, \gamma^*))$ *for some compact set* $H \subset I$ *and* $\rho^* \in H$*. Then a best-fitting parameter* ρ^*_α *exists. Suppose* ρ^*_α *is unique. Moreover assume* $0 < P_{1s} < \cdots < P_{(r-1)s} < 1$ *and* $0 < P_{r1} < \cdots < P_{r(s-1)} < 1$*. Then* $\tilde{\rho}_\alpha \xrightarrow{a.s.} \rho^*_\alpha$ *as* $n \to \infty$*.*

Proof. It is easy to see that the divergence as a map $\rho \mapsto d_\alpha(g, \pi(\rho, \gamma^*))$ is continuous at fixed γ^*, g. Since $\inf_{\rho \in I \setminus H} d_\alpha(g, \pi(\rho, \gamma^*)) > d_\alpha(g, \pi(\rho^*, \gamma^*))$ for some compact set $H \subset I$ and $\rho^* \in H$, there exists a minimizer ρ^*_α of $d_\alpha(g, \pi(\rho, \gamma^*))$ inside $H \subset I$. At fixed g and $\alpha > 0$, we similarly define

$$\hbar(\rho) = d_\alpha(g, \pi(\rho, \gamma^*)) \text{ and } \hbar_n(\rho) = d_\alpha(p, \pi(\rho, \tilde{\gamma})). \tag{4.8}$$

Observe that ρ_α^* and $\tilde{\rho}_\alpha$ respectively minimize $\hbar(\rho)$ and $\hbar_n(\rho)$. First, we shall show that $|\hbar(\tilde{\rho}_\alpha) - \hbar(\rho_\alpha^*)| \xrightarrow{a.s.} 0$ for $n \to \infty$. See that

$$|\hbar(\rho) - \hbar_n(\rho)| \leq \sum_{i,j} \left|\pi_{ij}^{1+\alpha}(\rho, \gamma^*) - \pi_{ij}^{1+\alpha}(\rho, \tilde{\gamma})\right| + \left(1 + \frac{1}{\alpha}\right) \sum_{i,j} \left|\pi_{ij}^{\alpha}(\rho, \gamma^*) g_{ij} - \pi_{ij}^{\alpha}(\rho, \tilde{\gamma}) p_{ij}\right| + \frac{1}{\alpha} \sum_{i,j} \left|g_{ij}^{1+\alpha} - p_{ij}^{1+\alpha}\right|. \tag{4.9}$$

Since $p_{ij} \xrightarrow{a.s} g_{ij}$ for all i, j, it implies that $\sum_{i,j} |p_{ij}^{1+\alpha} - g_{ij}^{1+\alpha}| \xrightarrow{a.s} 0$ as $n \to \infty$. We also know that Φ_1^{-1} is continuous, strictly increasing and bounded in $(0, 1)$. Therefore

$$-\infty < \eta_1 < \eta_2 < \cdots < \eta_{r-1} < \infty \iff 0 < G_{1s} < G_{2s} < \cdots < G_{(r-1)s} < 1. \tag{4.10}$$

It is sufficient to have $0 < P_{1s} < \cdots < P_{(r-1)s} < 1$ to make $P_{is} \xrightarrow{a.s.} G_{is}$ for $1 \leq i \leq r-1$, and similarly, the condition $0 < P_{r1} < \cdots < P_{r(s-1)} < 1$ would imply $P_{rj} \xrightarrow{a.s.} G_{rj}$ for $1 \leq j \leq s-1$. So we have $\tilde{\gamma} \xrightarrow{a.s.} \gamma^*$ for $n \to \infty$. Thus, using the continuity of the map $\theta \mapsto \pi(\theta)$, we arrive at

$$|\hbar(\rho) - \hbar_n(\rho)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty \text{ for each } \rho. \tag{4.11}$$

Also, see that

$$0 \leq \hbar_n(\tilde{\rho}_\alpha) - \hbar(\rho_\alpha^*) \leq \hbar_n(\rho_\alpha^*) - \hbar(\rho_\alpha^*) \text{ when } \hbar_n(\tilde{\rho}_\alpha) \geq \hbar(\rho_\alpha^*), \tag{4.12}$$

$$0 \leq \hbar(\rho_\alpha^*) - \hbar_n(\tilde{\rho}_\alpha) \leq \hbar(\tilde{\rho}_\alpha) - \hbar_n(\tilde{\rho}_\alpha) \text{ when } \hbar_n(\tilde{\rho}_\alpha) \leq \hbar(\rho_\alpha^*). \tag{4.13}$$

This implies that

$$|\hbar_n(\tilde{\rho}_\alpha) - \hbar(\rho_\alpha^*)| \leq |\hbar_n(\rho_\alpha^*) - \hbar(\rho_\alpha^*)| + |\hbar(\tilde{\rho}_\alpha) - \hbar_n(\tilde{\rho}_\alpha)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty. \tag{4.14}$$

Combining (4.11) with (4.14), we get

$$|\hbar(\tilde{\rho}_\alpha) - \hbar(\rho_\alpha^*)| \leq |\hbar(\tilde{\rho}_\alpha) - \hbar_n(\tilde{\rho}_\alpha)| + |\hbar_n(\tilde{\rho}_\alpha) - \hbar(\rho_\alpha^*)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty. \tag{4.15}$$

Notice that the map $\rho \mapsto \hbar(\rho)$ is continuous. When $\tilde{\rho}_\alpha$ is not consistent, (4.15) leads to a contradiction, because the best-fitting parameter is assumed to be unique. Thus, the consistency holds, and we have $\hbar(\tilde{\rho}_\alpha) \xrightarrow{a.s.} \hbar(\rho_\alpha^*)$ if and only if $\tilde{\rho}_\alpha \xrightarrow{a.s.} \rho_\alpha^*$ as $n \to \infty$. □

Remark 4.1. *Notice that the conditions* $0 < P_{1s} < P_{2s} < \cdots < P_{(r-1)s} < 1$ *and* $0 < P_{r1} < P_{r2} < \cdots < P_{r(s-1)} < 1$ *together imply the consistency of* $\tilde{\gamma}$ *for* γ^**. The statistical interpretation of this condition is that none of the rows and columns should be non-empty. In the subsequent results in the two-step scenario, we shall state simply that* $\tilde{\gamma} \xrightarrow{\mathbb{P}/a.s.} \gamma^*$ *as a condition.*

The assumption, that the best-fitting parameter exists uniquely, may be quite restrictive in many practical applications. Often, there exist multiple best-fitting parameters. In the next result, this uniqueness assumption is further relaxed, and under such circumstances, we seek to show consistency. This proof follows the approaches of Wald (1949).

Theorem 4.2. *Suppose* $\inf_{\rho \in I \setminus H} d_\alpha(g, \pi(\rho, \gamma^*)) > d_\alpha(g, \pi(\rho^*, \gamma^*))$ *for some compact set* $H \subset I$ *and* $\rho^* \in H$*. Also assume that for every open set* U *containing* ρ *such that* $diam(U) \downarrow 0+$*, the map* $Z_l \mapsto \inf_{\rho \in U} V(\rho, \tilde{\gamma}, Z_l)$ *is measurable when conditioned at* $\tilde{\gamma}$*, and* $\mathbb{E}_G\left[\inf_{\rho \in U} V(\rho, \gamma^*, Z_l)\right] > -\infty$*. Assume that* $\tilde{\gamma} \xrightarrow{\mathbb{P}} \gamma^*$ *as* $n \to \infty$*. Then for any estimator* $\tilde{\rho}_\alpha$ *satisfying* $d_\alpha(p, \pi(\tilde{\rho}_\alpha, \tilde{\gamma})) \leq d_\alpha(p, \pi(\rho_\alpha^*, \tilde{\gamma})) + o_{\mathbb{P}}(1)$*, the following holds*

$$\mathbb{P}\left\{\tilde{\rho}_\alpha \in H : \min_{\rho_\alpha^*} |\tilde{\rho}_\alpha - \rho_\alpha^*| \geq \epsilon\right\} \longrightarrow 0 \text{ as } n \to \infty \tag{4.16}$$

for every $\epsilon > 0$*.*

Proof. At fixed α and any $\epsilon > 0$, define the following set:

$$A = \left\{ \rho \in H : \min_{\rho^*_\alpha} |\rho - \rho^*_\alpha| \geq \epsilon \right\}. \tag{4.17}$$

Being a closed subset, A is also a compact set. Then an open cover $\{U_t \text{ open } : t \in A, diam(U_t) < 1/n\}$ of A contains a finite sub-cover $\cup_{i=1}^{k} U_{t_i}$ for some finite integer k. We know that

$$\begin{aligned} \inf_{\rho \in U_{t_i}} \mathbb{E}_{\mathbb{P}_n}\Big[V(\rho, \tilde{\gamma}, Z_l) \Big] &= \inf_{\rho \in U_{t_i}} \left[\frac{1}{n} \sum_{l=1}^{n} V(\rho, \tilde{\gamma}, Z_l) \right] \\ &\geq \frac{1}{n} \sum_{l=1}^{n} \inf_{\rho \in U_{t_i}} V(\rho, \tilde{\gamma}, Z_l) = \mathbb{E}_{\mathbb{P}_n} \left[\inf_{\rho \in U_{t_i}} V(\rho, \tilde{\gamma}, Z_l) \right] \end{aligned} \tag{4.18}$$

for $i = 1, \ldots, k$. Thus we have

$$\begin{aligned} \inf_{\rho \in A} \mathbb{E}_{\mathbb{P}_n}\Big[V(\rho, \tilde{\gamma}, Z_l) \Big] &\geq \min \left\{ \inf_{\rho \in U_{t_i}} \mathbb{E}_{\mathbb{P}_n}\Big[V(\rho, \tilde{\gamma}, Z_l) \Big]; i = 1, 2, \ldots, k \right\} \\ &\geq \min_{i=1,\ldots,k} \mathbb{E}_{\mathbb{P}_n}\Big[\inf_{\rho \in U_{t_i}} V(\rho, \tilde{\gamma}, Z_l) \Big] \\ &= \min_{i=1,\ldots,k} \mathbb{E}_{G}\Big[V(\rho_i, \gamma^*, Z_l) \Big] + o_{\mathbb{P}}(1) \\ &> \mathbb{E}_{G}\Big[V(\rho^*_\alpha, \gamma^*, Z_l) \Big] + o_{\mathbb{P}}(1) \text{ as } \rho^*_\alpha \neq A, \text{ and } \rho_1, \ldots, \rho_k \in A. \end{aligned} \tag{4.19}$$

When $\tilde{\rho}_\alpha \in A$, it is true that

$$\begin{aligned} \inf_{\rho \in A} d_\alpha(p, \pi(\rho, \tilde{\gamma})) &\leq d_\alpha(p, \pi(\tilde{\rho}_\alpha, \tilde{\gamma})) \\ &\leq d_\alpha(p, \pi(\rho^*_\alpha, \tilde{\gamma})) + o_{\mathbb{P}}(1) \text{ (by assumption)} \\ &= d_\alpha(g, \pi(\rho^*_\alpha, \gamma^*)) + o_{\mathbb{P}}(1) \text{ (by consistency of } \tilde{\gamma}). \end{aligned} \tag{4.20}$$

Thus

$$\mathbb{P}\Big\{\tilde{\rho}_\alpha \in A\Big\} \leq \mathbb{P}\Bigg\{\inf_{\rho\in A}\mathbb{E}_{\mathbb{P}_n}\Big[V(\rho,\tilde{\gamma},Z_l)\Big] \leq \mathbb{E}_G\Big[V(\rho^*_\alpha,\gamma^*,Z_l)\Big] + o_{\mathbb{P}}(1)\Bigg\} \longrightarrow 0 \text{ as } n\to\infty. \tag{4.21}$$

This completes the proof. □

Next, we shall show both these one-step and two-step estimates of the polychoric correlation converge to the same value when the model that truly generates data.

Theorem 4.3. *Suppose the true distribution belongs to the model family, and both $\hat{\rho}_\alpha$ and $\tilde{\rho}_\alpha$ are consistent. Then*

$$|\tilde{\rho}_\alpha - \hat{\rho}_\alpha| \xrightarrow{\mathbb{P}} 0 \text{ as } n\to\infty. \tag{4.22}$$

Proof. Suppose the true distribution generating the contingency table belongs to the model family. Then the one-step best-fitting parameter θ_α turns out to be the true value $\theta^0 = (\rho^0,\gamma^0)$ due to Fisher consistency. This also gives $\gamma^* = \gamma_0$. Consequently, the two-step best-fitting parameter ρ^*_α becomes ρ^0. Therefore we get that

$$|\hat{\rho}_\alpha - \tilde{\rho}_\alpha| \leq |\hat{\rho}_\alpha - \rho^0| + |\tilde{\rho}_\alpha - \rho^0| \longrightarrow 0 \text{ as } n\to\infty. \tag{4.23}$$

This completes the proof. □

Theorem 4.3 shows that both these approaches lead the estimators of the polychoric correlation to converge to a common limit for a data set truly generated by the model. This phenomenon will be further demonstrated in the simulation studies.

4.3.2 Asymptotic Normality

It is clear from the definition that the asymptotic distribution of the two-step estimator $\tilde{\rho}_\alpha$ invariably depends on $\tilde{\gamma}$. Though we derive the asymptotic distribution of $\tilde{\gamma}$, we find it intractable to derive the asymptotic covariance between $\tilde{\rho}_\alpha$ and $\tilde{\gamma}$. This leads to a bias term which appears in the normalizing of $\tilde{\rho}_\alpha$. To do that, let us define

$$\nu^2_{\rho,\gamma} = \frac{(1+\alpha)^2\Big\{\sum_{i,j}\pi^{2\alpha}_{ij}(\rho,\gamma)\Big[\frac{\partial \ln \pi_{ij}(\rho,\gamma)}{\partial\rho}\Big]^2 g_{ij} - \Big(\sum_{i,j}\pi^{\alpha}_{ij}(\rho,\gamma)\Big[\frac{\partial \ln \pi_{ij}(\rho,\gamma)}{\partial\rho}\Big]g_{ij}\Big)^2\Big\}}{\Big(\mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho^2}V(\rho,\gamma,Z_l)\Big]\Big)^2}, \tag{4.24}$$

$$\mathcal{B}_n = \frac{\partial^2}{\partial\rho^2}H_n(\rho^*_\alpha,\gamma^*) \text{ and } \mathcal{C} = \frac{\partial^2}{\partial\gamma\partial\rho}H(\rho^*_\alpha,\gamma^*). \tag{4.25}$$

The asymptotic normality result is the following.

Theorem 4.4. *Suppose* $\frac{\partial^2}{\partial\rho^2}H(\rho,\gamma) \neq 0$ *around* (ρ^*_α,γ^*)*, and* $\tilde{\gamma} \xrightarrow{\mathbb{P}} \gamma^*$*. Then a consistent sequence of roots* $\tilde{\rho}_\alpha$ *exists such that*

$$\sqrt{n}\Big(\tilde{\rho}_\alpha - \rho^*_\alpha - \mathcal{B}_n^{-1}(\tilde{\gamma}-\gamma^*)^T\mathcal{C}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \nu^2_{\rho^*_\alpha,\gamma^*}\Big) \text{ as } n\to\infty. \tag{4.26}$$

Proof. Expanding $\frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma})$ around (ρ^*_α,γ^*) up to second-order and rearranging them yields

$$\left\{\frac{\partial^2}{\partial\rho^2}H_n(\rho',\gamma')\right\}(\tilde{\rho}_\alpha-\rho^*_\alpha) = \frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma}) - \frac{\partial}{\partial\rho}H_n(\rho^*_\alpha,\gamma^*) - (\tilde{\gamma}-\gamma^*)^T\frac{\partial^2}{\partial\gamma\partial\rho}H_n(\rho^{'}_\alpha,\gamma^{'}) \tag{4.27}$$

where $\rho^{'}$ lies between $\tilde{\rho}_\alpha$ and ρ^*_α; similarly the components of γ' are defined. As (ρ',γ') lies in a neighbourhood of (ρ^*_α,γ^*), by assumption, it is true that $\left|\frac{\partial^2}{\partial\rho^2}H(\rho',\gamma')\right| > 0$. First we prove the consistency of $\tilde{\rho}_\alpha$. Choose $0 < \kappa < \left|\frac{\partial^2}{\partial\rho^2}H(\rho',\gamma')\right|$. From the estimating

equation it follows that $\frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma})=0$. Also we know that $\tilde{\gamma}\xrightarrow{\mathbb{P}}\gamma^*$. Further, see that

$$\frac{\partial}{\partial\rho}H_n(\rho^*_\alpha,\gamma^*)=\frac{1}{n}\sum_{l=1}^{n}\frac{\partial}{\partial\rho}V(\rho^*_\alpha,\gamma^*,Z_l)\xrightarrow{\mathbb{P}}\mathbb{E}_G\Big[\frac{\partial}{\partial\rho}V(\rho^*_\alpha,\gamma^*,Z_l)\Big]=0. \tag{4.28}$$

So we get $\left|\frac{\partial}{\partial\rho}H_n(\rho^*_\alpha,\gamma^*)\right|<\kappa$ and $\|\tilde{\gamma}-\gamma^*\|<\kappa$ with probability tending to 1. Also recall that all the second-order derivatives of $H(\rho,\gamma)$ with respect to each component of $\theta=(\rho,\gamma)$ are bounded as long as $|\rho|<1$. So there exists finite constants M_1,M_2 such that

$$\begin{aligned}\left\|\frac{\partial^2}{\partial\gamma\partial\rho}H_n(\rho^{'},\gamma^{'})\right\|&\le\left\|\frac{\partial^2}{\partial\gamma\partial\rho}H_n(\rho^{'},\gamma^{'})-\frac{\partial^2}{\partial\gamma\partial\rho}H(\rho^{'},\gamma^{'})\right\|+\left\|\frac{\partial^2}{\partial\gamma\partial\rho}H(\rho^{'},\gamma^{'})\right\|\le\kappa+M_1,\\ 0<M_2<\left|\frac{\partial^2}{\partial\rho^2}H(\rho',\gamma')\right|-\kappa&\le\left|\frac{\partial^2}{\partial\rho^2}H_n(\rho',\gamma')\right|\le\left|\frac{\partial^2}{\partial\rho^2}H(\rho',\gamma')\right|+\kappa\end{aligned} \tag{4.29}$$

with probability tending to 1. From (4.27) then it follows that

$$M_2|\tilde{\rho}_\alpha-\rho^*_\alpha|\le\left|\frac{\partial^2}{\partial\rho^2}H_n(\rho',\gamma')(\tilde{\rho}_\alpha-\rho^*_\alpha)\right|\le\kappa+\kappa(M_1+\kappa) \tag{4.30}$$

implying that $|\tilde{\rho}_\alpha-\rho^*_\alpha|\le\kappa\frac{1+M_1+\kappa}{M_2}$ with probability tending to 1. Finally, making $\kappa\downarrow 0$ proves the consistency of $\tilde{\rho}_\alpha$.

To prove the asymptotic normality, expand $\frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma})$ around (ρ^*_α,γ^*) up to second-order as

$$\begin{aligned}\frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma})=\mathcal{A}_n+(\tilde{\rho}_\alpha-\rho^*_\alpha)\left\{\mathcal{B}_n+\frac{1}{2}(\tilde{\rho}_\alpha-\rho^*_\alpha)\mathcal{D}_n+\frac{1}{2}(\tilde{\gamma}-\gamma^*)^T\mathcal{F}_n\right\}\\ +(\tilde{\gamma}-\gamma^*)^T\left\{\mathcal{C}_n+\frac{1}{2}\mathcal{E}_n(\tilde{\gamma}-\gamma^*)\right\},\end{aligned} \tag{4.31}$$

where

$$\mathcal{A}_n = \frac{\partial}{\partial\rho}H_n(\rho_\alpha^*,\gamma^*), \mathcal{B}_n = \frac{\partial^2}{\partial\rho^2}H_n(\rho_\alpha^*,\gamma^*), \mathcal{C}_n = \frac{\partial^2}{\partial\gamma\partial\rho}H_n(\rho_\alpha^*,\gamma^*), \tag{4.32}$$

$$\mathcal{D}_n = \frac{\partial^3}{\partial\rho^3}H_n(\rho_\alpha^{'},\gamma^{'}),\ \mathcal{E}_n = \frac{\partial^3}{\partial\gamma\partial\gamma^T\partial\rho}H_n(\rho^{'},\gamma^{'}), \mathcal{F}_n = \frac{\partial^3}{\partial\gamma\partial\rho^2}H_n(\rho^{'},\gamma^{'}), \tag{4.33}$$

and ρ', γ' are defined as earlier. Inserting $\frac{\partial}{\partial\rho}H_n(\tilde{\rho}_\alpha,\tilde{\gamma}) = 0$ and rearranging the terms give

$$\begin{aligned}\sqrt{n}\Big(\tilde{\rho}_\alpha - \rho_\alpha^* - \mathcal{B}_n^{-1}(\tilde{\gamma}-\gamma^*)^T\mathcal{C}\Big) = &-\frac{\sqrt{n}\mathcal{A}_n + \sqrt{n}(\tilde{\gamma}-\gamma^*)^T\left\{(\mathcal{C}_n-\mathcal{C}) + \frac{1}{2}\mathcal{E}_n(\tilde{\gamma}-\gamma^*)\right\}}{\mathcal{B}_n + \frac{1}{2}(\tilde{\rho}_\alpha-\rho_\alpha^*)\mathcal{D}_n + \frac{1}{2}(\tilde{\gamma}-\gamma^*)^T\mathcal{F}_n}\\ &-\sqrt{n}(\tilde{\gamma}-\gamma^*)^T\left\{\frac{\mathcal{C}}{\mathcal{B}_n} - \frac{\mathcal{C}+\frac{1}{2}\mathcal{E}_n(\tilde{\gamma}-\gamma^*)}{\mathcal{B}_n + \frac{1}{2}(\tilde{\rho}_\alpha-\rho_\alpha^*)\mathcal{D}_n + \frac{1}{2}(\tilde{\gamma}-\gamma^*)^T\mathcal{F}_n}\right\},\end{aligned} \tag{4.34}$$

where $\mathcal{C} = \frac{\partial^2}{\partial\gamma\partial\rho}H(\rho_\alpha^*,\gamma^*)$. See that

$$\mathcal{B}_n \xrightarrow{\mathbb{P}} \mathcal{B} = \frac{\partial^2}{\partial\rho^2}H(\rho_\alpha^*,\gamma^*) \neq 0,\ \mathcal{C}_n \xrightarrow{\mathbb{P}} \mathcal{C} \text{ and } \sqrt{n}(\tilde{\gamma}-\gamma^*) = \mathcal{O}_{\mathbb{P}}(1). \tag{4.35}$$

Since all the third-order derivatives are bounded, the consistency of $\tilde{\rho}_\alpha$ and $\tilde{\gamma}$ implies that

$$\sqrt{n}\Big(\tilde{\rho}_\alpha - \rho_\alpha^* - \mathcal{B}_n^{-1}(\tilde{\gamma}-\gamma^*)^T\mathcal{C}\Big) \overset{\mathcal{L}}{=} -\frac{\sqrt{n}\mathcal{A}_n}{\mathcal{B}} + o_{\mathbb{P}}(1). \tag{4.36}$$

To find the distribution of $\sqrt{n}\mathcal{A}_n$, we write

$$\sqrt{n}\mathcal{A}_n = \frac{1}{\sqrt{n}}\sum_{l=1}^{n}\frac{\partial}{\partial\rho}V(\rho_\alpha^*,\gamma^*,Z_l) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\mathcal{B}^2\nu^2_{\rho_\alpha^*,\gamma^*}). \tag{4.37}$$

Therefore $\sqrt{n}\Big(\tilde{\rho}_\alpha - \rho_\alpha^* - \mathcal{B}_n^{-1}(\tilde{\gamma}-\gamma^*)^T\mathcal{C}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\nu^2_{\rho_\alpha^*,\gamma^*})$. This completes the proof. □

Remark 4.2. *Note that the bias disappears when $\mathcal{C} = 0$. Also see that the bias $\mathcal{B}_n^{-1}(\tilde{\gamma} - \gamma^*)^T\mathcal{C}$ (excluding the factor $\sqrt{n}$) converges to 0 at a rate $n^{-1/2}$. To avoid further complications, we subsequently assume that $\mathcal{C} = 0$.*

For the sake of completeness, we present the asymptotic distribution of $\tilde{\gamma}$. Let us define

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}, \tag{4.38}$$

where the block matrices are given by

$$\Sigma_{11} = \begin{bmatrix} G_{1s}(1-G_{1s}) & G_{1s}(1-G_{2s}) & G_{1s}(1-G_{3s}) & \cdots G_{1s}(1-G_{(r-1)s}) \\ \cdot & G_{2s}(1-G_{2s}) & G_{2s}(1-G_{3s}) & \cdots G_{2s}(1-G_{(r-1)s}) \\ \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot G_{(r-1)s}(1-G_{(r-1)s}) \end{bmatrix}_{(r-1)\times(r-1)} \tag{4.39}$$

$$\Sigma_{22} = \begin{bmatrix} G_{r1}(1-G_{r1}) & G_{r1}(1-G_{r2}) & G_{r1}(1-G_{r3}) & \cdots G_{r1}(1-G_{r(s-1)}) \\ \cdot & G_{r2}(1-G_{r2}) & G_{r2}(1-G_{r2}) & \cdots G_{r2}(1-G_{r(s-1)}) \\ \vdots & \vdots & \vdots & \vdots \\ \cdot & \cdot & \cdot & \cdot G_{r(s-1)}(1-G_{r(s-1)}) \end{bmatrix}_{(s-1)\times(s-1)} \tag{4.40}$$

and

$$\Sigma_{12} = ((\sigma_{ij})) \text{ where } \sigma_{ij} = \sum_{l_1=1}^{i}\sum_{k_1=1}^{j} g_{l_1k_1}(1-g_{l_1k_1}) - \sum_{l_1=1}^{i}\sum_{l_2\neq l_1}^{r}\sum_{k_1=1}^{s}\sum_{k_2\neq k_1}^{j} g_{l_1k_1}g_{l_2k_2}. \tag{4.41}$$

Also, define the diagonal matrix

$$D = Diag\left(\frac{1}{\phi_1(\eta_1)},\cdots,\frac{1}{\phi_1(\eta_{r-1})},\frac{1}{\phi_1(\beta_1)},\cdots,\frac{1}{\phi_1(\beta_{s-1})}\right). \tag{4.42}$$

The asymptotic distribution of $\tilde{\gamma}$ is presented here.

Theorem 4.5. *Suppose* $0 < P_{1s} < \cdots < P_{(r-1)s} < 1$ *and* $0 < P_{r1} < \cdots < P_{r(s-1)} < 1$. *Then*

$$\sqrt{n}(\tilde{\gamma} - \gamma^*) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, D\Sigma D\Big) \text{ as } n \to \infty. \tag{4.43}$$

Proof. Let us define

$$a_i = \sqrt{n}\Big(\Phi_1^{-1}(P_{is}) - \Phi_1^{-1}(G_{is})\Big) \text{ and } b_i = \sqrt{n}\Big(\Phi_1^{-1}(P_{rj}) - \Phi_1^{-1}(G_{rj})\Big) \tag{4.44}$$

for all $i = 1,\ldots,r-1$ and $j = 1,\ldots,s-1$. Notice that $\frac{d}{dt}\Phi_1^{-1}(t) = \frac{1}{\phi_1(\Phi_1^{-1}(t))} > 0$, and it is bounded for $0 < t < 1$. Applying the mean value theorem, we obtain

$$a_i = \frac{\sqrt{n}(P_{is} - G_{is})}{\phi_1(\Phi_1^{-1}(t_{i,n}))} \text{ where } t_{i,n} = (1-\epsilon_i^*)P_{is} + \epsilon_i^* G_{is} \text{ with } 0 < \epsilon_i^* < 1. \tag{4.45}$$

Since $\phi_1(\Phi_1^{-1}(\cdot))$ is a continuous function, it follows that

$$\phi_1(\Phi_1^{-1}(t_{i,n})) \xrightarrow{\mathbb{P}} \phi_1(\Phi_1^{-1}(G_{is})) = \phi_1(\eta_i) \text{ as } n \to \infty. \tag{4.46}$$

Note that $\phi_1(\eta_i)$ is positive as long as $-\infty < \eta_i < \infty$ for all $i = 1,2,\ldots,r-1$. We also know that $nP_{is} \sim Bin(n, G_{is})$. Write

$$P_{is} = \frac{1}{n}\sum_{l=1}^{n} \mathbb{1}_{is}(l) \text{ where } \mathbb{1}_{is}(l) = \sum_{l'=1}^{i}\sum_{k=1}^{s}\delta_{l'k}(Z_l), \tag{4.47}$$

which yields

$$\sqrt{n}(P_{is} - G_{is}) = \frac{1}{\sqrt{n}} \sum_{l=1}^{n} A_{is}(l) \text{ where } A_{is}(l) = \left(\mathbb{1}_{is}(l) - G_{is} \right) \tag{4.48}$$

and $i = 1, 2, \ldots, r-1$. See that $\mathbb{E}_G(A_{is}(l)) = 0$ and

$$Var_G(A_{is}(l)) = Var_G(\sqrt{n}P_{is}) = G_{is}(1 - G_{is}). \tag{4.49}$$

See that $\delta_{ij}(Z_l)$ and $\delta_{ij}(Z_{l'})$ are independent, so are $A_{ij}(l)$s at fixed i, j. Hence

$$a_i = \frac{\sqrt{n}(P_{is} - G_{is})}{\phi_1(\Phi_1^{-1}(t_{i,n}))} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{G_{is}(1 - G_{is})}{\phi_1^2(\eta_i)}\right) \text{ as } n \to \infty \tag{4.50}$$

for all $i = 1, \ldots, r-1$. Similarly

$$b_j \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{G_{rj}(1 - G_{rj})}{\phi_1^2(\beta_j)}\right) \text{ as } n \to \infty \text{ for } j = 1, \ldots, s-1. \tag{4.51}$$

Now define $U = n(P_{1s}, P_{2s} - P_{1s}, P_{3s} - P_{2s}, \ldots, 1 - P_{(r-1)s})$. From the construction, it is clear that U is a multinomial distribution as $U \sim MN(n; G_{1s}, G_{2s} - G_{1s}, G_{3s} - G_{2s}, \ldots, 1 - G_{(r-1)s})$. Using the standard results we find that

$$\begin{aligned}
Cov_G(P_{1s}, P_{(l+1)s} - P_{ls}) &= -\frac{1}{n} G_{1s}(G_{(l+1)s} - G_{ls}) \text{ for } l = 1, \ldots, r-1, \\
Cov_G\Big(P_{2s}, P_{is} - P_{2s}\Big) &= Cov_G\Big(P_{1s} + (P_{2s} - P_{1s}), \sum_{l=2}^{i-1}(P_{(l+1)s} - P_{ls})\Big) \\
&= \sum_{l=2}^{i-1} Cov_G(P_{1s}, P_{(l+1)s} - P_{ls}) + \sum_{l=2}^{i-1} Cov_G\Big(P_{2s} - P_{1s}, P_{(l+1)s} - P_{ls}\Big) \\
&= -\frac{1}{n} \sum_{l=2}^{i-1} \Bigg[G_{1s}\Big(G_{(l+1)s} - G_{ls}\Big) + (G_{2s} - G_{1s})\Big(G_{(l+1)s} - G_{ls}\Big) \Bigg] \\
&= -\frac{1}{n} G_{2s}(G_{is} - G_{2s}) \text{ for } i > 2.
\end{aligned} \tag{4.52}$$

Thus

$$\begin{aligned}
Cov_G\Big(P_{1s}, P_{is}\Big) &= Cov_G\Big(P_{1s}, P_{1s} + \sum_{l=1}^{i-1}(P_{(l+1)s} - P_{ls})\Big) \\
&= Var_G(P_{1s}) + \sum_{l=1}^{i-1} Cov_G\Big(P_{1s}, P_{(l+1)s} - P_{ls}\Big) \\
&= \frac{1}{n}\Big\{G_{1s}(1 - G_{1s}) - \sum_{l=1}^{i-1} G_{1s}(G_{(l+1)s} - G_{ls})\Big\} \\
&= \frac{1}{n} G_{1s}\Big(1 - G_{is}\Big) \text{ for } i = 1, 2, \ldots, r-1, \\
Cov_G\Big(P_{2s}, P_{is}\Big) &= \frac{1}{n} G_{2s}(1 - G_{is}) \text{ for } i = 2, \ldots, r-1.
\end{aligned} \tag{4.53}$$

Similarly, we can obtain the covariances among P_{rj}s. Next, we find that

$$\begin{aligned}
Cov_G\Big(P_{is}, P_{rj}\Big) &= Cov_G\Big(\sum_{l_1=1}^{i}\sum_{k_1=1}^{s} p_{l_1k_1}, \sum_{l_2=1}^{r}\sum_{k_2=1}^{j} p_{l_2k_2}\Big) \\
&= \sum_{l_1=1}^{i}\sum_{k_1=1}^{j} Var_G(p_{l_1k_1}) + \sum_{l_1=1}^{i}\sum_{l_2\neq l_1}^{r}\sum_{k_1=1}^{s}\sum_{k_2\neq k_1}^{j} Cov_G\Big(p_{l_1k_1}, p_{l_2k_2}\Big) \\
&= \frac{1}{n}\Big\{\sum_{l_1=1}^{i}\sum_{k_1=1}^{j} g_{l_1k_1}(1 - g_{l_1k_1}) - \sum_{l_1=1}^{i}\sum_{l_2\neq l_1}^{r}\sum_{k_1=1}^{s}\sum_{k_2\neq k_1}^{j} g_{l_1k_1}g_{l_2k_2}\Big\}.
\end{aligned} \tag{4.54}$$

Now define

$$V = \sqrt{n}\Bigg[(P_{1s} - G_{1s}), \ldots, (P_{(r-1)s} - G_{(r-1)s}), (P_{r1} - G_{r1}), \ldots, (P_{r(s-1)} - G_{r(s-1)})\Bigg]. \tag{4.55}$$

See that $\mathbb{E}_G(V) = 0$ and $Var_G(V) = \Sigma$. We need to show that $V \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$. Let $L = (L_1, \ldots, L_{r+s-2}) \in \mathbb{R}^{r+s-2}$ be any vector. See that

$$L^T V = \frac{1}{\sqrt{n}}\sum_{l=1}^{n} B_l \text{ where } B_l = \sum_{t=1}^{r-1} L_t A_{ts}(l) + \sum_{t=r}^{r+s-2} L_t A_{r(t-r+1)}(l), \tag{4.56}$$

where B_ls are iid as $A_{(\cdots)}(l)$s are iid. Hence the linear combination $L^T V$ converges to a normal distribution with the mean and variance as

$$\mathbb{E}_G(L^T V) = 0, \text{ and } Var_G(L^T V) = L^T \Sigma L. \tag{4.57}$$

Hence $V \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$ by the Cramér-Wold device. Now express

$$\sqrt{n}(\tilde{\gamma} - \gamma^*) = D_n V, \text{ where } D_n = Diag\Big(\frac{1}{\phi_1(\Phi_1^{-1}(t_{1,n}))}, \cdots \frac{1}{\phi_1(\Phi_1^{-1}(t_{(r+s-2),n}))}\Big) \tag{4.58}$$

and

$$t_{i,n} = \begin{cases} (1-\epsilon_i^*)P_{is} + \epsilon_i^* G_{is} & \text{for } i = 1, \ldots, (r-1), \\ (1-\epsilon_i^*)P_{r(i-r+1)} + \epsilon_i^* G_{r(i-r+1)} & \text{for } i = r+1, \ldots, (r+s-2). \end{cases} . \tag{4.59}$$

Here ϵ^*s are fixed in $(0,1)$. As

$$D_n \xrightarrow{\mathbb{P}} D \text{ where } D = Diag\left(\frac{1}{\phi_1(\eta_1)}, \cdots, \frac{1}{\phi_1(\eta_{r-1})}, \frac{1}{\phi_1(\beta_1)}, \cdots, \frac{1}{\phi_1(\beta_{s-1})}\right), \tag{4.60}$$

we obtain $\sqrt{n}(\tilde{\gamma} - \gamma^*) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, D\Sigma D\Big)$ as $n \to \infty$. This completes the proof. □

4.3.3 Test Statistic

Now consider the problem of testing a simple null hypothesis against two-sided alternatives. The polychoric correlation functional in two-step approach is denoted by $T_\alpha^*(G) = \rho_\alpha^*$. We wish to test the following statistical hypothesis

$$\mathbb{H} : \rho = r_h \text{ V/s } \mathbb{K} : \rho \neq r_h \text{ for some fixed } r_h. \tag{4.61}$$

Let G_h be the true probability distribution under the null hypothesis $\mathbb{H}$ such that $T^*_\alpha(G_h) = r_h$. A family of Wald-type test statistics is given by

$$\widetilde{W}_\alpha = \frac{n(\tilde{\rho}_\alpha - r_h)^2}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}}. \tag{4.62}$$

We would reject $\mathbb{H}$ for large values of $\widetilde{W}_\alpha$ at fixed α. To find the critical value, it is required to find the asymptotic distribution of $\widetilde{W}_\alpha$ under the null hypothesis. In Theorem 4.4 we see the asymptotic distribution of $\tilde{\rho}_\alpha$ depends on a bias term. This yields an additional complexity in finding the asymptotic distribution of $\widetilde{W}_\alpha$. To simplify that we take $\mathcal{C} = 0$. Later, we shall see that the assumption $\mathcal{C} = 0$ does not entail much difference in simulation studies. Under that assumption it is easy to see that $\widetilde{W}_\alpha$ is central χ^2_1. Under two-sided alternatives, the power function at $100c\%$ nominal level of significance is given by

$$\widetilde{\xi}_\alpha(G) = \mathbb{P}\Big\{\widetilde{W}_\alpha > \chi^2_{1,c} \text{ under true } G\Big\}, \tag{4.63}$$

where $\chi^2_{1,c}$ is the upper $100c\%$ point of central χ^2_1 distribution. Using Theorem 4.4, it can be easily shown that

$$\Big|\widetilde{\xi}_\alpha(G) - \Big(1 - F_{\chi^2_1(\delta^{*2}_n)}(\chi^2_{1,c})\Big)\Big| \longrightarrow 0 \text{ for } n \to \infty, \tag{4.64}$$

under $\mathbb{K}$, where $F_{\chi^2(\delta^{*2}_n)}$ is the CDF of noncentral χ^2_1 distribution with noncentrality parameter (n.c.p) $\delta^{*2}_n = n\Big[\frac{(T^*_\alpha(G)-r_h)}{v_{\rho^*_\alpha,\gamma^*}}\Big]^2$. Under the null hypothesis see that $\delta^{*2}_n = 0$, hence $\lim_{n\to\infty}\widetilde{\xi}_\alpha(G_h) = c$. When the alternative hypothesis is true, the *n.c.p* depends on n and the polychoric correlation functional at unknown true G. When $\mathbb{K}$ is true, there exists a $r_k \neq r_h$ such that $T^*_\alpha(P) \xrightarrow{\mathbb{P}} T^*_\alpha(G_k) = r_k$ for some G_k. In the following theorem the power of $\widetilde{W}_\alpha$ is approximated which does not involve the n.c.p. Further a simplified result is presented under contiguous alternatives. As done previously, we define $q(\rho,\gamma) = \frac{(\rho-r_h)^2}{v^2_{\rho,\gamma}}$.

Theorem 4.6. *Suppose the assumptions of Theorem 4.4 are true, and $\mathcal{C} = 0$ under true distribution. Assume $\frac{\partial q(\rho,\gamma)}{\partial \rho} \neq 0$ and $\frac{\partial^2 q(\rho,\gamma)}{\partial \rho^2}$ is bounded in a small neighbourhood of r_k and $\tilde{\gamma}$.*

(i) Then the power of $\widetilde{W}_\alpha$ under the alternative $\mathbb{K} : \rho = r_k$ has the following limit

$$\left|\widetilde{\xi}_\alpha(G_k) - \left(1 - \Phi_1\Big(\frac{1}{\sigma(r_k)}\Big(\frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n}q(r_k,\tilde{\gamma}_\alpha)\Big)\Big)\right)\right| \longrightarrow 0 \text{ as } n \to \infty, \tag{4.65}$$

where $\sigma^2(r_k) = q'^2_*(r_k)v^2_{r_k,\gamma^*(G_k)}$ such that $q'_*(r_k) = \frac{\partial q(r_k,\gamma^*(G_k))}{\partial \rho}$.

(ii) Consider the sequence of contiguous alternatives $\mathbb{K}_n : \rho = r_h + n^{-1/2}d$, where $d \neq 0$. If $\mathbb{K}_n \longrightarrow \mathbb{H}$ implies that $G_{k_n} \longrightarrow G_h$, then

$$\widetilde{\xi}_\alpha(G_{k_n}) \longrightarrow 1 - F_{\chi^2_1(\delta^{*2})}(\chi^2_{1,c}), \tag{4.66}$$

where $\delta^{*2}_n = \frac{d^2}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}} \xrightarrow{\mathbb{P}} \frac{d^2}{v^2_{r_h,\gamma^*(G_h)}} = \delta^{*2}$ and G_{k_n} is the CDF associated with $\mathbb{K}_n$.

Proof. (i) See that $\widetilde{W}_\alpha = nq(\tilde{\rho}_\alpha,\tilde{\gamma})$. The power of the test statistic $\widetilde{W}_\alpha$ at $\mathbb{K} : \rho = r_k$ is given by

$$\widetilde{\xi}_\alpha(G_k) = \mathbb{P}_{G_k}\Big(\widetilde{W}_\alpha > \chi^2_{1,c}\Big) = \mathbb{P}_{G_k}\Big[\sqrt{n}\Big(q(\tilde{\rho}_\alpha,\tilde{\gamma}) - q(r_k,\tilde{\gamma})\Big) > \frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n}q(r_k,\tilde{\gamma})\Big]. \tag{4.67}$$

A first-order Taylor series expansion of $\sqrt{n}q(\tilde{\rho}_\alpha,\tilde{\gamma})$ around r_k gives

$$\sqrt{n}\Big(q(\tilde{\rho}_\alpha,\tilde{\gamma}) - q(r_k,\tilde{\gamma})\Big) = \sqrt{n}(\tilde{\rho}_\alpha - r_k)\Big[\frac{\partial q(r_k,\tilde{\gamma})}{\partial \rho}\Big] + R_n, \tag{4.68}$$

where the remainder is given by

$$R_n = \sqrt{n}\frac{(\tilde{\rho}_\alpha - r_k)^2}{2}\left[\frac{\partial^2}{\partial \rho^2}q(\rho,\tilde{\gamma})\right]_{\rho=\rho^*} \text{ as } \rho^* \text{ lies in } \tilde{\rho}_\alpha, r_k. \tag{4.69}$$

Note that $\tilde{\rho}_\alpha \xrightarrow{\mathbb{P}} r_k$ and $\tilde{\gamma}_\alpha \xrightarrow{\mathbb{P}} \gamma^*(G_k)$ under $\mathbb{K}$. Using the boundedness condition of the second-order derivative, we find that

$$R_n = (\tilde{\rho}_\alpha - r_k) \cdot \underbrace{\sqrt{n}(\tilde{\rho}_\alpha - r_k)}_{\mathcal{O}_{\mathbb{P}}(1)} \cdot \mathcal{O}_{\mathbb{P}}(1) = o_{\mathbb{P}}(1). \tag{4.70}$$

Also see that $\frac{\partial}{\partial \rho} q(r_k, \tilde{\gamma}) \xrightarrow{\mathbb{P}} q'_*(r_k)$. We have also assumed that $\mathcal{C} = 0$ under the true distribution, so

$$\sqrt{n}\Big(q(\rho^*_\alpha, \tilde{\gamma}) - q(r_k, \tilde{\gamma})\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \underbrace{q'_*(r_k)^2 \nu^2_{r_k, \gamma^*(G_k)}}_{\sigma^2(r_k)}\Big). \tag{4.71}$$

Therefore we obtain

$$\left| \widetilde{\xi}_\alpha(G_k) - \left\{ 1 - \Phi_1\Big(\frac{1}{\sigma(r_k)} \Big(\frac{\chi^2_{1,c}}{\sqrt{n}} - \sqrt{n} q(r_k, \tilde{\gamma}) \Big) \Big) \right\} \right| \longrightarrow 0 \text{ as } n \to \infty. \tag{4.72}$$

(ii) Denote $r_{k_n} = r_h + n^{-1/2} d$, so $r_{k_n} = T^*_\alpha(G_{k_n})$. Now see that

$$\begin{aligned} \widetilde{W}_\alpha = \frac{n(\tilde{\rho}_\alpha - r_h)^2}{\nu^2_{\tilde{\rho}_\alpha, \tilde{\gamma}}} &= \frac{n(\tilde{\rho}_\alpha - r_{k_n})^2}{\nu^2_{\tilde{\rho}_\alpha, \tilde{\gamma}}} + \frac{n(r_{k_n} - r_h)^2}{\nu^2_{\tilde{\rho}_\alpha, \tilde{\gamma}}} + 2\frac{n(\tilde{\rho}_\alpha - r_{k_n})(r_{k_n} - r_h)}{\nu^2_{\tilde{\rho}_\alpha, \tilde{\gamma}}} \\ &= \left(Z_n + \frac{d}{\nu_{\tilde{\rho}_\alpha, \tilde{\gamma}}} \right)^2 \text{ where } Z_n = \frac{\sqrt{n}(\tilde{\rho}_\alpha - r_{k_n})}{\nu_{\tilde{\rho}_\alpha, \tilde{\gamma}}}. \end{aligned} \tag{4.73}$$

Since we assume that $\mathbb{K}_n \to \mathbb{H}$ implies $G_{k_n} \to G_h$, so we have $(\tilde{\rho}_\alpha, \tilde{\gamma}) \xrightarrow{\mathbb{P}} (r_h, \gamma^*(G_h))$.

$$Z_n = \Big(\frac{\nu_{r_{k_n}, \gamma^*(G_{k_n})}}{\nu_{\tilde{\rho}_\alpha, \tilde{\gamma}}} \Big) \times U_n \text{ where } U_n = \frac{\sqrt{n}(\tilde{\rho}_\alpha - r_{k_n})}{\nu_{r_{k_n}, \gamma^*(G_{k_n})}}. \tag{4.74}$$

Using Theorem 4.4 along with $\mathcal{C} = 0$ gives

$$U_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ under } \mathbb{K}_n \text{ as } n \to \infty. \tag{4.75}$$

The coefficient associated with U_n converges to 1 in probability. So applying the Slutsky's theorem we obtain $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ under $\mathbb{K}_n$. Thus $\widetilde{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^{*2})$ as $n \to \infty$ where $\delta^{*2} = \frac{d^2}{v^2_{r_h,\gamma^*(G_h)}}$. This completes the proof.

□

The calculation of the power function under K_n still requires a heavy computational burden to carry out.

4.4 Robustness Study

In this section, we shall study the stability behaviour of the proposed estimator and the test statistic with the assumption that the true distribution becomes contaminated. This is mainly done by computing the influence functions (IF) of the polychoric correlation and the Wald-type test functionals. Unlike the earlier analysis, the influence functions in the two-step situation depend on that of γ^*.

4.4.1 Influence Function of the Polychoric Correlation Functional

Let the true density g be contaminated as before. The contaminated functionals are given by $\rho^*_{\epsilon,\alpha}$ and γ^*_ϵ. We know that

$$\mathbb{E}_G\Big[\frac{\partial}{\partial\rho} V(\rho^*_\alpha, \gamma^*, Z_l)\Big] = 0. \tag{4.76}$$

Substituting G_ϵ, $\rho^*_{\alpha,\epsilon}$ and $\gamma_{\epsilon_\epsilon}$ for the respective terms in the last equation gives

$$(1-\epsilon)\mathbb{E}_G\Big[\frac{\partial}{\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z_l)\Big]+\epsilon\frac{\partial}{\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z^o_l)=0. \tag{4.77}$$

Differentiating (4.77) with respect to ϵ gives

$$-\mathbb{E}_G\Big[\frac{\partial}{\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z_l)\Big] \tag{4.78}$$

$$+(1-\epsilon)\mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho^2}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z_l)\Big]\frac{\partial\rho^*_{\epsilon,\alpha}}{\partial\epsilon}+(1-\epsilon)\mathbb{E}_G\Big[\frac{\partial^2}{\partial\gamma\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z_l)\Big]\frac{\partial\gamma^*_\epsilon}{\partial\epsilon}^T$$

$$+\frac{\partial}{\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z^o_l)+\epsilon\frac{\partial^2}{\partial\rho^2}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z^o_l)\frac{\partial\rho^*_{\epsilon,\alpha}}{\partial\epsilon}+\epsilon\frac{\partial^2}{\partial\gamma\partial\rho}V(\rho^*_{\epsilon,\alpha},\gamma^*_\epsilon,Z^o_l)\frac{\partial\gamma^*_\epsilon}{\partial\epsilon}^T=0. \tag{4.79}$$

Evaluating the above expression at $\epsilon=0$ gives

$$\mathcal{IF}_1(T^*_\alpha,G,Z^o_l)=-\frac{\frac{\partial}{\partial\rho}V(\rho^*_\alpha,\gamma^*,Z^o_l)}{\mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho^2}V(\rho^*_\alpha,\gamma^*,Z_l)\Big]+\mathcal{C}^T\mathcal{IF}_1(\gamma^*,G,Z^o_l)}. \tag{4.80}$$

Note the influence function of T^*_α depends on the influence function of γ^*. Eliminating the effect γ^*, or taking $\mathcal{C}=0$ gives

$$\mathcal{IF}_1(T^*_\alpha,G,Z^o_l)=\frac{-\frac{\partial}{\partial\rho}V(\rho^*_\alpha,\gamma^*,Z^o_l)}{\mathbb{E}_G\Big[\frac{\partial^2}{\partial\rho^2}V(\rho^*_\alpha,\gamma^*,Z_l)\Big]}\text{ for all }Z^o_l. \tag{4.81}$$

Since it is required, we also compute the influence function of γ^* as

$$\mathcal{IF}_1(\gamma^*,G,Z^o_l)=\Big[\frac{\partial\gamma^*_\epsilon}{\partial\epsilon}\Big]_{\epsilon=0}$$

$$=\left(\frac{\Lambda_{1s}(Z^o_l)-G_{1s}}{\phi_1(\eta_1)},\ \cdots,\frac{\Lambda_{(r-1)s}(Z^o_l)-G_{(r-1)s}}{\phi_1(\eta_{r-1})},\frac{\Lambda_{r1}(Z^o_l)-G_{r1}}{\phi_1(\beta_1)},\ \cdots,\frac{\Lambda_{r(s-1)}(Z^o_l)-G_{r(s-1)}}{\phi_1(\beta_{s-1})}\right) \tag{4.82}$$

where $\Lambda_{1s}(Z^o_l)=\sum_{l_1=1}^{i}\sum_{k_1=1}^{j}\delta_{ij}(Z^o_l)$ for all i,j.

4.4.2 Influence Function of the Wald-type Test Functional

As in Subsection 4.4.1, now we will study the stability behaviour of the Wald-type test functional corresponding to the test statistic defined in (4.62). As before, the Wald-type test functional (ignoring the multiplier n) in this context may be given by

$$W_\alpha^*(G) = \left[\frac{T_\alpha^*(G) - r_h}{v_{\tilde{\rho}_\alpha,\tilde{\gamma}}}\right]^2. \tag{4.83}$$

At a fixed contamination proportion $\epsilon \in [0,1]$, the test functional is similarly defined as $W_\alpha^*(G_\epsilon)$. Differentiating $W_\alpha^*(G_\epsilon)$ with respect to ϵ, and evaluating at $\epsilon = 0$ gives

$$\mathcal{IF}_1(W_\alpha^*, G, Z_l^o) = 2\left[\frac{T_\alpha^*(G) - r_h}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}}\right]\mathcal{IF}_1(T_\alpha^*, G, Z_l^o) \text{ for all } Z_l^o, \tag{4.84}$$

which at the null hypothesis $\mathbb{H}$ becomes zero identically. Therefore a first-order IF of Wald-type test functional reveals no further information about robustness for whatever the tuning parameter α may be. To gain further insight, therefore, it is necessary to compute the second-order influence function. Simple calculations give

$$\begin{aligned}\mathcal{IF}_2(W_\alpha^*, G_h, Z_l^o) &= 2\left[\frac{\mathcal{IF}_1(T_\alpha^*, G_h, Z_l^o)}{v_{\rho_\alpha,\tilde{\gamma}}}\right]^2 \\ &= \frac{2}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}}\cdot\left[\frac{\frac{\partial}{\partial\rho}V(\rho_\alpha^*,\gamma^*,Z_l^o)}{\mathbb{E}_G\left[\frac{\partial^2}{\partial\rho^2}V(\rho_\alpha^*,\gamma^*,Z_l)\right] + \mathcal{C}^T\mathcal{IF}_1(\gamma^*, G_h, Z_l^o)}\right]^2 \text{ for all } Z_l^o,\end{aligned} \tag{4.85}$$

under the null hypothesis $\mathbb{H}$. Unlike the one-step approach, the influence function of the Wald-type test functional turns out to be more complicated in the two-step case.

In Figure 4.1, we plot the asymptotic variance, gross error sensitivity of the polychoric correlation, and Wald-type test functional. For an easy comparison, graphs related to

both these approaches are presented side-wise. As seen in this plot higher values of α increase stability in both the polychoric and test functionals across one-step and two-step methods. When choosing between these two methods, the one-step approach is preferred over the other, because it produces lower asymptotic variances over α. However, we should not overlook, the fact, that the difference between asymptotic variances corresponding to these methods becomes very close when $\alpha \to 0$, and widens only when α becomes larger. This difference may be due to not taking into consideration the asymptotic covariance between $\tilde{\rho}_\alpha$ and $\tilde{\gamma}$ as we find it mathematically intractable.

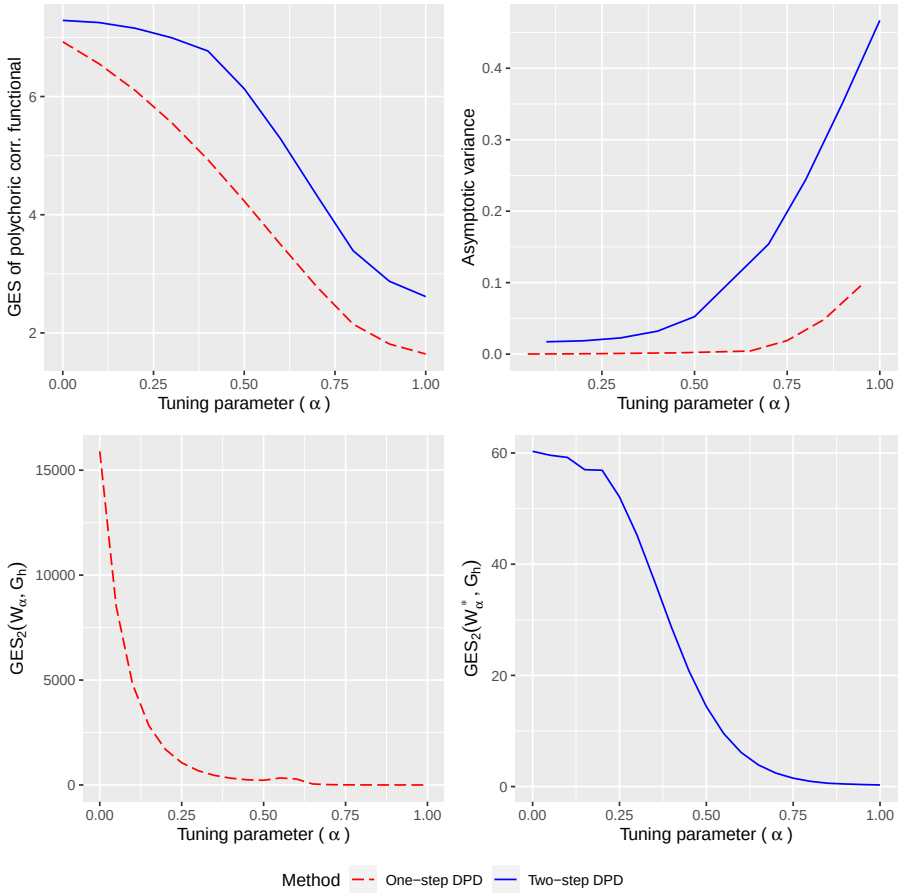

FIGURE 4.1: Comparison of GES and asymptotic variance for one-step and two-step (with $\mathcal{C} = 0$) MDPD estimates of the polychoric correlation when the latent vector (U, V) is generated through standard bivariate normal distribution with $\rho = 0.75$, and the cut-offs are considered as $\eta = (-\infty, 0.7, 1.25, \infty)$, $\beta = (-\infty, -0.67, 0.67, \infty)$.

When $G \neq G_h$ things get harder, because in those cases we do not have simplified expressions for the second-order influence functions as the alternative hypothesis $\mathbb{K}$ is a composite one. To study the stability of the tests at contiguous alternatives, we therefore calculate the level and power in the next subsection.

4.4.3 Level and Power Influence Functions

Now, we shall study the local stability of the type-I and type-II errors for the Wald-type test statistic. In particular, we shall calculate the first-order influence functions of the asymptotic level and power functions at the true distributions associated with the contaminated null and contaminated contiguous alternatives. As before, the sequence of contiguous alternatives is denoted by $\mathbb{K}_n : \rho = r_h + \frac{d}{\sqrt{n}}$ for some $d \neq 0$. All the notations from the previous chapter carry forward over here.

The level of $\tilde{W}_\alpha$ at the contaminated null, and its power at the sequence of contaminated contiguous alternatives are respectively given by

$$\alpha(G^L_{h,\epsilon_n}, Z^o_l) = \mathbb{P}_{G^L_{h,\epsilon_n}}\left\{\widetilde{W}_\alpha > \chi^2_{1,c}\right\} \text{ and } \pi(G^P_{k_n,\epsilon_n}, Z^o_l) = \mathbb{P}_{G^P_{k_n,\epsilon_n}}\left\{\widetilde{W}_\alpha > \chi^2_{1,c}\right\}. \tag{4.86}$$

The level and power IF based on $\widetilde{W}_\alpha$ are similarly given by

$$\mathcal{LIF}(\alpha, G_h, Z^o_l) = \lim_{n\to\infty}\left[\frac{\partial\alpha(G^L_{h,\epsilon_n}, Z^o_l)}{\partial\epsilon}\right]_{\epsilon=0} \text{ and } \mathcal{PIF}(\pi, G_h, Z^o_l) = \lim_{n\to\infty}\left[\frac{\partial\pi(G^P_{k_n,\epsilon_n}, Z^o_l)}{\partial\epsilon}\right]_{\epsilon=0}.$$

As before, we define $\delta^*_{d,\epsilon} = \frac{d+\epsilon\mathcal{IF}_1(T^*_\alpha, G_h, Z^o_l)}{v_{\rho^*_\alpha,\gamma(G_h)}}$. To calculate the values of $\mathcal{LIF}$ and $\mathcal{PIF}$, we need the following results.

Theorem 4.7. *Suppose T^*_α has a non-zero Hadamard derivative at G_{k_n}. Moreover, the assumptions of Theorem 4.4 are true along with $\mathcal{C} = 0$. Then we have*

$$\widetilde{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^{*2}_{d,\epsilon}) \text{ under } G^P_{k_n,\epsilon_n} \text{ as } n \to \infty, \tag{4.87}$$

for each $\epsilon > 0$. The limiting distribution is noncentral chi-squared with 1 df and the n.c.p. $\delta^{*2}_{d,\epsilon}$.

Proof. Let us express $\widetilde{W}_\alpha$ as

$$\widetilde{W}_\alpha = \frac{n(\tilde{\rho}_\alpha - r_h)^2}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}} = S_{1n} + S_{2n} + S_{3n}, \tag{4.88}$$

where

$$S_{1n} = \left[\frac{\sqrt{n}(\tilde{\rho}_\alpha - \rho^P_{n,\alpha})}{v_{\tilde{\rho}_\alpha,\tilde{\gamma}}}\right]^2, S_{2n} = \left[\frac{\sqrt{n}(\rho^{*P}_{n,\alpha} - r_h)}{v_{\tilde{\rho}_\alpha,\tilde{\gamma}}}\right]^2, S_{3n} = \frac{2n(\tilde{\rho}_\alpha - \rho^P_{n,\alpha})(\rho^{*P}_{n,\alpha} - r_h)}{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}}, \tag{4.89}$$

and $\rho^{*P}_{n,\alpha} = \arg\min_\rho d_\alpha(g^P_{k_n,\epsilon_n}, \pi(\rho, \gamma^*_\epsilon))$. When the true distribution is $G^P_{k_n,\epsilon_n}$, the term inside the squared bracket of S_{1n} is asymptotically $\mathcal{N}(0,1)$ by Theorem 4.6. So $S_{1n} \xrightarrow{\mathcal{L}} Z^2$, where $Z \sim \mathcal{N}(0,1)$ under $G^P_{k_n,\epsilon_n}$. The numerator of S_{2n} is expressed as

$$\begin{aligned}\left[\sqrt{n}(\rho^{*P}_{n,\alpha} - r_h)\right]^2 &= \left[\sqrt{n}\Big(T^*_\alpha(G^P_{k_n,\epsilon_n}) - T^*_\alpha(G_h)\Big)\right]^2 \\ &= \left[\underbrace{\sqrt{n}\Big(T^*_\alpha(G^P_{k_n,\epsilon_n}) - T^*_\alpha(G_{k_n})\Big)}_{S_{2na}} + \underbrace{\sqrt{n}\Big(T^*_\alpha(G_{k_n}) - T^*_\alpha(G_h)\Big)}_{S_{2nb}}\right]^2.\end{aligned} \tag{4.90}$$

Using the same argument as before, we see that $S_{2na} = \epsilon\mathcal{IF}_1(T^*_\alpha, G_h, Z^o_l) + o_{\mathbb{P}}(1)$. Also, see that $S_{2nb} = \sqrt{n}\Big(T^*_\alpha(G_{k_n}) - T^*_\alpha(G_h)\Big) = d$ as $\mathbb{K}_n : \rho = r_h + \frac{d}{\sqrt{n}}$. Further, see that

$$(G^P_{k_n,\epsilon_n} - G_{k_n}) \xrightarrow{\mathbb{P}} 0 \text{ and } (G_{k_n} - G_h) \longrightarrow 0 \Big[\text{ By assumption }\Big] \implies (G^P_{k_n,\epsilon_n} - G_h) \xrightarrow{\mathbb{P}} 0 \tag{4.91}$$

as $n \to \infty$. Also note that $\Big(P - G^P_{k_n,\epsilon_n}\Big) \xrightarrow{\mathbb{P}} 0$ under $G^P_{k_n,\epsilon_n}$, which implies $P \xrightarrow{\mathbb{P}} G_h$ for $n \to \infty$. Hence $\nu_{\tilde{\rho}_\alpha,\tilde{\gamma}} \xrightarrow{\mathbb{P}} \nu_{r_h,\gamma^*(G_h)}$, and finally $S_{2n} \xrightarrow{\mathbb{P}} \delta^{*2}_{d,\epsilon}$ for $n \to \infty$. Similarly, $S_{3n} \xrightarrow{\mathcal{L}} 2Z\delta^*_{d,\epsilon}$. Using Slutsky's theorem we finally get

$$\widetilde{W}_\alpha \xrightarrow{\mathcal{L}} (Z + \delta^*_{d,\epsilon})^2 \text{ as } n \to \infty. \tag{4.92}$$

However, we know that $(Z + \delta^*_{d,\epsilon})^2$ has a noncentral chi-squared distribution with 1 df and n.c.p $\delta^{*2}_{d,\epsilon}$. This completes the proof. □

Corollary 4.1. *Substituting $\epsilon = 0$ in Theorem 4.7, we get $\delta^{*2}_{d,0} = \delta^{*2}$ as defined in Theorem 4.6 (ii). Thus the power of the test under contiguous alternatives has the following limit*

$$\Big|\pi(G^P_{k_n,\epsilon_n}, Z^o_l) - \mathbb{P}\Big\{\chi^2_1(\delta^{*2}) > \chi^2_{1,c}\Big\}\Big| \longrightarrow 0, \textit{ for } n \to \infty. \tag{4.93}$$

Corollary 4.2. *When $d = 0$, we get $\sqrt{n}\Big(T^*_\alpha(G_{k_n}) - T^*_\alpha(G_h)\Big) = 0$. If the map $G \mapsto T^*_\alpha(G)$ is $1-1$, then $G_{k_n} \equiv G_h$. This implies that $G^P_{k_n,\epsilon_n} \equiv G^L_{h,\epsilon_n}$. Thus it easily follows form Theorem 4.7 that $\widetilde{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1(\delta^{*2}_{0,\epsilon})$ under G^L_{h,ϵ_n}. Moreover, if we make $\epsilon \downarrow 0+$ along with $d = 0$, we get $\widetilde{W}_\alpha \xrightarrow{\mathcal{L}} \chi^2_1$ under G^L_{h,ϵ_n}. In that case, it holds that*

$$\lim_{\epsilon \downarrow 0+} \Big|\alpha(G^L_{h,\epsilon_n}, Z^o_l) - c\Big| \longrightarrow 0 \textit{ as } n \to \infty. \tag{4.94}$$

Exact values of level and power under ϵ_n-contaminated versions of true null and contiguous distributions are not easy to compute. But, they can be approximated under certain conditions. However, when the order of limit and differentiation can be interchanged, the exact expressions of $\mathcal{PIF}$ and $\mathcal{LIF}$ can be obtained with the aid of Theorem 4.7. In the next result, we explicitly mention the condition under which these simplifications hold.

Theorem 4.8. *Suppose* $\left\{\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ *and* $\left\{\frac{\partial}{\partial\epsilon}\alpha(G^L_{h,\epsilon_n}, Z^o_l)\right\}$, *viewed as a function of* ϵ, *converge uniformly in* $[0,1]$ *at fixed* Z^o_l. *Further, assume that the conditions of Theorem 4.7 are true.*

(i) Then the power influence function is obtained as

$$\mathcal{PIF}(\pi, G_h, Z^o_l) = \frac{d\cdot\mathcal{IF}_1(T^*_\alpha, G_h, Z^o_l)}{v^2_{r_h,\gamma_\alpha(G_h)}} e^{-\frac{\delta^{*2}}{2}} \sum_{v=0}^{\infty} \frac{(\delta^{*2}/2)^{v-1}}{v!}\left(v-\frac{\delta^{*2}}{2}\right)\mathbb{P}\left\{\chi^2_{1+2v} > \chi^2_{1,c}\right\}. \tag{4.95}$$

(ii) The level influence function becomes $\mathcal{LIF}(\alpha, G_h, Z^o_l) \equiv 0$.

Proof. (i) When Z^o_l is fixed, $\pi(G^P_{k_n,\epsilon_n}, Z^o_l)$ is a sequence of functions on $[0,1]$. Using the result of Theorem 4.7, we know that the sequence $\pi(G^P_{k_n,\epsilon_n}, Z^o_l)$ is convergent to the following limit

$$\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n}, Z^o_l) = \sum_{v=0}^{\infty}\frac{e^{-\frac{\delta^{*2}_{d,\epsilon}}{2}}}{v!}\left(\frac{\delta^{*2}_{d,\epsilon}}{2}\right)^v \mathbb{P}\left\{\chi^2_{1+2v} > \chi^2_{1,c}\right\} \text{ for each } \epsilon\in[0,1]. \tag{4.96}$$

This along with the assumption of uniform convergence of $\left\{\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ imply the uniform convergence of the sequence of functions $\left\{\pi(G^P_{k_n,\epsilon_n}, Z^o_l)\right\}$ itself.

Hence the interchanging of limit and differentiation is permissible. So we get

$$\lim_{n\to\infty}\frac{\partial}{\partial\epsilon}\pi(G^P_{k_n,\epsilon_n},Z^o_l)=\frac{\partial}{\partial\epsilon}\Big[\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n},Z^o_l)\Big]$$
$$=\frac{\partial}{\partial\epsilon}\Bigg[\sum_{v=0}^{\infty}\frac{e^{-\frac{\delta^{*2}_{d,\epsilon}}{2}}}{v!}\Big(\frac{\delta^{*2}_{d,\epsilon}}{2}\Big)^v\mathbb{P}\Big\{\chi^2_{1+2v}>\chi^2_{1,c}\Big\}\Bigg]. \tag{4.97}$$

Evaluating the right-hand side of (4.97) at $\epsilon=0$, we obtain

$$\mathcal{PIF}(\pi,G_h,Z^o_l)=\left\{\frac{\partial}{\partial\epsilon}\Big[\lim_{n\to\infty}\pi(G^P_{k_n,\epsilon_n},Z^o_l)\Big]\right\}_{\epsilon=0}$$
$$=\frac{d\cdot\mathcal{IF}_1(T^*_\alpha,G_h,Z^o_l)}{v^2_{r_h,\gamma_\alpha(G_h)}}e^{-\frac{\delta^{*2}}{2}}\sum_{v=0}^{\infty}\frac{(\delta^{*2}/2)^{v-1}}{v!}\Big(v-\frac{\delta^{*2}}{2}\Big)\mathbb{P}\Big\{\chi^2_{1+2v}>\chi^2_{1,c}\Big\}. \tag{4.98}$$

(ii) Putting $d=0$ in (4.98) gives $\mathcal{LIF}(\alpha,G_h,Z^o_l)\equiv 0$.

□

Remark 4.3. *When $\mathcal{IF}_1(T^*_\alpha,G_h,Z^o_l)$ is bounded, so is $\mathcal{PIF}$. Also, note that $\mathcal{LIF}$ up to any order is always zero. Hence, the power and level of the test statistic based on the MDPDE are stable at the contaminated null and contaminated contiguous alternatives. Although, the $\mathcal{LIF}$ does not reveal any robustness features. So the proposed Wald-type test statistic is robust in those situations.*

4.5 Simulation Studies

Let the experiment for simulation studies be set up as in Chapter 3. In Figure 4.2 we see that small values of the tuning parameters produce estimates that are almost as good as the MLE in terms of having small bias and MSE. We also find that the observed

TABLE 4.1: Two-step estimates of the polychoric correlation

Methods	Pure data	5%Type 1 cont.	10%Type 1 cont.	5%Type 2 cont.	10%Type 2 cont.
$\hat{\rho}_0$	0.74765	0.48424	0.32478	0.66841	0.60920
$\hat{\rho}_{0.2}$	0.74482	0.54568	0.34951	0.68376	0.62663
$\hat{\rho}_{0.4}$	0.74386	0.61350	0.38600	0.69524	0.64185
$\hat{\rho}_{0.6}$	0.74315	0.66274	0.43855	0.70114	0.65210
$\hat{\rho}_{0.9}$	0.74170	0.69419	0.53176	0.70396	0.65934
HD	0.76880	0.63733	0.47710	0.70354	0.64108
NCS	0.736693	0.70581	0.57917	0.70103	0.66021
SCS	0.75747	0.67352	0.44893	0.70019	0.64133
NED	0.74924	0.64820	0.36322	0.69019	0.62982

levels and powers (albeit a bit lower for higher α) are quite stable across all the values of the tuning parameters. The 95% confidence intervals are plotted in Figure 4.3. Under different types of data contamination, higher value of α adds much stability to the performance of both the estimates (Figures 4.4, 4.5, 4.6) and the tests (Figures 4.7, 4.8).

To compare with the other robust methods such as– those coming out of minimising the SCS (symmetric chi-square), NCS (Neyman chi-square), HD (Hellinger distance), and NED (Negative exponential disparity), we present these estimates in Table 4.1. We see that those methods that perform better than MDPDE in contaminated data sets may work badly for pure data, or, the other way around. Thus MDPDE may win over these methods in terms of performing quite reasonably well in both the pure and contaminated data. Comparing Table 3.2 with Table 4.1, we see that both one-step and two-step estimates of polychoric correlation are almost the same for pure data. However, these estimates slightly differ in 10%Type 1 data contamination. From that, we speculate that one-step MDPDE performs marginally better than its two-step version, at higher data contamination, but at the cost of a huge computational burden.

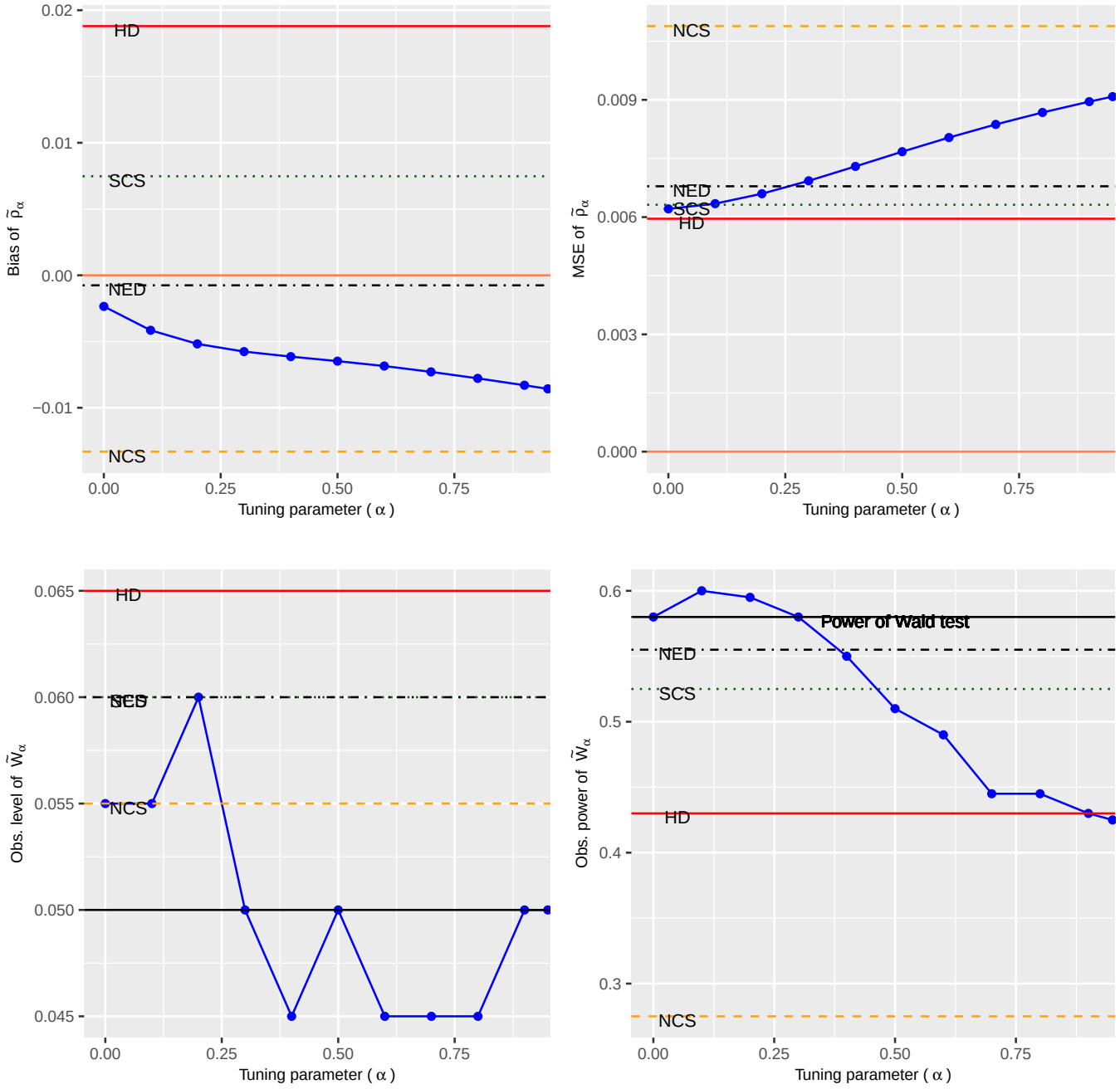

FIGURE 4.2: Plots under pure data

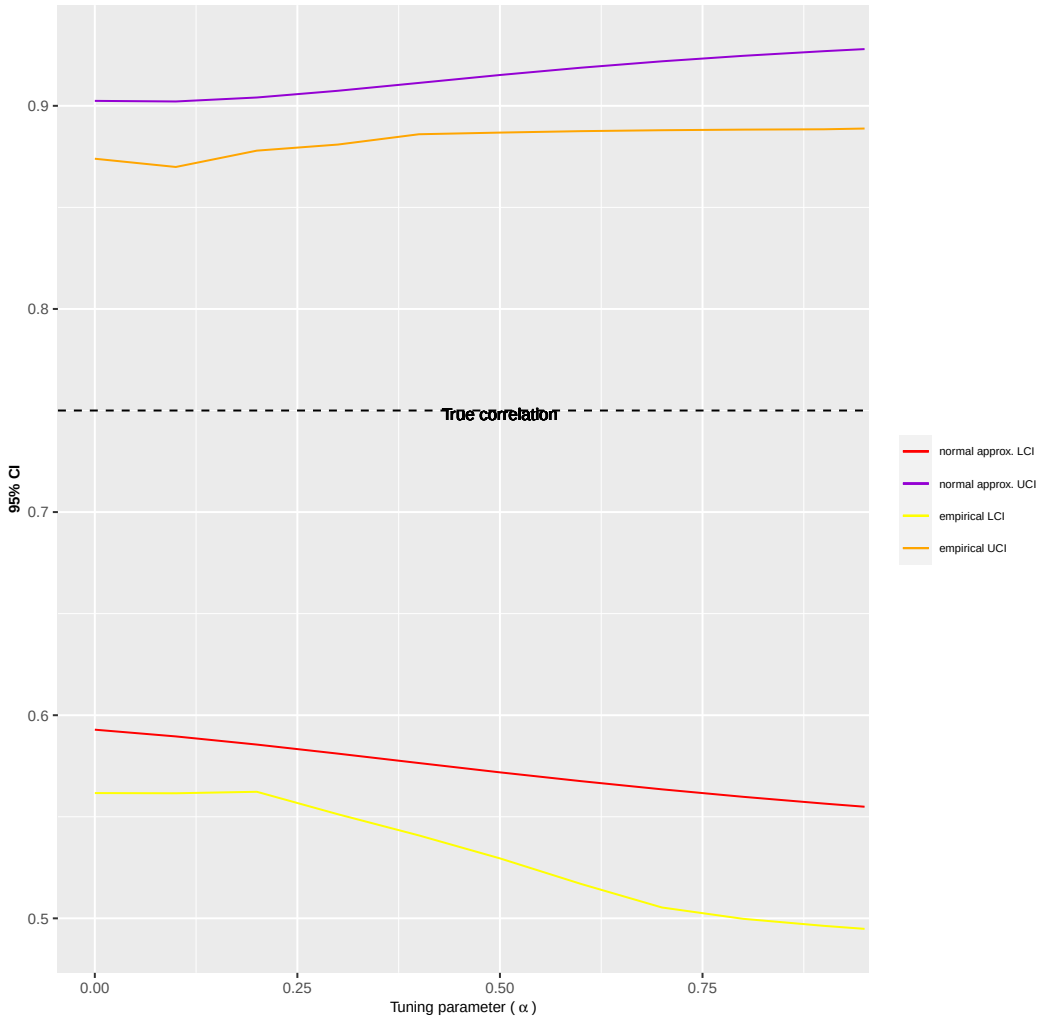

FIGURE 4.3: Plots of confidence intervals of $\tilde{\rho}_{\alpha}$ under pure data

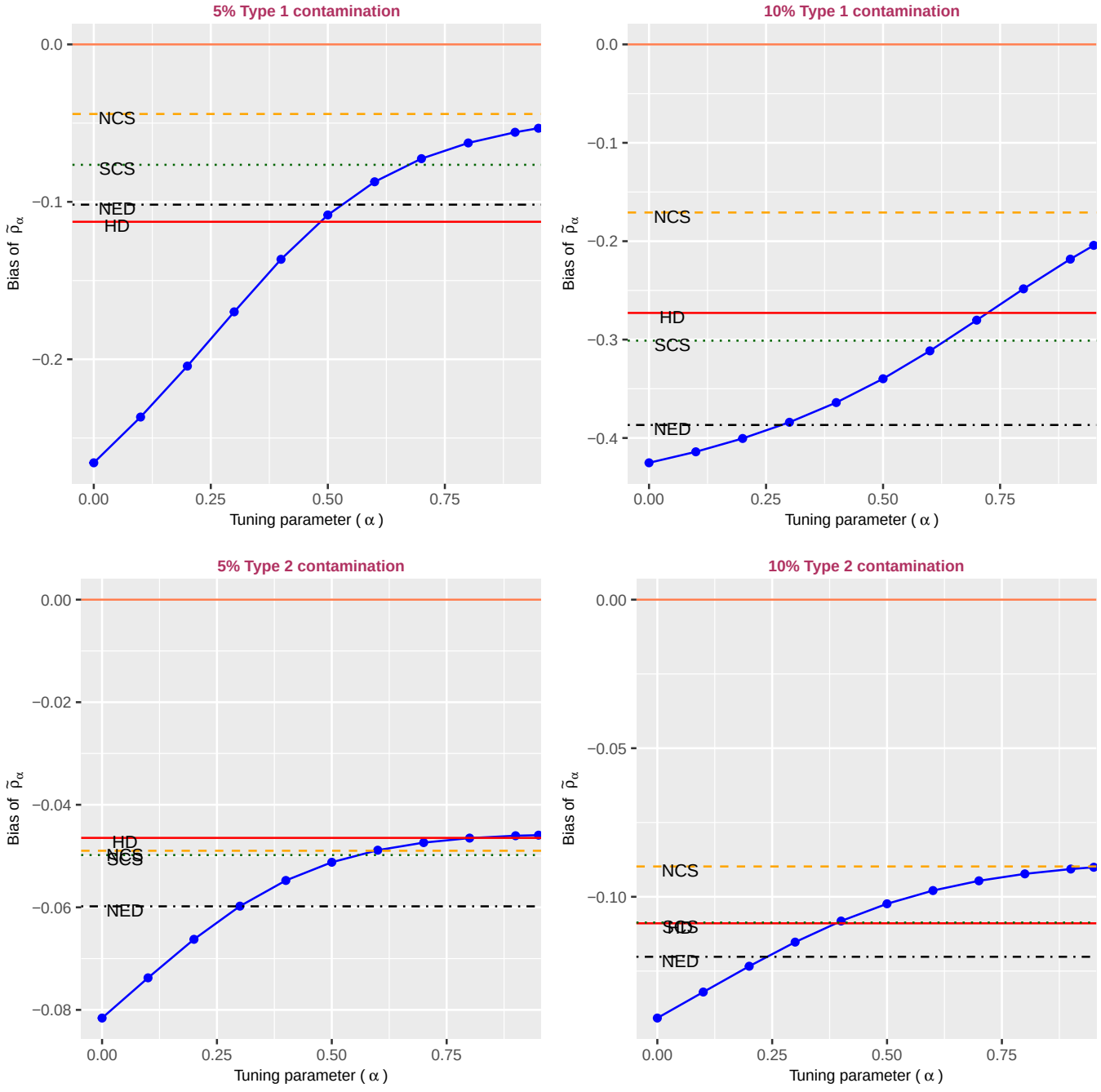

FIGURE 4.4: Bias in two-step estimates of the polychoric correlation under data contamination

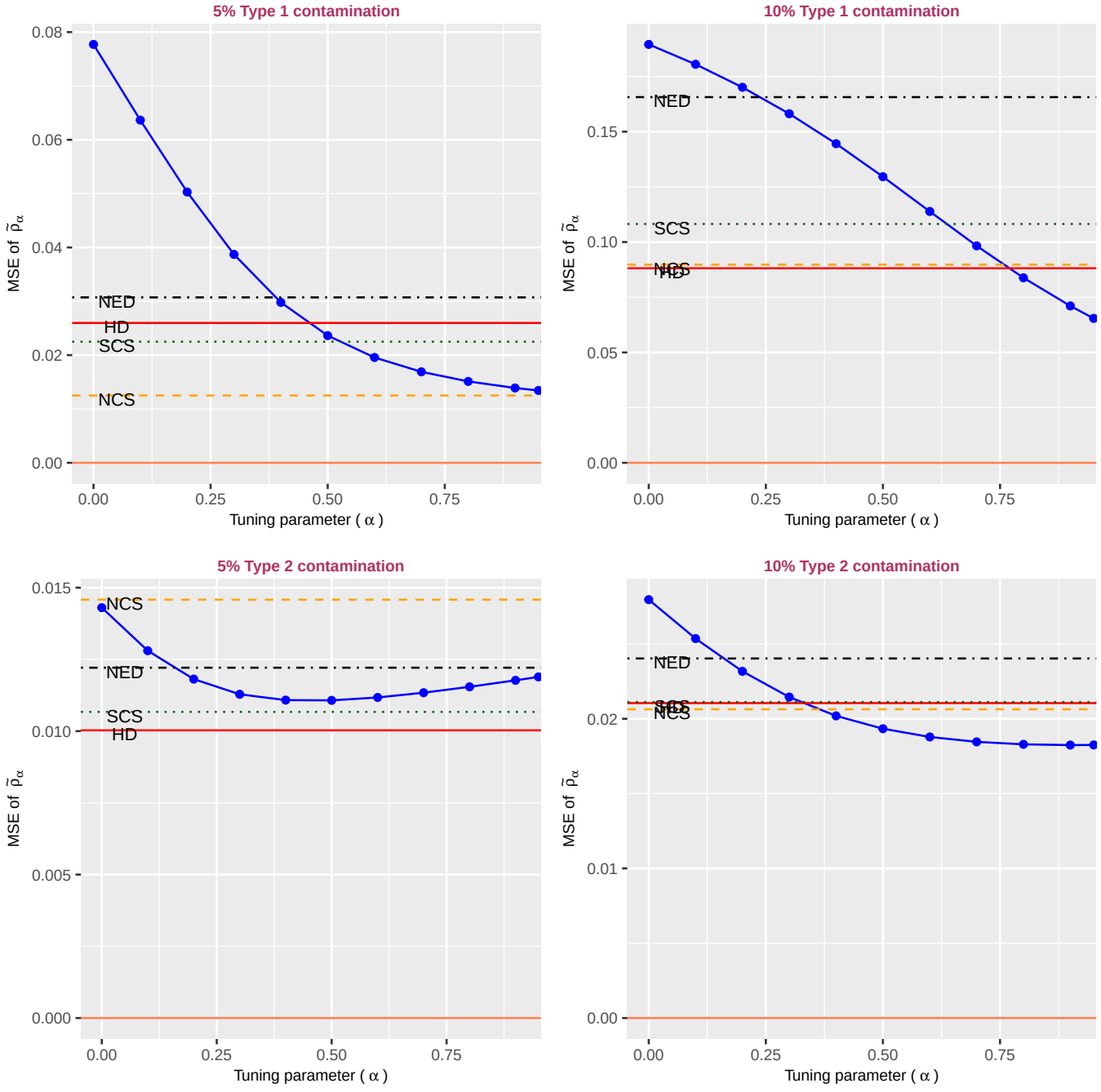

FIGURE 4.5: MSE in two-step estimates of the polychoric correlation under data contamination

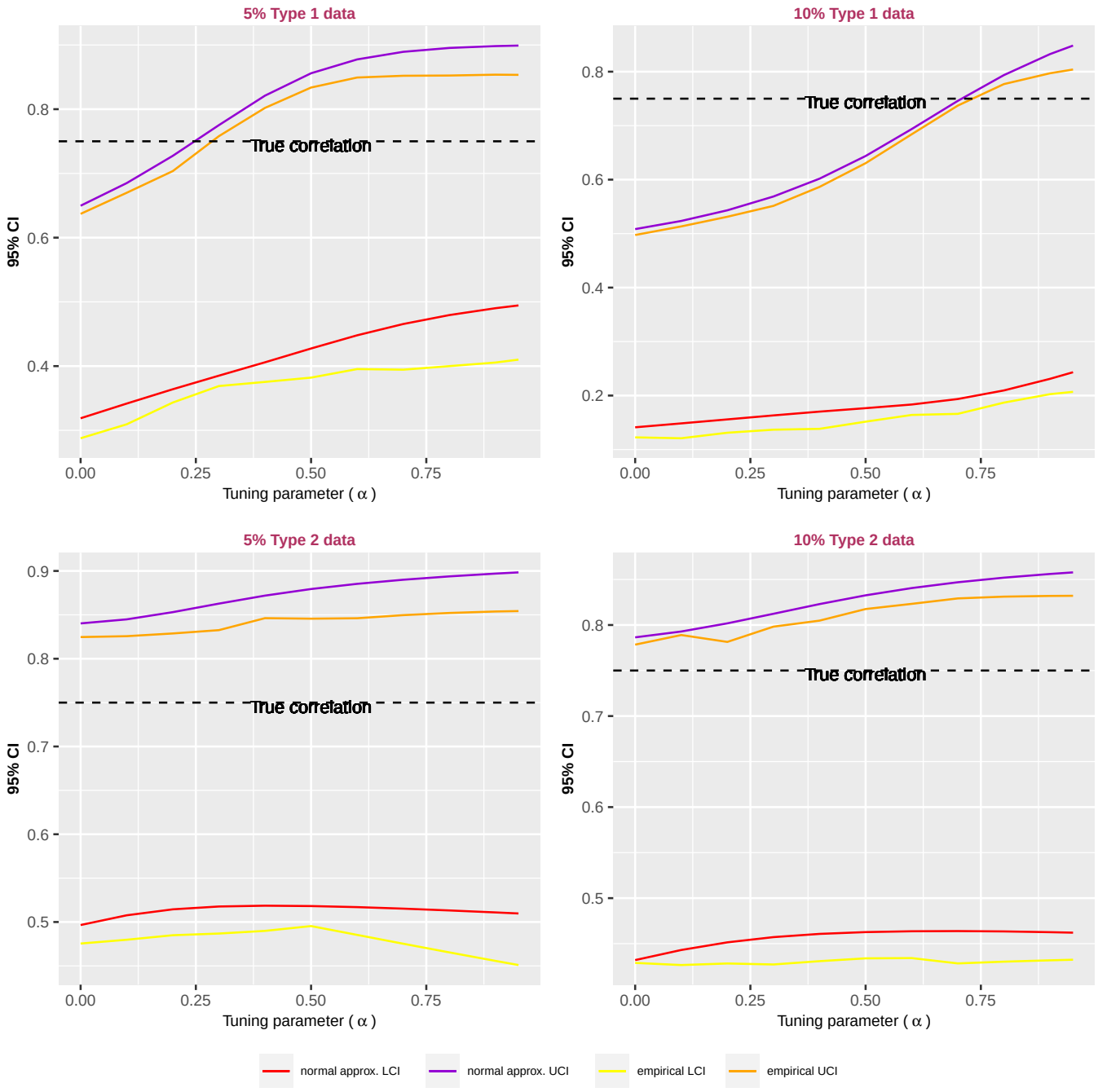

FIGURE 4.6: Confidence intervals (CI) of $\tilde{\rho}_{\alpha}$ under data contamination

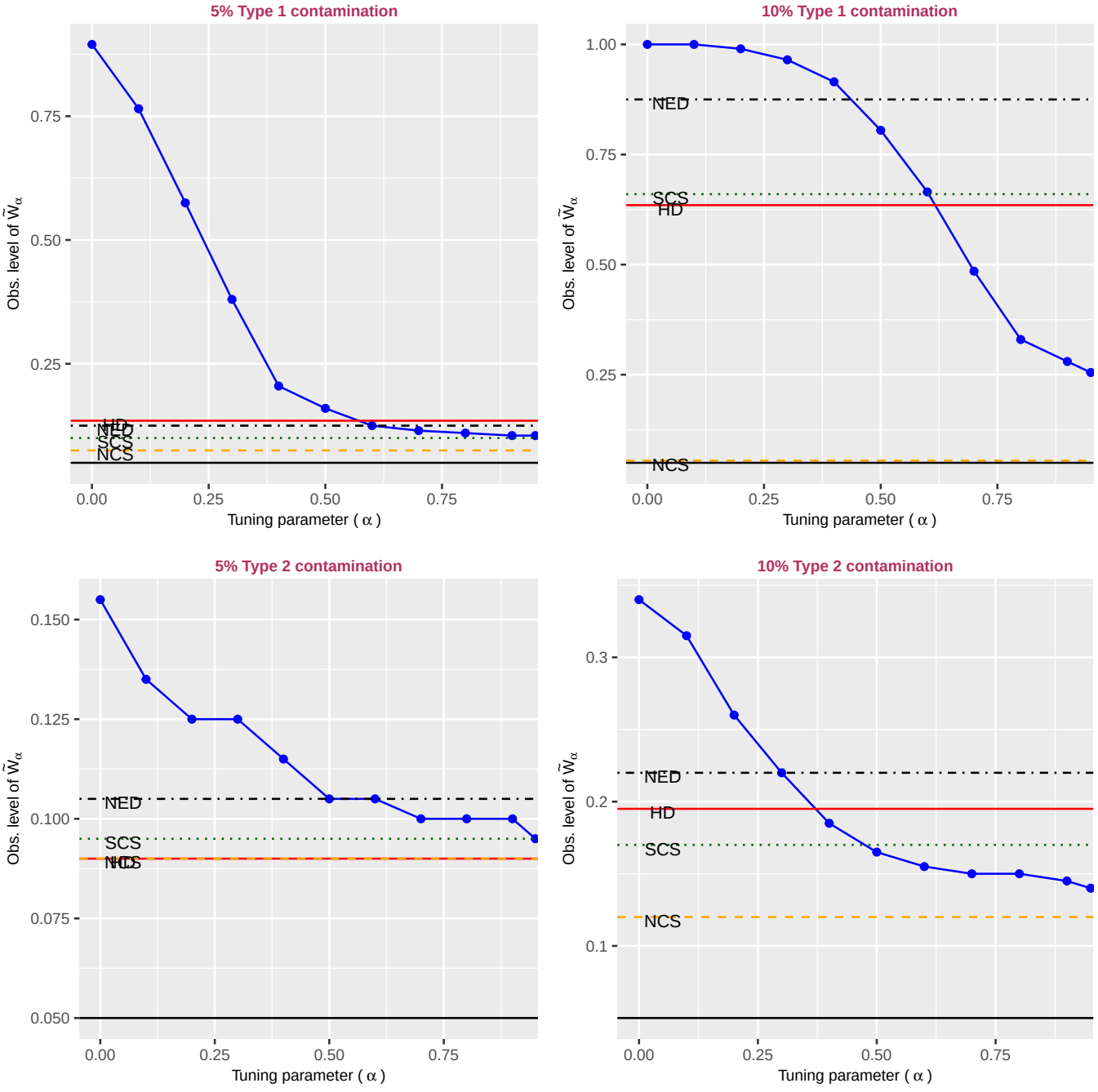

FIGURE 4.7: Obs. levels of the Wald-type test statistic under data contamination

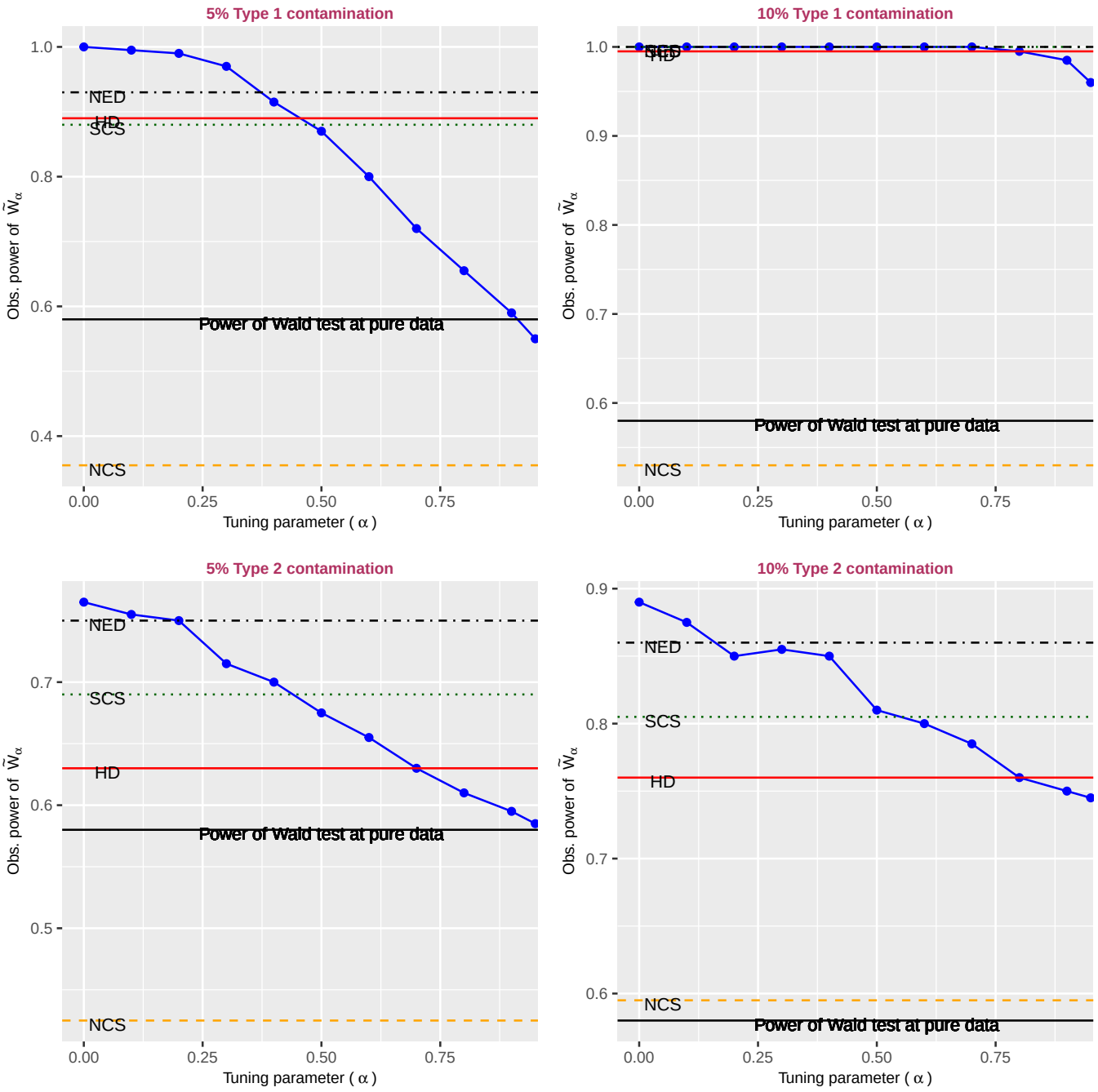

FIGURE 4.8: Obs. powers of the Wald-type test statistic

4.6 Data Driven Selection of Tuning Parameter

The tuning parameter selection criterion will be analogously modified in a two-stage scenario using an estimate of the asymptotic variance of $\tilde{\rho}_\alpha$. For simplicity, we assume that $\mathcal{C} = 0$. Then the criterion as in (1.75) gives

$$\widehat{MSE}_\alpha(\rho_P^*) = (\tilde{\rho}_\alpha - \rho_P^*)^2 + \frac{v^2_{\tilde{\rho}_\alpha,\tilde{\gamma}}}{n}, \tag{4.99}$$

where ρ_P^* is a pilot value. As before we shall try out different MDPD estimators for pilot values. This method is applied to the analyses of a couple of real data examples in Section 4.7. The estimate corresponding to the optimum α will be compared with the minimum Hellinger distance estimator.

4.7 Real Data Examples

First, we analyze the data set of Example 3.2. In Table 4.2 we report optimum tuning parameters corresponding to different pilot values along with the optimum MSE. See that $\alpha = 1$ works best as a pilot, and the optimum α also turns out to be 1. Since it requires a very high value of α to minimize the empirical MSE, the data seems to have a high amount of anomaly compared to the assumed model. The histograms presented in the previous chapter might validate this observation. In Table 4.3 the estimates of polychoric correlations (as the estimates of other parameters are the same across different α) are reported.

Next, we analyze the data of Example 3.3. As earlier we find in Table 4.4 that $\alpha = 1$ works best as a pilot. Consequently, the optimum tuning parameter is similarly obtained as $\alpha = 1$. The Two-step estimate of polychoric correlation is presented in Table 4.5. The estimates of polychoric correlation are very close in the two scenarios.

TABLE 4.2: Optimum α and MSE in Example 3.2 with different pilots ($\tilde{\rho}_\alpha$) for two-step estimates

Tuning parameter α for pilot estimator	Optimal α	Optimal $\widehat{MSE}$
0.0	1	0.1142309
0.1	1	0.07364852
0.2	1	0.04168934
0.3	1	0.02311389
0.4	1	0.01548869
0.5	1	0.01276258
0.6	1	0.01185987
0.7	1	0.01160595
0.8	1	0.01154675
0.9	1	0.01153621
1.0	1	0.01153525

TABLE 4.3: Two-step estimates of the polychoric correlation for Example 3.2

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.83	0.9	1.0	$MHDE$
$\tilde{\rho}_\alpha$	0.575	0.646	0.722	0.788	0.833	0.860	0.878	0.887	0.892	0.893	0.895	0.896	0.857

TABLE 4.4: Optimum α and MSE in Example 3.3 with different pilots ($\tilde{\rho}_\alpha$) for two-step estimates

Tuning parameter α for pilot estimator	Optimal α	Optimal $\widehat{MSE}$
0.0	0.72	0.01797999
0.1	0.75	0.01490853
0.2	0.8	0.01249426
0.3	0.86	0.01062642
0.4	0.92	0.009185058
0.5	0.98	0.008061663
0.6	1	0.007210099
0.7	1	0.006631211
0.8	1	0.006273705
0.9	1	0.006087842
1.0	1	0.006034087

TABLE 4.5: Two-step estimates of the polychoric correlation in Example 3.3 for different methods in two-step scenario

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	$MHDE$
$\tilde{\rho}_\alpha$	0.452	0.469	0.485	0.500	0.513	0.525	0.536	0.545	0.554	0.563	0.570	0.464

4.8 Conclusions

In Chapter 3 we estimate, and develop a Wald-type test for polychoric correlation using the density power divergence. This entails a heavy computational burden. To provide a simpler alternative we propose a two-step adaptation of the usual density power divergence. This provides a useful solution. Two-step estimates of the polychoric correlation turn out to be very close to one-step estimates for pure and mildly contaminated (5% – 10%) data. Moreover, they require much less computational time. Because of using $\tilde{\gamma}$ in the first step of estimation, a bias term appears in the asymptotic normality result. This causes a minor irritation, but this bias term decreases to 0 for a large sample size. Here we have performed the tuning parameter selection assuming $\mathcal{C} = 0$ which may be a little restrictive, but that does not change the overall inference regarding the data.

This page is intentionally left blank.

Chapter 5

Improving Bias and MSE in Two-Step Inference

5.1 A Related Divergence

In our previous work related to DPD (Basu et al., 1998) based estimation of the polychoric correlation, we have noticed that it requires high values of the tuning parameter α to make the MDPDE of polychoric correlation more stable in the presence of 10% outliers. As the contamination proportion increases, the range of stable estimators within the MDPDE class shrinks further. With this background, we try to modify the DPD in such a way that the minimum divergence estimator under the newly proposed divergence may tolerate even more data contamination than the usual DPD.

Now, we introduce a new family of divergence measures based on the density power divergence. We call it the symmetric density power divergence (SDPD) which is defined as

$$D_{\alpha,w}(g,f) = (1-w) \times d_{\alpha}(g,f) + w \times d_{1-\alpha}(f,g) \text{ where } w, \alpha \in [0,1], \tag{5.1}$$

and f, g are, respectively, the model and true probability density functions with respect to a common dominating measure. The values of this divergence at $\alpha = 0$ and $\alpha = 1$ are defined as continuous limits of the divergence $D_{\alpha,w}(g, f)$ by letting $\alpha \downarrow 0+$ and $\alpha \uparrow 1-$ respectively. For notational simplicity, we write $\overline{\alpha} = 1 - \alpha$ and $\overline{w} = 1 - w$, and also refer to $d_{1-\alpha}(f, g)$ as the conjugate of the usual density power divergence $d_\alpha(g, f)$. The divergence defined in (5.1) encompasses the DPD as its special case. It is a super-family that includes the following special cases as mentioned in Table 5.1.

TABLE 5.1: Values of $D_{\alpha,w}(g, f)$

	Weight(w)	
Tuning constant(α)	0	1
0	$LD(g, f)$	$L_2(g, f)$
1	$L_2(g, f)$	$KLD(g, f)$

where KLD is defined in (1.37). We call the new class of divergences symmetric for the following reason:

$$D_{\alpha,w}(g, f) = D_{\overline{\alpha},\overline{w}}(f, g) \text{ for all } w, \alpha \in [0, 1]. \tag{5.2}$$

Similarly, the best-fitting parameter is given by

$$T_{\alpha,w}(G) := \arg\min_{\rho} D_{\alpha,w}(g, \pi(\rho)) \text{ at fixed } w, \alpha \in I, \tag{5.3}$$

where the symbols are already defined in Chapter 3. The minimum SDPD estimator can be routinely obtained via minimization of its empirical version $D_{\alpha,w}(p, \pi(\rho))$ over the parameter space when α, w are fixed. As in the case of simultaneous overall minimization of the DPD to obtain the polychoric correlation, this optimization problem also may pose a considerable computational challenge, even when the number of cut-offs is moderately large. To alleviate this problem, we will adopt a two-step approach where the cut-offs are considered as nuisance parameters because estimating the polychoric

correlation is of primary interest to us. As done previously in Chapter 4, the cut-offs are estimated from the marginal cumulative frequencies, and thereafter ρ will be estimated as a minimizer of the empirical version of the proposed divergence measure that uses the estimates of the cut-offs. A two-stage adaptation of the original problem reduces the computational burden to a great extent. Empirical evidence suggests that estimates coming from both methods are quite similar. Thus the empirical version of the divergence, that we consider, is given by

$$D_{\alpha,w}(p,\pi(\rho,\tilde{\gamma})) \text{ where } \pi(\rho,\tilde{\gamma}) = \Big(\big(\pi_{ij}(\rho,\tilde{\gamma})\big)\Big)_{r\times s}. \tag{5.4}$$

The minimum SDPD estimator of polychoric correlation is therefore given by

$$T^*_{\alpha,w}(P) := \arg\min_{\rho} D_{\alpha,w}(p,\pi(\rho,\tilde{\gamma})) \text{ at fixed } \alpha, w. \tag{5.5}$$

The population analogue of $T^*_{\alpha,w}(P)$ in a two-stage scenario is similarly given by

$$T^*_{\alpha,w}(G) = \arg\min_{\rho} D_{\alpha,w}(g,\pi(\rho,\gamma^*)) \text{ at fixed } \alpha, w. \tag{5.6}$$

The symbols $\tilde{\gamma}$ and γ^* are as already defined in Chapter 4.

5.1.1 A Relation to the Weighted Likelihood Estimating Equation

The best-fitting parameter $\rho^*_{\alpha,w}$ (= $T^*_{\alpha,w}(G)$) solves the estimating equation

$$\frac{\partial}{\partial\rho} D_{\alpha,w}(g,\pi(\rho,\gamma^*)) = \sum_{i,j} W_{ij}\Big(\pi(\rho,\gamma^*),g\Big)\frac{\partial}{\partial\rho}\pi_{ij}(\rho,\gamma^*) = 0, \tag{5.7}$$

where

$$W_{ij}\Big(\pi(\rho,\gamma^*),g\Big) = \overline{w}(1+\alpha)\pi_{ij}^{\alpha-1}(\rho,\gamma^*)(\pi_{ij}(\rho,\gamma^*)-g_{ij}) + w(1+1/\overline{\alpha})(\pi_{ij}^{\overline{\alpha}}(\rho,\gamma^*)-g_{ij}^{\overline{\alpha}}) \text{ for } \alpha \neq 1. \tag{5.8}$$

If $w = 0$ the last term in (5.8) vanishes. When $w > 0$, we apply the first-order Taylor series approximation to the last factor of the second term in (5.8) and get

$$\pi_{ij}^{\overline{\alpha}}(\rho,\gamma^*) - g_{ij}^{\overline{\alpha}} = \overline{\alpha}(\pi_{ij}(\rho,\gamma^*)-g_{ij})g_{ij}^{-\alpha} + o(\underbrace{|\pi_{ij}(\rho,\gamma^*)-g_{ij}|}_{\mathcal{O}(1)}) \tag{5.9}$$

when $g_{ij}^{-\alpha-1}$ is bounded and $g_{ij} > a$ for some $a > 0$ for all i, j. See that (5.9) leads to

$$W_{ij}\Big(\pi(\rho,\gamma^*),g\Big) = \underbrace{\left\{\overline{w}(1+\alpha)\pi_{ij}^{\alpha}(\rho,\gamma^*) + w(1+\overline{\alpha})\cdot\frac{\pi_{ij}(\rho,\gamma^*)}{g_{ij}}g_{ij}^{\overline{\alpha}}\right\}}_{w_{ij}(\pi(\rho,\gamma^*),g)}\underbrace{\left(1-\frac{g_{ij}}{\pi_{ij}(\rho,\gamma^*)}\right)}_{\delta_{ij}^{P}(\pi(\rho,\gamma^*),g)} + o(1) \tag{5.10}$$

when $w > 0$. Notice that w_{ij}s are always non-negative and therefore can be normalised as $w_{ij}(\pi(\rho,\gamma^*),g) \mapsto \frac{w_{ij}(\pi(\rho,\gamma^*),g)}{\sum_{i,j} w_{ij}(\pi(\rho,\gamma^*),g)}$. As the weights are functions of unknown parameters, this normalization would, in fact, affect only the residual term due to the linear approximation in (5.10). Without loss of generality, we may consider $w_{ij}(\pi(\rho,\gamma^*),g)$ as a normalised quantity. The right-hand side of (5.7) may be equivalently expressed as

$$\frac{\partial}{\partial\rho}D_{\alpha,w}(g,\pi(\rho,\gamma^*)) = \begin{cases} \sum_{i,j}(1+\alpha)\Big(\pi_{ij}^{\alpha}(\rho,\gamma^*) - \pi_{ij}^{\alpha-1}(\rho,\gamma^*)g_{ij}\Big)\frac{\partial}{\partial\rho}\pi_{ij}(\rho,\gamma^*) & \text{when } w = 0, \\ \sum_{i,j} w_{ij}\big(\pi(\rho,\gamma^*),g\big)\delta_{ij}^{P}\big(\pi(\rho,\gamma^*),g\big)\frac{\partial}{\partial\rho}\pi_{ij}(\rho,\gamma^*) + o(1) & \text{when } w > 0. \end{cases} \tag{5.11}$$

See that (5.11) (for $w > 0$) excluding the residual term, has a striking resemblance to the estimating equations related to a weighted likelihood function. From this perspective, the minimum SDPD estimator (for $w > 0$) can be interpreted as an approximated weighted MLE ignoring the $o(1)$ term with weights being chosen as $((w_{ij}))$. The estimator $\widetilde{\rho}_{\alpha,w}(= T_{\alpha,w}(P))$ similarly satisfies the estimating equations for ρ,

$$\sum_{i,j}\Big(\pi_{ij}^{\alpha}(\rho,\tilde{\gamma}) - \pi_{ij}^{\alpha-1}(\rho,\tilde{\gamma})p_{ij}\Big)\frac{\partial}{\partial\rho}\pi_{ij}(\rho,\tilde{\gamma}) = 0 \text{ when } w = 0, \tag{5.12}$$

$$\sum_{i,j} w_{ij}\big(\pi(\rho,\tilde{\gamma}),p\big)\delta_{ij}^{P}\big(\pi(\rho,\tilde{\gamma}),p\big)\frac{\partial}{\partial\rho}\pi_{ij}(\rho,\tilde{\gamma}) \approx 0 \text{ when } w > 0. \tag{5.13}$$

We have shown that the minimum SDPD estimator may be approximately considered as a weighted MLE. However, we shall not solve the estimating equation as in (5.12) but obtain our estimator by minimizing $D_{\alpha,w}(p,\pi(\rho,\tilde{\gamma}))$ over ρ.

5.2 Asymptotic Properties

5.2.1 Consistency

In this section, we will prove the consistency of $T^*_{\alpha,w}(P)$. This proof follows the approach of Beran (1977a).

Theorem 5.1. *Suppose that the models are conditionally identifiable in the sense that* $\pi_{ij}(\rho_1,\gamma^*) \neq \pi_{ij}(\rho_2,\gamma^*)$ *for all* i,j *and* $\rho_1 \neq \rho_2$. *Also* $\inf_{\rho\in(-1,1)\setminus H} D_{\alpha,w}(g,\pi(\rho,\gamma^*)) > D_{\alpha,w}(g,\pi(\rho^*,\gamma^*))$ *for some compact set* $H \subset (-1,1)$ *and* $\rho^* \in H$. *Then the following statements are true.*

(C1) *A best-fitting parameter* $T^*_{\alpha,w}(G)$ *exists.*

(C2) *Suppose* $T^*_{\alpha,w}(G)$ *is unique. Also, assume that* $\tilde{\gamma} \xrightarrow{\mathbb{P}} \gamma^*$. *Then* $T^*_{\alpha,w}(P) \xrightarrow{\mathbb{P}} T^*_{\alpha,w}(G)$ *as* $n \to \infty$.

(C3) *$T^*_{\alpha,w}(G)$ is Fisher consistent at the model.*

Proof.

(a) Let us define the map $t \mapsto h(t) := D_{\alpha,w}(g, \pi(t,\gamma^*))$ for fixed density g, and tuning parameters α, w. Since $\pi_{ij}(t,\gamma^*)$s are continuous in t, so is $t \mapsto h(t)$. Hence under the condition $\inf_{\rho \in I \setminus H} D_{\alpha,w}(g, \pi(\rho,\gamma^*)) > D_{\alpha,w}(g, \pi(\rho^*,\gamma^*))$ where H is compact set and $\rho^* \in H$, a minimizer $T^*_{\alpha,w}(G)$ of $h(t)$ exists in $(-1,1)$. Note that $\pi(t,\gamma^*)$ does involve the cut-offs (η_i, β_j)s that are again functions of true distribution G, because

$$G_{is} = \mathbb{P}(X \le i) = \sum_{l=1}^{i} \mathbb{P}(\eta_{l-1} < U \le \eta_l) = \Phi(\eta_i) \implies \eta_i = \Phi^{-1}(G_{is}), \tag{5.14}$$

and similarly $\beta_j = \Phi^{-1}(G_{rj})$ for all i, j.

(b) We define $\tilde{h}_n(t) = D_{\alpha,w}(p, \pi(t,\tilde{\gamma}))$ where $0 \le w \le 1$ and $0 < \alpha < 1$. Recall that $\rho^*_{\alpha,w} = T^*_{\alpha,w}(G)$, and let $\tilde{\rho}_{\alpha,w}$ minimizes $\tilde{h}_n(t)$. First, we will show that $h(\widetilde{\rho}_{\alpha,w}) \xrightarrow{\mathbb{P}} h(\rho^*_{\alpha,w})$ as $n \to \infty$. See that

$$\begin{aligned}
\Big|h(t) - \tilde{h}_n(t)\Big| = \Bigg| \overline{w} \sum_{i,j} \Bigg\{ & \Big(\pi_{ij}^{1+\alpha}(t,\gamma^*) - \pi_{ij}^{1+\alpha}(t,\tilde{\gamma})\Big) + (1+1/\alpha)\Big(\pi_{ij}^{\alpha}(t,\tilde{\gamma}) p_{ij} - \pi_{ij}^{\alpha}(t,\gamma^*) g_{ij}\Big) \\
& + \frac{1}{\alpha}(g_{ij}^{1+\alpha} - p_{ij}^{1+\alpha}) \Bigg\} \\
& + w \sum_{i,j} \Bigg\{ (g_{ij}^{1+\overline{\alpha}} - p_{ij}^{1+\overline{\alpha}}) + (1+1/\overline{\alpha})\Big(p_{ij}^{\overline{\alpha}} \pi_{ij}(t,\tilde{\gamma}) - g_{ij}^{\overline{\alpha}} \pi_{ij}(t,\gamma^*)\Big) \\
& + \frac{1}{\overline{\alpha}}\Big(\pi_{ij}^{1+\overline{\alpha}}(t,\gamma^*) - \pi_{ij}^{1+\overline{\alpha}}(t,\tilde{\gamma})\Big) \Bigg\} \Bigg|.
\end{aligned} \tag{5.15}$$

We know that $p_{ij} \xrightarrow{\mathbb{P}} g_{ij}$ as $n \to \infty$ for all i, j. We have also assumed that $\tilde{\gamma} \xrightarrow{\mathbb{P}} \gamma^*$. Using the continuity of $t \mapsto \pi(t, \cdot)$, one can easily get

$$\left|h(t) - \tilde{h}_n(t)\right| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty \text{ for each } t \in (-1, 1). \tag{5.16}$$

Now, assume that $0 \le \tilde{h}_n(\tilde{\rho}_{\alpha,w}) \le h(\rho^*_{\alpha,w})$. Then

$$0 \le h(\rho^*_{\alpha,w}) - \tilde{h}_n(\tilde{\rho}_{\alpha,w}) \le h(\tilde{\rho}_{\alpha,w}) - \tilde{h}_n(\tilde{\rho}_{\alpha,w}). \tag{5.17}$$

Similarly, when $0 \le h(\rho^*_{\alpha,w}) \le \tilde{h}_n(\tilde{\rho}_{\alpha,w})$ we get

$$0 \le \tilde{h}_n(\tilde{\rho}_{\alpha,w}) - h(\rho^*_{\alpha,w}) \le \tilde{h}_n(\rho^*_{\alpha,w}) - h(\rho^*_{\alpha,w}). \tag{5.18}$$

Combining (5.17) and (5.18), and then applying (5.16) gives

$$0 \le \left|\tilde{h}_n(\tilde{\rho}_{\alpha,w}) - h(\rho^*_{\alpha,w})\right| \le \left|h(\tilde{\rho}_{\alpha,w}) - \tilde{h}_n(\tilde{\rho}_{\alpha,w})\right| + \left|\tilde{h}_n(\rho^*_{\alpha,w}) - h(\rho^*_{\alpha,w})\right| \xrightarrow{\mathbb{P}} 0 \tag{5.19}$$

as $n \to \infty$. Thus, applying the triangle inequality we obtain

$$0 \le \left|h(\tilde{\rho}_{\alpha,w}) - h(\rho^*_{\alpha,w})\right| \le \left|h(\tilde{\rho}_{\alpha,w}) - \tilde{h}_n(\tilde{\rho}_{\alpha,w})\right| + \left|\tilde{h}_n(\tilde{\rho}_{\alpha,w}) - h(\rho^*_{\alpha,w})\right|. \tag{5.20}$$

Applying (5.16) together with (5.19) on the right-hand side of (5.20) implies that $h(\tilde{\rho}_{\alpha,w}) \xrightarrow{\mathbb{P}} h(\rho^*_{\alpha,w})$ when $n \to \infty$. Further, we have assumed that the best-fitting parameter $\rho^*_{\alpha,w}$ is unique. As before, using the continuity of $\rho \mapsto h(\rho)$ and $h(\tilde{\rho}_{\alpha,w}) \xrightarrow{\mathbb{P}} h(\rho^*_{\alpha,w})$ we would immediately get that $\widetilde{\rho}_{\alpha,w} \xrightarrow{\mathbb{P}} \rho^*_{\alpha,w}$ as $n \to \infty$.

(c) Since the model family is conditionally identifiable, we have $\pi_{ij}(\rho_1, \gamma^*) \neq \pi_{ij}(\rho_2, \gamma^*)$ for all $\rho_1 \neq \rho_2$ and i, j. Suppose the true distribution belongs to the model family, i.e., $g_{ij} = \pi_{ij}(\rho^*, \gamma^*)$ for all i, j and some fixed ρ^*. Then the divergence $h(t) =$

$D_{\alpha,w}(\pi(\rho^*,\gamma^*),\pi(t,\gamma^*))$ attains its minimum value zero only at $t=\rho^*$. Therefore, $T^*_{\alpha,w}$ is Fisher consistent.

□

5.3 Simulation Studies

We consider the simulation setup as in Chapter 3. The estimates of the polychoric correlation are reported in Table 5.2. We see that all the estimates are very close, except perhaps, sometimes when both w and α are quite high. The resulting bias and MSE in the estimates of polychoric correlation are respectively plotted in Figure 5.1 and Figure 5.2. As we see, the bias remains reasonably low up to the point $\alpha = 0.6$ when $w \leq 0.5$ as compared to the MDPDE. However, the MSE for other values of α, w never increases in comparison to the MDPD estimates. From this, we can conclude that we can find good estimates of the polychoric correlation for all $w \leq 0.5$ and $\alpha \leq 0.6$ in the pure data. In Figure 5.3 the 95% empirical confidence intervals are plotted. Here, 2.5% and 97.5% quantiles are respectively called the lower and upper empirical confidence intervals. We find that all these confidence limits contain the true value of the polychoric correlation.

Next, we contaminate the pure data with 10% and 15% levels of contamination. The levels of contamination are chosen quite high. Here, we only consider the Type 1 contamination. We find in Table 5.3 and Table 5.4 that when $w > 0$ it adds more stability at fixed α. Therefore we achieve better stability for $w > 0$. This is observed also in Figure 5.4. As for the confidence intervals as plotted in Figure 5.5 and Figure 5.5, we find that as the levels of contamination increase the MDPD-based confidence intervals do not contain the true parameter of the polychoric correlation. However, with $w > 0$, the gap between the confidence interval and true parameter decreases and eventually

contains higher α values. This gives an edge to the SDPD over the DPD in estimating the polychoric correlation at higher levels of contamination.

TABLE 5.2: Estimates of polychoric correlation in pure data

w \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.74765	0.74585	0.74482	0.74424	0.7438	0.74352	0.74315	0.74271	0.74222	0.74170
0.1	0.74750	0.74572	0.74475	0.74428	0.74415	0.74441	0.74545	0.74837	0.75601	0.77684
0.2	0.74732	0.74557	0.74466	0.74432	0.74446	0.74528	0.74745	0.75245	0.76345	0.78841
0.3	0.74709	0.7454	0.74457	0.74437	0.74479	0.74614	0.7492	0.75555	0.76819	0.79444
0.4	0.74682	0.74519	0.74446	0.74443	0.74515	0.74698	0.75076	0.75799	0.77151	0.79817
0.5	0.74646	0.74493	0.74434	0.74451	0.74553	0.74781	0.75215	0.75998	0.77397	0.80073

TABLE 5.3: Estimates of polychoric correlation in 10% Type 1 contaminated data

w \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.32478	0.33596	0.34951	0.36597	0.38600	0.41020	0.43855	0.46975	0.50157	0.53176
0.1	0.32697	0.33903	0.35388	0.37230	0.39530	0.42403	0.45944	0.50276	0.55854	0.64101
0.2	0.32964	0.34274	0.35909	0.37963	0.40555	0.43807	0.47798	0.52643	0.58748	0.66965
0.3	0.33299	0.34733	0.3654	0.3882	0.4169	0.45231	0.49457	0.54439	0.60525	0.68182
0.4	0.33730	0.35316	0.37318	0.39836	0.42953	0.46676	0.50953	0.55850	0.61729	0.68883
0.5	0.34307	0.36078	0.38305	0.41060	0.44365	0.48141	0.52309	0.56992	0.62596	0.69347

TABLE 5.4: Estimates of polychoric correlation in 15% Type 1 contaminated data

w \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.19614	0.19137	0.18713	0.18339	0.18010	0.17723	0.17471	0.17247	0.17034	0.16801
0.1	0.19586	0.19128	0.18759	0.18510	0.18460	0.18773	0.19812	0.22395	0.28499	0.43429
0.2	0.19552	0.19118	0.18811	0.18704	0.18947	0.19832	0.21919	0.26266	0.34831	0.50921
0.3	0.1951	0.19106	0.18874	0.18926	0.19476	0.209	0.23828	0.29278	0.38889	0.53905
0.4	0.19456	0.19090	0.18948	0.19182	0.20054	0.21978	0.25566	0.31692	0.41726	0.55605
0.5	0.19384	0.19068	0.19039	0.19483	0.20690	0.23069	0.27158	0.33675	0.44103	0.56712

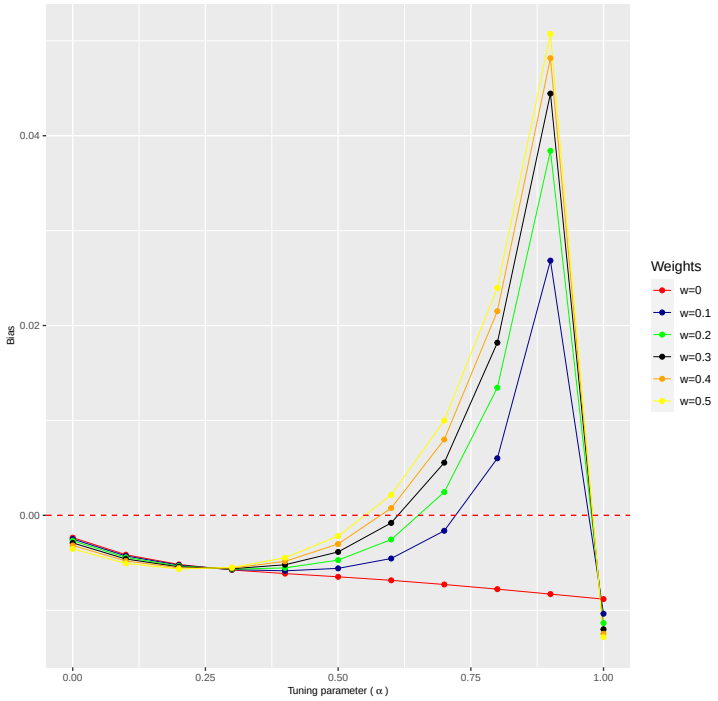

FIGURE 5.1: Bias in the estimates of the polychoric correlation.

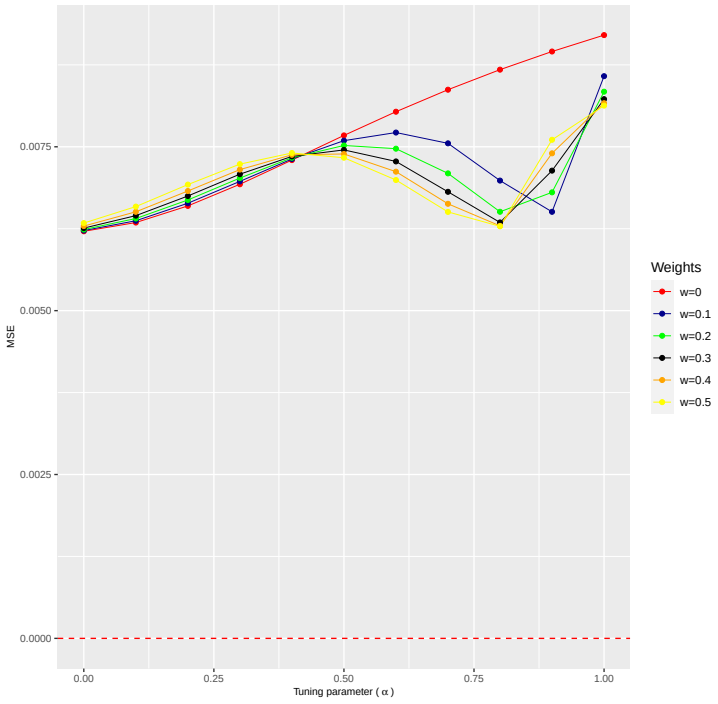

FIGURE 5.2: MSE in the estimates of the polychoric correlation.

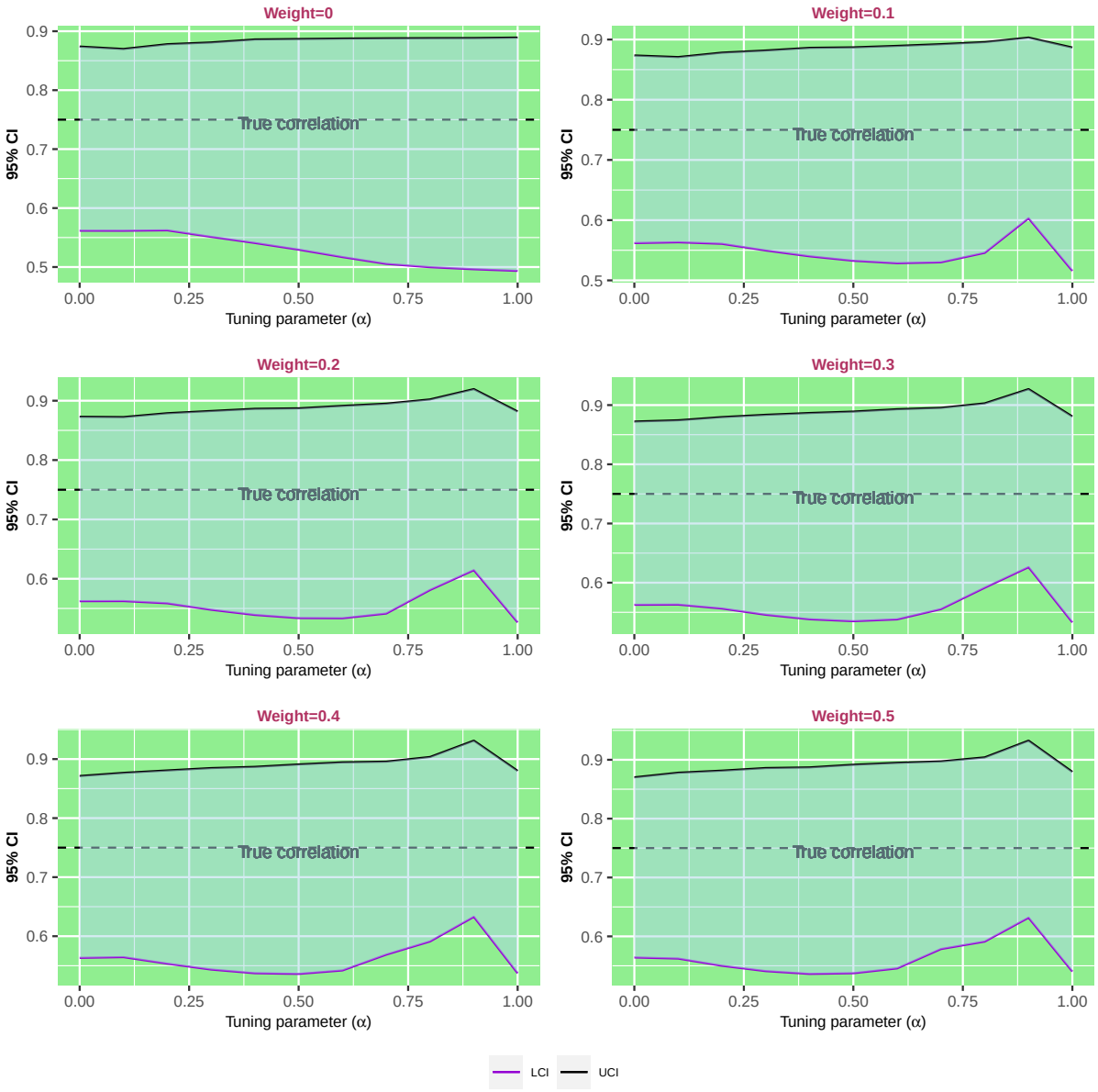

FIGURE 5.3: 95% empirical confidence intervals of the polychoric correlation at pure data.

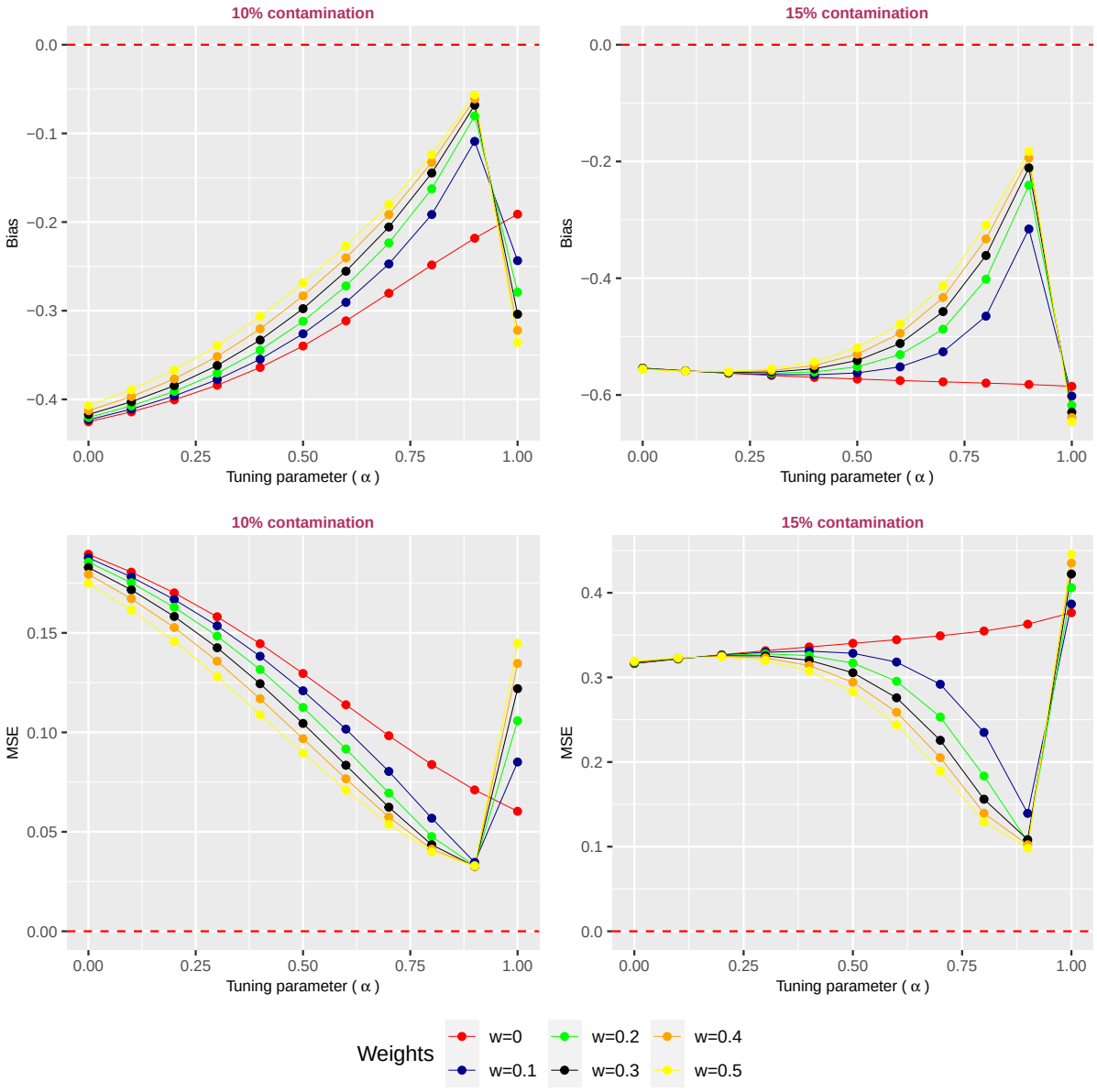

FIGURE 5.4: Bias and MSE in the estimates of the polychoric correlation in Type 1 contaminated data.

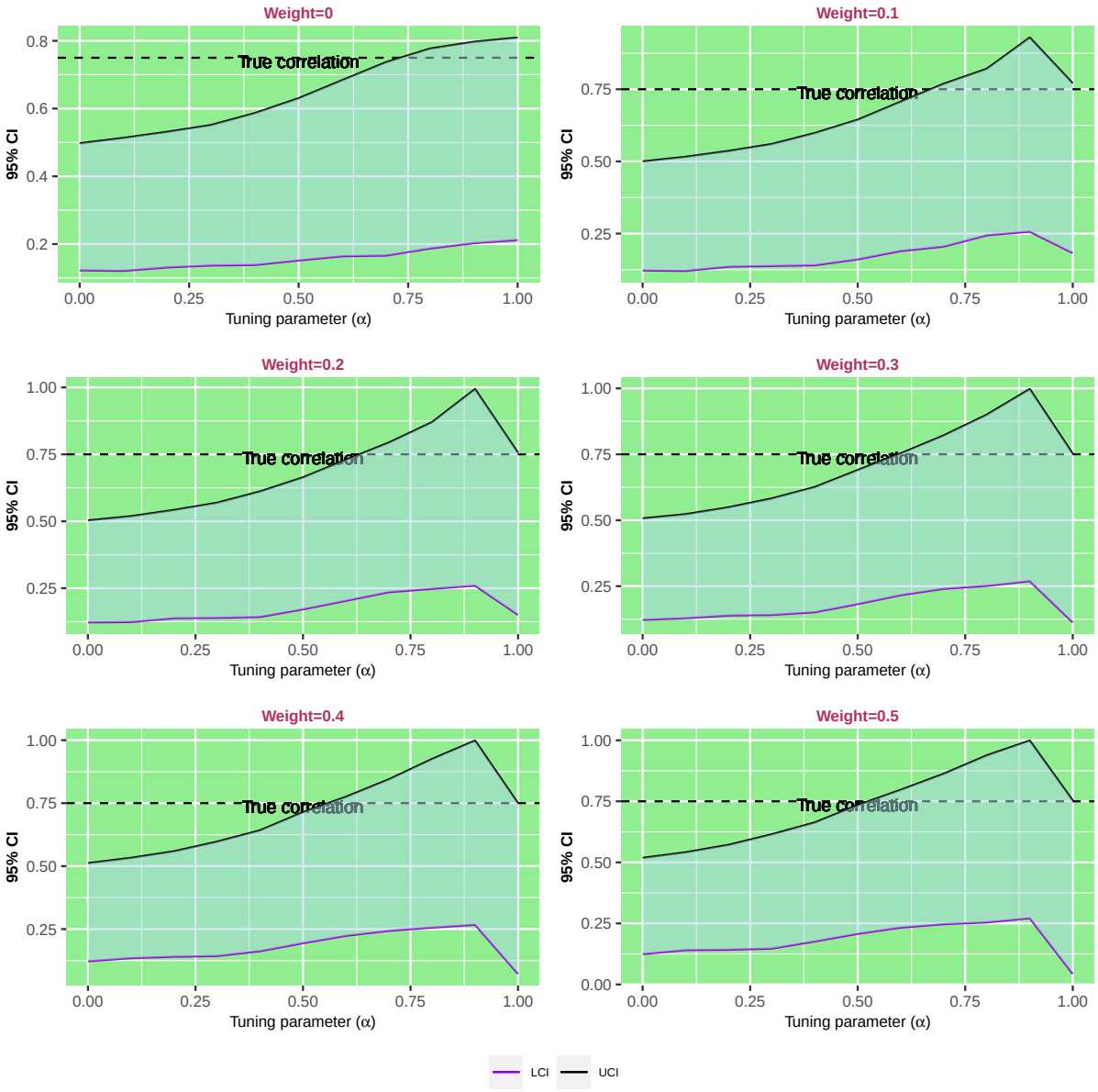

FIGURE 5.5: 95% empirical confidence intervals of the polychoric correlation at 10% Type 1 contaminated data.

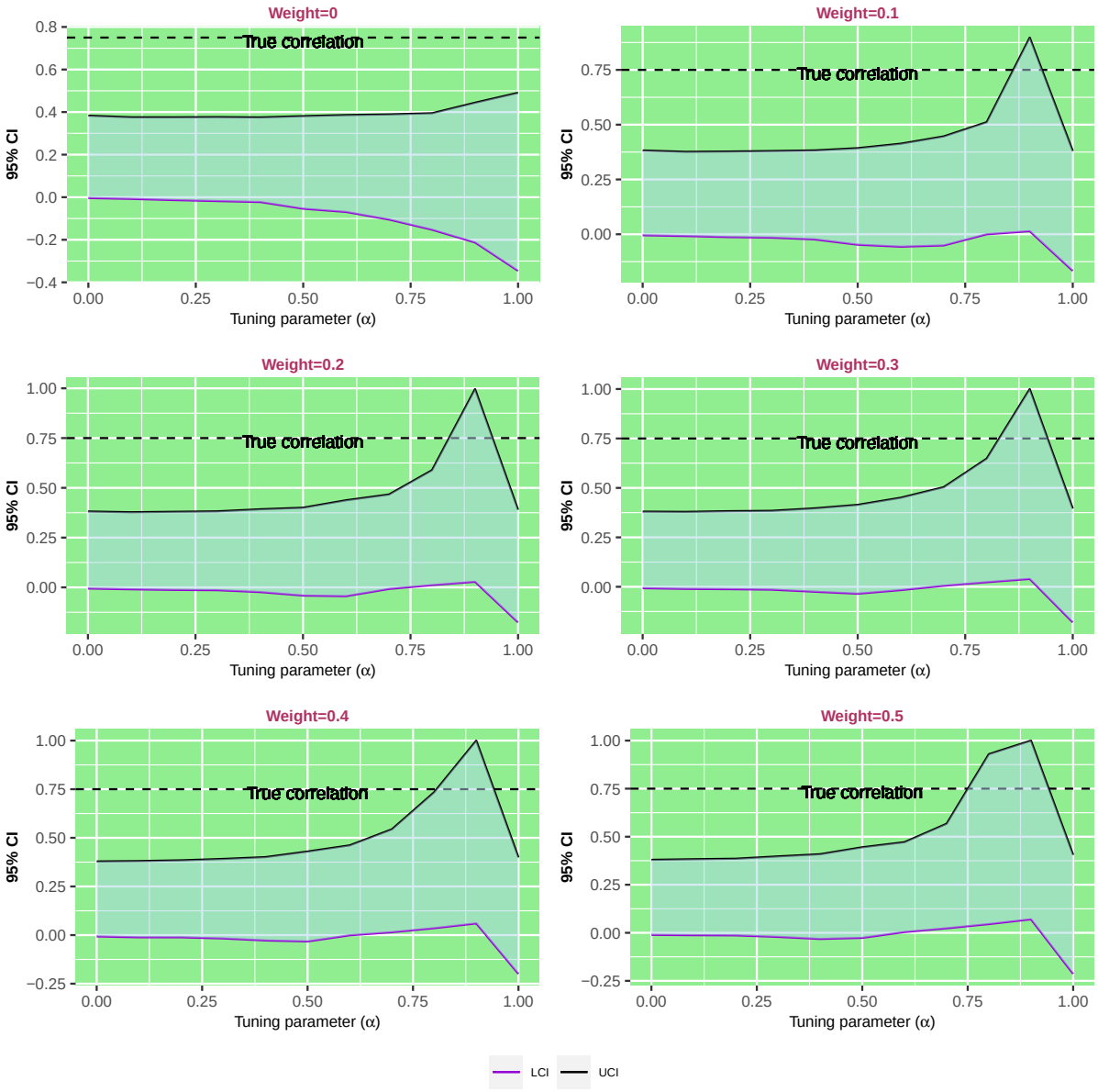

FIGURE 5.6: 95% empirical confidence intervals of the polychoric correlation at 15% Type 1 contaminated data.

5.4 Conclusions

We have seen in Chapter 3 and Chapter 4 it requires higher values of the tuning parameter α to increase the stability of the estimates of the polychoric correlation. Here, we propose a new divergence measure called the SDPD. It combines the DPD and its conjugate version through an added tuning parameter called the weight. The SDPD is, in fact, a superfamily that includes the DPD as its special case. Though it is not apparent, this simple modification makes the estimates more robust. In fact, now, in this case, we are left with more freedom to choose the estimate of the polychoric correlation that is more robust than the usual DPD, yet the loss of efficiency as compared to the DPD in pure data is minimal. Along the way, we prove the consistency of the minimum SDPD estimates when a two-step approach is adapted. The performances of the minimum SDPD estimates of the polychoric correlation are very good up to reasonably high values of the weight parameter. At higher levels of contamination, our proposed estimator performs substantially better than the DPD-based estimates. This is a key highlight of this chapter.

This page is intentionally left blank.

Chapter 6

A Two-Sample Non-parametric Test using the Extended Bregman Divergence: General Theory

6.1 Introduction

Up to this point, we have discussed the application of the density power divergence measure in parameter estimation and hypothesis testing problems related to mixed data. Now, we move on to the next part of this thesis. In this chapter, we develop a class of the two-sample non-parametric tests using different divergence measures. In particular, this chapter focuses on the general class of extended Bregman divergences. Special families of this class will be considered for discussion in the next couple of chapters.

In order to discuss the development of the present theory in the context of the extended Bregman divergence, it is necessary to give a little perspective on how this divergence

measure fits into the discussion of robust inference. We already know that the Cressie-Read family of power divergence (Cressie and Read, 1984) plays an important role in robust inference. Many prominent divergences such as the Hellinger distance (HD), the Likelihood disparity (LD) and many other chi-square type distances belong to the class of power divergence family. Moreover, the power divergence family itself belongs to a more generalized class of statistical distances– namely the ϕ-divergence family (Csiszár, 1964; Ali and Silvey, 1966). In the context of parametric estimation, all members of the ϕ-divergence family yield fully asymptotically efficient estimators at the true model. Moreover, many of them exhibit strong robustness features. Disparity difference tests based on the ϕ-divergence family similarly possess many desirable properties. On a different note, Micheas and Zografos (2006) use the ϕ-divergence to measure stochastic dependency among a finite number of random variables.

Bregman divergence (Bregman, 1967) is another important class of statistical distances which includes the density power divergence (Basu et al., 1998) and Bregman exponential divergence (BED) measure (Mukherjee et al., 2019). Like the ϕ-divergence, the Bregman divergence class produces strongly robust estimators coupled with high asymptotic efficiencies in parametric estimation problems. However, unlike the ϕ-divergence family, the latter class does not require kernel smoothing in constructing the empirical divergence even for the continuous models. This saves substantial additional complexities that the ϕ-divergence family inherits.

The likelihood disparity is the only common member belonging to both these classes. Ghosh et al. (2017) define the S-divergence class that smoothly connects the power divergence family to the density power divergence family; and in this process, generates many other important new divergences. Furthermore, the generalized S-Bregman divergence (Basak and Basu, 2022) unifies both the S-divergence and the BED family with

the introduction of a new tuning parameter in the ordinary Bregman divergence family. Although the Cressie-Read power divergence family does not belong to the class of ordinary Bregman divergences, this extension enables us to view the power divergence as a member of the extended Bregman divergence family.

Inferential methodologies emerging from statistical divergences are primarily limited to parametric models. Extension of them to the non-parametric domain is not rather commonplace in this trade. In the context of non-parametric inference, empirical likelihood plays a pivotal role which gains popularity mainly through the works of Owen (2001). The empirical likelihood has been adopted for the ϕ-divergence family ever since. Contributions of Morales et al. (2001), Balakrishnan et al. (2015; 2017), Jing (1995), Qin and Zhao (2000), Liu et al. (2008), Wu and Yan (2012), Balakrishnan et al. (2017) in and around the two-sample testing problem are worth mentioning here.

In most of these earlier works specific to two-sample problems, the authors consider the null hypothesis of equality in terms of some summary measures (viz., mean, etc.). Guha and Chothia (2014) use mutual information (MI) to test the equality of two completely unstructured absolutely continuous distributions without taking recourse via any particular summary measure. They propose a two-sample non-parametric test that makes use of kernel density estimates. Later on, Guha et al. (2021) robustify the LD-based MI using ϕ-divergences and present simulation studies for the power divergence family. However, we see further that the asymptotic null distribution of the empirical mutual information based on the ϕ-divergence requires the support of the continuous random variables to be bounded. This stringent requirement, in a way, comes as a serious impediment to practitioners who might want to use the asymptotic null distribution to calculate the critical region for a test when the underlying continuous random variables have unbounded supports. To get around this roadblock, Guha et al. (2021) suggest using a permutation test, but it is computationally quite expensive.

In this chapter, we propose a two-sample non-parametric test for the equality between two completely unstructured absolutely continuous distributions using the extended Bregman divergence that partially solves this issue.

Remark 6.1. *When the underlying distributions are continuous, several well-known non-parametric tests are available, many of which are based on the empirical distribution function. Popular examples include the Kolmogorov-Smirnov (KS) test, the Wilcoxon test, the Anderson-Darling (AD) test, and the Cramér-von Mises (CVM) test. A two-sample t-test and Wilcoxon test are commonly used to detect differences in location. The t-test and Wilcoxon test are less effective for distributions that are matched in location but differ in shape, skewness, or kurtosis. While the AD, CVM and KS tests perform reasonably well in these situations, Guha and Chothia (2014) demonstrate that tests based on Mutual Information (MI) often outperform them. Further, Guha et al. (2021) improve the robustness of these tests using the ϕ-divergence, particularly the power divergence family. This chapter introduces the extended Bregman divergence, a general statistical distance measure encompassing the power divergence family, to develop a new class of two-sample non-parametric tests which subsumes the tests of Guha and Chothia (2014). Simulation studies reveal the existence of highly robust tests outside the power divergence family.*

The key takeaways from this chapter are outlined as follows.

(a) We propose a general definition of mutual information based on the extended Bregman divergence. Needless to say, this definition extends the scope of using MI for many other divergences over and above the power divergence family.

(b) A class of consistent non-parametric tests based on the generalized MI is proposed to test the equality between two completely unstructured absolutely continuous distributions.

(c) We establish the asymptotic normality of the test statistic under the null hypothesis and its contiguous alternatives. In the next couple of chapters, we will consider two specific families– namely the generalized S-Bregman divergence (Basak and Basu, 2022) and the Exponential-Polynomial divergence (Singh et al., 2021). For both these families, asymptotic null distribution can be used to find the critical value of the test even if the supports of continuous random variables are unbounded except for the cases when only one tuning parameter α becomes zero. However, this is not the case for the members of the ϕ-divergence family.

(d) The influence function of the generalized MI functional is calculated to study its Infinitesimal stability behaviour. Also, the *level influence function* (LIF) and *power influence function* (PIF) of the test functional are computed along the way.

(e) Under certain regularity conditions, we compute the asymptotic breakdown point of the generalized mutual information functional.

In this chapter, we develop a general theory of two-sample non-parametric tests applicable to all members of the extended Bregman divergence class and save the numerical illustrations against the backdrop of two specific examples in the next chapters. As compared to Guha et al. (2021), derivation of the asymptotic distribution of the test statistic under contiguous alternatives, computations of the influence functions and the breakdown points, and an algorithm for tuning parameter selection in real data examples are new over here.

6.2 A Generalized MI using the Extended Bregman Divergence

6.2.1 The Class of Extended Bregman Divergences

Let $\phi : \mathcal{S} \to \mathbb{R}$ be a strictly convex differentiable function, where $\mathcal{S}$ is a convex subset of $\mathbb{R}^p$. The Bregman divergence (Bregman, 1967) between two points $x, y \in \mathcal{S}$ generated by ϕ is defined as a difference between $\phi(x)$ and its first-order Taylor series approximation at y as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \phi'(y), x - y \rangle, \tag{6.1}$$

where $\phi'(y)$ is the gradient of ϕ evaluated at y, and $\langle s, t \rangle$ denotes an appropriate inner product between s, t. The convexity of ϕ ensures that D_ϕ is non-negative and $D_\phi(x, y) = 0$ if and only if $x = y$. A Bregman divergence is similar to a metric, but in general, it satisfies neither the triangle inequality nor is it symmetric. However, it satisfies a generalization of the Pythagorean theorem.

When points are interpreted as probability density functions such as g and f (both defined on a common support χ with respect to a common σ-finite measure μ), we take the Bregman divergence to be

$$D_\phi(g, f) = \int_\chi \Big\{ \phi(g) - \phi(f) - (g - f)\phi'(f) \Big\} d\mu. \tag{6.2}$$

Later on the symbols χ, μ will be kept implicit as it is clearly understood from the context. The Likelihood disparity and squared L_2 distance are two well-known divergences belonging to this class which are generated respectively by $\phi(t) = t \log t$ and $\phi(t) = t^2$.

An important family in this class is the density power divergence (DPD) (Basu et al., 1998) which is generated by $\phi(t) = \frac{t^{1+\alpha}-t}{\alpha}$ with $\alpha > 0$. Mukherjee et al. (2019) define another important family in this class, known as the Bregman exponential divergence that is given by

$$BED_\beta(g,f) = \frac{2}{\beta}\int \left\{e^{\beta f}\left(f-\frac{1}{\beta}\right) - e^{\beta f}g + \frac{e^{\beta g}}{\beta}\right\} \text{ for } \beta \neq 0. \tag{6.3}$$

This family is generated by $\phi(t) = \frac{2(e^{\beta t}-\beta t-1)}{\beta^2}$ with $\beta \neq 0$. Similarly, the Bregman exponential divergence at $\beta = 0$ is defined as $BED_0(g,f) = \lim_{\beta\to 0} BED_\beta(g,f)$ which turns out to be the squared L_2 distance.

On the other hand, another important class of divergences, the ϕ-divergence is defined as

$$d_\phi(g,f) = \int \phi\left(\frac{g}{f}\right) f. \tag{6.4}$$

Here the convex function ϕ is additionally required to satisfy the following conditions– $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$ and $0\phi(f/0) = f\lim_{u\to\infty}\frac{\phi(u)}{u}$. This class includes the well-known power divergence family which is given by

$$PD_\lambda(g,f) = \frac{1}{\lambda(1+\lambda)}\int g\left\{\left(\frac{g}{f}\right)^\lambda - 1\right\} \text{ for } \lambda \neq -1, 0. \tag{6.5}$$

The divergence at $\lambda = -1, 0$ are similarly defined as continuous limits of $PD_\lambda(g,f)$ as $\lambda \to -1, 0$. Ghosh et al. (2017) define the S-divergence which integrates the power divergence family and the density power divergence in a unified framework. The S-divergence involving two tuning parameters $\alpha \in [0,1]$ and $\lambda \in \mathbb{R}$, is defined as

$$SD_{\alpha,\lambda}(g,f) = \int \left\{\frac{1}{B}(g^{A+B} - f^{A+B}) - (g^A - f^A)\frac{A+B}{AB}f^B\right\} \text{ for } AB \neq 0, \tag{6.6}$$

where $A = 1 + \lambda(1-\alpha)$, $B = \alpha - \lambda(1-\alpha)$. As before the divergence is defined as a continuous limit at $A = 0$ (or, and $B = 0$). When $\lambda = 0$, the S-divergence $SD_{\alpha,\lambda}(g, f)$ becomes the density power divergence $d_\alpha(g, f)$ as in (1.54), and it yields the power divergence $PD_\lambda(g, f)$ as in (6.5) when $\alpha = 0$. An important subfamily of (6.6) is the S-Hellinger distance (SHD) which is given by

$$SHD(g, f) := SD_{\alpha,\lambda=-0.5}(g, f) = \frac{2}{1+\alpha} \int \left(g^{\frac{1+\alpha}{2}} - f^{\frac{1+\alpha}{2}} \right)^2 \text{ for } \alpha \in [0, 1]. \tag{6.7}$$

Notice that the family of S-divergence does not belong to the class of ordinary Bregman divergences. The likelihood disparity is the only common member between them. To unify both these families, Basak and Basu (2022) define a ϕ-generated extended Bregman divergence with a positive index k as

$$D_\phi^{(k)}(g, f) = \int \left\{ \phi(g^k) - \phi(f^k) - (g^k - f^k)\phi'(f^k) \right\}. \tag{6.8}$$

The choice $k = 1$ recovers the ordinary Bregman divergence. In the present discussion, we assume that the convex ϕ is four times continuously differentiable. Notice that S-divergence may be generated from (6.8) by $\phi(t) = \frac{t^{1+\frac{B}{A}}}{B}$ with $k = A$ such that $AB \neq 0$. Furthermore, Basak and Basu (2022) consider another important family by choosing the following convex function

$$\phi(t) = e^{\beta t} + \frac{t^{1+\frac{B}{A}}}{B} \text{ for } A, B \neq 0, \tag{6.9}$$

where A, B are defined as in (6.6) with $\alpha \geq -1$ and $\lambda, \beta \in \mathbb{R}$. The function ϕ in (6.9) together with the exponent $k = A > 0$ generates the divergence

$$D_*(g,f) = \int \left\{ e^{\beta f^A}(\beta f^A - \beta g^A - 1) + e^{\beta g^A} + \frac{1}{B}(g^{A+B} - f^{A+B}) - (g^A - f^A)\frac{A+B}{AB} f^B \right\} \tag{6.10}$$

for $A, B \neq 0$. This divergence at $A = 0$ and (or) $B = 0$ are similarly defined as the continuous limits of the (6.10) when $A, B \to 0$. This family is called the generalized S-Bregman (GSB) divergence, which becomes the S-divergence when $\beta = 0$. A scaled BED family $BED^s_\beta(g,f) = \frac{\beta^2}{2} \times BED_\beta(g,f)$ is generated for $\alpha = -1, \lambda = 0, \beta \neq 0$. Apart from unification, we will see later that the GSB family may be utilized to produce many other consistent and stable tests.

6.2.2 A Generalized Mutual Information and its Properties

Definition 6.1. *($\mathscr{B}$-MI) The mutual information based on a ϕ-generated extended Bregman divergence for a set of random variables $\{X_1, \ldots, X_m\}$ is defined as a map $I_{D^{(k)}_\phi} : \otimes_{i=1}^m \chi_i \longrightarrow [0, \infty]$ such that*

$$I_{D^{(k)}_\phi}(X_1, X_2, \ldots, X_m) = D^{(k)}_\phi(f_{X_1, X_2, \ldots, X_m}, f_{X_1} \cdots f_{X_m}), \tag{6.11}$$

where f denotes the density of the indicated random variable(s), $f_{X_1} \cdots f_{X_m}$ is the product of marginal densities, and χ_i is the sample space of X_i for $i = 1, 2, \ldots, m$.

Given a set of random variables, the expression in (6.11) is uniquely defined up to a set of probability measure zero. The generalized mutual information $I_{D^{(k)}_\phi}$ may be interpreted as a measure of association among a set of random variables. Substituting the function ϕ from (6.9) in (6.11), gives the explicit form of the MI based on the GSB divergence. Additionally when $\alpha = \beta = 0$, $I_{D^{(k)}_\phi}$ becomes I_λ (Guha et al., 2021). Mutual

information based on the DPD may be similarly obtained for $\alpha \geq 0$ with $\lambda = \beta = 0$. The expression of MI for several other divergences may be similarly obtained for different choices of ϕ-function.

An extended Bregman divergence generated through a convex function ϕ will be referred to as the ϕ-generated extended Bregman divergence in this thesis. Let the common members between the ϕ-divergences and ϕ-generated extended Bregman divergences be denoted as $C_{B\phi} = \left\{\phi \text{ for some } k > 0 : d_\phi(g,f) = D_\phi^{(k)}(g,f) \forall f,g\right\}$. Micheas and Zografos (2006) explore a measure of dependence for multivariate data by constructing the ϕ-divergence between the joint density of random variables and the product of its marginals. We shall discuss similar useful properties of the mutual information based on the ϕ-generated extended Bregman divergence. First we define $\phi_0 = \phi(0)$, $\phi_2 = \phi_0 + \lim_{u\to\infty} \frac{\phi(u)}{u}$ and $c_L = \int_{f>0} \frac{k^2 f^{2k} \phi''(f^k)}{\phi''(1)}$ where $f = f_{X_1} f_{X_2}, \ldots, f_{X_m}$. Often for a ϕ-divergence, the convex function ϕ is assumed to satisfy $\phi''(1) = 1$. Some mathematical properties of the generalized mutual information $I_{D_\phi^{(k)}}$ are presented in the following proposition.

Proposition 6.1. *Mutual information based on the ϕ-generated extended Bregman divergence has the following properties.*

(P1) $I_{D_\phi^{(k)}}(X_1, \ldots, X_m) = 0$ *if and only if* $f_{X_1, X_2, \ldots, X_m} = f_{X_1} f_{X_2} \cdots f_{X_m}$ *a.s.*.

(P2) $I_{D_\phi^{(k)}}(X_1, \ldots, X_m) = I_{D_\phi^{(k)}}(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$ *for any permutation* $(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$ *of* $(X_1, X_2, \ldots, X_m)$.

(P3) $I_{D_\phi^{(k)}}(T_1(X_1), \ldots, T_m(X_m)) = I_{D_\phi^{(k)}}(X_1, \ldots, X_m)$ *for any one-one onto transformation* $(X_1, \ldots, X_m) \mapsto (T_1(X_1), T_2(X_2), \ldots, T_m(X_m))$. *Consequently,* $I_{D_\phi^{(k)}}$ *is invariant under strictly increasing transformations.*

(P4) *Let $\phi_2 = \infty$ and the joint distribution is singular to the product of its marginals. Then $I_{D_\phi^{(k)}}(X_1, \ldots, X_m) = \infty$.*

(P5) *Suppose ϕ is strictly convex, twice-continuously differentiable such that $\phi(1) = \phi'(1) = 0$ and $\phi_2, c_L < \infty$. Then $I_{D_\phi^{(k)}}(X_1, \ldots, X_m) = c_L \phi_2$ if and only if the joint distribution is singular to the product of its marginals for all $\phi \notin C_{B\phi}$.*

Proof.

(P1) Suppose the random variables are independent. Then it follows from Definition 6.1 that $I_{D_\phi^{(k)}} = 0$. The converse follows from the strict convexity of ϕ.

(P2) Consider any permutation π of the random vector $(X_1, \ldots, X_m)$ as a map $\pi(X_1, \ldots, X_m) = (X_{i_1}, \ldots, X_{i_m})$. Since $\pi(X_1, \ldots, X_m) = \pi(Y_1, \ldots, Y_m)$ implies that $X_{i_j} = Y_{i_j}$ for $j = 1, 2, \ldots, m$, the map π is one-to-one. Also, π is a onto map. Thus π is bijective, therefore π^{-1} exists. Thus we get

$$\begin{aligned} f_{\pi(X_1,\ldots,X_m)}(\pi(x_1, \ldots, x_m)) &= f_{X_1,\ldots,X_m}(\pi^{-1}\pi(x_1, \ldots, x_m))|\pi| \\ &= f_{X_1,\ldots,X_m}(x_1, \ldots, x_m)|\pi|, \end{aligned} \tag{6.12}$$

where $|\pi|$ denotes the Jacobian of this map, which turns out to be 1. Therefore we have

$$I_{D_\phi^{(k)}}(X_{i_1}, \ldots, X_{i_m}) = I_{D_\phi^{(k)}}(\pi(X_1, \ldots, X_m), \tag{6.13}$$

for any permutation π.

(P3) Since $(X_1, \ldots, X_m) \mapsto (T_1(X_1), \ldots, T_m(X_m))$ is a one-one and onto transformation, it is a permutation. Thus we can use the same argument as before to get the result.

(P4) Choose a finite constant $c\phi'(t) \neq k^2t^{2k-1}\phi''(t)$ for all t. Define

$$R(s,t) = c\phi\Big(\frac{s}{t}\Big)t - \Big\{\phi(s^k) - \phi(t^k) - (s^k - t^k)\phi'(t^k)\Big\} \text{ for } s,t > 0. \tag{6.14}$$

Note that $R(s,s) = 0$. See that

$$\frac{\partial}{\partial s}R(s,t) = c\phi'(s/t) - ks^{k-1}\Big\{\phi'(s^k) - \phi'(t^k)\Big\}, \tag{6.15}$$

$$\frac{\partial^2}{\partial s^2}R(s,t) = c\phi''(s/t)\frac{1}{t} - k(k-1)s^{k-2}\Big\{\phi'(s^k) - \phi'(t^k)\Big\} - (ks^{k-1})^2\phi''(s^k) \neq 0 \text{ at } s \neq t. \tag{6.16}$$

When t is fixed, $\frac{\partial}{\partial s}R(s,t) = 0$ gives a stationary point at $s = t$. See that $R(s,t)$ attains minimum or maximum at $s = t$ according to

$$\Big[\frac{\partial^2}{\partial s^2}R(s,t)\Big]_{s=t} \gtrless 0 \iff c \gtrless \frac{k^2t^{2k-1}\phi''(t^k)}{\phi''(1)}. \tag{6.17}$$

Since $s = t$ is a stationary point, $R(s,t) \gtrless R(t,t) = 0$ for all s,t. Substituting the functions $g(x)$ and $f(x)$ respectively for s,t, we get that $R(g(x),f(x)) \gtrless 0$ when

$$cf(x) \gtrless \frac{k^2f(x)^{2k}\phi''(f(x)^k)}{\phi''(1)} = \tilde{c}_L(x) \text{ (say)} \tag{6.18}$$

for each point x such that $f(x), g(x) > 0$. Since ϕ is twice continuously differentiable and strictly convex, $\phi''(t) > 0$ for all $t > 0$. Therefore $\tilde{c}_L(x) > 0$ on the common support of g, f. Define $c_L := \int_{f>0}\tilde{c}_L(x)dx$. It is easy to see that $c \lesseqgtr c_L$ when $c \lesseqgtr \tilde{c}_L(x)$ for all x such that $f, g > 0$. Further, integrating $R(g(x),f(x))$ over the common support of f, g, we obtain

$$c_1d_\phi(g,f) \leq D_\phi^{(k)}(g,f) \leq c_2d_\phi(g,f) \text{ for all } 0 < c_1 < c_L < c_2. \tag{6.19}$$

Take $g = f_{X_1,X_2,\ldots,X_m}$ and $f = f_{X_1}\cdots f_{X_m}$. Note that c_L may be ∞.

Assuming $\phi_2 = \infty$, Vajda (1972) shows that $d_\phi(g, f) = \infty$ when f, g are singular. Thus (6.19) implies that $D_\phi^{(k)}(g, f) = \infty$.

(P5) Assume $\phi_2, c_L < \infty$ and $\phi \notin C_{B\phi}$. Then $d_\phi(g, f) = \phi_2$ (Vajda, 1972) also $D_\phi^{(k)}(g, f) < \infty$. At fixed finite constants c_1, c_2, choose two convergent sequences $\{c_{1n}\}$ and $\{c_{2n}\}$ such that

$$0 < c_1 \leq c_{1n} < c_{1(n+1)} < \cdots < c_L < \cdots < c_{2(n+1)} < c_{2n} \leq c_2. \tag{6.20}$$

It is easy to see that (6.19) holds for these two sequences. Define a sequence of closed and bounded intervals $C_n := \left[D_\phi^{(k)}/c_{2n}, D_\phi^{(k)}/c_{1n}\right]$ where $d_\phi := d_\phi(g, f)$ and $D_\phi^{(k)} := D_\phi^{(k)}(g, f)$. Clearly

$$d_\phi, \frac{D_\phi^{(k)}}{c_L} \in C_n \text{ and } C_{n+1} \subset C_n \text{ for all } n. \tag{6.21}$$

So $C := \cap_{n \geq 1} C_n$ will be a non-empty singleton set. Since $d_\phi, \frac{D_\phi^{(k)}}{c_L} \in C$, they must be the same. Hence we have $D_\phi^{(k)}(g, f) = c_L d_\phi(g, f)$. In this condition, Vajda (1972) shows that

$$f \perp g \iff d_\phi(g, f) = \phi_2 \iff D_\phi^{(k)}(g, f) = c_L \phi_2. \tag{6.22}$$

Hence the result follows. □

(P1) states that the generalized MI is zero if and only if the random variables are independent. (P2) says that the generalized MI remains unchanged for any permutation of its arguments. We show in (P3) that the generalized mutual information has the invariance property. So it cannot be increased by any one-one onto transformation of $(X_1, \ldots, X_m)$. In (P4) we investigate an upper bound of the generalized mutual information when the joint distribution is singular to the product of its marginals. There exists

an almost sure relationship among $(X_1, \ldots, X_m)$ when $I_{D_\phi^{(k)}}$ attains its upper bound. Micheas and Zografos (2006) show that mutual information based on the ϕ-divergence satisfies many other important properties which also carry over into the power divergence family, i.e., $\phi \in C_{B\phi}$. In (P5) we prove a converse result of (P4) under certain conditions.

Assume that X is a binary $0-1$ variable and Y is a continuous random variable. Then the joint and conditional densities at a point $(x, y) \in \{0, 1\} \times \mathbb{R}$ are defined as

$$f_{x,y}dy = \mathbb{P}\Big\{X = x, y < Y < y + dy\Big\}, \tag{6.23}$$

$$f_{y|x}dy = \mathbb{P}\Big\{y < Y < y + dy | X = x\Big\}. \tag{6.24}$$

The marginal densities of X and Y are, respectively, denoted by $f_x = \mathbb{P}[X = x]$ and $f_y = \sum_{x=0}^{1} f_{x,y}$. For notational convenience, we denote $f_{x_0} = \mathbb{P}[X = 0]$ and $f_{x_1} = \mathbb{P}[X = 1]$. In this hybrid setup the $\mathscr{B}$-MI becomes

$$I_{D_\phi^{(k)}}(X, Y) = \sum_{x=0}^{1} \int_{f_y>0} \Big\{\phi(f_{x,y}^k) - \phi(f_x^k f_y^k) - (f_{x,y}^k - f_x^k f_y^k)\phi'(f_x^k f_y^k)\Big\}dy. \tag{6.25}$$

6.3 A Two-Sample Test based on $\mathscr{B} - MI$

In this section we propose a class of two-sample non-parametric tests based on a ϕ-generated extended Bregman divergence for unstructured comparison of two independent random samples $Y_0 = (Y_{01}, \ldots, Y_{0n_0})$ and $Y_1 = (Y_{11}, Y_{12}, \ldots, Y_{1n_1})$. We assume that Y_0 and Y_1 are respectively drawn from absolutely continuous distribution functions F_0

and F_1. The following hypothesis specifies the unstructured two-sample testing problem

$$\mathbb{H} : F_0 = F_1 \text{ against } \mathbb{K} : F_0 \neq F_1. \tag{6.26}$$

Let f_0, f_1 be the probability density functions corresponding to F_0, F_1. Following Guha and Chothia (2014) and Guha et al. (2021) these two samples Y_0 and Y_1 are combined into $Y := (Y_0, Y_1)$. Further, define a 0-1 binary vector $X = (X_{01}, \ldots, X_{0n_0}, X_{11}, \ldots X_{1n_1})$. The components of X are defined as

$$X_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \in \{Y_{01}, \ldots, Y_{0n_0}\}, \\ 1 & \text{if } Y_{ij} \in \{Y_{11}, \ldots, Y_{1n_1}\} \end{cases} \tag{6.27}$$

for $i = 0, 1$ and $j = 1, \ldots, n_i$. Observe that X is a vector of n_0 zeros followed by n_1 unities. However, the following test will not depend on the values we assign for these two groups. When the null hypothesis $\mathbb{H}$ is true, all the components of the combined sample $Y = (Y_0, Y_1)$ come from the same distribution; hence it will have no bearing on the values that the components of X may take. So the null hypothesis $\mathbb{H}$ is equivalent to that X and Y are independent. In terms of the mutual information $I_{D_\phi^{(k)}}(X, Y)$ as in hybrid setup (6.25), the unstructured testing problem in (6.26) may be restated as

$$\mathbb{H} : f_{x,y} = f_x f_y \text{ for all } (x, y) \iff I_{D_\phi^{(k)}}(X, Y) = 0, \tag{6.28}$$

$$\text{against } \mathbb{K} : f_{x,y} \neq f_x f_y \text{ for at least one } (x, y) \iff I_{D_\phi^{(k)}}(X, Y) > 0. \tag{6.29}$$

To carry out this test, we therefore need to estimate $I_{D_\phi^{(k)}}(X, Y)$. Also, its asymptotic distribution under independence is required to find the critical value of the test statistic. Given a data set the null hypothesis $\mathbb{H}$ is rejected if the estimated MI exceeds the critical

value, and we fail to reject otherwise. For simplicity, we restrict our discussions only to the two-sample problems. However, a generalization for more than two samples can be easily achieved with higher-order kernels because the curse of dimensionality usually haunts any non-parametric density estimation. Later on, for brevity, we shall interchangeably use the notation $I_{D_\phi^{(k)}}$ for $I_{D_\phi^{(k)}}(X, Y)$. In the next subsection, we give a plug-in estimate of $I_{D_\phi^{(k)}}(X, Y)$ based on kernel density estimates.

6.3.1 A Non-parametric Estimate of $\mathscr{B} - MI$ and its Asymptotics

For simplicity, denote $Y = (Y_1, \ldots, Y_{n_0}, Y_{n_0+1}, \ldots, Y_n)$ and $X = (X_1, \ldots, X_{n_0}, X_{n_0+1}, \ldots, X_n)$ where $n = n_0 + n_1$. The generalized mutual information $I_{D_\phi^{(k)}}$ in this hybrid setup may be estimated using the following density estimates at a point $(x, y) \in \{0, 1\} \times \mathbb{R}$ as

$$\widehat{f_x} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{ix}\,, \widehat{f_y} = \frac{1}{nh_n}\sum_{i=1}^{n} K_{h_n i}(y) \text{ and } \widehat{f_{x,y}} = \frac{1}{nh_n}\sum_{i=1}^{n} K_{h_n i}(y)\mathbb{1}_{ix}, \tag{6.30}$$

where $K_{h_n i}(y) = K\left(\frac{Y_i - y}{h_n}\right)$, $\mathbb{1}_{ix} = \mathbb{1}\{X_i = x\}$. Here K denotes a suitable kernel function and $\{h_n\}$ is a bandwidth sequence. The estimated generalized MI is denoted by $\widehat{I}_{D_\phi^{(k)}}$.

Next, we define $Z_{ix} = \mathbb{1}_{ix} - f_x$ that will be used in further discussions. Following Fernandes and Néri (2009) the estimated MI $\widehat{I}_{D_\phi^{(k)}}$ can be expanded up to an error term of appropriate order. To do that, let us make the following assumptions.

(A1) The convex function $\phi(t)$ as in (6.8) is four times differentiable. These derivatives are bounded by integrable functions uniformly for all t.

(A2) The probability density functions $f_{x,y}, f_y$ are continuously twice differentiable with respect to y for $x = 0, 1$. These derivatives and the continuous densities are assumed to be bounded.

(A3) **(i)** The kernel K is symmetric about zero and bounded, i.e.,

$$K(u) = K(-u) \text{ for all } u \in \mathbb{R}, \text{ and } ||K|| := \sup_{u \in \mathbb{R}} |K(u)| < \infty. \tag{6.31}$$

(ii) The kernel K has a bounded support.

(A4) The bandwidth sequence $\{h_n\}$ satisfies the following conditions

$$h_n \longrightarrow 0, nh_n^2 \longrightarrow \infty, \text{ and } nh_n^4 \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{6.32}$$

Assumption (A1) allows a second-order Taylor series expansion of $\widehat{I}_{D_\phi^{(k)}}$ up to a remainder term. Assumption (A2) requires that the joint and marginal densities of Y are smooth enough to admit functional expansions. Assumptions (A1) and (A2) together imply that the remainder term in the expansion of $\widehat{I}_{D_\phi^{(k)}}$ is bounded in probability. Assumption (A3) imposes some conditions on kernels to reduce the bias in the kernel density estimation. Assumption (A4) is a set of technical conditions on the bandwidth sequence that are required to establish the asymptotic normality of $\widehat{I}_{D_\phi^{(k)}}$. Assumption (A3) (i) and (A4) together ensure that the convergences $\widehat{f}_{x,y} \xrightarrow{\mathbb{P}} f_{x,y}$ and $\widehat{f}_y \xrightarrow{\mathbb{P}} f_y$ are uniform in x, y with an appropriate order. Assumption (A3) (ii) restricts the use of a kernel with unbounded support. As the kernel density estimates are generally robust with the choice of kernels and the class of kernels with bounded supports are rich enough, this may not pose a serious problem in practical applications. However Assumption (A3) (ii) may be alleviated with conditions on the finiteness of complicated integrals involving the kernels which we defer for further research. This asymptotic expansion will derive the asymptotic distribution under the null hypothesis.

Lemma 6.1. *Suppose the Assumptions (A1) - (A3) (i) and (A4) are true. Then under the null hypothesis* $\mathbb{H}$*, the estimator* $\widehat{I}_{D_\phi^{(k)}}$ *in hybrid setup admits the following expansion*

$$\widehat{I}_{D_\phi^{(k)}} = \frac{1}{2}\int_{f_y>0}\sum_{x=0}^{1}\left[k^2(f_xf_y)^{2k}\phi''(f_x^kf_y^k)\left(\frac{h_x}{f_x}+\frac{h_y}{f_y}-\frac{h_{x,y}}{f_xf_y}\right)^2\right]dy + o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right), \tag{6.33}$$

where $h_x = \widehat{f}_x - f_x$, $h_y = \widehat{f}_y - f_y$, $h_{x,y} = \widehat{f}_{x,y} - f_{x,y}$ *for all* x, y.

Proof. Define $g_{x,y}(\lambda) = (f_x + \lambda h_x)(f_y + \lambda h_y)$ as a function of $\lambda \in [0,1]$. See that

$$g_{x,y}(0) = f_xf_y, \text{ and } g_{x,y}(1) = \widehat{f}_x\widehat{f}_y. \tag{6.34}$$

Consider the ϕ-generated extended Bregman divergence with the following arguments

$$\Lambda(\lambda) = \int_{f_y>0}\sum_{x=0}^{1}\left\{\phi\Big((f_{x,y}+\lambda h_{x,y})^k\Big) - \phi\Big(g_{x,y}(\lambda)^k\Big) - \Big((f_{x,y}+\lambda h_{x,y})^k - g_{x,y}(\lambda)^k\Big)\phi'\Big(g_{x,y}(\lambda)^k\Big)\right\}dy. \tag{6.35}$$

See that

$$\Lambda(1) = D_\phi^{(k)}(\widehat{f}_{XY}, \widehat{f}_x\widehat{f}_y) = \widehat{I}_{D_\phi^{(k)}}(X,Y), \tag{6.36}$$

$$\Lambda(0) = D_\phi^{(k)}(f_{XY}, f_Xf_Y) = I_{D_\phi^{(k)}}(X,Y). \tag{6.37}$$

Since ϕ is assumed to be four times differentiable, we can expand $\Lambda(1)$ around $\lambda = 0$ up to second-order term with the Lagrange's form of the remainder as

$$\Lambda(1) = \Lambda(0) + \frac{\partial\Lambda(0)}{\partial\lambda} + \frac{1}{2}\cdot\frac{\partial^2\Lambda(0)}{\partial\lambda^2} + \underbrace{\frac{1}{6}\cdot\frac{\partial^3\Lambda(\lambda^*)}{\partial\lambda^3}}_{R_n} \text{ where } \lambda^* \in (0,1). \tag{6.38}$$

It is implicitly assumed that the partial derivatives with respect to λ are interchangeable with integration over y. It is easy to verify that

$$\frac{\partial g_{x,y}(0)}{\partial \lambda} = f_x h_y + f_y h_x, \tag{6.39}$$

$$\frac{\partial^2 g_{x,y}(0)}{\partial \lambda^2} = 2 h_x h_y. \tag{6.40}$$

When $\mathbb{H}$ is true $f_{x,y} = f_x f_y$ for all (x, y). Then we know that $I_{D_\phi^{(k)}} = 0$,

$$\begin{aligned}\frac{\partial \Lambda(\lambda)}{\partial \lambda} = \int_{f_y>0} \sum_{x=0}^{1} \Bigg[& k(f_{x,y} + \lambda h_{x,y})^{k-1} h_{x,y} \phi'\Big((f_{x,y} + \lambda h_{x,y})^k\Big) - k g_{x,y}(\lambda)^{k-1} \frac{\partial g_{x,y}(\lambda)}{\partial \lambda} \phi'\Big(g_{x,y}(\lambda)^k\Big) \\ & - \Big((f_{x,y} + \lambda h_{x,y})^k - (g_{x,y}(\lambda))^k\Big) k g_{x,y}(\lambda)^{k-1} \frac{\partial g_{x,y}(\lambda)}{\partial \lambda} \phi''\Big(g_{x,y}(\lambda)^k\Big) \\ & - \Big(k(f_{x,y} + \lambda h_{x,y})^{k-1} h_{x,y} - k(g_{x,y}(\lambda))^{k-1} \frac{\partial g_{x,y}(\lambda)}{\partial \lambda}\Big) \phi'\Big(g_{x,y}(\lambda)^k\Big) \Bigg] dy = 0 \text{ at } \lambda = 0,\end{aligned} \tag{6.41}$$

$$\begin{aligned}\frac{\partial^2 \Lambda(0)}{\partial \lambda^2} = \int_{f_y>0} \sum_{x=0}^{1} \Bigg[& k(k-1) f_{x,y}^{k-2} h_{x,y}^2 \phi'(f_{x,y}^k) + k^2 f_{x,y}^{2k-2} h_{x,y}^2 \phi''(f_{x,y}^k) \\ & - k(k-1)(f_x f_y)^{k-2} (f_x h_y + f_y h_x)^2 \phi'\Big(f_x^k f_y^k\Big) - k(f_x f_y)^{k-1} (2 h_x h_y) \phi'\Big(f_x^k f_y^k\Big) \\ & - k^2 (f_x f_y)^{2k-2} (f_x h_y + f_y h_x)^2 \phi''\Big(f_x^k f_y^k\Big) \\ & - \Big\{ k f_{x,y}^{k-1} h_{x,y} - k(f_x f_y)^{k-1} (f_x h_y + f_y h_x) \Big\} k (f_x f_y)^{k-1} (f_x h_y + f_y h_x) \phi''\Big(f_x^k f_y^k\Big) \\ & - \Big\{ k f_{x,y}^{k-1} h_{x,y} - k(f_x f_y)^{k-1} (f_x h_y + f_y h_x) \Big\} k (f_x f_y)^{k-1} (f_x h_y + f_y h_x) \phi''\Big(f_x^k f_y^k\Big) \\ & - \Big\{ k(k-1) f_{x,y}^{k-2} h_{x,y}^2 - k(f_x f_y)^{k-1} (2 h_x h_y) - k(k-1)(f_x f_y)^{k-2} (f_x h_y + f_y h_x)^2 \Big\} \phi'\Big(f_x^k f_y^k\Big) \Bigg] dy.\end{aligned} \tag{6.42}$$

Upon algebraic simplification we obtain

$$\frac{\partial^2\Lambda(0)}{\partial\lambda^2} = \int_{f_y>0}\sum_{x=0}^{1} k^2(f_x f_y)^{2k}\phi''(f_x^k f_y^k)\left(\frac{h_x}{f_x}+\frac{h_y}{f_y}-\frac{h_{x,y}}{f_x f_y}\right)^2 dy. \tag{6.43}$$

The expression of the remainder R_n may be obtained from (6.124) with $\Delta_{x,y}, \Delta_x, \Delta_y, \theta, t$ in that expression being replaced by $h_{x,y}, h_x, h_y, \lambda, 1$ respectively. Since f_x is a probability mass function, $|h_x|$ is always bounded in x. When $\sup_y f_y < \infty$, we get

$$|g'_{x,y}(\lambda)| < \sup_y |h_y|(1+2\sup_x |h_x|) + \sup_y(f_y)\times\sup_x |h_x| = \mathcal{O}_{\mathbb{P}}(\sup_y |h_y|),$$
$$|g''_{x,y}(\lambda)| = \mathcal{O}_{\mathbb{P}}(\sup_y |h_y|^2)$$

for all $0 < \lambda < 1$. Later we shall see that $\sup_x |h_x|$, $\sup_y |h_y|$ and $\sup_{x,y} |h_{x,y}|$ are of the same stochastic order. Since $f_{x,y}, f_y$ are assumed to have bounded first derivatives they are Lipschitz functions. Hence they are uniformly continuous. In the view of (6.124) and Assumption (A1), we see that $|R_n|$ can be bounded by the product of $\mathcal{O}_{\mathbb{P}}(\sup_y |h_y|^3)$ and an integral of finite integrand. Hence we get $R_n = \mathcal{O}_{\mathbb{P}}(\sup_y |h_y|^3)$.

From Fernandes and Neŕi (2009) we know that $\sup_y |h_y|^3 = o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right)$ and $\sup_{x,y} |h_{x,y}|^3 = o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right)$. We already know that $h_x = n^{-1}\sum_{i=1}^n Z_{ix}$. Using the Markov's inequality, we get

$$\mathbb{P}\left\{(nh_n^{1/2})^{1/3}|h_x| > \epsilon\right\} \le \frac{1}{\epsilon^4}\mathbb{E}\left\{(nh_n^{1/2})^{1/3}|h_x|\right\}^4 = \frac{1}{\epsilon^4}\frac{h_n^{2/3}}{n^{5/3}}\mathbb{E}(Z_{1x})^4 \longrightarrow 0 \text{ as } n\to\infty \tag{6.44}$$

for fix $\epsilon > 0$. This gives $|h_x|^3 = o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right)$ for all x. Since x takes only $0-1$ values $\sup_x |h_x| = \max\{|h_0|, |h_1|\} = o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right)$. This implies that $R_n = o_{\mathbb{P}}\left(\frac{1}{nh_n^{1/2}}\right)$. Hence the proof is complete. □

It is clear from Lemma 6.1 that the limiting distribution of $\widehat{I}_{D_\phi^{(k)}}$ (after proper normalization) is driven by the first non-degenerate functional derivative of $\widehat{I}_{D_\phi^{(k)}}$. As it turns out, the first two terms in the expansion are singular, and the asymptotic distribution is only determined by its second derivative. In the following discussions, it is assumed that the mathematical expectations can be done under the integral sign.

Lemma 6.2. *Suppose the Assumptions (A1) - (A4) are true. Then* $\widehat{I}_{D_\phi^{(k)}} \xrightarrow{\mathbb{P}} I_{D_\phi^{(k)}}$ *at the true joint density of* (X, Y).

Proof. Recall that the true joint density at a point (x, y) is given by $f_{X,Y}(x, y) \coloneqq f_{x,y}$. As in the proof of Lemma 6.1, we expand $\Lambda(1)$ around $\lambda = 0$ up to first-order term as

$$\Lambda(1) = \Lambda(0) + \Lambda'(0) + \underbrace{\frac{1}{2}\Lambda''(\lambda^*)}_{R_n} \text{ for } \lambda^* \in (0,1) \tag{6.45}$$

with R_n being the remainder term. We know that $\Lambda(1) = \widehat{I}_{D_\phi^{(k)}}$, $\Lambda(0) = I_{D_\phi^{(k)}}$ and

$$\begin{aligned}\frac{\partial}{\partial\lambda}\Lambda(0) = \int_{f_y>0}\sum_{x=0}^{1}\Bigg[&k(f_{x,y})^k\Big\{\phi'\Big(f_{x,y}^k\Big) - \phi'\Big((f_x f_y)^k\Big\}\frac{h_{x,y}}{f_{x,y}} \\ &- \Big(f_{x,y}^k - (f_x f_y)^k\Big)k(f_x f_y)^k\left(\frac{h_x}{f_x} + \frac{h_y}{f_y}\right)\phi''\Big((f_x f_y)^k\Big)\Bigg]dy.\end{aligned} \tag{6.46}$$

Write $\widehat{I}_{D_\phi^{(k)}} - I_{D_\phi^{(k)}} = \frac{1}{n}\sum_{i=1}^n V_i + R_n = \overline{V}_n + R_n$ where

$$\begin{aligned}V_i = \frac{1}{2}\int_{f_y>0}\sum_{x=0}^{1}\Bigg[&k(f_{x,y})^k\Big\{\phi'(f_{x,y}^k) - \phi'\Big((f_x f_y)^k\Big)\Big\}\left(\frac{K_{h_n i}(y)\mathbb{1}_{ix}}{h_n f_{x,y}} - 1\right) \\ &- \Big(f_{x,y}^k - (f_x f_y)^k\Big)k(f_x f_y)^k\phi''\Big((f_x f_y)^k\Big)\left(\frac{\mathbb{1}_{ix}}{f_x} + \frac{K_{h_n i}(y)}{h_n f_y} - 2\right)\Bigg]dy,.\end{aligned} \tag{6.47}$$

See that

$$\begin{aligned}
\mathbb{E}_{f_{X,Y}}\left(\frac{K_{h_n i}(y)\mathbb{1}_{ix}}{h_n f_{x,y}}\right) &= \frac{1}{h_n f_{x,y}} \sum_{t=0}^{1} \int_w K\Big(\frac{w-y}{h_n}\Big)\mathbb{1}(t=x) f_{X,Y}(t,w)dw \\
&= \frac{1}{h_n f_{x,y}} \int_u K(u) f_{X,Y}(x, y+uh_n) h_n du \\
&= \frac{1}{f_{x,y}} \int K(u)\big[f_{x,y} + \mathcal{O}(h_n)\big]du \ \Big(\text{Assumption (A2)}\Big) \\
&= 1 + \mathcal{O}(h_n) \ \Big(\text{Assumption (A3)}\Big). \qquad (6.48)
\end{aligned}$$

Similarly, we have

$$\mathbb{E}_{f_{X,Y}}\left(\frac{K_{h_n i}(y)}{h_n f_y}\right) = 1 + \mathcal{O}(h_n) \qquad (6.49)$$

which yields $\mathbb{E}_{f_{X,Y}}(V_i) = \mathcal{O}(h_n)$. See that V_is are independent because the pairs (Y_i, X_i)s are independent. So the weak laws of large numbers imply that $\overline{V}_n \xrightarrow{\mathbb{P}} 0$. The remainder term R_n involves the terms which are cross products of $h_x, h_y, h_{x,y}$ of order 2, multiplied as factors, with ϕ'', ϕ''' inside an integration. This implies that $R_n = o_{\mathbb{P}}\Big(\frac{1}{(nh_n^{1/2})^{2/3}}\Big)$, and finally

$$\widehat{I}_{D_\phi^{(k)}} = I_{D_\phi^{(k)}} + o_{\mathbb{P}}(1) + o_{\mathbb{P}}\left(\frac{1}{(nh_n^{1/2})^{2/3}}\right) \longrightarrow I_{D_\phi^{(k)}} \text{ as } n \to \infty. \qquad (6.50)$$

This completes the proof. □

Although Assumption (A2) makes derivatives up to second-orders are bounded, restricting them up to first-order is enough to carry out this proof.

Remark 6.2. *Parzen (1962) shows that the kernel density estimates are weakly consistent when $nh_n \to \infty$. See Wied and Weißbach (2012) for more details about stronger versions of consistency results.*

6.3.2 Asymptotic Normality of $\widehat{I}_{D_\phi^{(k)}}$ under Independence

In this subsection, we shall formally state the asymptotic distribution of $\widehat{I}_{D_\phi^{(k)}}$ under the null hypothesis $\mathbb{H}$. To do that we define

$$\mu_\phi = \frac{1}{2}\int_u K^2(u)du \int_{f_y>0}\left[\sum_{x=0}^{1} k^2(f_x f_y)^{2k-1}\phi''(f_x^k f_y^k)(1-f_x)\right]dy, \tag{6.51}$$

$$\sigma_\phi^2 = \frac{1}{2}\int_u \left(\int_z K(z)K(z+u)dz\right)^2 du \int_{f_y>0}\left[\sum_{x=0}^{1} k^2(f_x f_y)^{2k-1}\phi''\left(f_x^k f_y^k\right)(1-f_x)\right]^2 dy. \tag{6.52}$$

In connection with that, we make the following assumption.

(A5) Both μ_ϕ and σ_ϕ are finite. Moreover σ_ϕ should be positive.

Later, we shall see that $\frac{\mu_\phi}{nh_n}$ and $\frac{\sigma_\phi}{nh_n^{1/2}}$ will be used, respectively, as the centring and scaling sequences in normalizing $\widehat{I}_{D_\phi^{(k)}}$ to derive its asymptotic distribution under independence. To see that Assumption (A5) is not superfluous, consider the case when

$$\sum_{x=0}^{1} k^2(f_x f_y)^{2k-1}\phi''\left(f_x^k f_y^k\right)(1-f_x) \tag{6.53}$$

remains constant as a function of y. In addition to that, let us consider that Y has an unbounded support. In this case, Assumptions (A1) and (A3) alone would not imply that both μ_ϕ and σ_ϕ are finite. So Assumption (A5) needs to be separately assumed. Note that (6.53) is satisfied for the power divergence family, therefore it is required that both μ_ϕ and σ_ϕ^2 be finite to hold the asymptotic normality result under the null hypothesis. In this case, Assumption (A5) is equivalent to assuming that the continuous random variables have bounded support, thus restricting the scope of this result in practical applications. However, we can still use that result for the generalized S-Bregman divergence family (See Chapter 7) and the Exponential-Polynomial divergence family (See

Chapter 8) even with unbounded support when $\alpha \neq 0$. Now we start with the following lemma that essentially verifies the conditions of Hall and Heyde (2014) (Lemma 3.1, p.57), and will be further used in the proof of CLT in the hybrid setup. Let us define $T_i = \frac{2}{nh_n^{3/2}} \sum_{j<i} V_{ij}$ where

$$V_{ij} = \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k-2} \phi''(f_x^k f_y^k) Z_{ix} Z_{jx} K_{h_n i}(y) K_{h_n j}(y) dy = V_{ji}. \tag{6.54}$$

Let $Z_n := \prod_{j=1}^{n} (1 + itT_j)$ be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for any $t \in \mathbb{R}$.

Lemma 6.3. *Denote the sum of random variables as $S_n = \sum_{i=1}^{n} T_i$, and $\mathcal{F}_m := \sigma(T_1, \ldots, T_m)$ be an increasing sequence of σ-fields generated by $\{(T_1, \ldots, T_m) : m \leq n\}$. Under the null hypothesis $\mathbb{H}$ and the Assumptions (A1) - (A5), the following statements are true:*

(a) $\sum_{i=1}^{n} T_i^2 \xrightarrow{\mathbb{P}} 4\sigma_\phi^2$,

(b) $\max_{1 \leq i \leq n} |T_i| \xrightarrow{\mathbb{P}} 0$,

(c) $Z_n \longrightarrow 1$ *(weakly in L_1), i.e.,* $\mathbb{E}\Big[Z_n \mathbb{1}(E)\Big] \longrightarrow \mathbb{P}(E)$ *for all $E \in \mathcal{F}$ as $n \to \infty$.*

Proof. The null hypothesis $\mathbb{H}$ is implicitly assumed throughout this proof. We see that $\mathbb{E}(S_n) = 0$ for all n, and $\mathbb{E}\Big(S_n | \mathcal{F}_m\Big) = S_m$ a.s.-$\mathcal{F}_m$ for all $m < n$. So $\{S_n, \mathcal{F}_n : n \geq 1\}$ is a zero-mean, square-integrable martingale and $\{T_i : i \geq 1\}$ being the martingale difference. The conditional variance $v_n^2 = \sum_{i=1}^{n} \mathbb{E}(T_i^2 | \mathcal{F}_{i-1})$ is closely tied to the behaviour of S_n. In the first result, we establish that the conditional variance v_n^2 may be well approximated by the squared variation $u_n^2 := \sum_{i=1}^{n} T_i^2$. The second result implies that the tails of S_n are asymptotically negligible.

(a) First we shall show that $\mathbb{E}\Big(\sum_{i=1}^{n} T_i^2\Big) \to 4\sigma_\phi^2$. See that

$$\sum_{i=1}^{n} \mathbb{E}(T_i^2) = \frac{4}{n^2 h_n^3}\left[\sum_{i=1}^{n}\sum_{j<i} \mathbb{E}(V_{ij}^2) + 2\sum_{i=1}^{n}\sum_{j_1<j_2<i} \mathbb{E}(V_{ij_1}V_{ij_2})\right] = \frac{4}{n^2 h_n^3}\sum_{i=1}^{n}\sum_{j<i} \mathbb{E}(V_{ij}^2), \tag{6.55}$$

as the second term is 0 by independence under $\mathbb{H}$. Therefore

$$\sum_{i=1}^{n} \mathbb{E}(T_i^2) = \frac{4}{n^2 h_n^3}\sum_{i=1}^{n}\sum_{j<i} \mathbb{E}(V_{ij}^2) = \frac{4}{n^2 h_n^3}\binom{n}{2}\mathbb{E}(V_{ij}^2) = \frac{\mathbb{E}(C_n^2)}{n^2 h_n^3} = 4\sigma_\phi^2 + o(1), \tag{6.56}$$

where $C_n = nh_n^{3/2}S_n$ as in Theorem 6.3.2. Next we show that $Var(\sum_{i=1}^{n} T_i^2) \longrightarrow 0$. Observe that

$$Var(\sum_{i=1}^{n} T_i^2) = \sum_{i=1}^{n} Var(T_i^2) + 2\sum_{j<k} Cov(T_j^2, T_k^2), \tag{6.57}$$

where

$$\begin{aligned} Cov(T_j^2, T_k^2) &= \frac{2^4}{n^4 h_n^6} Cov\Big(\sum_{l<j} V_{lj}^2, \sum_{m<k} V_{mk}^2\Big) \\ &= \frac{2^4}{n^4 h_n^6}\left[4\sum_{l_1<l_2<j}\sum_{m_1<m_2<k} Cov(V_{l_1 j}V_{l_2 j}, V_{m_1 k}V_{m_2 k}) + \sum_{l<j}\sum_{m<k} Cov(V_{lj}^2, V_{mk}^2)\right]. \end{aligned} \tag{6.58}$$

From (6.100) we will know that

$$\mathbb{E}V_{ij}^2 = \mathbb{E}\left[\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-2}\phi''(f_x^k f_y^k) Z_{ix}Z_{jx}K_{h_n i}(y)K_{h_n j}(y)dy\right]^2 = \mathcal{O}(h_n^3), \tag{6.59}$$

and

$$\begin{aligned}
&\mathbb{E}(V_{lj}V_{mk})^2 \\
&= \mathbb{E}\Bigg[\sum_x \int C_{x,y} Z_{lx} Z_{jx} K_{h_n l}(y) K_{h_n j}(y)dy \times \sum_x \int C_{x,y} Z_{mx} Z_{kx} K_{h_n m}(y) K_{h_n k}(y)dy\Bigg]^2 \\
&= \mathbb{E}\Bigg[\sum_{x_1,x_2} \int C_{x_1,y_1} C_{x_2,y_2} (Z_{lx_1} Z_{jx_1} Z_{kx_2} Z_{mx_2}) K_{h_n l}(y_1) K_{h_n j}(y_1) K_{h_n m}(y_2) K_{h_n k}(y_2) dy_1 dy_2\Bigg]^2 \\
&= \mathbb{E}\Bigg[\sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4,y_4} (Z_{lx_1} Z_{lx_3})(Z_{jx_1} Z_{jx_3})(Z_{kx_2} Z_{kx_4})(Z_{mx_2} Z_{mx_4}) \\
&\times (K_{h_n l}(y_1) K_{h_n l}(y_3))(K_{h_n j}(y_1) K_{h_n j}(y_3))(K_{h_n m}(y_2) K_{h_n m}(y_4))(K_{h_n k}(y_2) K_{h_n k}(y_4)) dy_1 dy_2 dy_3 dy_4\Bigg] \\
&= \Bigg[\sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4,y_4} \mathbb{E}^2(Z_{lx_1} Z_{lx_3}) \mathbb{E}^2(Z_{kx_2} Z_{kx_4}) \\
&\times \mathbb{E}^2\big(K_{h_n l}(y_1) K_{h_n l}(y_3)\big) \mathbb{E}^2\big(K_{h_n m}(y_2) K_{h_n m}(y_4)\big) dy_1 dy_2 dy_3 dy_4\Bigg], \qquad (6.60)
\end{aligned}$$

where $C_{x,y} = k^2 (f_x f_y)^{2k-2} \phi''(f_x^k f_y^k)$. See that

$$\begin{aligned}
dy_3 \mathbb{E}^2\Big(K_{h_n l}(y_1) K_{h_n l}(y_3)\Big) &= dy_3 \Bigg[\int K\Big(\frac{w - y_1}{h_n}\Big) K\Big(\frac{w - y_3}{h_n}\Big) f_Y(w) dw\Bigg]^2 \\
&= -h_n dz \Bigg[\int K(u) K(u+z) f_Y(y_1 + u h_n) h_n du\Bigg]^2 \\
&= -h_n^3 dz \Bigg[\int K(u) K(u+z) f_Y(y_1 + u h_n) du\Bigg]^2 \\
&= -f_{y_1}^2 h_n^3 dz \Bigg[\int K(u) K(u+z) du\Bigg]^2 + o(h_n^3) \Big\{\text{Assumptions (A2) and (A3)}\Big\}
\end{aligned}$$

(6.61)

by transforming $u = \frac{w-y_1}{h_n}$, $u+z = \frac{w-y_3}{h_n}$, $y_3 = y_1 - zh_n$. Similarly,

$$dy_2\Bigg[\int K\Big(\frac{w-y_2}{h_n}\Big)K\Big(\frac{w-y_4}{h_n}\Big)f_Y(w)dw\Bigg]^2 = -f_{y_4}^2 h_n^3 dz\Bigg[\int K(u)K(u+z)du\Bigg]^2 + o(h_n^3). \tag{6.62}$$

Also, see that

$$\begin{aligned}\mathbb{E}(V_{lj}V_{mk})^2 &= \sum_{x_1,x_2,x_3x_4}\int C_{x_1,y_1}C_{x_2y_4+zh_n}C_{x_3y_1-zh_n}C_{x_4,y_4}\mathbb{E}^2(Z_{lx_1}Z_{lx_3})\mathbb{E}^2(Z_{kx_2}Z_{kx_4})\\ &\times f_{y_1}^2 f_{y_4}^2 h_n^6\Bigg(\int\Big[\int K(u)K(u+z)du\Big]^2dz\Bigg)^2 dy_1dy_4\\ &= h_n^6\sum_{x_1,x_2,x_3x_4}\int C_{x_1,y_1}C_{x_2y_4}C_{x_3y_1}C_{x_4,y_4}f_{y_1}^2f_{y_4}^2dy_1dy_4\\ &\times\mathbb{E}^2(Z_{lx_1}Z_{lx_3})\mathbb{E}^2(Z_{kx_2}Z_{kx_4})\Bigg(\int\Big[\int K(u)K(u+z)du\Big]^2dz\Bigg)^2+\mathcal{O}(h_n^8),\end{aligned} \tag{6.63}$$

by expanding $C_{x_2,y_4+zh_n}, C_{x_3,y_1-zh_n}$ respectively around y_4 and y_1. Thus

$$\begin{aligned}\frac{1}{n^4h_n^6}\sum_{j<k}\sum_{l<j}\sum_{m<k}Cov(V_{lj}^2,V_{mk}^2) &= \frac{1}{n^4h_n^6}\binom{n}{3}Cov(V_{lj}^2,V_{mk}^2)\\ &= \frac{1}{n^4h_n^6}\binom{n}{3}\big[\mathbb{E}(V_{lj}V_{mk})^2-\mathbb{E}^2(V_{lj}^2)\big]\\ &= \frac{1}{nh_n^6}\Big(\mathcal{O}(h_n^6)+\mathcal{O}(h_n^6)\Big)\longrightarrow 0.\end{aligned} \tag{6.64}$$

The order of $Cov(V_{l_1 j}V_{l_2 j}, V_{m_1 k}V_{m_2 k})$ depends on the order of $\mathbb{E}(V_{l_1 j}V_{l_2 j}V_{m_1 k}V_{m_2 k})$ which depends on the following term

$$
\begin{aligned}
&= dy_1 dy_2 dy_3 dy_4 \mathbb{E}\Bigg[K_{h_n l_1}(y_1) K_{h_n j}(y_1) K_{h_n l_2}(y_2) K_{h_n j}(y_2) K_{h_n m_1}(y_3) K_{h_n k}(y_3) K_{h_n m_2}(y_4) K_{h_n k}(y_4) \Bigg] \\
&= dy_1 \mathbb{E}\Bigg[K_{h_n j}(y_1) K_{h_n j}(y_2) \Bigg] \times dy_3 \mathbb{E}\Bigg[K_{h_n k}(y_3) K_{h_n k}(y_4) \Bigg] \\
&\times dy_2 dy_4 \mathbb{E}\Big[K_{h_n l_1}(y_1) \Big] \mathbb{E}\Big[K_{h_n l_2}(y_2) \Big] \mathbb{E}\Big[K_{h_n m_1}(y_3) \Big] \mathbb{E}\Big[K_{h_n m_2}(y_4) \Big] \Bigg] \\
&= h_n^4 f_{y_1} f_{y_3} \Big[\iint K(u) K(u+z) du dz \Big]^2 \times f_{y_1} f_{y_2} f_{y_3} f_{y_4} h_n^4 \\
&= h_n^8 (f_{y_2 + z h_n} f_{y_4 + z h_n})^2 f_{y_2} f_{y_4} \Big[\iint K(u) K(u+z) du dz \Big]^2 dy_2 dy_4 \\
&= h_n^8 (f_{y_2} f_{y_4})^3 \Big[\iint K(u) K(u+z) du dz \Big]^2 dy_2 dy_4 + \mathcal{O}(h_n^{10}). \qquad (6.65)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{2^4}{n^4 h_n^6} \sum_{l_1 < l_2 < j} \sum_{m_1 < m_2 < k} Cov(V_{l_1 j}V_{l_2 j}, V_{m_1 k}V_{m_2 k}) &= \frac{2^4}{n^4 h^6} Cov(V_{l_1 j}V_{l_2 j}, V_{m_1 k}V_{m_2 k}) \sum_{l_1 < l_2 < j} \sum_{m_1 < m_2 < k} 1 \\
&= \frac{2^4}{n^4 h^6} \binom{n}{4} Cov(V_{l_1 j}V_{l_2 j}, V_{m_1 k}V_{m_2 k}) \\
&= \mathcal{O}(h_n^2). \qquad (6.66)
\end{aligned}
$$

So we have $\sum_{j<k} Cov(T_j^2, T_k^2) \to 0$ when $n \to \infty$. For $j < i$, observe that

$$\mathbb{E}(V_{ij}^4) = \mathbb{E}\Bigg[\sum_x \int C_{x,y} Z_{ix} Z_{jx} K_{h_n i}(y) K_{h_n j} dy\Bigg]^4$$

$$= \mathbb{E}\Bigg[\sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4,y_4} (Z_{ix_1} Z_{jx_1})(Z_{ix_2} Z_{jx_2})(Z_{ix_3} Z_{jx_3})(Z_{ix_4} Z_{jx_4})$$

$$\times \big(K_{h_n i}(y_1) K_{h_n j}(y_1)\big)\big(K_{h_n i}(y_2) K_{h_n j}(y_2)\big)\big(K_{h_n i}(y_3) K_{h_n j}(y_3)\big)\big(K_{h_n i}(y_4) K_{h_n j}(y_4)\big)\Bigg] dy_1 dy_2 dy_3 dy_4$$

$$= \sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4,y_4} \mathbb{E}^2\big(Z_{ix_1} Z_{ix_2} Z_{ix_3} Z_{ix_4}\big)$$

$$\times \mathbb{E}^2\big[K_{h_n i}(y_1) K_{h_n i}(y_2) K_{h_n i}(y_3) K_{h_n i}(y_4)\big] dy_1 dy_2 dy_3 dy_4, \tag{6.67}$$

where

$$dy_1 dy_2 dy_3 dy_4 \mathbb{E}^2\big[K_{h_n i}(y_1) K_{h_n i}(y_2) K_{h_n i}(y_3) K_{h_n i}(y_4)\big]$$

$$= dy_1 dy_2 dy_3 dy_4 \Bigg[\int K\Big(\frac{y_1 - w}{h_n}\Big) K\Big(\frac{y_2 - w}{h_n}\Big) K\Big(\frac{y_3 - w}{h_n}\Big) K\Big(\frac{y_4 - w}{h_n}\Big) f_Y(w) dw\Bigg]^2. \tag{6.68}$$

Transforming $u_1 = \frac{y_1 - w}{h_n}, u_2 = \frac{y_2 - w}{h_n}, u_3 = \frac{y_3 - w}{h_n}$ and $u_4 = \frac{y_4 - w}{h_n}$, the above expression simplifies to

$$h_n^3 dy_1 du_2 du_3 du_4 \Bigg[\int K(u_1) K(u_2) K(u_3) K(u_4) f_Y(y_1 - u_1 h_n) h_n du_1\Bigg]^2$$

$$= h_n^5 \Bigg[\int K(u_1) K(u_2) K(u_3) K(u_4) (f_{y_1} + \mathcal{O}(h_n)) du_1\Bigg]^2 du_2 du_3 dy_4 dy_1$$

$$= h_n^5 \Bigg[\int K(u_1) K(u_2) K(u_3) K(u_4) du_1\Bigg]^2 du_2 du_2 dy_3 \times f_{y_1}^2 dy_1 + \mathcal{O}(h_n^6). \tag{6.69}$$

So, we get $\mathbb{E}(V_{ij}^4) = \mathcal{O}(h_n^5)$ for all $j < i$. Similarly when $j_1 < j_2 < i$, we get

$$\begin{aligned}
&\mathbb{E}\big(V_{ij_1}^3 V_{ij_2}\big) \\
&= \mathbb{E}\Bigg[\sum_x \int C_{x,y} Z_{ix} Z_{j_1 x} K_{h_n i}(y) K_{h_n j_1}(y)\Bigg]^3 \Bigg[\sum_x \int C_{x,y} Z_{ix} Z_{j_2 x} K_{h_n i}(y) K_{h_n j_2}(y)\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{x_1,x_2,x_3} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} (Z_{ix_1} Z_{j_1 x_1})(Z_{ix_2} Z_{j_1 x_2})(Z_{ix_3} Z_{j_1 x_3}) \\
&\times \Big(K_{h_n i}(y_1) K_{h_n j_1}(y_1)\Big)\Big(K_{h_n i}(y_2) K_{h_n j_1}(y_2)\Big)\Big(K_{h_n i}(y_3) K_{h_n j_1}(y_3)\Big)\Bigg] \\
&\times \Bigg[\sum_x \int C_{x,y} Z_{ix} Z_{j_2 x} K_{h_n i}(y) K_{h_n j_2}(y)\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4 y_4} (Z_{ix_1} Z_{j_1 x_1})(Z_{ix_2} Z_{j_1 x_2})(Z_{ix_3} Z_{j_1 x_3})(Z_{ix_4} Z_{j_2 x_4}) \\
&\times \Big(K_{h_n i}(y_1) K_{h_n j_1}(y_1)\Big)\Big(K_{h_n i}(y_2) K_{h_n j_1}(y_2)\Big)\Big(K_{h_n i}(y_3) K_{h_n j_1}(y_3)\Big)\Big(K_{h_n i}(y_4) K_{h_n j_2}(y_4)\Big)\Bigg].
\end{aligned} \tag{6.70}$$

This becomes

$$\begin{aligned}
&\sum_{x_1,x_2,x_3,x_4} \int C_{x_1,y_1} C_{x_2,y_2} C_{x_3,y_3} C_{x_4 y_4} \mathbb{E}(Z_{ix_1} Z_{ix_2} Z_{ix_3} Z_{ix_4}) \mathbb{E}(Z_{j_1 x_1} Z_{j_1 x_2} Z_{j_1 x_3}) \underbrace{\mathbb{E}(Z_{j_2 x_4})}_{=0} \\
&\times \mathbb{E}\Big(K_{h_n i}(y_1) K_{h_n i}(y_2) K_{h_n i}(y_3) K_{h_n i}(y_4)\Big) \mathbb{E}\Big(K_{h_n j_1}(y_1) K_{h_n j_1}(y_2) K_{h_n j_1}(y_3)\Big) \mathbb{E}\Big(K_{h_n j_2}(y_4)\Big)\Bigg] = 0.
\end{aligned} \tag{6.71}$$

The remaining cross-product terms will be 0 similarly. This yields

$$\sum_{i=1}^n \mathbb{E}(T_i^4) = \frac{2^4}{n^4 h_n^6} \sum_{i=1}^n \sum_{j<i} \mathbb{E}(V_{ij}^4) = \frac{2^4}{n^4 h_n^6} \times \binom{n}{2} \mathcal{O}(h_n^5) = \mathcal{O}\Big(\frac{1}{n^2 h_n}\Big) = o(1), \tag{6.72}$$

as $nh_n^{1/2} \to \infty$. Therefore, we get $\sum_{i=1}^n Var(T_i^2) \leq \sum_{i=1}^n \mathbb{E}(T_i^4) \to 0$ as $n \to \infty$. Finally applying Chebyshev's inequality, we get

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n T_i^2 - 4\sigma_\phi^2 \right| > \epsilon \right\} \leq \frac{Var\left(\sum_{i=1}^n T_i^2 \right)}{\epsilon^2} \longrightarrow 0 \text{ for any } \epsilon > 0. \tag{6.73}$$

(b) For any $\epsilon > 0$,

$$\mathbb{P}\Big\{ \max_{1\leq i\leq n} |T_i| > \epsilon \Big\} \leq \frac{\mathbb{E}\big(\max_{1\leq i\leq n} |T_i|^4 \big)}{\epsilon^4} \leq \frac{\sum_{i=1}^n \mathbb{E}(T_i^4)}{\epsilon^4} = \mathcal{O}\Big(\frac{1}{\epsilon^4 n^2 h_n} \Big) \longrightarrow 0, \tag{6.74}$$

as $nh_n^{1/2} \to \infty$ for $n \to \infty$.

(c) We already know that $\sum_{i=1}^n T_i^2 \xrightarrow{\mathbb{P}} 4\sigma_\phi^2$ and $\max_{1\leq i\leq n} |T_i| \xrightarrow{\mathbb{P}} 0$. To prove the third part, first, we shall prove that $Z_n = \prod_{j=1}^n (1 + itT_j)$ is uniformly integrable.

Fix any M such that $0 < 4\sigma_\phi^2 < M$, and let $t_n := \sum_{i=1}^n T_i^2$. Then there exists a subsequence $\{t_{k_n} : k_n \geq n\}$ such that $t_{k_n} \xrightarrow{a.s.} 4\sigma_\phi^2 < M$. Define $A_n = \Big\{ \omega : t_n \leq M \Big\}$. Note that $A_n^c \subseteq A_{n+1}^c$ as $t_n \leq t_{n+1}$ for all $n \geq 1$. Notice that

$$\mathbb{P}(A_n^c) \leq \mathbb{P}\Big\{ t_n > M \text{ infinitely often } \Big\} \leq \mathbb{P}\Big\{ t_{k_n} > M \text{ infinitely often } \Big\} = 0$$

for all $n \geq 1$. Using the Cauchy-Schwarz inequality we obtain

$$R_n := \mathbb{E}\left(\prod_{j=1}^n (1 + t^2 T_j^2) \mathbb{1}(A_n^c) \right) \leq \sqrt{ \mathbb{E}\left(\prod_{j=1}^n (1 + t^2 T_j^2)^2 \right) \mathbb{P}(A_n^c) } = 0. \tag{6.75}$$

Now, see that

$$\begin{aligned}
\mathbb{E}|Z_n|^2 &= \mathbb{E}\left(\prod_{j=1}^{n}(1+t^2T_j^2)\mathbb{1}(A_n)\right) + \underbrace{\mathbb{E}\left(\prod_{j=1}^{n}(1+t^2T_j^2)\mathbb{1}(A_n^c)\right)}_{R_n} \\
&\leq \mathbb{E}\Big[\Big\{e^{t^2\sum_{j=1}^{n-1}T_j^2}\Big\}\Big(1+t^2T_n^2\Big)\mathbb{1}(A_n)\Big] \text{ (as } e^x \geq 1+x) \\
&\leq \Big\{e^{t^2M}\Big\}(1+t^2\mathbb{E}T_n^2) \\
&< \Big\{e^{t^2M}\Big\}\Big(1+t^2\sum_{n=1}^{n}\mathbb{E}T_n^2\Big) \\
&= \Big\{e^{t^2M}\Big\}\Big(1+t^24\sigma_\phi^2\Big) + o(1) < \infty \text{ uniformly in } n
\end{aligned} \tag{6.76}$$

by (6.56). So Z_n is uniformly integrable (UI). Next we shall prove that $\mathbb{E}\Big(Z_n\mathbb{1}(E)\Big) \longrightarrow \mathbb{P}(E)$ for any $E \in \mathcal{F}$. Define

$$J_n = \begin{cases} \min\Big\{m \leq n : \sum_{i=1}^{m} T_i^2 > 2M\Big\} & \text{if } \sum_{i=1}^{n} T_i^2 > 2M, \\ n & \text{Otherwise.} \end{cases}$$

See that $J_n \leq n$. Recall that $\mathbb{E}(S_n|\mathcal{F}_m) = S_m$ for all $m \leq n$, so

$$\mathbb{E}(T_j|\mathcal{F}_{j-1}) = \mathbb{E}(S_j - S_{j-1}|\mathcal{F}_{j-1}) \ = \mathbb{E}(S_j|\mathcal{F}_{j-1}) - S_{j-1} = S_{j-1} - S_{j-1} = 0.$$

See that

$$\begin{aligned}
\mathbb{E}\Big[Z_n\mathbb{1}(E)\Big] &= \mathbb{E}\left\{\mathbb{1}(E)\prod_{j=1}^{J_n}(1+itT_j)\prod_{j=J_n+1}^{n}\Big(1+it\mathbb{E}\Big(T_j|\mathcal{F}_{j-1}\Big)\Big)\right\} \\
&= \mathbb{E}\left\{\mathbb{1}(E)\prod_{j=1}^{J_n}(1+itT_j)\right\} \\
&= \mathbb{P}(E) + R_n',
\end{aligned} \tag{6.77}$$

where the remainder term R'_n consists of at most $(2^{J_n} - 1)$ terms of the following form

$$\mathbb{E}\Big[\mathbb{1}(E)(it)^r T_{i_1} T_{i_2} \cdots T_{i_r}\Big] \tag{6.78}$$

such that $1 \le r \le J_n$ and $1 \le i_1 \le i_2 \le \cdots \le i_r \le J_n$. Since

$$\begin{aligned}
&\left|T_{i_1}^2 T_{i_2}^2 \cdots T_{i_{r-1}}^2\right|^{\frac{1}{r-1}} \le \frac{1}{r-1} \sum_{t=1}^{r-1} T_{i_t}^2 \le \sum_{t=1}^{r-1} T_{i_t}^2 \le \sum_{t=1}^{J_n - 1} T_{i_t}^2 \\
\implies &\left|T_{i_1}^2 T_{i_2}^2 \cdots T_{i_{r-1}}^2\right| \le \left(\sum_{t=1}^{J_n-1} T_{i_t}^2\right)^{r-1} \\
\implies &\left|T_{i_1}^2 T_{i_2}^2 \cdots T_{i_r}^2\right| \le \left(\sum_{t=1}^{J_n-1} T_{i_t}^2\right)^{r-1} \left(\max_{1\le i\le n} T_i^2\right) \le (2M)^{r-1} \left(\max_{1\le i\le n} T_i^2\right) \\
\implies &\mathbb{E}\left|T_{i_1} T_{i_2} \cdots T_{i_r}\right| \le (2M)^{\frac{r-1}{2}} \mathbb{E}\left(\max_{1\le i\le n} |T_i|\right),
\end{aligned} \tag{6.79}$$

it follows that

$$|R'_n| \le (2^{J_n} - 1)(2M)^{\frac{J_n}{2}} \mathbb{E}\Big(\max_{1\le i\le n} |T_i|\Big) \tag{6.80}$$

as the remainder term contains at most $(2^{J_n} - 1)$ terms. But, for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{E}\left(\max_{1\le i\le n} |T_i|\right) &\le \epsilon + \mathbb{E}\left[\max_{1\le i\le n} |T_i| \mathbb{1}\left(|T_i| > \epsilon\right)\right] \\
&\le \epsilon + \left[\mathbb{E}\left(\max_{1\le i\le n} T_i^2\right) \mathbb{P}\left(\max_{1\le i\le n} |T_i| > \epsilon\right)\right]^{\frac{1}{2}} \text{ (Cauchy-schwarz inequality)} \\
&\longrightarrow \epsilon \text{ when } n \to \infty,
\end{aligned} \tag{6.81}$$

as (b) holds and $\mathbb{E}(T_i^2)$ is finite for all i. Since $\epsilon > 0$ is arbitrary, $\mathbb{E}\Big(\max_{1\le i\le n} |T_i|\Big) \to 0$. So we get $|R'_n| \le (2^{J_n} - 1)(2M)^{\frac{J_n}{2}} \epsilon$ with $J_n \le n$ which yields $R'_n \to 0$ as $n \to \infty$. Hence we get $\mathbb{E}(Z_n \mathbb{1}(E)) \longrightarrow \mathbb{P}(E)$ for any $E \in \mathcal{F}$, so $Z_n \longrightarrow 1$ weakly in L_1. □

Now we present the asymptotic normality result.

Theorem 6.1. *Suppose the Assumptions (A1) - (A5) are true. Then under the null hypothesis* $\mathbb{H}$*, it holds that*

$$nh_n^{1/2}\Big(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \sigma_\phi^2\Big) \text{ as } n \to \infty. \tag{6.82}$$

Proof. Recall the notations $K_{h_n i}(y)$, Z_{ix} and $\mathbb{1}_{ix}$. Let us define

$$I_n = \frac{1}{2}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k}\phi''(f_x^k f_y^k)\left(\frac{h_x}{f_x} + \frac{h_y}{f_y} - \frac{h_{x,y}}{f_x f_y}\right)^2 dy, \tag{6.83}$$

$$\widetilde{I}_n = \frac{1}{2}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k}\phi''(f_x^k f_y^k)\left(\frac{h_{x,y}}{f_x f_y} - \frac{h_y}{f_y}\right)^2 dy. \tag{6.84}$$

From Lemma 6.1 we know that

$$\widehat{I}_{D_\phi^{(k)}} = I_n + o_{\mathbb{P}}\Big(\frac{1}{nh_n^{1/2}}\Big) \text{ under } \mathbb{H}. \tag{6.85}$$

First we shall show that $nh_n^{1/2}|I_n - \widetilde{I}_n| \xrightarrow{\mathbb{P}} 0$ which implies that I_n and $\widetilde{I}_n$ have the same asymptotic distribution under the null hypothesis $\mathbb{H}$. See that

$$\begin{aligned}
|I_n - \widetilde{I}_n| &\le \frac{1}{2}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k}\Big|\phi''(f_x^k f_y^k)\Big|\left|\left(\frac{h_{x,y}}{f_x f_y} - \frac{h_x}{f_x} - \frac{h_y}{f_y}\right)^2 - \left(\frac{h_{x,y}}{f_x f_y} - \frac{h_y}{f_y}\right)^2\right| dy \\
&\le \sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-1}\Big|\phi''(f_x^k f_y^k)\Big|\frac{|h_x|}{f_x}\Big|2h_{x,y} - h_x f_y - 2h_y f_x\Big| dy \\
&\le \frac{\sup_x |h_x|}{\min_x f_x}\sup_{x,y}\Big|2h_{x,y} - h_x f_y - 2h_y f_x\Big| \times \sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-1}\Big|\phi''(f_x^k f_y^k)\Big| dy \\
&\le \frac{2\sup_x |h_x|}{\min_x f_x} \times \left\{\sup_{x,y}\Big|h_{x,y}\Big| + \sup_x |h_x| \sup_y |f_y| + \sup_y |h_y| \sup_x |f_x|\right\} \\
&\times \sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-1}\Big|\phi''(f_x^k f_y^k)\Big| dy.
\end{aligned} \tag{6.86}$$

We know that $\sup_y |h_y|^3 = o_{\mathbb{P}}\Big(\frac{1}{nh_n^{1/2}}\Big)$ and $\sup_y |h_{x,y}|^3 = o_{\mathbb{P}}\Big(\frac{1}{nh_n^{1/2}}\Big)$ uniformly. Further, by Assumption (A1) , $\sum_{x=0}^{1} k^2(f_x f_y)^{2k-1}\phi''(f_x^k f_y^k)$ is uniformly bounded by integrable function. Also, see that

$$\mathbb{P}\Big\{(nh_n^{1/2})^{2/3}|h_x| \geq \epsilon\Big\} \leq \frac{1}{\epsilon^4}\mathbb{E}\Big((nh_n^{1/2})^{2/3}h_x\Big)^4 = \frac{1}{\epsilon^4}\cdot\frac{h_n^{4/3}}{n^{1/3}}\cdot\mathbb{E}[Z_{1x}]^4 \to 0 \text{ as } n\to\infty, \tag{6.87}$$

for any fixed $\epsilon > 0$. Thus we get $(nh_n^{1/2})^{2/3}|h_x| = o_{\mathbb{P}}(1)$ and $nh_n^{1/2}|I_n - \widetilde{I}_n| \xrightarrow{\mathbb{P}} 0$.

Next, we shall find the asymptotic distribution of $\widetilde{I}_n$. See that

$$\begin{aligned}
\widetilde{I}_n &= \frac{1}{2}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k}\phi''(f_x^k f_y^k)\left[\frac{1}{nh_n}\sum_{i=1}^{n}\left\{\frac{K_{h_n i}(y)1_{ix}}{f_x f_y} - \frac{K_{h_n i}(y)}{f_y}\right\}\right]^2 dy \\
&= \frac{1}{2(nh_n)^2}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-2}\phi''(f_x^k f_y^k)\left[\sum_{i=1}^{n} Z_{ix}^2 K_{h_n i}^2(y) + 2\sum_{i=1}^{n}\sum_{j<i} Z_{ix}Z_{jx}K_{h_n i}(y)K_{h_n j}(y)\right]dy \\
&= \frac{B_n + C_n}{2(nh_n)^2},
\end{aligned} \tag{6.88}$$

where

$$B_n = \sum_{i=1}^{n}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-2}\phi''(f_x^k f_y^k)Z_{ix}^2 K_{h_n i}^2(y)dy, \tag{6.89}$$

$$C_n = 2\sum_{i=1}^{n}\sum_{j<i}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k-2}\phi''(f_x^k f_y^k)Z_{ix}Z_{jx}K_{h_n i}(y)K_{h_n j}(y)dy. \tag{6.90}$$

Notice that under the null hypothesis $\mathbb{E}(Z_{ix}^2) = f_x(1-f_x)$ and

$$\begin{aligned}
\mathbb{E}(K_{h_n i}^2(y)) &= \int_w K^2\Big(\frac{w-y}{h_n}\Big)f_Y(w)dw \\
&= h_n\int_u K^2(u)f_Y(y+uh_n)du \\
&= h_n\Big\{f_y\int_u K^2(u)du + \mathcal{O}(h_n)\Big\} \Big\{ \text{ Assumptions (A2) and (A3) } \Big\} \\
&= h_n f_y\int_u K^2(u)du + \mathcal{O}(h_n^2).
\end{aligned} \tag{6.91}$$

Thus we find that

$$\mathbb{E}(B_n) = nh_n \int_u K^2(u)du \int_{f_y>0} \sum_{x=0}^{1} k^2 (f_x f_y)^{2k-1} \phi''(f_x^k f_y^k)(1-f_x)dy + \mathcal{O}(nh_n^2) \tag{6.92}$$

$$= 2nh_n \mu_\phi + o(nh_n). \tag{6.93}$$

Similarly, under independence,

$$\begin{aligned} Var(B_n) &= \sum_{i=1}^{n} Var\left[\sum_{x=0}^{1} \int_{f_y>0} \underbrace{k^2 (f_x f_y)^{2k-2} \phi''(f_x^k f_y^k)}_{C_{x,y}} Z_{ix}^2 K_{h_n i}^2(y)dy \right] \\ &\le \sum_{i=1}^{n} \mathbb{E}\left[\sum_{x=0}^{1} \int C_{x,y} Z_{ix}^2 K_{h_n i}^2(y)dy \right]^2 \\ &= n\mathbb{E}\left[\sum_{x=0}^{1} \int C_{x,y} Z_{ix}^2 K_{h_n i}^2(y)dy \right]^2 \\ &= n\mathbb{E}\left[\sum_{x_1,x_2} \iint C_{x_1,y_1} C_{x_2,y_2} Z_{ix_1}^2 K_{h_n i}^2(y_1) Z_{ix_2}^2 K_{h_n i}^2(y_2) dy_1 dy_2 \right] \\ &= n \sum_{x_1,x_2} \iint C_{x_1,y_1} C_{x_2,y_2} \mathbb{E}\Big(Z_{ix_1}^2 Z_{ix_2}^2 \Big) \mathbb{E}\Big(K_{h_n i}^2(y_1) K_{h_n i}^2(y_2) \Big) dy_1 dy_2. \end{aligned} \tag{6.94}$$

Also, it holds that

$$\begin{aligned} dy_1 dy_2 \mathbb{E}\Big(K_{h_n i}^2(y_1) K_{h_n y_2}^2 \Big) &= dy_1 dy_2 \int \left[K^2\Big(\frac{w-y_1}{h_n} \Big) K^2\Big(\frac{y_2 - w}{h_n} \Big) f_Y(w)dw \right] \\ &= h_n dz dy_1 \Big[\int K^2(u) K^2(u+z) f_Y(y_1 + uh_n) h_n du \Big] \\ &= h_n^2 (f_{y_1} dy_1) \times \Big(\int K^2(u) K^2(u+z) du \Big) dz + \mathcal{O}(h_n^3) \end{aligned} \tag{6.95}$$

by Assumption (A3). Thus we have

$$\begin{aligned} Var(B_n) \le nh_n^2 \sum_{x_1,x_2} \mathbb{E}(Z_{ix_1}^2 Z_{ix_2}^2) \int C_{x_1,y_1} C_{x_2,y_1+h_n(u+z)} f_{y_1} dy_1 \\ \times \iint K^2(u) K^2(u+z) du dz = \mathcal{O}(nh_n^2). \end{aligned} \tag{6.96}$$

Next see that $\mathbb{E}(C_n) = 0$ and

$$\mathbb{E}(C_n^2) = 4\binom{n}{2}\mathbb{E}\left[\sum_{x=0}^{1}\int k^2(f_x f_y)^{2k-2}\phi''(f_x^k f_y^k)Z_{ix}Z_{jx}K_{h_n i}(y)K_{h_n j}(y)dy\right]^2$$
$$= 4\binom{n}{2}k^4 \iint \left[\sum_{x=0}^{1}(f_x f_{y_1})^{2k-2}(f_x f_{y_2})^{2k-2}\phi''(f_x^k f_{y_1}^k)\phi''(f_x^k f_{y_2}^k)\mathbb{E}^2(Z_{ix}^2)\right.$$
$$\left.+2(f_{x_0} f_{y_1})^{2k-2}(f_{x_1} f_{y_2})^{2k-2}\phi''(f_{x_0}^k f_{y_1}^k)\phi''(f_{x_1}^k f_{y_2}^k)\mathbb{E}^2(Z_{ix_0}Z_{ix_1})\right]\mathbb{E}^2\Big(K_{h_n i}(y_1)K_{h_n i}(y_2)\Big)dy_1 dy_2. \tag{6.97}$$

Observe that $Z_{ix_1} = -Z_{ix_2}$ when $x_1 \neq x_2$, and

$$\mathbb{E}\Big(Z_{ix_1}Z_{ix_2}\Big) = \begin{cases} -f_{x_1}(1-f_{x_1}) & \text{if } x_1 \neq x_2, \\ f_{x_1}(1-f_{x_1}) & \text{if } x_1 = x_2. \end{cases} \tag{6.98}$$

Thus $\mathbb{E}^2\Big[Z_{ix_1}Z_{ix_2}\Big] = f_{x_1}^2(1-f_{x_1})^2$. See that

$$dy_2\mathbb{E}^2\Big(K_{h_n i}(y_1)K_{h_n i}(y_2)\Big) = dy_2\left[\int_w K\Big(\frac{y_1-w}{h_n}\Big)K\Big(\frac{y_2-w}{h_n}\Big)f_Y(w)dw\right]^2. \tag{6.99}$$

Substituting $\frac{y_1-w}{h_n} = z$, $\frac{y_2-w}{h_n} = z+u$ and $y_2 = y_1 + uh_n$ in (6.99), we obtain

$$h_n du\left[h_n\int_w K(z)K(z+u)f_Y(y_1-zh_n)dz\right]^2$$
$$= h_n^3 du\left[\int_w K(z)K(z+u)(f_{y_1}+\mathcal{O}(h_n))dz\right]^2 \left\{\text{ Assumption (A3) }\right\}$$
$$= f_{y_1}^2 h_n^3 du\left[\int_w K(z)K(z+u)dz\right]^2 + o(h_n^3). \tag{6.100}$$

This yields that

$$\mathbb{E}(C_n^2) = 2n^2k^4h_n^3 \int \Bigg[\sum_{x=0}^{1} (f_x f_{y_1})^{2k-2} (f_x f_{y_1+uh_n})^{2k-2} \phi''(f_x^k f_{y_1}^k) \phi''(f_x^k f_{y_1+uh_n}^k) f_x^2 (1-f_x)^2 f_{y_1}^2$$
$$+2(f_{x_0} f_{y_1})^{2k-2} (f_{x_1} f_{y_1+uh_n})^{2k-2} \phi''(f_{x_0}^k f_{y_1}^k) \phi''(f_{x_1}^k f_{y_1+uh_n}^k) f_{x_0}^2 f_{x_1}^2 f_{y_1}^2 \Bigg] dy_1$$
$$\times \int \Big(\int K(z)K(z+u)dz \Big)^2 du + o(n^2h_n^3). \tag{6.101}$$

Since $f_{y_1+uh_n} = f_{y_1} + \mathcal{O}(h_n)$ by Assumption (A2), we can approximate $f_{y_1+uh_n} \approx f_{y_1}$ for sufficiently large n and the remainder term will be $o(n^2h_n^3)$. Using this approximation we get

$$\mathbb{E}(C_n^2) = 2n^2h_n^3 \int \Bigg[\sum_{x=0}^{1} k^2 (f_x f_y)^{2k-1} \phi''\Big(f_x^k f_y^k\Big)(1-f_x) \Bigg]^2 dy \int \Big(\int K(z)K(z+u)dz \Big)^2 du + o(n^2h_n^3)$$
$$= 4n^2h_n^3\sigma_\phi^2 + o(n^2h_n^3). \tag{6.102}$$

Also, see that

$$\frac{1}{(nh_n)^2} |Cov(B_n, C_n)| \leq \frac{1}{(nh_n)^2} \sqrt{Var(B_n) Var(C_n)}$$
$$= \frac{1}{(nh_n)^2} \sqrt{\mathcal{O}(nh_n^2) \times \mathcal{O}(n^2h_n^3)}$$
$$= \mathcal{O}(n^{-1/2}h_n^{1/2}) \to 0 \text{ as } n \to \infty. \tag{6.103}$$

So it follows that

$$\mathbb{E}(\widetilde{I}_n) \approx \frac{\mu_\phi}{nh_n} \text{ and } Var(\widetilde{I}_n) \approx \frac{\sigma_\phi^2}{n^2h_n}. \tag{6.104}$$

We see that the asymptotic distribution of $\widetilde{I}_n$ is determined by C_n. We already know that $S_n = \sum_{i=1}^{n} T_i$ where

$$T_i = \frac{2}{nh_n^{3/2}} \sum_{j<i} \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k-2} \phi''(f_x^k f_y^k) Z_{ix} Z_{jx} K_{h_n i}(y) K_{h_n j}(y) dy. \tag{6.105}$$

Lemma 6.3 verifies the condition of of Hall and Heyde (2014) (Theorem 3.2) which gives

$$\frac{C_n}{nh_n^{3/2}} = S_n = \sum_{i=1}^{n} T_i \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 4\sigma_\phi^2\right) \text{ as } n \to \infty. \tag{6.106}$$

Substituting $C_n = nh_n^{3/2} S_n$ in the expression of $\widetilde{I}_n$, we get

$$nh_n^{1/2}\left(\widetilde{I}_n - \frac{\mu_\phi}{nh_n}\right) \overset{\mathcal{L}}{=} nh_n^{1/2}\left(\widetilde{I}_n - \frac{B_n}{2(nh_n)^2}\right) = \frac{S_n}{2} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_\phi^2\right) \text{ as } n \to \infty. \tag{6.107}$$

This completes the proof. □

The presence of a bias, which grows as the order of $h_n^{-1/2}$ in the hybrid setup, is one of the known problems of the MI estimates. It renders the bias estimation essential to apply Theorem 6.1. The simulation studies presented in Chapter 7 show that the empirical powers are slightly lower in the GSB divergence for $\beta < 0$. This may be because the convergence in distribution is rather slow in these cases. However, this can be improved with larger sample sizes. This result can be easily extended to multi-sample problems with higher-order kernels.

6.3.3 Consistency and Power under Contiguous Alternatives

Suppose the conditions of Theorem 6.1 are satisfied, and μ_ϕ, σ_ϕ are known. Then the null hypothesis $\mathbb{H}$ as in (6.28) is rejected at $100c\%$ nominal level of significance when

$\widehat{I}_{D_\phi^{(k)}}$ exceeds the critical point t_c. Simple calculation shows that $t_c = \frac{\tau_c \sigma_\phi}{nh_n^{1/2}} + \frac{\mu_\phi}{nh_n}$ for sufficiently large n, where τ_c is the upper-100c% point of $\mathcal{N}\Big(0,1\Big)$. Using Corollary 6.2, it can be easily shown that $\mathbb{P}_{\mathbb{K}}\Big\{\widehat{I}_{D_\phi^{(k)}} > \frac{\tau_c \sigma_\phi}{nh_n^{1/2}} + \frac{\mu_\phi}{nh_n}\Big\} \longrightarrow 1$ as $n \to \infty$. So the class of tests is consistent. In practice μ_ϕ, σ_ϕ are generally unknown, which need to be estimated to carry out the testing procedure. However μ_ϕ and σ_ϕ can be estimated using $\widehat{f}_X(x)$ and $\widehat{f}_Y(y)$. The normalized test statistic and its empirical version are given by

$$T_{D_\phi^{(k)}} = \frac{nh_n^{1/2}(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n})}{\sigma_\phi} \text{ and } \widehat{T}_{D_\phi^{(k)}} = \frac{nh_n^{1/2}(\widehat{I}_{D_\phi^{(k)}} - \frac{\widehat{\mu_\phi}}{nh_n})}{\widehat{\sigma}_\phi}. \tag{6.108}$$

In the next result we establish that both $\widehat{T}_{D_\phi^{(k)}}$ and $T_{D_\phi^{(k)}}$ converge to the same distribution under the null hypothesis $\mathbb{H}$, also the class of tests based on $\widehat{T}_{D_\phi^{(k)}}$ is consistent.

Theorem 6.2. *Under the Assumptions (A1) - (A5), the following results are true.*

(i) $\Big|T_{D_\phi^{(k)}} - \widehat{T}_{D_\phi^{(k)}}\Big| \xrightarrow{\mathbb{P}} 0$. *Consequently,* $\widehat{T}_{D_\phi^{(k)}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0,1\Big)$ *under the null hypothesis* $\mathbb{H}$.

(ii) Tests based on $\widehat{T}_{D_\phi^{(k)}}$ *are consistent.*

Proof. (i) As the Kernel density estimates are consistent so are $\widehat{\mu}_\phi, \widehat{\sigma}_\phi$. Define

$$T(\lambda) = \frac{nh_n^{1/2}\Big(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi + \lambda(\hat{\mu}_\phi - \mu_\phi)}{nh_n}\Big)}{\sigma_\phi + \lambda(\widehat{\sigma}_\phi - \sigma_\phi)} \text{ for } \lambda \in [0,1]. \tag{6.109}$$

See that $T(1) = \widehat{T}_{D_\phi^{(k)}}$ and $T(0) = T_{D_\phi^{(k)}}$. Expanding $T(1)$ around $\lambda = 0$ we get

$$T(1) = T(0) + \frac{1}{2}T'(\lambda^{**}) \text{ where } \lambda^{**} \in (0,1), \tag{6.110}$$

where

$$T'(\lambda^{**}) = \frac{\mu_\phi - \hat{\mu}_\phi}{h_n^{1/2}(\sigma_\phi + \lambda^{**}(\widehat{\sigma}_\phi - \sigma_\phi))} - \underbrace{T(\lambda^{**})}_{\mathcal{O}_{\mathbb{P}}(1)} \times \underbrace{\frac{(\widehat{\sigma}_\phi - \sigma_\phi)}{\sigma_\phi + \lambda^{**}(\widehat{\sigma}_\phi - \sigma_\phi)}}_{o_{\mathbb{P}}(1)}. \tag{6.111}$$

Define

$$\mu_\phi(\lambda) = \frac{1}{2}\int_u K^2(u)du \int_{f_y>0} \sum_{x=0}^{1} k^2 \xi_{xy}^{2k-1} \phi''\Big(\xi_{xy}^k\Big)(1 - f_x - \lambda h_x)dy, \tag{6.112}$$

where $\xi_{xy} = (f_x + \lambda h_x)(f_y + \lambda h_y)$. Observe that $\mu_\phi(0) = \mu_\phi$ and $\mu_\phi(1) = \hat{\mu}_\phi$. Expanding $\mu_\phi(1)$ around $\lambda = 0$ up to first-order term we get $\mu_\phi(1) = \mu_\phi(0) + \frac{\mu'_\phi(\lambda^*)}{2}$ where $\lambda^* \in (0,1)$. See that

$$\begin{aligned}
\left|\frac{\mu'_\phi(\lambda^*)}{0.5\int_u K^2(u)du}\right| &= \left| \int \sum_{x=0}^{1} k^2 \xi_{xy}^{2k-1} \Bigg\{ -\phi''(\xi_{xy}^k)h_x \right. \\
&\quad + k\xi^k \phi'''(\xi_{xy}^k)\Big(\frac{h_y}{f_y + \lambda^* h_y} + \frac{h_x}{f_x + \lambda^* h_x}\Big)(1 - f_x - \lambda^* h_x) \\
&\quad \left. + (2k-1)\Big(\frac{h_y}{f_y + \lambda^* h_y} + \frac{h_x}{f_x + \lambda^* h_x}\Big)\phi''(\xi_{xy}^k)(1 - f_x - \lambda^* h_x) \Bigg\} dy \right| \\
&\leq \sup_x |h_x| \int \sum_{x=0}^{1} \left| 3k^2 \xi_{xy}^{2k-1} \left\{ -\frac{\phi''(\xi_{xy}^k)}{3} + \frac{k\xi^k \phi'''(\xi_{xy}^k) + (2k-1)\phi''(\xi_{xy}^k)}{f_x + \lambda^* h_x} \right\} \right| dy \\
&\quad + \sup_y |h_y| \int \sum_{x=0}^{1} \left| 3k^2 \xi_{xy}^{2k-1} \left\{ \frac{k\xi^k \phi'''(\xi_{xy}^k) + (2k-1)\phi''(\xi_{xy}^k)}{f_y + \lambda^* h_y} \right\} \right| dy \\
&\leq \mathcal{O}_{\mathbb{P}}\Big(\sup_x |h_x|\Big) + \mathcal{O}_{\mathbb{P}}\Big(\sup_y |h_y|\Big).
\end{aligned} \tag{6.113}$$

As f_x is trapped between $[0,1]$, the term $|1 - f_x - \lambda^* h_x|$ is bounded by 3, which together with Assumptions (A1) and (A2) imply that the above integrations are bounded. We also know that $\sup_x |h_x|$ and $\sup_y |h_y|$ are $o_{\mathbb{P}}\Big(\frac{1}{(nh_n^{1/2})^{1/3}}\Big)$, so $|\mu'_\phi(\lambda)| =$

$o_{\mathbb{P}}\Big(\frac{1}{(nh_n^{1/2})^{1/3}}\Big)$ uniformly in x, y. This gives $\hat{\mu}_\phi = \mu_\phi + o_{\mathbb{P}}\Big(\frac{1}{(nh_n^{1/2})^{1/3}}\Big)$. Thus we get

$$|\widehat{T}_{D_\phi^{(k)}} - T_{D_\phi^{(k)}}| = \frac{1}{2}|T'(\lambda^{**})| \leq \frac{1}{|\sigma_\phi + \lambda^{**}(\widehat{\sigma}_\phi - \sigma_\phi)|} o_{\mathbb{P}}\Big(\frac{1}{(nh_n^2)^{1/3}}\Big) + o_{\mathbb{P}}(1) \tag{6.114}$$

by Assumption (A4). Hence both $\widehat{T}^{(n)}_{D_\phi^{(k)}}$ and $T_{D_\phi^{(k)}}$ converge to the same limit in distribution. Consequently, $\widehat{T}^{(n)}_{D_\phi^{(k)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ under the null hypothesis $\mathbb{H}$.

(ii) Corollary 6.2 along with the consistency of $\widehat{\mu_\phi}, \widehat{\sigma}_\phi$ imply that

$$(nh_n^{1/2})^{-1}\widehat{T}_{D_\phi^{(k)}} = \frac{\widehat{I}_{D_\phi^{(k)}} - \frac{\widehat{\mu_\phi}}{nh_n}}{\widehat{\sigma}_\phi} \xrightarrow{\mathbb{P}} \frac{I_{D_\phi^{(k)}}}{\sigma_\phi} = \begin{cases} 0 & \text{if null hypothesis is true ,} \\ t & \text{if alternative hypothesis is true} \end{cases} \tag{6.115}$$

for some fixed $t > 0$. Therefore $\widehat{T}_{D_\phi^{(k)}}$ diverges in probability when the alternative hypothesis $\mathbb{K}$ is true. This proves the consistency of $\widehat{T}_{D_\phi^{(k)}}$.

□

When the kernel density estimates are uniformly consistent, so are the MI based on them. Then it is pertinent to investigate the behaviour of its power function when the alternatives are contiguous to the null hypothesis. Consider the following sequence of contiguous or local alternatives

$$\mathbb{K}_n : (X, Y) \sim f^{(n)} \text{ where } f^{(n)}_{x,y} = f_x f_y + \frac{d}{\sqrt{nh_n^{1/2}}}\Delta_{x,y} \text{ where } n = 1, 2, \ldots, \tag{6.116}$$

for some fixed $d \geq 0$. Here $\Delta_{x,y}$ is such that it is independent of the sample size and $\sum_{x=0}^{1} \int_{f_y>0} \Delta_{x,y} dy = 0$. As $f^{(n)}_{x,y}$ is a contiguous sequence to the density under $\mathbb{H}$, so will its marginals with the same sequence of contaminating proportions. In this connection,

the contaminating sequences of the marginals are given by

$$f_x^{(n)} = \int_{f_y>0} f_{x,y}^{(n)} dy = f_x + \frac{d}{\sqrt{nh_n^{1/2}}}\Delta_x \text{ and } f_y^{(n)} = \sum_{x=0}^{1} f_{x,y}^{(n)} = f_y + \frac{d}{\sqrt{nh_n^{1/2}}}\Delta_y, \tag{6.117}$$

where $\Delta_x = \int_{f_y>0} \Delta_{x,y} dy$ and $\Delta_y = \sum_{x=0}^{1} \Delta_{x,y}$. Note that the constant $d \geq 0$ is to be so chosen such that the contiguous sequence remains the probability density. See that the contiguous sequence converges to the density under the null hypothesis at a rate $\frac{1}{\sqrt{nh_n^{1/2}}}$ which is slower than $\frac{1}{nh_n^{1/2}}$. This rate is required to stabilize a bias term which would invariably shift the asymptotic null distribution when $\mathbb{K}_n$ is true. So the choice of such contaminating sequence plays a crucial role in deriving the asymptotic distribution of $T_{D_\phi^{(k)}}$ under $\mathbb{K}_n$. Notice that when $d = 0$, the joint density under $\mathbb{K}_n$ becomes $f_{x,y}^{(n)} = f_x f_y$ for all x, y. This would imply that $\mathbb{K}_n = \mathbb{H}$ in that case. Before starting our next result, it is useful to establish some notation. Let us define

$$I_{D_\phi^{(k)}}^{(n)} = \sum_{x=0}^{1} \int_{f_y>0} \left[\phi\Big((f_{x,y}^{(n)})^k\Big) - \phi\Big((f_x^{(n)} f_y^{(n)})^k\Big) - \Big\{\Big(f_{x,y}^{(n)}\Big)^k - \Big(f_x^{(n)} f_y^{(n)}\Big)^k\Big\} \phi'\Big((f_x^{(n)} f_y^{(n)})^k\Big) \right] dy. \tag{6.118}$$

Next, we shall approximate $I_{D_\phi^{(k)}}^{(n)}$ up to a second-order term that will be further useful to derive the asymptotic distribution of $T_{D_\phi^{(k)}}$ under the contiguous alternatives $\mathbb{K}_n$.

Lemma 6.4. *Suppose the Assumption (A1) is true. Then it holds that*

$$I_{D_\phi^{(k)}}^{(n)} = \frac{d^2}{2nh_n^{1/2}} \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''(f_x^k f_y^k) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy + o\left(\frac{d^2}{nh_n^{1/2}} \right). \tag{6.119}$$

Proof. Let us consider

$$\zeta(t) = \sum_{x=0}^{1} \int_{f_y>0} \left[\phi\Big((f_x f_y + t\Delta_{x,y})^k \Big) - \phi\Big(g_{x,y}(t)^k \Big) - \Big\{ (f_x f_y + t\Delta_{x,y})^k - g_{x,y}(t)^k \Big\} \phi'\Big(g_{x,y}(t)^k \Big) \right] dy \tag{6.120}$$

where $g_{x,y}(t) = (f_x + t\Delta_x)(f_y + t\Delta_y)$ with $t = \frac{d}{\sqrt{nh_n^{1/2}}}$. See that

$$f_{x,y}^{(n)} = f_x f_y + t\Delta_{x,y},\ f_x^{(n)} = f_x + t\Delta_x \text{ and } f_y^{(n)} = f_y + t\Delta_y. \tag{6.121}$$

Notice that $\zeta(0) = \zeta'(0) = 0$. Expanding $\zeta(t)$ around $t = 0$ up to second-order gives

$$\zeta(t) = \zeta(0) + t\zeta'(0) + \frac{t^2}{2}\zeta''(0) + \frac{t^3}{6}\zeta'''(\theta) \text{ for } 0 < \theta < t. \tag{6.122}$$

Simple calculations give

$$\begin{aligned}
\zeta''(t) = \sum_{x=0}^{1} \int_{f_y>0} \Bigg[& k^2 (f_x f_y + t\Delta_{x,y})^{2k-2} \Delta_{x,y}^2 \phi''\Big((f_x f_y + t\Delta_{x,y})^k \Big) \\
& + k(k-1)(f_x f_y + t\Delta_{x,y})^{k-2} \Delta_{x,y}^2 \phi'\Big((f_x f_y + t\Delta_{x,y})^k \Big) \\
& - k^2 (f_x f_y + t\Delta_{x,y})^{k-1} \Delta_{x,y} g_{x,y}^{k-1}(t) g'_{x,y}(t) \phi''\Big(g_{x,y}^k(t) \Big) \\
& - k(k-1)(f_x f_y + t\Delta_{x,y})^{k-2} \Delta_{x,y}^2 \phi'\Big(g_{x,y}^k(t) \Big) \\
& - \Big\{ (f_x f_y + t\Delta_{x,y})^k - g_{x,y}^k(t) \Big\} k^2 g_{x,y}^{2k-2}(t) (g'_{x,y}(t))^2 \phi'''\Big(g_{x,y}^k(t) \Big) \\
& - \Big\{ (f_x f_y + t\Delta_{x,y})^k - g_{x,y}^k(t) \Big\} k g_{x,y}^{k-1}(t) g''_{x,y}(t) \phi''\Big(g_{x,y}^k(t) \Big) \\
& - \Big\{ (f_x f_y + t\Delta_{x,y})^k - g_{x,y}^k(t) \Big\} k(k-1) g_{x,y}^{k-2}(t) (g'_{x,y}(t))^2 \phi''\Big(g_{x,y}^k(t) \Big) \\
& - \Big\{ k(f_x f_y + t\Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(t) g'_{x,y}(t) \Big\} k g_{x,y}^{k-1}(t) g'_{x,y}(t) \phi''\Big(g_{x,y}^k(t) \Big) \Bigg] dy \\
= & \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''(f_x^k f_y^k) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy \text{ at } t = 0,
\end{aligned} \tag{6.123}$$

$$
\begin{aligned}
\zeta'''(\theta) = \sum_{x=0}^{1} \int_{f_y>0} \Bigg[& k^3 (f_x f_y + \theta \Delta_{x,y})^{3k-3} \Delta_{x,y}^3 \phi''' \Big((f_x f_y + \theta \Delta_{x,y})^k \Big) \\
& + k^2 (2k-2) (f_x f_y + \theta \Delta_{x,y})^{2k-3} \Delta_{x,y}^3 \phi'' \Big((f_x f_y + \theta \Delta_{x,y})^k \Big) \\
& + k^2 (k-1) (f_x f_y + \theta \Delta_{x,y})^{2k-3} \Delta_{x,y}^3 \phi'' \Big((f_x f_y + \theta \Delta_{x,y})^k \Big) \\
& + k(k-1)(k-2) (f_x f_y + \theta \Delta_{x,y})^{k-3} \Delta_{x,y}^3 \phi' \Big((f_x f_y + \theta \Delta_{x,y})^k \Big) \\
& - k^3 (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} g_{x,y}^{2k-2}(\theta) (g'_{x,y}(\theta))^2 \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - k^2 (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} g_{x,y}^{k-1}(\theta) g''_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - k^2 (k-1) (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} g_{x,y}^{k-2}(\theta) (g'_{x,y}(\theta))^2 \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - k^2 (k-1) (f_x f_y + \theta \Delta_{x,y})^{k-2} \Delta_{x,y}^2 g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - k^2 (k-1) (f_x f_y + \theta \Delta_{x,y})^{k-2} \Delta_{x,y}^2 g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - k(k-1)(k-2) (f_x f_y + \theta \Delta_{x,y})^{k-3} \Delta_{x,y}^3 \phi' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k^3 g_{x,y}^{3k-3}(\theta) (g'_{x,y}(\theta))^3 \phi'''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k^2 g_{x,y}^{2k-2}(\theta) 2 g'_{x,y}(\theta) g''_{x,y}(\theta) \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k^2 (2k-2) g_{x,y}^{2k-3}(\theta) (g'_{x,y}(\theta))^3 \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k^2 g_{x,y}^{2k-2}(\theta) (g'_{x,y}(\theta))^2 \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k^2 g_{x,y}^{2k-2}(\theta) g'_{x,y}(\theta) g''_{x,y}(\theta) \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k(k-1) g_{x,y}^{k-2}(\theta) g'_{x,y}(\theta) g''_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k g_{x,y}^{k-1}(\theta) g''_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k^2 (k-1) g_{x,y}^{2k-3}(\theta) (g'_{x,y}(\theta))^3 \phi''' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k(k-1) g_{x,y}^{k-2}(\theta) 2 g'_{x,y}(\theta) g''_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ (f_x f_y + \theta \Delta_{x,y})^k - g_{x,y}^k(\theta) \Big\} k(k-1)(k-2) g_{x,y}^{k-3}(\theta) (g'_{x,y}(\theta))^3 \phi \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k(k-1) g_{x,y}^{k-2}(\theta) (g'_{x,y}(\theta))^2 \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k^2 g_{x,y}^{2k-2}(\theta) (g'_{x,y}(\theta))^2 \phi \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k g_{x,y}^{k-1}(\theta) g''_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k (f_x f_y + \theta \Delta_{x,y})^{k-1} \Delta_{x,y} - k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \Big\} k(k-1) g_{x,y}^{k-2}(\theta) (g'_{x,y}(\theta))^2 \phi'' \Big(g_{x,y}^k(\theta) \Big) \\
& - \Big\{ k(k-1) (f_x f_y + \theta \Delta_{x,y})^{k-2} \Delta_{x,y}^2 - k g_{x,y}^{k-1}(\theta) g''_{x,y}(\theta) - k(k-1) g_{x,y}^{k-2}(\theta) (g'_{x,y}(\theta))^2 \Big\} \\
& \times k g_{x,y}^{k-1}(\theta) g'_{x,y}(\theta) \phi'' \Big(g_{x,y}^k(\theta) \Big) \Bigg] dy. \qquad (6.124)
\end{aligned}
$$

Assumption (A1) ensures that $|\zeta'''(\theta)| = \mathcal{O}(1)$. Therefore the remainder in the expansion of $\zeta(t)$ in (6.122) will be $\mathcal{O}(t^3) = o(t^2)$ when $t \to 0$. Substituting $t = \frac{d}{\sqrt{nh_n^{1/2}}}$ in (6.122) gives

$$I_{D_\phi^{(k)}}^{(n)} = \zeta(t) = \frac{d^2}{2(nh_n^{1/2})} \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''(f_x^k f_y^k) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy + o\left(\frac{d^2}{nh_n^{1/2}} \right). \tag{6.125}$$

This completes the proof. □

Next, we will see that a scaled version of the second-order term in (6.119) adds a location-shift to the null distribution of $\widehat{I}_{D_\phi^{(k)}}$ under the contiguous alternatives. Now we are ready to derive the asymptotic distribution of $T_{D_\phi^{(k)}}$ under the contiguous alternatives $\mathbb{K}_n$.

Theorem 6.3. *Suppose that the Assumptions (A1) - (A5) are true, and* $0 \le d \le C \cdot \sup_{x,y} |f_{x,y} - f_x f_y|$ *for a constant* $C > 0$. *Then the following result is true:*

$$T_{D_\phi^{(k)}} \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''(f_x^k f_y^k) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy + \mathcal{N}(0,1) \tag{6.126}$$

as $n \to \infty$ *under the contiguous alternatives* $\mathbb{K}_n$.

Proof. Now see that

$$\begin{aligned} T_{D_\phi^{(k)}} &= nh_n^{1/2} \sigma_\phi^{-1} \left(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n} \right) \\ &= nh_n^{1/2} \sigma_\phi^{-1} \left(\widehat{I}_{D_\phi^{(k)}} - I_{D_\phi^{(k)}}^{(n)} - \frac{\mu_\phi}{nh_n} \right) + nh_n^{1/2} \sigma_\phi^{-1} I_{D_\phi^{(k)}}^{(n)} \\ &= nh_n^{1/2} \sigma_\phi^{-1} \left(\widehat{I}_{D_\phi^{(k)}} - I_{D_\phi^{(k)}}^{(n)} - \frac{\mu_\phi}{nh_n} \right) \\ &\quad + \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''(f_x^k f_y^k) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy + o(1). \end{aligned} \tag{6.127}$$

Let us define

$$U_n = nh_n^{1/2}\sigma_\phi^{-1}\Big(\widehat{I}_{D_\phi^{(k)}} - I^{(n)}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n}\Big),\ W_n(t) = \mathbb{1}\big\{U_n \le t\big\} \text{ and } L_n = \prod_{i=1}^{n}\frac{f^{(n)}_{X_i,Y_i}(x_i,y_i)}{f_{X_i}(x_i)f_{Y_i}(y_i)} \tag{6.128}$$

for any fixed $t \in \mathbb{R}$. The probability density function under the null hypothesis is $f_{x,y} = f_x f_y$ for all x, y. Let $F^{(n)}$ and F_0 be the distribution functions associated with contiguous alternatives and null hypothesis. Since $f^{(n)}_{x,y} \to f_x f_y$ pointwise, it follows from the theorem of Scheffè that $F^{(n)} \xrightarrow{\mathcal{L}} F_0$. Then an application of the Portmanteau lemma implies that

$$\limsup_{n\to\infty}\mathbb{P}_{\mathbb{K}_n}\big\{U_n \le t\big\} \le \limsup_{n\to\infty}\mathbb{P}_{\mathbb{H}}\big\{U_n \le t\big\} = \Phi_1(t), \tag{6.129}$$

as $d = 0$ under the null hypothesis $\mathbb{H}$. In the other direction, see that

$$\begin{aligned}\mathbb{P}_{\mathbb{K}_n}\big\{U_n \le t\big\} = \mathbb{E}_{\mathbb{K}_n}\Big[W_n(t)\Big] &\ge \sum\int_{\prod_{i=1}^n f_{x_i}f_{y_i}>0} W_n(t)\prod_{i=1}^{n} f^{(n)}_{X_iY_i}(x_i,y_i)dy_i \\ &= \mathbb{E}_{\mathbb{H}}\Big[W_n(t)L_n\Big] \text{ for all } n. \end{aligned}\tag{6.130}$$

Thus we have $\liminf_{n\to\infty}\mathbb{P}_{\mathbb{K}_n}\big\{U_n \le t\big\} \ge \liminf_{n\to\infty}\mathbb{E}_{\mathbb{H}}\Big[W_n(t)L_n\Big]$. Under the null hypothesis $\mathbb{H}$, see that

$$L_n = 1 \text{ and } W_n(t) = \mathbb{1}\left\{nh_n^{1/2}\sigma_\phi^{-1}\left(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n^{1/2}}\right) \le t\right\} \xrightarrow{\mathcal{L}} T, \tag{6.131}$$

where $T \sim Bernoulli(\Phi_1(t))$. This gives $W_n(t)L_n \xrightarrow{\mathcal{L}} T$ for any $t \in \mathbb{R}$ under the null hypothesis. Since W_n is bounded, it is uniformly integrable. This results into $\mathbb{E}_{\mathbb{H}}\Big(W_n(t)L_n\Big) \longrightarrow$

$\mathbb{E}(T) = \Phi_1(t)$, and consequently we obtain

$$\mathbb{P}_{\mathbb{K}_n}\left\{ nh_n^{1/2}\sigma_\phi^{-1}\Big(\widehat{I}_{D_\phi^{(k)}} - I^{(n)}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n}\Big) \le t \right\} \longrightarrow \Phi_1(t) \text{ for all } t \in \mathbb{R} \tag{6.132}$$

when $n \to \infty$. Putting all these pieces together results into

$T_{D_\phi^{(k)}} \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma_\phi}\sum_{x=0}^{1}\int_{f_y>0} k^2(f_x f_y)^{2k}\phi''(f_x^k f_y^k)\Big(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y}\Big)^2 dy + \mathcal{N}(0,1)$ under $\mathbb{K}_n$. This completes the proof. □

Remark 6.3. *In particular, suppose we take $\Delta_{x,y} = (\delta_{x_0}(x)\delta_{y_0}(y) - f_x f_y)$ where $\delta_{x_0}(x)\delta_{y_0}(y)$ is a probability density function degenerate at a point $t_0 = (x_0, y_0)$. Then we obtain*

$$T_{D_\phi^{(k)}} \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma_\phi}\mathcal{IF}_2\Big(I_{D_\phi^{(k)}}, f_X f_Y, t_0\Big) + \mathcal{N}(0,1) \text{ under } \mathbb{K}_n, \tag{6.133}$$

where $\mathcal{IF}_2$ is the second-order influence function of $I_{D_\phi^{(k)}}$ under the null distribution at the point t_0. Later we will see that $\mathcal{IF}_2$ measures the infinitesimal stability behaviour of $I_{D_\phi^{(k)}}$ under the null hypothesis $\mathbb{H}$ at a point t_0. Computation of $\mathcal{IF}_2$ is deferred to the next section. The asymptotic local power at contiguous alternatives $\mathbb{K}_n$ is therefore given by

$$\lim_{n\to\infty}\mathbb{P}_{\mathbb{K}_n}\Big[T_{D_\phi^{(k)}} > \tau_c\Big] = 1 - \Phi_1(\tau_c - d^2 S_\phi) \text{ where } S_\phi := \frac{\mathcal{IF}_2(I_{D_\phi^{(k)}}, f_X f_Y, t_0)}{2\sigma_\phi}. \tag{6.134}$$

We may call S_ϕ as a slope function associated with the asymptotic power of $T_{D_\phi^{(k)}}$ at local alternatives. Observe that the local asymptotic power of $T_{D_\phi^{(k)}}$ varies as the slope function S_ϕ changes with the ϕ-function itself. We can study S_ϕ to compare the power of the standardized test statistic $T_{D_\phi^{(k)}}$ at such contiguous alternatives for different members of the ϕ-generated extended Bregman divergence family. Also, note that S_ϕ is positive for all strictly convex and at least twice differentiable ϕ-function. The larger the slope S_ϕ becomes, the faster the asymptotic contiguous power increases towards one from the nominal level "c" with the increment of d^2. So

the second-order influence function $\mathcal{IF}_2$, which is a measure of robustness for $I_{D_\phi^{(k)}}$, is directly linked to the local asymptotic power of $T_{D_\phi^{(k)}}$ at contiguous alternatives $\mathbb{K}_n$ in an interesting way.

6.4 Robustness Studies

The robustness of the test statistics depends on the stability of $I_{D_\phi^{(k)}}$ itself. This is the topic of our next discussion. We already know that the influence function (IF) is one of the popular measures of robustness. However, it is a local measure and sometimes it may fail to reveal additional robustness features. In such cases, we need to study the higher-order influence functions.

6.4.1 Influence Function Analysis of $I_{D_\phi^{(k)}}$ under Independence

Let the joint density of (X, Y) be ϵ-contaminated at a point $t_0 = (x_0, y_0) \in \{0, 1\} \times \mathbb{R}$ as

$$f^{\epsilon}_{X,Y}(x, y) = (1 - \epsilon) f_{x,y} + \epsilon \delta_{x_0}(x) \delta_{y_0}(y), \tag{6.135}$$

where $\delta_{z_0}(z) = \mathbb{1}\{z = z_0\}$ is the Dirac delta function and $0 \leq \epsilon \leq 1$. We know that $\int \delta_{z_0}(dz) = 1$, or $\sum \delta_{z_0}(z) = 1$ according to z is continuous or discrete. A convenient abuse of notation for the integration of the Dirac delta function is $\int \delta_{z_0}(z)dz = 1$ which is understood in the former sense. The marginals corresponding to the ϵ-contaminated joint density are given by

$$f^{\epsilon}_X(x) = (1 - \epsilon) f_x + \delta_{x_0}(x) \text{ and } f^{\epsilon}_Y(y) = (1 - \epsilon) f_y + \epsilon \delta_{y_0}(y). \tag{6.136}$$

Simple calculations show that the first-order influence function of the generalized mutual information functional $I_{D_\phi^{(k)}}$ at the point $t_0 = (x_0, y_0)$ is given by

$$\mathcal{IF}_1(I_{D_\phi^{(k)}}, f_{XY}, t_0) = \sum_{x=0}^{1} \int_{f_y>0} \Bigg[k(f_{x,y})^{k-1} \Big(\delta_{x_0}(x)\delta_{y_0}(y) - f_{x,y}\Big) \Big\{ \phi'\Big(f_{x,y}^k\Big) - \phi'\Big(f_x^k f_y^k\Big) \Big\}$$

$$-k(f_x f_y)^{k-1} \Big(f_{x,y}^k - (f_x f_y)^k\Big) \Big(f_x \Big(\delta_Y(y_0) - f_y\Big) + f_y(\delta_X(x_0) - f_x)\Big) \phi''\Big(f_x^k f_y^k\Big) \Bigg] dy. \tag{6.137}$$

Under the null hypothesis $\mathbb{H}$, this turns out to be

$$\mathcal{IF}_1(I_{D_\phi^{(k)}}, f_X f_Y, t_0) = 0 \text{ for all } t_0 \in \{0,1\} \times \mathbb{R}. \tag{6.138}$$

See that the first-order influence function fails to reveal any robustness feature of the functional $I_{D_\phi^{(k)}}$ under independence for whatever the ϕ-function may be. To get further insights, we look up to the second-order influence function which is defined in a similar way as

$$\mathcal{IF}_2(I_{D_\phi^{(k)}}, f_{X,Y}, t_0) = \left[\frac{\partial^2 D_\phi^{(k)}(f_{X,Y}^\epsilon, f_X^\epsilon f_Y^\epsilon)}{\partial \epsilon^2} \right]_{\epsilon=0}. \tag{6.139}$$

An explicit expression of the second-order influence function under the null hypothesis is given in the next result.

Theorem 6.4. *Under the null hypothesis $\mathbb{H}$, the second-order influence function of $I_{D_\phi^{(k)}}$ at a point t_0 is given by*

$$\mathcal{IF}_2(I_{D_\phi^{(k)}}, f_X f_Y, t_0) = \sum_{x=0}^{1} \int_{f_y>0} k^2 (f_x f_y)^{2k} \phi''\Big(f_x^k f_y^k\Big) \left[\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right]^2 dy \tag{6.140}$$

for $t_0 \in \{0,1\} \times \mathbb{R}$, where $\Delta_x = \delta_{x_0}(x) - f_x$, $\Delta_y = \delta_{y_0}(y) - f_y$ and $\Delta_{x,y} = \delta_{x_0}(x)\delta_{y_0}(y) - f_x f_y$.

The proof of this result easily follows from (6.123). It is clear from (6.140) that the values of the second-order influence function are always positive. The higher the absolute values of the influence function, the better the test statistics' stability becomes.

To get further simplification for computational convenience, it is useful to use the following property of the Dirac delta function: $\int_{-\infty}^{\infty} g(z)\delta_{z_0}(dz) = g(z_0)$ for a "well-behaved" function g. Later in Chapter 7 and Chapter 8 we shall plot the influence functions for the generalized S-Bregman and the Exponential-Polynomial divergence families, and discuss their properties in greater detail. As we find, there exist members from both these families corresponding to which the influence function yields much lower values when compared to the power divergence family. This indicates the existence of more robust but consistent tests outside the power divergence family. Also, we will show that the influence function will remain invariant across all members of the power divergence family. This means that the power divergence family is as stable as the Kullback-Leibler divergence in producing a test as long as the influence function is concerned. However, in practice, we know that the robustness of the power divergence family-based MI depends on its tuning parameter. This is one of the limitations of the influence function that often it might fail to reveal the robustness features.

Remark 6.4. *In parametric estimation, it is generally assumed that a contaminated density should not belong to the parametric family, or it should be geometrically separated from the true distribution. A similar interpretation may be drawn in this case as well. Notice that the null hypothesis* $\mathbb{H}$ *defines the class of densities* $\mathcal{H} := \{f_{x,y} : f_{x,y} = f_x f_y \text{ for all } x, y\}$. *The joint density under null gets contaminated as*

$$f^{\epsilon}_{x,y} = (1-\epsilon) f_x f_y + \epsilon \delta_{x_0}(x)\delta_{y_0}(y), \tag{6.141}$$

which no longer belongs to the class $\mathcal{H}$. *A non-robust distance such as the Kullback-Leibler*

divergence, which is very sensitive to contamination proportion, enlarges the dissimilarity between the contaminated and the true density (under null). Consequently, the level gets inflated. However, using robust statistical distances mitigates this dissimilarity. Therefore, the levels become stable. The role of the influence function in revealing the robustness of these non-parametric tests will be further cleared through the plots in later chapters.

Next, we compute the influence function of the test functional. Excluding the rate $nh_n^{1/2}$ the test functional may be considered as

$$W_{D_\phi^{(k)}} = \left[\frac{I_{D_\phi^{(k)}} - \frac{\widehat{\mu}_\phi}{nh_n}}{\widehat{\sigma}_\phi}\right]. \tag{6.142}$$

Simple calculations show that

$$\mathcal{IF}_1(W_{D_\phi^{(k)}}, f_X f_Y, t_0) = 0 \text{ and } \mathcal{IF}_2(W_{D_\phi^{(k)}}, f_X f_Y, t_0) = \frac{1}{\widehat{\sigma}_\phi}\mathcal{IF}_2(I_{D_\phi^{(k)}}, f_X f_Y, t_0) \tag{6.143}$$

under the null hypothesis $\mathbb{H}$. The stability behaviour of the test functional $W_{D_\phi^{(k)}}$ is determined by $I_{D_\phi^{(k)}}$ as their second-order influence functions are proportional to each other.

Unfortunately, things get much harder when the null hypothesis is not true. In those cases, we do not have simplified expressions for the second-order influence functions as the alternative hypothesis $\mathbb{K}$ is a composite one. An alternative approach is to study its stability at the contiguous alternatives using the level and power influence functions, which will be discussed in the next subsection.

6.4.2 Level and Power Influence Functions

In Subsection 6.4.1, we discuss the robustness of $I_{D_\phi^{(k)}}$ and $T_{D_\phi^{(k)}}$ through the second-order influence function analysis when the null hypothesis is true. But, how does the

level and power of this class of tests get affected when the distributions under the null and contiguous alternatives become contaminated? The answer is not a simple one. One way of finding the answer is to calculate the influence functions of the level and power functionals for this class of tests. But, the difficulty lies in the fact that closed-form expressions of the size and power of these tests are hard to compute at those contaminated distributions. Of course, one can make certain approximations under different conditions. In this section, we will explore them in further detail.

Let the densities under the null and contiguous alternatives be contaminated in the following way

$$f^{\epsilon}_{x,y} = f_x f_y + \frac{\epsilon}{\sqrt{nh_n^{1/2}}}\Delta_{x,y} \text{ and } f^{(P_n)}_{x,y} = f^{(n)}_{x,y} + \frac{\epsilon}{\sqrt{nh_n^{1/2}}}\Delta'_{x,y} \text{ for all } x, y \tag{6.144}$$

respectively, where $\epsilon \geq 0$ and $\sum_{x=0}^{1}\int_{f_y>0}\Delta'_{x,y}dy = 0$. Here we allow $\Delta'_{x,y}$ to depend on the sample size n. The densities $f^{(n)}_{x,y}$ under the contiguous alternatives are defined as in (6.116). When $\epsilon = 0$, no such contamination happens in either situation. In particular, when $\Delta_{x,y} = \Delta'_{x,y}$ for all x, y along with $d = 0$, this implies that $f^{(P_n)}_{x,y} = f^{\epsilon}_{x,y}$ for x, y. Now consider the point-mass contamination in either situation such as $\Delta_{x,y} = (\delta_{x_0}(x)\delta_{y_0}(y) - f_x f_y)$ and $\Delta'_{x,y} = (\delta_{x_1}(y)\delta_{y_1}(y) - f^{(n)}_{x,y})$ with $t_0 = (x_0, y_0)$ and $t_1 = (x_1, y_1)$. In these cases, the level and power functionals become

$$\alpha(f^{\epsilon}_{X,Y}, t_0) = \mathbb{P}_{f^{\epsilon}_{X,Y}}\left\{T_{D^{(k)}_{\phi}} > \tau_c\right\} \text{ and } \pi(f^{(P_n)}_{X,Y}, t_1) = \mathbb{P}_{f^{(P_n)}_{X,Y}}\left\{T_{D^{(k)}_{\phi}} > \tau_c\right\} \tag{6.145}$$

respectively. As in Hampel (1986), the level and power influence functions are therefore given by

$$\mathcal{LIF}(\alpha, t_0) = \lim_{n\longrightarrow\infty}\left[\frac{\partial\alpha(f^{(\epsilon)}_{X,Y}, t_0)}{\partial\epsilon}\right]_{\epsilon=0} \text{ and } \mathcal{PIF}(\pi, t_1) = \lim_{n\longrightarrow\infty}\left[\frac{\partial\pi(f^{(P_n)}_{X,Y}, t_1)}{\partial\epsilon}\right]_{\epsilon=0}. \tag{6.146}$$

To compute the level and power influence functions as in (6.146) we require the asymptotic distributions of $T_{D_\phi^{(k)}}$ under both the contaminated null and contaminated contiguous alternatives. The former is already derived in Theorem 6.3. The derivation of the asymptotic distribution of $T_{D_\phi^{(k)}}$ under $f_{X,Y}^{(P_n)}$ is the next thing we want to do. To do that let us define

$$I_{D_\phi^{(k)}}^{(P_n)} = D_\phi^{(k)}(f_{X,Y}^{(P_n)}, f_X^{(P_n)} f_Y^{(P_n)}) \tag{6.147}$$

where $f_X^{(P_n)}$ and $f_Y^{(P_n)}$ are the marginal densities of the joint density $f_{X,Y}^{(P_n)}$ as they are clearly understood from the context. The next lemma establishes an asymptotic expression for $I_{D_\phi^{(k)}}^{(P_n)}$.

Lemma 6.5. *Under the Assumption (A1) it holds that*

$$I_{D_\phi^{(k)}}^{(P_n)} = \frac{k^2}{2nh_n^{1/2}} \sum_{x=0}^{1} \int (f_x f_y)^{2k} \left\{ d\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y} \right) + \epsilon \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y} \right) \right\}^2 \phi''\left(f_x^k f_y^k \right) dy + o\left(\frac{(d+\epsilon)^2}{nh_n^{1/2}} \right). \tag{6.148}$$

Proof. Write

$$\Gamma(s,t) = \sum_{x=0}^{1} \int \left\{ \phi\left((f_{x,y}^{(P_n)})^k \right) - \phi\left((f_x^{(P_n)})^k (f_y^{(P_n)})^k \right) - \left((f_{x,y}^{(P_n)})^k - (f_x^{(P_n)})^k (f_y^{(P_n)})^k \right) \phi'\left((f_x^{(P_n)})^k (f_y^{(P_n)})^k \right) \right\} dy$$

where $s = \frac{\epsilon}{\sqrt{nh_n^{1/2}}}$ and $t = \frac{d}{\sqrt{nh_n^{1/2}}}$. Expanding $\Gamma(s,t)$ around $s = 0$ upto second-order gives

$$\Gamma(s,t) = \Gamma(0,t) + s\Gamma^{(10)}(0,t) + \frac{s^2}{2}\Gamma^{(20)}(0,t) + \frac{s^3}{3}\Gamma^{(30)}(s^*,t) \text{ with } 0 < s^* < s. \tag{6.149}$$

Here the partial derivatives are denoted as $\Gamma^{(r_1 r_2)}(s_1,t_1) = \frac{\partial^{r_1+r_2}}{\partial s^{r_1}\partial s^{r_2}}\Gamma(s,t)\Big|_{s=s_1,t=t_1}$, $r_1, r_2 = 1,2,\ldots$. We get $f_{x,y}^{(P_n)} = f_{x,y}^{(n)}$ for all x,y when $s=0$. This gives $\Gamma(0,t) = I^{(n)}_{D_\phi^{(k)}}$ as in (6.118). From Lemma 6.4 we know that

$$\Gamma(0,t) = I^{(n)}_{D_\phi^{(k)}} = \frac{d^2}{2nh_n^{1/2}}\sum_{x=0}^{1}\int k^2(f_x f_y)^{2k}\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)^2 \phi''\left(f_x^k f_y^k\right)dy + o\left(\frac{d^2}{nh_n^{1/2}}\right). \tag{6.150}$$

Simple calculation yields that

$$\begin{aligned}\Gamma^{(10)}(0,t) = \sum_{x=0}^{1}\int_{f_y>0}\Bigg[&k\left(f_{x,y}^{(n)}\right)^k\frac{\Delta'_{x,y}}{f_{x,y}^{(n)}}\left\{\phi'\left((f_{x,y}^{(n)})^k\right) - \phi'\left((f_x^{(n)}f_y^{(n)})^k\right)\right\}\\ &- k\left(\frac{\Delta'_x}{f_x^{(n)}} + \frac{\Delta'_y}{f_y^{(n)}}\right)\left((f_{x,y}^{(n)})^k - (f_x^{(n)}f_y^{(n)})^k\right)\left(f_x^{(n)}f_y^{(n)}\right)^k\phi''\left((f_x^{(n)}f_y^{(n)})^k\right)\Bigg]dy.\end{aligned} \tag{6.151}$$

Expanding $\Gamma^{(10)}(0,t)$ around $t=0$ upto gives

$$\Gamma^{(10)}(0,t) = \Gamma^{(10)}(0,0) + t\Gamma^{(11)}(0,0) + \frac{t^2}{2}\Gamma^{(12)}(0,t^*) \text{ where } 0 < t^* < t. \tag{6.152}$$

Simple calculations show that $\Gamma^{(10)}(0,0) = 0$ and

$$\Gamma^{(11)}(0,0) = \sum_{x=0}^{1}\int k^2(f_x f_y)^{2k}\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y}\right)\phi''\left(f_x^k f_y^k\right)dy. \tag{6.153}$$

Using the same argument as before one can show that the remainder term is $o(t)$. Thus we obtain

$$s\Gamma^{(10)}(0,t) = \frac{d\epsilon}{nh_n^{1/2}}\sum_{x=0}^{1}\int k^2(f_x f_y)^{2k}\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y}\right)\phi''\left(f_x^k f_y^k\right)dy + o\left(\frac{d\epsilon}{nh_n^{1/2}}\right). \tag{6.154}$$

Next, see that

$$\begin{aligned}\Gamma^{(20)}(0,t) = \sum_{x=0}^{1} \int \Bigg\{ & k(k-1)(f_{x,y}^{(n)})^k \left(\frac{\Delta'_{x,y}}{f_{x,y}^{(n)}}\right)^2 \left\{\phi'\left(f_{x,y}^{(n)}\right)^k - \phi'\left(f_x^{(n)} f_y^{(n)}\right)^k\right\} \\ & + k^2 (f_{x,y}^{(n)})^{2k} \left(\frac{\Delta'_{x,y}}{f_{x,y}^{(n)}}\right)^2 \phi''\left((f_{x,y}^{(n)})^k\right) \\ & - k^2 (f_{x,y}^{(n)} f_x^{(n)} f_y^{(n)})^k \frac{\Delta'_{x,y}}{f_{x,y}^{(n)}} \left(\frac{\Delta'_x}{f_x^{(n)}} + \frac{\Delta'_y}{f_y^{(n)}}\right) \phi''\left((f_x^{(n)} f_y^{(n)})^k\right) \\ & - k^2 (f_x^{(n)} f_y^{(n)})^{2k} \left\{\left(\frac{f_{x,y}^{(n)}}{f_x^{(n)} f_y^{(n)}}\right)^k \frac{\Delta'_{x,y}}{f_{x,y}^{(n)}} - \frac{\Delta'_x}{f_x^{(n)}} - \frac{\Delta'_y}{f_y^{(n)}}\right\} \left(\frac{\Delta'_x}{f_x^{(n)}} + \frac{\Delta'_y}{f_y^{(n)}}\right) \phi''\left((f_x^{(n)} f_y^{(n)})^k\right) \\ & - \left((f_{x,y}^{(n)})^k - (f_x^{(n)} f_y^{(n)})^k\right) (f_x^{(n)} f_y^{(n)})^k \Bigg\{ k(k-1) \left(\frac{\Delta'_x}{f_x^{(n)}}\right)^2 + 2k^2 \left(\frac{\Delta'_x \Delta'_y}{f_x^{(n)} f_y^{(n)}}\right) \\ & + k(k-1) \left(\frac{\Delta'_y}{f_y}\right)^2 \Bigg\} \phi''\left(\left(f_x^{(n)} f_y^{(n)}\right)^k\right) \\ & - k \left\{(f_{x,y}^{(n)})^k - (f_x^{(n)} f_y^{(n)})^k\right\} \left(\frac{\Delta'_x}{f_x^{(n)}} + \frac{\Delta'_y}{f_y^{(n)}}\right)^2 \phi'''\left((f_x^{(n)} f_y^{(n)})^k\right) \Bigg\} dy. \end{aligned} \tag{6.155}$$

As before, we expand $\Gamma^{(20)}(0,t)$ around $t = 0$ and get

$$\Gamma^{(20)}(0,t) = \Gamma^{(20)}(0,0) + t\Gamma^{(21)}(0,t^{**}) \text{ where } 0 < t^{**} < t, \tag{6.156}$$

where

$$\Gamma^{(20)}(0,0) = \sum_{x=0}^{1} \int k^2 (f_x f_y)^{2k} \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y}\right)^2 \phi''\left(f_x^k f_y^k\right) dy. \tag{6.157}$$

Thus we obtain

$$\frac{s^2}{2}\Gamma^{(20)}(0,t) = \frac{\epsilon^2}{2nh_n^{1/2}} \sum_{x=0}^{1} \int k^2 (f_x f_y)^{2k} \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y}\right)^2 \phi''\left(f_x^k f_y^k\right) dy + O\left(\frac{s^2 t}{2}\right). \tag{6.158}$$

Combining (6.150), (6.154) and (6.158) we get

$$I^{(P_n)}_{D^{(k)}_\phi} = \frac{k^2}{2nh_n^{1/2}} \sum_{x=0}^{1} \int (f_x f_y)^{2k} \left\{ d\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y} \right) + \epsilon \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y} \right) \right\}^2 \phi''\left(f_x^k f_y^k \right) dy + o\left(\frac{(d+\epsilon)^2}{nh_n^{1/2}} \right).$$

This completes the proof. □

Now, we proceed to find the asymptotic distribution of $I^{(P_n)}_{D^{(k)}_\phi}$ under $f^{(P_n)}_{X,Y}$.

Theorem 6.5. *Suppose the Assumptions (A1) - (A5) are true and* $0 \leq \max\{d, \epsilon\} \leq C^* \sup_{x,y} |f_{x,y} - f_x f_y|$ *for some* $C^* > 0$. *Then it follows that*

$$T_{D^{(k)}_\phi} - \frac{k^2}{2\sigma_\phi} \sum_{x=0}^{1} \int (f_x f_y)^{2k} \left\{ d\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y} \right) + \epsilon \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y} \right) \right\}^2 \phi''\left(f_x^k f_y^k \right) dy \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \tag{6.159}$$

under $f^{(P_n)}_{X,Y}$ *as* $n \to \infty$.

Proof. Using Lemma 6.5 we see that

$$\begin{aligned} T_{D^{(k)}_\phi} &= nh_n^{1/2} \sigma_\phi^{-1} \left(\widehat{I}_{D^{(k)}_\phi} - \frac{\mu_\phi}{nh_n^{1/2}} - I^{(P_n)}_{D^{(k)}_\phi} \right) + nh_n^{1/2} \sigma_\phi^{-1} I^{(P_n)}_{D^{(k)}_\phi} \\ &= U_n + \frac{k^2}{2\sigma_\phi} \sum_{x=0}^{1} \int (f_x f_y)^{2k} \left\{ d\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y} \right) + \epsilon \left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y} \right) \right\}^2 \phi''\left(f_x^k f_y^k \right) dy + o(1), \end{aligned}$$

where $U_n = nh_n^{1/2} \sigma_\phi^{-1} \left(\widehat{I}_{D^{(k)}_\phi} - \frac{\mu_\phi}{nh_n^{1/2}} - I^{(P_n)}_{D^{(k)}_\phi} \right)$. See that $f^{(P_n)}_{x,y} \longrightarrow f_x f_y$ for all x, y, when $n \to \infty$. So $f^{(P_n)}_{X,Y}$ is contiguous to the joint density under the null hypothesis. Applying the

results of Portmanteau as in Theorem 6.3, we find that

$$\lim_{n\to\infty} \mathbb{P}_{f_{X,Y}^{(P_n)}}\Big[U_n \le t\Big] = \lim_{n\to\infty} \mathbb{E}_{f_{X,Y}^{(P_n)}}\Big(\mathbb{1}\{U_n \le t\}\Big) = \lim_{n\to\infty} \mathbb{E}_{\mathbb{H}}\Big(\mathbb{1}\{U_n \le t\}L_n\Big) \text{ for fixed } t \in \mathbb{R}, \tag{6.160}$$

where $L_n = \prod_{i=1}^{n} \frac{f_{x,y}^{(P_n)}}{f_x f_y}$. From the condition $0 \le \max\{d, \epsilon\} \le C^* \sup_{x,y} |f_{x,y} - f_x f_y|$ we find that $d = \epsilon = 0$, and consequently $f_{x,y}^{(P_n)} = f_x f_y$ for all x, y when the null hypothesis is true. Thus we get $L_n = 1$ and

$$\mathbb{1}\Big\{U_n \le t\Big\} = \mathbb{1}\left\{nh_n^{1/2}\Big(\widehat{I}_{D_\phi^{(k)}} - \frac{\mu_\phi}{nh_n}\Big) \le t\right\} \xrightarrow{L} T \text{ under } \mathbb{H}, \tag{6.161}$$

where $T \sim Bernoulli(\Phi_1(t))$. Since $\mathbb{1}\Big\{U_n \le t\Big\}$ is bounded, it is uniformly intergrable. Then it follows that

$$\mathbb{P}_{f_{X,Y}^{(P_n)}}\Big[U_n \le t\Big] \longrightarrow \mathbb{E}_{\mathbb{H}}(T) = \Phi_1(t) \tag{6.162}$$

for all $t \in \mathbb{R}$. Thus it follows that

$$T_{D_\phi^{(k)}} - \frac{k^2}{2\sigma_\phi} \sum_{x=0}^{1} \int (f_x f_y)^{2k} \left\{ d\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right) + \epsilon\left(\frac{\Delta'_{x,y}}{f_x f_y} - \frac{\Delta'_x}{f_x} - \frac{\Delta'_y}{f_y}\right)\right\}^2 \phi''\Big(f_x^k f_y^k\Big) dy \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

under $f_{X,Y}^{(P_n)}$ when $n \to \infty$. This completes the proof. □

Remark 6.5. *The distribution of the test statistics under $f_{X,Y}^{(P_n)}$ becomes as in Theorem 6.3 when $\Delta_{x,y} = \Delta'_{x,y}$ for all x, y with $\epsilon = 0$.*

Now we are, in the position, to obtain the expressions of $\mathcal{LIF}$ and $\mathcal{PIF}$. As already seen in the previous chapters, exact values of level and power under ϵ-contaminated versions of true null and contiguous distributions are very difficult to compute. Only

limiting values of them can be obtained under certain conditions. To be able to do that we need to interchange the order of limit and differentiation which is true under certain conditions. These results are presented in the next theorem.

To compute the power influence function we need to take

$$\Delta'_{x,y} = \delta_{x_1}(x)\delta_{y_1}(y) - f^{(P_n)}_{x,y}. \tag{6.163}$$

Theorem 6.6. *Suppose that the sequence* $\left\{\frac{\partial}{\partial\epsilon}\pi(f^{(P_n)}_{X,Y}, t_1)\right\}$ *converges uniformly for all* $\epsilon \in [0,1]$ *at fixed* $t_1 = (x_1, y_1)$*. Further assume that the conditions of Theorem 6.5 is true. Then we have the following results.*

(a) The power influence function is given by

$$\begin{aligned}\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi}\sum_{x=0}^{1}\int k^2(f_x f_y)^{2k}\left(\frac{\Delta^*_{x,y}}{f_x f_y} - \frac{\Delta^*_x}{f_x} - \frac{\Delta^*_y}{f_y}\right)\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\phi''\left(f^k_x f^k_y\right)dy \\ \times \phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi}\mathcal{IF}_2(I_{D^{(k)}_\phi}, f_X f_Y, t_0)\right)\end{aligned} \tag{6.164}$$

where $\Delta^*_{x,y} = (\delta_{x_1}(x)\delta_{y_1}(y) - f_x f_y)$*,* $\Delta^*_x = (\delta_{x_1}(x) - f_x)$ *and* $\Delta^*_y = (\delta_{y_1}(y) - f_y)$*.*

(b) When $t_0 = t_1$*, we get*

$$\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi}\mathcal{IF}_2(I_{D^{(k)}_\phi}, f_X f_Y, t_1)\cdot\phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi}\mathcal{IF}_2(I_{D^{(k)}_\phi}, f_X f_Y, t_1)\right). \tag{6.165}$$

(c) The level influence function is given by

$$\mathcal{LIF}(\pi, t_0) \equiv 0. \tag{6.166}$$

Proof. (a) Let us denote

$$r_n = \frac{k^2}{2\sigma_\phi}\sum_{x=0}^{1}\int (f_x f_y)^{2k}\left\{d\left(\frac{\Delta_{x,y}}{f_x f_y}-\frac{\Delta_x}{f_x}-\frac{\Delta_y}{f_y}\right)+\epsilon\left(\frac{\Delta'_{x,y}}{f_x f_y}-\frac{\Delta'_x}{f_x}-\frac{\Delta'_y}{f_y}\right)\right\}^2 \phi''\left(f_x^k f_y^k\right)dy. \tag{6.167}$$

See that $r_n \to r_*$ where

$$r_* = \frac{k^2}{2\sigma_\phi}\sum_{x=0}^{1}\int (f_x f_y)^{2k}\left\{d\left(\frac{\Delta_{x,y}}{f_x f_y}-\frac{\Delta_x}{f_x}-\frac{\Delta_y}{f_y}\right)+\epsilon\left(\frac{\Delta^*_{x,y}}{f_x f_y}-\frac{\Delta^*_x}{f_x}-\frac{\Delta^*_y}{f_y}\right)\right\}^2 \phi''\left(f_x^k f_y^k\right)dy. \tag{6.168}$$

It is known from Theorem 6.5 that

$$\left|\mathbb{P}_{f_{X,Y}^{(P_n)}}\left\{T_{D_\phi^{(k)}} - t_n > \tau_c - r_n\right\} - (1-\Phi_1(\tau_c - r_n))\right| \longrightarrow 0 \text{ when } n \to \infty. \tag{6.169}$$

A first-order Taylor series expansion gives

$$\Phi_1(\tau_c - r_n) = \Phi_1(\tau_c - r_*) + (r_* - r_n)\phi_1(\tau_c - r'_*) \tag{6.170}$$

where r'_* is an intermediate point between r_n, r_*. Since ϕ_1 is bounded,

$$\mathbb{P}_{f_{X,Y}^{(P_n)}}\left\{T_{D_\phi^{(k)}} - r_n > \tau_c - r_n\right\} \longrightarrow 1-\Phi_1(\tau_c - r_*) \text{ when } n \to \infty. \tag{6.171}$$

The assumption of uniformly convergence of $\left\{\frac{\partial}{\partial\epsilon}\pi(f_{X,Y}^{(P_n)}, t_1)\right\}$ implies that differentiation and limit can be interchanged. This, in turn, gives the power influence function as

$$\mathcal{PIF}(\pi, t_1) = \left[\frac{\partial}{\partial\epsilon}\left\{\lim_{n\longrightarrow\infty}\pi(f_{X,Y}^{(P_n)}, t_1)\right\}\right]_{\epsilon=0} = \left[\frac{\partial r_*}{\partial\epsilon}\cdot\phi_1(\tau_c - r_*)\right]_{\epsilon=0}. \tag{6.172}$$

When $\epsilon = 0$, we find that $r_* = \frac{d^2}{2\sigma_\phi}\mathcal{IF}_2(I_{D_\phi^{(k)}}, f_X f_Y, t_1)$ (see Theorem 6.4) and

$$\left[\frac{\partial r_*}{\partial \epsilon}\right]_{\epsilon=0} = \frac{d}{\sigma_\phi}\sum_{x=0}^{1}\int k^2(f_x f_y)^{2k}\left(\frac{\Delta^*_{x,y}}{f_x f_y} - \frac{\Delta^*_x}{f_x} - \frac{\Delta^*_y}{f_y}\right)\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\phi''\left(f_x^k f_y^k\right)dy. \tag{6.173}$$

This gives the desired result.

(b) When $t_0 = t_1$, we get $\Delta_{x,y} = \Delta^*_{x,y}$ for all x, y. Thus the result follows.

(c) See that $f_{x,y}^{(P_n)} = f^{(n)}$ for all x, y when $d = 0$ and $t_0 = t_1$. Substituting $d = 0$ and $t_0 = t_1$ in the expression of $\mathcal{PIF}$ as in (b) gives that $\mathcal{LIF}(\alpha, t_0) \equiv 0$.

This completes the proof. □

Remark 6.6. *The stability of the power influence function is proportional to the second-order influence function of $I_{D_\phi^{(k)}}$ under the null distribution. However, $\mathcal{LIF}$ is zero identically. As in the earlier chapters, the level influence function also fails to reveal the robustness of the type-I error, which is not the case in reality. Because, in simulation studies, we show that the robustness of the levels of this class of tests depends on the choice of the ϕ-function as well.*

6.4.3 Asymptotic Breakdown Point of $I_{D_\phi^{(k)}}$

Earlier we studied the robustness of $I_{D_\phi^{(k)}}$ through influence function analysis. Influence function being defined as a directional derivative in the direction of a point mass describes the infinitesimal stability of the functional at that point. Although a useful concept, it sometimes fails to reveal the desired robustness of a functional that may be otherwise apparent. Since the influence function is entirely a local concept, it should be complemented by a global measure of robustness. One such global measure is the breakdown point that quantifies the maximum proportion of outlying observations in

a data set that a statistical functional may tolerate before it gives erratic values. The notion of outlying observations in this non-parametric setup is already discussed in Remark 6.4.

Sometimes it is intractable to compute the actual breakdown point using the definitions given in Hampel (1986). However asymptotic breakdown point is often easier to calculate in many situations.

Before starting our working definition, we introduce some useful notations. Let $k_{XY,m}(x,y) = k_{xy,m}$ be a sequence of contaminating densities, and $k_{x,m} = \int_{f_y>0} k_{xy,m}dy$, $k_{y,m} = \sum_{x=0}^{1} k_{xy,m}$ be its marginals. The true densities $f_{x,y}$, f_x and f_y are respectively contaminated as

$$\begin{aligned} f^m_{x,y} &= (1-\epsilon)f_{x,y} + \epsilon k_{xy,m}, \\ f^m_x &= (1-\epsilon)f_x + \epsilon k_{x,m}, \\ f^m_y &= (1-\epsilon)f_y + \epsilon k_{y,m} \end{aligned} \tag{6.174}$$

with $0 \le \epsilon \le 1$. Often we express $f^m_{x,y} = f^m_x f^m_{y|x}$ for all x, y where $f^m_{y|x}$ is the conditional density of y given x obtained from $f^m_{x,y}$. Similar notations are used for the other densities as well. Recall that $I_{D^{(k)}_\phi} = D^{(k)}_\phi(f_{X,Y}, f_X f_Y)$, and further define

$$I_m = D^{(k)}_\phi(f^m_{X,Y}, f^m_X f^m_Y) \text{ and } I'_m = D^{(k)}_\phi(f^m_{X,Y}, f_X f_Y). \tag{6.175}$$

The definition of the asymptotic breakdown point used so far in a parametric setup would not quite fit here with the different viewpoints. Here, the problem is more like– given a joint density, we are eager to calculate the mutual information. Of course, we need some kernel density estimates to be plugged into the mutual information to be

applicable in practical situations. When robust at a particular sequence of contaminating densities, the generalised mutual information should not be seemingly erratic with changing values as we move along that sequence of contaminating densities. In the view of this, contaminated mutual information I_m as in (6.175) should not break down at a contamination proportion ϵ when $\lim_{m\to\infty} I_m$ exists.

With this background, we give our working definition of a breakdown point in the present context.

Definition 6.2. *Given the true density $f_{X,Y}$, the generalized mutual information functional $I_{D_\phi^{(k)}}$ will breakdown asymptotically at ϵ if its contaminated version I_m does not converge when $I_{D_\psi^{(k)}} < \infty$. Then ϵ^* is said to be the asymptotic breakdown point of $I_{D_\phi^{(k)}}$ if*

$$\epsilon^* = \inf\left\{\epsilon : I_m \text{ does not converge but } I_{D_\psi^{(k)}} < \infty\right\}. \tag{6.176}$$

In the view of Definition 6.2, we know that $I_{D_\phi^{(k)}}$ may breakdown at ϵ only when it is bounded. The exclusion of the divergences with $I_{D_\phi^{(k)}} = \infty$ may somewhat limit the full generality of this definition of the breakdown point. It does not anyway affect much as it only leaves out some specific cases that are not statistically very interesting.

Following Park and Basu (2004) and Roy et al. (2023) we take a similar approach to establish the asymptotic breakdown point of $I_{D_\phi^{(k)}}$ but using Definition 6.2. Let us make the following assumptions:

(BP1) $\int \min\{f_{y|x}, k_{y|x,m}\}dy \longrightarrow 0$ as $m \to \infty$ for $x = 0, 1$,

(BP2) $\int \min\{f_{y|x}, f_y^m\}dy \longrightarrow 0$ as $m \to \infty$ for $x = 0, 1$,

(BP3) $\int \min\{f_y, k_{y|x,m}\}dy \longrightarrow 0$ as $m \to \infty$ for $x = 0, 1$,

(BP4) $\phi(0), \phi'(0)$ are finite and Y has a bounded support,

(BP5) $I_m \leq I'_m$ for all $m \geq 1$,

(BP6) there exists $\tilde{\epsilon} \in [0, \frac{1}{2}]$ such that for all $\epsilon < \tilde{\epsilon}$,

$$\begin{aligned}\liminf_{m\to\infty} D_\phi^{(k)}(\epsilon k_{XY,m}, f_X^m f_Y^m) > \limsup_{m\to\infty} \sum \int \Big[& \phi((\epsilon k_{xy,m})^k) - (\epsilon k_{xy,m})^k \phi'(0) \Big] dy \\ & + \sum \int \Big[((1-\epsilon) f_{xy})^k \phi'(0) - \phi((f_x f_y)^k) \\ & - \Big\{ ((1-\epsilon) f_{xy})^k - (f_x f_y)^k \Big\} \phi'((f_x f_y)^k) \Big] dy. \end{aligned} \tag{6.177}$$

The first Assumption (BP1) ensures that the conditional contaminating densities $k_{y|x,m}$ are asymptotically singular to the conditional density $f_{y|x}$ for $x = 0, 1$. Assumption (BP2) ensures that the contaminating densities f_y^m are also asymptotically singular to the conditional density $f_{y|x}$ for all x. Similarly, Assumption (BP3) makes the conditional contaminating densities $k_{y|x,m}$ asymptotically singular to the marginal density f_y for all x. Through these three conditions, we have rather assumed the worst possible scenarios that may drive I_m not to converge for sufficiently large m. In Assumption (BP4) certain technical conditions are imposed on the ϕ-function. These are true in most of the cases. In Assumption (BP4), it is also assumed that the support of Y is bounded. This is required to have a finite value of the integration in some degenerate case when a density is replaced by *zero* in this divergence measure. Some technical conditions are assumed in Assumption (BP5) because we do not have proof. It means that the $\mathscr{B}$-MI between contaminated joint density and uncontaminated marginals should be at least as the $\mathscr{B}$-MI between contaminated joint and its marginals. Note that, if Assumption (BP5) is negated, we arrive at a contradiction when X, Y are independent under contaminated joint densities $f_{x,y}^m$. Assumption (BP6) states the extremity of contamination

that could be handled in the next result to be true. Notice that if X, Y are independent Assumptions (BP1) and (BP3) become identical.

Theorem 6.7. *Suppose the Assumptions (BP1) - (BP6) are true with* $I_{D_\phi^{(k)}} < \infty$. *Then the asymptotic breakdown point of generalized mutual information* $I_{D_\phi^{(k)}}$ *is at least* $\min\left\{\frac{1}{2}, \tilde{\epsilon}\right\}$, *where* $\tilde{\epsilon}$ *is defined in Assumption (BP6).*

Proof. See that

$$I_m = \int \sum \left\{\phi((f^m_{x,y})^k) - \phi((f^m_x f^m_y)^k) - \left((f^m_{x,y})^k - (f^m_x f^m_y)^k\right)\phi'((f^m_x f^m_y)^k)\right\}. \tag{6.178}$$

Define $A_m = \left\{(x,y) : f_{x,y} > \max\{k_{xy,m}, f^m_x f^m_y\}\right\}$. The summation over "$x$" and integration over "$y$" are implicit for simplicity. Note that, as $m \to \infty$,

$$\begin{aligned}\int \sum_{A_m} k_{xy,m} = \int \sum_{A_m} \min\{f_{x,y}, k_{xy,m}\} &\leq \int \sum \min\{f_{x,y}, k_{xy,m}\} \\ &= \int \sum \min\{f_x f_{y|x}, k_{x,m} k_{y|x,m}\} \\ &\leq \sum \max\{f_x, k_{x,m}\} \int \min\{f_{y|x}, k_{y|x,m}\} \to 0 \end{aligned} \tag{6.179}$$

by Assumption (BP1) as $f_x, k_{x,m}$ are discrete and bounded in $[0,1]$. So we get $k_{xy,m} \to 0$, and subsequently $f^m_{x,y} \to (1-\epsilon) f_{x,y}$ on A_m as $m \to \infty$. Similarly, we also have

$$\begin{aligned}\int \sum_{A_m} f^m_x f^m_y = \int \sum_{A_m} \min\{f_{x,y}, f^m_x f^m_y\} &\leq \int \sum_{A_m} \max\{f_x, f^m_x\} \min\{f_{y|x}, f^m_y\} \\ &\leq \sum \max\{f_x, f^m_x\} \int \min\{f_{y|x}, f^m_y\} \to 0 \text{ as } m \to \infty, \end{aligned} \tag{6.180}$$

by Assumption (BP2). Thus we get $f_x^m f_y^m \to 0$ on A_m for $m \to \infty$. Now see that

$$\begin{aligned}\max\{k_{xy,m}, f_x^m f_y^m\} &\geq \min\{k_{x,m}, f_x^m\} \min\{k_{y|x,m}, f_y^m\} \\ &\geq \min\{k_{x,m}, f_x^m\} \min\{k_{y|x,m}, f_{y|x}, f_y^m\} \\ &= \min\{k_{x,m}, f_x^m\} \min\Big\{ \min\{k_{y|x,m}, f_{y|x}\}, \min\{f_{y|x}, f_y^m\}\Big\}. \end{aligned} \tag{6.181}$$

The last factor tends to 0 by Assumptions (BP1) and (BP2) when $m \to \infty$. This implies that $A_m \to \{(x,y) : f_{x,y} > 0\}$. Therefore we get

$$\begin{aligned}&\Bigg| \int \sum_{A_m} \Big\{\phi((f_{x,y}^m)^k) - \phi((f_x^m f_y^m)^k) - \Big((f_{x,y}^m)^k - (f_x^m f_y^m)^k\Big)\phi'((f_x^m f_y^m)^k)\Big\} \\ &\qquad - \int \sum \Big\{\phi\Big((1-\epsilon)^k f_{x,y}^k\Big) - \phi(0) - \Big((1-\epsilon)^k f_{x,y}^k\Big)\phi'(0)\Big\}\Bigg| \\ = &\left| \int \sum_{A_m} \Big\{\phi((f_{x,y}^m)^k) - \phi((f_x^m f_y^m)^k) - \Big((f_{x,y}^m)^k - (f_x^m f_y^m)^k\Big)\phi'((f_x^m f_y^m)^k)\Big\} - D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big)\right| \to 0 \end{aligned} \tag{6.182}$$

as $m \to \infty$, because $\phi(0), \phi'(0)$ are assumed to be finite in Assumption (BP4). Next, see that

$$\begin{aligned}\int \sum_{A_m^c} f_{x,y} &\leq \int \sum \min\{f_{x,y}, k_{xy,m}\} + \int \sum \min\{f_{x,y}, f_x^m f_y^m\} \\ &\leq \sum \max\{f_x, k_{x,m}\} \int \min\{f_{y|x}, k_{y|x,m}\} \\ &\quad + \sum \max\{f_x, f_x^m\} \int \min\{f_{y|x}, f_y^m\}. \end{aligned} \tag{6.183}$$

The first factor with each integral is bounded. So Assumptions (BP1) and (BP2) together imply that the above integration goes to 0 as $m \to \infty$. So A_m^c is asymptotically a null set

under $f_{X,Y}$. Therefore we get

$$\Bigg| \int \sum_{A_m^c} \Big\{ \phi((f_{x,y}^m)^k) - \phi((f_x^m f_y^m)^k) - \Big((f_{x,y}^m)^k - (f_x^m f_y^m)^k\Big)\phi'((f_x^m f_y^m)^k) \Big\}$$
$$- \underbrace{\sum \int \Big\{ \phi((\epsilon k_{xy,m})^k) - \phi((f_x^m f_y^m)^k) - \Big((\epsilon k_{xy,m})^k - (f_x^m f_y^m)^k\Big)\phi'((f_x^m f_y^m)^k) \Big\}}_{D_\phi^{(k)}(\epsilon k_{XY,m}, f_X^m f_Y^m)} \Bigg| \to 0 \quad (6.184)$$

for $m \to \infty$. Combining (6.182) and (6.184) gives

$$\left| I_m - D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big) - D_\phi^{(k)}\Big(\epsilon k_{XY,m}, f_X^m f_Y^m\Big) \right| \to 0 \text{ as } m \to \infty. \quad (6.185)$$

Define

$$a_1(\epsilon) = D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big) + \liminf_{m\to\infty} D_\phi^{(k)}\Big(\epsilon k_{XY,m}, f_X^m f_Y^m\Big). \quad (6.186)$$

Next, see that

$$\begin{aligned} \liminf_{m\to\infty} I_m &= \liminf_{m\to\infty} \left[D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big) + D_\phi^{(k)}\Big(\epsilon k_{XY,m}, f_X^m f_Y^m\Big) \right] \\ &\geq D_\phi^{(k)}((1-\epsilon) f_{X,Y}, 0) + \liminf_{m\to\infty} D_\phi^{(k)}(\epsilon k_{XY,m}, f_X^m f_Y^m) = a_1(\epsilon). \end{aligned} \quad (6.187)$$

Also, see that $\limsup_{m\to\infty} I_m \leq \limsup_{m\to\infty} I'_m$ as $I_m \leq I'_m$ for all $m \geq 1$ by Assumption (BP5). Consider

$$B_m = \Big\{ (x,y) : k_{xy,m} > \max\{f_{x,y}, f_x f_y\} \Big\}. \quad (6.188)$$

Similarly as before, B_m is asymptotically null set under both $f_{X,Y}$ and $f_X f_Y$. Also note that B_m^c is asymptotically null under $k_{XY,m}$. Using the same argument as before we

obtain

$$\left| I'_m - D_\phi^{(k)}(\epsilon k_{XY,m}, 0) - D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, f_X f_Y\Big) \right| \to 0 \text{ for } m \to \infty. \tag{6.189}$$

Next, see that

$$\begin{aligned} \limsup_{m\to\infty} I_m \le \limsup_{m\to\infty} I'_m &= \limsup_{m\to\infty} \left[D_\phi^{(k)}(\epsilon k_{XY,m}, 0) + D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, f_X f_Y\Big) \right] \\ &\le \limsup_{m\to\infty} D_\phi^{(k)}(\epsilon k_{XY,m}, 0) + D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, f_X f_Y\Big) = a_2(\epsilon) \text{ (say).} \end{aligned} \tag{6.190}$$

Suppose $a_2(\epsilon) < a_1(\epsilon)$, then $\limsup_{m\to\infty} I_m < \liminf_{m\to\infty} I_m$. So $\lim_{m\to\infty} I_m$ exists, i.e., asymptotically there will be no breakdown, when $a_2(\epsilon) < a_1(\epsilon)$ which is the same as

$$D_\phi^{(k)}(\epsilon k_{XY,m}, 0) + D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, f_X f_Y\Big) < D_\phi^{(k)}(\epsilon k_{XY,m}, f_X^m f_Y^m) + D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big) \tag{6.191}$$

for sufficiently large m. This condition is equivalent to

$$\begin{aligned} \liminf_{m\to\infty} D_\phi^{(k)}(\epsilon k_{XY,m}, f_X^m f_Y^m) &> \limsup_{m\to\infty} D_\phi^{(k)}(\epsilon k_{XY,m}, 0) + D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, f_X f_Y\Big) - D_\phi^{(k)}\Big((1-\epsilon) f_{X,Y}, 0\Big) \\ &= \limsup_{m\to\infty} \sum \int \Big[\phi((\epsilon k_{xy,m})^k) - (\epsilon k_{xy,m})^k \phi'(0) \Big] \\ &\quad + \sum \int \left(((1-\epsilon) f_{xy})^k \phi'(0) - \phi(f_x^k f_y^k) \right. \\ &\quad \left. - \Big\{ ((1-\epsilon) f_{xy})^k - (f_x f_y)^k \Big\} \phi'(f_x^k f_y^k) \right] dy. \end{aligned} \tag{6.192}$$

Clearly, Assumption (BP6) ensures that the (6.192) holds for $\epsilon < \tilde{\epsilon}$. This completes the proof. □

Theorem 6.7 depends heavily on Assumption (BP6), and the asymptotic breakdown point can be found as an implicit solution of ϵ that satisfies Assumption (BP6). This is not easy in practice. However, we can do better. We can find a closed-form expression for the asymptotic breakdown point albeit with further conditions. Next, we shall state a set of sufficient conditions for Assumption (BP6) in the spirit of Roy et al. (2023). To do this, firstly, we define $M_{g,f} = \sum \int g^k \phi'(f^k)$, and start with the following lemma that provides a lower (or, upper) bound for the ϕ-generated extended Bregman divergence. This lemma will be further used to provide a second version of the asymptotic breakdown point of $I_{D_\phi^{(k)}}$.

Lemma 6.6. *Assume that* $M_{f,f} \gtreqless M_{g,f}$, *where* $g = f_{X,Y}$ *and* $f = f_X f_Y$. *Then* $D_\phi^{(k)}(\epsilon g, f) \gtreqless D_\phi^{(k)}(\epsilon g, g)$ *for* $\epsilon^k \lesseqgtr 1 + \frac{\sum \int [\phi(g^k) - \phi(f^k)]}{M_{f,f} - M_{g,g}}$ *when* $g \neq f$.

Proof. Let us assume $M_{f,f} \geq M_{g,f}$. See that

$$
\begin{aligned}
& D_\phi^{(k)}(\epsilon g, f) - D_\phi^{(k)}(\epsilon g, g) \\
&= \sum \int \left[\phi((\epsilon g)^k) - \phi(f^k) - \left\{ (\epsilon g)^k - f^k \right\} \phi'(f^k) - \phi((\epsilon g)^k) + \phi(g^k) + \left\{ (\epsilon g)^k - g^k \right\} \phi'(g^k) \right] \\
&= - \sum \int \phi(f^k) - \epsilon^k M_{g,f} + M_{f,f} + \sum \int \phi(g^k) + (\epsilon^k - 1) M_{g,g} \\
&\geq \sum \int [\phi(g^k) - \phi(f^k)] - \epsilon^k M_{f,f} + M_{f,f} + (\epsilon^k - 1) M_{g,g} \\
&= \sum \int [\phi(g^k) - \phi(f^k)] - (\epsilon^k - 1) M_{f,f} + (\epsilon^k - 1) M_{g,g} \\
&= \sum \int [\phi(g^k) - \phi(f^k)] - (\epsilon^k - 1)(M_{f,f} - M_{g,g}) \geq 0
\end{aligned} \tag{6.193}
$$

when

$$
\epsilon^k \leq 1 + \frac{\sum \int [\phi(g^k) - \phi(f^k)]}{(M_{f,f} - M_{g,g})} \text{ for } g \neq f. \tag{6.194}
$$

The reverse inequality can be similarly proved when $M_{f,f} \leq M_{g,f}$. This completes the proof. □

When the divergence is defined as a limit for some tuning parameters, the same limit is taken over the appropriate assumptions, e.g., $M_{f,f} \leq M_{g,f}$ and the upper bound of ϵ as well. Now, we state some additional sufficient conditions for Assumption (BP6).

(BP7) Let the densities $g = f_{X,Y}$, $f = f_X f_Y$, $f^m = f_X^m f_Y^m$ and $k_m = K_{XY,m}$ satisfy

$$M_{f,f} \leq M_{g,f} \text{ and } \liminf_{m\to\infty} M_{f^m,f^m} \geq \limsup_{m\to\infty} M_{k_m,f^m}. \tag{6.195}$$

Also, X, Y are not independent for any joint distribution under consideration.

(BP8) For all $\epsilon < \tilde{\epsilon}$, it is true that

$$\begin{aligned}\liminf_{m\to\infty} D_\phi^{(k)}(\epsilon k_m, k_m) > \limsup_{m\to\infty} D_\phi^{(k)}(\epsilon k_m, 0) \\ + D_\phi^{(k)}((1-\epsilon)f_{X,Y}, f_{X,Y}) - D_\phi^{(k)}((1-\epsilon)f_{X,Y}, 0).\end{aligned} \tag{6.196}$$

In the next result, we shall see that the Assumptions (BP7) and (BP8) will together imply Assumption (BP6). Further, the bounds of the asymptotic breakdown point will be derived using Lemma 6.6. Let us define the following quantities:

$$\epsilon_1 = \left[1 + \limsup_{m\to\infty} \frac{\sum\int \left\{\phi[(k_{xy,m})^k] - \phi[(f_x^m f_y^m)^k]\right\}}{(M_{f^m,f^m} - M_{k_m,k_m})}\right]^{1/k}, \tag{6.197}$$

$$\epsilon_2 = 1 - \left[1 + \frac{\sum\int \left[\phi(f_{x,y}^k) - \phi(f_x^k f_y^k)\right]}{M_{f_X f_Y, f_X f_Y} - M_{f_{X,Y}, f_{X,Y}}}\right]^{1/k}. \tag{6.198}$$

Theorem 6.8. *Suppose the Assumptions (BP1) - (BP5) and (BP7) - (BP8) are true. Then the asymptotic breakdown point of* $I_{D_\phi^{(k)}}$ *is atleast* $\min\left\{\epsilon_1, \epsilon_2, \frac{1}{2}\right\}$.

Proof. An application of Lemma 6.6 in combination of Assumption (BP7) gives

$$D_\phi^{(k)}(\epsilon k_m, f_X^m f_Y^m) \geq D_\phi^{(k)}(\epsilon k_m, k_m) \text{ for } \epsilon \leq \left[1 + \frac{\sum\int\left[\phi[(k_{xy,m})^k] - \phi\left[(f_x^m f_y^m)^k\right]\right.}{(M_{f^m,f^m} - M_{k_m,k_m})}\right]^{1/k} \leq \epsilon_1. \tag{6.199}$$

Similarly, we also find that

$$D_\phi^{(k)}((1-\epsilon)f_{X,Y}, f_{X,Y}) \geq D_\phi^{(k)}((1-\epsilon)f_{X,Y}, f_X f_Y) \text{ for } \epsilon \leq 1 - \left[1 + \frac{\sum\int[\phi(f_{xy}^k) - \phi(f_x^k f_y^k)]}{M_{f_X f_Y, f_X f_Y} - M_{f_{X,Y}, f_{X,Y}}}\right]^{1/k} = \epsilon_2. \tag{6.200}$$

Combining these results, we further obtain that, for sufficiently large m,

$$\begin{aligned} D_\phi^{(k)}(\epsilon k_m, f_X^m f_Y^m) &\geq D_\phi^{(k)}(\epsilon k_m, k_m) \left[\text{ by (6.199) }\right] \\ &> D_\phi^{(k)}(\epsilon k_m, 0) + D_\phi^{(k)}((1-\epsilon)f_{X,Y}, f_{X,Y}) - D_\phi^{(k)}((1-\epsilon)f_{X,Y}, 0) \\ &\left[\text{ by Assumption (BP8) and then (6.200) }\right] \\ &\geq D_\phi^{(k)}(\epsilon k_m, 0) + D_\phi^{(k)}((1-\epsilon)f_{X,Y}, f_X f_Y) - D_\phi^{(k)}((1-\epsilon)f_{X,Y}, 0) \end{aligned} \tag{6.201}$$

when $\epsilon \leq \min\left\{\epsilon_1, \epsilon_2, \frac{1}{2}\right\}$. It is also clear from (6.201), that the Assumption (BP6) is also satisfied along the way under such a choice of ϵ. So the Assumptions (BP7) and (BP8) together work as a sufficient condition for Assumption (BP6), and therefore Theorem 6.7 is applicable. Hence, the asymptotic breakdown point is atleast $\min\left\{\epsilon_1, \epsilon_2, \frac{1}{2}\right\}$. □

6.5 Conclusions

This chapter defines generalized mutual information based on the class of ϕ-generated extended Bregman divergences. Using this idea we develop a class of two-sample non-parametric tests for the equality between two completely unstructured absolutely continuous distributions. These tests are consistent. Important theoretical properties such as asymptotics and robustness, of this class of tests, are studied rigorously in greater detail for this general class with the necessary detail. To our knowledge, some such properties have never been studied before in this context. These results may be extended in many different directions and can be further deduced for special families belonging to this class. The results obtained in this chapter may turn out to be quite helpful to the practitioners who might wish to apply them in two-sample testing in real-life situations.

This page is intentionally left blank.

Chapter 7

Example I: The Generalized S-Bregman Divergence

7.1 Introduction

In Chapter 6 we develop the general theory of a class of non-parametric tests using the extended Bregman divergence. This includes many important families of divergence measures such as the generalized S-Bregman divergence (Basak and Basu, 2022). As mentioned before, it is known that the GSB divergence contains many important sub-families, viz. power divergence, S-Hellinger distance, density power divergence, and Bregman exponential divergence, to name a few. In this chapter, we will discuss the general theory in the context of these special families along with numerical illustrations. The highlights of this chapter are the following.

(a) All the formulae specific to the GSB divergence are deduced from the general theory developed earlier. The effect of the tuning parameters on the tests will become prominent over time.

(b) The influence functions of the mutual information are plotted across different

combinations of the tuning parameters. All the pointers indicate that we gain robustness as the tuning parameter α increases. Moreover, some divergences still exist outside the power divergence family, producing tests more stable at the contamination of true distributions.

(c) Extensive simulation studies are presented under different setups. All these tests yield very high empirical powers except sometimes when $\beta < 0$. Nonetheless, this improves as the sample size increases. A comparative study shows that tests outside the power divergence family are generally more stable at higher contamination. Also, using the asymptotic distribution takes much less computational time than the permutation tests.

(d) A data-driven scheme for selecting tuning parameters is proposed.

(e) A couple of real data examples are presented.

7.2 Mutual Information based on the GSB Divergence

The generalized S-Bregman divergence, which motivates the formulation of the extended Bregman divergence, has already been introduced in Chapter 6. A form of the generalized mutual information may be obtained using the GSB divergence as in (6.10). The GSB divergence satisfies (P1) - (P4) of Proposition 6.1. However the ϕ-function as in (6.9) does not satisfy the conditions– $\phi(1) = \phi'(1) = 0$ of (P5). Therefore (P5) does not hold in this case. We import all the notations here from Chapter 6 unless otherwise specified. In a hybrid setup when X takes $0-1$ and Y is a continuous random variable,

the mutual information based on the GSB divergence becomes

$$I_{D^*}(X,Y) = D^*(f_{X,Y}, f_X f_Y)$$
$$= \sum_{x=0}^{1} \int_{f_y>0} \Big\{ e^{\beta(f_x f_y)^A} (\beta (f_x f_y)^A - \beta f_{x,y}^A - 1) + e^{\beta (f_{x,y})^A} + \frac{1}{B}(f_{x,y}^{A+B} - (f_x f_y)^{A+B}) \\ -(f_{x,y}^A - (f_x f_y)^A) \frac{A+B}{AB} (f_x f_y)^B \Big\} dy \text{ for } A, B \neq 0 \tag{7.1}$$

where $A = 1 + \lambda(1-\alpha)$ and $B = \alpha - \lambda(1-\alpha)$. If $A = 0$ but $B \neq 0$, the expression in (7.1) is defined as

$$\lim_{A \downarrow 0+} I_{D^*}(X,Y) = \sum_{x=0}^{1} \int_{f_y>0} \left\{ (f_x f_y)^{1+\alpha} \ln \left(\frac{f_x f_y}{f_{x,y}} \right) - \frac{(f_x f_y)^{1+\alpha} - f_{x,y}^{1+\alpha}}{1+\alpha} \right\} dy, \tag{7.2}$$

and similarly, it is given by

$$\lim_{B \to 0} I_{D^*}(X,Y) = \sum_{x=0}^{1} \int_{f_y>0} \left\{ e^{\beta (f_x f_y)^{1+\alpha}} \left(\beta (f_x f_y)^{1+\alpha} - \beta f_{x,y}^{1+\alpha} - 1 \right) + e^{\beta (f_{x,y})^{1+\alpha}} \right. \\ \left. + f_{x,y}^{1+\alpha} \ln \left(\frac{f_{x,y}}{f_x f_y} \right) - \frac{f_{x,y}^{1+\alpha} - (f_x f_y)^{1+\alpha}}{1+\alpha} \right\} dy \tag{7.3}$$

for $A \neq 0, B = 0$ and $\alpha > -1$. In the Equations (7.2) and (7.3), the case $\alpha = -1$ is similarly defined as $\alpha \to -1$. This limit turns out to be 0 in both cases. Substituting $\beta = 0$ in the above expressions gives the MI for the S-divergence. The MI based on the other subfamilies of the GSB divergence may be similarly obtained.

7.3 Asymptotic Results

Assume that the derivatives up to the fourth-order of the ϕ-function, that are given by

$$\phi'(t) = \beta e^{\beta t} + \Big(1 + \frac{B}{A}\Big)\frac{t^{\frac{B}{A}}}{B}, \tag{7.4}$$

$$\phi''(t) = \beta^2 e^{\beta t} + \Big(1 + \frac{B}{A}\Big)\frac{B}{A}\cdot\frac{t^{\frac{B}{A}-1}}{B}, \tag{7.5}$$

$$\phi'''(t) = \beta^3 e^{\beta t} + \Big(1 + \frac{B}{A}\Big)\cdot\frac{B}{A}\cdot\Big(\frac{B}{A}-1\Big)\frac{t^{\frac{B}{A}-2}}{B}, \tag{7.6}$$

$$\phi''''(t) = \beta^4 e^{\beta t} + \Big(1 + \frac{B}{A}\Big)\cdot\frac{B}{A}\cdot\Big(\frac{B}{A}-1\Big)\Big(\frac{B}{A}-2\Big)\frac{t^{\frac{B}{A}-3}}{B}, \tag{7.7}$$

are uniformly bounded by integrable functions. The boundedness assumption would invariably put some restrictions on the tuning parameters involved. Under the assumptions of Theorem 6.1, we get

$$nh_n^{1/2}\Big(\widehat{I}_{D_*} - \frac{\mu_*}{nh_n}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_*^2) \text{ as } n \to \infty \tag{7.8}$$

under the null hypothesis $\mathbb{H}$, where

$$\mu_* = \frac{1}{2}\int_u K^2(u)du \int_{f_y>0}\left[\sum_{x=0}^{1}(f_x f_y)^{2A-1}\Big[(A\beta)^2 e^{\beta(f_x f_y)^A} + (A+B)(f_x f_y)^{B-A}\Big](1-f_x)\right]dy, \tag{7.9}$$

$$\sigma_*^2 = \frac{1}{2}\int\left(\int K(u)K(u+z)dz\right)^2 du \times \int_{f_y>0}\left[\sum_{x=0}^{1}(f_x f_y)^{2A-1}\Big[(A\beta)^2 e^{\beta(f_x f_y)^A} + (A+B)(f_x f_y)^{B-A}\Big](1-f_x)\right]^2 dy. \tag{7.10}$$

Next, we deduce the asymptotic mean and variance for the following subfamilies.

Corollary 7.1. *(Power divergence family) Here $\beta = 0$, $A + B = 1$, so we get*

$$\mu_* = \frac{1}{2}\int_u K^2(u)du \int_{f_y>0} dy, \tag{7.11}$$

$$\sigma_*^2 = \frac{1}{2}\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du \int_{f_y>0} dy. \tag{7.12}$$

Corollary 7.2. *(S-divergence) Here $\beta = 0$, $A + B = 1 + \alpha$, so we get*

$$\mu_* = \frac{(1+\alpha)}{2}\int_u K^2(u)du\Big(f_{x_0}^{\alpha} f_{x_1} + f_{x_1}^{\alpha} f_{x_0}\Big) \int_{f_y>0} f_y^{\alpha} dy, \tag{7.13}$$

$$\sigma_*^2 = \frac{(1+\alpha)^2}{2}\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du\Big(f_{x_0}^{\alpha} f_{x_1} + f_{x_1}^{\alpha} f_{x_0}\Big)^2 \int_{f_y>0} f_y^{2\alpha} dy. \tag{7.14}$$

See that (7.13) and (7.14) do not depend on the tuning parameter λ. Therefore we get the same expressions of the asymptotic mean and variance for both the S-Hellinger and the density power divergence family.

Corollary 7.3. *(Squared L_2 distance) Here $\beta = 0$, $A + B = 2$, so we get*

$$\mu_* = 2f_{x_0} f_{x_1}\int_u K^2(u)du, \tag{7.15}$$

$$\sigma_*^2 = 8(f_{x_0} f_{x_1})^2\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du \int_{f_y>0} f_y^2 dy. \tag{7.16}$$

Corollary 7.4. *(BED) We know that $\beta \in \mathbb{R} \setminus \{0\}$ and $A + B = 0, A = 1$ give the scaled BED family. We can get the expressions of μ_* and σ_*^2 for the scaled BED family, and adjust them for the BED family by multiplying the asymptotic mean and asymptotic variance of the scaled BED family with the factors $\frac{2}{\beta^2}$ and $\frac{4}{\beta^4}$ respectively. This gives*

$$\mu_* = (f_{x_0} f_{x_1})\int_u K^2(u)du \int_{f_y>0}\Big(e^{\beta(f_{x_0} f_y)} + e^{\beta(f_{x_1} f_y)}\Big) f_y dy, \tag{7.17}$$

and

$$\sigma_*^2 = 2(f_{x_0} f_{x_1})^2 \int_u \left(\int K(z)K(z+u)dz \right)^2 du \int_{f_y>0} \left(e^{\beta(f_{x_0} f_y)} + e^{\beta(f_{x_1} f_y)} \right)^2 f_y^2 dy. \quad (7.18)$$

In this family, the degenerate case $\beta = 0$ is defined as continuous limit $\beta \to 0$.

The asymptotic mean and variance for the S-divergence family depend only on the tuning parameter $\alpha \in [0,1]$. Empirical evidence suggests that the asymptotic mean and variance decrease when α increases towards 1. The mutual information based on the S-divergence also decreases with the increment of α at a fixed λ. This is a crucial observation because perhaps, it partly explains why the test statistics are more stable against contamination at higher values of α. On the contrary, the robustness for the BED family increases when β decreases in the real line.

Further see that if $\alpha = 0$, the support of Y needs to be bounded otherwise the asymptotic mean and variance given in (7.9) and (7.10) would become infinite. This happens irrespective of the values the tuning parameters may take except when $\alpha = 0$. In these cases, we cannot use the asymptotic normality results. Therefore, a permutation algorithm should be invoked to calculate the empirical level and power. We observe the permutation test entails a heavy computational burden. The members of the GSB divergence family other than the cases with $\alpha = 0$ do not face this issue. This is the one, perhaps very useful, among many other advantages for advocating the use of the GSB divergence with an exception for $\alpha = 0$. This latter class includes the power divergence family. Sometimes the asymptotic normality result may require a large sample size to be useful in practical applications; this primarily happens in the case with $\beta < 0$.

Corollary 7.5. *MI based on the Itakura-Saito distance is obtained from extended Bregman divergence for* $\phi(t) = -\frac{\log(t)}{2\pi}$ *with* $k = 1$. *So*

$$\mu_* = \frac{1}{4\pi}\int_u K^2(u)du\left[\frac{f_{x_0}}{f_{x_1}} + \frac{f_{x_1}}{f_{x_0}}\right]\int_{f_y>0}\frac{dy}{f_y}, \tag{7.19}$$

$$\sigma_*^2 = \frac{1}{8\pi^2}\int_u\left(\int K(z)K(z+u)dz\right)^2 du\left[\frac{f_{x_1}}{f_{x_0}} + \frac{f_{x_0}}{f_{x_1}}\right]^2\int_{f_y>0}\frac{dy}{f_y^2}. \tag{7.20}$$

We will not present simulation results for this divergence measure.

It follows from Theorem 6.2 that the class of tests defined in (6.108) based on the GSB family are consistent. Let the contiguous alternatives be $\mathbb{K}_n$ as in (6.116). Also $\Delta_{x,y}, \Delta_x, \Delta_y$ are defined as before. Then it follows from Theorem 6.3 that

$$T_{D_*} \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma_*}\sum_{x=0}^{1}\int\left[(A\beta)^2 e^{\beta(f_xf_y)^A}(f_xf_y)^{2A} + (1+\alpha)(f_xf_y)^{1+\alpha}\right]\left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_xf_y}\right)^2 dy + \mathcal{N}(0,1) \tag{7.21}$$

as $n \to \infty$ under the contiguous alternatives $\mathbb{K}_n$. When $\alpha = -1$ and $\beta = 0$ the asymptotic distributions of T_{D^*} become identical under both $\mathbb{K}_n$ and $\mathbb{H}$.

Corollary 7.6. *(S-divergence) Here we have* $\beta = 0$, *therefore*

$$T_{D_*} \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma_*}\sum_{x=0}^{1}\int(1+\alpha)(f_xf_y)^{1+\alpha}\left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_xf_y}\right)^2 dy + \mathcal{N}(0,1) \tag{7.22}$$

as $n \to \infty$ *under* $\mathbb{K}_n$.

Corollary 7.7. *(BED) Substituting* $\alpha = -1$ *and transforming the* ϕ*-function by* $\frac{2}{\beta^2}$ *gives*

$$T_{D_*} \xrightarrow{\mathcal{L}} \frac{d^2}{\sigma_*} \sum_{x=0}^{1} \int e^{\beta(f_x f_y)} (f_x f_y)^2 \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2 dy + \mathcal{N}(0,1) \tag{7.23}$$

as $n \to \infty$ *under* $\mathbb{K}_n$.

7.4 Robustness Studies

This section will discuss the stability behaviour of I_{D_*}. First, we shall study its influence function. After that, the asymptotic breakdown point of I_{D_*} will be computed.

7.4.1 Influence Functions

It follows from Theorem 6.4 that the second-order influence function of I_{D_*} at a point $t_0 = (x_0, y_0)$ under the null hypothesis is given by

$$\mathcal{IF}_2(I_{D_*}, f_X f_Y, t_0) = \sum_{x=0}^{1} \int \left\{ A^2 \beta^2 e^{\beta (f_x f_y)^A} (f_x f_y)^{2A-1} + (1+\alpha)(f_x f_y)^{\alpha} \right\} U_{xy} dy, \tag{7.24}$$

where $\Delta_x = \delta_{x_0}(x) - f_x$, $\Delta_y = \delta_{y_0}(y) - f_y$, $\Delta_{x,y} = \delta_{x_0}(x)\delta_{y_0}(y) - f_x f_y$ and

$$U_{xy} = (f_x f_y) \left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y} \right)^2. \tag{7.25}$$

In particular, see that

$$\mathcal{IF}_2(I_{D_*}, f_X f_Y, t_0) = \begin{cases} \sum_{x=0}^{1} \int A^2 \beta^2 e^{\beta (f_x f_y)^A} (f_x f_y)^{2A-1} U_{xy} dy & \text{if } \alpha = -1, \\ (1+\alpha) \sum_{x=0}^{1} \int (f_x f_y)^{\alpha} U_{xy} dy & \text{if } \beta = 0. \end{cases} \tag{7.26}$$

Notice that when $\alpha = -1$ and $A = 1$, we obtain the expression for a scaled BED family which needs to be multiplied by $2/\beta^2$ (with $\beta \neq 0$) to get the $\mathcal{IF}_2$ for the BED family.

The stability of the influence function essentially depends on controlling the term U_{xy} inside the summation and integration in (7.24). Since f_y is a probability density function of a continuous random variable, its boundedness may be an issue. However, a sufficient condition such as the density of Y has a continuously bounded derivative ensures that its density becomes uniformly bounded. When f_y is uniformly bounded, U_{xy} can still be unbounded. However in most situations, we see that the factor $(f_x f_y)^\tau$ adds stability to U_{xy} for $\tau > 0$. The higher the τ is, the greater downweighting is achieved for the influence function at extreme outliers. From that observation, we see that the first term of $\mathcal{IF}_2$ in (7.24) becomes more stable for $2A - 1 > 0$, i.e, $A > \frac{1}{2}$. Similarly, the second term becomes more stable when $\alpha > 0$. We also note that $e^{\beta(f_x f_y)}$ always stays bounded for any finite β and bounded density f_y.

We study the influence function in the following regions as in Basak and Basu (2022).

(1) We know that $\beta = 0$ gives the S-divergence family. In addition, if $\alpha = 0$ we obtain the expression of $\mathcal{IF}_2$ for the power divergence family, which is independent of λ. From the earlier discussions and the expression in (7.26), we know that $\mathcal{IF}_2$ of the S-divergence family will be more stable only when $\alpha > 0$. It becomes quite unstable at $\alpha = 0$, i.e., for the power divergence family. This region is denoted by $\mathbb{S}_1 = \Big\{(\alpha, \lambda, \beta) : \alpha > 0, \lambda \in \mathbb{R}, \beta = 0\Big\}$.

(2) When $\beta \neq 0$ and $A = 0$, the first term in (7.24) drops out and the non-vanishing second term becomes more stable for $\alpha > 0$. However, $A = 0$ implies that $\lambda = \frac{1}{\alpha - 1}$, and with that the entire expression becomes independent of β. Under this setup, the allowable region turns out to be $\mathbb{S}_2 = \Big\{(\alpha, \lambda, \beta) : \alpha > 0, \lambda = \frac{1}{\alpha-1}, \beta \neq 0\Big\}$. Notice that when $\alpha < 1$ the tuning parameter $\lambda \in \mathbb{S}_2$ becomes negative.

(3) When $(A+B) = 0$, or $\alpha = -1$, the second term in (7.24) vanishes. Then the first term becomes more stable for $A > \frac{1}{2}$, i.e., $\lambda > -\frac{1}{4}$ and $\beta \neq 0$. In this case, $\mathcal{IF}_2$ will be more stable in $\mathbb{S}_3 = \left\{(\alpha, \lambda, \beta) : \alpha = -1, \lambda > -\frac{1}{4}, \beta \neq 0\right\}$.

(4) In the fourth case, all of $\beta, A, (A+B)$ should be non-zero subject to the constraint that $\tau > 0$ for $(f_x f_y)^\tau U_{xy}$. This yields the allowable region as $\mathbb{S}_4 = \left\{(\alpha, \lambda, \beta) : \alpha > 0, \lambda(1-\alpha) > -\frac{1}{2}, \beta \neq 0\right\}$.

Combining all these cases, we conclude that the infinitesimal effect of an outlier on the MI, as measured by $\mathcal{IF}_2$, is visibly downweighted when the tuning parameters belong to $\mathbb{S} := \cup_{i=1}^{4} \mathbb{S}_i$. Outside the region $\mathbb{S}$, the influence function becomes quite unstable. It is worth noticing that the well-known PD family comes out of this unstable region. To explore further, we discuss the following example in this context.

Example 7.1. *Let us generate* (X, Y) *such that* $X \sim Bernoulli(0.5)$ *and* $Y \sim \mathcal{N}(0,1)$. *The random variables* X, Y *are independently generated under the null hypothesis* $\mathbb{H}$. *In Figure 7.1, 7.2 and 7.3, we plot the second-order gross error sensitivity (GES) that is defined as* $GES_2 := \sup_y \mathcal{IF}_2(\cdots, (0, y))$ *fixing* $x = 0$. *For practical purposes, we have taken the supremum over* $[-20, 20]$ *which is quite a large interval for outliers to be added to the continuous distribution* $\mathcal{N}(0,1)$. *From Figure 7.1, it is clear that all members of the power divergence family* ($\alpha = \beta = 0$) *are equivalent in terms of second-order gross error sensitivity. Therefore outliers coming from the extreme region of the true continuous distribution, or even lying far outside the most probable region can make the tests based on the PD family very unstable. However, their robustness increases substantially when* $\alpha > 0$. *In other figures, we plot the stable GES curves when the tuning parameters belong to* $\mathbb{S}_i, i = 2, 3$, *and some unbounded curves outside those regions. In Figure 7.4 we plot the* $\mathcal{IF}_2$ *and some appropriately identified unbounded curves outside that specified region.*

Empirically we can suggest that the members outside the PD family generally exhibit

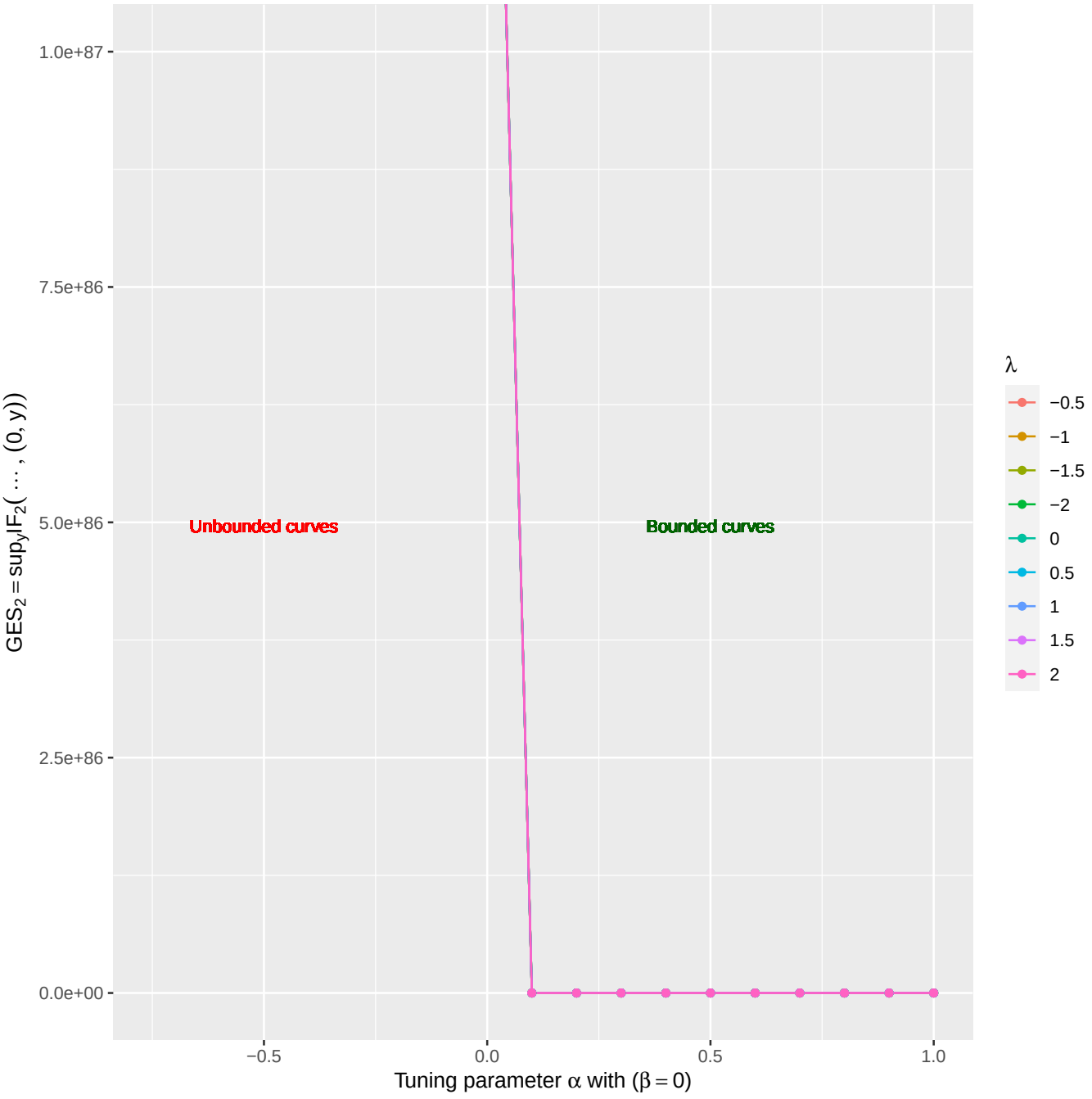

FIGURE 7.1: Stable GES curves when $(\alpha, \lambda, \beta) \in \$_1$, and unbounded curves as $\alpha \leq 0$.

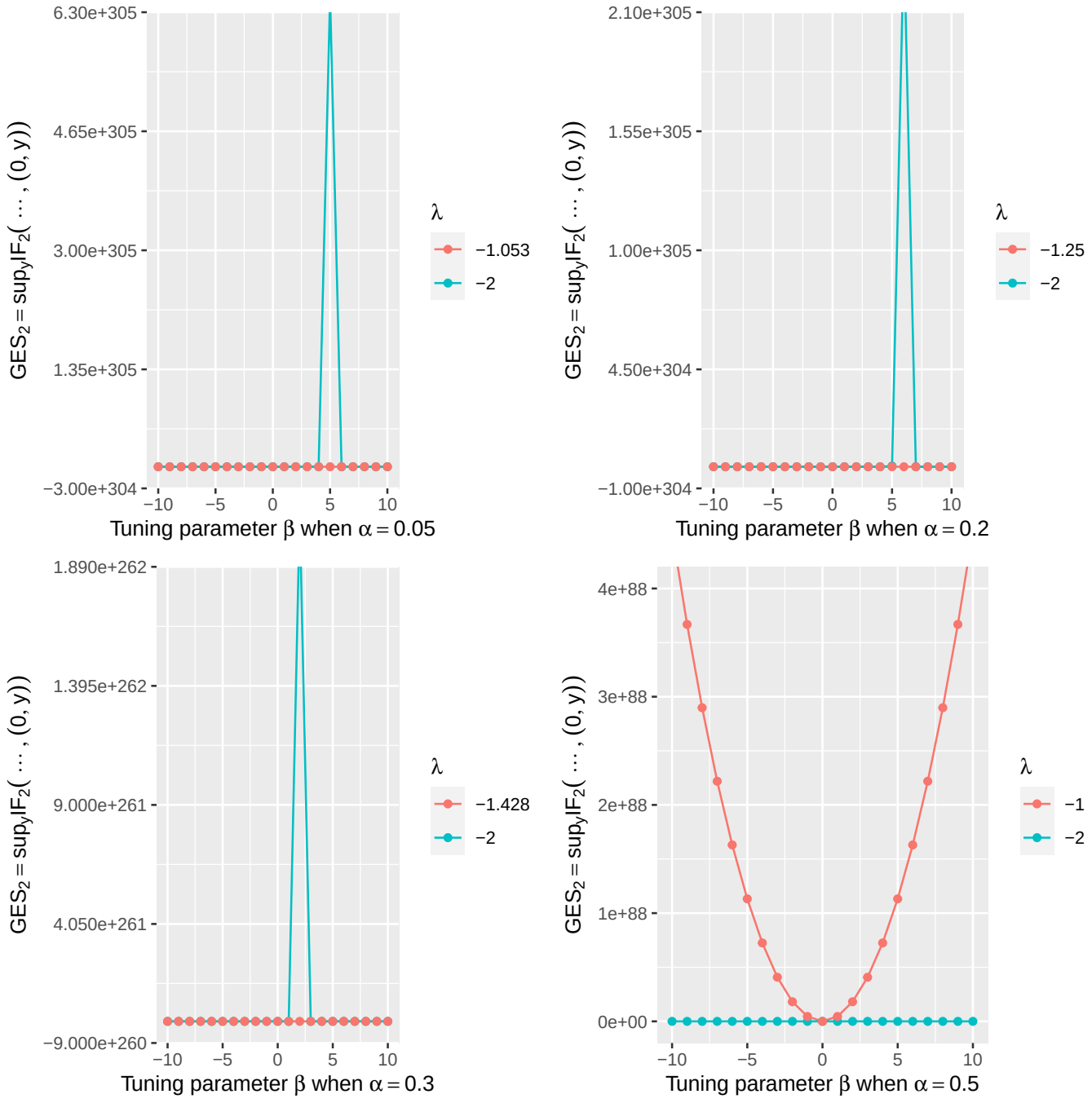

FIGURE 7.2: GES curves when $(\alpha, \lambda, \beta) \in \mathbb{S}_2$, and some curves as $\lambda \neq \frac{1}{\alpha-1}$.

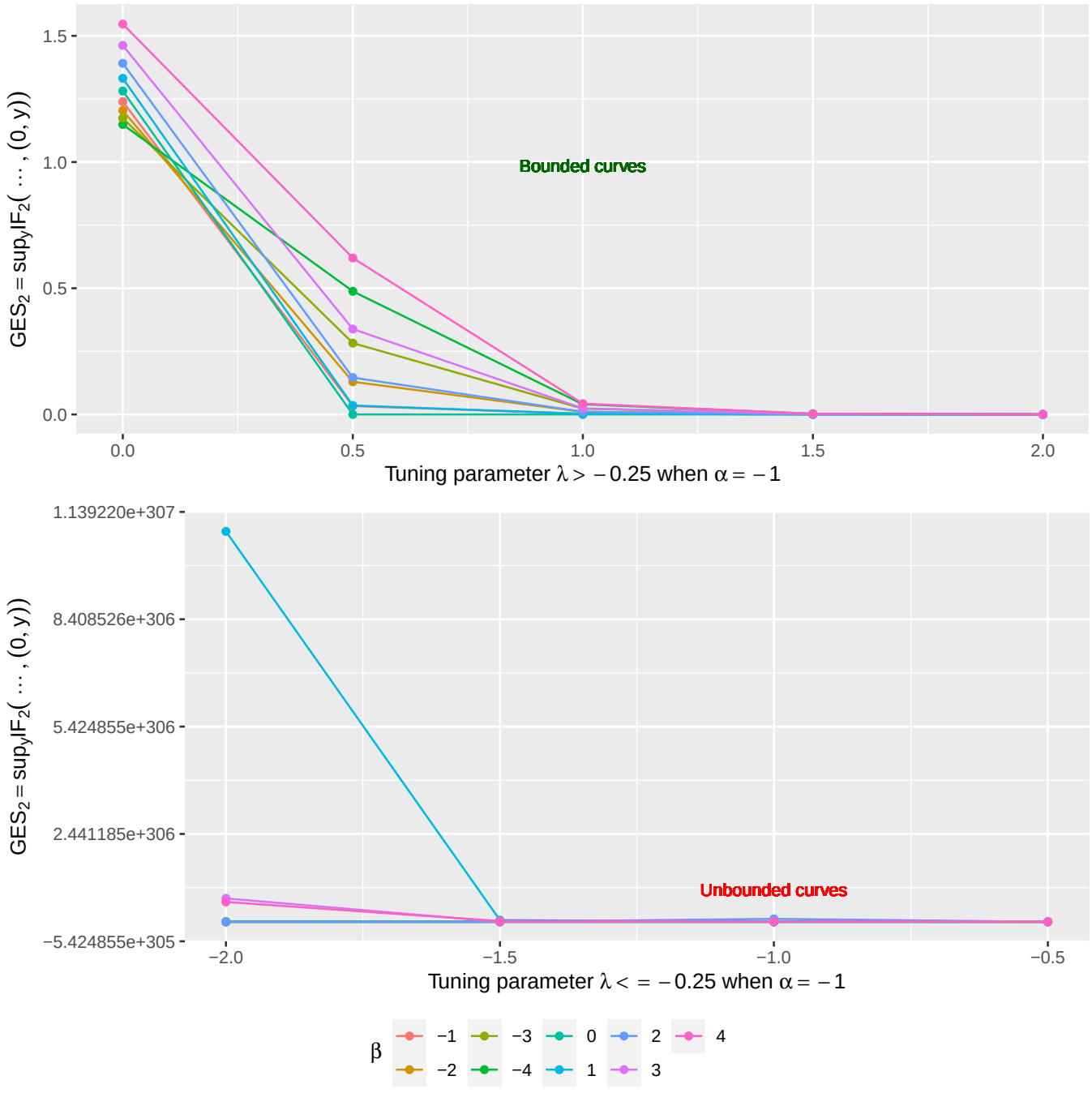

FIGURE 7.3: Bounded GES curves when $(\alpha, \lambda, \beta) \in \mathbb{S}_3$, and some unbounded curves as $\lambda \leq -0.25$.

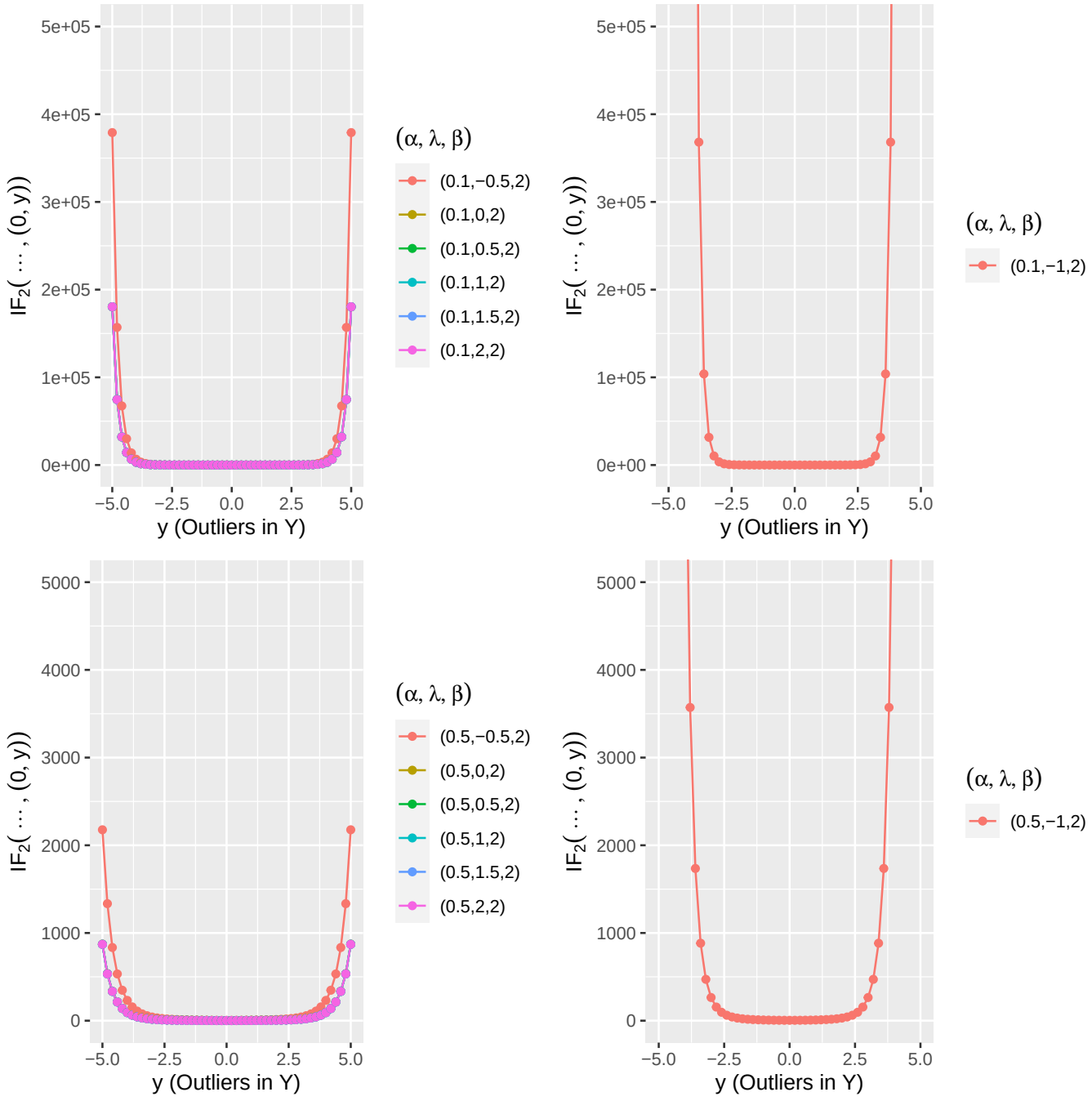

FIGURE 7.4: Bounded $\mathcal{IF}_2$ when $(\alpha, \lambda, \beta) \in \mathbb{S}_4$ in the left panel, and some unbounded curves as $\lambda(1-\alpha) \leq -0.5$ in the right panel.

better stability upon extreme contamination at a point mass. When $\lambda \in \mathbb{R}$, better stability is usually achieved for higher values of α and negative β. For completeness, we write down the expressions of $\mathcal{LIF}$ and $\mathcal{PIF}$ in long hand using Theorem 6.6 as follows.

(a) The power influence function is given by

$$\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi} \cdot \phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int \left\{A^2\beta^2 e^{\beta(f_x f_y)^A}(f_x f_y)^{2A-1} + (1+\alpha)(f_x f_y)^\alpha\right\} U_{xy} dy\right)$$
$$\times \sum_{x=0}^{1} \int (f_x f_y)\left(\frac{\Delta^*_{x,y}}{f_x f_y} - \frac{\Delta^*_x}{f_x} - \frac{\Delta^*_y}{f_y}\right)\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\left[A^2\beta^2 e^{\beta(f_x f_y)^A}(f_x f_y)^{2A-1} + \Big(1+\alpha\Big)(f_x f_y)^\alpha\right] dy. \tag{7.27}$$

(b) When $t_0 = t_1$, we get

$$\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi} \cdot \phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int \left\{A^2\beta^2 e^{\beta(f_x f_y)^A}(f_x f_y)^{2A-1} + (1+\alpha)(f_x f_y)^\alpha\right\} U_{xy} dy\right)$$
$$\times \sum_{x=0}^{1} \int \left[A^2\beta^2 e^{\beta(f_x f_y)^A}(f_x f_y)^{2A-1} + \Big(1+\alpha\Big)(f_x f_y)^\alpha\right] U_{xy} dy. \tag{7.28}$$

(c) The level influence function is given by

$$\mathcal{LIF}(\pi, t_0) \equiv 0. \tag{7.29}$$

It is clear that the stability of $\mathcal{PIF}$ is proportional to the influence function of I_{D_*} itself. See that $\mathcal{LIF}$ up to any order is always zero. This is, in fact, one limitation of the influence function, which cannot capture the robustness of the type-I error of this class of tests. However, we will see in the simulation studies the levels of the tests at the contaminated null are susceptible to the choice of tuning parameters.

7.4.2 Asymptotic Breakdown Point of I_{D_*}

Theorem 6.7 can be easily translated for the GSB divergence. Since it heavily depends on Assumption (BP6), we give a set of sufficient conditions that imply Assumption (BP6). In the view of Lemma 6.6 we find that $D_*(\epsilon g, f) \geq D_*(\epsilon g, g)$ for

$$\epsilon^A \leq 1 + \frac{\sum \int \left[AB(e^{\beta g^A} - e^{\beta f^A}) + A(g^{1+\alpha} - f^{1+\alpha}) \right]}{\sum \int \left[AB\{(\beta f^A)e^{\beta f^A} - (\beta g^A)e^{\beta g^A}\} + (1+\alpha)(f^{1+\alpha} - g^{1+\alpha}) \right]}, \tag{7.30}$$

provided it is true that

$$\sum \int \left[AB\beta e^{\beta f^A}(f^A - g^A) + (1+\alpha)(f^A - g^A)f^B \right] \geq 0 \text{ with } B > 0, \tag{7.31}$$

where $g = f_{X,Y}, f = f_X f_Y$, and X, Y are not independent. When the divergence is defined as a limit, all the conditions as in (7.30) and (7.31) will be similarly restated using the appropriate limit(s). If $A = 0$ the expression in (7.30) will become independent of ϵ, hence may be discarded. So, without loss of generality, we will implicitly assume that $A > 0$ without bothering about the degenerate case. The other degenerate case occurs when $B \downarrow 0+$. Further, if $B \downarrow 0+$ but $\alpha > -1, A > 0$, $\lim_{B\downarrow 0+} D_*(\epsilon g, f) \geq \lim_{B\downarrow 0+} D_*(\epsilon g, g)$ for $\epsilon = 0$, when $\sum \int (f^{1+\alpha} - g^{1+\alpha}) \geq 0$ and X, Y are not independent. In this case, the tuning parameters belong to a sufficiently small neighbourhood containing $\mathbb{S}_5 = \left\{ (\alpha, \frac{\alpha}{1-\alpha}, \beta) : \alpha > -1, \beta \in \mathbb{R} \right\}$. Similarly, $\alpha = -1$ and $B \downarrow 0+$ imply that $A \downarrow 0+$. Consequently, the expression in (7.30) will become independent of ϵ, hence will be discarded.

Corollary 7.8. *When $\beta = 0, \alpha \geq 0$, GSB divergence becomes the S-divergence. In that case $D_*(\epsilon g, f) \geq D_*(\epsilon g, g)$ for $\epsilon^A \leq \frac{B}{1+\alpha}$, provided*

$$\sum \int \left\{ f^{1+\alpha} - g^A f^B \right\} \geq 0 \text{ with } B > 0, \tag{7.32}$$

where $g = f_{X,Y}, f = f_X f_Y$ *and* X, Y *are not independent. If we assume that* $\sum \int f^{1+\alpha} \geq \sum \int g^{1+\alpha}$, *then (7.32) is implied by a simple application of Hőlder's inequality. The degenerate case of* $B \downarrow 0+$ *for the S-divergence may be similarly characterized as before.*

Corollary 7.9. *For the density power divergence,* $A = 1, B = \alpha$. *So* $D_*(\epsilon g, f) \geq D_*(\epsilon g, g)$ *for* $\epsilon \leq \frac{\alpha}{1+\alpha}$ *provided*

$$\sum \int \left\{ f^{1+\alpha} - g^{1+\alpha} \right\} \geq 0, \tag{7.33}$$

where $g = f_{X,Y}, f = f_X f_Y$ *and* X, Y *are not independent.*

Corollary 7.10. *When* $A = 1 + \lambda, B = -\lambda$, *it becomes the power divergence family. Since* $A > 0$, *i.e.,* $\lambda > -1$. *Combining the degenerate case* $\lambda = -1$, *we get* $D_*(\epsilon g, f) \geq D_*(\epsilon g, g)$ *when*

$$\epsilon \leq (-\lambda)^{\frac{1}{1+\lambda}} \text{ for } -1 \leq \lambda \leq 0, \tag{7.34}$$

such that X, Y *are not independent. Since the conditions as in Lemma 6.6 are trivially true, we do not need any further restrictions.*

Corollary 7.11. *When* $\alpha = -1, \lambda = 0$, *it becomes a scaled BED family with tuning parameter* β. *Here* $A = 1, B = -1$. *Suppose* $\beta \neq 0$ *then* $D_*(\epsilon g, f) \geq D_*(\epsilon f, f)$ *such that*

$$\epsilon \leq 1 + \frac{\sum \int \left(e^{\beta g} - e^{\beta f} \right)}{\beta \sum \int \left(f e^{\beta f} - g e^{\beta g} \right)} \text{ for } \beta \sum \int \left(e^{\beta f} (f - g) \right) \geq 0, \tag{7.35}$$

and X, Y *are not independent. When* $\beta = 0$, *Lemma 6.6 holds in limit as* $\beta \to 0$ *for* $\epsilon \leq \frac{1}{2}$ *such that* $\sum \int f^2 > \sum \int g^2$ *and* X, Y *are not independent.*

In Chapter 6 we also present a second version of the breakdown point for the generalized MI functional. In Theorem 6.8 we find that the asymptotic breakdown point of I_{D_*}

is at least $\min\{\epsilon_1, \epsilon_2, \frac{1}{2}\}$ where

$$\epsilon_1 = \left(1 + \limsup_{m\to\infty} \frac{\sum\int\left[AB\{e^{\beta k^A_{xy,m}} - e^{\beta(f^m_x f^m_y)^A}\} + A(k^{1+\alpha}_{xy,m} - (f^m_x f^m_y)^{1+\alpha})\right]}{\sum\int\left[AB\beta\{(f^m_x f^m_y)^A e^{\beta(f^m_x f^m_y)^A} - (k^A_{xy,m})e^{\beta k^A_{xy,m}}\} + (1+\alpha)((f^m_x f^m_y)^{1+\alpha} - k^{1+\alpha}_{xy,m})\right]}\right)^{1/A}, \tag{7.36}$$

$$\epsilon_2 = 1 - \left(1 + \frac{\sum\int\left[AB\{e^{\beta f^A_{x,y}} - e^{\beta(f_x f_y)^A}\} + A(f^{1+\alpha}_{x,y} - (f_x f_y)^{1+\alpha})\right]}{\sum\int\left[AB\{\beta(f_x f_y)^A e^{\beta(f_x f_y)^A} - (\beta f^A_{x,y})e^{\beta f^A_{x,y}}\} + (1+\alpha)((f_x f_y)^{1+\alpha} - f^{1+\alpha}_{x,y})\right]}\right)^{1/A}. \tag{7.37}$$

As it turns out the result of the asymptotic breakdown point of the GSB divergence family is valid for $A > 0$ and $B > 0$, i.e., $-\frac{1}{1-\alpha} \le \lambda \le \frac{1}{1-\alpha}$ and $\alpha \ge -1$.

Corollary 7.12. *For the S-divergence family $\beta = 0$, and*

$$\epsilon_1 = \left(\frac{B}{1+\alpha}\right)^{1/A} \text{ and } \epsilon_2 = 1 - \left(\frac{B}{1+\alpha}\right)^{1/A}. \tag{7.38}$$

So the asymptotic breakdown point of I_{D_*} will be

$$\min\left\{\left(\frac{B}{1+\alpha}\right)^{1/A}, 1 - \left(\frac{B}{1+\alpha}\right)^{1/A}, \frac{1}{2}\right\} \text{ with } A > 0, B > 0, \tag{7.39}$$

under the assumptions of Theorem 6.8.

Corollary 7.13. *Next, we consider the power divergence family, i.e., $\alpha = \beta = 0$. Then under the assumptions of Theorem 6.8, the asymptotic breakdown point of I_{D_*} will be*

$$\min\left\{(-\lambda)^{1/(1+\lambda)}, 1 - (-\lambda)^{1/(1+\lambda)}, \frac{1}{2}\right\} \text{ for } -1 \le \lambda \le 0. \tag{7.40}$$

Corollary 7.14. *When $\lambda = \beta = 0$ it becomes the DPD. Then under the assumptions of Theorem 6.8 the asymptotic breakdown point of I_{D_*} becomes*

$$\min\left\{\frac{\alpha}{1+\alpha}, \frac{1}{1+\alpha}, \frac{1}{2}\right\}. \tag{7.41}$$

It is evident that as α increases, the asymptotic breakdown point of the generalized MI based on the DPD increases considerably while increasing λ should have the opposite effect of reducing the asymptotic breakdown point. Figure 1 of Roy et al. (2023) depicts a wide range of (α, λ) where I_{D_*} based on the S-divergence would be highly robust. Observe that if $\beta \neq 0$, the asymptotic breakdown point of I_{D_*} will depend on both the densities of Y and its contaminating sequences. Hence it is hard to characterize the tuning parameters producing highly asymptotic breakdown points, in general situations, unless specific examples are considered. The important thing to note here is that the asymptotic breakdown points do not depend on the dimension of Y.

7.5 Numerical Studies

7.5.1 Simulation Results

(A) In this section, we report the simulation results for the following sets of models.

Model 0 : (Null Model) $Y_0 \sim \mathcal{N}(0,1)$,

Model 1 : $Y_1 \sim \mathcal{N}(0,(1+a)^2)$ with $a = 0.75$,

Model 2 : $Y_1 \sim (1-a)\mathcal{N}(-1,1) + a\mathcal{N}(1,1)$ with $a = 0.6$.

The hypotheses of our interest are given in (6.28).

(B) To study the robustness of type-I error of the tests based on the GSB divergence the null model is contaminated as $(1-\epsilon)\mathcal{N}(0,1)+\epsilon\mathcal{N}(5,2)$ with $\epsilon = 0.05, 0.10, 0.12, 0.15$ when the second sample comes from $\mathcal{N}(0,1)$.

Samples of sizes $n_0 = 100$ and $n_1 = 100$ are drawn from the null and alternatives models, respectively, over 500 replications. The tests are conducted at 5% nominal level of significance with the tuning parameters chosen as $\alpha = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)^T$, $\beta = (0, -0.05)^T$ and $\lambda = (-0.5, -0.3, -0.2, -0.1, 0, 0.25, 0.50, 1)^T$.

Since the distributions in this setup have unbounded supports, we cannot use the asymptotic null distribution to simulate level and powers when $\alpha = 0$. Therefore we need to use the permutation test using the algorithm of Guha et al. (2021). In these cases, we consider 500 permutations. However, the unbounded support of the continuous distribution should not cause a problem in using the asymptotic null distribution for other members, i.e., outside $\alpha = 0$, of the generalized S-Bregman divergence family. Using the asymptotic null distribution saves a lot of computational burden. Throughout the numerical studies, the kernels and the bandwidth sequence are chosen as

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\left\{|u| \le 1\right\} \text{ and } h_n = 1.06 \times sd(Y) n^{-1/5}, \tag{7.42}$$

where $n = n_0 + n_1$ being the combined sample size. This kernel is called the Epanechnikov kernel and it satisfies Assumption (A3) with $\|K\| = \frac{3}{4}$. As it is well-known that kernel density estimates are generally robust to the choice of kernels, it is sometimes useful to use bounded and symmetric kernels. The optimum choice of bandwidth sequence for kernel density estimates depends on the problem at hand. However a *rule-of-thumb* optimal bandwidth sequence h_n considered here is best suited when the original distribution is Gaussian or symmetric. Also, it works fairly well even if the distribution is not heavily skewed. See Silverman (2018) for more details about such discussions.

In Table 7.1 and Table 7.2 the observed levels of the tests under pure models are reported. We see that when $\beta = -0.05$, sometimes the tests become conservative in the sense that they produce very low observed levels. The null hypothesis $\mathbb{H}$ is rejected if the first sample comes from Model 0 but the second sample comes from Model 1 or Model 2. The observed proportions of rejections (i.e., empirical power) in such cases are reported through Table 7.3 to Table 7.6. Sometimes we observe that the empirical powers are comparatively low for $\beta < 0$, which may be due to the slow rate of convergence of distributions. However, as we increase the sample sizes, the empirical powers improve. When the null model is contaminated, the observed levels increase along with the amount of contamination. However, all members of the GSB divergence family except $\alpha = 0$ exhibit stale type-I errors. These values are reported in Table 7.7 to Table 7.14.

TABLE 7.1: Proportion of Rejections when both the samples are generated through Model 0 and $\beta = 0$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	0.058	0.054	0.040	0.034	0.034	0.036	0.032	0.026	0.026	0.022	0.024
−0.3	0.058	0.040	0.034	0.034	0.032	0.034	0.030	0.026	0.026	0.022	0.024
−0.2	0.058	0.036	0.030	0.032	0.032	0.034	0.030	0.026	0.026	0.022	0.024
−0.1	0.058	0.034	0.030	0.032	0.032	0.034	0.030	0.026	0.026	0.022	0.024
0.0	0.040	0.036	0.030	0.030	0.032	0.032	0.030	0.026	0.024	0.022	0.024
0.25	0.042	0.098	0.096	0.028	0.032	0.030	0.028	0.026	0.024	0.022	0.024
0.50	0.042	0.092	0.084	0.078	0.032	0.026	0.028	0.026	0.022	0.022	0.024
1.0	0.042	0.092	0.082	0.076	0.076	0.098	0.024	0.026	0.022	0.022	0.024

TABLE 7.2: Proportion of Rejections when both the samples are generated through Model 0 and $\beta = -0.05$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	0.028	0.054	0.040	0.034	0.034	0.036	0.032	0.026	0.026	0.022	0.024
−0.3	0.018	0.012	0.008	0.008	0.006	0.004	0.004	0.004	0.004	0.004	0.024
−0.2	0.012	0.008	0.008	0.006	0.004	0.004	0.004	0.004	0.004	0.004	0.024
−0.1	0.012	0.008	0.006	0.006	0.004	0.004	0.004	0.004	0.004	0.004	0.024
0.0	0.036	0.010	0.006	0.006	0.004	0.004	0.004	0.004	0.004	0.004	0.024
0.25	0.034	0.036	0.042	0.006	0.004	0.004	0.004	0.004	0.004	0.004	0.024
0.50	0.034	0.036	0.036	0.040	0.004	0.004	0.004	0.004	0.004	0.004	0.024
1.0	0.054	0.040	0.036	0.036	0.036	0.052	0.004	0.004	0.004	0.004	0.024

TABLE 7.3: Proportion of Rejections when the first and second samples are respectively generated through Model 0 and Model 1, and $\beta = 0$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.992	0.980	0.958	0.936	0.914	0.886	0.862	0.846	0.822	0.806
−0.3	1	0.986	0.966	0.946	0.928	0.904	0.886	0.858	0.838	0.822	0.806
−0.2	1	0.980	0.960	0.942	0.924	0.898	0.878	0.858	0.838	0.822	0.806
−0.1	1	0.978	0.954	0.938	0.924	0.896	0.874	0.858	0.838	0.822	0.806
0.0	1	0.984	0.948	0.932	0.918	0.894	0.874	0.854	0.838	0.822	0.806
0.25	1	0.986	0.964	0.932	0.912	0.888	0.866	0.848	0.836	0.822	0.806
0.50	1	0.982	0.974	0.964	0.910	0.884	0.864	0.846	0.832	0.822	0.806
1.0	1	0.986	0.968	0.960	0.942	0.964	0.860	0.842	0.830	0.816	0.806

TABLE 7.4: Proportion of Rejections when the first and second sample are respectively generated through Model 0 and Model 1, and $\beta = -0.05$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.992	0.980	0.958	0.936	0.914	0.886	0.862	0.846	0.822	0.806
−0.3	1	0.952	0.924	0.902	0.866	0.840	0.792	0.768	0.736	0.708	0.806
−0.2	1	0.944	0.922	0.892	0.860	0.832	0.792	0.764	0.736	0.708	0.806
−0.1	1	0.940	0.912	0.888	0.856	0.826	0.792	0.764	0.734	0.708	0.806
0.0	1	0.948	0.906	0.880	0.854	0.818	0.790	0.762	0.734	0.708	0.806
0.25	1	0.958	0.938	0.870	0.840	0.802	0.786	0.750	0.732	0.708	0.806
0.50	1	0.950	0.942	0.926	0.840	0.792	0.772	0.744	0.732	0.706	0.806
1.0	1	0.950	0.940	0.924	0.910	0.942	0.762	0.740	0.724	0.702	0.806

TABLE 7.5: Proportion of Rejections when the first and second samples are respectively generated through Model 0 and Model 2, and $\beta = 0$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.868	0.822	0.780	0.720	0.674	0.634	0.604	0.586	0.568	0.544
−0.3	1	0.830	0.796	0.754	0.694	0.668	0.628	0.600	0.586	0.562	0.544
−0.2	1	0.820	0.786	0.738	0.690	0.664	0.628	0.600	0.586	0.562	0.544
−0.1	1	0.816	0.774	0.724	0.680	0.652	0.622	0.596	0.586	0.562	0.544
0.0	0.962	0.882	0.766	0.718	0.672	0.650	0.618	0.592	0.584	0.562	0.544
0.25	0.980	0.914	0.898	0.710	0.666	0.638	0.608	0.590	0.580	0.562	0.544
0.50	0.980	0.910	0.896	0.886	0.674	0.630	0.604	0.584	0.578	0.562	0.544
1.0	0.980	0.914	0.896	0.880	0.864	0.900	0.594	0.582	0.572	0.560	0.544

TABLE 7.6: Proportion of Rejections when the first and second samples are respectively generated through Model 0 and Model 2, and $\beta = -0.05$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.868	0.822	0.780	0.720	0.674	0.634	0.604	0.586	0.568	0.544
−0.3	1	0.614	0.552	0.500	0.454	0.420	0.400	0.382	0.372	0.360	0.544
−0.2	1	0.598	0.532	0.488	0.442	0.414	0.398	0.380	0.372	0.358	0.544
−0.1	1	0.586	0.516	0.476	0.432	0.414	0.398	0.376	0.372	0.356	0.544
0.0	0.962	0.594	0.506	0.462	0.426	0.410	0.396	0.376	0.370	0.356	0.544
0.25	0.980	0.796	0.756	0.450	0.416	0.394	0.382	0.374	0.366	0.354	0.544
0.50	0.980	0.780	0.760	0.724	0.422	0.392	0.374	0.368	0.366	0.354	0.544
1.0	0.980	0.778	0.746	0.718	0.694	0.764	0.374	0.366	0.364	0.350	0.544

TABLE 7.7: Proportion of Rejections under the null hypothesis with $\beta = 0$ when 5% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.148	0.094	0.078	0.056	0.048	0.044	0.040	0.038	0.036	0.038
−0.3	1	0.094	0.076	0.058	0.052	0.046	0.040	0.040	0.038	0.036	0.038
−0.2	1	0.084	0.072	0.056	0.048	0.046	0.040	0.040	0.038	0.036	0.038
−0.1	1	0.084	0.070	0.056	0.048	0.046	0.040	0.040	0.038	0.036	0.038
0.0	0.996	0.124	0.064	0.054	0.046	0.044	0.040	0.036	0.036	0.036	0.038
0.25	0.998	0.226	0.222	0.056	0.046	0.044	0.038	0.036	0.036	0.036	0.038
0.50	0.998	0.208	0.184	0.164	0.048	0.044	0.038	0.036	0.036	0.036	0.038
1.0	0.998	0.260	0.188	0.162	0.146	0.258	0.038	0.036	0.036	0.036	0.038

TABLE 7.8: Proportion of Rejections under the null hypothesis with $\beta = -0.05$ when 5% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.148	0.094	0.078	0.056	0.048	0.044	0.040	0.038	0.036	0.038
−0.3	1	0.024	0.016	0.014	0.014	0.008	0.006	0.004	0.004	0.004	0.038
−0.2	1	0.020	0.016	0.014	0.010	0.008	0.006	0.004	0.004	0.004	0.038
−0.1	1	0.018	0.014	0.012	0.010	0.008	0.006	0.004	0.004	0.004	0.038
0.0	0.996	0.024	0.012	0.012	0.010	0.008	0.006	0.004	0.004	0.004	0.038
0.25	0.998	0.074	0.102	0.012	0.010	0.008	0.004	0.004	0.004	0.004	0.038
0.50	0.998	0.070	0.062	0.060	0.010	0.008	0.004	0.004	0.004	0.004	0.038
1.0	0.998	0.086	0.060	0.058	0.056	0.140	0.004	0.004	0.004	0.004	0.038

TABLE 7.9: Proportion of Rejections under the null hypothesis with $\beta = 0$ when 10% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.770	0.510	0.308	0.214	0.148	0.122	0.106	0.100	0.092	0.088
−0.3	1	0.578	0.390	0.250	0.180	0.136	0.120	0.106	0.100	0.092	0.088
−0.2	1	0.522	0.338	0.236	0.168	0.134	0.118	0.106	0.100	0.092	0.088
−0.1	1	0.488	0.310	0.220	0.162	0.132	0.116	0.106	0.100	0.092	0.088
0.0	1	0.522	0.300	0.212	0.146	0.130	0.116	0.104	0.098	0.092	0.088
0.25	1	0.664	0.402	0.220	0.146	0.122	0.114	0.104	0.098	0.092	0.088
0.50	1	0.632	0.520	0.428	0.158	0.120	0.110	0.104	0.096	0.092	0.088
1.0	1	0.668	0.522	0.424	0.336	0.502	0.108	0.100	0.096	0.090	0.088

TABLE 7.10: Proportion of Rejections under the null hypothesis with β = −0.05 when 10% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.770	0.510	0.308	0.214	0.148	0.122	0.106	0.100	0.092	0.088
−0.3	1	0.214	0.130	0.090	0.064	0.054	0.042	0.036	0.036	0.032	0.088
−0.2	1	0.182	0.114	0.084	0.060	0.052	0.040	0.036	0.036	0.032	0.088
−0.1	1	0.172	0.100	0.076	0.058	0.048	0.040	0.036	0.036	0.032	0.088
0.0	1	0.198	0.096	0.066	0.058	0.048	0.040	0.036	0.036	0.032	0.088
0.25	1	0.356	0.246	0.070	0.054	0.046	0.040	0.036	0.036	0.032	0.088
0.50	1	0.316	0.248	0.204	0.054	0.040	0.038	0.036	0.034	0.032	0.088
1.0	1	0.354	0.254	0.196	0.156	0.308	0.036	0.034	0.034	0.032	0.088

TABLE 7.11: Proportion of Rejections under the null hypothesis with β = 0 when 12% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.932	0.792	0.540	0.360	0.260	0.206	0.168	0.148	0.134	0.128
−0.3	1	0.848	0.648	0.434	0.300	0.236	0.194	0.164	0.146	0.134	0.128
−0.2	1	0.792	0.588	0.408	0.292	0.230	0.192	0.164	0.142	0.134	0.128
−0.1	1	0.750	0.540	0.392	0.280	0.228	0.184	0.164	0.142	0.134	0.128
0.0	1	0.754	0.524	0.364	0.268	0.226	0.184	0.162	0.142	0.134	0.128
0.25	1	0.882	0.516	0.350	0.246	0.210	0.182	0.160	0.140	0.134	0.128
0.50	1	0.858	0.756	0.650	0.260	0.200	0.176	0.160	0.140	0.134	0.128
1.0	1	0.844	0.748	0.618	0.506	0.606	0.170	0.154	0.140	0.134	0.128

TABLE 7.12: Proportion of Rejections under the null hypothesis with β = −0.05 when 12% obs. of the first sample come from $\mathcal{N}(5,2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.932	0.792	0.540	0.360	0.260	0.206	0.168	0.148	0.134	0.128
−0.3	1	0.492	0.290	0.168	0.124	0.094	0.076	0.070	0.068	0.064	0.128
−0.2	1	0.414	0.236	0.156	0.118	0.094	0.074	0.070	0.066	0.064	0.128
−0.1	1	0.364	0.214	0.150	0.114	0.092	0.072	0.070	0.066	0.064	0.128
0.0	1	0.392	0.200	0.140	0.104	0.090	0.072	0.070	0.066	0.064	0.128
0.25	1	0.610	0.318	0.130	0.096	0.088	0.070	0.068	0.066	0.062	0.128
0.50	1	0.546	0.428	0.332	0.104	0.076	0.066	0.066	0.066	0.062	0.128
1.0	1	0.582	0.430	0.318	0.262	0.446	0.068	0.066	0.066	0.062	0.128

TABLE 7.13: Proportion of Rejections under the null hypothesis with $\beta = 0$ when 15% obs. of the first sample come from $\mathcal{N}(5, 2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.994	0.972	0.892	0.710	0.530	0.394	0.300	0.254	0.234	0.232
−0.3	1	0.978	0.934	0.802	0.626	0.472	0.350	0.290	0.252	0.234	0.232
−0.2	1	0.972	0.906	0.778	0.600	0.446	0.344	0.284	0.252	0.234	0.232
−0.1	1	0.962	0.880	0.740	0.590	0.436	0.334	0.280	0.250	0.234	0.232
0.0	1	0.954	0.854	0.708	0.560	0.424	0.334	0.276	0.248	0.234	0.232
0.25	1	0.980	0.652	0.654	0.522	0.404	0.314	0.272	0.244	0.234	0.232
0.50	1	0.974	0.954	0.906	0.498	0.384	0.306	0.268	0.240	0.234	0.232
1.0	1	0.974	0.952	0.892	0.796	0.764	0.296	0.266	0.240	0.234	0.232

TABLE 7.14: Proportion of Rejections under the null hypothesis with $\beta = -0.05$ when 15% obs. of the first sample come from $\mathcal{N}(5, 2)$

λ \ α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
−0.5	1	0.994	0.972	0.892	0.710	0.530	0.394	0.300	0.254	0.234	0.232
−0.3	1	0.872	0.694	0.456	0.288	0.216	0.184	0.158	0.142	0.132	0.232
−0.2	1	0.810	0.604	0.388	0.262	0.208	0.182	0.158	0.142	0.132	0.232
−0.1	1	0.760	0.554	0.352	0.248	0.200	0.180	0.154	0.142	0.132	0.232
0.0	1	0.742	0.506	0.332	0.232	0.194	0.174	0.152	0.142	0.132	0.232
0.25	1	0.910	0.472	0.290	0.220	0.186	0.172	0.148	0.140	0.132	0.232
0.50	1	0.886	0.776	0.662	0.216	0.180	0.166	0.148	0.138	0.132	0.232
1.0	1	0.858	0.758	0.618	0.538	0.608	0.166	0.146	0.134	0.132	0.232

7.5.2 Tuning Parameter Selection

In the simulation studies, we have seen that the choice of tuning parameters plays an important role in the level and power of the tests. It is therefore required to have a data-driven scheme for optimum tuning parameter selection. To the best of our knowledge, no such satisfactory theories exist in the context of hypothesis testing. Here we propose an algorithm for optimum tuning parameter selection. We start with the test function

that is given by

$$\psi(Y) = \begin{cases} 1 & \text{if } \widehat{T}_{D_*} > \tau_c, \\ 0 & \text{otherwise} \end{cases} \tag{7.43}$$

where $Y = (Y_0, Y_1)$ being the combined sample. The $0-1$ loss function for the *decision rule* $\psi(\cdot)$ is defined as

$$L(I_{D_*}, \psi(Y)) = \begin{cases} 1 & \text{if true } I_{D_*} = 0 \text{ and } \widehat{T}_{D_*} > \tau_c, \\ 1 & \text{if true } I_{D_*} > 0 \text{ and } \widehat{T}_{D_*} \leq \tau_c, \\ 0 & \text{otherwise.} \end{cases} \tag{7.44}$$

The risk function associated with the loss function is given by

$$R(I_{D_*}, \psi) = \mathbb{E}_Y\Big[L(I_{D_*}, \psi(Y))\Big] = \begin{cases} \mathbb{P}\Big\{\widehat{T}_{D_*} > \tau_c\Big\} & \text{when true } I_{D_*} = 0, \\ \\ 1 - \mathbb{P}\Big\{\widehat{T}_{D_*} > \tau_c\Big\} & \text{when true } I_{D_*} > 0. \end{cases} \tag{7.45}$$

The risk function $R(I_{D_*}, \psi)$ plays a similar role in the hypothesis testing problems as the *mean squared error* (MSE) do in the context of estimation. In the view of the decision theory, an optimum (*admissible*) set of tuning parameters would be $(\alpha_*, \lambda_*, \beta_*)$ if

$$(\alpha_*, \lambda_*, \beta_*) := \arg\min_{\alpha,\lambda,\beta} R(I_{D_*}, \psi) \text{ for all true } I_{D_*}. \tag{7.46}$$

Notice that the true value of I_{D_*} is unknown to us, also the optimum tuning parameters may not be unique. For multiple optimizers, all the tests are said to be *risk equivalent*. As a function of tuning parameters (α, λ, β), the risk function may be thought of as a map such that $R(I_{D_*}, \psi) : [-1, 1] \times \mathbb{R}^2 \mapsto [0, 1]$ whose minimizer may not always exist.

Given a data set, we need a *suitable* estimate of the risk function defined in (7.45). The probability is estimated using re-samples from the combined data Y. The acceptance or rejection of the null hypothesis may be decided by p-value using *true tuning parameters* which are also unknown to us. To do that we need to start with some *robust pilot tuning parameters*. Taking a cue from the simulation studies, we know that a robust pilot (α, λ, β) may lead to reasonably low type-I and type-II errors even if the data set is pure or contaminated. The algorithm for tuning parameter selection is stated as follows.

Algorithm 7.1.

Step 1 *The first sample Y_0 of size n_0 is combined with the second sample Y_1 of size n_1 as $Y = (Y_0, Y_1)$. Define a $0-1$ valued dummy variable X as in (6.27).*

Step 2 *Take a robust pilot $(\alpha_1, \lambda_1, \beta_1)$, and compute the p-value (p_1) using Theorem 6.3.2. The pilot is chosen such that the p-value should be consistent with the data.*

Step 3 *Reject the null hypothesis $\mathbb{H}$ at $100c\%$ nominal level if $p_1 \leq c$ or accept otherwise.*

Step 4 *If $p_1 \leq c$ draw independent random resamples Y_{0b} and Y_{1b} respectively from Y_0 and Y_1; otherwise draw both Y_{0b} and Y_{1b} from $Y := (Y_0, Y_1)$ where $b = 1, 2, \ldots, B$. Using the resamples, the test statistics are computed as $T^{(b)}_{D_*}, b = 1, \ldots, B$.*

Step 5 *Calculate*

$$\hat{P} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{ T^{(b)}_{D_*} > \tau_c \right\}, \tag{7.47}$$

where τ_c being the $100(1-c)\%$ point of $\mathcal{N}(0,1)$.

Step 6 *Define*

$$\widehat{R}(I_{D_*}, \psi) = \begin{cases} \hat{P} & \text{if } p_1 > c, \\ 1 - \hat{P} & \text{if } p_1 \leq c. \end{cases} \tag{7.48}$$

Step 7 *Compute the optimum tuning parameters as* $(\alpha_2, \lambda_2, \beta_2) := \arg\min_{\alpha,\lambda,\beta} \widehat{R}(I_{D_*}, \psi)$.

It may be easy to see that when the combined sample Y is fixed

$$\widehat{R}(I_{D_*}, \psi) \xrightarrow{\mathbb{P}} R(I_{D_*}, \psi) \text{ as } B \to \infty \text{ at a true pilot } (\alpha_1, \lambda_1, \beta_1). \tag{7.49}$$

So $\widehat{R}(I_{D_*}, \psi)$ is a good proxy for $R(I_{D_*}, \psi)$. An empirical version of the risk function $R(I_{D_*}, \psi)$ is minimized over a fine grid of tuning parameters as a function of suitable robust pilot tuning parameters. Though Algorithm 7.1 depends heavily on the choice of robust pilot tuning parameters, nevertheless it will not cause a serious problem if they lead to a reasonable optimum set of parameters, consistent with the data set itself.

7.5.3 Real Data Examples

Here we take up two real data examples and consider the problem of choosing 'optimal' tuning parameters using Algorithm 7.1. Here the primary aim is to show that Algorithm 7.1 leads to a set of tuning parameters that are at least as good as the power divergence family, if not better. The number of resamples B is chosen to be 200. In the following examples, the continuous random variables cannot be unbounded from practical considerations. So it is safe to use the asymptotic null distributions for all combinations of the tuning parameters.

Example 7.2. *(Data science salaries in* 2023*) This data set contains salaries (in USD) of the*

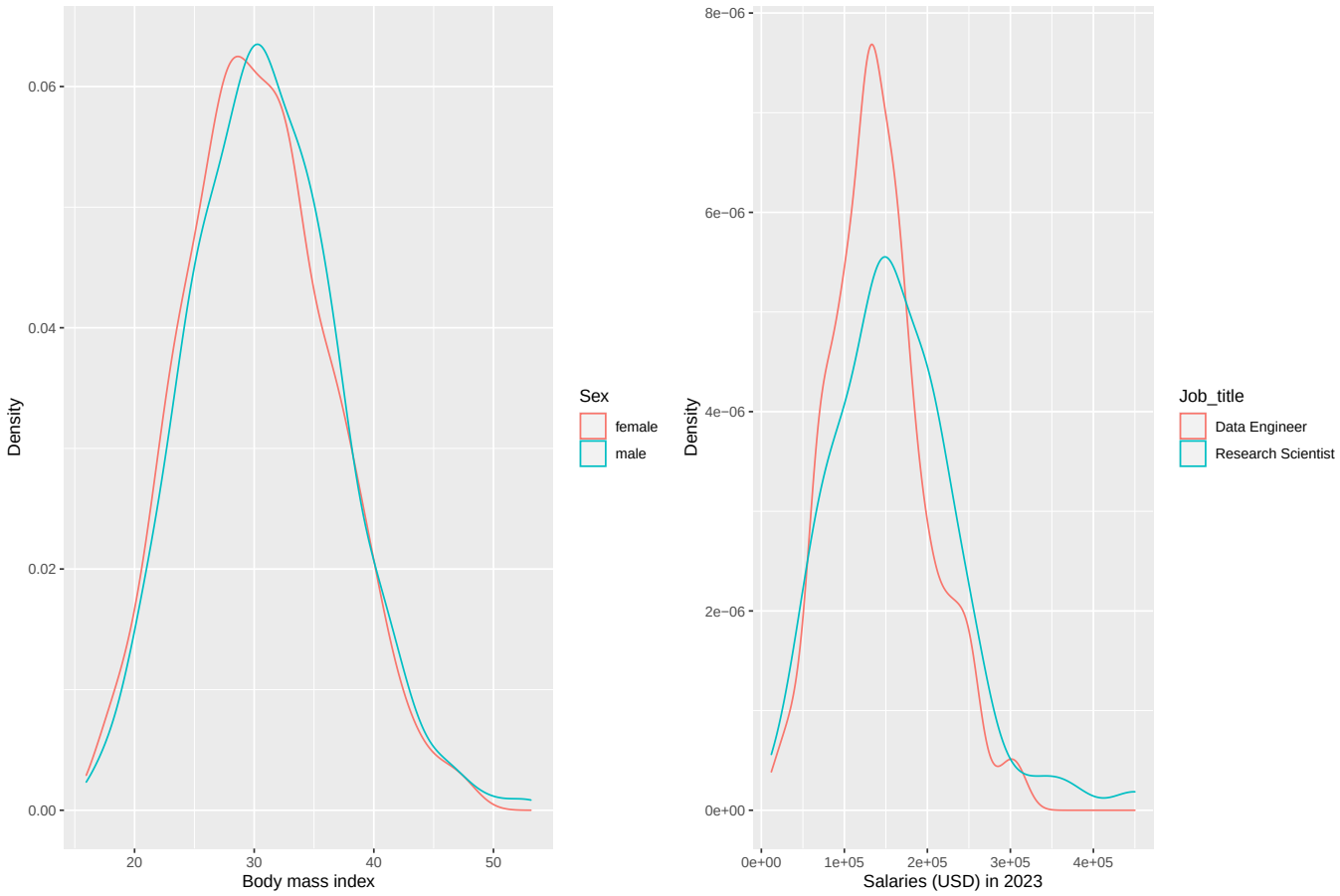

FIGURE 7.5: Plots of data sets.

employees in the "data science" profession in the year 2023 *across different companies and countries. We are interested to see if the distributions of salaries among the "Research scientists" differ significantly from that of the "Data engineers". We plot this data set in the right panel of Figure 7.5. There are* 1040 *data engineers and* 82 *research scientists in the data set. We start with the pilot tuning parameters* $(\alpha, \lambda, \beta) = (0.1, 0.5, 0)$ *which gives the p-value as* $0.01739 < 0.05$*. So both distributions are different at* 5% *level of significance, and this conclusion is consistent with Figure 7.5. In Table 7.15 the lowest risks are highlighted in green, and the tuning parameters corresponding to them are color-coded in blue. Here the set of optimum tuning parameters does not come from the PD family, but outside of them.*

Example 7.3. *(USA health insurance data set) This data set contains information about the*

TABLE 7.15: Comparison of risks for different methods in Example 7.2

Method	(α, λ, β)	Risk	Power
PD	$(0, -0.5, 0)$	0.640	0.360
	$(0, -0.3, 0)$	0.690	0.310
	$(0, -0.2, 0)$	0.665	0.335
	$(0, -0.1, 0)$	0.580	0.420
	$(0, 0, 0)$	0.425	0.575
	$(0, 0.25, 0)$	0.380	0.620
	$(0, 0.5, 0)$	0.285	0.715
	$(0, 1, 0)$	0.105	0.895
GSB	$(-0.1718, 1.0348, -0.0204)$	0.040	0.960

TABLE 7.16: Comparison of risks for different methods in Example 7.3

Method	(α, λ, β)	Risk	Pr. truly accept $\mathbb{H}$
PD	$(0, -0.5, 0)$	0.020	0.080
	$(0, -0.3, 0)$	0.010	0.090
	$(0, -0.2, 0)$	0.010	0.090
	$(0, -0.1, 0)$	0.015	0.985
	$(0, 0, 0)$	0.020	0.980
	$(0, 0.25, 0)$	0.020	0.980
	$(0, 0.5, 0)$	0.020	0.980
	$(0, 1, 0)$	0.050	0.950
GSB	$(0.5968, 1.2306, -0.5153)$	0.005	0.995
	$(0.1248, 0.0802, -0.4720)$	0.005	0.995

insurance charges based on different categories such as– age, sex, and BMI among many others. An important question is to test the equality of BMI between males and females who buy some insurance policies. Distributions of BMI are plotted for males and females in the left panel of Figure 7.5. Here the Pilot is taken as $(\alpha, \lambda, \beta) = (0, 0.8, 0)$*, which gives p-value as* $0.86274 > 0.05$*. In Table 7.16 the lowest risks are highlighted in green, and the tuning parameters corresponding to them are color-coded in blue. Here the set of optimum tuning parameters does not belong to the PD family, but outside of them.*

7.6 Conclusions

From the discussions, we see that the tests based on the generalized S-Bregman divergence produce very high empirical powers across different choices of tuning parameters. Also, the empirical levels stay near the 5% nominal level of significance. However, contamination in one data set may push the type-I error unacceptably high. In the face of contamination, the members of the GSB divergence generally exhibit much better robustness when chosen outside the power divergence family. Finally, a data-driven algorithm for optimum tuning parameter selection is proposed for the application of this two-sample test in real data examples. We hope this will turn out quite helpful to the applied scientists.

Data availability statement

The data sets that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023?resource=download and https://www.kaggle.com/code/teertha/us-health-insurance-eda/data.

This page is intentionally left blank.

Chapter 8

Example II: The Exponential-Polynomial Divergence

8.1 Introduction

Earlier we discussed a two-sample non-parametric test using the generalized S-Bregman divergence. In this chapter we shall discuss the same in the context of the Exponential-Polynomial (EP) divergence (Singh et al., 2021). This is an important family in the class of Bregman divergences because it unifies the DPD and BED families through a continuous extension. The key highlights of this chapter are the following.

(a) All the formulae specific to the Exponential-Polynomial divergence are deduced from the general theory developed earlier.

(b) The influence functions of the mutual information functional are plotted across different combinations of the tuning parameters.

(c) Through extensive simulation studies we find that some divergences outside the DPD and BED families are very competitive for producing very high empirical

powers and stable type-I errors. Sometimes the empirical power becomes lower for $\beta < 0$. Nonetheless, this improves with larger sample sizes.

(d) We also apply the algorithm of tuning parameter selection in a couple of real data examples.

8.2 Mutual Information based on the EP Divergence

8.2.1 The Class of EP Divergence

Singh et al. (2021) define the Exponential-Polynomial divergence that uses the ϕ-functions of both the BED and DPD families through a convex combination as

$$\phi(t) = \gamma \frac{2(e^{\beta t} - \beta t - 1)}{\beta^2} + (1-\gamma)\frac{t^{1+\alpha} - t}{\alpha} \text{ for } \alpha > 0, \beta \neq 0, \gamma \in [0,1]. \tag{8.1}$$

Consequently, the Exponential-Polynomial divergence unifies both these divergences in the following way:

$$d^{EP}_{\alpha,\beta,\gamma}(g,f) = \gamma \times BED_\beta(g,f) + (1-\gamma) \times d_\alpha(g,f) \text{ for } \alpha > 0, \beta \neq 0, \gamma \in [0,1]. \tag{8.2}$$

The EP divergence at the degenerate cases (i.e., $\alpha = 0$ and, or $\beta = 0$) are defined as continuous limit(s) of the expression in (8.2) for appropriate tuning parameter(s). See that the EP divergence becomes the DPD for $\gamma = 0$. When $\gamma = 1$, it yields the BED family. Notice that Singh et al. (2021) originally define the ϕ-function for the EP divergence without the factor 2 as present in the first term of (8.1). We make this minor adjustment in our current discussion to match its form with the BED family. However such change will not have any impact in the context of parametric estimation. A moment's reflection shows that the EP divergence can be embedded into the extended Bregman divergence

family (Basak and Basu, 2022) with the index $k = 1$. Hence the theory developed in Chapter 6 may be mimicked over here.

8.2.2 Mutual Information in a Hybrid Setup

A form of generalized mutual information based on the EP divergence satisfies (P1) - (P4) of Proposition 6.1. However it does not satisfy the conditions– $\phi(1) = \phi'(1) = 0$ of (P5). Hence (P5) does not hold here. All the notations of Chapter 6 are imported here unless otherwise specified. In a hybrid setup as mentioned before, the MI becomes

$$I_{EP}(X,Y) = \sum_{x=0}^{1} \int \left[\gamma \left\{ \frac{2}{\beta} e^{\beta(f_x f_y)} \left(f_x f_y - \frac{1}{\beta} \right) - e^{\beta(f_x f_y)} f_{x,y} + \frac{1}{\beta} e^{\beta(f_{x,y})} \right\} \right.$$
$$\left. + (1-\gamma) \left\{ (f_x f_y)^{1+\alpha} - \left(1 + \frac{1}{\alpha} \right) (f_x f_y)^{\alpha} f_{x,y} + \frac{1}{\alpha} f_{x,y}^{1+\alpha} \right\} \right] dy \text{ for } \alpha > 0, \beta \neq 0. \tag{8.3}$$

If $\alpha = 0$ but $\beta \neq 0$, the MI becomes

$$\lim_{\alpha \to 0} I_{EP}(X,Y) = \sum_{x=0}^{1} \int \left[\gamma \left\{ \frac{2}{\beta} e^{\beta(f_x f_y)} \left(f_x f_y - \frac{1}{\beta} \right) - e^{\beta(f_x f_y)} f_{x,y} + \frac{1}{\beta} e^{\beta(f_{x,y})} \right\} + (1-\gamma) f_{x,y} \ln \frac{f_{x,y}}{f_x f_y} \right] dy. \tag{8.4}$$

Similarly, it turns out that

$$\lim_{\beta \to 0} I_{EP}(X,Y) = \sum_{x=0}^{1} \int \left[\gamma (f_{x,y} - f_x f_y)^2 + (1-\gamma) \left\{ (f_x f_y)^{1+\alpha} - \left(1 + \frac{1}{\alpha} \right) (f_x f_y)^{\alpha} f_{x,y} + \frac{1}{\alpha} f_{x,y}^{1+\alpha} \right\} \right] dy \tag{8.5}$$

when $\alpha > 0, \beta = 0$. Also, we obtain

$$\lim_{\alpha, \beta \to 0} I_{EP}(X,Y) = \sum_{x=0}^{1} \int \left[\gamma (f_{x,y} - f_x f_y)^2 + (1-\gamma) f_{x,y} \ln \frac{f_{x,y}}{f_x f_y} \right] dy \tag{8.6}$$

for $\alpha = \beta = 0$.

8.3 Asymptotic Results

Let us assume that the derivatives up to the fourth-order of ϕ that are given by

$$\begin{aligned}
\phi'(t) &= 2\gamma\frac{(e^{\beta t}-1)}{\beta} + (1-\gamma)\frac{(\alpha+1)t^{\alpha}-1}{\alpha},\\
\phi''(t) &= 2\gamma e^{\beta t} + (1-\gamma)(1+\alpha)t^{\alpha-1},\\
\phi'''(t) &= 2\gamma\beta e^{\beta t} + (1-\gamma)(\alpha+1)(\alpha-1)t^{\alpha-2},\\
\phi''''(t) &= 2\gamma\beta^2 e^{\beta x} + (1-\gamma)(\alpha+1)(\alpha-1)(\alpha-2)t^{\alpha-3}
\end{aligned} \tag{8.7}$$

are uniformly bounded by integrable functions. Under the assumptions of Theorem 6.1 it follows that

$$nh_n^{1/2}\Big(\widehat{I}_{EP} - \frac{\mu}{nh_n}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2) \text{ as } n \to \infty \tag{8.8}$$

when the null hypothesis $\mathbb{H}$ is true, where

$$\mu = \frac{1}{2}\int_u K^2(u)du \int \sum_{x=0}^{1}\Big[2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^{\alpha}\Big](1-f_x)dy, \tag{8.9}$$

$$\sigma^2 = \frac{1}{2}\int_u\Big(\int_z K(z)K(z+u)dz\Big)^2 du \int\Big\{\sum_{x=0}^{1}\Big[2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^{\alpha}\Big](1-f_x)\Big\}^2 dy. \tag{8.10}$$

As before, $\widehat{I}_{EP}$ is a plug-in estimator of I_{EP}. The asymptotic mean and variance for different special cases are computed as follows.

Corollary 8.1. *(Likelihood disparity) Here $\alpha = \gamma = 0$ and we get*

$$\mu = \frac{1}{2}\int_u K^2(u)du \int dy, \tag{8.11}$$

$$\sigma^2 = \frac{1}{2}\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du \int dy. \tag{8.12}$$

Corollary 8.2. *(Density power divergence) Here $\alpha \geq 0$, $\gamma = 0$ and we get*

$$\mu = \frac{(1+\alpha)}{2}\int_u K^2(u)du\Big(f_{x_0}^{\alpha} f_{x_1} + f_{x_1}^{\alpha} f_{x_0}\Big) \int f_y^{\alpha} dy, \tag{8.13}$$

$$\sigma^2 = \frac{(1+\alpha)^2}{2}\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du\Big(f_{x_0}^{\alpha} f_{x_1} + f_{x_1}^{\alpha} f_{x_0}\Big)^2 \int f_y^{2\alpha} dy. \tag{8.14}$$

Corollary 8.3. *(Bregman exponential disparity) In this case $\beta \in \mathbb{R}$, $\gamma = 1$, this yields*

$$\mu = (f_{x_0} f_{x_1})\int_u K^2(u)du \int \left(e^{\beta(f_{x_0} f_y)} + e^{\beta(f_{x_1} f_y)}\right) f_y dy, \tag{8.15}$$

$$\sigma^2 = 2(f_{x_0} f_{x_1})^2 \int_u \left(\int K(z)K(z+u)dz\right)^2 du \int \left(e^{\beta(f_{x_0} f_y)} + e^{\beta(f_{x_1} f_y)}\right)^2 f_y^2 dy. \tag{8.16}$$

Corollary 8.4. *(Squared L_2 distance) In this case $\alpha = 1$, $\gamma = 0$ and we get*

$$\mu = 2 f_{x_0} f_{x_1} \int_u K^2(u)du, \tag{8.17}$$

$$\sigma^2 = 8(f_{x_0} f_{x_1})^2 \int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du \int f_y^2 dy. \tag{8.18}$$

Corollary 8.5. *When* $\alpha = 0$ *it becomes*

$$\mu = \frac{1}{2}\int_u K^2(u)du \int \sum_{x=0}^{1}\left[2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)\right](1-f_x)dy, \tag{8.19}$$

$$\sigma^2 = \frac{1}{2}\int_u \Big(\int_z K(z)K(z+u)dz\Big)^2 du \int \left\{\sum_{x=0}^{1}\left[2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)\right](1-f_x)\right\}^2 dy. \tag{8.20}$$

Suppose $0 \leq \gamma < 1, \alpha = 0$ and the support of Y is unbounded. In this case, we must implement the permutation algorithm to compute the empirical power and levels as we cannot readily use Theorem 6.1. As noted the permutation test takes a lot of computational burden. However, this can be completely avoided for other members of the EP divergence when $\alpha \neq 0$ even if support of Y is unbounded.

Let us define $T = nh_n^{1/2}\sigma^{-1}(\widehat{I}_{EP} - \frac{\mu}{nh_n})\sigma$. By Theorem 6.2, we already know that the tests are consistent. Also, it follows from Theorem 6.3 that

$$T \xrightarrow{\mathcal{L}} \frac{d^2}{2\sigma}\sum_{x=0}^{1}\int \left\{2\gamma(f_x f_y)^2 e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^{\alpha+1}\right\}\left(\frac{\Delta_x}{f_x} + \frac{\Delta_y}{f_y} - \frac{\Delta_{x,y}}{f_x f_y}\right)^2 dy + \mathcal{N}(0,1) \tag{8.21}$$

as $n \to \infty$ under the sequence of contiguous alternatives $\mathbb{K}_n$ given in (6.116).

8.4 Robustness Studies

This section discusses the stability behaviour of I_{EP}. Firstly, we shall plot the influence function. After that, its asymptotic breakdown point will be computed.

8.4.1 Influence Functions

It follows from Theorem 6.4 that the second-order influence function of I_{EP} at $t_0 = (x_0, y_0)$ under the null hypothesis is given by

$$\mathcal{IF}_2(I_{EP}, f_X f_Y, t_0) = \sum_{x=0}^{1} \int \left\{ 2\gamma (f_x f_y) e^{\beta (f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^{\alpha} \right\} U_{xy} dy, \tag{8.22}$$

where U_{xy} is defined in (7.24). In particular

$$\mathcal{IF}_2(I_{EP}, f_X f_Y, t_0) = \begin{cases} (1+\alpha) \sum_{x=0}^{1} \int (f_x f_y)^{\alpha} U_{xy} dy & \text{if } \gamma = 0, \\ \sum_{x=0}^{1} \int 2 (f_x f_y) e^{\beta (f_x f_y)} U_{xy} dy & \text{if } \gamma = 1. \end{cases} \tag{8.23}$$

Similarly, as before, the boundedness of the influence function essentially depends on controlling the term U_{xy} inside the summation and integration in (8.22). Since f_y is a probability density function of a continuous random variable, its boundedness may be an issue. However, a sufficient condition such as the density of Y has a continuously bounded derivative ensures that its density is uniformly bounded. When f_y is uniformly bounded U_{xy} can still be unbounded. However, we will see further that it is the factor $(f_x f_y)^{\alpha}$ that adds higher stability to the influence function at extreme outliers with increasing α. Also the term $e^{\beta (f_x f_y)}$ always stays bounded for any finite β when f_y is bounded. Taking a cue from earlier discussions, we shall study the behaviour of the influence functions separately for each of the following regions.

(1) When $\gamma = 0$ the first term in (8.22) drops out and the stability of the non-vanishing second term increases with α. In this case, the entire expression becomes independent of β. We denote this region by $\mathbb{S}_6 = \left\{ (\alpha, \beta, \gamma) : \alpha \geq 0, \beta \in \mathbb{R}, \gamma = 0 \right\}$.

(2) When $\gamma = 1$ the second term in (8.22) vanishes. Consequently, the influence function becomes bounded for any finite β. In this case, the expression becomes independent of α. This region is denoted by $\mathbb{S}_7 = \Big\{(\alpha, \beta, \gamma) : \alpha \geq 0, \beta \in \mathbb{R}, \gamma = 1\Big\}$.

(3) In the third case neither of these two terms in (8.22) drops out. This region is denoted by $\mathbb{S}_8 = \Big\{(\alpha, \beta, \gamma) : \alpha \geq 0, \beta \in \mathbb{R}, 0 < \gamma < 1\Big\}$.

We consider Example 7.1 from Chapter 7 for numerical illustrations.

It is clear in Figure 8.1 that the values of the influence functions are substantially down-weighted as α increases from zero in the set $\mathbb{S}_6$. The influence functions become bounded for whatever the finite choice of β may be at $\gamma = 1$ in the set $\mathbb{S}_7$. This is evident in Figure 8.2. Similarly in Figure 8.3, we see that the influence functions in $\mathbb{S}_8$ tend to become bounded for increasing α and γ.

For completeness, we calculate the $\mathcal{LIF}$ and $\mathcal{PIF}$ using Theorem 6.6 as the following.

(a) The power influence function is given by

$$\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi} \cdot \phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int \left\{2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^\alpha\right\} U_{xy} dy\right)$$
$$\times \sum_{x=0}^{1} \int (f_x f_y)\left(\frac{\Delta^*_{x,y}}{f_x f_y} - \frac{\Delta^*_x}{f_x} - \frac{\Delta^*_y}{f_y}\right)\left(\frac{\Delta_{x,y}}{f_x f_y} - \frac{\Delta_x}{f_x} - \frac{\Delta_y}{f_y}\right)\left\{2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^\alpha\right\} dy. \tag{8.24}$$

(b) When $t_0 = t_1$, we get

$$\mathcal{PIF}(\pi, t_1) = \frac{d}{\sigma_\phi} \cdot \phi_1\left(\tau_c - \frac{d^2}{2\sigma_\phi} \sum_{x=0}^{1} \int \left\{2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^\alpha\right\} U_{xy} dy\right)$$
$$\times \sum_{x=0}^{1} \int \left\{2\gamma(f_x f_y)e^{\beta(f_x f_y)} + (1-\gamma)(1+\alpha)(f_x f_y)^\alpha\right\} U_{xy} dy. \tag{8.25}$$

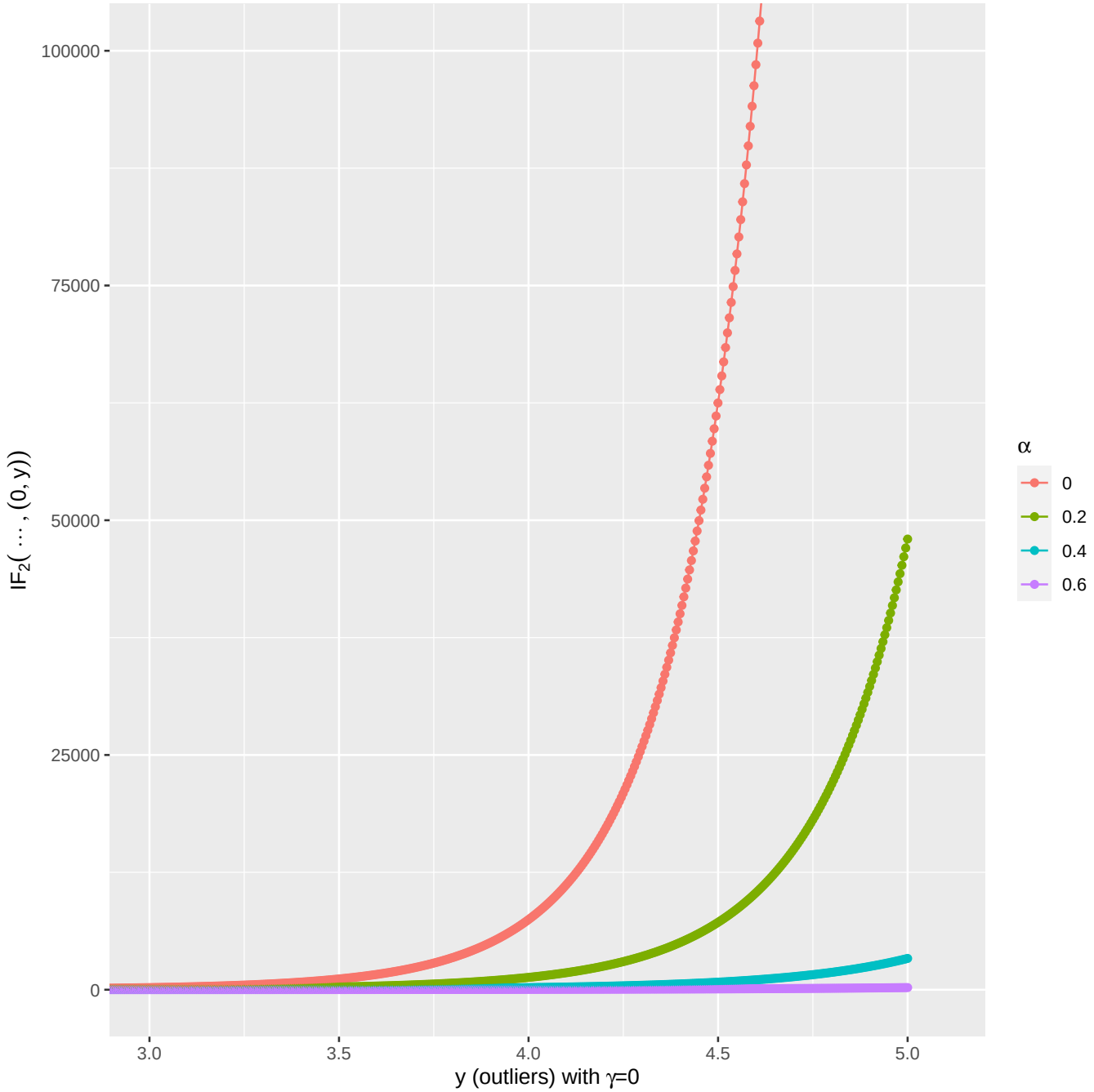

FIGURE 8.1: Influence function when the tuning are in $\$_6$ region.

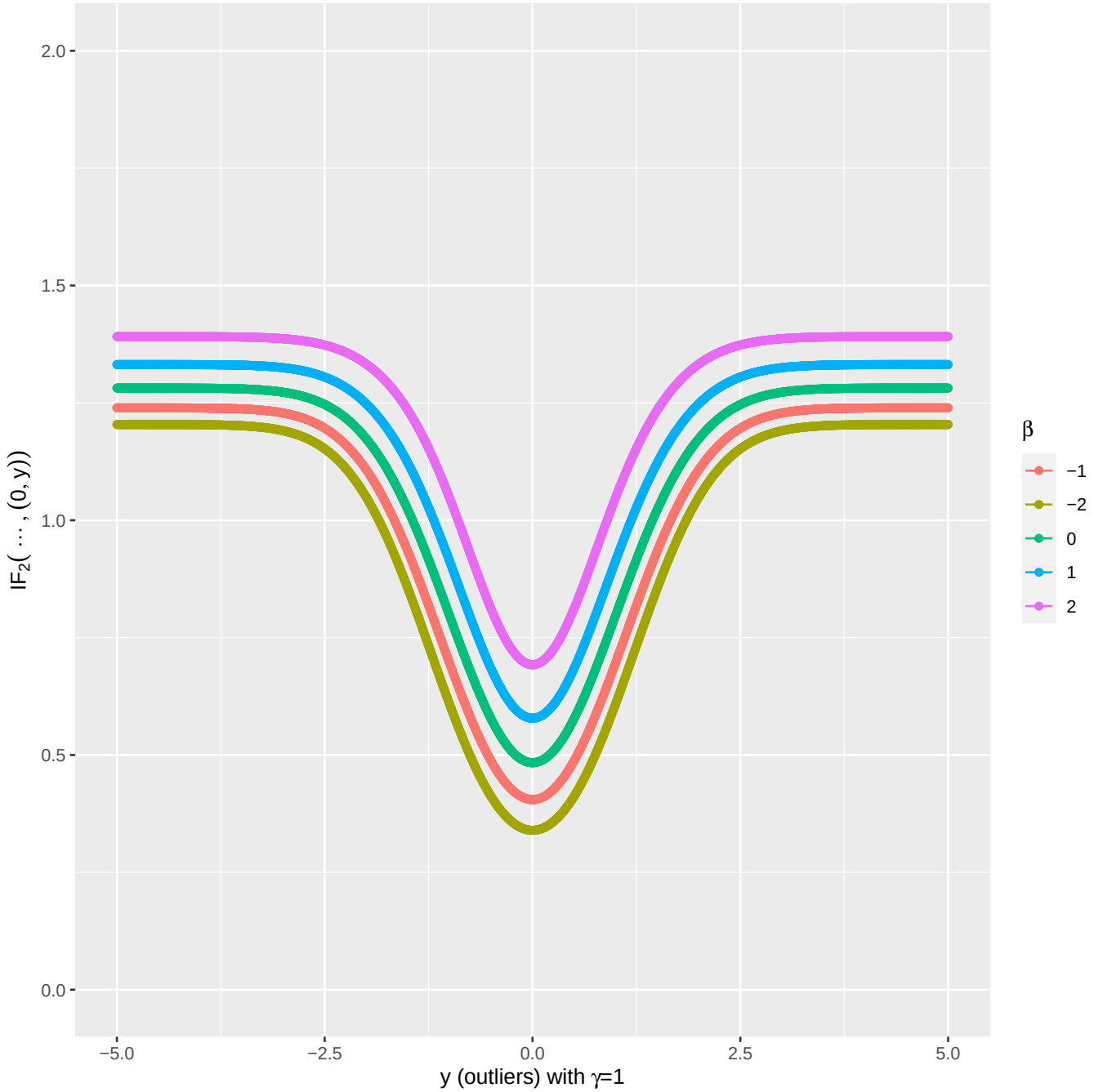

FIGURE 8.2: Influence function when the tuning are in $\$_7$ region.

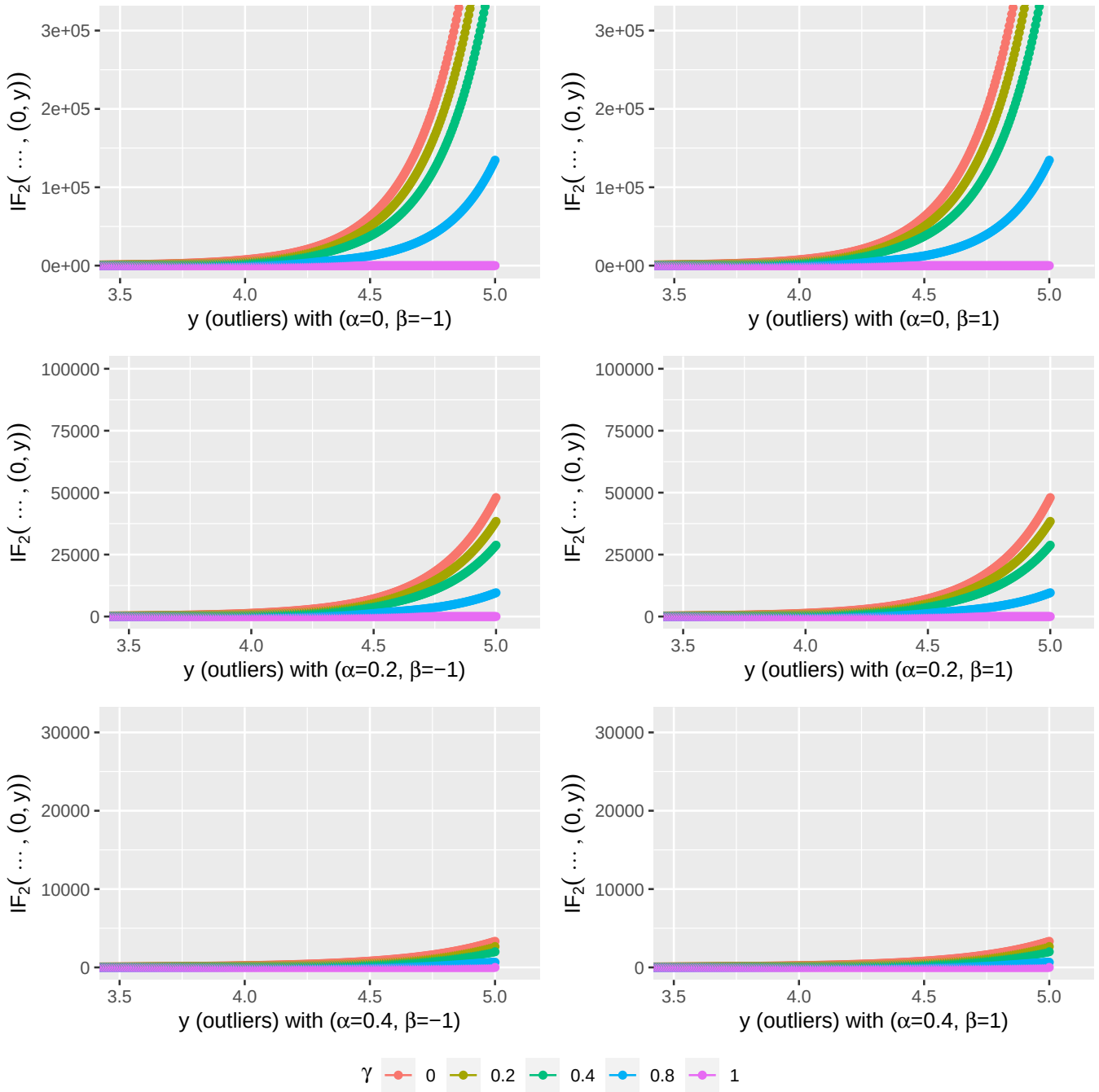

FIGURE 8.3: Influence functions when the tuning are in $\$_8$ region.

(c) The level influence function is given by

$$\mathcal{LIF}(\pi, t_0) \equiv 0. \tag{8.26}$$

We see that $\mathcal{PIF}$ becomes stable whenever the influence function of I_{EP} is stable. Also, notice that $\mathcal{LIF}$ up to any order is always zero. This is, in fact, one limitation of the influence function that it cannot capture the robustness of the type-I error even when it is not hard to make out in the simulation studies at different choices of tuning parameters.

8.4.2 Asymptotic Breakdown Point of I_{EP}

We use Theorem 6.7 to compute the asymptotic breakdown point of the MI based on the EP divergence. This result heavily depends on Assumption (BP6). Later we give sufficient conditions for Assumption (BP6). Let us define

$$M_{g,f} = \sum \int g \left\{ \frac{2\gamma}{\beta}(e^{\beta f} - 1) + (1-\gamma)\frac{(1+\alpha)f^{\alpha} - 1}{\alpha} \right\}. \tag{8.27}$$

It follows from Lemma 6.6 that $d^{EP}_{\alpha,\beta,\gamma}(\epsilon g, f) \geq d^{EP}_{\alpha,\beta,\gamma}(\epsilon g, g)$ when

$$\epsilon \leq 1 + \frac{\sum \int \left[2\alpha\gamma(e^{\beta g} - e^{\beta f}) + (1-\gamma)\beta^2(g^{1+\alpha} - f^{1+\alpha}) \right]}{\sum \int \left[2\alpha\beta\gamma(f e^{\beta f} - g e^{\beta g}) + (1-\gamma)(1+\alpha)\beta^2(f^{1+\alpha} - g^{1+\alpha}) \right]} \tag{8.28}$$

provided $M_{f,f} \geq M_{g,f}$, $g \neq f$ and $\alpha, \beta \neq 0$. When the divergence is defined as a limit for some tuning parameters, the same limit is passed over the assumptions and the upper bound of ϵ.

Corollary 8.6. *When $\gamma = 0$, the EP divergence becomes the density power divergence. In this case, we have $d_\alpha(\epsilon g, f) \geq d_\alpha(\epsilon g, g)$ for $\epsilon \leq \frac{\alpha}{1+\alpha}$ when*

$$\sum \int \left\{ f^{1+\alpha} - g^{1+\alpha} \right\} \geq 0, \tag{8.29}$$

provided $g = f_{X,Y}, f = f_X f_Y$ and X, Y are not independent.

Corollary 8.7. *When $\gamma = 1$, it becomes the BED family with the tuning parameter β. Let us assume that $\beta \neq 0$. Then we have $BED_\beta(\epsilon g, f) \geq BED_\beta(\epsilon f, f)$ for*

$$\epsilon \leq 1 + \frac{\sum \int \left(e^{\beta g} - e^{\beta f} \right)}{\beta \sum \int \left(f e^{\beta f} - g e^{\beta g} \right)} \text{ when } \sum \int (f - g) \frac{(e^{\beta f} - 1)}{\beta} \geq 0, \tag{8.30}$$

and X, Y are not independent. At $\beta = 0$, the condition holds in the limit as $\beta \to 0$. In this case this inequality holds for $\epsilon \leq \frac{1}{2}$ when $\sum \int f^2 > \sum \int g^2$ and X, Y are not independent.

A second version of the breakdown point is also presented in Chapter 6. It follows from Theorem 6.8 that the asymptotic breakdown point of I_{EP} is at least $\min\{\epsilon_1, \epsilon_2, \frac{1}{2}\}$ where

$$\epsilon_1 = 1 + \limsup_{m \to \infty} \frac{\sum \int \left[2\alpha\gamma (e^{\beta k_{xy,m}} - e^{\beta (f_x^m f_y^m)}) + (1-\gamma)\beta^2 (k_{xy,m}^{1+\alpha} - (f_x^m f_y^m)^{1+\alpha}) \right]}{\sum \int \left[2\alpha\beta\gamma ((f_x^m f_y^m) e^{\beta (f_x^m f_y^m)} - k_{xy,m} e^{\beta k_{xy,m}}) + (1-\gamma)(1+\alpha)\beta^2 ((f_x^m f_y^m)^{1+\alpha} - k_{xy,m}^{1+\alpha}) \right]}, \tag{8.31}$$

$$\epsilon_2 = \frac{\sum \int \left[2\alpha\gamma (e^{\beta f_{x,y}} - e^{\beta (f_x f_y)}) + (1-\gamma)\beta^2 (f_{x,y}^{1+\alpha} - (f_x f_y)^{1+\alpha}) \right]}{\sum \int \left[2\alpha\beta\gamma (f_{x,y} e^{\beta f_{x,y}} - (f_x f_y) e^{\beta (f_x f_y)}) + (1-\gamma)(1+\alpha)\beta^2 (f_{x,y}^{1+\alpha} - (f_x f_y)^{1+\alpha}) \right]}. \tag{8.32}$$

Corollary 8.8. *For the density power divergence (The EP divergence with $\gamma = 0$), we obtain*

$$\epsilon_1 = \left(\frac{\alpha}{1+\alpha} \right) \text{ and } \epsilon_2 = \left(\frac{1}{1+\alpha} \right). \tag{8.33}$$

The asymptotic breakdown point becomes

$$\min\left\{\frac{\alpha}{1+\alpha}, \frac{1}{1+\alpha}, \frac{1}{2}\right\} \text{ with } \alpha \geq 0, \tag{8.34}$$

under the assumptions of Theorem 6.8.

It is evident that when α increases, the asymptotic breakdown point of MI based on the DPD increases considerably. However, a simplified expression is not obtained when $\gamma \neq 0$. In that case the asymptotic breakdown point of I_{EP} depends on the densities of Y and the contaminating sequences. Hence it is hard to characterize the tuning parameters producing highly asymptotic breakdown points, in general, unless specific examples are considered.

8.5 Numerical Studies

8.5.1 Simulation Results

(A) Here we perform simulation studies for the following sets of models.

Model 0 : (Null Model) $Y_0 \sim \mathcal{N}(0,1)$,

Model 1 : $Y_1 \sim \mathcal{N}(0,(1+a)^2)$ with $a = 0.75$,

Model 2 : $Y_1 \sim (1-a)\mathcal{N}(-1,1) + a\mathcal{N}(1,1)$ with $a = 0.6$,

Model 3 : $f_{Y_1}(y) = \Big(1 - \sin(\pi y/\sqrt{2})a\Big)\phi(y)$ with $a = 0.6$.

(B) To study the robustness of type-I error of the proposed tests, the first sample is drawn from $(1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(6,2)$ with $\epsilon = 0.05, 0.065, 0.08, 0.10$ while the second sample comes from $\mathcal{N}(0,1)$.

Samples of sizes $n_0 = 100$ and $n_1 = 100$ are drawn from the distributions under null and alternative models, respectively, over 500 replications. The tests are conducted at 5% nominal level of significance. We consider the following combinations of the tuning parameters: $\alpha = (0, 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, 1)^T$, $\gamma = (0, 0.25, 0.5, 0.75, 1)^T$ and $\beta = (0, -1)^T$. Since the models have unbounded supports, the asymptotic null distribution cannot be used for $0 \leq \gamma < 1$ with $\alpha = 0$. In that case, we use permutation tests using the algorithm of Guha et al. (2021) over 500 permutations. However, the unbounded support of the continuous distribution will not cause a problem in using the asymptotic null distribution for the members of the EP divergence family other than $0 \leq \gamma < 1$ with $\alpha = 0$. This relieves a lot of computational burden. The kernels and the bandwidth sequence to be used are chosen as in (7.42).

In the following simulation studies, we see that the observed levels of the tests are fairly reasonable for all our chosen tuning parameters. These tests become slightly conservative for β smaller than -1. Observed levels are reported in Table 8.1 and Table 8.2. The null hypothesis $\mathbb{H}$ is rejected when both the models differ. In these cases, we report the empirical powers of the tests in Table 8.3 to Table 8.8 which turn out to be fairly competitive against both the DPD and BED families for our choice of tuning parameters. The null model is misspecified when the first sample is slightly contaminated but, as usual, the second sample comes from $\mathcal{N}(0, 1)$. Empirical levels are reported in Table 8.9 to Table 8.16 when the null model is contaminated at different strengths. We find that the robustness of the proposed tests increases as α tends towards 1 at fixed β and γ. Lower values of β yield better robustness when $\gamma = 1$. A similar trend is also noticed for increasing γ when α, β are held fixed.

TABLE 8.1: Observed level when both samples are generated through Model 0 with $\beta = 0$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.040	0.036	0.030	0.030	0.032	0.030	0.026	0.024	0.024
0.25	0.040	0.034	0.030	0.032	0.028	0.026	0.026	0.022	0.024
0.50	0.040	0.032	0.032	0.032	0.028	0.026	0.024	0.022	0.024
0.75	0.040	0.034	0.034	0.030	0.026	0.022	0.022	0.022	0.024
1.0	0.024	0.024	0.024	0.024	0.024	0.024	0.024	0.024	0.024

TABLE 8.2: Observed level when both samples are generated through Model 0 with $\beta = -1$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.040	0.036	0.030	0.030	0.032	0.030	0.026	0.024	0.024
0.25	0.040	0.034	0.030	0.032	0.030	0.030	0.026	0.024	0.024
0.50	0.040	0.032	0.032	0.032	0.030	0.026	0.026	0.022	0.022
0.75	0.040	0.032	0.032	0.032	0.026	0.026	0.024	0.022	0.022
1.0	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022

TABLE 8.3: Observed power when the first and second samples are respectively generated through Model 0 and Model 1 with $\beta = 0$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.984	0.948	0.932	0.894	0.874	0.854	0.838	0.806
0.25	1	0.984	0.948	0.930	0.886	0.864	0.846	0.832	0.806
0.50	1	0.968	0.942	0.922	0.878	0.856	0.838	0.824	0.806
0.75	1	0.952	0.926	0.896	0.854	0.838	0.826	0.816	0.806
1.0	0.806	0.806	0.806	0.806	0.806	0.806	0.806	0.806	0.806

TABLE 8.4: Observed power when the first and second samples are respectively generated through Model 0 and Model 1 with $\beta = -1$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.984	0.948	0.932	0.894	0.874	0.854	0.838	0.806
0.25	1	0.984	0.948	0.930	0.888	0.864	0.846	0.832	0.806
0.50	1	0.970	0.944	0.924	0.876	0.858	0.842	0.828	0.812
0.75	1	0.954	0.930	0.900	0.858	0.844	0.830	0.822	0.816
1.0	0.816	0.816	0.816	0.816	0.816	0.816	0.816	0.816	0.816

TABLE 8.5: Observed power when the first and second samples are respectively generated through Model 0 and Model 2 with $\beta = 0$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.962	0.822	0.766	0.718	0.650	0.618	0.592	0.584	0.544
0.25	0.962	0.818	0.762	0.700	0.634	0.608	0.586	0.578	0.544
0.50	0.962	0.808	0.732	0.680	0.614	0.594	0.586	0.576	0.544
0.75	0.962	0.756	0.686	0.646	0.590	0.582	0.576	0.560	0.544
1.0	0.544	0.544	0.544	0.544	0.544	0.544	0.544	0.544	0.544

TABLE 8.6: Observed power when the first and second samples are respectively generated through Model 0 and Model 2 with $\beta = -1$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.962	0.822	0.766	0.718	0.650	0.618	0.592	0.584	0.544
0.25	0.962	0.818	0.760	0.700	0.636	0.608	0.586	0.576	0.546
0.50	0.962	0.808	0.738	0.682	0.622	0.600	0.582	0.574	0.548
0.75	0.962	0.764	0.694	0.656	0.596	0.584	0.576	0.566	0.552
1.0	0.556	0.556	0.556	0.556	0.556	0.556	0.556	0.556	0.556

TABLE 8.7: Observed power when the first and second samples are respectively generated through Model 0 and Model 3 with $\beta = 0$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.338	0.204	0.190	0.192	0.192	0.188	0.176	0.174	0.168
0.25	0.338	0.204	0.196	0.198	0.188	0.180	0.174	0.172	0.168
0.50	0.338	0.202	0.200	0.194	0.186	0.176	0.174	0.170	0.168
0.75	0.338	0.208	0.194	0.186	0.174	0.170	0.170	0.168	0.168
1.0	0.168	0.168	0.168	0.168	0.168	0.168	0.168	0.168	0.168

TABLE 8.8: Observed power when the first and second samples are respectively generated through Model 0 and Model 3 with $\beta = -1$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	0.338	0.204	0.190	0.192	0.192	0.188	0.176	0.174	0.168
0.25	0.338	0.204	0.194	0.198	0.190	0.182	0.176	0.174	0.168
0.50	0.338	0.206	0.192	0.198	0.190	0.176	0.174	0.174	0.168
0.75	0.338	0.200	0.194	0.190	0.176	0.174	0.174	0.172	0.170
1.0	0.172	0.172	0.172	0.172	0.172	0.172	0.172	0.172	0.172

TABLE 8.9: Observed level under the null hypothesis with $\beta = 0$ when 5% observations of first sample is $\mathcal{N}(6, 2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.156	0.082	0.070	0.052	0.054	0.050	0.052	0.048
0.25	1	0.148	0.080	0.066	0.054	0.052	0.052	0.052	0.048
0.50	1	0.134	0.078	0.062	0.054	0.052	0.052	0.054	0.048
0.75	1	0.102	0.068	0.058	0.052	0.052	0.052	0.050	0.048
1.0	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048

TABLE 8.10: Observed level under the null hypothesis with $\beta = -1$ when 5% observations of first sample is $\mathcal{N}(6,2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.156	0.082	0.070	0.052	0.054	0.050	0.052	0.048
0.25	1	0.152	0.082	0.066	0.052	0.052	0.052	0.052	0.050
0.50	1	0.142	0.080	0.064	0.054	0.050	0.052	0.052	0.048
0.75	1	0.106	0.072	0.058	0.052	0.052	0.052	0.052	0.048
1.0	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

TABLE 8.11: Observed level under the null hypothesis with $\beta = 0$ when 6.5% observations of the first sample is $\mathcal{N}(6,2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.236	0.112	0.086	0.072	0.070	0.062	0.058	0.050
0.25	1	0.216	0.108	0.086	0.072	0.066	0.060	0.052	0.050
0.50	1	0.186	0.100	0.084	0.068	0.064	0.058	0.050	0.050
0.75	1	0.138	0.094	0.076	0.062	0.052	0.050	0.050	0.050
1.0	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

TABLE 8.12: Observed level under the null hypothesis with $\beta = -1$ when 6.5% observations of the first sample $\mathcal{N}(6,2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.236	0.112	0.086	0.072	0.070	0.062	0.058	0.050
0.25	1	0.216	0.108	0.086	0.072	0.066	0.060	0.058	0.050
0.50	1	0.188	0.102	0.086	0.070	0.062	0.058	0.054	0.050
0.75	1	0.140	0.092	0.074	0.062	0.058	0.058	0.052	0.048
1.0	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

TABLE 8.13: Observed level under the null hypothesis with $\beta = 0$ when 8% observations of the first sample $\mathcal{N}(6, 2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.446	0.220	0.144	0.108	0.100	0.100	0.096	0.082
0.25	1	0.414	0.198	0.140	0.108	0.100	0.100	0.098	0.082
0.50	1	0.366	0.176	0.130	0.102	0.102	0.100	0.092	0.082
0.75	1	0.266	0.142	0.118	0.102	0.100	0.092	0.082	0.082
1.0	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.082	0.082

TABLE 8.14: Observed level under the null hypothesis with $\beta = -1$ when 8% observations of the first sample $\mathcal{N}(6, 2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.446	0.220	0.144	0.108	0.100	0.100	0.096	0.082
0.25	1	0.416	0.202	0.140	0.108	0.100	0.100	0.096	0.082
0.50	1	0.372	0.182	0.132	0.100	0.100	0.098	0.094	0.082
0.75	1	0.282	0.148	0.120	0.100	0.098	0.096	0.086	0.080
1.0	0.078	0.078	0.078	0.078	0.078	0.078	0.078	0.078	0.078

TABLE 8.15: Observed level under the null hypothesis with $\beta = 0$ when 10% observations of the first sample $\mathcal{N}(6, 2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.680	0.422	0.290	0.184	0.160	0.146	0.146	0.136
0.25	1	0.644	0.394	0.266	0.176	0.156	0.146	0.144	0.136
0.50	1	0.600	0.342	0.250	0.166	0.146	0.146	0.142	0.136
0.75	1	0.446	0.268	0.206	0.152	0.146	0.142	0.140	0.136
1.0	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136	0.136

TABLE 8.16: Observed level under the null hypothesis with $\beta = -1$ when 10% observations of the first sample $\mathcal{N}(6,2)$

γ \ α	0	0.1	0.2	0.3	0.5	0.6	0.7	0.8	1
0.0	1	0.680	0.422	0.290	0.184	0.160	0.146	0.146	0.136
0.25	1	0.650	0.398	0.270	0.176	0.158	0.146	0.146	0.136
0.50	1	0.610	0.360	0.254	0.166	0.146	0.146	0.144	0.134
0.75	1	0.478	0.274	0.222	0.150	0.146	0.144	0.142	0.136
1.0	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132	0.132

8.5.2 Real Data Examples

Like most other robust statistical methods, the proposed tests based on the EP divergence depend heavily on the choice of tuning parameters. To be able to apply this test in real-life applications, we should have a data-driven methodology to choose optimum tuning parameters without having any prior information about the following two cases– both these samples come from the same distribution, and both come from different distributions. Algorithm 7.1 partially solves this issue. We will adopt it over here.

We take up two real data examples and find the optimal tuning parameters using Algorithm 7.1. The primary aim here is to show that the algorithm leads to a set of tuning parameters that are at least as good as the DPD and BED families, if not better. The number of resamples is chosen to be 200. In the following examples, the continuous random variables– price and BMI cannot be unbounded from practical considerations. So we can safely use the asymptotic null distributions for all combinations of the tuning parameters.

Example 8.1. *(Automobile data set) This data set is picked up from* 1985 *Ward's Automotive Yearbook. The engines of automobiles are suited to use either diesel or gas as fuel. We are interested to see if the distributions of prices vary significantly according to their fuel types.*

This data set is plotted in the left panel of Figure 8.4. We start with the pilot tuning parameters $(\alpha, \gamma, \beta) = (0, 0.5, -1)$ *that gives the p-value as* $0.029 < 0.05$*. So both distributions are different at* 5% *level of significance. This conclusion is consistent with the data set. In Table 8.17 the lowest risks are highlighted in green, and the tuning parameters corresponding to them are color-coded in blue. Here the set of optimum tuning parameters come from neither the DPD nor the BED family, but outside of them.*

Example 8.2. *(USA health insurance data set) This data set is the same as in Example 7.3. The distributions of BMI are plotted for males and females in the right panel of Figure 8.4. Here the pilot is taken as* $(\alpha, \gamma, \beta) = (0, 0, -1)$ *which gives the p-value as* $0.93 > 0.05$*. Again, this conclusion is consistent with Figure 8.4 which says that both the samples come from the same distribution. In Table 8.18 the lowest risks are highlighted in green, and the tuning parameters corresponding to the lower risks are color-coded in blue. The optimum choice of tuning parameters is not unique in this example. They come from DPD, BED, and also outside of them.*

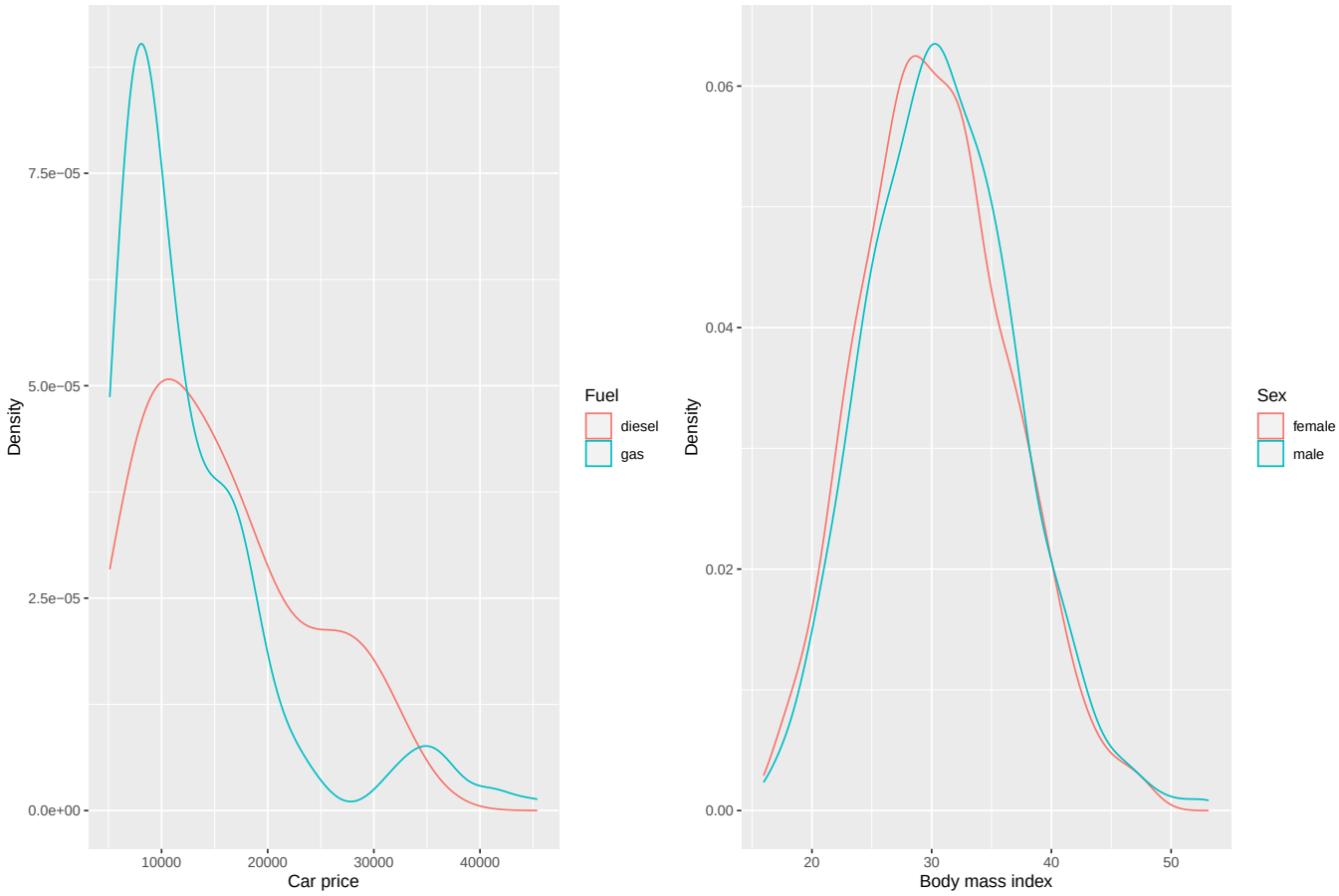

FIGURE 8.4: Plots of data sets.

TABLE 8.17: Comparison of risks for different methods for Example 8.1

Method	(α, γ, β)	Risk	Observed power
DPD	$(0, 0, 0)$	0.340	0.660
	$(0.1, 0, 0)$	0.430	0.570
	$(0.2, 0, 0)$	0.450	0.550
	$(0.3, 0, 0)$	0.475	0.525
	$(0.4, 0, 0)$	0.535	0.465
	$(0.5, 0, 0)$	0.590	0.410
	$(0.6, 0, 0)$	0.640	0.360
	$(0.7, 0, 0)$	0.695	0.305
	$(0.8, 0, 0)$	0.730	0.270
	$(0.9, 0, 0)$	0.765	0.235
	$(1.0, 0, 0)$	0.805	0.195
BED	$(0, 1, -1)$	0.805	0.195
	$(0, 1, -2)$	0.805	0.195
	$(0, 1, -3)$	0.805	0.195
	$(0, 1, -4)$	0.805	0.195
	$(0, 1, -5)$	0.805	0.195
Pilot	$(0, 0.5, -1)$	0.340	0.660
EPD	$(0.0046, 0.5656, -0.9055)$	0.315	0.685

TABLE 8.18: Comparison of risks for different methods in Example 8.2

Method	(α, γ, β)	Risk	Pr. truly acc. $\mathbb{H}$
DPD	$(0, 0, 0)$	0.020	0.980
	$(0.1, 0, 0)$	0.020	0.980
	$(0.2, 0, 0)$	0.020	0.980
	$(0.3, 0, 0)$	0.015	0.985
	$(0.4, 0, 0)$	0.015	0.985
	$(0.5, 0, 0)$	0.010	0.990
	$(0.6, 0, 0)$	0.010	0.990
	$(0.7, 0, 0)$	0.010	0.990
	$(0.8, 0, 0)$	0.010	0.990
	$(0.9, 0, 0)$	0.010	0.990
	$(1.0, 0, 0)$	0.010	0.990
BED	$(0, 1, -1)$	0.010	0.990
	$(0, 1, -2)$	0.010	0.990
	$(0, 1, -3)$	0.010	0.990
	$(0, 1, -4)$	0.010	0.990
	$(0, 1, -5)$	0.010	0.990
Pilot	$(0, 0, -1)$	0.020	0.980
EPD	$(0.5941, 0.4435, 0.4754)$	0.010	0.990

8.6 Conclusions

From the discussions, we see that the tests based on the MI using the EP divergence produce robust and consistent tests for the equality between two completely unstructured absolutely continuous distributions. Sometimes contamination in the null model may push the type-I error unacceptably high for $\alpha = 0$. In the face of contamination, the members of the EP divergence exhibit much better robustness for higher values of α and γ. Also, the empirical powers do not vary much across different choices of tuning parameters except sometimes for very small values of β. Finally, a couple of real data examples are analyzed by applying the tuning parameter selection algorithm in the context of EP divergence. This may be quite helpful to the applied scientists.

Data availability statement

The data sets that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/toramky/automobile-dataset and https://www.kaggle.com/code/teertha/us-health-insurance-eda/data.

This page is intentionally left blank.

Chapter 9

Epilogue

This thesis primarily focuses on the application of the density power divergence to the studies of some mixed data problems. Also, we develop a unified theory of two-sample non-parametric tests for a general class of divergence measures. The main part of this thesis can be roughly divided into three parts. Starting with a brief introduction of the role of the minimum distance methodologies in robust inference in Chapter 1, we quickly move on to our main work.

In the first part as in Chapter 2, we make a detailed study of the estimation of the parameters in ordinal response models. These data sets are quite common in many areas of scientific studies. In the future, we would like to develop robust Wald-type and score-type tests for testing statistical hypotheses in the same setup.

The second part, which starts from Chapter 3 and continues up to Chapter 5, deals with the estimation and development of Wald-type tests regarding the polychoric correlation. In continuation, later, we would like to study how the efficiency and robustness may vary as the order of the contingency table– $r_1 \times r_2 \times r_3 \times \cdots$ increases. Also, we would like to develop robust and efficient estimators and tests when both the polychoric and polyserial correlation are involved through the parametric model.

In the third part, we propose a class of two-sample non-parametric tests based on the class of extended Bregman divergences for testing the equality between two completely unstructured absolutely continuous distributions in Chapter 6. In the next couple of chapters, i.e., Chapter 7 and Chapter 8, this theory is fully illustrated with extensive numerical studies with applications to the specific classes of divergences, namely, the GSB and EP divergences. Later, we would like to use the extended Bregman divergence measures to study the association among different components of a time series data. Also, we would like to study the asymptotic breakdown point of the minimu m extended Bregman divergence estimator in parametric estimation.

This page is intentionally left blank.

List of Papers

Published:

- Pyne A, Roy S, Ghosh A and Basu A (2024), Robust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence, In: *Statistics*, pp. 1-40.

Ongoing Papers:

- Pyne A, Ghosh A and Basu A (2022), One-step Inference of Polychoric Correlation using the Density Power Divergence.
- Pyne A, Ghosh A and Basu A (2022), Two-step Inference of Polychoric Correlation using the Density Power Divergence.
- Pyne A, Ghosh A and Basu A (2022), Improving Bias and MSE over Two-step DPD-based Estimation of Polychoric Correlation.
- Pyne A, Ghosh A and Basu A (2023), Two-Sample Nonparametric Tests using the Extended Bregman Divergence with Applications of the Generalized S-Bregman Divergence.
- Pyne A, Ghosh A and Basu A (2023), A Class of Two-Sample Nonparametric Tests using the Exponential-Polynomial Divergence.

Conference Presentations:

- "Roust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence" at **ISI-ISM-ISSAS Conference 2023**, Indian Statistical Institute, Kolkata

- "Roust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence" at **IMS-APRM 2024**, University of Melbourne, Melbourne, Australia.

- "Roust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence" at **IMS-Bernoulli 2024 World Congress in Probability and Statistics**, Ruhr University, Bochum, Germany.

- "One-step Inference of Polychoric Correlation using the Density Power Divergence" at **IISA 2024**, Cochin University of Science and Technology, Cochin, India.

This page is intentionally left blank.

Bibliography

Aldrich J (1997). “RA Fisher and the making of maximum likelihood 1912-1922”. In: *Statistical science* 12.3, pp. 162–176.

Ali SM and Silvey SD (1966). “A general class of coefficients of divergence of one distribution from another”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 28.1, pp. 131–142.

Anderson TW and Darling DA (1952). “Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes”. In: *The annals of mathematical statistics*, pp. 193–212.

— (1954). “A test of goodness of fit”. In: *Journal of the American statistical association* 49.268, pp. 765–769.

Balakrishnan N and Lai CD (2009). *Continuous bivariate distributions*. Springer Science & Business Media.

Balakrishnan N, Martin N, and Pardo L (2017). “Empirical phi-divergence test statistics for the difference of means of two populations”. In: *AStA Advances in Statistical Analysis* 101, pp. 199–226.

— (2015). “Empirical phi-divergence test statistics for testing simple and composite null hypotheses”. In: *Statistics* 49.5, pp. 951–977.

Basak S and Basu A (2022). “The extended Bregman divergence and parametric estimation”. In: *Statistics*, pp. 1–20.

Basak S, Basu A and Jones MC (2021). "On the 'optimal'density power divergence tuning parameter". In: *Journal of Applied Statistics* 48.3, pp. 536–556.

Basu A, A Ghosh, et al. (2017). "A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator". In: *Electronic Journal of Statistics* 11.2, pp. 2741–2772.

Basu A, Harris IR, Hjort NL, and Jones MC (1998). "Robust and efficient estimation by minimising a density power divergence". In: *Biometrika* 85.3, pp. 549–559.

Basu A and Lindsay BG (1994). "Minimum disparity estimation for continuous models: efficiency, distributions and robustness". In: *Annals of the Institute of Statistical Mathematics* 46.4, pp. 683–705.

Basu A, Shioya H and Park C (2011). *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC.

Beran R (1977a). "Minimum Hellinger distance estimates for parametric models". In: *The annals of Statistics*, pp. 445–463.

— (1977b). "Robust location estimates". In: *The Annals of Statistics*, pp. 431–444.

Bernoulli D (1777). "Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda". In: *Acta Acad. Sci. Petropolit* 1, pp. 3–33.

Bickel P (1965). "On some robust estimates of location". In: *The Annals of Mathematical Statistics*, pp. 847–858.

Boos DD (1981). "Minimum distance estimators for location and goodness of fit". In: *Journal of the American Statistical association* 76.375, pp. 663–670.

— (1982). "Minimum anderson-darling estimation". In: *Communications in Statistics-Theory and Methods* 11.24, pp. 2747–2774.

Boscovich RJ (1757). "De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impessa".

In: *Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii* 4, pp. 353–396.

Box GEP (1953). "Non-normality and tests on variances". In: *Biometrika* 40.3/4, pp. 318–335.

Bregman Lev M (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR computational mathematics and mathematical physics* 7.3, pp. 200–217.

Burbea J and Rao CR (1982). "Entropy differential metric, distance and divergence measures in probability spaces: A unified approach". In: *Journal of Multivariate Analysis* 12.4, pp. 575–596.

Chauvenet W (1863). "Method of least squares". In: *Appendix to manual of Spherical and Practical Astronomy* 2, pp. 469–566.

Collins J and Wiens D (1989). "Minimax properties of M-, R-and L-estimators of location in Levy neighbourhoods". In: *The Annals of Statistics*, pp. 327–336.

Cox DR and Hinkley DV (1974). *Theoretical Statistics Chapman and Hall*.

Cramér H (1946). "A contribution to the theory of statistical estimation". In: *Scandinavian Actuarial Journal* 1946.1, pp. 85–94.

Cressie N and Read TRC (1984). "Multinomial goodness-of-fit tests". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 46.3, pp. 440–464.

Croux C, Flandre C, and Haesbroeck G (2002). "The breakdown behavior of the maximum likelihood estimator in the logistic regression model". In: *Statistics & Probability Letters* 60.4, pp. 377–386.

Croux C, Haesbroeck G, and Ruwet C (2013). "Robust estimation for ordinal regression". In: *Journal of Statistical Planning and Inference* 143.9, pp. 1486–1499.

Csiszár I (1964). "Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten". In: *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8, pp. 85–108.

Fernandes M and Néri B (2009). "Nonparametric entropy-based tests of independence between stochastic processes". In: *Econometric Reviews* 29.3, pp. 276–306.

Fernholz LT (2012). *Von Mises calculus for statistical functionals*. Vol. 19. Springer Science & Business Media.

Fisher RA (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 222.594-604, pp. 309–368.

— (1925). "Theory of statistical estimation". In: *Mathematical proceedings of the Cambridge philosophical society*. Vol. 22. 5. Cambridge University Press, pp. 700–725.

— (1934). "Two new properties of mathematical likelihood". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 144.852, pp. 285–307.

— (1935). "The logic of inductive inference". In: *Journal of the royal statistical society* 98.1, pp. 39–82.

Fujisawa H and Eguchi S (2006). "Robust estimation in the normal mixture model". In: *Journal of Statistical Planning and Inference* 136.11, pp. 3989–4011.

Gauss KF (1821). "Theoria combinationis obsercationunt erronbus minimis obnoxiae". In: *An English translation can be found in Gauss's work (1803-1826) on the Theory of Least Squares*.

Ghosh A and Basu A (2013). "Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression". In: *Electronic Journal of statistics* 7, pp. 2420–2456.

— (2015). "Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach". In: *Journal of Applied Statistics* 42.9, pp. 2056–2072.

Ghosh A. Harris IR, Maji A, Basu A, and Pardo L (2017). "A generalized divergence for statistical inference". In: *Bernoulli* 23.4A, pp. 2746–2783.

Guha A, Biswas A and Ghosh A (2021). "A nonparametric two-sample test using a general φ-divergence-based mutual information". In: *Statistica Neerlandica* 75.2, pp. 180–202.

Guha A and Chothia T (2014). "A two sample test based on mutual information". In: *Calcutta Statistical Association Bulletin* 66.1-2, pp. 39–54.

Hald A (1998). *A History of Mathematical Statistics from 1750 to 1930*. Vol. 314. Wiley-Interscience.

Hall P and Heyde CC (2014). *Martingale limit theory and its application*. Academic press.

Hamdan MA (1968). "On the structure of the tetrachoric series". In: *Biometrika* 55.1, pp. 261–262.

— (1970). "The equivalence of tetrachoric and maximum likelihood estimates of P in $2\times$ 2 tables". In: *Biometrika* 57.1.

— (1971). "On the polychoric series method for estimation of ρ in contingency tables". In: *Psychometrika* 36.3, pp. 253–259.

Hampel FR (1968). *Contributions to the theory of robust estimation*. University of California, Berkeley.

— (1971). "A general qualitative definition of robustness". In: *The annals of mathematical statistics* 42.6, pp. 1887–1896.

— (1974). "The influence curve and its role in robust estimation". In: *Journal of the american statistical association* 69.346, pp. 383–393.

Hampel FR. Ronchetti EM, Rousseeuw PJ and Stahel WA (1986). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

— (2011). *Robust statistics: the approach based on influence functions*. Vol. 196. John Wiley & Sons.

Hong C and Kim Y (2001). "Automatic Selection of the Turning Parametter in the Minimum Density Power Divergence Estimation". In: *Journal of the Korean Statistical Society* 30.3, pp. 453–465.

Huber PJ (1964). "Robust estimation of a location parameter". In: *The Annals of Mathematical Statistics*, pp. 73–101.

— (1965). "A robust version of the probability ratio test". In: *The Annals of Mathematical Statistics*, pp. 1753–1758.

— (1968). "Robust confidence limits". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 10.4, pp. 269–278.

— (1992). "Robust estimation of a location parameter". In: *Breakthroughs in statistics: Methodology and distribution*. Springer, pp. 492–518.

— (2011). "Robust statistics". In: *International encyclopedia of statistical science*. Springer, pp. 1248–1251.

Huzurbazar VS (1947). "The likelihood equation, consistency and the maxima of the likelihood function". In: *Annals of Eugenics* 14.1, pp. 185–200.

Iannario M, Monti AC and Piccolo D (2016). "Robustness issues for cub models". In: *Test* 25.4, pp. 731–750.

Iannario M and Piccolo D (2016). "A generalized framework for modelling ordinal data". In: *Statistical Methods & Applications* 25.2, pp. 163–189.

Iannario M. Monti AC, Piccolo D and Ronchetti EM (2017). "Robust inference for ordinal response models". In: *Electronic Journal of Statistics* 11.2, pp. 3407–3445.

Jing BY (1995). "Two-sample empirical likelihood method". In: *Statistics & probability letters* 24.4, pp. 315–319.

Jöreskog KG (1994). "On the estimation of polychoric correlations and their asymptotic covariance matrix". In: *Psychometrika* 59.3, pp. 381–389.

Kac M, Kiefer J and Wolfowitz J (1955). "On tests of normality and other tests of goodness of fit based on distance methods". In: *The Annals of Mathematical Statistics*, pp. 189–211.

Kiefer J (1959). "K-sample analogues of the Kolmogorov-Smirnov and Cramér-V. Mises tests". In: *The Annals of Mathematical Statistics*, pp. 420–447.

Kolmogorov A (1933). "Sulla determinazione empirica di una lgge di distribuzione". In: *Inst. Ital. Attuari, Giorn.* 4, pp. 83–91.

Kullback S and Leibler RA (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Lancaster HO and Hamdan MA (1964). "Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters". In: *Psychometrika* 29.4, pp. 383–391.

Le Cam L (1953). "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates". In: *Univ. Calif. Publ. in Statist.* 1, pp. 277–330.

Lee SY (1985). "Maximum likelihood estimation of polychoric correlations in r× s× t contingency tables". In: *Journal of Statistical Computation and Simulation* 23.1-2, pp. 53–67.

Lee SY and Chiu YM (1990). "Analysis of multivariate polychoric correlation models with incomplete data". In: *British Journal of Mathematical and Statistical Psychology* 43.1, pp. 145–154.

Lee SY and Lam ML (1988). "Estimation of polychoric correlation with elliptical latent variables". In: *Journal of Statistical Computation and Simulation* 30.3, pp. 173–188.

Lehmann EL and Casella G (2006). *Theory of point estimation*. Springer Science & Business Media.

Leung KM (1990). "ESTIMATION OF MULTIVARIATE POLYSERIAL AND POLYCHORIC CORRELATIONS VITH INCOMPLETE DATA". PhD thesis. The Chinese University of Hong Kong.

Lilliefors HW (1967). "On the Kolmogorov-Smirnov test for normality with mean and variance unknown". In: *Journal of the American statistical Association* 62.318, pp. 399–402.

Lindsay BG (1994). "Efficiency versus robustness: the case for minimum Hellinger distance and related methods". In: *The annals of statistics* 22.2, pp. 1081–1114.

Liu Y, Zou C, and Zhang R (2008). "Empirical likelihood for the two-sample mean problem". In: *Statistics & Probability Letters* 78.5, pp. 548–556.

Loeve M (1977). "Elementary probability theory". In: *Probability theory i*. Springer, pp. 1–52.

Mansuy R (2005). "An interpretation and some generalizations of the Anderson–Darling statistics in terms of squared Bessel bridges". In: *Statistics & probability letters* 72.2, pp. 171–177.

Markatou, M, Basu A, and Lindsay BG (1998). "Weighted likelihood equations with bootstrap root search". In: *Journal of the American Statistical Association* 93.442, pp. 740–750.

Maronna RA, Martin RD, Yohai VJ, and Salibián-Barrera M (2019). *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

Martinson EO and Hamdan MA (1972). "Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables". In: *Journal of Statistical Computation and Simulation* 1.1, pp. 45–54.

— (1975). "Algorithm AS 87: Calculation of the polychoric estimate of correlation in contingency tables". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.2, pp. 272–278.

McCullagh P (1980). "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.

McCullagh P and Nelder JA (2019). *Generalized linear models*. Routledge.

Micheas AC and Zografos K (2006). "Measuring stochastic dependence using ϕ-divergence". In: *Journal of Multivariate Analysis* 97.3, pp. 765–784.

Morales D, Pardo L, and Pardo MC (2001). "Likelihood divergence statistics for testing hypotheses about multiple population". In: *Communications in Statistics-Simulation and Computation* 30.4, pp. 867–884.

Moustaki I (2000). "A latent variable model for ordinal variables". In: *Applied psychological measurement* 24.3, pp. 211–223.

— (2003). "A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables". In: *British journal of mathematical and statistical psychology* 56.2, pp. 337–357.

Moustaki I and Victoria-Feser MP (2004). "Bounded-Bias Robust Estimation in Generalized Linear Latent Variable Models". In: *Available at SSRN 1763238*.

— (2006). "Bounded-influence robust estimation in generalized linear latent variable models". In: *Journal of the American Statistical Association* 101.474, pp. 644–653.

Mukherjee T, Mandal A and Basu A (2019). "The B-exponential divergence and its generalizations with applications to parametric estimation". In: *Statistical Methods & Applications* 28, pp. 241–257.

Müller CH and Neykov N (2003). "Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models". In: *Journal of Statistical Planning and inference* 116.2, pp. 503–519.

Nelder JA and Wedderburn R (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384.

Olsson U (1979). "Maximum likelihood estimation of the polychoric correlation coefficient". In: *Psychometrika* 44.4, pp. 443–460.

Owen A (2001). "Empirical likelihood. Chapman and Hall/CRC". In.

Öztürk Ö and Hettmansperger TP (1996). "Almost fully efficient and robust simultaneous estimation of location and scale parameters: A minimum distance approach". In: *Statistics & probability letters* 29.3, pp. 233–244.

Pardo L (2006). *Statistical Inference Based on Divergence Measures. Chapman Hall/CRC*.

Park C and Basu A (2004). "Minimum disparity estimation: Asymptotic normality and breakdown point results". In: *Bulletin of Informatics and Cybernetics* 36, pp. 19–33.

Parr WC (1981). "Minimum distance estimation: a bibliography". In: *Communications in Statistics-Theory and Methods* 10.12, pp. 1205–1224.

Parr WC and De Wet T (1981). "On minimum Cramer-von Mises-norm parameter estimation". In: *Communications in Statistics-Theory and Methods* 10.12, pp. 1149–1166.

Parr WC and Schucany WR (1980). "Minimum distance and robust estimation". In: *Journal of the American Statistical Association* 75.371, pp. 616–624.

— (1982). "Minimum Distance Estimation and Components of Goodness-Of-Fit Statistics". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 178–189.

Parzen E (1962). "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3, pp. 1065–1076.

Patra S, Maji A, Basu A, and Pardo L (2013). "The power divergence and the density power divergence families: the mathematical connection". In: *Sankhya B* 75.1, pp. 16–28.

Pearson ES and Sekar CC (1936). "The efficiency of statistical tools and a criterion for the rejection of outlying observations". In: *Biometrika* 28.3/4, pp. 308–320.

Pearson K (1900a). "I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195.262-273, pp. 1–47.

— (1900b). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302, pp. 157–175.

— (1920). "Notes on the history of correlation". In: *Biometrika* 13.1, pp. 25–45.

Peirce B (1852). "Criterion for the rejection of doubtful observations". In: *The Astronomical Journal* 2, pp. 161–163.

Piccolo D (2003). "On the moments of a mixture of uniform and shifted binomial random variables". In: *Quaderni di Statistica* 5.1, pp. 85–104.

Poon WY and Lee SY (1987). "Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients". In: *Psychometrika* 52.3, pp. 409–430.

Quiroga AM (1994). "Studies of the polychoric correlation and other correlation measures for ordinal variables." In.

Ritchie-Scott A (1918). "The correlation coefficient of a polychoric table". In: *Biometrika* 12.1/2, pp. 93–133.

Ronchetti EM and Huber PJ (2009). *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA.

Roscino A and Pollice A (2006). "A Generalization of the Polychoric Correlation Coefficient". In: *Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6-8, 2005*. Springer Science & Business Media, p. 135.

Rousseeuw PJ, Hampel FR, Ronchetti EM, and Stahel WA (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Roy S, Sarkar A, Ghosh A, and Basu A (2023). "Breakdown Point Analysis of the Minimum S-Divergence Estimator". In: *arXiv preprint arXiv:2304.07466*.

Ruckstuhl AF and Welsh AH (2001). "Robust fitting of the binomial model". In: *The Annals of Statistics* 29.4, pp. 1117–1136.

Savage LJ (1976). "On rereading RA Fisher". In: *The Annals of Statistics*, pp. 441–500.

Scalera V, Iannario M and Monti AC (2021). "Robust link functions". In: *Statistics* 55.4, pp. 963–977.

Scholz FW and Stephens MA (1987). "K-sample Anderson–Darling tests". In: *Journal of the American Statistical Association* 82.399, pp. 918–924.

Shapiro SS. and Wilk MB (1965). "An analysis of variance test for normality (complete samples)". In: *Biometrika* 52.3/4, pp. 591–611.

Silverman BW (2018). *Density estimation for statistics and data analysis*. Routledge.

Simpson DG (1987). "Minimum Hellinger distance estimation for the analysis of count data". In: *Journal of the American statistical Association* 82.399, pp. 802–807.

— (1989). "Hellinger deviance tests: efficiency, breakdown points, and examples". In: *Journal of the American Statistical Association* 84.405, pp. 107–113.

Singh P, Mandal A, and Basu A (2021). "Robust Inference Using the Exponential-Polynomial Divergence". In: *Journal of Statistical Theory and Practice* 15, pp. 1–22.

Stephens MA (1974). "EDF statistics for goodness of fit and some comparisons". In: *Journal of the American statistical Association* 69.347, pp. 730–737.

Stigler SM (1986a). "Laplace's 1774 memoir on inverse probability". In: *Statistical Science* 1.3, pp. 359–363.

— (1986b). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Student (1927). "Errors of routine analysis". In: *Biometrika*, pp. 151–164.

Sungur EA (1990). "Dependence information in parameterized copulas". In: *Communications in Statistics-Simulation and Computation* 19.4, pp. 1339–1360.

Tallis GM (1962). "The maximum likelihood estimation of correlation from contingency tables". In: *Biometrics* 18.3, pp. 342–353.

Tamura RN and Boos DD (1986). "Minimum Hellinger distance estimation for multivariate location and covariance". In: *Journal of the American Statistical Association* 81.393, pp. 223–229.

Thas O and Ottoy JP (2003). "Some generalizations of the Anderson–Darling statistic". In: *Statistics & probability letters* 64.3, pp. 255–261.

Timofeeva AY and Khailenko EA (2016). "Generalizations of the polychoric correlation approach for analyzing survey data". In: *2016 11th International Forum on Strategic Technology (IFOST)*. IEEE, pp. 254–258.

Tukey JW (1960). "A survey of sampling from contaminated distributions". In: *Contributions to probability and statistics*, pp. 448–485.

Vajda I (1972). "On the f-divergence and singularity of probability measures". In: *Periodica Mathematica Hungarica* 2.1-4, pp. 223–234.

— (1989). *Theory of statistical inference and information*. Springer.

van der Vaart AW (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Victoria-Feser MP and Ronchetti E (1997). "Robust estimation for grouped data". In: *Journal of the American Statistical Association* 92.437, pp. 333–340.

von Mises R (1939). "Sur les fonctions statistiques". In: *Bulletin de la Société Mathématique de France* 67, pp. 177–184.

— (1947). "On the asymptotic distribution of differentiable statistical functions". In: *The annals of mathematical statistics* 18.3, pp. 309–348.

Wald A (1949). "Note on the consistency of the maximum likelihood estimate". In: *The Annals of Mathematical Statistics* 20.4, pp. 595–601.

Warwick J and Jones MC (2005). "Choosing a robustness tuning parameter". In: *Journal of Statistical Computation and Simulation* 75.7, pp. 581–588.

Wied D and Weißbach R (2012). "Consistency of the kernel density estimator: a survey". In: *Statistical Papers* 53, pp. 1–21.

Wolfowitz J (1952). "Consistent estimators of the parameters of a linear structural relation". In: *Scandinavian Actuarial Journal* 1952.3-4, pp. 132–151.

— (1953). "Estimation by the minimum distance method". In: *Annals of the institute of Statistical Mathematics* 5.1, pp. 9–23.

Wolfowitz J (1954). "Estimation by the minimum distance method in nonparametric stochastic difference equations". In: *The Annals of Mathematical Statistics* 25.2, pp. 203–217.

— (1957). "The minimum distance method". In: *The Annals of Mathematical Statistics*, pp. 75–88.

Wu C and Yan Y (2012). "Empirical likelihood inference for two-sample problems". In: *Statistics and its Interface* 5.3, pp. 345–354.

Yong SQ and Zhao LC (2000). "Empirical likelihood ratio confidence intervals for various differences of two populations". In: *Journal of Systems Science and Complexity* 13.1, p. 23.