Indian Statistical Institute

Doctoral Dissertation

Some Contributions to Inference and Model/Variable Selection in High-Dimensional Problems

A thesis submitted to the Indian Statistical Institute in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics

Author: **Sayantan Paul**

Supervisor: **Dr. Arijit Chakrabarti**

Applied Statistics Unit (ASU)

Statistical Sciences Division (StatSD)

Acknowledgements

My doctoral thesis would not have been completed within the stipulated time without the help and support of several people whom I came across during my Ph.D life. Hence, it would be my great privilege to express my gratitude towards them.

First and foremost, I consider myself to be indebted to my thesis advisor, Professor Arijit Chakrabarti. It would have been an inhuman task for me to complete my thesis without his profound advice and continued support throughout my Ph.D. life. Discussions with him on different problems and his guidance, followed by thoughtful explanations, have helped me in several cases. I wish good health for him and his family and many more years of research in the days to come. I am also grateful to Dr. Prasenjit Ghosh of Texas A&M University, USA, whose thoughtful comments helped me to improve my overall quality of work to a large extent. I want to convey thanks to Dr. Jyothiska Datta of Department of Statistics, Virginia Polytechnic Institute and State University, for providing many useful comments and valuable suggestions on different occasions.

During the start of my Ph.D. journey, I had the opportunity to be present in the classes of Professor Tapas Samanta, Professor Subir Kumar Bhandari, Professor Ayanendranath Basu, Professor Atanu Biswas, Professor Anil Kumar Ghosh, Professor Kiranmoy Das, Professor Sourabh Bhattacharya. Their punctuality, way of teaching, humble behaviour towards the students, being involved in discussions before and after the classes, are all the features that should be treat to watch for a young researcher so that he/she can carry forward these in his/her future life. My sincere thanks to Professor Smarajit Bose and Professor Tapas Samanta for giving me the opportunity to be the teaching assistant of the courses they taught. I am also particularly grateful to Professor Subir Kumar Bhandari to giving me the opportunity to be the co-teacher of one of the courses he taught. I really enjoyed my time during the discussions with the students related to different topics. A lot of thanks to them, too.

I want to thank the present and former Deans of studies (Professor Biswabrata Pradhan and Professor Gopal Krishna Basak, respectively) and the Dean's office for helping me on different occasions. I want to convey my sincere thanks to the research fellow advisory committee, Professor Ayanendranath Basu and Dr. Abhik Ghosh, for providing helpful comments and suggestions throughout this journey. I also want to thank the PhD-DSc committee for helping whenever required.

A long journey like Ph.D sometimes becomes boring without the presence of a person with whom you can share your day-to-day experiences, ups and downs of the life. I consider myself to be fortunate enough in this case. Sayan, Biswadeep, Anik, Monitirtha, Meghna Di, Arnab, Amarnath Da, Mriganka, Subhasish Da, Sourabh, Chirayata, Arijit, Anirban, Soutik, Javed Da, Sampurna, Soumya Da, Subhankar Da, Subhajit Da, Rahul Da, Sujay Da, Arijit Da-all of you have now become an integral part of my life. Saying a thanks will probably be an injustice to the joyful and emotional memories you have given to me, which I would love to cherish for the rest of my life.

Many thanks to the Editors, Associate Editors, and the anonymous reviewers of various journals for raising many interesting questions and providing insightful suggestions, which helped me to improve the quality of my work. The Indian Statistical Institute provided me a research fellowship which helped me to be financially stable and also provided computing facilities throughout my journey-thanks for all of these. In order to obtain numerical results, I used the R, WinBUGS and Stan software. My sincere thanks to their contributors, who have kept them freely accessible.

Last, but not the least, my journey of Ph.D would never be easy if I would not get care, love and support from my family. It is not expressible within words to describe the sacrifice they have made through years to pursue my dream. I want to dedicate my thesis to my father Rintu Kumar Paul and my mother Sampa Paul. I find myself to be really privileged to have such a caring family. Finally, I want to thank God for keeping my family, my friends, my teachers, my near and dear ones safe and sound. Otherwise, it would be difficult for me to complete my thesis.

Sayantan Paul

Publications and Preprints

- A version of Chapter 2 of this thesis has been published online as the following:
  Paul, S. and Chakrabarti, A. (2025). Posterior contraction rate and asymptotic Bayes optimality for one group global-local shrinkage priors in sparse normal means problem. Ann Inst Stat Math (2025), 77:787-819. https://doi.org/10.1007/s10463-025-00932-1

- Chapter 3 is based on the following preprint:
  Paul, S., Ghosh, P. and Chakrabarti, A. (2025a). Sharp Asymptotic Minimaxity for One-Group Priors in Sparse Normal Means Problem. arXiv:2505.16428
  https://doi.org/10.48550/arXiv.2505.16428

- A version of Chapter 4 of this thesis has been published online as the following:
  Paul, S., Ghosh, P. and Chakrabarti, A. (2026). Consistent group selection using global-local shrinkage priors in sparse normal linear regression. Ann Inst Stat Math(2026), 78: 399-433.
  https://doi.org/10.1007/s10463-025-00950-z

- Chapter 5 is based on the following preprint:
  Paul, S. and Chakrabarti, A. (2024). Asymptotic Bayes Optimality for Sparse Count Data. arXiv:2401.05693 https://doi.org/10.48550/arXiv.2401.05693

Contents

Notations

For any two sequences of non-negative real numbers $\{a_n\}$ and $\{b_n\}$ with $b_n \neq 0$ for all sufficiently large n, we write the following:

- $a_n \sim b_n$ implies $\lim_{n\to\infty} a_n/b_n = 1$.
- By $a_n = O(b_n)$, and $a_n = o(b_n)$ we denote $|a_n/b_n| < M$ for all sufficiently large n, and $\lim_{n\to\infty} a_n/b_n = 0$, respectively, $M > 0$ being a global constant independent of n.
- We write $a_n \asymp b_n$ to denote that there exist two constants c_1 and c_2 such that $0 < c_1 \leq a_n/b_n \leq c_2 < \infty$ for sufficiently large n.

Likewise, for any two positive real-valued functions $f_1(\cdot)$ and $f_2(\cdot)$ with a common domain of definition that is unbounded to the right $f_1(x) \sim f_2(x)$ denotes $\lim_{x\to\infty} f_1(x)/f_2(x) = 1$.
Throughout this article, the indicator function of any set A will always be denoted $I\{A\}$.

- $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cumulative distribution function of a standard normal distribution, respectively.
- In the full Bayes procedure of the normal means model, we use the notation $\tau_n(q_n)$ which denotes $(q_n/n)\sqrt{\log(n/(q_n))}$.
- Let $G_{\mathcal{A}_n}$ and G_n denote the number of active groups and the total number of groups, respectively, with $G_{\mathcal{A}_n} \leq G_n \leq n$. Since we are interested in the sparse situation, we assume that $G_{\mathcal{A}_n} = o(G_n)$.
- $\boldsymbol{\beta}_g^0$ denote the true value of the vector of unknown coefficients $\boldsymbol{\beta}_g$.
- Let A matrix $\mathbf{A}$ of order $n \times m$ is said to be block orthogonal if for any two sub-matrices $\mathbf{A}_i$ (of order $n \times m_i$) and $\mathbf{A}_j$ (of order $n \times m_j$), we have $\mathbf{A}_i^{\mathrm{T}}\mathbf{A}_j = \mathbf{0}$ for all $i \neq j$.
- For any square matrix $\mathbf{A}$, $e_{\min}\mathbf{A}$ and $e_{\max}\mathbf{A}$ denote the minimum and the maximum eigenvalues of $\mathbf{A}$, respectively.
- For any square matrix $\mathbf{A}$, $\mathbf{A}^{\frac{1}{2}}$ is defined as $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$.

Throughout this article, we use the notation $\mathcal{D}$ to denote the data $\mathcal{D} = \{\mathbf{y}\}$.

Abstract

In today's world, we often come across situations where we need to infer about several parameters simultaneously based on a single dataset. This is known as the problem of simultaneous statistical inference and the need for such inference arises for data from the fields of biology, astronomy, genomics, bioinformatics, medicine, economics, finance, and image processing, just to name a few. As the name suggests, instead of being concerned about the optimality of inference for the individual parameters, the main interest here is in proposing procedures which can provide satisfactory results in the overall inferential problem involving all parameters of interest. The need for simultaneous inference may arise in the contexts of hypothesis testing, point estimation of several parameters, and building confidence intervals. Such inference becomes challenging when the number of parameters increases with sample size at the same or at a higher rate, i.e., when the problem becomes the high-dimensional. Equally challenging are the problems of model selection or variable selection for data of such kind.

In this thesis, our interest is in simultaneous statistical inference and model/variable selection in the high-dimensional setting. One of our main goals would be to propose new inference and model/variable selection procedures and study their properties through theory and simulations. We would also employ some of our proposed methods in suitable real-life examples to understand how they perform in actual applications. We would also like to theoretically address interesting questions about existing procedures widely used in such contexts. An important focus of this work will be the special situations when the number of significant (which may mean nonzero or of sufficiently large magnitude in appropriate contexts) parameters is small compared to the total number of parameters under consideration. These are the so-called "sparse" situations. The assumption of sparsity is natural and very common in the literature, e.g., in the high-dimensional regression setting and in the problem of inference on a high dimensional mean vector. Our approaches here will be built under the Bayesian setting where the unknown parameters are assumed to follow some prior probability distributions. A big focus of our theoretical work will be on studying the optimality (in the frequentist or Bayesian decision theoretic sense) of the new methods proposed or of existing methods.

A natural Bayesian approach under sparse settings is to model the parameters by two-group *spike-and-slab* priors. These priors are expressed as mixtures of a distribution degenerate at 0 (or highly concentrated near 0) and a distribution with high spread. However, as noted by many authors, inference with such priors can be computationally very challenging in high-dimensional situations and complex parametric frameworks. As an alternative to two-group mixture pri-

ors, there have been proposals in the literature to consider unimodal, continuous priors having sufficient mass around zero while possessing sufficiently heavy tails. Such priors are named as one-group "global-local" priors, the name "global-local" originating from the use of two sets of parameters to simultaneously enforce and modulate shrinkage at the global level and at the levels of individual parameters respectively. These are called the global shrinkage parameter and the local shrinkage parameters respectively.

Although serious efforts have gone into studying optimality of inference using one-group priors in sparse parametric settings, several interesting questions in this connection are still unanswered and several areas somewhat less trodden in the current literature. This thesis is a modest attempt to address some of these in the context of specific models. A large part of our work will focus on inference on the famous normal means model and the linear regression model. The thesis concludes with a study of inference on non-normal count data. Our study is based on a broad class of one-group global-local shrinkage priors covering a lot of popularly used priors including the horseshoe.

When the mean parameter of the sparse normal means model is modeled by a spike-and-slab prior, it has been shown in the literature that the corresponding posterior distribution contracts around the truth at a near minimax rate when the level of sparsity is unknown. This raises the question of whether the same phenomenon works when one uses a one-group prior instead of its two-group counterpart for the same problem. We provide an answer to this question in Chapter 2 of the thesis. In order to handle the unknown level of sparsity, we either estimate the global shrinkage parameter based on the data, an empirical Bayes approach, or model it by a non-degenerate absolutely continuous prior distribution on a suitable support in a full Bayes approach. We establish that the posterior distributions of the mean parameter vector, when modelled by broad classes of one-group priors, contract around the truth at a near minimax rate, for both the empirical Bayes and full Bayes approaches. Another interesting question is whether one-group priors can be used to form a good decision rule for the simultaneous testing problem of whether the means are zero or not when the means are truly generated from a two-group prior. In the latter half of this chapter, we are interested in answering this question, assuming that the level of sparsity is unknown. Considering a full Bayes approach, we are to establish that the Bayes risk of a decision rule using a broad class of one-group priors attains the Bayes risk of the optimal rule for the two-group settings asymptotically for a wide range of sparsity levels. The loss function considered is the additive symmetric $0-1$ loss measuring the number of misclassifications made by a multiple testing rule.

One of the main goals in a multiple hypothesis testing problem is to propose a decision rule which can control some overall measure of type I error rate, e.g., the False Discovery Rate (FDR). Given that a testing rule controls the FDR at some desired level, the next obvious question is whether the same rule can provide any control over the False Negative Rate (FNR). In this context, researchers are often interested in the optimal multiple testing rules in terms of controlling sum of certain type I and type II error measures. This can be answered by studying the minimax risk of multiple testing rules with respect to the corresponding loss functions. Very recently, for the normal means model, the expressions for the minimax risks based on misclassification (or Hamming) loss and the loss defined as the sum of FDP and FNP have been derived. It has also

been proved that the famous Benjamini-Hochberg (BH) procedure and an $\ell-$value based procedure using spike-and-slab priors attain the minimax risk asymptotically adaptively over broad sparsity levels. This motivates us to study whether decision rules based on one-group priors, if any, can enjoy such asymptotic optimality. When the level of sparsity is known, by choosing the global shrinkage parameter appropriately based on the knowledge of sparsity, we prove that the corresponding decision rule based on the broad class of one-group priors mentioned earlier achieves the minimax risk for both loss functions stated earlier. When the level of sparsity is unknown, some empirical Bayes and full Bayes versions of our decision rules can also attain the minimax risk. These results are proved in Chapter 3 of the thesis.

Another important problem of interest is variable/model selection in a high-dimensional normal linear regression model. In this thesis, we are interested in a situation where covariates under study in a linear regression model form groups due to their inherent similarities. This is known as the grouped regression model. In Chapter 4 of this thesis, motivated by the existing literature, we are interested in proposing a decision rule and resulting estimators, based on a general class of global-local priors, which have the "Oracle property", in the sense that, they achieve variable selection consistency and optimal estimation rate, respectively. We propose a decision rule which declares a group to be active if the ratio of the ℓ_2 norm of the posterior mean of the group regression coefficient to that of the least square estimate exceeds half. When the design matrix is block-orthogonal, we are able to establish that the global shrinkage parameter can be chosen in such a way that our proposed inference procedures have both selection consistency and optimal estimation rate, provided the level of sparsity is known. Even if the sparsity pattern is unknown, our modified decision rules, by either estimating the global shrinkage parameter from the data or by modeling a prior on it, can still enjoy the oracle property. In the simulation studies, our rules perform favorably compared to many existing methods in a variety of sparsity settings. Our methods, when applied to real datasets, also return encouraging results.

In Chapters 2 through 4 of the thesis, our main areas of interest were the situations when the observed data are generated from normal distributions of various parametric forms. However, depending on the problem of interest, the Gaussianity assumption is not always proper. In Chapter 5 of this thesis, we are interested in one such example where the data consists of counts of events, most counts being close to zero while some are moderate or large. The data is modelled by Poisson distribution with unknown means. Clearly the natural prior for such means would be a two-group mixture with large mass on the component concentrated near zero. Our interest is in finding at if one-group modelling is still a good alternative in this context, as seen in previously in Chapters 2 through 5 of this thesis. Specifically, one of the questions this leads us to is whether any decision rule for multiple testing based on one-group priors can approximate the optimal rule with respect to two-group priors in terms of risk when the sample size gets large. Towards that we first obtain the asymptotic expression for the optimal Bayes risk under two-group prior in appropriate asymptotic framework. The loss is taken to be additive symmetric $0-1$ loss. Next, irrespective of the sparsity pattern to be known or unknown, we establish that the Bayes risks corresponding to our proposed decision rules based on one-group priors attain the optimal Bayes risk, up to some multiplicative constant. Finally, the theoretical results are verified using simulation studies followed by a real data analysis. Many of the theo-

retical results derived in this thesis are the first of their kind in the literature in their specific contexts. Last but not the least, our theoretical results, simulations and to some extent real data analyses reinforce the logic of using appropriately chosen one-group priors as alternatives to their two-group counterparts in high-dimensional sparse parametric settings.

List of Tables

Chapter 1

Introduction

In today's world, we often come across situations where we need to infer about several parameters simultaneously based on a single dataset. This is known as the problem of simultaneous statistical inference and the need for such inference arises for data from the fields of biology, astronomy, genomics, bioinformatics, medicine, economics, finance, and image processing, just to name a few. As the name suggests, instead of being concerned about the optimality of inference for the individual parameters, the main interest here is in proposing procedures which can provide satisfactory results in the overall inferential problem involving all parameters of interest. The need for simultaneous inference may arise in the contexts of hypothesis testing, point estimation of several parameters, and building confidence intervals. Such inference becomes challenging when the number of parameters increases with sample size at the same or at a higher rate, i.e., when the problem becomes the high-dimensional. Equally challenging are the problems of model selection or variable selection for data of such kind.

In this thesis, our interest is in simultaneous statistical inference and model/variable selection in the high-dimensional setting. One of our main goals would be to propose new inference and model/variable selection procedures and study their properties through theory and simulations. We would also employ some of our proposed methods in suitable real-life examples to understand how they perform in actual applications. We would also like to theoretically address interesting questions about existing procedures widely used in such contexts. An important focus of this work will be the special situations when the number of significant (which may mean nonzero or of sufficiently large magnitude in appropriate contexts) parameters is small compared to the total number of parameters under consideration. These are the so-called "sparse" situations. The assumption of sparsity is natural and very common in the literature, e.g., in the high-dimensional regression setting and in the problem of inference on a high-dimensional mean vector. See, for example, Bühlmann and Van De Geer (2011). Our approaches here will be built under the Bayesian setting where the unknown parameters are assumed to follow some prior probability distributions. A big focus of our theoretical work will be on studying the optimality (in the frequentist or Bayesian decision theoretic sense) of the new methods proposed or of existing methods.

At first, we will review in Section 1.1 the relevant existing literature in this domain. This

will be followed (in Section 1.2) by descriptions and motivations of the specific problems we tried to address in different chapters of this work and our contributions.

1.1 Review of Literature

In this section, we present a review of the existing literature on simultaneous inference and variable/model selection in high-dimensional settings. Our focus will be on the most relevant parts (in the context of our work) of this vast literature. We first describe the famous normal means model, questions of interest related to it, and highlight existing approaches in dealing with them. These are presented in Subsections 1.1.2 through 1.1.6. Subsection 1.1.7 comprises similar discussions in the context of the widely used linear regression model.

1.1.1 The Normal Means Model

A canonical example of a high-dimensional model, where the number of parameters equals the number of observations, is the *normal means model* or the *sequence model*. Suppose we have an observation vector $\mathbf{X} \in \mathbb{R}^n$, $\mathbf{X} = (X_1, X_2, \cdots, X_n)$, where each observation X_i is expressed as,

$$X_i = \theta_i + \epsilon_i, i = 1, 2, \cdots, n, \tag{1.1}$$

where $\theta_1, \theta_2, \cdots, \theta_n$ are the unknown mean parameters and ϵ_i's are the random errors present in the model. Generally, ϵ_i's are assumed to be independent $\mathcal{N}(0, \sigma^2)$ random variables. Usually σ^2 is assumed to be known in the literature and hence, without loss of generality, a typical choice of σ^2 is $\sigma^2 = 1$. It is interesting to observe that these X_i's above need not be raw observations always. These might be suitably transformed versions (to validate the normality assumption) of some original observations.

In the context of this model, several interesting problems described below have been studied in the literature in some depth:

1. Simultaneous testing problem: we want to test simultaneously $H_{0i} : \theta_i = 0$ against $H_{1i} : \theta_i \neq 0$ for $i = 1, 2, \cdots, n$.
2. Estimation problem: we want to estimate the mean vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.
3. Uncertainty quantification for Bayesian credible sets for $\boldsymbol{\theta}$.

In Subsections 1.1.2 and 1.1.3, we discuss two prominent Bayesian approaches for modelling the unknown parameter vector $\boldsymbol{\theta}$. Subsections 1.1.4 through 1.1.6 review relevant existing approaches in the literature vis-a-vis these three problems.

1.1.2 Two-group priors for the normal means model

In the Bayesian paradigm, a popular way to model θ_i in (1.1), is through a mixture of two distributions given by,

$$\theta_i \overset{iid}{\sim} (1-p)\delta_0 + pF, i = 1, 2, \cdots, n, \tag{1.2}$$

where δ_0 denotes the Dirac measure which is degenerate at 0 and F denotes a non-degenerate absolutely continuous distribution over $\mathbb{R}$. According to this model, each θ_i is 0 with probability $(1-p)$ and with probability p it is non-zero, having distribution F. The distribution F in (1.2) is chosen in such a way that it can be used to model the intermediate and large values of θ_i. In a sparse situation, where most of the θ_i's are zero, p is assumed to be small. This type of modeling is widely studied in the literature, see, for e.g., Mitchell and Beauchamp (1988), Johnstone and Silverman (2004), just to name a few. The resulting marginal density of X_i is a mixture of the form,

$$X_i \overset{ind}{\sim} (1-p)\phi(x) + p\int_{\mathbb{R}} \phi(x-\theta)F(\theta)d\theta, i = 1, 2, \cdots, n. \tag{1.3}$$

Statistical models of the form (1.3) are popularly known as the two-group models and priors which can be represented as (1.2) are called the two-group or *spike-and-slab* priors. Here, the name *spike* comes from the degeneracy (and the resulting jump in probability at a particular point) of the prior distribution of θ_i under null and F is referred to as the *slab* part. Two-group models as above are considered the gold standard for modelling sparse high-dimensional data and they have drawn the attention of researchers, resulting in a rich literature on their optimality. Generally speaking, the role of the spike part is to capture the overall sparsity and shrink observations corresponding to noises towards zero. The slab part helps in allowing the model to account for possible large or moderately large signals. Hence, it is recommended to use a prior for the slab part which has a heavy (or slowly varying) tail.

Instead of modeling the "spike" as a Dirac measure at 0 (as given in (1.2)), one can alternatively consider a prior distribution which is highly concentrated around zero. An example of such a prior may be a normal distribution with mean 0 and a small variance ψ_0^2 with $\psi_0^2 > 0$. In such a case, for each $i = 1, \ldots, n$, θ_i is typically modeled as,

$$\theta_i \overset{iid}{\sim} (1-p)\mathcal{N}(0, \psi_0^2) + p\mathcal{N}(0, \psi^2), \tag{1.4}$$

where ψ_0^2 and ψ^2 are such that $\psi_0^2 \ll \psi^2$. So, one possible choice of ψ^2 is $\psi^2 = \psi_0^2 + \psi_1^2$, ψ_1^2 being very large compared to ψ_0^2. Such a choice of the variance parameter is essential for differentiating zero or small values of θ_i from those of moderate or large values. See in this context Bogdan et al. (2011) and the references therein. One can further generalize the formulation (1.4) by considering prior distributions on ψ_0^2, ψ_1^2 and p. See Scott and Berger (2006) in this context. Another variant of the spike-and-slab prior, known as *spike-and-slab* LASSO, was introduced by Ročková and George (2018), where they proposed to model the mean parameters by a mixture of Laplace densities. The hyperparameters corresponding to Laplace densities were chosen in such a way that the variance for the spike part is much smaller than that of the slab part. As mentioned by Banerjee et al. (2021), the whole class of spike-and-slab priors can be divided into two major categories; the *hard-spike-and-slab* prior, where the spike density is the Dirac measure at 0, and the *soft-spike-and-slab* prior, where the softness comes from the spike density which is not degenerate at 0, but is concentrated around zero.

1.1.3 One-group priors for the normal means model

Two-group priors, although most natural in sparse scenarios and studied extensively for the last few decades, can be computationally very challenging in problems with large model spaces, and specifically in high-dimensional situations. In particular, the rate of convergence of Markov Chain Monte Carlo (MCMC) algorithms can become very slow while exploring a large model space, see Narisetty and He (2014), Bhattacharya et al. (2015), Yang and Narisetty (2020) in this context. The high computational burden is mainly because of the fact that the two-group priors are a mixture of two distributions. In order to deal with this problem, one may search for "non-mixture" priors which are computationally more amenable but at the same time can mimic the two-group mixture priors in the sense that they have sufficient mass around zero and thick tails to allow large signals also. A major breakthrough in this direction came in Carvalho et al. (2009), where the famous horseshoe prior was introduced. Horseshoe prior is an example of a continuous "one-group" prior, having the following hierarchical formulation:

$$\begin{aligned} \theta_i|\lambda_i &\overset{ind}{\sim} \mathcal{N}(0, \lambda_i^2) \\ \lambda_i|\tau &\overset{ind}{\sim} C^+(0, \tau) \\ \tau|\sigma &\sim C^+(0, \sigma), \end{aligned} \tag{1.5}$$

where $C^+(0, a)$ denotes the half-cauchy distribution (with scale parameter a) supported on the positive part of the real line. In this paper, motivated by the work of Gelman (2006), they modeled the error variance σ^2 as $\pi(\sigma^2) \propto 1/\sigma^2$. Because of the nature of the prior, the horseshoe turns out to be computationally much easier to handle (see Carvalho et al. (2009)) and it also has many attractive properties for modelling sparse data. As established by Carvalho et al. (2010), it is unbounded at the origin and has sufficient mass for large values. Thus when the mean parameter is modeled by the horseshoe prior, the noise observations are shrunk towards zero while large observations remain almost unshrunk. The latter property is known as *tail robustness*.

As alluded to earlier, it is sometimes also possible that most of the parameters are close to zero, but not necessarily exactly zero. In such cases, two-group priors with a Dirac measure at 0 are not the ideal choice. Hence, in order to model unknown mean parameters, one can use either a two-group prior with both mixtures being continuous distributions or a unimodal continuous prior which can mimic its two-group counterpart, as mentioned before. However, due to the computational simplicity, a continuous prior having sufficient prior mass near the origin may be a more appropriate option to model this type of sparsity.

Note that, in the hierarchical formulation (1.5), the posterior mean of θ_i is obtained as

$$\mathbb{E}(\theta_i|\mathbf{X}, \tau, \sigma) = (1 - \mathbb{E}(\kappa_i|X_i, \tau, \sigma))X_i, \tag{1.6}$$

where $\kappa_i = 1/(1 + \lambda_i^2\tau^2)$. As a result, in the literature of one-group priors, $\mathbb{E}(\kappa_i|X_i, \tau, \sigma)$ is interpreted as a shrinkage coefficient and hence such priors are also known as one-group shrinkage priors. Of the two sets of hyperparameters (λ_i and τ) present in all these hierarchies,

λ_i's are called *local shrinkage parameters* and τ is called the *global shrinkage parameter*. Hence, alternatively, one-group priors are also called global-local shrinkage priors.

A broad class of one-group priors available in the literature can be expressed as scale mixtures of normals in the following hierarchical formulation:

$$\begin{aligned} \theta_i|\lambda_i,\tau &\overset{ind}{\sim} \mathcal{N}(0,\lambda_i^2\sigma^2\tau^2) \\ \lambda_i^2 &\overset{ind}{\sim} \pi_1(\lambda_i^2). \end{aligned} \tag{1.7}$$

Different choices of $\pi_1(\cdot)$ proposed by several authors have enriched the literature over the years. Some examples of such priors are the t-prior (Tipping (2001)), the Laplace prior (Park and Casella (2008)), the normal-exponential-gamma prior (Griffin and Brown (2005)), the horseshoe prior (Carvalho et al. (2009), Carvalho et al. (2010), Polson and Scott (2010), Polson and Scott (2012)), the three parameter beta normal priors (Armagan et al. (2011)), the generalized double Pareto priors (Armagan et al. (2013a)), the Dirichlet-Laplace prior (Bhattacharya et al. (2015)), and the horseshoe+ prior (Bhadra et al. (2017)). On the other hand, the global parameter τ is either used as a tuning parameter (Datta and Ghosh (2013), Ghosh et al. (2016), van der Pas et al. (2016)) or estimated using an empirical Bayes approach (van der Pas et al. (2014), van der Pas et al. (2017a), Ghosh and Chakrabarti (2017)) or modelled by an absolutely continuous prior on it (van der Pas et al. (2017a)). The error variance σ^2 is usually kept fixed for the theoretical study and for application it is modeled by Jeffreys prior as mentioned before.

In order to mimic the properties of their two-group counterparts, a class of one-group priors should possess the feature of being able to model the overall sparsity and simultaneous occurrence of a small fraction of true (and potentially large) signals. In this hierarchical formulation, the global shrinkage parameter τ is used to account for the overall (hence global) sparsity and hence a small value of it ensures a large mass around the origin. On the other hand, λ_i's are intended to act at the local (i.e. the individual) levels and hence they should have thick tails in order to ensure that the observations corresponding to true signals are left unshrunk but the noises are indeed shrunk towards zero.

Assuming $\sigma^2 = 1$, Theorem 1 of Polson and Scott (2010) consider the case when $\pi_1(\lambda_i^2)$ in (1.7) above is chosen as follows,

$$\pi_1(\lambda_i^2) \sim (\lambda_i^2)^{-a-1} e^{-\eta\lambda_i^2} L(\lambda_i^2) \text{ as } \lambda_i^2 \to \infty, \tag{1.8}$$

where $a > 0, \eta \geq 0$ and $L : (0,\infty) \to (0,\infty)$ is a measurable non-constant slowly varying function in Karamata's sense (see Bingham et al. (1987)), that is, $\frac{L(\alpha x)}{L(x)} \to 1$ as $x \to \infty$, for any $\alpha > 0$. They prove that under these conditions

$$\lim_{X_i \to \infty} |X_i - \mathbb{E}(\theta_i|\mathbf{X},\tau,\sigma)| = 2\eta. \tag{1.9}$$

As a consequence of (1.9), one-group priors whose local shrinkage parameter is modeled by (1.8) with $\eta > 0$, will not be able to shrink the observations towards the origin even if they are large. Examples of such priors are the Laplace or double exponential, the half-normal, the normal-gamma, the normal-inverse Gaussian prior distributions, i.e., the priors which have exponential

or lighter tails. Since the phenomenon of overshrinkage of large signals is not preferable in sparse high-dimensional situations, motivated by Polson and Scott (2010), using $\eta = 0$ in (1.8), Ghosh et al. (2016) considered a general class of priors of the form

$$\pi_1(\lambda_i^2) = K(\lambda_i^2)^{-a-1}L(\lambda_i^2), \tag{1.10}$$

where $K > 0$ is the constant of proportionality and a is any positive real number and L is a slowly varying function as defined earlier. They established that for different values of a and different functions $L(\cdot)$, the class of priors they consider contains among others, three parameter beta normal priors introduced by Armagan et al. (2011) and the generalized double Pareto prior due to Armagan et al. (2013a). Three parameter beta normal is a rich class containing the horseshoe (Carvalho et al. (2009)), the Strawderman-Berger prior, the Normal-exponential-gamma priors. More about one-group priors in the context of specific problems of interest will be discussed later in this chapter.

1.1.4 Simultaneous hypothesis testing for the sparse normal means model

Multiple hypothesis testing is an old problem in statistics. Research in this field, both in the frequentist and Bayesian domains, received a huge impetus after the publication of the seminal work of Benjamini and Hochberg (1995). Multiple testing has wide scope of applications in several fields like genomics, bioinformatics, medicine, economics, finance; particularly in the era of high-dimensional data.

As discussed earlier, for the sparse normal means model, each mean parameter θ_i is popularly modeled by a two-group prior as given in (1.2) (or (1.4)) with a small p. Consider the simultaneous testing problem $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ for $i = 1, 2, \cdots, n$. In this context, the model (1.2) naturally arises as follows. Consider latent (unobservable) random variables $\nu_i, i = 1, 2, \cdots, n$, with $\nu_i = 0$ indicating that H_{0i} is true while $\nu_i = 1$ indicating that H_{0i} is false. Here ν_i's are i.i.d. Bernoulli(p) random variables, for some $p \in (0, 1)$. Given $\nu_i = 0$, θ_i is assumed to have a unit mass at 0, and given $\nu_i = 1$, θ_i is assumed to follow a non-degenerate distribution F. This implies, marginally, each θ_i follows a two-group spike and slab distribution as given in (1.2). With $F = \mathcal{N}(0, \psi^2)$ for $\psi^2 > 0$, the marginal distribution of X_i is obtained as,

$$X_i \overset{ind}{\sim} (1-p)\mathcal{N}(0, \sigma^2) + p\mathcal{N}(0, \sigma^2 + \psi^2), i = 1, 2, \cdots, n. \tag{1.11}$$

When $\nu_i = 0$, $\theta_i = 0$ with probability 1, i.e., H_{0i} is true. On the other hand, when $\nu_i = 1$, i.e., H_{1i} is true, it is assumed that $\theta_i \sim \mathcal{N}(0, \psi^2)$, i.e., $P_{H_{1i}}(\theta_i \neq 0) = 1$. Hence, the testing problem is equivalent to testing simultaneously

$$H_{0i} : \nu_i = 0 \text{ versus } H_{1i} : \nu_i = 1 \text{ for } i = 1, 2, \cdots, n. \tag{1.12}$$

Using the same latent variable interpretation, one can simultaneously test between two "theories" or "beliefs" about the distribution of θ_i, where under theory 1 or $H_{0i}(\nu_i = 0)$, θ_i has a distribution highly concentrated around 0, and under theory 2 or $H_{1i}(\nu_i = 1)$, θ_i has a distribution with

larger spread, as in (1.4). We are interested in an asymptotic regime where the parameters p and ψ^2 vary with n. Since the focus is on the sparse situations, it is assumed that p is small and that it converges to zero as the number of hypotheses n diverges to infinity. On the other hand, ψ^2 is typically taken to be large in order to keep detecting true signals present in the model. So, we are looking at situations where the small fraction of (large) signals need to be picked from a large pool of noises. The oldest reference, known to us, of modeling this kind is due to Mitchell and Beauchamp (1988). In this work, instead of the normal prior, the "slab" part was modelled by a uniform prior with large variance. For the sparse normal means model, both empirical Bayes and full Bayes approaches in the context of multiple testing (using spike-and-slab priors) have been studied in the literature. See, e.g., Efron (2004), Efron (2008), Storey (2003), Bogdan et al. (2008), Scott and Berger (2006), just to name a few.

Under a two-group formulation with θ_i as in (1.4), with the number of hypotheses growing to infinity, Bogdan et al. (2011) obtained an expression for the optimal Bayes risk based on an additive loss function (allowing different losses for type I and type II errors), under conditions on the model parameters p and ψ and ratio of losses for type I and type II errors. A multiple testing rule is defined as *Asymptotically Bayes Optimal under Sparsity* (ABOS) if ratio of its risk to that of the Bayes rule (named as Bayes Oracle) converges to 1 as the number of hypotheses n tends to infinity. They established that under very mild conditions, the popular multiple testing procedures of Benjamini and Hochberg (1995) and Bonferroni enjoy the ABOS property.

Using an additive $0-1$ loss function (with equal losses for both type I and type II errors), the optimal rule (Bayes oracle of Bogdan et al. (2011)) is given by:

$$\text{Reject } H_{0i} \text{ if } \quad w_i = w_i(X_i) > \frac{1}{2}, i = 1, 2, \cdots, n, \tag{1.13}$$

where w_i is the posterior inclusion probability for θ_i, i.e., $P(\theta_i \neq 0|\text{data})$, namely the posterior probability of H_{1i} being true. It can be shown that for a two-group model, for large values of ψ^2, the posterior mean of θ_i can be expressed as,

$$\mathbb{E}(\theta_i|X_i, p, \psi, \sigma) \approx w_i(X_i)X_i, \tag{1.14}$$

Recalling from (1.6) the expression for the posterior mean of θ_i, and comparing with (1.14), it appears that the shrinkage factor $(1 - \mathbb{E}(\kappa_i|X_i, \tau, \sigma))$ is playing a role similar to the posterior inclusion probability $w_i(X_i)$ corresponding to the two-group model. Inspired by this observation, Carvalho et al. (2010) conducted a simulation study assuming that observations are coming from a sparse normal means model with different sparsity levels while the true signals are fixed at equispaced points ranging from 0.5 to 5.0. Analysis was done by using both horseshoe and two-group prior on the same data sets. Shrinkage weights arising from horseshoe prior, $\mathbb{E}(1 - \kappa_i|X_i, \tau, \sigma)$ were compared with the posterior inclusion probabilities, w_i for the two-group model. An interesting phenomenon observed in this study was that the shrinkage weights resulting from the horseshoe prior were very close to the inclusion probabilities corresponding to their two-group counterparts for a wide range of sparsity patterns. This implies that though the shrinkage coefficient can not be interpreted as the posterior inclusion probabilities in a one-group model (since the hierarchical formulation doesn't model the zero and non-zero means

separately), it nonetheless clearly mimics the role of the w_i of the two-group framework if one-group prior is applied in such a sparse setting.

Based on these simulation results, and keeping in mind the optimal rule (1.13) for two-group models, Carvalho et al. (2010) proposed the following intuitive decision rule for the simultaneous testing problem ($H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ for $i = 1, 2, \cdots, n$):

$$\text{reject } H_{0i} \text{ if } \quad \mathbb{E}(1 - \kappa_i | X_i, \tau, \sigma) > \frac{1}{2}, i = 1, 2, \cdots, n. \tag{1.15}$$

A natural question was to establish theoretically whether (1.15) can approximate the performance of the optimal Bayes rule when applied to situations where the data comes from a two-group mixture. Inspired by the simulation studies of Carvalho et al. (2010), Datta and Ghosh (2013) first investigated the asymptotic optimality of the decision rule (1.15) from a formal decision theoretic viewpoint. They established that the Bayes risk of rule (1.15) corresponding to the horseshoe prior attains the optimal Bayes risk (up to a constant) of its two-group counterpart provided the global shrinkage parameter τ is chosen appropriately based on the level of sparsity, i.e., it is approximately ABOS in the setting of Bogdan et al. (2011). Later Ghosh et al. (2016) extended their work by studying the same problem when the mean parameter is modeled by a broad class of global-local priors containing horseshoe and established the same result for their general class of priors. These works will be discussed in more detail later in this chapter and Chapter 2.

Over the past several decades, many different multiple testing procedures have been proposed in the literature, primarily aimed at controlling some overall measure of type I error at a predetermined level $\alpha \in (0, 1)$. The family-wise error rate (FWER), defined as the probability of making at least one false rejection, is a type I error measure which has been known and used for a very long time. The popular Bonferroni correction is the most widely used FWER controlling procedure. However, the FWER criterion can be too stringent, especially when the number of hypotheses being tested is very large. A more pragmatic approach is to try to control the rate of erroneous rejections instead, which is measured by the False Discovery Rate (FDR). Introduced by Benjamini and Hochberg (1995), FDR is defined as the expected proportion of incorrectly rejected null hypotheses among all rejections, with the proportion set to zero when there are no rejections. Under the assumption of independence of the test statistics, Benjamini and Hochberg (1995) produced a specific step-up multiple testing procedure that can control FDR at the desired tolerance level $\alpha \in (0, 1)$. Following this seminal work, numerous other FDR-controlling procedures have been proposed in the literature. Some relevant references include works by Benjamini and Yekutieli (2001), Storey (2002), Storey (2003), Benjamini et al. (2006), Romano and Shaikh (2006), Sarkar (2007), Sarkar (2008), Sun and Tony Cai (2009), Sarkar and Guo (2009) and Guo et al. (2014), among others.

Formally, a multiple testing procedure $\boldsymbol{\psi}$ is a measurable function of the data, that is, $\psi \colon \mathbf{X} \in \mathbb{R}^n \mapsto (\psi_i(\mathbf{X}))_{i=1,\ldots,n} \in \{0, 1\}^n$, where $\psi_i \equiv \psi_i(\mathbf{X}) = 1$ implies rejecting the i^{th} null hypothesis, H_{0i}. Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and any testing procedure $\boldsymbol{\psi}$, the FDR of $\boldsymbol{\psi}$ at $\boldsymbol{\theta}$

is defined as

$$FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_\theta \left[\frac{\sum_{i:\theta_i=0} \psi_i}{max(\sum_{i=1}^n \psi_i, 1)} \right]. \tag{1.16}$$

Given that a testing rule controls the FDR at some desired level, the next obvious question is whether the same rule can provide any control over the type II error, which would imply power properties for it. This probability of error is measured by the False Negative Rate (FNR). For a testing rule $\boldsymbol{\psi}$, FNR at $\boldsymbol{\theta}$ is of the form

$$FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_\theta \left[\frac{\sum_{i:\theta_i\neq 0} (1 - \psi_i)}{max(\sum_{i=1}^n 1_{\theta_i\neq 0}, 1)} \right]. \tag{1.17}$$

A question that naturally arises in this context is the optimality of a multiple testing procedure in terms of controlling the sum of type I and type II error measures. In the case of a simple hypothesis testing problem involving one test, this question has been explored in depth by Ingster and Suslina (2012) and Giné and Nickl (2021). Recently, similar questions have been studied in relation to multiple hypothesis testing. The multiple testing risk at $\boldsymbol{\theta}$ for a testing procedure $\boldsymbol{\psi}$ takes the following form:

$$R(\boldsymbol{\theta}, \boldsymbol{\psi}) = FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) + FNR(\boldsymbol{\theta}, \boldsymbol{\psi}). \tag{1.18}$$

Given the risk function, a natural problem is to try to get an optimal rule with respect to some benchmark criterion. We will be focusing on the widely used minimaxity criterion defined as

$$R(\boldsymbol{\Theta}) = \inf_{\boldsymbol{\psi}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R(\boldsymbol{\theta}, \boldsymbol{\psi}), \tag{1.19}$$

where the infimum is taken over all multiple testing rules and $\boldsymbol{\Theta}$ is the parameter set.

Our interest is in an asymptotic framework where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is assumed to be sparse, that is, it belongs to the set

$$\ell_0[q_n] = \{\theta \in \mathbb{R}^n : ||\boldsymbol{\theta}||_0 \leq q_n\}, \text{ with } ||\boldsymbol{\theta}||_0 := \#\{1 \leq i \leq n : \theta_i \neq 0\},$$

consisting of vectors that have at most q_n many nonzero coordinates, where $q_n \to \infty$ as $n \to \infty$ such that $q_n = o(n)$. In the following references $\boldsymbol{\Theta}$ is taken as some subset of $\ell_0[q_n]$. Typically, one takes the parameter space $\boldsymbol{\Theta}$ to be *as large* as possible, while maintaining that the minimax risk $R(\boldsymbol{\Theta})$ is strictly less than 1, due to the fact that the naive testing rule $\psi_i = 0$ for all i, which fails to reject any null hypotheses regardless of the data, results in a minimax risk of exactly 1. Therefore, it is crucial to include more sophisticated testing procedures that yield a minimax risk $R(\boldsymbol{\Theta})$ that is strictly less than 1. At the same time, it is essential to ask how large the minimum magnitude of the signals should be in order to effectively distinguish them from the noise present in the data. A natural condition in this context is the "beta-min" condition (which will be formally defined in Chapter 3), ensuring that the magnitudes of the non-null θ_i values exceed a specific threshold. This has not been considered in the literature only until recently; see, for instance Arias-Castro and Chen (2017) and Salomond (2017) in this context.

Arias-Castro and Chen (2017) was the first to study this problem under the beta-min condition, assuming that all non-null signals have the same magnitude, denoted $M \equiv M(n, q_n)$, in the sparse normal means model. They established that if $M = a\sqrt{2\log\left(\frac{n}{q_n}\right)}$ for some $a < 1$, then there does not exist any thresholding multiple testing rule (such as the BH method or variants thereof) that yields a non-trivial minimax risk uniformly across the corresponding parameter space $\boldsymbol{\Theta}(M) \subset \ell_0[q_n]$. They further showed that if $M = a\sqrt{2\log\left(\frac{n}{q_n}\right)}$ for some $a > 1$, the tolerance parameter α corresponding to the Benjamini-Hochberg (BH) procedure can be appropriately chosen so that the minimax risk based on the BH procedure converges to 0 as the dimension n increases towards infinity. As highlighted by Abraham et al. (2024), these two results serve as a significant foundation for confirming the importance of the beta-min condition in achieving a non-trivial minimax risk in this context. In other words, the results indicate that for a multiple testing rule based on thresholding, the detection boundary for signals should be close to the universal threshold $\sqrt{2\log\left(\frac{n}{q_n}\right)}$. However this work does not give the precise asymptotic boundary for the signal strength M at which $R(\boldsymbol{\Theta})$ varies within from 0 to 1.

Abraham et al. (2024) identified and fully characterized the phase transition of multiple testing performance in sparse high-dimensional models, showing both fundamental limits and how practical procedures (BH and empirical-Bayes ℓ-values) sharply attain these limits-even adaptively. The authors of this paper focused on identifying sharp boundaries that separate the region where reliable multiple testing is possible from where it is impossible, under the minimax criterion. They derive precise asymptotic expressions for the minimax risk (1.19). Assuming a beta-min condition on the signals, and a sparsity framework that allows for an arbitrary level of sparsity, they showed that, for the sparse Gaussian sequence model, if the signal magnitudes are larger than the phase transition boundary, namely,

$$a_b = \sqrt{2\log\left(\frac{n}{q_n}\right)} + b, \text{ for } b \in \mathbb{R},$$

the minimax risk in (1.19) over the corresponding parameter space $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}(b, q_n)$, defined in Section 3.2 (equation (3.6)), can be expressed in the form

$$R(\boldsymbol{\Theta}) = 1 - \Phi(b) + o(1), \tag{1.20}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. The aforesaid result also extends to cases when $b \equiv b_n \to \infty$ and $b \equiv b_n \to -\infty$. This expression establishes a phase transition: testing becomes nearly impossible below this threshold, and achievable above it.

Abraham et al. (2024) also considered the classification loss or, the Hamming loss and evaluated the closed-form expression of the corresponding asymptotic minimax risk. They showed that classical procedures—such as the Benjamini-Hochberg (BH) method with a carefully tuned level and an empirical Bayes approach using ℓ-values - can attain this sharp boundary adaptively, that is, without prior knowledge of the sparsity level q_n. Beyond the classical beta-min

assumption, the authors extend their results to more general signal distributions and error measures, confirming the robustness of the boundary phenomenon. To the best of our knowledge, the only work related to the minimax risk based on one-group priors is due to Salomond (2017). This work will be discussed in detail in Subsection 1.2.2 and Chapter 3.

1.1.5 Minimax Estimation in the sparse normal means problem

In this subsection our interest is in the simultaneous estimation of the vector $\boldsymbol{\theta}$ of mean parameters in the normal means model (1.1). A popular benchmark for evaluation of the performance of an estimator is the minimax risk (with respect to appropriate loss function). One of the most popular loss functions is the squared ℓ_2 norm, defined as $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{n}(\widehat{\theta}_i - \theta_i)^2$. There have been attempts to derive minimax risks w.r.t. several parameter spaces. Here we are interested in the situation where the mean vector $\boldsymbol{\theta}$ is assumed to lie in $\ell_0[q_n]$, where $q_n \to \infty$ such that $q_n = o(n)$.

Donoho et al. (1992) established that for sparse normal mean model the minimax risk over the parameter space $\boldsymbol{\Theta} = \ell_0[q_n]$ is obtained as,

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}_0 \in \ell_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} ||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2 = 2q_n \log\left(\frac{n}{q_n}\right)(1 + o(1)). \tag{1.21}$$

An estimator $\widehat{\boldsymbol{\theta}}$ satisfying

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} ||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2 \leq C q_n \log\left(\frac{n}{q_n}\right)(1 + o(1)), \tag{1.22}$$

for some global constant $C(\geq 2)$, is called an asymptotically minimax "rate optimal" estimator of $\boldsymbol{\theta}$ over $\ell_0[q_n]$. If the estimator satisfies

$$\sup_{\boldsymbol{\theta}_0 \in \ell_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} ||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2 \sim 2q_n \log\left(\frac{n}{q_n}\right)(1 + o(1)), \tag{1.23}$$

it is called an asymptotically minimax estimator over $\ell_0[q_n]$, where $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. Using the penalized least squares approach, Donoho et al. (1992) proposed a hard thresholding estimator of $\boldsymbol{\theta}$ and proved that it is asymptotically minimax (i.e. satisfies (1.23)) for estimating $\boldsymbol{\theta}$ over $\ell_0[q_n]$. Donoho and Johnstone (1994) proposed a soft thresholding estimate of $\boldsymbol{\theta}$ and proved that estimator enjoys near-minimax property asymptotically. Based on the work of Benjamini and Hochberg (1995), Abramovich et al. (2006) proposed an adaptive thresholding estimator and proved the asymptotic minimaxity of the estimator under various loss functions and different parameter spaces, including $\ell_0[q_n]$.

In addition to such frequentist results, there are several Bayesian approaches in the literature leading to the estimates of $\boldsymbol{\theta}$ and their evaluation vis-a-vis the minimax criterion. Johnstone and Silverman (2004) used two-group spike and slab priors to estimate $\boldsymbol{\theta}$. They considered the coordinate-wise posterior median of $\boldsymbol{\theta}$ and used an empirical Bayesian way to estimate p by employing the Marginal Maximum Likelihood Estimate(MMLE) of p. Abramovich et al. (2007)

proposed a broad class of MAP estimators and proved their asymptotic optimality in minimax sense under the usual quadratic loss function and different levels of sparsity. Based on a data-dependent two-group formulation, Martin and Walker (2014) proved that the posterior mean of $\boldsymbol{\theta}$ attains the minimax risk (1.21), up to some multiplicative constant. Several other works related to empirical Bayes procedures are Johnstone and Silverman (2005), Yuan and Lin (2005), Zhang (2005), Jiang and Zhang (2009), Brown and Greenshtein (2009), Koenker and Mizera (2014), Martin et al. (2017) and references therein. The first work in a full Bayes approach in this context was due to Castillo and Van Der Vaart (2012). They established that for their chosen class of two-group priors, the posterior distribution of $\boldsymbol{\theta}$ contracts around the truth at the minimax rate w.r.t. squared ℓ_2 norm. The exact definition of posterior contraction rate will be given in Chapter 2. Modeling the mean parameter by continuous priors is a relatively new approach in this context. The first work in this field using one-group priors was done by Bhattacharya et al. (2012). Using Dirichlet-Laplace prior to model the mean parameters, they established the resultant posterior distribution of $\boldsymbol{\theta}$ contracts around the truth at the minimax rate (using squared ℓ_2 loss) with proper choice of hyperparameters. Later van der Pas et al. (2014) proved the same rate of contraction of the mean vector when it is modeled by the Horseshoe prior with appropriately chosen τ. This result was generalized by Ghosh and Chakrabarti (2017) for the broad class of priors mentioned in (1.7) and (1.10). Recently, Song (2020) studied how, in the normal means model, the tail of the prior distribution of the mean parameter quantifies the spread of its posterior distribution. They considered priors (given the global parameter τ) for which the tail decays polynomially with order $\alpha > 1$. They proved that the polynomial order α directly affects the multiplicative constant of the posterior contraction rate in the sense that if the polynomial order is sufficiently close to 1 and the global shrinkage parameter τ is chosen properly, then the posterior distribution of $\boldsymbol{\theta}$ concentrates around the true value at the minimax rate when the level of sparsity is known. When the level of sparsity is unknown, they proposed a full Bayesian approach using a Beta prior on τ and derived similar results in terms of posterior concentration rate. The importance of the choice of α was also confirmed in the numerical studies, as it produced better results (in terms of posterior contraction rate) for small values of α (e.g., $\alpha = 1.1$) compared to the larger ones. They illustrated the adaptiveness of their method on a real dataset by analyzing gene expression data for cancer research. However, in our study, we were mainly motivated by the works of van der Pas et al. (2014), Ghosh and Chakrabarti (2017) and van der Pas et al. (2017a) and hence did not consider the priors studied by Song (2020). These works will be discussed in detail in Chapter 2.

1.1.6 Uncertainty quantification

One important related question of serious interest for researchers is whether a given Bayesian confidence set has both the right size(corresponding to optimal minimax contraction rate) and also the right coverage probability (i.e., having at least $1 - \alpha$ coverage probability over the parameter space). The impossibility theorem (e.g., Li (1989)) says that irrespective of the approach, whether frequentist or Bayesian, there does not exist any confidence set which is adaptive both ways, in the sense that the size corresponds to the minimax rate for the

unknown sparsity level and is honest, in the sense that it attains the nominal frequentist coverage probability, uniformly over all $\boldsymbol{\theta}_0$, in the parameter space, when $\boldsymbol{\theta_0}$ is the vector of true means. See in this context van der Pas et al. (2017b) and the discussion thereof Martin et al. (2017). Therefore it is of interest to quantify or describe the subset of parameters where the best of both worlds occur simultaneously.

To the best of our knowledge, the only work on uncertainty quantification for the sparse normal means model based on a continuous prior is due to van der Pas et al. (2017b). The unknown mean parameter vector $\boldsymbol{\theta}$ was modelled by the famous horseshoe prior. Under the assumption that the level of sparsity is unknown, they considered two situations, empirical Bayes and full Bayes procedures. For the empirical Bayes approach, motivated by the results of their previous work (van der Pas et al. (2017a)), they estimated the sparsity level by the MMLE of τ, the global shrinkage parameter. On the other hand, they also studied the situation when τ is modeled by a non-degenerate prior satisfying conditions similar to those of van der Pas et al. (2017a).

Given the hyperparameter τ, they considered a credible ball which is of the form

$$\widehat{C}_n(\tilde{L}, \tau) = \{\theta : \left\|\theta - \widehat{\theta}(\tau)\right\|_2 \leq \tilde{L}\widehat{r}(\alpha, \tau)\}, \tag{1.24}$$

where $\widehat{\theta}(\tau)$ is the posterior mean of θ, $\tilde{L}$ is a positive constant. For any given $\alpha \in (0, 1)$, $\widehat{r}(\alpha, \tau)$ is obtained as

$$\Pi(\theta : \left\|\theta - \widehat{\theta}(\tau)\right\|_2 \leq \widehat{r}(\alpha, \tau)|\mathbf{X}, \tau) = 1 - \alpha.$$

For the empirical Bayes approach, $\widehat{\theta}(\tau)$ and $\widehat{r}(\alpha, \tau)$ in (1.24) are replaced by $\widehat{\theta}(\hat{\tau})$ and $\widehat{r}(\alpha, \widehat{\tau})$, respectively, where $\widehat{\tau}$ is the MMLE of τ. Similarly, for the full Bayes procedure, those are replaced by $\widehat{\theta}$ and $\widehat{r}$, respectively, where $\widehat{\theta}$ denotes the posterior mean of θ and $\widehat{r}$ is obtained as

$$\Pi(\theta : \left\|\theta - \widehat{\theta}\right\|_2 \leq \widehat{r}(\alpha)|\mathbf{X}) = 1 - \alpha.$$

They also considered marginal credible intervals of individual coordinates of θ_{0i}, but we are not discussing it further for the review. The adaptive versions of the credible interval are defined similarly as done for those of credible balls. They proved that though adaptive versions of the credible sets are not honest for the full parameter space, they have good frequentist coverage properties and optimal size simultaneously for parameters $\boldsymbol{\theta}_0$ if a fraction of its non-zero coordinates exceed the threshold of detectability $\sqrt{2\log\left(\frac{n}{q_n}\right)}$ or more generally, if the squared l_2 norm of the vector of "non-detectable" non-zero coordinates is not too large.

For the same problem, Belitser and Nurushev (2020) introduced a condition named *Excessive Bias Restriction*(EBR) that restricts bias by a multiple of variability at a parameter value. Based on this condition, they obtained credible regions having frequentist coverage and also having an adaptive optimal size on some restricted parameter space. As already mentioned, it is impossible to obtain honest coverage with adaptive size ball, irrespective of the method, Bayesian or not. Hence, restriction of the parameter space becomes essential, and Excessive Bias Restriction condition comes into play. Later, Castillo and Szabó (2020) studied the same problem

and modeled the mean parameter under the alternative by a heavy tailed distribution. They estimated the sparsity level p by the marginal maximum likelihood empirical Bayes approach (MMLE) and obtained similar coverage results to those of Belitser and Nurushev (2020) for adaptive credible sets. They also showed that the EBR condition is necessary in a minimax sense.

1.1.7 Inference and variable selection for the linear Regression model

A linear regression model consisting of n observations and p covariates is formally expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1.25}$$

Here $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ design matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients. We further assume that the unobservable error term $\boldsymbol{\epsilon}$ has a $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ distribution. Selecting the most relevant predictors in a regression model and also estimating the corresponding regression coefficients are classical problems in statistics. In this subsection, our focus will be on two broad classes of methods available in the literature in this context, with special emphasis on sparse problems and grouped regression (to be introduced shortly). The first class is based on of certain penalized likelihood methods where the log-likelihood is penalized by some measure of the magnitude of the vector of regression coefficients. This makes full sense when the proportion of regressors with non-null (or sufficiently large) effects is known to be small, i.e., the sparse setting. Assuming σ^2 as known, these methods estimate $\boldsymbol{\beta}$ by minimizing an objective function of the following form

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|),$$

where $p_{\lambda_j}(|\beta_j|)$ is an appropriately chosen penalty function, and $\lambda_j > 0$ is the penalty parameter. By choosing $\lambda_j = \lambda$ for all j, $p_\lambda(|\beta_j|) = \lambda\beta_j^2$ corresponds to the Ridge regression of Hoerl and Kennard (1970), while $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ corresponds to the LASSO estimator of Tibshirani (1996). The LASSO is very popular due to its ability to perform both estimation and variable selection simultaneously (by estimating many components of $\boldsymbol{\beta}$ as exactly zero). We also mention the SCAD of Fan and Li (2001), the elastic net of Zou (2006)), the MCP of Zhang (2010) and the adaptive LASSO as examples of methods of the first kind.

The second class consists of Bayesian methods obtained from two key approaches. In the first approach, a Bayesian starts by assigning a prior probability distribution to the model space, namely a collection of models, where, under each model, a subset of the regression coefficients (β_i's) are small (or exactly zero). In contrast, the remaining β_i's are large (or nonzero). Under each model, a prior quantifying the uncertainty about the parameters (regression coefficients) is then specified. The model with the highest posterior probability (the Bayes rule with respect to 0-1 loss) is typically the model of choice. Variable selection is carried out using the model, thus chosen, by only keeping those variables/predictors in the linear regression whose regression coefficients are supposed to be large (or nonzero) according to this model. See, for instance,

George and McCulloch (1993), George and McCulloch (1997), Scott and Berger (2010), Bayarri et al. (1997), Maruyama and George (2011), Mukhopadhyay et al. (2015) and Liang et al. (2008), in this connection. Two popular choices of priors on the regression coefficients in the above references are (i) the g-prior of Zellner (1986) (or variants of it) on the nonzero regression coefficients under a given model and (ii) independent spike and slab priors on the regression coefficients. In the second Bayesian approach, one directly specifies a prior distribution on the parameters of the full model, namely the model with all the possible parameters/regressors. Model/variable selection is performed by appropriately using the posterior distribution of the parameter vector. See, for instance, Polson and Scott (2012), Tang et al. (2018), Li and Lin (2010), Bhattacharya et al. (2015) for examples of this kind. The priors used therein are the continuous one-group priors, which have been introduced before.

A natural modelling for regression, namely, two-group spike and slab priors on regression coefficients is obtained by associating a latent random vector $(\gamma_1, \ldots, \gamma_p)$ with $(\beta_1, \ldots, \beta_p)$ such that γ_i are independent and identically distributed with $\gamma_i = 0$ with probability $(1 - \nu)$ and $\gamma_i = 1$ with probability ν, and $\beta_i|\gamma_i = 0$ has the spike distribution and $\beta_i|\gamma_i = 1$ has the slab distribution. Several choices of *spike-and-slab* priors have been proposed in the literature, in the context of regression, including those by Mitchell and Beauchamp (1988), George and McCulloch (1993), Geweke (1996), Narisetty and He (2014), and Ročková and George (2018), among others.

As already mentioned in the discussion on the normal means model, over the past fifteen years, the literature has presented various proposals for modeling sparsity through hierarchical one-group global-local "shrinkage" priors. In the context of (1.25), such priors model β_j as follows:

$$\beta_j \mid \lambda_j^2, \sigma^2, \tau^2 \overset{ind}{\sim} \mathcal{N}(0, \lambda_j^2\sigma^2\tau^2), \quad \lambda_j^2 \overset{ind}{\sim} \pi(\lambda_j^2), \quad (\tau, \sigma^2) \sim \pi(\tau, \sigma^2). \tag{1.26}$$

It is important to note that some penalized likelihood estimates can be derived as *maximum a posteriori* (MAP) estimates corresponding to appropriate one-group shrinkage priors for the unknown regression coefficients. For example, Park and Casella (2008) observed while proposing the Bayesian LASSO, that the usual LASSO estimator can be thought of as a MAP estimator corresponding to double exponential prior for the regression coefficients. The double-exponential prior, in fact, can be expressed as a scale mixture of normals which can be written in the form of a one-group prior as above. Two important works in this context are due to Hans (2009) and Hans (2011). Hans (2009) provides a direct characterization of the posterior distribution of the regression coefficients, which makes the computation quite straightforward. They also emphasize point estimation using the posterior mean, which facilitates the prediction of future observations via the posterior predictive distribution. This idea was extended in their later work, namely Hans (2011). This article broadens the scope of the Bayesian connection by providing a complete characterization of a class of prior distributions that generate the elastic net estimate as the posterior mode. They also provided MCMC methods and an R package related to posterior inference in this context.

Unlike a spike-and-slab prior, the one-group priors don't model the important and the non-important variables separately. Hence, it is of importance to form a decision rule for the selection of active covariates when one-group priors are being used. One of the first works in this context

is due to Li and Lin (2010), where they modeled the regression vector $\boldsymbol{\beta}$ (conditioned on σ^2) as $\pi(\boldsymbol{\beta}|\sigma^2) \sim \exp\{-\frac{1}{2\sigma^2}(\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2)\}$ followed by a Jeffreys prior for σ^2 ($\pi(\sigma^2) \sim \frac{1}{\sigma^2}$). With this, the posterior mode of $\boldsymbol{\beta}$ becomes the elastic net estimator of $\boldsymbol{\beta}$. Li and Lin (2010) proposed a variable selection rule using the Bayesian elastic net credible interval criterion depending on whether the credible interval of β_i contains zero or not. On the other hand, for the same problem, Bhattacharya et al. (2015) proposed Dirichlet-Laplace prior distributions and provided a heuristic argument for variable selection based on an approximate measure of model size. It was noted earlier that for the normal means model, the decision rule (1.15) is really based on the ratio of the posterior mean (under one-group model) of the parameter of interest and the observation. For the regression problem, this idea was extended by Tang et al. (2018) by proposing a rule that selects a variable if the ratio of the posterior mean of its regression coefficient to the corresponding ordinary least squares estimate is greater than half. This rule is named as the half-thresholding rule (HT).

Very recently, Song and Liang (2023) used shrinkage priors in the problem of variable selection in the linear regression framework and established that if a shrinkage prior has a dominating peak near zero and a heavy tail, it achieves nearly optimal L_1 and L_2 posterior contraction rates. Later, they introduced a prior-dependent hard-thresholding rule and proved that this rule can consistently recover the true sparse model under high-dimensional settings. The numerical results indicate that their proposed decision rule based on continuous shrinkage priors often outperform standard regularization techniques like Lasso and SCAD in terms of false positive rate. This work emphasizes that shrinkage priors provide a powerful bridge in high-dimensional statistics, offering the theoretical rigor of hierarchical spike-and-slab models with the computational speed necessary for modern big-data applications.

Standard priors often assume regression coefficients are independent. The paper of Griffin and Brown (2017) argues regarding the validity of this assumption by providing examples where the importance of one variable is naturally linked to another. For example, in models with interactions, it is common to assume that an interaction term should only be included if its corresponding main effects are also present. Another example is the generalized additive model (GAM) which represents the mean of the response as a linear combination of potentially non-linear functions of each variable. In order to deal with these situations, they introduced two continuous one-group priors, namely, Normal-Gamma (NG) Prior: a scale mixture of normals where the variance follows a Gamma distribution, and Normal-Gamma-Gamma (NGG) Prior: also known as a generalized beta mixture, which adds another level to the hierarchy to allow for heavier tails and more flexible shrinkage. Both priors use a parameter (λ) to control behavior near zero, determining how aggressively small coefficients are shrunk. They assumed that the regression coefficients can be arranged into L levels. Coefficients at higher levels (e.g., complex interactions) add more model complexity and are typically shrunk more aggressively than those at lower levels. The proposed priors can be implemented using standard Markov chain Monte Carlo (MCMC) methods, specifically Gibbs sampling, which ensures computational efficiency. Overall, their work provides a way to incorporate prior structural knowledge into high-dimensional regression while maintaining the benefits of continuous Bayesian regularization. Now we move towards another direction in the regression setup when the design

matrix (or the Gram matrix) is block-diagonal. In this situation, providing a scalable algorithm to select important variables is natural to ask about. Papaspiliopoulos and Rossell (2017) introduced a scalable framework for performing Bayesian variable selection and model averaging in high-dimensional linear models where the predictors exhibit a block-diagonal structure. In block-diagonal designs, their proposed approach returns the most probable model of any given size without resorting to numerical integration. When the design matrix is not block-diagonal, they proposed to use spectral clustering to approximate the Gram matrix by a block-diagonal matrix and suggested using an iterative algorithm that capitalizes on the block-diagonal algorithm to explore the model space efficiently. Finally, in order to implement their proposed approaches, they created an R package named **mombf**.

There are also few works based on posterior post-processing; one of them is due to Hahn and Carvalho (2015). They denoted their method as "Decoupling Shrinkage and Selection"(DSS) which is based on minimizing an expected loss function that balances two competing goals: minimizing the mean squared prediction error relative to the full posterior mean and penalizing the number of non-zero coefficients (using an l_0 penalty) to ensure the final model is easy to communicate and interpret. Through real data sets, they illustrated that even if the regressors are strongly correlated, results obtained through DSS are similar to those of median probability model (MPM) and the highest probability model (HPM). Motivated by this work, Ray and Bhattacharya (2018) proposed another posterior post-processing algorithm, namely "Signal Adaptive Variable Selector" (SAVS) approach using the horseshoe prior. Their proposed approach post-processes a point estimate such as the posterior mean to group the variables into signals and nulls. Unlike other approaches, it is free from any tuning parameter. Through different simulations and a genomic dataset with more than 20,000 covariates, they compared the performance of the SAVS approach with other frequentist penalization procedures and Bayesian model selection procedures and found that SAVS is highly competitive across all the settings considered.

We now turn our discussion to the grouped regression model. Model (1.25) is a particular case of that and it will be evident shortly. Potential predictors or regressors are often inherently linked, forming clusters or groups. For instance, gene expression data may have among the predictors groups of genes controlling similar phenotypical traits. Grouping structure is also seen, for example, in multifactor ANOVA and nonparametric additive models. See, for instance, the discussions in Yuan and Lin (2006), Yang and Narisetty (2020), Wang and Leng (2008), and Wei and Huang (2010), in this connection. There are situations, where it becomes important to identify the important groups rather than examining individual variables within groups. For example, Huang et al. (2012) noted that when a continuous factor is represented through a set of basis functions, the individual variables are artificially constructed. Thus, rather than selecting significant individual members, determining which groups are significant overall is the priority. In such situations, it becomes crucial to have rules which can effectively select relevant groups of regressors as a whole. Consequently, variable selection essentially transforms into a problem of group selection. This issue has received significant attention from researchers over the past couple of decades.

Formally, a grouped linear regression model containing G groups of potential regressors or

predictors is defined as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}. \tag{1.27}$$

Here, $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}_g$ is an $n \times m_g$ design matrix and $\boldsymbol{\beta}_g$ is an $m_g \times 1$ vector of unknown regression coefficients for the g^{th} group, where $g \in \{1, \ldots, G\}$ and $\sum_{g=1}^{G} m_g = p$. We further assume that the vector of unobserved residuals $\boldsymbol{\epsilon}$ has a $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ distribution. Our interest is in a sparse asymptotic setting where $G \equiv G_n$ increases at the same rate as the sample size n and the number of *active* groups $G_{\mathcal{A}_n}$ grows at a slower rate than G_n, an active group being one with a true non-zero group regression coefficient vector.

Here the target is to propose a decision rule which can detect truly important groups present in the model. Several approaches used for the ungrouped setup can be extended to the group selection problem. One such example is the group LASSO of Yuan and Lin (2006), which is defined as the minimizer of the following objective function

$$\mathrm{S}(\boldsymbol{\beta}) = \frac{1}{2}\left(\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\right)^{\mathrm{T}} \left(\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\right) + \lambda \sum_{g=1}^{G} \sqrt{m_g} \sqrt{\boldsymbol{\beta}_g{}^{\mathrm{T}} \boldsymbol{\beta}_g} \ .$$

The adaptive group LASSO is intended to be an improvement over the group LASSO by using separate data-based penalties for each group, similar to the adaptive lasso. Wang and Leng (2008) proved asymptotic optimality properties of the adaptive group LASSO estimator and the consistency of the adaptive group LASSO for group selection. For group selection, researchers have also used a multivariate-Laplacian one-group prior distribution to model the unknown group coefficients. This method is called the Bayesian group LASSO, and the corresponding MAP estimator is known as the Bayesian group LASSO estimator (Raman et al. (2009), and Casella et al. (2010)).

Two-group priors can also be used in the group selection problem by using a mixture of degenerate measure at $\mathbf{0}$ and a heavy-tailed absolutely continuous distribution F over $\mathbb{R}^{m_g}$ on the group coefficients $\boldsymbol{\beta}_g$'s independently for $g = 1, \ldots, G$. One such example is the Bayesian group LASSO with spike and slab priors (BGL-SS), introduced by Xu and Ghosh (2015). For the same problem, Yang and Narisetty (2020) modified the BGL-SS by introducing a binary latent variable for each group to indicate the activeness or otherwise of the group and proposing spike and slab priors on group coefficients depending on the latent variables. These works along with the hierarchical formulation, are discussed in detail in Chapter 4. In order to maintain the continuity and overall readability of the thesis, works done on the group selection problem using one-group shrinkage priors have been discussed in Section 1.2.3.

1.2 Our contributions in this thesis

We are now in a position to describe and motivate the specific problems we tried to address in this thesis and our main contributions. These are described in the next few subsections by focusing on each chapter individually.

1.2.1 Motivation and Contributions of Chapter 2

In Chapter 2, our focus is on the famous sparse normal means model (1.1) where the sample size n is assumed to be large. Our interest here is in inference (using one-group priors) on the mean vector $\boldsymbol{\theta}$, specifically, estimation and simultaneous hypothesis testing on whether the i-th coordinate of $\boldsymbol{\theta}$ is zero or not for $i = 1, 2, \cdots, n$. These problems and different approaches in dealing with them that exist in the literature have already been broadly discussed in several subsections of this introduction. We first highlight our main contributions in this chapter. This is followed by a discussion on some key references which motivated us to investigate the specific problems addressed in this chapter.

In this chapter, we made an attempt to provide some answers from decision-theoretic viewpoint to some questions related to estimation and testing. This is done by modelling the mean parameter by a broad class of one-group priors satisfying (1.7) and (1.10). At first, we state our results related to the optimal posterior contraction rate of empirical Bayes and full Bayes procedures. We first establish the near-minimax contraction rate (in Theorem 2.2.1) for empirical Bayes posterior distributions for a class of priors using the plug-in estimate of τ (defined in (2.4)) of van der Pas et al. (2014). We are able to prove that a near-minimax contraction rate (in Theorem 2.2.7) result also holds for the empirical Bayes posterior using the MMLE of τ (defined in (2.9)), when the mean vector is modelled by the three parameter beta normal priors with $a = \frac{1}{2}$. We further show that for the full Bayes procedure, too, such optimality results related to the posterior contraction rate (in Theorem 2.2.11) indeed hold for the three parameter beta normal class of priors with $a = \frac{1}{2}$, under conditions on the prior density of τ similar to those of van der Pas et al. (2017a). In the latter half of this chapter, we focus on the asymptotic decision theoretic performance of simultaneous testing rules using one-group priors in a full Bayes approach, when the data is generated from a two-group framework. Towards this, we consider arbitrary priors on τ, supported on carefully chosen small intervals. Our theoretical calculations reveal that our proposed condition on the prior on τ is sufficient for the Bayes risk of the induced decision rule (defined in (2.24) in Section 2.3) to attain the Optimal Bayes risk up to a multiplicative constant. We further show that for problems where sparsity is pronounced, for "horseshoe-type" priors, the decision rule asymptotically attains the Bayes Oracle exactly. Our results are adaptive and hold true for a wide range of sparsity levels. They also highlight the fact that one has the liberty to use any non-degenerate prior on τ supported on our proposed range for ensuring optimal/near-optimal results using one-group priors. To the best of our knowledge, these are the first optimality results of their kind in the literature in the full Bayes approach in this hypothesis testing problem, which is one of the statistical takeaways of this chapter. We expect that researchers from the applied field will get more confidence to use one-group priors instead of the 'gold-standard' two-groups in several problems related to sparsity due to presence of strong theoretical background of one-group priors. The technical details which were fundamental for establishing these results are discussed in detail in Remarks 2.2.2, 2.2.4, 2.2.6 and 2.2.10, respectively in Chapter 2. We discuss in the next few paragraphs the main existing results which motivated us to undertake this study.

We first focus on the problem of estimation of $\boldsymbol{\theta}$. It may be interesting to emphasize first

that the work of Castillo and Van Der Vaart (2012) established the following fact. When a two-group prior is used to model $\boldsymbol{\theta}$, the tails of the prior for modelling the non-zero means need to be at least as heavy as that of the Laplace distribution for the resulting posterior distribution to contract around the truth at the minimax rate. Martin and Walker (2014) established the interesting result that when the prior on the slab part is modelled by a normal distribution with mean fixed as X_i and variance is large, fixed value, the posterior contracts at the minimax rate. Hence, the prior on the mean parameters incorporates the sparsity assumption present in the model. They also mentioned that the data-dependent choice of the mean of the slab prior does not affect the conjugacy. As a result, generating samples from the posterior distribution becomes relatively simple. See also Martin et al. (2017) in this context. van der Pas et al. (2014) modelled the mean vector by the horseshoe prior. They showed that if $\boldsymbol{\theta}$ is estimated by the corresponding Bayes estimate, then it attains the minimax risk possibly up to a multiplicative constant (with respect to squared ℓ_2 loss) over the class $\ell_0[q_n]$ when the global shrinkage parameter τ is chosen based on the known level of sparsity. They went on to show that with such choice of τ, the corresponding posterior distribution contracts around the truth and the Bayes estimate at the minimax rate. In case the proportion of non-zero means is unknown, it was also proved that when τ is estimated using a plug-in empirical Bayes estimate (introduced in that paper), the resulting empirical Bayes posterior mean asymptotically achieves the minimax risk up to a logarithmic factor, i.e., it achieves a "near-minimax rate" to be explained in Chapter 2. As remarked by these authors, their proofs did not make it clear if a pole near zero like the horeshoe, is necessary for optimal performance or a heavy tailed prior with sufficient mass near zero would also enjoy such optimality. This question was taken up for further study by Ghosh and Chakrabarti (2017), who modelled $\boldsymbol{\theta}$ by a broad class of priors satisfying (1.7) and (1.10). These priors are heavy tailed and assign sufficient prior mass around zero, but do not necessarily have a pole near zero. By choosing τ appropriately based on the proportion of non-zero means, they showed that the Bayes estimates corresponding to this class of priors achieve the minimax risk up to a constant that equals 1 for horseshoe-type priors. The corresponding posterior distribution was shown to have minimax contraction rate around the truth. When the proportion of non-zero means is unknown, they generalized the result of van der Pas et al. (2014) by showing that the empirical Bayes version of the posterior mean using the estimate of τ proposed by van der Pas et al. (2014) attains the minimax risk under the squared error loss, up to some multiplicative constant when $q_n \sim n^{\beta}$, for some $0 < \beta < 1$. One question that these results leave unanswered is whether the empirical Bayes posterior (using the estimate of τ as in van der Pas et al. (2014)) enjoys minimax contraction rate around the truth for the class of priors considered by Ghosh and Chakrabarti (2017). This question is the key motivation behind Theorem 2.2.1 described in Chapter 2.

It may be noted that a more natural way to estimate the global shrinkage parameter τ is by using the maximum likelihood estimate of it obtained from the marginal distribution of $\mathbf{X}$ given τ. Such an estimate is popularly known as the MMLE of τ. van der Pas et al. (2017a) studied the posterior contraction properties of the empirical Bayesian posterior distribution of the Horseshoe where τ is estimated by the MMLE. Since in the definition of MMLE, the maximization is over $\tau \in [\frac{1}{n}, 1]$, it ensures that, the MMLE estimator does not collapse to zero.

While studying the optimality property based on the horseshoe prior, they first established that the MMLE corresponding to the horseshoe prior lies in a small interval whose upper limit is of the order of the optimal choice of τ when treated as a tuning parameter. See Theorem 3.1 of van der Pas et al. (2017a) in this context. Later, this crucial result helped them to prove that the empirical Bayes posterior distribution using the MMLE of τ contracts around the truth at the near-minimax rate. They remarked that the plug-in estimator introduced in van der Pas et al. (2014) also satisfies Theorem 3.1 of their work, implying that the empirical Bayes estimate using the plug-in estimator of τ, too, contracts around the truth at the near-minimax rate. Hence, it becomes quite natural to question whether a similar phenomenon holds for empirical Bayesian treatment using MMLE when the mean parameter is modelled by a more general class of one group priors containing horseshoe as a special case. We address this question in Theorem 2.2.7 in Chapter 2.

An alternative approach for modeling the global parameter is the full Bayes approach where τ is modeled by a non-degenerate prior distribution. van der Pas et al. (2017a) also considered the full Bayes procedure using the horseshoe prior and proposed two conditions (on the prior on τ) under which the full Bayes posterior contracts around the truth at a "near-minimax" rate. A main step towards obtaining the result was to prove that the region in which the posterior distribution of τ puts most of its mass is the same region where the empirical Bayes versions of it resides with high probability. This might drive one's interest to investigate whether similar results hold for a more general class of priors in a full Bayes approach. Theorem 2.2.11 of Chapter 2 provides a positive answer regarding this.

We now focus on the problem of simultaneous testing on the coordinates of $\boldsymbol{\theta}$ as mentioned before. Assuming that the true means are generated from a spike and slab prior, Bogdan et al. (2011) proved that the Bayes risk of the optimal rule (with respect to an additive misclassification loss penalizing both types of errors equally), namely the Bayes Oracle for this problem, is asymptotically of the form $np\left[2\Phi(\sqrt{C})-1\right](1+o(1))$ where n denotes the number of hypotheses, p is the proportion of non-zero means and C is a constant defined in Assumption 2 of Chapter 2. There have been several attempts in the literature to investigate asymptotic optimality of multiple testing rules based on one group global-local priors when the true means are generated by a spike-and-slab prior. In this direction, the results in Datta and Ghosh (2013), Ghosh et al. (2016) and Ghosh and Chakrabarti (2017) mentioned before used the global shrinkage parameter τ either as a tuning parameter or it was estimated in an empirical Bayesian way and it was established that the Bayes risk of the Oracle is matched at least up to a constant by multiple testing rules based on several global-local priors. It is worth mentioning that Ghosh and Chakrabarti (2017) established that this constant can be shown to be exactly 1 for a "horseshoe-type" priors to be defined in Chapter 2. This is an important subclass of their proposed class of priors containing horseshoe, strawderman-berger, normal-exponential-gamma, and standard double Pareto priors as special cases. This means that appropriately chosen one-group priors can literally mimic their two-group counterparts in sparse situation and may be used as a viable alternative. Ghosh et al. (2016) reported a simulation study using a full Bayes version of decision rule (1.15) based on several choices of one-group priors when a non-degenerate prior was placed on τ. The results were promising, in that, for small values of p, the full Bayes rule practically

matches the Bayes Oracle in terms of risk. This raises the natural question of whether a full Bayes approach for multiple testing using suitable one-group priors can be theoretically shown to match the Bayes Oracle in terms of Bayes risk. We provide a positive answer to this question in Theorem 2.3.4 of Chapter 2.

1.2.2 Motivation and Contributions of Chapter 3

In Chapter 3, we explore the multiple hypothesis testing problem in the sparse Gaussian sequence model. We focus on the same broad class of priors previously mentioned and consider the thresholding multiple testing rules originally proposed by Carvalho et al. (2010) and later examined by Ghosh et al. (2016), Ghosh and Chakrabarti (2017), and Paul and Chakrabarti (2025). Our aim is to investigate whether these multiple testing rules, based on this diverse array of shrinkage priors, achieve the minimax risk stated in equation (1.20) under appropriate conditions. To this end, we adopt the asymptotic framework outlined by Abraham et al. (2024). Our theoretical results indicate that one-group shrinkage priors of the form described in (1.7), which satisfy (1.10), cannot achieve the desired minimax risk—regardless of the loss functions mentioned—if $a > 0.5$. Therefore, we restrict our focus to the *horseshoe-type* priors, with $a = 0.5$, which include a large collection of important priors described in Subsection 1.2.1. When the number of non-zero signals is known, we theoretically demonstrate that the decision rule based on horseshoe-type priors can asymptotically attain the minimax risk with a suitable choice of the global shrinkage parameter, dependent on the proportion of non-zero means, for both the FDP + FNP and misclassification loss functions examined by Abraham et al. (2024). In cases where the level of sparsity is unknown, we consider empirical Bayes and full Bayes decision rules based on the horseshoe-type priors and establish that these adaptive decision rules achieve the corresponding minimax risks asymptotically, as the dimension grows to infinity. Our findings provide important guidelines for selecting priors, showing that beyond a certain subclass of our chosen class—specifically, the horseshoe-type priors—the minimax risk cannot be achieved for either of the two loss functions discussed. To the best of our knowledge, these findings represent the first results in the literature demonstrating that global-local priors can guarantee achieving exact asymptotic minimax risk in the context of simultaneous testing problems. In the next paragraph, we discuss the main existing results which motivated us to undertake this study.

As discussed in Subsection 1.1.4, Abraham et al. (2024) obtained the minimax risk for the sparse normal means model for both the misclassification loss and the loss defined as the sum of FDP and FNP, assuming the beta-min condition on the non-zero means. Here FDP and FNP denote False Discovery proportions and False Negative proportions, respectively. Later, they established that the BH procedure of Benjamini and Hochberg (1995) and the ℓ- value approach based on two-group priors achieve the minimax risk, even if the level of sparsity is unknown. It is mentionworthy that, prior to their work, Salomond (2017) investigated the asymptotic behavior of Bayesian multiple-testing procedures that utilize thresholding rules using a broad class of Gaussian scale-mixture priors. Drawing inspiration from the findings of van der Pas et al. (2016), they modeled each mean parameter using a broad category of one-group shrink-

age priors, specifically scale mixtures of normals. Under conditions on the size of the non-zero means, they derived an upper bound for the minimax risk, corresponding to the sum of the false discovery rate (FDR) and the false non-discovery rate (FNR). Their results indicated that if the non-zero signals are sufficiently large—referred to as the "separation condition"—the minimax risk approaches zero asymptotically, which suggests consistent detection capability. However, the detection boundary for non-zero signals identified by Salomond (2017) is larger than the oracle threshold proposed by Abraham et al. (2024). This raises the question of whether any one-group prior can asymptotically achieve the minimax risk under the beta-min condition put forth by Abraham et al. (2024) for non-zero signals, a question that remains unresolved. Additionally, Salomond (2017) did not examine the scenario involving misclassification loss. When the level of sparsity is unknown, one potential solution is to use the full Bayesian approach, which their paper did not address. We provide a modest attempt to address these questions in Theorems 3.3.2, 3.3.4, 3.3.6, 3.3.7, 3.3.9 and 3.3.11 in Chapter 3.

1.2.3 Motivation and Contributions of Chapter 4

In Chapter 4, our interest is in group selection and estimation in normal linear regression models with grouped predictors. Grouped regression has been introduced and the importance of grouped variable selection has been highlighted already in Subsection 1.1.7 along with a broad review of the relevant literature. Our focus in this chapter is on the application of one-group priors in this context, a path somewhat less trodden in the literature. First, we highlight our main contributions in this chapter. Next, we provide a discussion on some key references which motivated us to investigate the specific problems addressed in this chapter.

We propose in this chapter a thresholding rule in the group selection problem and also an estimator of the active groups based on a very broad class of one-group shrinkage priors described as:

$$\boldsymbol{\beta}_g \mid \lambda_g^2, \sigma^2, \tau^2 \sim \mathcal{N}_{m_g}(\mathbf{0}, \lambda_g^2\sigma^2\tau^2(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1}), \quad \pi(\lambda_g^2) \propto (\lambda_g^2)^{-a-1}L(\lambda_g^2), \tag{1.28}$$

where a is a positive real number and $L : (0,\infty) \to (0,\infty)$ is a measurable non-constant slowly varying function in Karamata's sense (see Bingham et al. (1987)), that is, $\frac{L(\alpha x)}{L(x)} \to 1$ as $x \to \infty$, for any $\alpha > 0$. Tang et al. (2018) also considered a prior similar to $\pi(\cdot)$ as defined in (1.28) in their paper, although in an ungrouped regression, and referred to such priors as having "polynomial tails". We assume the error variance σ^2 to be known throughout the theoretical analysis. For our simulations, we assume the usual Jeffreys prior for σ^2 given by, $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$. On the other hand, depending on whether the level of sparsity is known or not, τ is either treated as a tuning parameter or it is treated in empirical Bayesian or full Bayesian ways.

Our proposed rule, also referred to as the half-thresholding (HT) rule (defined formally in (4.8)), declares a group to be active if the ratio of the ℓ_2 norm of the posterior mean of the regression coefficients to that of the ordinary least squares estimate of the corresponding coefficient vector exceeds half. Consequently, our proposed half-thresholding (HT) reduces to that of Tang et al. (2018) when the group size is unity.

Our contributions in this chapter are as follows. Firstly, we propose a new general class of one-group global-local shrinkage priors in the case of grouped regression. This prior explicitly takes into account the grouping structure by bringing the additional $(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1}$ term. Secondly, we propose a half-thresholding rule that can be easily implemented regardless of whether the underlying sparsity level is known or not. We first show that when the proportion of active groups is known, the global shrinkage parameter τ can be appropriately chosen based on this proportion so that the resulting decision rule enjoys the oracle property mentioned before. When the proportion of active groups is unknown, we propose to use an empirical Bayes estimator of the global shrinkage component. This estimator generalizes the empirical Bayes estimator proposed by van der Pas et al. (2014) in the context of normal means model. We show that using this estimator, the resulting data-adaptive half-thresholding rule enjoys the oracle optimality properties under very mild conditions. This is the first result of this kind in the literature. As an immediate consequence, our result settles the optimality of an empirical Bayes version of the HT rule of Tang et al. (2018) left open in that paper. Last but not the least, our theoretical results hold when the number of true active groups grows substantially with n. It is important to note that we need to develop novel and rigorous analytical techniques to theoretically establish these properties. Finally, in our simulation studies, we use both empirical Bayes and full Bayes versions of our proposed decision rule and demonstrate that our methods compare favorably to well-known group selection methods for a wide range of settings, to be made precise in Chapter 4. The performance of our method, when applied to real datasets, is also quite satisfactory and noteworthy. We now describe the key references and some questions left open in these works, which motivated us to conduct the study of problems discussed in this chapter.

Grouped variable selection and estimation are generalizations of the corresponding concepts in the traditional "ungrouped" regression problem. We first discuss Tang et al. (2018), an article focusing on the setting of traditional ungrouped regression. Under the assumption of an orthogonal design, Theorem 1 of Tang et al. (2018) shows that for a broad class of one-group shrinkage priors used for modelling λ_i^2 in (1.10), their proposed "half-thresholding" (HT) variable selection rule and the resulting estimators enjoy the oracle property (see Fan and Li (2001), Zou (2006), Zou and Zhang (2009) in this context) in the sense that both variable selection consistency and optimal estimation rate are achieved. The last two concepts mentioned will be described formally in Chapter 4. Tang et al. (2018) proved these results under the assumptions that (a) the number of active regressors is fixed as sample size increases and (b) the level of sparsity is known. This optimality hinges upon the choice of the global shrinkage parameter in terms of the level of sparsity. Both (a) and (b) are hard to verify in many examples and therefore it is worth investigating if similar results are obtained when the number of active regressors does not necessarily remain bounded and the level of sparsity is unknown. The latter point will be addressed to an extent if an empirical Bayesian version of the procedure of Tang et al. (2018) can also be shown to have optimality properties, a problem left open by these authors. We have explored these issues within the broader context of grouped regression and investigated that the half-thresholding (HT) technique can be extended in this setting. It is also of interest to ask if the oracle property holds without assumptions (a) and (b) in the grouped problems also. We discussed these in Theorems 4.3.8 and 4.3.10 in Chapter 4. We now recall some crucial works in

the field of grouped regression to underscore the relevance of the problems investigated in this chapter, given the existing state of knowledge.

Towards that, we first discuss Xu et al. (2016). These authors proposed variants of the horseshoe prior to model grouped regression coefficient vectors in regression models. They used two sets of local shrinkage parameters to simultaneously control the shrinkage at the group level and within the groups. They used the resulting estimators in a problem of prediction and demonstrated superiority of their proposed priors compared to the usual horseshoe prior. But this paper did not provide any rule for selection of active groups or any rule for bi-level selection, which, after selection of active groups digs further to select important regressors within the selected group. Very recently, Boss et al. (2023) proposed another one-group global-local prior to model the individual regression coefficients in a grouped regression context, which is structurally similar to that in Xu et al. (2016), in that there is a single global shrinkage parameter and two sets of local shrinkage parameters acting on the group level and on the individual regression coefficients within a group. The resulting full posterior distribution was shown to be consistent. Results corresponding to posterior concentration properties of shrinkage parameters demonstrated the overriding effect of the local shrinkage parameter within a group to ensure that solitary large signals within a group are not shrunk, even if the group level shrinkage warrants heavy shrinkage overall for a group with very few signals. Boss et al. (2023) also did not provide any rule for selection of variables/groups of variables. Note that in both the works mentioned here, authors were interested in modelling not only groups of covariates, but also the individual variables present in the group. On the contrary, we are interested in situations where groups as a whole are important. There are situations, already discussed briefly before, where it is important to identify the important groups rather than examining individual variables within groups. For example, Huang et al. (2012) noted that when a continuous factor is represented through a set of basis functions, the individual variables are artificially constructed. Thus, rather than selecting significant individual members, discovering groups that are significant overall is the priority. Group selection also becomes important in the seemingly unrelated regressions (SUR) model proposed by Zellner (1962) or the multitask learning model within machine learning put forth by Caruana (1997), and further discussed by Argyriou et al. (2008). Hence it is of keen interest to propose a decision rule to select relevant groups based on one-group priors and come up with corresponding estimators which enjoy the Oracle property in the grouped regression framework. To the best of our knowledge, the only Bayesian work in the grouped regression setting where the above mentioned oracle property has been studied is Xu and Ghosh (2015), though under the assumption that the number of groups is fixed. Yang and Narisetty (2020) studied group selection consistency results but not the asymptotic optimality of their proposed estimator. Nevertheless, it is important to note that both of these works are under a two-group framework. In a nutshell, in the context of linear regression with grouped predictors, many questions related to the optimality of decision rule based on one-group global-local priors have not been addressed till now. We hope that Theorems 4.3.1, 4.3.4, 4.3.7, 4.3.10 and 4.3.12 in Chapter 4 enlighten us to some extent in this direction. To the best of our knowledge, this is the first study from decision theoretic viewpoint of one-group priors in group selection problems.

1.2.4 Motivation and Contributions of Chapter 5

In Chapters 2 through 4 of this thesis, our focus is on the normal means model and the normal linear regression model in the high-dimensional situations. However, the assumption of Gaussianity is not reasonable for all types of data and one may need to model data through other appropriate probability distributions. Our work in Chapter 5 is focused on an interesting situation of this kind as described below.

Suppose one is interested in modeling the counts of an event, for example, a terrorist attack, happening in several areas of a country or several countries in a continent. First thing to observe is that the assumption of Gaussianity is completely inappropriate in such cases. It may also be noted that the rates of occurrences are small in most of the areas while some areas may be much more prone to such attacks. Our work in this chapter is motivated by Datta and Dunson (2016) who studied such a problem. The focus of Datta and Dunson (2016) is inference on "quasi-sparse" count data having a lot of counts near zero but not necessarily exactly zero with some large and moderate counts. They assumed that the i-th observation follows a Poisson distribution with mean θ_i, i.e., if the observation vector is $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$, then $Y_i \overset{ind}{\sim} Poi(\theta_i)$, where $Poi(\theta_i)$ denotes a Poisson distribution with mean θ_i.

The most challenging part in this situation is to be able to simultaneously model the abundance of small counts and also the moderate and large counts present in the data. The standard priors for "zero-inflation" would lead to a mixture of a degenerate distribution at 0 with a Poisson or negative binomial distribution and for the marginal distribution of Y_i. As commented by Datta and Dunson (2016), may be computationally unstable. Datta and Dunson (2016) instead proposed a model to capture data of such kind through a continuous prior on θ_i. Specifically, they considered the following hierarchical prior:

$$\theta_i|\lambda_i, \tau \overset{ind}{\sim} Ga(\alpha, \lambda_i^2\tau^2), \lambda_i^2 \overset{ind}{\sim} \pi_1(\lambda_i^2), \tau^2 \sim \pi_2(\tau^2), \tag{1.29}$$

where $Ga(\alpha_1, \alpha_2)$ denotes a Gamma distribution with shape parameter α_1 and scale parameter α_2 with p.d.f. as $f(x|\alpha_1, \alpha_2) = \frac{1}{\Gamma(\alpha_1)\alpha_2^{\alpha_1}} x^{\alpha_1 - 1} e^{-x/\alpha_2}, x > 0, \alpha_1 > 0, \alpha_2 > 0$ and $\pi_1(\cdot), \pi_2(\cdot)$ are densities for λ_i^2 and τ^2, respectively. For the theoretical development, they either fixed τ or let $\tau \to 0$ as $n \to \infty$. This is an example of application of a global-local shrinkage prior in the context of count data. Now integrating out θ_i and defining $\kappa_i = 1/(1 + \lambda_i^2\tau^2)$, we have,

$$f(Y_i|\kappa_i) \propto (1 - \kappa_i)^{Y_i} \kappa_i^{\alpha}.$$

This shows that marginally each Y_i follows a negative binomial distribution with parameters α and κ_i. We note that $\theta_i|Y_i, \kappa_i \sim Ga(Y_i + \alpha, 1 - \kappa_i)$, and $E(\theta_i|Y_i, \kappa_i) = (1 - \kappa_i)(Y_i + \alpha)$. Then $(1 - \kappa_i)$ is the factor by which the posterior mean shrinks $(Y_i + \alpha)$ towards zero. Datta and Dunson (2016) proposed to apply Gauss Hypergeometric (GH) prior on κ_i given τ given as

$$\pi(\kappa_i|a_1, a_2, \tau, \gamma) = C_2 \kappa_i^{a_1 - 1}(1 - \kappa_i)^{a_2 - 1}[1 - (1 - \tau^2)\kappa_i]^{-\gamma}, 0 < \kappa_i < 1, a_1, a_2, \tau, \gamma > 0\,, \tag{1.30}$$

where $C_2^{-1} = Beta(a_1, a_2)_2F_1(\gamma, a_1, a_1+a_2, 1-\tau^2)$ is the normalizing constant, where $Beta(a_1, a_2) = \int_0^1 t^{a_1-1}(1-t)^{a_2-1}dt$ the beta function and ${}_2F_1$ the Gauss hypergeometric function. For their theoretical analysis they further fixed $a_1 = a_2 = \frac{1}{2}$. The choice of the GH prior is motivated by the additional flexibility in the prior mass (of κ_i) brought in the regions near 0 and 1 leading to flexible shrinkage; see Datta and Dunson (2016) for more detail. Part of this can be explained by the flexibility in choosing γ. They established that for their proposed hierarchical form (1.29) with κ_i modelled as (1.30), the marginal prior density of θ_i given τ is unbounded at origin which justifies the abundance of zero and near zero counts, i.e., θ_i has a *pole at zero*. This ensures high probability at zero of the marginal distribution of Y_i. They also showed that, for their proposed class of priors, the posterior distribution of κ_i will concentrate near 0 or 1 depending on whether the observation is large or small. They then remarked that one natural choice of a two-group model to incorporate quasi-sparsity is:

$$\theta_i \overset{iid}{\sim} (1-p)Ga(\alpha, \beta) + pGa(\alpha, \beta+\delta), i = 1, 2, \cdots, n. \tag{1.31}$$

By keeping α and β small one can ensure high concentration of a $Ga(\alpha, \beta)$ distribution near zero while letting $\delta >> \beta$ ensures that $Ga(\alpha, \beta+\delta)$ is flatter than the $Ga(\alpha, \beta)$ distribution. For such choices of α, β and δ, $Ga(\alpha, \beta)$ is a natural prior for modelling the small θ_i's while $Ga(\alpha, \beta+\delta)$ is a natural prior for modelling the large or moderate θ_i's. Note that, the marginal distribution of Y_i is then obtained as a mixture of two negative binomial distributions with the following form:

$$Y_i \overset{iid}{\sim} (1-p)NB(\alpha, \frac{1}{\beta+1}) + pNB(\alpha, \frac{1}{\beta+\delta+1}), i = 1, 2, \cdots, n, . \tag{1.32}$$

Here $NB(\alpha, p)$ refers to a negative binomial distribution denoting the number of failures before the α-th success of independent trials and p being the probability of success in each trial. It may be observed that by introducing latent variables $\nu_i, i = 1, 2, \cdots, n$, such that $\nu_i = 0$ indicates $H_{0i} : \theta_i \sim Ga(\alpha, \beta)$ and $\nu_i = 1$ indicates $H_{1i} : \theta_i \sim Ga(\alpha, \beta+\delta)$ with $P(\nu_i = 1) = p$ and $P(\nu_i = 0) = 1-p$, one gets (1.31) as the marginal prior on θ_i. Datta and Dunson (2016) wanted to investigate the performance of a multiple testing rule that rejects H_{0i} if $1-E(\kappa_i|Y_i, \alpha, \tau, \gamma) > \zeta$, where ζ is a suitably chosen threshold, if the true θ_i's are generated following (1.31). This rule was proposed by comparing the expressions of the posterior means under the two-group model and their one-group prior. In a simulation study, they showed that the performance of GH prior with $\gamma \neq 1, a_1 = a_2 = \frac{1}{2}$ is better than that of the horseshoe prior (with $\gamma = 1, a_1 = a_2 = \frac{1}{2}$) in terms of the misclassification probability. They also obtained an upper bound on the type I error probability of the resulting rule. However, they did not obtain any expression for the type II error. They also did not study the optimal Bayes risk in the two-group formulation (1.31) based on any decision theoretic framework and for that matter, this study has not been undertaken in the literature till now by any other authors, in the context of count data, as far as we know.

The above discussion brings forth many interesting questions which warrant further investigation. Our work is a modest attempt to address some of these. First and foremost, it would be of interest to find expressions for the optimal Bayes risk with respect to some natural loss func-

tions for multiple hypothesis testing if θ_i's are indeed coming from a two-group gamma mixture as in (1.31). Secondly, it would be of keen interest to investigate if other choices of one-group shrinkage priors can be employed for simultaneous testing on quasi-sparse count data and if they can match (or better) the good performance/properties of inference using GH prior, as shown in this context in Datta and Dunson (2016). Most importantly, it will be very interesting to settle, if one-group priors can indeed be proxies of their two-group counterparts when evaluated in a decision theoretic sense and if they can attain the corresponding optimal Bayes risk, even in the context of count data of this kind.

Building an asymptotic framework inspired by the work of Bogdan et al. (2011) for the normal means model, in Chapter 5 we obtain an expression for the optimal Bayes risk (with respect to additive $0-1$ loss) in the context of quasi-sparse count data. The conditions on the parameters in this framework are somewhat similar in spirit to those introduced by Bogdan et al. (2011). However, since both the level of sparsity p and the scale parameter β under the null go to zero, an extra assumption on the relative growth of $\frac{\beta}{p}$ becomes necessary. These conditions are discussed in detail in Section 5.2. Next, inspired by the results of Ghosh et al. (2016) and Datta and Dunson (2016), in this work, we have considered a decision rule for performing simultaneous testing on the θ_i's based on a broad class of global-local shrinkage priors of the form (1.29) with the local shrinkage parameter λ_i^2 modelled as some proper subclass of (1.10) to be made precise in Chapter 5. We have used τ either as a tuning parameter or have treated it in an empirical Bayesian way. Our chosen class of priors contains, among others, big subclasses of the three parameter beta normal priors of Armagan et al. (2011), the generalized double Pareto priors due to Armagan et al. (2013a). These priors are discussed in more details in Chapter 5. When the level of sparsity p is known, we are able to prove that the Bayes risk of our decision rule based on our chosen class of priors attains the optimal Bayes risk, up to some multiplicative constant. On the other hand, when the underlying level of sparsity is unknown, we provide an empirical Bayes estimate of τ motivated by the work of Yano et al. (2021). In this context, too, we have been able to establish that our proposed empirical Bayes procedure based on our class of global-local shrinkage priors attains the optimal Bayes risk, up to some multiplicative constant. Part of these results is based on a series of lemmas about several quantities related to the posterior distribution of the shrinkage coefficient. As far as we know, this is the first formal study of decision theoretic optimality of one-group priors in the context of sparse count data. Nevertheless, we should mention that since GH prior is not included in our chosen class of priors, our results do not provide any theoretical guarantee that the decision rule based on GH prior also attains the optimal Bayes risk, up to some multiplicative constant. However, in our simulation studies, we have observed that both the performance of our prior and that of GH prior are quite similar to that of the two-group prior in terms of misclassification probability for a wide range of sparsity levels, specifically for small values of p. Our method has also been applied for the analysis of a real data set and has returned pretty satisfactory results. We hope that the simulation results, followed by real data analysis along with theoretical guarantees, will help the researchers to use one-group priors in problems related to quasi-sparse count data, which is a key statistical takeaway of our work considered in this chapter.

Chapter 2

Posterior Contraction Rate and Asymptotic Bayes Optimality for One Group Global-Local Shrinkage Priors in Sparse Normal Means Problem

2.1 Introduction

Consider an observation vector $\mathbf{X} \in \mathbb{R}^n$, $\mathbf{X} = (X_1, \cdots, X_n)$ from the normal means model (introduced in Chapter 1) where each observation X_i can be written as

$$X_i = \theta_i + \epsilon_i, i = 1, 2, \cdots, n, \tag{2.1}$$

where $(\theta_1, \theta_2, \cdots, \theta_n)$ are the unknown means and ϵ_i's are independent $\mathcal{N}(0, 1)$ random variables. In this chapter, our interest is in inference on $\boldsymbol{\theta}$ from the above model using one-group global-local shrinkage priors when $\boldsymbol{\theta}$ is sparse in appropriate senses.

We have already discussed in Chapter 1 modelling $\boldsymbol{\theta}$ by spike-and-slab priors, widely considered as the gold standard in sparse Bayesian modelling and also the motivation behind introduction of one-group continuous shrinkage priors for modelling sparse parameter vectors. It may be recalled that one-group shrinkage priors may be described as follows:

$$\theta_i | \lambda_i, \tau \overset{ind}{\sim} \mathcal{N}(0, \lambda_i^2 \tau^2), \lambda_i^2 \overset{ind}{\sim} \pi_1(\lambda_i^2), \tau \sim \pi_2(\tau), \tag{2.2}$$

where λ_i, $i = 1, \ldots, n$ are the *local shrinkage* parameters and τ is the *global shrinkage* parameter. Our focus in this chapter is the investigation of some questions on asymptotic optimality of inference on $\boldsymbol{\theta}$ by modelling it using a broad class of one-group priors to be specified shortly. As

already discussed in some detail in Chapter 1, similar questions have been researched in some depth in the literature when the global shrinkage parameter τ is used as a tuning parameter to be set using the knowledge of the level of sparsity if it is known. This is technically same as starting with $\pi_2(\cdot)$ being degenerate at point τ to be tuned appropriately using the information on sparsity. When the level of sparsity is unknown, there have been limited works in the literature by treating τ in empirical Bayesian or some full Bayesian ways. But this still leaves some interesting questions which are hitherto unanswered in the literature in this domain. This chapter of the thesis is a modest attempt to answer some of these questions and they have been already described broadly in Subsection 1.2.1. We have come up with totally new results for answering some of our questions, while the rest can be considered as generalizations/extensions of some existing works requiring novel technical arguments. In the next paragraph, we describe our contribution in the order they appear in this chapter.

To address the situation when the proportion of non-zero means is unknown, van der Pas et al. (2014) proposed a simple estimator $\widehat{\tau}$ of τ (defined in equation (2.4)) and showed that the horseshoe estimator with their proposed estimator of τ attains a near-minimax rate for squared ℓ_2 loss. Ghosh and Chakrabarti (2017) extended this result for the class of priors satisfying (2.2) where the local shrinkage parameter is modeled as

$$\pi_1(\lambda_i^2) = K(\lambda_i^2)^{-a-1}L(\lambda_i^2), \tag{2.3}$$

where K and a are same as before and $L : (0,\infty) \to (0,\infty)$ is measurable non-constant slowly varying function (defined in Section 2.2) satisfying Assumption 1. In a follow-up paper, van der Pas et al. (2017a) proved that the empirical Bayes (using the simple estimator of τ and MMLE of τ defined in (2.9)) and full Bayes posterior distributions (using a prior on τ with their proposed conditions) corresponding to the horseshoe contract around the truth at a near minimax rate. The natural question on whether this optimality holds for the broader class of priors satisfying (2.2) and (2.3) is still not addressed in the literature. We are able to establish (through Theorem 2.2.1) that a near-minimax contraction rate holds for the empirical Bayes posterior corresponding to this general class of priors. In Theorem 2.2.7, we have proved that a near-minimax contraction rate result also holds for the empirical Bayes posterior using the MMLE of τ for the three parameter beta normal priors. We have established by proving Theorem 2.2.11 that such optimality indeed holds for the three parameter beta normal class of priors, under conditions on the prior density of τ similar to those of van der Pas et al. (2017a). The technical innovations required to come up with these results are summarized in several remarks following statements of the main results. Next we consider the problem of testing simultaneously whether each coordinate of $\boldsymbol{\theta}$ is zero or not when the θ_i's are actually generated from a two-group spike-and-slab distribution. based on the observations coming from (2.1) in the sparse regime. Motivated by the simulation study of Ghosh et al. (2016) already described in Subsection 1.2.1, we consider a simultaneous multiple testing rule based on the class of one-group priors satisfying (2.2) and (2.3), where the prior on τ is supported on a carefully chosen small interval. The details about the support on the prior on τ considered by us along with the intuition for considering such support, are given in (C4) and in the remark following (C4) of Section 2.3. Our theoretical

calculations establish the crucial fact that condition (C4) is sufficient for the Bayes risk of the induced decision rule (defined in (2.24) in Section 2.3) to attain the Optimal Bayes risk up to a multiplicative constant, for a wide range of sparsity levels. These results substantially reinforce the intuition that suitably chosen one-group priors can be meaningful alternatives to two-group priors for sparse Bayesian modelling.

The rest of the chapter is organized as follows. Section 2.2 discusses the main results on (near) optimal posterior concentration rates. In Subsection 2.2.1, we discuss the corresponding results using the empirical Bayes estimate of τ as proposed by van der Pas et al. (2014). Subsection 2.2.2 also contains the results using the empirical Bayes estimate of τ, namely MMLE due to van der Pas et al. (2017a). Results of the full Bayes approach are presented in Subsection 2.2.3. In Section 2.3, optimality results in the problem of simultaneous hypothesis testing are discussed. Section 2.4 contains a discussion and proposes possible extensions of our work. Proofs of all the results can be found in Section 2.5.

This chapter is based on Paul and Chakrabarti (2025).

2.2 Contraction rate of posterior distribution using global-local shrinkage priors

In this section, we discuss our results on the posterior contraction rate for certain classes of global-local shrinkage priors we consider, when the level of sparsity is assumed unknown. The results of the empirical Bayes and the full Bayes approaches are presented in three separate subsections.

We start with some preliminary definitions and assumptions will be used throughout. Recall that a function $L(\cdot) : (0, \infty) \to (0, \infty)$ is said to be slowly varying, if for any $\alpha > 0$, $\frac{L(\alpha x)}{L(x)} \to 1$ as $x \to \infty$. For the theoretical development of the Chapter, we consider slowly varying functions that satisfy Assumption 1 below.

Assumption 1.

(A1) There exists some $c_0(> 0)$ such that $L(t) \geq c_0 \ \forall t \geq t_0$, for some $t_0 > 0$, which depends on both L and c_0.

(A2) There exists some $M \in (0, \infty)$ such that $\sup_{t\in(0,\infty)} L(t) \leq M$.

These assumptions are similar to those of Ghosh and Chakrabarti (2017).

The minimax rate is considered to be a useful benchmark for the rate of contraction of posterior distributions. According to Theorem 2.5 of Ghosal et al. (2000), posterior distributions cannot contract around the truth faster than the minimax risk. Thus it is of considerable interest to investigate if the posterior distribution generated by a Bayesian procedure contracts around the truth at the minimax rate or not. When $\boldsymbol{\theta}_0 \in l_0[q_n]$, Donoho et al. (1992) proved that the minimax risk under the usual squared L_2 loss $||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 = \sum_{i=1}^{n}(\tilde{\theta}_i - \theta_i)^2$ is asymptotically of the form

$$\inf_{\tilde{\theta}} \sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} ||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 = 2q_n \log\left(\frac{n}{q_n}\right)(1 + o(1)),$$

when $q_n \to \infty$ as $n \to \infty$ and $q_n = o(n)$. Many results on the posterior concentration rate available in the literature are in terms of $q_n \log n$. See, e.g., van der Pas et al. (2017a) and some references therein. This is referred to as the "near-minimax" rate in this literature. It may be however noted that the near-minimax rate $q_n \log n$ is equivalent to the minimax rate if $q_n \sim n^\beta$ for some $0 < \beta < 1$. In both the empirical Bayes and full Bayes approaches, our results in the next subsections show that under the broad classes of priors, the posterior distributions of $\boldsymbol{\theta}$ contract around the truth at a near minimax rate, which substantially generalize the results of van der Pas et al. (2014) and van der Pas et al. (2017a) proved for the horseshoe.

2.2.1 Empirical Bayes Procedure-plug-in Estimator

When the proportion of non-zero means is unknown, van der Pas et al. (2014) proposed using an empirical Bayes estimate of τ. The estimator is defined as

$$\widehat{\tau} = \max\left\{\frac{1}{n}, \frac{1}{c_2 n}\sum_{i=1}^{n} 1\{|X_i| > \sqrt{c_1 \log n}\}\right\}, \tag{2.4}$$

where $c_1 \geq 2$ and $c_2 \geq 1$ are constants. Observe that $\widehat{\tau}$ always lies between $\frac{1}{n}$ and 1. Let $T_{\widehat{\tau}}(\mathbf{X})$ be the Bayes estimate of $\boldsymbol{\theta}$ given τ, namely $T_\tau(\mathbf{X}) = E(\boldsymbol{\theta}|\mathbf{X}, \tau)$ evaluated at $\tau = \widehat{\tau}$. van der Pas et al. (2014) showed that the empirical Bayes estimate $T_{\widehat{\tau}}(\mathbf{X})$ corresponding to the horseshoe prior (by taking $a = 0.5$ and $L(t) = t/(t+1)$ in (2.3)) asymptotically attains the near minimax L_2 risk. This fact was generalized by Ghosh and Chakrabarti (2017) who established the same asymptotic rate for the class of priors (2.2) satisfying (2.3) when $a \geq 0.5$ provided $q_n \propto n^\beta, 0 < \beta < 1$. van der Pas et al. (2017a) also proved that the empirical Bayes posterior (i.e., the posterior distribution of the mean vector given τ evaluated at $\tau = \widehat{\tau}$) for the horseshoe contracts around the truth at a near minimax rate. Our first result, namely Theorem 2.2.1 presented below, is a generalization of this result. The proof of this result is given in Section 2.5.

Theorem 2.2.1. *Suppose* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\theta}_0, \mathbf{I}_n)$ *where* $\boldsymbol{\theta}_0 \in l_0[q_n]$ *with* $q_n \propto n^\beta$ *where* $0 < \beta < 1$. *Consider the class of priors* (2.2) *satisfying* (2.3) *with* $a \geq \frac{1}{2}$ *where* $L(\cdot)$ *satisfies Assumption 1. Then the empirical Bayes posterior contracts around the true parameter* $\boldsymbol{\theta}_0$ *at the near minimax rate, namely,*

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} \Pi_{\widehat{\tau}}\Big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}\Big) \to 0\,,$$

for any $M_n \to \infty$.

Remark 2.2.2. *For proving Theorem 2.2.1, first we need to show that, the total posterior variance using the empirical Bayes estimator of* τ *corresponding to this class of priors satisfies, as* $n \to \infty$,

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} \sum_{i=1}^{n} Var(\theta_i | X_i, \widehat{\tau}) \lesssim q_n \log n, \tag{2.5}$$

where $Var(\theta_i|X_i, \widehat{\tau})$ *denotes* $Var(\theta_i|X_i, \tau)$ *evaluated at* $\tau = \widehat{\tau}$. *In order to obtain* (2.5), *we divide our calculations for the non-zero and zero means, namely in Step 1 and Step 2, respectively. For*

both cases, we need to invoke substantially new technical arguments to achieve our results. For the non-zero means, using an interesting division of the range of $|X_i|$ (keeping in mind the minimax rate) and the monotonicity of the posterior mean of the shrinkage coefficients, we are able to exploit some known results of Ghosh and Chakrabarti (2017) for the case when τ is used as a tuning parameter. For Step 2, we divide our calculations based on $a \in [\frac{1}{2}, 1)$ and $a \geq 1$. When $a \in [\frac{1}{2}, 1)$, using some novel handling of cases for different ranges of $|X_i|$ and $\widehat{\tau}$, we are again able to exploit the monotonicity of the expected posterior shrinkage coefficients and known results for the fixed τ cases. However, for $a \geq 1$, we need to substantially modify our arguments, since using similar arguments as before produces an upper bound on the posterior variance which exceeds the near minimax rate. To take care of this, We need to come up with a sharper upper bound on the posterior shrinkage coefficient compared to that in Ghosh and Chakrabarti (2017) in Lemma 2.2.3 presented below and employ a very handy upper bound on the posterior variance (given in equation (2.42) of the Section 2.5).

We end this subsection by stating and proving an important lemma related to the upper bound on the posterior mean of the shrinkage coefficient which is crucial for the proof of the Theorem 2.2.1.

Lemma 2.2.3. *Suppose $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\theta}_0, \mathbf{I}_n)$. Consider the class of priors (2.2) satisfying (2.3) with $a \geq 1$. Then for any $\tau \in (0,1)$ and any $x_i \in \mathbb{R}$,*

$$E(1-\kappa_i|x_i,\tau) \leq \left(\tau^2 e^{\frac{x_i^2}{4}} + K\tau^2 \int_1^\infty \frac{1}{(1+t\tau^2)^{\frac{3}{2}}} t^{-a} L(t) e^{\frac{x_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}} dt\right)(1+o(1)), \tag{2.6}$$

where the term $o(1)$ depends only on τ and $\lim_{\tau\to 0} o(1) = 0$ and $\int_0^\infty t^{-a-1}L(t)dt = K^{-1}$.

Proof of Lemma 2.2.3. Posterior distribution of κ_i given x_i and τ is,

$$\pi(\kappa_i|x_i,\tau) \propto \kappa_i^{a-\frac{1}{2}}(1-\kappa_i)^{-a-1} L(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)) \exp\left\{\frac{(1-\kappa)x_i^2}{2}\right\}, 0<\kappa_i<1.$$

Using the transformation $t = \frac{1}{\tau^2}(\frac{1}{\kappa}-1)$, $E(1-\kappa_i|x_i,\tau)$ becomes

$$E(1-\kappa_i|x_i,\tau) = \frac{\tau^2 \int_0^\infty (1+t\tau^2)^{-\frac{3}{2}} t^{-a} L(t) e^{\frac{x_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}} dt}{\int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t) e^{\frac{x_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}} dt}.$$

Note that,

$$\int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t) e^{\frac{x_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}} dt \geq \int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t) dt = K^{-1}(1+o(1)),$$

where $o(1)$ depends only on τ and tends to zero as $\tau \to 0$. The equality in the last line follows due to the Dominated Convergence Theorem. Hence

$$E(1-\kappa_i|x_i,\tau) \leq K(A_1+A_2)(1+o(1)), \tag{2.7}$$

where

$$A_1 = \int_0^1 \frac{t\tau^2}{1+t\tau^2} \cdot \frac{1}{\sqrt{1+t\tau^2}} t^{-a-1} L(t) e^{\frac{x_i^2}{2} \cdot \frac{t\tau^2}{1+t\tau^2}} dt \, .$$

and $A_2 = \int_1^\infty \frac{t\tau^2}{1+t\tau^2} \cdot \frac{1}{\sqrt{1+t\tau^2}} t^{-a-1} L(t) e^{\frac{x_i^2}{2} \cdot \frac{t\tau^2}{1+t\tau^2}} dt$. Since, for any $t \leq 1$ and $\tau \leq 1$, $\frac{t\tau^2}{1+t\tau^2} \leq \frac{1}{2}$, using the fact $\int_0^\infty t^{-a-1} L(t) dt = K^{-1}$ (obtained from (2.3)),

$$A_1 \leq K^{-1} \tau^2 e^{\frac{x_i^2}{4}} \, . \tag{2.8}$$

Using the (2.7) and (2.8) along with the form of A_2, for any $\tau \in (0,1)$ and $x_i \in \mathbb{R}$, we get the result of the form (2.6). □

Remark 2.2.4. *Lemma 2.2.3 is a refinement of Lemma 2 of Ghosh and Chakrabarti (2017). The proof is obtained by a careful inspection and division of the range of the integral in* $(0,1)$ *and* $[1,\infty)$*. Note that, the upper bound of* $E(1-\kappa_i|x_i,\tau)$ *within the interval* $(0,1)$ *is sharper than that of Ghosh and Chakrabarti (2017). This order is important for obtaining the optimal posterior contraction rate for the empirical Bayes treatment when* $a \geq 1$ *and hence is a key ingredient for proving the Theorem 2.2.1. In this proof, the second term in the r.h.s. of (2.6) is either used unchanged or a further upper bound on it needs to be employed as may be seen in the proof of* ***Case-2*** *for Theorem 2.2.1 given in Section 2.5.*

2.2.2 Empirical Bayes Procedure-Maximum Marginal Likelihood Estimator(MMLE)

As an alternative to the simple estimator in van der Pas et al. (2014), van der Pas et al. (2017a) proposed the Maximum Marginal Likelihood Estimator (MMLE) of τ, defined as

$$\widehat{\tau}_n = \underset{\tau \in [\frac{1}{n}, 1]}{\operatorname{argmax}} \prod_{i=1}^n \int_{-\infty}^{\infty} \phi(x_i - \theta_i) g_\tau(\theta_i) d\theta_i. \tag{2.9}$$

Here $\phi(\cdot)$ denotes the density of a standard normal random variable and $g_\tau(\theta_i)$ is the marginal pdf of θ_i given τ defined as

$$g_\tau(\theta_i) = \int_0^\infty \frac{1}{\lambda_i \tau} \phi(\frac{\theta_i}{\lambda_i \tau}) \pi_1(\lambda_i^2) d\lambda_i^2.$$

van der Pas et al. (2017a) studied MMLE when the unknown mean vector is modeled by the horseshoe prior. They proved that the empirical Bayes posterior distribution of the mean vector (with τ estimated by the MMLE) contracts around the truth at the near-minimax rate. We want to investigate if this result can be generalized when the mean vector is modeled by a broader class of one-group priors of which the horseshoe is a member. We recall at this point that in the literature of global-local priors used for sparse modelling, the parameter τ has to play in a sense the role of the fraction of non-zero parameters and is usually of the order of the level of sparsity or thereabouts. See, for e.g., Datta and Ghosh (2013), van der Pas et al. (2014), Ghosh and Chakrabarti (2017) etc. This suggests that any empirical Bayes estimate that turns out to

be close to such an optimal value with high probability might turns out to be also optimal. We also note that, in sparse situation the fraction tends to zero. So, it means the maximization can be restricted in an interval which contains the optimal value. This motivates us to redefine the MMLE by the maximization w.r.t. τ in an interval $[\frac{1}{n}, \frac{1}{\log n}]$, i.e.

$$\widehat{\tau}_n = \underset{\tau \in [\frac{1}{n}, \frac{1}{\log n}]}{\operatorname{argmax}} \prod_{i=1}^{n} \int_{-\infty}^{\infty} \phi(x_i - \theta_i) g_\tau(\theta_i) d\theta_i, \tag{2.10}$$

This also ensures that the optimal value of τ is of a smaller order than $\frac{1}{\log n}$. We are able to establish in our Theorem 2.2.7 (stated at the end of this section) that the empirical Bayes posterior distribution of the mean vector corresponding to the MMLE as given in (2.10) contracts around the truth at the near-minimax rate for the class of three parameter beta normal priors (with $a = \frac{1}{2}$ and $\alpha \geq \frac{1}{2}$). This class of priors was introduced by Armagan et al. (2011) and is defined as

$$\begin{aligned} \theta_i | \lambda_i^2, \tau^2 &\overset{ind}{\sim} \mathcal{N}(0, \lambda_i^2 \tau^2), \\ \lambda_i^2 &\overset{ind}{\sim} \pi_1(\lambda_i^2), \\ \tau &\sim \pi_2(\tau), \end{aligned} \tag{2.11}$$

where

$$\pi_1(\lambda_i^2) = \frac{\Gamma(a+\alpha)}{\Gamma(a)\Gamma(\alpha)} (\lambda_i^2)^{\alpha-1} (1+\lambda_i^2)^{-(\alpha+a)}$$

for $\alpha > 0, a > 0$. Ghosh et al. (2016) has shown that the above class of priors can be represented as (2.2) where $\pi_1(.)$ satisfies (2.3) with $L(\lambda_i^2) = (1 + \frac{1}{\lambda_i^2})^{-(a+\alpha)}$ with $a > 0$ and $\alpha > 0$ and $K = \frac{\Gamma(a+\alpha)}{\Gamma(a)\Gamma(\alpha)}$. Three parameter beta normal priors is a rich class of priors containing horseshoe for $a = \frac{1}{2} = \alpha$, the Strawderman–Berger prior with $\alpha = 1, a = \frac{1}{2}$ and the normal–exponential–gamma priors with $\alpha = 1, a > 0$ as some special cases.

The proof of Theorem 2.2.7 crucially depends on proving that the MMLE based on the three parameter beta normal family satisfies condition (C1) below. This condition is exactly the same as that of van der Pas et al. (2017a), which was proved in that paper to be satisfied by the MMLE based on the horseshoe. It may be recalled from the results and discussions of van der Pas et al. (2014) and Ghosh and Chakrabarti (2017) that $\tau_n(q_n)$ is an optimal order of τ which ensures that the posterior distribution of the mean vector given τ contracts around the truth at the minimax rate when τ is fixed at $\tau_n(q_n)$ as a tuning parameter. Thus satisfying condition (C1) by an empirical Bayes estimator of τ ensures that the estimator is at most of an optimal order and it may be expected that the empirical Bayes posterior using such an estimator might also have similar optimal behaviour. Interestingly, we have observed while proving Theorem 2.2.1 that the simple estimator of van der Pas et al. (2014) also satisfies a similar condition. We now state condition (C1).

(C1) There exists a constant $C > 0$ such that the empirical Bayes estimator of τ, say $\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]$, with P_{θ_0}-probability tending to one, uniformly in $\theta_0 \in l_0[q_n]$.

Lemma 2.2.5, presented below, shows that the MMLE satisfies condition (C1). Proof of this

lemma is given in Section 2.5.

Lemma 2.2.5. *The MMLE defined in* (2.10) *satisfies (C1).*

Remark 2.2.6. *In order to prove Lemma 2.2.5, we need to prove in a series of lemmas (Lemmas 2.5.2-2.5.9 of Section 2.5, similar to those of van der Pas et al. (2017a)) results related to the upper bound of the derivative (with respect to τ) of the logarithm of the marginal likelihood function of $\mathbf{X}$ given τ. Proofs of all of these lemmas are given in Section 2.5. The proof starts by showing that the derivative can be expressed as $\frac{1}{\tau}\sum_{i=1}^{n} m_\tau(x_i)$ of suitably defined function $m_\tau(x_i)$. The next step is to decompose the sum based on x_i coming from zero and non-zero means and to study their behaviours separately. This study reveals that the expectations of $m_\tau(x_i)$ corresponding to zero means are strictly negative and bounded away from zero for $\tau \geq \epsilon$, given any $\epsilon > 0$. We have also noticed that the average of $m_\tau(x_i)$ terms for zero means behaves like its expectation, uniformly in τ. On the other hand, the function $m_\tau(x_i)$ is uniformly bounded by a constant, which is used in the case of non-zero means. All of these arguments are essential for proving Lemma 2.2.5. van der Pas et al. (2017a) had shown that all of these arguments hold for horseshoe prior and stated a condition exactly as same as (C1). However, our calculations show that the same holds even for the three parameter beta normal class of priors containing horseshoe as a special case. In spite of the lemmas of van der Pas et al. (2017a) being available for the horseshoe prior, it is a non-trivial task to extend their results to the three parameter beta normal class of priors due to the more general form of the marginal density of $\mathbf{X}$ given τ. In our case the $m_\tau(x_i)$ function is significantly different from that of van der Pas et al. (2017a) and it requires new technical observations and manipulations to obtain results similar to those of van der Pas et al. (2017a) in this context. In this way, Lemma 2.2.5 generalizes Theorem 3.1 of van der Pas et al. (2017a).*

Theorem 2.2.7. *Suppose $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\theta}_0, \mathbf{I}_n)$ where $\boldsymbol{\theta}_0 \in l_0[q_n]$ with $q_n \propto n^\beta$ where $0 < \beta < 1$. Consider the class of priors* (2.2) *satisfying* (2.3) *where $L(t) = (1 + 1/t)^{-(\frac{1}{2}+\alpha)}$ with $\alpha \geq \frac{1}{2}$. Then, for any empirical Bayes estimator of τ, say $\widehat{\tau}_n$ that satisfies (C1), the empirical Bayes posterior contracts around the true parameter $\boldsymbol{\theta}_0$ at the near minimax rate, namely,*

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} \Pi_{\widehat{\tau}_n}\left(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}\right) \to 0 \,,$$

for any $M_n \to \infty$.

The proof of the theorem is given in Section 2.5.

Remark 2.2.8. *Proof of Theorem 2.2.7 starts with the division of the range of $\widehat{\tau}_n$ as mentioned in Lemma 2.2.5 and next uses Lemma 2.2.5. After that, we use Markov's inequality to show that both the expected value of the posterior variance and the square of the L_2 loss based on MMLE of τ are bounded by $q_n \log n$, for both zero and non-zero means. For non-zero means, we use the same technique as used in Theorem 2.2.1 there. For zero means, arguments similar to Theorem 2.2.1 and Theorem 2 of Ghosh and Chakrabarti (2017) become useful.*

2.2.3 A Full Bayes Approach

In this section, we consider a full Bayes approach where we assume a non-degenerate prior $\pi_{2n}(\tau)$ on the global shrinkage parameter τ. We are making explicit the dependence on n of the prior on τ.

Motivated by van der Pas et al. (2017a), we consider priors $\pi_{2n}(\tau)$ which satisfy the conditions (C2) and (C3) below.

(C2) $\pi_{2n}(.)$ is supported on $[\frac{1}{n}, \frac{1}{\log n}]$.

(C3) Let $t_n = \frac{\sqrt{\pi}}{K}\tau_n(q_n)$ where $K = \frac{\Gamma(\frac{1}{2}+\alpha)}{\sqrt{\pi}\Gamma(\alpha)}$ as given in the previous subsection. Then $\pi_{2n}(.)$ satisfies

$$\left(\frac{q_n}{n}\right)^{M_1} \int_{\frac{t_n}{2}}^{t_n} \pi_{2n}(\tau)d\tau \gtrsim e^{-cq_n}, \text{ for some } c \leq \frac{1}{2}, M_1 \geq 1 \,.$$

It may be noted that the empirical Bayes estimates of τ considered in (2.4) and (2.10) always lie within $[\frac{1}{n}, 1]$ and we have proved in Theorem 2.2.1 of Subsection 2.2.1 and Theorem 2.2.7 of Subsection 2.2.2 that the empirical Bayes posterior contracts around the truth at a near minimax rate. This is a motivation behind (C2) i.e. the restriction of the support of $\pi_{2n}(\cdot)$ to the interval $[\frac{1}{n}, \frac{1}{\log n}]$. (C3) is a slightly modified version of the condition therein but is asymptotically equivalent to it.

When q_n is known, van der Pas et al. (2014) proved that the horseshoe posterior attains the minimax rate when the global shrinkage parameter τ is chosen of the order of $\tau_n(q_n) = (q_n/n)\sqrt{\log(n/(q_n))}$. van der Pas et al. (2014) also stated that a choice of τ that goes to zero slower than $\tau_n(q_n)$, say the $(\frac{q_n}{n})^{\alpha}, 0 < \alpha < 1$, will lead to a sub-optimal rate of posterior contraction in the squared L_2 sense. Ghosh and Chakrabarti (2017) had the same observations for a broad class of priors. Condition (C3) guarantees that there will be sufficient prior mass around the *optimal* value of τ so that the full Bayes posterior distribution may be expected to have a high probability around the optimal values and may lead to a near-optimal rate of contraction. This condition is satisfied by many prior densities, except in the very sparse cases when $q_n \lesssim \log n$. Using the same arguments as in van der Pas et al. (2017a), it can be shown that the half Cauchy distribution supported on $[\frac{1}{n}, 1]$ having density $\pi_{2n}(\tau) = (\arctan(1) - \arctan(1/n))^{-1}(1+\tau^2)^{-1}\mathbf{1}_{\tau\in[\frac{1}{n},1]}$ and uniform prior on $[\frac{1}{n}, 1]$ with density $\pi_{2n}(\tau) = n/(n-1)\mathbf{1}_{\tau\in[\frac{1}{n},1]}$ satisfy condition (C3) provided $q_n \gtrsim \log n$.

Note that, using (C2), the posterior mean of θ_i in the full Bayes approach can be written as

$$\begin{aligned}
\widehat{\theta_i} = E(\theta_i|\mathbf{X}) &= \int_{\frac{1}{n}}^{\frac{1}{\log n}} E(\theta_i|\mathbf{X}, \tau)\pi_{2n}(\tau|\mathbf{X})d\tau \\
&= \int_{\frac{1}{n}}^{\frac{1}{\log n}} E(1-\kappa_i|X_i, \tau)X_i \cdot \pi_{2n}(\tau|\mathbf{X})d\tau \\
&= \int_{\frac{1}{n}}^{\frac{1}{\log n}} T_\tau(X_i)\pi_{2n}(\tau|\mathbf{X})d\tau. \qquad (2.12)
\end{aligned}$$

Thus the full Bayes posterior mean is expressed as a weighted average of the posterior mean given τ, the weight being the posterior density of τ given data.

The next lemma is crucial for deriving the optimal posterior contraction rate for the three parameter beta normal priors with $a = \frac{1}{2}$. This states that the full Bayes posterior of τ asymptotically concentrates its mass in a neighbourhood of zero of size at most a constant multiple of t_n away from zero.

Lemma 2.2.9. *If (C2) and (C3) are satisfied, then*

$$\sup_{\theta_0 \in l_0[q_n]} E_{\theta_0} \Pi(\tau \geq 5t_n | \mathbf{X}) \to 0.$$

Proof of Lemma 2.2.9. Using a similar set of arguments used in Lemma 3.6 of van der Pas et al. (2017a), with P_{θ_0}-probability tending to one, we have

$$\frac{d}{d\tau} M_\tau(\mathbf{X}) < \begin{cases} \frac{q_n}{t_n/2} & , \text{if } t_n/2 \leq \tau \leq t_n \\ 0 & , \text{if } t_n \leq \tau \leq 2t_n \\ -\frac{q_n}{2t_n} & , \text{if } \tau \geq 2t_n. \end{cases} \tag{2.13}$$

As a result of this, using (2.13) for $\tau \geq 2t_n$, we have

$$\begin{aligned} M_\tau(\mathbf{X}) - M_{\tau_{min}}(\mathbf{X}) &= [\int_{\tau_{min}}^{t_n} + \int_{t_n}^{2t_n} + \int_{2t_n}^{\tau}] \frac{d}{ds} M_s(\mathbf{X}) ds \\ &\leq -\frac{q_n(\tau - 4t_n)}{2t_n}, \end{aligned} \tag{2.14}$$

where $\tau_{min} = \text{argmin}_{\tau \in [\frac{t_n}{2}, t_n]} M_\tau(\mathbf{X})$. Also note that, for $\tau \geq 5t_n$, the r.h.s. of (2.14) can be further bounded above by $-\frac{q_n}{2}$. Since, $\pi(\tau|\mathbf{X}) \propto \pi(\tau) e^{M_\tau(\mathbf{X})}$ and $M_\tau(\mathbf{X}) \leq M_{\tau_{min}}(\mathbf{X}) - \frac{q_n}{2}$, for $\tau \geq 5t_n$, with P_{θ_0}-probability tending to one, one can show that

$$\Pi(\tau \geq 5t_n | \mathbf{X}) \lesssim \frac{e^{-0.5q_n}}{\int_{\frac{t_n}{2}}^{t_n} \pi_{2n}(\tau) d\tau}, \tag{2.15}$$

which tends to zero using (C3). As a result, for $b_n = (\frac{q_n}{n})^{M_1}$ and $h(\mathbf{X}) = \Pi(\tau \geq 5t_n | \mathbf{X})$,

$$\begin{aligned} E_{\theta_0} h(\mathbf{X}) &= E_{\theta_0}[h(\mathbf{X}) 1_{\{h(\mathbf{X}) \lesssim b_n\}}] + E_{\theta_0}[h(\mathbf{X}) 1_{\{h(\mathbf{X}) \gtrsim b_n\}}] \\ &\lesssim b_n + P_{\theta_0}[h(\mathbf{X}) \gtrsim b_n]. \end{aligned}$$

The proof follows from this since by definition the first term goes to zero and on the other hand, the second term goes to zero due to (2.13). □

Remark 2.2.10. *For obtaining arguments similar to Lemma 3.6 of van der Pas et al. (2017a), we need to come up with an upper bound on the derivative of the log-likelihood of* $\mathbf{X}$ *given* τ*. This is obtained with the help of Lemmas 2.5.2-2.5.9 given in the Section 2.5. A brief discussion regarding the importance of these lemmas and how these help us to establish Lemma 2.2.9 is also given in Remark 2.2.6.*

The next theorem proves that the hierarchical Bayes posterior estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ also attains near minimax L_2 risk. The proof of the theorem is given in Section 2.5.

Theorem 2.2.11. *Suppose* $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\theta}_0, \mathbf{I}_n)$ *where* $\boldsymbol{\theta}_0 \in l_0[q_n]$ *with* $q_n \propto n^\beta$ *where* $0 < \beta < 1$. *Consider the class of priors* (2.2) *satisfying* (2.3) *where* $L(t) = (1 + 1/t)^{-(\frac{1}{2}+\alpha)}$ *with* $\alpha \geq \frac{1}{2}$. *If the prior on* τ *satisfies conditions (C2) and (C3), then the full Bayes posterior contracts around the true parameter* $\boldsymbol{\theta}_0$ *at the near minimax rate, namely,*

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} \Pi\Big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}\Big) \to 0,$$

for any $M_n \to \infty$.

Remark 2.2.12. *For proving Theorem 2.2.11, first we use the fact that* $\pi(\theta|\mathbf{X}) = \int_{\frac{1}{n}}^{\frac{1}{\log n}} \pi(\theta|\mathbf{X}, \tau)\pi(\tau|\mathbf{X})d\tau$. *Next, we divide the range of* τ *in order to make use of Lemma 2.2.9. After that, we use Markov's inequality to show that both the expected value of the posterior variance and the squared* L_2 *loss are bounded by* $q_n \log n$, *for both zero and non-zero means. For non-zero means, some techniques used in Theorem 2.2.1 and for zero means, some arguments of Ghosh and Chakrabarti (2017) become handy.*

2.3 Asymptotic optimality of a full Bayes approach in a problem of hypothesis testing

Suppose we have observations $X_i, i = 1, 2, \cdots, n$, modeled through (2.1) and our problem is to test simultaneously the hypotheses $H_{0i} : \theta_i = 0$ against $H_{1i} : \theta_i \neq 0$ for $i = 1, 2, \cdots, n$. Our interest is in the sparse asymptotic regime when n is large and the proportion of non-zero means is expected to be small. Formally, we will be using the asymptotic framework of Bogdan et al. (2011) in this section. As in Bogdan et al. (2011), we introduce a set of iid binary latent variables $\nu_1, \nu_2, \cdots, \nu_n$ where $\nu_i = 0$ refers to the event that H_{0i} is true and $\nu_i = 1$ denotes that H_{1i} is true with $P(\nu_i = 1) = p$ for $i = 1, 2, \cdots, n$. Given $\nu_i = 0$, θ_i has a degenerate distribution at 0 and when $\nu_i = 1$, θ_i is assumed to have a $\mathcal{N}(0, \psi^2)$ distribution with $\psi^2 > 0$. In this Bayesian setting, the priors on θ_i's are given by

$$\theta_i \overset{iid}{\sim} (1 - p)\delta_0 + p\mathcal{N}(0, \psi^2), i = 1, 2, \cdots, n. \tag{2.16}$$

The marginal distribution of X_i for two-group model is thus of the form

$$X_i \overset{iid}{\sim} (1 - p)\mathcal{N}(0, 1) + p\mathcal{N}(0, 1 + \psi^2), i = 1, 2, \cdots, n. \tag{2.17}$$

Our testing problem is equivalent to test simultaneously

$$H_{0i} : \nu_i = 0 \text{ versus } H_{1i} : \nu_i = 1, \text{ for } i = 1, 2, \cdots, n. \tag{2.18}$$

The overall loss in a multiple testing rule is defined as the total number of misclassifications made by the test which is nothing but a sum $0-1$ losses corresponding to each individual test. If t_{1i} and t_{2i} denote the probabilities of type I and type II errors respectively for the i^{th} test, the Bayes risk R of a multiple testing procedure is given by

$$R = \sum_{i=1}^{n}[(1-p)t_{1i} + pt_{2i}]. \tag{2.19}$$

Bogdan et al. (2011) showed that the multiple testing rule minimizing the Bayes risk (2.19) rejects H_{0i} if

$$\pi(\nu_i = 1|X_i) > \frac{1}{2}, \text{ i.e. } X_i^2 > c^2 \tag{2.20}$$

and accepts H_{0i} otherwise, for each $i = 1, 2, \cdots n$. Here $\pi(\nu_i = 1|X_i)$ denotes the posterior probability of H_{1i} to be true and $c^2 \equiv c_{p,\psi}^2 = \frac{1+\psi^2}{\psi^2}(\log(1+\psi^2) + 2\log f)$ with $f = \frac{1-p}{p}$. Due to the presence of unknown model parameters p and ψ^2 in (2.20), this rule is termed the Bayes Oracle. For defining their asymptotic framework, Bogdan et al. (2011) introduced two parameters $u \equiv u_n = \psi_n^2$ and $v \equiv v_n = u_n f_n^2$ and also assumed the following.

Assumption 2. $p_n \to 0$, $u_n = \psi_n^2 \to \infty$ and $\frac{\log v_n}{u_n} \to C \in (0, \infty)$ as $n \to \infty$.

For simplicity of notation, we will henceforth remove the dependence of the model parameters on n. Under Assumption 2, the asymptotic expression for the Bayes risk of the Bayes Oracle (2.20), denoted by $R_{\text{Opt}}^{\text{BO}}$ is of the form

$$R_{\text{Opt}}^{\text{BO}} = np\left[2\Phi(\sqrt{C}) - 1\right](1 + o(1)). \tag{2.21}$$

A multiple testing rule is defined to Asymptotically Bayes Optimal under Sparsity (ABOS) if the ratio of its Bayes risk and $R_{\text{Opt}}^{\text{BO}}$ converges to 1 under the asymptotic framework of Assumption 2.

As mentioned in the introduction, Datta and Ghosh (2013) first studied from a decision theoretic viewpoint, the asymptotic optimality of a testing rule, based on the horseshoe prior in this problem. Treating the global parameter τ as a tuning parameter, they studied the following decision rule for horseshoe prior,

$$\text{reject } H_{0i} \text{ if } E(1 - \kappa_i|X_i, \tau) > \frac{1}{2}, i = 1, 2, \cdots, n. \tag{2.22}$$

They showed that under Assumption 2, their proposed decision rule (2.22) attains the optimal Bayes risk (2.21) up to some multiplicative constant if $\tau = O(p)$. Later Ghosh et al. (2016) obtained similar results (under the condition that $\lim_{n\to\infty} \frac{\tau}{p} \in (0, \infty)$) using the same testing rule as above but modeling θ_i by a general class of one-group global-local shrinkage priors where τ is treated as a tuning parameter. By choosing $\tau \sim p$ and considering a prior of the form (2.2) satisfying (2.3) with $a = \frac{1}{2}$ with $L(\cdot)$ satisfying Assumption 1, Ghosh and Chakrabarti (2017) proved that the induced decision rule is ABOS for the subclass of priors, named as horseshoe-

type priors by Ghosh and Chakrabarti (2017). This result demonstrates that appropriately chosen one-group prior can be a reasonable alternative to the two-group prior in this problem.

When p is unknown, Ghosh and Chakrabarti (2017) considered an empirical Bayes version of (2.22) for the same class of priors as above and used the following decision rule,

$$\text{reject } H_{0i} \text{ if } E(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, i = 1, 2, \cdots, n, \tag{2.23}$$

where $\widehat{\tau}$ is defined in (2.4). Under Assumption 2, they proved that for $p_n \propto n^{-\epsilon}$, $0 < \epsilon < 1$ and for $a = \frac{1}{2}$ with $L(\cdot)$ defined in (1.10) satisfying Assumption 1, the decision rule (2.23) is also ABOS.

We are interested in a full Bayes approach to the same problem where one puts a non-degenerate prior on τ. In that case the decision rule modifies to

$$\text{reject } H_{0i} \text{ if } E(1-\kappa_i|\mathbf{X}) > \frac{1}{2}, i = 1, 2, \cdots, n. \tag{2.24}$$

Note that, the rule (2.24) depends on the entire data set and is clearly different from (2.22) and (2.23). The study of asymptotic optimality of a rule like (2.24) is still missing in the literature, even for the horseshoe prior. Ghosh et al. (2016) had done some simulation studies using the decision rule (2.24) when a standard half Cauchy prior was used for τ. Their results obtained through simulations motivate us to study the decision rule (2.24) in a formal decision-theoretic manner. Our goal is to find a general class of priors on τ under which the decision rule (2.24) behaves in an optimal or near-optimal manner in terms of Bayes risk.

We start with a general class of priors $\pi_{2n}(\cdot)$ satisfying the following condition:

(C4) $\int_{\frac{1}{n}}^{\alpha_n} \pi_{2n}(\tau)d\tau = 1$ for some α_n such that $\alpha_n \to 0$ and $n\alpha_n \to \infty$ as $n \to \infty$.

It may be recalled that for proving the optimality of the decision rule based on (2.22), a choice of $\tau = O(p)$ is sufficient. In the present situation, p is unknown, so intuition suggests that it might be a good idea to start with a prior on τ which gives sufficient prior probability around such values of τ. We were initially interested in situations where the unknown $p \propto n^{-\epsilon}, 0 < \epsilon \leq 1$. So the choice of the lower boundary of the support as $\frac{1}{n}$ at least ensures positive prior mass around $\frac{1}{n}$ and $\frac{1}{n} = O(p)$ as $n \to \infty$ when $p \propto n^{-\epsilon}, 0 < \epsilon \leq 1$. Later, it turns out that instead of assuming a specific order of p (e.g., $p \propto n^{-\epsilon}, 0 < \epsilon \leq 1$), we can assume more general conditions on p for studying the optimality result based on the decision rule (2.24). Given this, the specific choices of α_n that ensure asymptotic optimality become clear from Theorems 2.3.1-2.3.4 and the discussions thereafter. In order to keep the support of τ as broad as possible, we assume $n\alpha_n \to \infty$ as $n \to \infty$. The proofs of these theorems are given in Section 2.5.

Theorem 2.3.1. *Let $X_1, X_2, \cdots, X_n$ be iid with distribution (2.17). Suppose we test (2.18) using decision rule (2.24) induced by the class of priors (2.2) with $\pi_{2n}(\cdot)$ satisfying (C4) and $\pi_1(\cdot)$ satisfying (2.3) with $a \in (0, 1)$, where $L(\cdot)$ satisfies Assumption 1. Then the probability of*

type-I error of the decision rule (2.24), *denoted* t_{1i}^{FB}, *satisfies*

$$t_{1i}^{FB} \lesssim \frac{\alpha_n^{2a}}{\sqrt{\log\left(\frac{1}{\alpha_n^2}\right)}}(1+o(1)),$$

where $o(1)$ *term is independent of* i *and depends only on* n *such that* $\lim_{n\to\infty} o(1) = 0$.

Let us now state a result on the upper bound on the probability of type II error induced by the decision rule (2.24).

Theorem 2.3.2. *Consider the set-up of Theorem 2.3.1. Assume that* $p_n \to 0$ *as* $n \to \infty$ *such that* $\frac{\log\left(\frac{1}{p_n}\right)}{\log n} \to C_1$, *for some constant* $C_1 > 0$. *Then under Assumption 2, for any* $\rho > 2$, *the probability of type-II error induced by the decision rule* (2.24), *denoted* t_{2i}^{FB} *satisfies,*

$$t_{2i}^{FB} \leq \left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1)).$$

where the $o(1)$ *term is independent of* i *and satisfies* $\lim_{n\to\infty} o(1) = 0$.

Remark 2.3.3. *Combining Theorems 2.3.1 and 2.3.2, we immediately see that the Bayes risk of the decision rule* (2.24), *denoted* R_{OG}^{FB} *satisfies, for any* $\rho > 2$,

$$R_{OG}^{FB} \leq n\left[(1-p)\frac{\alpha_n^{2a}}{\sqrt{\log\left(\frac{1}{\alpha_n^2}\right)}}(1+o(1)) + p\left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1))\right],$$

where $o(1)$ *term is independent of* i *and is such that* $\lim_{n\to\infty} o(1) = 0$. *We are now in a position to choose* α_n, p *and* a *appropriately so that this upper bound can be made of the same order as* R_{Opt}^{BO} *in* (2.21). *The result is stated in Theorem 2.3.4 below.*

Theorem 2.3.4. *Let* $X_1, X_2, \cdots, X_n$ *be iid with distribution* (2.17). *Suppose we test* (2.18) *using decision rule* (2.24) *induced by the class of priors* (2.2) *with* $\pi_{2n}(\cdot)$ *satisfying (C4) and* $\pi_1(\cdot)$ *satisfying* (2.3) *with* $a \in [0.5, 1)$, *where* $L(\cdot)$ *satisfies Assumption 1. Assume* $\log\left(\frac{1}{\alpha_n}\right) = \log n - \frac{1}{2}\log\log n + c_n$, *where* $c_n \to \infty$ *such that* $c_n = o(\log\log n)$ *as* $n \to \infty$. *Also assume* $p_n \to 0$ *as* $n \to \infty$ *such that* $\underline{\lim}\, np_n > 0$ *and* $\frac{\log\left(\frac{1}{p_n}\right)}{\log n} \to C_1$, *for some constant* $0 < C_1 \leq 1$. *Then under Assumption 2, for any* $\rho > 2$, *the Bayes risk of the decision rule* (2.24), *denoted* R_{OG}^{FB} *satisfies*

$$R_{OG}^{FB} \leq np\left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1)), \tag{2.25}$$

where $o(1)$ *term is independent of* i *and is such that* $\lim_{n\to\infty} o(1) = 0$.
For $C_1 = 1$ *with* $a = 0.5$,

$$\lim_{n\to\infty} \frac{R_{OG}^{FB}}{R_{Opt}^{BO}} = 1,$$

i.e. the rule (2.24) *is ABOS.*

2.4 Concluding remarks and scope for future work

In this chapter, we study inference on a high-dimensional sparse normal means model when the proportion of non-zero means is unknown. In this work, we have generalized the results of van der Pas et al. (2014) and van der Pas et al. (2017a) by establishing that for both the empirical Bayes and full Bayes approaches, the posterior distributions of the mean vector contract around the truth at a near minimax rate for broad classes of priors containing horseshoe as a special one. We then study the asymptotic properties of a decision rule in the full Bayes approach based on global-local priors in a problem of simultaneous hypothesis testing. Here we can assign arbitrary priors on the global shrinkage parameter τ provided its support is suitably chosen, but still ensure optimal performance as described in the following discussion. Within the asymptotic framework of Bogdan et al. (2011), we have been able to prove that the Bayes risk induced by this broad class of priors attains the optimal Bayes risk, up to some multiplicative constant. Moreover, when the underlying model is extremely sparse, the decision rule is indeed ABOS for horseshoe-type priors. This is the first result of its kind in the literature. We hope that our results in this chapter will reinforce the logic for using the global-local priors in sparse situations as an alternative to their two-group counterparts.

Our chosen classes of priors are rich enough to include many interesting priors, but due to the assumption of boundedness on the slowly varying function, Horseshoe+ prior is not included in this class. It does not appear that our techniques can be easily extended for the Horseshoe+ due to Bhadra et al. (2017). van der Pas et al. (2016) proved that the optimal posterior concentration result also holds for Horseshoe+ using appropriate choice of τ. No such result is still available in the literature for the same when either the empirical Bayes estimate of τ based on maximizing the marginal likelihood function is used or a full Bayes procedure is employed. We would like to consider this problem elsewhere.

It is worth mentioning that although most of our results related to optimal contraction rate and asymptotic Bayes optimality are valid for a broad class of global-local priors, the results related to the posterior concentration rate either based on MMLE or the full Bayes approach, have been proved only for the three parameter beta normal class of priors. The techniques used in this context do not seem to be applicable, e.g., for the Generalized double Pareto priors. Thus a question which is still unanswered is whether the same results hold for the Generalized double Pareto priors, or, for that matter, for a more general class of priors that includes both the three parameter beta normal and the Generalized double Pareto. This might be considered as an interesting problem to be studied in the future.

Throughout this work, we use the concept of slowly varying functions. van der Pas et al. (2016) considered a broader class of global-local priors that contains our chosen class of priors (2.2) satisfying (2.3), by using the concept of "uniformly regularly varying functions". They established optimal posterior contraction rate for this class of priors when the level of sparsity is known, by choosing the global shrinkage parameter τ appropriately based on the level of sparsity. To the best of our knowledge, no such result is available in the literature when the level of sparsity is unknown. These priors have not been studied in the context of multiple testing either when the proportion of non-null means is unknown. These will be interesting

problems to consider and we expect that some of the techniques employed in this work will be helpful in these studies also.

It is important to highlight that the results of Bogdan et al. (2011) for sparse normal means problem have been recently extended by Roy and Bhandari (2024) when the test statistics are dependent. In the asymptotic framework of Bogdan et al. (2011), sufficient conditions were established in Roy and Bhandari (2024) under which different popular multiple testing procedures are ABOS when the test statistics under study follow an equicorrelated multivariate normal distribution having equal variance under null and alternative hypotheses. Since the asymptotic Bayes optimality studied in the latter half of our work is also under the assumptions of Bogdan et al. (2011), some comments are in order on the possibility of extending our work for the equicorrelated normal setting. Towards that, we first recall a result from Roy and Bhandari (2024). Suppose $\mathbf{X} \sim \mathcal{MVN}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma}_\rho)$, where σ^2 is the usual error variance and $\boldsymbol{\Sigma}_\rho$ is the correlation matrix having non-diagonal entries to be equal to ρ. Lemma 1 of Roy and Bhandari (2024) shows that there exist V and $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_n)$ such that for each $i = 1, 2, \cdots, n$, $X_i = V + Z_i$, where Z_i is a mixture of two normally distributed random variables, V is also normally distributed and V is independent of Z_i. This also implies, V is a good approximation of $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$ in the sense that $\overline{X} - V \xrightarrow{a.s.} 0$. This observation helps the authors of Roy and Bhandari (2024) in constructing a decision rule based on $\widehat{Z}_i$, where $\widehat{Z}_i = X_i - \overline{X}$. Roy and Bhandari (2024) established that instead of using entire $\mathbf{X}$, a decision rule based on an estimate of pseudo observations, $\widehat{Z}_i$ will also be ABOS. In our problem, when $\boldsymbol{X}$ is drawn from a multivariate normal distribution with the above mentioned dependency structure present therein, we can use the ABOS rule as our benchmark instead of that based on $\mathbf{X}$ which is intractable. In order to prove ABOS property of a rule based on one-group prior, we can analogously define a decision rule using the posterior distribution of k_i given $\widehat{Z}_i$ and τ in place of k_i given Z_i and τ. But the theoretical study of this rule vis-a-vis the ABOS property will be technically very involved. This is because one needs to establish results similar to Theorems 2.3.1-2.3.4 of our work using observations which are now dependent. The posterior distribution based on one-group prior in such a scenario will not be easily tractable. This leaves an interesting problem for future work.

2.5 Proofs

2.5.1 Results related to the contraction rate of Empirical Bayes Procedure– plug-in Estimator

Proof of Theorem 2.2.1. For proving Theorem 2.2.1, first we need to show that, the total posterior variance using the empirical Bayes estimator of τ corresponding to this class of priors satisfies, as $n \to \infty$,

$$\sup_{\boldsymbol{\theta}_0 \in l_0[q_n]} \mathbb{E}_{\boldsymbol{\theta}_0} \sum_{i=1}^{n} Var(\theta_i | X_i, \widehat{\tau}) \lesssim q_n \log n, \tag{2.26}$$

where $Var(\theta_i|X_i,\widehat{\tau})$ denotes $Var(\theta_i|X_i,\tau)$ evaluated at $\tau=\widehat{\tau}$. We make use of Markov's inequality along with (2.26) and Theorem 2 of Ghosh and Chakrabarti (2017).
Now we move towards establishing (2.26). Let us define $\tilde{q}_n=\sum_{i=1}^n 1_{\{\theta_{0i}\neq 0\}}$. Thus, $\tilde{q}_n\leq q_n$. We then note

$$\mathbb{E}_{\boldsymbol{\theta}_0}\sum_{i=1}^n Var(\theta_i|X_i,\widehat{\tau})=\sum_{i:\theta_{0i}\neq 0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})+\sum_{i:\theta_{0i}=0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})\,. \tag{2.27}$$

We now prove that $\sum_{i:\theta_{0i}\neq 0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})\lesssim \tilde{q}_n\log n$ and $\sum_{i:\theta_{0i}=0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})\lesssim q_n\log n$ in **Step-1** and **Step-2** below respectively. Combining these results we get the final result. Now let us prove **Step-1** and **Step-2**.
Proof of Step-1: Fix any $c>1$ and choose $\rho>c$. Define $r_n=\sqrt{4a\rho^2\log n}$. Fix any i such that $\theta_{0i}\neq 0$. We split $\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})$ as

$$\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})=\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|\leq r_n\}}]+\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|>r_n\}}]\,. \tag{2.28}$$

Since for any fixed $x_i\in\mathbb{R}$ and $\tau>0$, $Var(\theta_i|x_i,\tau)\leq 1+x_i^2$.

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|\leq\sqrt{4a\rho^2\log n}\}}]\leq 1+4a\rho^2\log n\,. \tag{2.29}$$

Note that, for any fixed $x_i\in\mathbb{R}$, $x_i^2E(\kappa_i^2|x_i,\tau)=x_i^2E(\frac{1}{(1+\lambda_i^2\tau^2)^2}|x_i,\tau)$ is non-increasing in τ. Also using Lemma A.1 in Ghosh and Chakrabarti (2017), $Var(\theta_i|x_i,\tau)\leq 1+x_i^2E(\kappa_i^2|x_i,\tau)$. Using these two results,

$$Var(\theta_i|x_i,\widehat{\tau})\leq 1+x_i^2E(\kappa_i^2|x_i,\widehat{\tau})\leq 1+x_i^2E(\kappa_i^2|x_i,\frac{1}{n})\ [\text{Since }\widehat{\tau}\geq\frac{1}{n}]\,.$$

Using arguments similar to Lemma 3 of Ghosh and Chakrabarti (2017), one can show that for any fixed $\eta\in(0,1)$ and $\delta\in(0,1)$, the r.h.s. above can be bounded by a non-negative and measurable real-valued function $\tilde{h}(x_i,\tau,\eta,\delta)$ satisfying $\tilde{h}(x_i,\tau,\eta,\delta)=1+\tilde{h}_1(x_i,\tau)+\tilde{h}_2(x_i,\tau,\eta,\delta)$ with

$$\tilde{h}_1(x_i,\tau)=C_{**}\left[x_i^2\int_0^{\frac{x_i^2}{1+t_0}}\exp\Big(-\frac{u}{2}\Big)u^{a+\frac{1}{2}-1}du\right]^{-1},$$

where C_{**} is a global constant independent of both x_i and τ and

$$\tilde{h}_2(x_i,\tau,\eta,\delta)=x_i^2\frac{H(a,\eta,\delta)}{\Delta(\tau^2,\eta,\delta)}\tau^{-2a}e^{-\frac{\eta(1-\delta)x_i^2}{2}},$$

where $H(a,\eta,\delta)=\frac{(a+\frac{1}{2})(1-\eta\delta)^a}{K(\eta\delta)^{(a+\frac{1}{2})}}$ and $\Delta(\tau^2,\eta,\delta)=\frac{\int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta}-1)}^{\infty}t^{-(a+\frac{3}{2})}L(t)dt}{(a+\frac{1}{2})^{-1}(\frac{1}{\tau^2}(\frac{1}{\eta\delta}-1))^{-(a+\frac{1}{2})}}$.
Since, $\tilde{h}_1(x_i,\tau)$ is strictly decreasing in $|x_i|$ for any fixed τ, one has,

$$\sup_{|x_i|>\sqrt{4a\rho^2\log n}}\tilde{h}_1(x_i,\frac{1}{n})\leq C_{**}\left[\rho^2\log n\int_0^{\frac{4a\rho^2\log n}{1+t_0}}\exp\Big(-\frac{u}{2}\Big)u^{a+\frac{1}{2}-1}du\right]^{-1}\lesssim\frac{1}{\log n}\,.$$

Also noting that $\tilde{h}_2(x_i, \tau, \eta, \delta)$ is strictly decreasing in $|x_i| > C_1 = \sqrt{\frac{2}{\eta(1-\delta)}}$, for any $\rho > C_1$,

$$\sup_{|x_i|>\sqrt{4a\rho^2 \log n}} \tilde{h}_2(x_i, \frac{1}{n}, \eta, \delta) \leq 4a\rho^2 \frac{H(a,\eta,\delta)}{\Delta(\frac{1}{n^2}, \eta, \delta)} \log n \cdot n^{-2a(\frac{2\rho^2}{C_1^2}-1)} .$$

Using the definition of the slowly varying function $L(\cdot)$ and assumption (A1) in the right-hand side of above, we have,

$$\sup_{|x_i|>\sqrt{4a\rho^2 \log n}} \tilde{h}_2(x_i, \frac{1}{n}, \eta, \delta) = o(1) \text{ as } n \to \infty.$$

Therefore, we have,

$$\sup_{|x_i|>\sqrt{4a\rho^2 \log n}} \tilde{h}(x_i, \frac{1}{n}, \eta, \delta) \leq 1 + \sup_{|x_i|>\sqrt{4a\rho^2 \log n}} \tilde{h}_1(x_i, \frac{1}{n}) + \sup_{|x_i|>\sqrt{4a\rho^2 \log n}} \tilde{h}_2(x_i, \frac{1}{n}, \eta, \delta) \lesssim 1 \cdot$$

Using above arguments, as $n \to \infty$,

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{|X_i|>\sqrt{4a\rho^2 \log n}\}}] \lesssim 1 . \tag{2.30}$$

Combining (2.28),(2.29) and (2.30), we obtain

$$\mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i, \widehat{\tau}) \lesssim \log n . \tag{2.31}$$

Noting that $\lesssim$ relation proved here actually holds uniformly in i's such that $\theta_{0i} \neq 0$, in that the corresponding constants in the upper bounds can be chosen to be the same for each i, we have

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i, \widehat{\tau}) \lesssim \tilde{q}_n \log n . \tag{2.32}$$

Proof of Step-2: Fix any i such that $\theta_{0i} = 0$. Define $v_n = \sqrt{c_1 \log n}$, where c_1 is defined in (2.4).

Case-1 First we consider the case when $a \in [\frac{1}{2}, 1)$. Again, we split $\mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i, \widehat{\tau})$ as

$$\mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i, \widehat{\tau}) = \mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{|X_i|>v_n\}}] + \mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{|X_i|\leq v_n\}}] \cdot \tag{2.33}$$

Using $Var(\theta_i|x_i, \widehat{\tau}) \leq 1 + x_i^2$ and the identity $x^2\phi(x) = \phi(x) - \frac{d}{dx}[x\phi(x)]$, we obtain,

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{|X_i|>v_n\}}] \leq 2 \int_{\sqrt{c_1 \log n}}^{\infty} (1 + x^2)\phi(x)dx \lesssim \sqrt{\log n} \cdot n^{-\frac{c_1}{2}} . \tag{2.34}$$

Now let us choose some $\gamma > 1$ such that $c_2\gamma - 1 > 1$. Next, we decompose the second term as follows:

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{|X_i|\leq v_n\}}] = \mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{\widehat{\tau}>\gamma\frac{q_n}{n}\}} 1_{\{|X_i|\leq v_n\}}] +$$
$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i, \widehat{\tau}) 1_{\{\widehat{\tau}\leq\gamma\frac{q_n}{n}\}} 1_{\{|X_i|\leq v_n\}}] . \tag{2.35}$$

Note that

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}>\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq v_n\}}] \leq (1+c_1\log n)\mathbb{P}_{\theta_0}[\widehat{\tau}>\gamma\frac{q_n}{n},|X_i|\leq v_n]$$

$$\leq (1+c_1\log n)\mathbb{P}_{\theta_0}[\frac{1}{c_2 n}\sum_{j=1(\neq i)}^{n}1_{\{|X_j|>\sqrt{c_1\log n}\}}>\gamma\frac{q_n}{n}] \lesssim \frac{q_n}{n}\log n\,. \tag{2.36}$$

Inequality in the last line is due to employing similar arguments used for proving Lemma A.7 in van der Pas et al. (2014).

We will now bound $\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}\leq\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq\sqrt{c_1\log n}\}}]$. Since for any fixed $x_i\in\mathbb{R}$ and $\tau>0$, $Var(\theta_i|x_i,\tau)\leq E(1-\kappa_i|x_i,\tau)+J(x_i,\tau)$ where $J(x_i,\tau)=x_i^2E[(1-\kappa_i)^2|x_i,\tau]$. Since $E(1-\kappa_i|x_i,\tau)$ is non-decreasing in τ, so, $E(1-\kappa_i|x_i,\widehat{\tau})\leq E(1-\kappa_i|x_i,\gamma\frac{q_n}{n})$ whenever $\widehat{\tau}\leq\gamma\frac{q_n}{n}$. Using Lemma 2 and A.2 of Ghosh and Chakrabarti (2017),

$$\begin{aligned}\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}\leq\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq\sqrt{c_1\log n}\}}] &\lesssim (\frac{q_n}{n})^{2a}\int_0^{\sqrt{c_1\log n}}e^{\frac{x^2}{2}}e^{-\frac{x^2}{2}}dx\\ &= (\frac{q_n}{n})^{2a}\sqrt{c_1\log n}\,.\end{aligned} \tag{2.37}$$

Note that all these preceding arguments hold uniformly in i such that $\theta_{0i}=0$. Combining all these results, for $a\in[\frac{1}{2},1)$ using (2.33)-(2.37), we have,

$$\begin{aligned}\sum_{i:\theta_{0i}=0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau}) &\lesssim (n-\tilde{q}_n)[\sqrt{\log n}\cdot n^{-\frac{c_1}{2}}+\frac{q_n}{n}\cdot\log n+(\frac{q_n}{n})^{2a}\sqrt{\log n}]\\ &\lesssim q_n\log n\,.\end{aligned} \tag{2.38}$$

The second inequality follows due to the fact that $\tilde{q}_n\leq q_n$ and $q_n=o(n)$ as $n\to\infty$.

Case-2 Now we assume $a\geq 1$ and split $\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})$ as

$$\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})=\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|>v_n\}}]+\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|\leq v_n\}}], \tag{2.39}$$

where v_n is the same as defined in **Case-1**. Using exactly the same arguments used for the case $a\in[\frac{1}{2},1)$, we have the following for $a\geq 1$,

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{|X_i|>\sqrt{c_1\log n}\}}]\lesssim\sqrt{\log n}\cdot n^{-\frac{c_1}{2}}\,. \tag{2.40}$$

and

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}>\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq\sqrt{c_1\log n}\}}]\lesssim\frac{q_n}{n}\log n\,. \tag{2.41}$$

For the part $\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}\leq\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq\sqrt{c_1\log n}\}}]$, note that for fixed $x_i\in\mathbb{R}$ and any $\tau>0$,

$$\begin{aligned}Var(\theta_i|x_i,\tau) &\leq E(1-\kappa_i|x_i,\tau)+x_i^2E[(1-\kappa_i)^2|x_i,\tau]\\ &\leq E(1-\kappa_i|x_i,\tau)+x_i^2E(1-\kappa_i|x_i,\tau)\\ &\leq E(1-\kappa_i|x_i,\tau)1_{\{|x_i|\leq 1\}}+2x_i^2E(1-\kappa_i|x_i,\tau)\,.\end{aligned} \tag{2.42}$$

Since for fixed $x_i \in \mathbb{R}$ and any $\tau > 0$, $E(1-\kappa_i|x_i,\tau)$ is non-decreasing in τ, we have,

$$\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}\le\gamma\frac{q_n}{n}\}}1_{\{|X_i|\le\sqrt{c_1\log n}\}}] \le \mathbb{E}_{\theta_{0i}}[E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\le 1\}}] + 2\mathbb{E}_{\theta_{0i}}[X_i^2 E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\le\sqrt{c_1\log n}\}}] \,. \tag{2.43}$$

For bounding the first term in the r.h.s. of (2.43), we use Lemma 2.2.3. We note that for any $\tau \in (0,1)$, $\frac{t\tau^2}{1+t\tau^2}\cdot\frac{1}{\sqrt{1+t\tau^2}}t^{-a-1} \le \tau t^{-(a+\frac{1}{2})}$ and that $L(t)$ is bounded. Using the fact that the second term in the upper bound in Lemma 2.2.3 can be bounded

$$A_2 \le \frac{2M}{(2a-1)}\tau e^{\frac{x_i^2}{2}} \,,$$

where A_2 has been defined in the proof of Lemma 2.2.3. So, we have

$$\mathbb{E}_{\theta_{0i}}[E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\le 1\}}] \lesssim \frac{q_n}{n}\int_0^1 e^{-\frac{x^2}{4}}dx + \frac{q_n}{n} \,.$$

Hence,

$$\mathbb{E}_{\theta_{0i}}[E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\le 1\}}] \lesssim \frac{q_n}{n} \,. \tag{2.44}$$

For the second term in the r.h.s. of (2.43) we shall use the upper bound of $E(1-\kappa_i|x_i,\tau)$ of the form (2.6) and hence,

$$\mathbb{E}_{\theta_{0i}}[X_i^2 E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\le\sqrt{c_1\log n}\}}] \lesssim \frac{q_n}{n}\int_0^{\sqrt{c_1\log n}} x^2 e^{\frac{x^2}{4}}\phi(x)dx$$

$$+\int_0^{\sqrt{c_1\log n}}\int_1^\infty \frac{t(\gamma\frac{q_n}{n})^2}{1+t(\gamma\frac{q_n}{n})^2}\cdot\frac{1}{\sqrt{1+t(\gamma\frac{q_n}{n})^2}}t^{-a-1}L(t)e^{\frac{x^2}{2}\cdot\frac{t(\gamma\frac{q_n}{n})^2}{1+t(\gamma\frac{q_n}{n})^2}}x^2\phi(x)dtdx \,. \tag{2.45}$$

Note that the first integral is bounded by a constant. Using Fubini's theorem and the transformation $y = \frac{x}{\sqrt{1+t(\gamma\frac{q_n}{n})^2}}$, the second integral becomes

$$\frac{1}{\sqrt{2\pi}}\int_1^\infty t^{-a-1}L(t)t(\gamma\frac{q_n}{n})^2\Big(\int_0^{\sqrt{\frac{c_1\log n}{1+t(\gamma\frac{q_n}{n})^2}}} y^2 e^{-\frac{y^2}{2}}dy\Big)dt \,.$$

We handle the above integral separately for $a=1$ and $a>1$. For $a>1$, using the boundedness of $L(t)$ it is easy to show that,

$$\frac{1}{\sqrt{2\pi}}\int_1^\infty t^{-a-1}L(t)t(\gamma\frac{q_n}{n})^2\Big(\int_0^{\sqrt{\frac{c_1\log n}{1+t(\gamma\frac{q_n}{n})^2}}} y^2 e^{-\frac{y^2}{2}}dy\Big)dt \lesssim \frac{q_n}{n} \,. \tag{2.46}$$

For $a=1$ note that,

$$\int_1^\infty t^{-a-1}L(t)t(\gamma\frac{q_n}{n})^2\Big(\int_0^{\sqrt{\frac{c_1\log n}{1+t(\gamma\frac{q_n}{n})^2}}} y^2 e^{-\frac{y^2}{2}}dy\Big)dt$$

$$\leq (\gamma\frac{q_n}{n})^2\sqrt{2\pi}\int_1^{\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}}\frac{1}{t}L(t)dt+(\sqrt{c_1\log n})^3(\gamma\frac{q_n}{n})^2\int_{\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}}^{\infty}\frac{L(t)}{t}\cdot\frac{1}{(t(\gamma\frac{q_n}{n})^2)^{\frac{3}{2}}}dt\,. \tag{2.47}$$

Here the division in the range of t in (2.47) occurs due to the fact the integral $(\int_0^{\sqrt{\frac{c_1\log n}{t(\gamma\frac{q_n}{n})^2}}}y^2e^{-\frac{y^2}{2}}dy)$ can be bounded by $(\frac{c_1\log n}{t(\gamma\frac{q_n}{n})^2})^{\frac{3}{2}}$ when $t\geq\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}$ and by $\sqrt{2\pi}$ when $t\leq\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}$. For the first term in (2.47) with the boundedness of $L(t)$,

$$(\gamma\frac{q_n}{n})^2\int_1^{\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}}\frac{1}{t}L(t)dt\leq(\gamma\frac{q_n}{n})^2M\log\left(\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}\right).$$

Hence for sufficiently large n with $q_n\propto n^{\beta},0<\beta<1$,

$$(\gamma\frac{q_n}{n})^2\int_1^{\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}}\frac{1}{t}L(t)dt\lesssim\frac{q_n}{n}\sqrt{\log n}\,. \tag{2.48}$$

Now for the second term in (2.47) again using the boundedness of $L(t)$,

$$(\sqrt{c_1\log n})^3(\gamma\frac{q_n}{n})^2\int_{\frac{c_1\log n}{(\gamma\frac{q_n}{n})^2(2\pi)^{\frac{1}{3}}}}^{\infty}t\cdot t^{-2}L(t)\cdot\frac{1}{(t(\gamma\frac{q_n}{n})^2)^{\frac{3}{2}}}dt\lesssim\frac{q_n}{n}\,. \tag{2.49}$$

So, using (2.47)-(2.49), for $a=1$,

$$\int_1^{\infty}t^{-a-1}L(t)t(\gamma\frac{q_n}{n})^2\left(\int_0^{\sqrt{\frac{c_1\log n}{1+t(\gamma\frac{q_n}{n})^2}}}y^2e^{-\frac{y^2}{2}}dy\right)dt\lesssim\frac{q_n}{n}\sqrt{\log n}\,. \tag{2.50}$$

With the help of (2.46) and (2.50), we have, for $a\geq1$

$$\int_1^{\infty}t^{-a-1}L(t)t(\gamma\frac{q_n}{n})^2\left(\int_0^{\sqrt{\frac{c_1\log n}{1+t(\gamma\frac{q_n}{n})^2}}}y^2e^{-\frac{y^2}{2}}dy\right)dt\lesssim\frac{q_n}{n}\sqrt{\log n}\,. \tag{2.51}$$

Combining these facts, we finally have

$$\mathbb{E}_{\theta_{0i}}[X_i^2E(1-\kappa_i|X_i,\gamma\frac{q_n}{n})1_{\{|X_i|\leq\sqrt{c_1\log n}\}}]\lesssim\frac{q_n}{n}\sqrt{\log n}\,. \tag{2.52}$$

Note that all these preceding arguments hold uniformly in i such that $\theta_{0i}=0$. Combining all these results, for $a\geq1$, using (2.39)-(2.52), we have,

$$\begin{aligned}\sum_{i:\theta_{0i}=0}\mathbb{E}_{\theta_{0i}}Var(\theta_i|X_i,\widehat{\tau})&\lesssim(n-\tilde{q}_n)[\sqrt{\log n}\cdot n^{-\frac{c_1}{2}}+\frac{q_n}{n}\cdot\log n+\frac{q_n}{n}\sqrt{\log n}]\\&\lesssim q_n\log n\,.\end{aligned} \tag{2.53}$$

Using (2.38) and (2.53), for $a \geq \frac{1}{2}$, we have,

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}) \lesssim q_n \log n \,. \tag{2.54}$$

Now using (2.27), (2.32) and (2.54), for sufficiently large n,

$$\mathbb{E}_{\boldsymbol{\theta}_0} \sum_{i=1}^{n} Var(\theta_i|X_i,\widehat{\tau}) \lesssim q_n \log n \,.$$

Finally, taking supremum over all $\boldsymbol{\theta}_0 \in l_0[q_n]$, the result is obtained.

□

Remark 2.5.1. *Note that, we have used different bounds on $Var(\theta_i|X_i,\tau)$ for two different ranges of a. When $a \in [\frac{1}{2},1)$, we have used the upper bound of $J(X_i,\tau)$ provided by Ghosh and Chakrabarti (2017). However using the same arguments when $a \geq 1$ yields an upper bound on $Var(\theta_i|X_i,\tau)$ such that $\mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau})1_{\{\widehat{\tau}\leq\gamma\frac{q_n}{n}\}}1_{\{|X_i|\leq\sqrt{c_1\log n}\}}]$ exceeds near minimax rate. As a result, we need to come up with a sharper upper bound, and hence (2.42) and Lemma 2.2.3 come in very handy.*

2.5.2 Lemmas related to MMLE

The logarithm of the marginal likelihood function $M_\tau(\mathbf{X})$ is of the form

$$M_\tau(\mathbf{X}) = \sum_{i=1}^{n} \log\Big(\phi(x_i-\theta_i)g_\tau(\theta_i)d\theta_i\Big),$$

where $g_\tau(\theta_i)$ is defined before.

Lemma 2.5.2. *The derivative of the log-likelihood function is of the form*

$$\frac{d}{d\tau}M_\tau(\mathbf{X}) = \frac{1}{\tau}\sum_{i=1}^{n} m_\tau(x_i), \tag{2.55}$$

where

$$m_\tau(x) = x^2\frac{J_{\alpha+1,a}(x)-J_{\alpha+2,a}(x)}{I_{\alpha,a}(x)} - \frac{J_{\alpha+1,a}(x)}{I_{\alpha,a}(x)}, \tag{2.56}$$

and $I_{\alpha,a}(x)$ and for $k=1,2$, $J_{\alpha+k,a}(x)$ are defined as follows

$$I_{\alpha,a}(x) = \int_0^1 e^{\frac{x^2z}{2}} z^{\alpha-1}(1-z)^{a-\frac{1}{2}}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{a+\alpha} dz \tag{2.57}$$

and

$$J_{\alpha+k,a}(x) = \int_0^1 e^{\frac{x^2z}{2}} z^{\alpha+k-1}(1-z)^{a-\frac{1}{2}}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{a+\alpha} dz. \tag{2.58}$$

Proof. Since, $X_i|\theta_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1)$ and $\theta_i|\lambda_i, \tau \overset{ind}{\sim} \mathcal{N}(0, \lambda_i^2\tau^2)$, the marginal density of X_i given τ is of this form

$$\begin{aligned}\psi_\tau(x) &= K\int_0^\infty \frac{e^{-\frac{1}{2}\frac{x^2}{(1+\lambda^2\tau^2)}}}{\sqrt{1+\lambda^2\tau^2}\sqrt{2\pi}}(\lambda^2)^{-a-1}L(\lambda^2)d\lambda^2 \\ &= K\frac{\tau^{2a}}{\sqrt{2\pi}}\int_0^1 e^{-\frac{x^2(1-z)}{2}}z^{-a-1}(1-z)^{a-\frac{1}{2}}L(\frac{1}{\tau^2}\frac{z}{1-z})dz,\end{aligned}$$

where equality in the second line follows due to the substitution $1-z = (1+\lambda^2\tau^2)^{-1}$. Next, noting that $L(t) = (1+\frac{1}{t})^{-(a+\alpha)}$,

$$\begin{aligned}\psi_\tau(x) &= K\tau^{2a}\int_0^1 \frac{e^{-\frac{x^2(1-z)}{2}}}{\sqrt{2\pi}}z^{\alpha-1}(1-z)^{a-\frac{1}{2}}[\frac{1}{\tau^2+(1-\tau^2)z}]^{a+\alpha}dz \\ &= K\tau^{2a}I_{\alpha,a}(x)\phi(x),\end{aligned} \tag{2.59}$$

where $\phi(x)$ denotes the standard normal density. Using (2.59),

$$\begin{aligned}\frac{\dot{\psi}_\tau}{\psi_\tau} &= \frac{2aI_{\alpha,a}(x)+\tau\dot{I}_{\alpha,a}(x)}{\tau I_{\alpha,a}(x)} \\ &= \frac{\int_0^1 e^{\frac{x^2z}{2}}z^{\alpha-1}(1-z)^{a-\frac{1}{2}}(\frac{1}{N(z)})^{a+\alpha+1}[2aN(z)-2\tau^2(\alpha+a)(1-z)]dz}{\tau I_{\alpha,a}(x)},\end{aligned} \tag{2.60}$$

where $N(z) = \tau^2 + (1-\tau^2)z$. On the other hand, using integration by parts,

$$\begin{aligned}&x^2(J_{\alpha+1,a}(x) - J_{\alpha+2,a}(x)) = \int_0^1 x^2e^{\frac{x^2z}{2}}z^\alpha(1-z)^{a+\frac{1}{2}}(\frac{1}{N(z)})^{a+\alpha}dz \\ &= \int_0^1 e^{\frac{x^2z}{2}}\frac{z^{\alpha-1}(1-z)^{a-\frac{1}{2}}}{(N(z))^{a+\alpha+1}}[N(z)\{-2\alpha(1-z)+2(a+\frac{1}{2})z\}+2z(1-z)(1-\tau^2)(\alpha+a)]dz\end{aligned}$$

As a consequence of this,

$$\begin{aligned}&x^2(J_{\alpha+1,a}(x) - J_{\alpha+2,a}(x)) - J_{\alpha+1,a}(x) \\ &= \int_0^1 e^{\frac{x^2z}{2}}\frac{z^{\alpha-1}(1-z)^{a-\frac{1}{2}}}{(N(z))^{a+\alpha+1}}[N(z)\{-2\alpha(1-z)+2az\}+2(\alpha+a)(1-z)(N(z)-\tau^2)]dz.\end{aligned} \tag{2.61}$$

On Simplification the r.h.s. of (2.61) matches with the numerator of (2.60) and completes the proof. □

Lemma 2.5.3. *Let κ_τ be the solution to the equation $\frac{e^{x^2/2}}{x^2/2} = \frac{1}{\tau}$ where $0 < \tau < 1$. Choose any $B \geq 1$. There exist functions R_τ with $\sup_x |R_\tau(x)| = O(\tau^{\frac{1}{2}})$ as $\tau \to 0$, such that, for $\alpha \geq \frac{1}{2}$,*

$$I_{\alpha,\frac{1}{2}}(x) = \begin{cases}\left(\frac{K^{-1}}{\tau} + \left(\frac{x^2}{2}\right)^{\frac{1}{2}}\int_1^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}}dv\right)(1+R_\tau(x)), \text{ uniformly in } |x| \leq B\kappa_\tau, \\ \left(\frac{x^2}{2}\right)^{\frac{1}{2}}\int_1^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}}dv(1+R_\tau(x)), \text{ uniformly in } |x| > B\kappa_\tau.\end{cases} \tag{2.62}$$

Further, given $\epsilon_\tau \to 0$, *there exist functions* S_τ *with* $\sup_{x \geq 1/\epsilon_\tau} |S_\tau(x)| = O(\tau^{\frac{1}{2}} + \epsilon_\tau^2)$, *such that as* $\tau \to 0$,

$$I_{\alpha,\frac{1}{2}}(x) = \frac{e^{x^2/2}}{x^2/2}(1 + S_\tau(x)), \tag{2.63}$$

where $K = \frac{\Gamma(\frac{1}{2}+\alpha)}{\sqrt{\pi}\Gamma(\alpha)}$.

Proof. We consider the cases separately when $|x| \leq B\kappa_\tau$ and $|x| > B\kappa_\tau$. Note, as observed in van der Pas et al. (2017a), $\kappa_\tau \sim \zeta_\tau + \frac{2\log\zeta_\tau}{\zeta_\tau}$ as $\tau \to 0$ where $\zeta_\tau = \sqrt{2\log(\frac{1}{\tau})}$.
Case 1:- When $|x| \leq B\kappa_\tau$, the range of the integration in $I_{\alpha,\frac{1}{2}}(x)$ is divided into three parts, namely, $I_1 = \int_0^\tau e^{\frac{x^2 z}{2}} z^{\alpha-1}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz$, $I_2 = \int_\tau^{(\frac{2}{x^2})\wedge 1} e^{\frac{x^2 z}{2}} z^{\alpha-1}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz$ and $I_3 = \int_{(\frac{2}{x^2})\wedge 1}^1 e^{\frac{x^2 z}{2}} z^{\alpha-1}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz$, where $y_1 \wedge y_2$ denotes the minimum of y_1 and y_2. Next, making the substitution $z = u\tau^2$ in I_1, we have

$$I_1 = \tau^{-1}\int_0^{\frac{1}{\tau}} u^{\alpha-1}(1 + u(1-\tau^2))^{-(\frac{1}{2}+\alpha)} e^{\frac{x^2\tau^2 u}{2}} du. \tag{2.64}$$

Next, define $I_1^* = \tau^{-1}\int_0^{\frac{1}{\tau}} u^{\alpha-1}(1 + u(1-\tau^2))^{-(\frac{1}{2}+\alpha)} du$. Our target is to show that $\frac{I_1 - I_1^*}{I_1^*} \to 0$ as $\tau \to 0$. Now following the argument same as that used in Lemma C.9 of van der Pas et al. (2017a), for $|x| \leq B\kappa_\tau$, the exponent in the integral tends to 1, uniformly in $u \leq \frac{1}{\tau}$. Since, for any $y \geq 0$, $e^y - 1 \leq ye^y$,

$$\begin{aligned} I_1 - I_1^* &= \tau^{-1}\int_0^{\frac{1}{\tau}} u^{\alpha-1}(1 + u(1-\tau^2))^{-(\frac{1}{2}+\alpha)}[e^{\frac{x^2\tau^2 u}{2}} - 1] du \\ &\lesssim x^2\tau e^{\frac{x^2\tau}{2}} \int_0^{\frac{1}{\tau}} u^{-\frac{1}{2}} du \lesssim \tau^{\frac{1}{2}} \log\left(\frac{1}{\tau}\right)(1 + o(1)). \end{aligned}$$

Now we want to find the asymptotic order of I_1^*. Towards this, observe that, replacing $\frac{1}{1+(1-\tau^2)u}$ by $\frac{1}{(1+u)(1-\tau^2)}$ provides a multiplicative error of the order $1+O(\tau^2)$, i.e., $\frac{1}{1+(1-\tau^2)u} = \frac{1}{(1+u)(1-\tau^2)}(1+O(\tau^2))$. Since,

$$\int_0^\infty u^{\alpha-1}(1+u)^{-(\alpha+\frac{1}{2})} du = \int_0^\infty u^{-\frac{3}{2}}(1+\frac{1}{u})^{-\frac{1}{2}-\alpha} du = K^{-1} \tag{2.65}$$

and

$$\int_{\frac{1}{\tau}}^\infty u^{\alpha-1}(1+u)^{-(\alpha+\frac{1}{2})} du \lesssim \tau^{\frac{1}{2}}, \tag{2.66}$$

we have as $\tau \to 0$, $I_1^* = \frac{K^{-1}}{\tau}[1 + O(\tau^{\frac{1}{2}})]$, uniformly in $|x| \leq B\kappa_\tau$. This implies, $0 < \frac{I_1 - I_1^*}{I_1^*} \lesssim \tau^{\frac{3}{2}}\log(\frac{1}{\tau})(1 + o(1))$, and hence $\frac{I_1 - I_1^*}{I_1^*} \to 0$ as $\tau \to 0$. Combining all these arguments along with (2.65) and (2.66), we obtain

$$I_1 = \frac{K^{-1}}{\tau}[1 + O(\tau^{\frac{1}{2}})], \tag{2.67}$$

uniformly in $|x| \leq B\kappa_\tau$. Moving towards the second integral, first, we make a transformation

$\frac{x^2 z}{2} = v$ and hence

$$I_2 = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_{\frac{x^2\tau}{2}}^{\frac{x^2}{2}\wedge 1} e^v v^{\alpha-1}\left(\frac{\tau^2 x^2}{2} + (1-\tau^2)v\right)^{-(\frac{1}{2}+\alpha)} dv.$$

Now, we bound $\frac{\tau^2 x^2}{2} + (1-\tau^2)v$ below by $(1-\tau^2)v$ and observe that the upper limit of the range of integration can be bounded by 1 irrespective of whether $\frac{x^2}{2} \leq 1$ or not. Hence, we can show that $I_2 \lesssim \tau^{-\frac{1}{2}}$, and this contributes negligibly compared to I_1. Finally, for I_3, again after the same transformation

$$I_3 = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_{1}^{\frac{x^2}{2}} e^v v^{\alpha-1}\left(\frac{\tau^2 x^2}{2} + (1-\tau^2)v\right)^{-(\frac{1}{2}+\alpha)} dv.$$

Here observe that, the integral contributes nothing when $\frac{x^2}{2} \leq 1$, hence we are only interested when $\frac{x^2}{2} > 1$. Next, we define

$$I_3^* = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_{1}^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}} dv.$$

Now our target is to show that the difference between I_3 and I_3^* is negligible compared to I_3^* as $\tau \to 0$. In order to prove that, first note,

$$\left(\frac{1}{\frac{\tau^2 x^2}{2} + (1-\tau^2)v}\right)^{\frac{1}{2}+\alpha} \leq \left(\frac{1}{v}\right)^{\frac{1}{2}+\alpha}\left[1 + \frac{\tau^2(v+x^2)}{v(1-\tau^2)}\right]^{\frac{1}{2}+\alpha}. \tag{2.68}$$

Now, first, consider the case when $\alpha + \frac{1}{2}$ is a positive integer. Hence using the Binomial theorem, we have

$$\left(\frac{1}{\frac{\tau^2 x^2}{2} + (1-\tau^2)v}\right)^{\frac{1}{2}+\alpha} - v^{-(\frac{1}{2}+\alpha)} \leq v^{-(\frac{1}{2}+\alpha)} \sum_{j=1}^{\frac{1}{2}+\alpha} \binom{\frac{1}{2}+\alpha}{j} \left[\frac{\tau^2}{(1-\tau^2)}\left(1 + \frac{x^2}{v}\right)\right]^j$$

Next, observing that for $1 \leq v \leq \frac{x^2}{2}, 2 \leq \frac{x^2}{v} \leq x^2$, for $|x| \leq B\kappa_\tau$, the difference between I_3 and I_3^* can be bounded as

$$I_3 - I_3^* \lesssim \left(\frac{\tau^2}{1-\tau^2}\right) \log\left(\frac{1}{\tau}\right)(1+o(1))\left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_{1}^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}} dv,$$

which implies

$$I_3 - I_3^* \lesssim \left(\frac{\tau^2}{1-\tau^2}\right) \log\left(\frac{1}{\tau}\right)(1+o(1)) I_3^*.$$

On the other hand, observe that,

$$I_3 - I_3^* = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_1^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}} \left[(\frac{\tau^2 x^2}{2} + 1 - \tau^2)^{-(\alpha+\frac{1}{2})} - 1 \right] dv$$
$$\gtrsim \left[(\frac{\tau^2 \zeta_\tau^2}{2}(1+o(1)) + 1 - \tau^2)^{-(\alpha+\frac{1}{2})} - 1 \right] I_3^*.$$

These two bounds ensure $\frac{I_3 - I_3^*}{I_3^*} \to 0$ as $\tau \to 0$ since $I_3^* \sim \frac{e^{x^2/2}}{x^2/2}$ by Lemma C.8 of van der Pas et al. (2017a). Next, consider the case when $\alpha + \frac{1}{2}$ is a fraction. When $\alpha + \frac{1}{2}$ is a fraction, then there exists another fraction $b > 0$ such that $\alpha + \frac{1}{2} + b$ is a positive integer. Hence, in this case, $\left[1 + \frac{\tau^2(v+x^2)}{v(1-\tau^2)}\right]^{\frac{1}{2}+\alpha} \leq \left[1 + \frac{\tau^2(v+x^2)}{v(1-\tau^2)}\right]^{\alpha+\frac{1}{2}+b}$. Now, applying exactly the same set of arguments on $\alpha + \frac{1}{2} + b$ in place of $\alpha + \frac{1}{2}$, we again can show that, $\frac{I_3 - I_3^*}{I_3^*} \to 0$ as $\tau \to 0$. This completes the proof for $|x| \leq B\kappa_\tau$.

Case 2:- When $|x| > B\kappa_\tau$, choose any $A \in (0,1)$. In this case, the range of the integration in $I_{\alpha,\frac{1}{2}}(x)$ is divided into two parts, namely, $I_4 = \int_0^A e^{\frac{x^2 z}{2}} z^{\alpha-1} \left(\frac{1}{\tau^2 + (1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz$ and $I_5 = \int_A^1 e^{\frac{x^2 z}{2}} z^{\alpha-1} \left(\frac{1}{\tau^2 + (1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz$. Note that

$$I_4 \leq e^{\frac{x^2 A}{2}} \left(\frac{1}{\tau^2}\right)^{\frac{1}{2}+\alpha} \int_0^A z^{\alpha-1} dz \lesssim \tau^{-1-2\alpha} e^{\frac{x^2 A}{2}}.$$

One can choose $B \geq 1$ and $A \in (0,1)$ such that $B^2(1-A) > 2\alpha + \frac{3}{2}$, which implies $I_4 \ll \tau^{\frac{1}{2}} \frac{e^{x^2/2}}{x^2/2}$, and hence the contribution is negligible compared to the second term in the expression of $I_{\alpha,\frac{1}{2}}(x)$ as given in (2.62). Finally, for I_5, we use that, for $z \geq A$, $\frac{1}{\tau^2 + (1-\tau^2)z} = \frac{1}{z}[1 + O(\tau^2)]$. This implies I_5 is of the form

$$I_5 = \int_A^1 z^{-\frac{3}{2}} e^{\frac{x^2 z}{2}} dz [1 + O(\tau^2)]^{\frac{1}{2}+\alpha}. \tag{2.69}$$

Next, note that, with the transformation $\frac{x^2 z}{2} = v$,

$$\int_A^1 z^{-\frac{3}{2}} e^{\frac{x^2 z}{2}} dz = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_{\frac{x^2 A}{2}}^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}} dv = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \left[\int_1^{\frac{x^2}{2}} - \int_1^{\frac{x^2 A}{2}} \right] e^v v^{-\frac{3}{2}} dv. \tag{2.70}$$

Now, using the first assertion of Lemma C.8 of van der Pas et al. (2017a), the second integral is bounded above by a multiple of $(x^2/2)^{-1} e^{x^2 A/2}$, which is negligible compared to the first (this is of the order of $(x^2/2)^{-1} e^{x^2/2}$). Hence, combining (2.69) and (2.70), we immediately have

$$I_5 = \left(\frac{x^2}{2}\right)^{\frac{1}{2}} \int_1^{\frac{x^2}{2}} e^v v^{-\frac{3}{2}} dv \big[1 + O(\tau^2)\big]. \tag{2.71}$$

Combining (2.67) and (2.71), the r.h.s. of (2.62) is established.

On the other hand, expanding the integral given in (2.62) with the help of Lemma C.8 of van der

Pas et al. (2017a) provides (2.63). □

Lemma 2.5.4. *There exist functions $R_{\tau,1}$ with $\sup_x |R_{\tau,1}(x)| = O(\tau^{\frac{1}{2}})$ as $\tau \to 0$, such that for $\alpha \geq \frac{1}{2}$,*

$$J_{\alpha+1,\frac{1}{2}}(x) = \left(\frac{x^2}{2}\right)^{-\frac{1}{2}} \int_0^{\frac{x^2}{2}} e^v v^{-\frac{1}{2}} dv(1 + R_{\tau,1}(x)) \lesssim (1 \wedge x^{-2}) e^{\frac{x^2}{2}} \tag{2.72}$$

and

$$J_{\alpha+1,\frac{1}{2}}(x) - J_{\alpha+2,\frac{1}{2}}(x) = \left(\frac{x^2}{2}\right)^{-\frac{1}{2}} \int_0^{\frac{x^2}{2}} e^v v^{-\frac{1}{2}} \left(1 - \frac{2v}{x^2}\right) dv(1 + R_{\tau,1}(x)) \lesssim (1 \wedge x^{-4}) e^{\frac{x^2}{2}}. \tag{2.73}$$

Proof. Recall that $J_{\alpha+1,\frac{1}{2}}(x)$ is defined as

$$J_{\alpha+1,\frac{1}{2}}(x) = \int_0^1 e^{\frac{x^2 z}{2}} z^\alpha \left(\frac{1}{\tau^2 + (1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz. \tag{2.74}$$

Next, we split the range of the integration into $[0, \tau]$ and $[\tau, 1]$. Note that the contribution of the first integral obtained from (2.74) is bounded by

$$\int_0^\tau e^{\frac{x^2 z}{2}} z^\alpha \left(\frac{1}{\tau^2 + (1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz \lesssim e^{\frac{x^2 \tau}{2}} \tau^{\frac{1}{2}}. \tag{2.75}$$

On the other hand, note that, when $z \geq \tau$, $\frac{1}{\tau^2+(1-\tau^2)z} = \frac{1}{z}[1 + O(\tau)]$. Therefore, we have

$$\begin{aligned} \int_\tau^1 e^{\frac{x^2 z}{2}} z^\alpha \left(\frac{1}{\tau^2 + (1-\tau^2)z}\right)^{\frac{1}{2}+\alpha} dz &= \int_\tau^1 e^{\frac{x^2 z}{2}} z^{-\frac{1}{2}} dz [1 + O(\tau)]^{\frac{1}{2}+\alpha} \\ &\gtrsim \exp\left(\frac{x^2}{2}\tau\right)(1 - \tau^{\frac{1}{2}})[1 + O(\tau)]. \end{aligned} \tag{2.76}$$

Combining (2.75) and (2.76) we see that the contribution of the first integral is negligible compared to the second one. Hence, one has

$$J_{\alpha+1,\frac{1}{2}}(x) = \int_\tau^1 e^{\frac{x^2 z}{2}} z^{-\frac{1}{2}} dz(1 + O(\tau) + O(\tau^{\frac{1}{2}})).$$

Also note that, $\tau = O(\tau^{\frac{1}{2}})$ as $\tau \to 0$. Observe that, (2.75) and (2.76) also imply that,

$$\int_\tau^1 e^{\frac{x^2 z}{2}} z^{-\frac{1}{2}} dz = \int_0^1 e^{\frac{x^2 z}{2}} z^{-\frac{1}{2}} dz[1 + O(\tau^{\frac{1}{2}})] \to 0 \text{ as } \tau \to 0.$$

As a result, $J_{\alpha+1,\frac{1}{2}}(x) = \int_0^1 e^{\frac{x^2 z}{2}} z^{-\frac{1}{2}} dz(1 + O(\tau^{\frac{1}{2}}))$. The equality in (2.72) follows due to the change of variable $\frac{x^2 z}{2} = v$. The second assertion in (2.72) is due to exactly the same set of arguments used in Lemma C.10 of van der Pas et al. (2017a).
Proof of (2.73) follows using a similar set of arguments. □

Lemma 2.5.5. *The function $x \mapsto m_\tau(x)$ is symmetric about 0 and non-decreasing in $[0, \infty)$ with*

(i) $-2\alpha \le m_\tau(x) \le 2a$, *for all* $x \in \mathbb{R}$ *and* $\tau \in (0,1)$.
(ii) $|m_\tau(x)| \lesssim \tau e^{x^2/2}(x^{-2} \wedge 1)$, *as* $\tau \to 0$, *for every* x *and* $a \in [\frac{1}{2}, 1)$.

Proof. The symmetric behavior follows from the definition of $m_\tau(x)$ as given in (2.56). For monotonicity, using (2.60) of Lemma 2.5.2, it readily follows that

$$\begin{aligned} m_\tau(x) &= 2a + \tau \frac{\dot{I}_{\alpha,a(x)}}{I_{\alpha,a}(x)} \\ &= 2a - 2\tau^2(a+\alpha)\frac{\int_0^1 e^{\frac{x^2 z}{2}} z^{\alpha-1}(1-z)^{a+\frac{1}{2}}(\frac{1}{N(z)})^{a+\alpha+1}dz}{\int_0^1 e^{\frac{x^2 z}{2}} z^{\alpha-1}(1-z)^{a-\frac{1}{2}}(\frac{1}{N(z)})^{a+\alpha}dz} \\ &= 2a + 2\tau^2(a+\alpha)\int_0^1 \left(\frac{z-1}{\tau^2+(1-\tau^2)z}\right) g_x(z)dz, \end{aligned} \tag{2.77}$$

where $z \mapsto g_x(z)$ is the probability density function on $[0,1]$ with $g_x(z) \propto z^{\alpha-1}e^{\frac{x^2 z}{2}}(1-z)^{a-\frac{1}{2}}\big(\frac{1}{N(z)}\big)^{a+\alpha}$. Next, following the same set of arguments as used in Lemma C.7 of van der Pas et al. (2017a).
(i) The upper bound is obvious by using (2.77). For the lower bound, note that

$$\tau \frac{\dot{I}_{\alpha,a(x)}}{I_{\alpha,a}(x)} = -2(\alpha+a)\int_0^1 \left(\frac{(1-z)\tau^2}{\tau^2+(1-\tau^2)z}\right) g_x(z)dz.$$

Since, for any $0 \le z \le 1$, $\tau^2 + (1-\tau^2)z \ge \tau^2$ implies $\frac{(1-z)\tau^2}{\tau^2+(1-\tau^2)z} \le 1$, the lower bound follows from it.
(ii) First, using the definition of $m_\tau(x)$ as given in (2.56) and then followed by the triangle inequality, an upper bound on $|m_\tau(x)|$ is obtained as

$$|m_\tau(x)| \le x^2 \frac{|J_{\alpha+1,a}(x) - J_{\alpha+2,a}(x)|}{I_{\alpha,a}(x)} + \frac{|J_{\alpha+1,a}(x)|}{I_{\alpha,a}(x)}.$$

The assertion is proved by using (2.72) and (2.73) of Lemma 2.5.4, (2.62) of Lemma 2.5.3 and noting that $a \in [\frac{1}{2}, 1)$. □

As an immediate consequence of Lemma 2.5.5, we have the following corollary.

Corollary 2.5.6. *Let* $X \sim \mathcal{N}(\theta, 1)$. *Then, as* $\tau \to 0$,

$$E_\theta m_\tau^2(X) = \begin{cases} o(\tau\zeta_\tau^{-2}), |\theta| \lesssim \zeta_\tau^{-1}, \\ o(\tau^{\frac{1}{16}}\zeta_\tau^{-2}), |\theta| \le \frac{\zeta_\tau}{4}. \end{cases}$$

Proof. Noting that the upper bound of the absolute value of $m_\tau(x)$ as obtained in (ii) of Lemma 2.5.5 matches with that of (vii) of Lemma C.7 of van der Pas et al. (2017a), the proof is immediate using the same set of arguments used in Lemma C.5 of van der Pas et al. (2017a). □

Lemma 2.5.7. *Let $X \sim \mathcal{N}(\theta, 1)$. For $|\theta| \lesssim \zeta_\tau^{-1}$, and $\tau \leq \tau_1 < \tau_2$ and $\tau_2 \to 0$,*

$$E_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1}m_{\tau_1}(X) - \frac{\zeta_{\tau_2}}{\tau_2}m_{\tau_2}(X)\right)^2 \lesssim (\tau_2 - \tau_1)^2\tau_1^{-3}. \tag{2.78}$$

Further, for $|\theta| \leq \frac{\zeta_\tau}{4}$, and $\epsilon = \frac{1}{16}$, and $\tau \leq \tau_1 < \tau_2$ and $\tau_2 \to 0$,

$$E_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1^\epsilon}m_{\tau_1}(X) - \frac{\zeta_{\tau_2}}{\tau_2^\epsilon}m_{\tau_2}(X)\right)^2 \lesssim (\tau_2 - \tau_1)^2\tau_1^{-2-\epsilon}. \tag{2.79}$$

Proof. Using Lemma C.11 of van der Pas et al. (2017a) with $V_\tau = \frac{\zeta_\tau}{\tau}m_\tau(X)$, the l.h.s. of (2.78) can be upper bounded as

$$\begin{aligned} &E_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1}m_{\tau_1}(X) - \frac{\zeta_{\tau_2}}{\tau_2}m_{\tau_2}(X)\right)^2 \leq (\tau_2 - \tau_1)^2 \sup_{\tau\in[\tau_1,\tau_2]} E_\theta\left(\frac{\zeta_\tau}{\tau}\dot{m}_\tau(X) - \frac{\zeta_\tau + \zeta_\tau^{-1}}{\tau^2}m_\tau(X)\right)^2 \\ &\leq 2(\tau_2 - \tau_1)^2\left[\sup_{\tau\in[\tau_1,\tau_2]} E_\theta\left(\frac{\zeta_\tau}{\tau}\dot{m}_\tau(X)\right)^2 + \sup_{\tau\in[\tau_1,\tau_2]} E_\theta\left(\frac{\zeta_\tau + \zeta_\tau^{-1}}{\tau^2}m_\tau(X)\right)^2\right]. \end{aligned} \tag{2.80}$$

Note that, with the help of the first part of Corollary 2.5.6, the second term in the r.h.s. of (2.80) is bounded above by a constant times $\sup_{\tau\in[\tau_1,\tau_2]}\tau^{-3} \lesssim \tau_1^{-3}$. Hence, in order to show that (2.78) holds, it is enough to show that the first term in the r.h.s. of (2.78) is bounded above by a constant times τ_1^{-3}.
For the first term, observe that using (2.56) with $a = \frac{1}{2}$,

$$\dot{m}_\tau(x) = (x^2 - 1)\frac{\dot{J}_{\alpha+1,\frac{1}{2}}(x)}{I_{\alpha,\frac{1}{2}}(x)} - x^2\frac{\dot{J}_{\alpha+2,\frac{1}{2}}(x)}{I_{\alpha,\frac{1}{2}}(x)} - \frac{\dot{I}_{\alpha,\frac{1}{2}}(x)}{I_{\alpha,\frac{1}{2}}(x)}m_\tau(x). \tag{2.81}$$

Now, note that, using the definition of $J_{\alpha+1,\frac{1}{2}}(x)$, $\dot{J}_{\alpha+1,\frac{1}{2}}(x) = 2\tau(\alpha+\frac{1}{2})(H_{\alpha+2,\frac{1}{2}}(x) - H_{\alpha+1,\frac{1}{2}}(x))$ and $\dot{J}_{\alpha+2,\frac{1}{2}}(x) = 2\tau(\alpha+\frac{1}{2})(H_{\alpha+3,\frac{1}{2}}(x) - H_{\alpha+2,\frac{1}{2}}(x))$ where $H_{\alpha+k,\frac{1}{2}}(x) = \int_0^1 e^{\frac{x^2 z}{2}}z^{\alpha+k-1}(1-z)^{a-\frac{1}{2}}\left(\frac{1}{\tau^2+(1-\tau^2)z}\right)^{\alpha+\frac{3}{2}}dz$. Next, note that, $H_{\alpha+k,\frac{1}{2}}(x)$ is a decreasing function of k. Also, observe that, $H_{\alpha+1,\frac{1}{2}}(x) \leq I_{\alpha,\frac{1}{2}}(x)$. Finally, for the third term in the r.h.s. of (2.81), the definition of $I_{\alpha,\frac{1}{2}}(x)$ implies $-\frac{\dot{I}_{\alpha,\frac{1}{2}}(x)}{I_{\alpha,\frac{1}{2}}(x)} \leq 2\tau(\alpha+\frac{1}{2}).\frac{1}{\tau^2}$. Combining all these arguments provides an upper bound for the r.h.s. of (2.81) as

$$\dot{m}_\tau(x) \leq 2\tau(\alpha + \frac{1}{2})[1 + x^2 + \frac{1}{\tau^2}m_\tau(x)].$$

As a consequence of this,

$$E_\theta\dot{m}_\tau^2(X) \lesssim \tau^2[1 + E_\theta X^4 + \frac{1}{\tau^4}E_\theta m_\tau^2(X)]. \tag{2.82}$$

Note that, in this case, $E_\theta X^4$ is bounded by a constant and from the first part of Corollary 2.5.6, $E_\theta m_\tau^2(X)$ is bounded by $\tau\zeta_\tau^{-2}$. This shows that $(\frac{\zeta_\tau}{\tau})^2E_\theta\dot{m}_\tau^2(X)$ is bounded above by a multiple of τ^{-3} and (2.78) is established.

For proving (2.79), we again use Lemma C.11 of van der Pas et al. (2017a), but with $V_\tau = \frac{\zeta_\tau}{\tau^\epsilon} m_\tau(X)$. As a result, we have

$$E_\theta\left(\frac{\zeta_{\tau_1}}{\tau_1^\epsilon} m_{\tau_1}(X) - \frac{\zeta_{\tau_2}}{\tau_2^\epsilon} m_{\tau_2}(X)\right)^2$$
$$\leq 2(\tau_2 - \tau_1)^2 \Bigg[\sup_{\tau\in[\tau_1,\tau_2]} E_\theta\left(\frac{\zeta_\tau}{\tau^\epsilon}\dot{m}_\tau(X)\right)^2 + \sup_{\tau\in[\tau_1,\tau_2]} E_\theta\left(\frac{\epsilon\zeta_\tau + \zeta_\tau^{-1}}{\tau^{1+\epsilon}} m_\tau(X)\right)^2\Bigg].$$

Using the second part of Corollary 2.5.6, the second term in the r.h.s. of the above inequality is bounded above by a constant times $\tau^{-2-\epsilon}$. Next, we follow the same steps as before and obtain an expression same as (2.82). In this case, $E_\theta X^4$ is bounded by ζ_τ^4 and from second part of Corollary 2.5.6, $E_\theta m_\tau^2(X)$ is bounded by $\tau^\epsilon \zeta_\tau^{-2}$. This implies that $(\frac{\zeta_\tau}{\tau^{1+\epsilon}})^2 E_\theta \dot{m}_\tau^2(X)$ is bounded above by a multiple of $\tau^{-2-\epsilon}$ completing the proof of (2.79). □

Next, we provide a corollary which becomes very important to study the relationship between $m_\tau(X_i)$ and its expectation for zero means.

Corollary 2.5.8. *If the cardinality of* $I_0 := \{i : \theta_{0,i} = 0\}$ *tends to infinity, then*

$$\sup_{\frac{1}{n}\leq\tau\leq\frac{1}{\log n}} \frac{1}{|I_0|}\Big|\sum_{i\in I_0} m_\tau(X_i) - \sum_{i\in I_0} E_{\theta_0} m_\tau(X_i)\Big| \xrightarrow{P_{\theta_0}} 0.$$

Proof. The proof of this result follows using a similar set of arguments to those of Lemma C.6 of van der Pas et al. (2017a) with some modifications. The main difference lies in calculating the entropy of the process $G_n(\tau) = |I_0|^{-1}\sum_{i\in I_0} m_\tau(X_i)$. Here, instead of covering the interval $[\frac{1}{n}, 1]$ as given in Lemma C.6 of van der Pas et al. (2017a), we need to cover the interval $[\frac{1}{n}, \frac{1}{\log n}]$. We use dyadic rationals $[\frac{2^i}{n}, \frac{2^{i+1}}{n}]$ to cover this interval with $i = 0, 1, 2, \cdots, [\log_2(\frac{n}{\log n})]$. Next, we follow steps similar to Lemma C.6 of van der Pas et al. (2017a) along with Lemma 2.5.7. □

Lemma 2.5.9. *Let* $X \sim \mathcal{N}(\theta, 1)$. *Then as* $\tau \to 0$

$$E_\theta m_\tau(X) = \begin{cases} -\frac{2}{K^{-1}\sqrt{2\pi}}\frac{\tau}{\zeta_\tau}(1 + o(1)), & |\theta| = o(\zeta_\tau^{-2}), \\ o(\tau^{\frac{1}{16}}\zeta_\tau^{-1}), & |\theta| \leq \frac{\zeta_\tau}{4}. \end{cases}$$

Proof. For proving the first assertion, following the steps of Proposition C.2 of van der Pas et al. (2017a), we can show that both $\int_{|x|\geq\kappa_\tau} m_\tau(x)\phi(x-\theta)dx = o(\frac{\tau}{\zeta_\tau})$ and $\int_{\zeta_\tau\leq|x|\leq\kappa_\tau} m_\tau(x)\phi(x-\theta)dx = o(\frac{\tau}{\zeta_\tau})$, where $\kappa_\tau \sim \zeta_\tau + \frac{2\log\zeta_\tau}{\zeta_\tau}$ as $\tau \to 0$, where κ_τ is defined in Lemma 2.5.3.
Note that, from (2.62) of Lemma 2.5.3, when $\frac{x^2}{2} \leq 1$, $I_{\alpha,\frac{1}{2}}(x) = \frac{K^{-1}}{\tau}[1 + O(\sqrt{\tau})]$. On the other hand, since, $\frac{e^{\frac{x^2}{2}}}{x^2}$ is increasing for large values of x and attains the value $\frac{\tau}{\zeta_\tau^2}$ at $x = \zeta_\tau$, by (2.62) of Lemma 2.5.3, $I_{\alpha,\frac{1}{2}}(x) = \frac{K^{-1}}{\tau}[1 + O(\frac{1}{\zeta_\tau^2})]$, when $1 \leq \frac{x^2}{2} \leq \log(\frac{1}{\tau})$. Combining these two facts,

we get that, $I_{\alpha,\frac{1}{2}}(x) = \frac{K^{-1}}{\tau}[1 + O(\frac{1}{\zeta_\tau^2})]$, uniformly in $x \in (0, \zeta_\tau)$. Hence,

$$\int_{|x|\le\zeta_\tau} m_\tau(x)\phi(x-\theta)dx = \int_0^{\zeta_\tau} \frac{x^2(J_{\alpha+1,\frac{1}{2}}(x) - J_{\alpha+2,\frac{1}{2}}(x)) - J_{\alpha+1,\frac{1}{2}}(x)}{\frac{K^{-1}}{\tau}}\phi(x)dx + R_\tau, \quad (2.83)$$

where the absolute value of R_τ is bounded by $\int_0^{\zeta_\tau} |x^2(J_{\alpha+1,\frac{1}{2}}(x) - J_{\alpha+2,\frac{1}{2}}(x)) - J_{\alpha+1,\frac{1}{2}}(x)|\phi(x)dx$ times $\sup_{|x|\le\zeta_\tau} |\frac{\phi(x-\theta)}{I_{\alpha,\frac{1}{2}}(x)\phi(x)} - \frac{K}{\tau^{-1}}|$. By using Lemma 2.5.4, the integrand is bounded above by a constant for x near 0 and by a multiple of x^{-2} otherwise, which makes the integral bounded. Next, observe that,

$$\begin{aligned}\sup_{|x|\le\zeta_\tau} |\frac{\phi(x-\theta)}{I_{\alpha,\frac{1}{2}}(x)\phi(x)} - \frac{K}{\tau^{-1}}| &= \sup_{|x|\le\zeta_\tau} \frac{\tau K}{I_{\alpha,\frac{1}{2}}(x)}|\tau^{-1}K^{-1}(e^{x\theta-\frac{\theta^2}{2}}) - I_{\alpha,\frac{1}{2}}(x)| \\ &= \sup_{|x|\le\zeta_\tau} \tau K|\frac{e^{x\theta-\frac{\theta^2}{2}}}{1+O(\frac{1}{\zeta_\tau^2})} - 1|\end{aligned}$$

Observe that, for $|\theta| = o(\zeta_\tau^{-2}), \zeta_\tau|\theta| - \frac{\theta^2}{2} = o(\zeta_\tau^{-1})$ and using the fact $e^y - 1 \sim y$ as $y \to 0$, we have

$$\sup_{|x|\le\zeta_\tau} |\frac{\phi(x-\theta)}{I_{\alpha,\frac{1}{2}}(x)\phi(x)} - \frac{K}{\tau^{-1}}| \lesssim \tau[\frac{1}{\zeta_\tau^2} + o(\zeta_\tau^{-1})] = o(\frac{\tau}{\zeta_\tau}).$$

These two arguments show that R_τ is negligible compared to $\frac{\tau}{\zeta_\tau}$.
Next, using Fubini's Theorem, the integral in (2.83) can be rewritten as

$$\begin{aligned}&\int_0^{\zeta_\tau} \frac{x^2(J_{\alpha+1,\frac{1}{2}}(x) - J_{\alpha+2,\frac{1}{2}}(x)) - J_{\alpha+1,\frac{1}{2}}(x)}{\frac{K^{-1}}{\tau}}\phi(x)dx \\ &= K\tau\int_0^1\int_0^{\zeta_\tau} z^\alpha\left(\frac{1}{N(z)}\right)^{\frac{1}{2}+\alpha} e^{\frac{x^2 z}{2}}[x^2(1-z)-1]\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx\,dz \quad (2.84) \\ &= K\tau\int_0^1 z^\alpha\left(\frac{1}{N(z)}\right)^{\frac{1}{2}+\alpha}\int_0^{\zeta_\tau}[x^2(1-z)-1]\frac{e^{-\frac{x^2(1-z)}{2}}}{\sqrt{2\pi}}dxdz.\end{aligned}$$

Note that the inner integral becomes zero if the range of integration is $(0,\infty)$ instead of $(0,\zeta_\tau)$. Hence, we have

$$\begin{aligned}&\int_0^{\zeta_\tau} \frac{x^2(J_{\alpha+1,\frac{1}{2}}(x) - J_{\alpha+2,\frac{1}{2}}(x)) - J_{\alpha+1,\frac{1}{2}}(x)}{\frac{K^{-1}}{\tau}}\phi(x)dx \\ &= -K\tau\int_0^1 z^\alpha\left(\frac{1}{N(z)}\right)^{\frac{1}{2}+\alpha}\frac{\zeta_\tau}{\sqrt{2\pi}}e^{-\frac{\zeta_\tau^2(1-z)}{2}}dz.\end{aligned}$$

In the last line above, we use the fact $\int_y^\infty[(vb)^2-1]\phi(vb)dv = y\phi(yb)$. Next, similar to Proposition C.2 of van der Pas et al. (2017a), we split the range of integration in $(0, \frac{1}{2}]$ and $(\frac{1}{2}, 1)$. When

$0 \leq z \leq \frac{1}{2}$, the absolute value of the integral is bounded above by

$$K\tau \int_0^{\frac{1}{2}} z^{\alpha} \left(\frac{1}{N(z)}\right)^{\frac{1}{2}+\alpha} \frac{\zeta_\tau}{\sqrt{2\pi}} e^{-\frac{\zeta_\tau^2(1-z)}{2}} dz \lesssim K \frac{e^{-\frac{\zeta_\tau^2}{4}} \tau\zeta_\tau}{\sqrt{2\pi}(1-\tau^2)^{\frac{1}{2}+\alpha}} \int_0^{\frac{1}{2}} z^{-\frac{1}{2}} dz$$
$$= O(e^{-\frac{\zeta_\tau^2}{4}} \tau\zeta_\tau) = o(\frac{\tau}{\zeta_\tau}).$$

On the other hand, when $\frac{1}{2} \leq z \leq 1$, we again use, $\frac{1}{\tau^2+(1-\tau^2)z} = \frac{1}{z}[1+O(\tau^2)]$. This implies

$$\begin{aligned}
&- K\tau \int_{\frac{1}{2}}^{1} z^{\alpha} \left(\frac{1}{N(z)}\right)^{\frac{1}{2}+\alpha} \frac{\zeta_\tau}{\sqrt{2\pi}} e^{-\frac{\zeta_\tau^2(1-z)}{2}} dz \\
&= -\frac{K\tau\zeta_\tau}{\sqrt{2\pi}} \int_{\frac{1}{2}}^{1} z^{-\frac{1}{2}} e^{-\frac{\zeta_\tau^2(1-z)}{2}} dz[1+O(\tau^2)] \\
&= -\frac{K}{\sqrt{2\pi}} \frac{\tau}{\zeta_\tau} \int_0^{\frac{\zeta_\tau^2}{2}} e^{-\frac{u}{2}} \frac{1}{(1-\frac{u}{\zeta_\tau^2})^{\frac{1}{2}}} du[1+O(\tau^2)],
\end{aligned}$$

where the equality is due to the substitution $\zeta_\tau^2(1-z) = u$. Finally, following the same argument as used in Proposition C.2 of van der Pas et al. (2017a), the integral tends to $\int_0^{\infty} e^{-\frac{u}{2}} du = 2$, completing the proof of the first assertion.
For proving the second statement, we follow the steps mentioned in Proposition C.2 of van der Pas et al. (2017a) and use that $e^{|\theta|2\zeta_\tau - \frac{\theta^2}{2}} \leq \tau^{-\frac{15}{16}}$, when $|\theta| \leq \frac{\zeta_\tau}{4}$. □

Proof of Theorem 2.2.7. For proving Theorem 2.2.7, with the use of (C1), note that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0} \Pi_{\widehat{\tau}_n}(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}) \Big) \\
&= \mathbb{E}_{\boldsymbol{\theta}_0} \left[\Pi_{\widehat{\tau}_n}(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}) 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \right] \\
&+ \mathbb{E}_{\boldsymbol{\theta}_0} \left[\Pi_{\widehat{\tau}_n}(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}) 1_{\{\widehat{\tau}_n \notin [\frac{1}{n}, C\tau_n(q_n)]\}} \right] \\
&\leq \mathbb{E}_{\boldsymbol{\theta}_0} \left[\Pi_{\widehat{\tau}_n}(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}) 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \right] + o(1). \qquad (2.85)
\end{aligned}$$

Next, applying Markov's inequality to the first term in the r.h.s. of (2.85), it is sufficient to show that

$$\mathbb{E}_{\theta_0} \sum_{i=1}^{n} Var(\theta_i | X_i, \widehat{\tau}_n) 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \lesssim q_n \log n \qquad (2.86)$$

and

$$\mathbb{E}_{\theta_0} ||T_{\widehat{\tau}_n}(\mathbf{X}) - \boldsymbol{\theta}_0||^2 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \lesssim q_n \log n. \qquad (2.87)$$

In order to prove (2.86), we need to show that

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} \lesssim \tilde{q}_n \log n \tag{2.88}$$

and

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} \lesssim q_n \log n. \tag{2.89}$$

Step-1 Fix any i such that $\theta_{0i} \neq 0$. Note that while proving the contraction rate of the empirical Bayes estimator of τ as proposed by van der Pas et al. (2014), we need to prove a statement same as that of (2.88). In that case, we used only that $\widehat{\tau} \geq \frac{1}{n}$ and some monotonicity of the shrinkage coefficient, which is true in this case, too. Hence, using the same arguments as used in Theorem 2.2.1 related to the non-null means, we can show (2.88) holds.
Step-2 Fix any i such that $\theta_{0i} = 0$. Next, moving towards proving (2.89), using (2.37) of Theorem 2.2.1 with $v_n = \sqrt{2\log n}$, we have,

$$\mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} 1_{\{|X_i|\leq v_n\}} \lesssim \frac{q_n}{n}\log n. \tag{2.90}$$

On the other hand, using (2.36) of Theorem 2.2.1,

$$\begin{aligned} &\mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} 1_{\{|X_i|> v_n\}} \\ &\leq \mathbb{E}_{\theta_{0i}}[Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{|X_i|> v_n\}}] \lesssim \frac{\sqrt{\log n}}{n}. \end{aligned} \tag{2.91}$$

Combining (2.90) and (2.91), we can conclude that

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}} Var(\theta_i|X_i,\widehat{\tau}_n) 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} \lesssim (n-\tilde{q}_n)[\frac{\sqrt{\log n}}{n} + \frac{q_n}{n}\cdot\log n] \lesssim q_n \log n\,.$$

This completes the proof of (2.89) and eventually (2.86) is also established.
Next, for (2.87), we aim to prove that

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}} (T_{\widehat{\tau}_n}(X_i) - \boldsymbol{\theta}_{0i})^2 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} \lesssim q_n \log n \tag{2.92}$$

and

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}} (T_{\widehat{\tau}_n}(X_i) - \boldsymbol{\theta}_{0i})^2 1_{\{\widehat{\tau}_n\in[\frac{1}{n},C\tau_n(q_n)]\}} \lesssim q_n \log n. \tag{2.93}$$

Step-1 Fix any i such that $\theta_{0i} \neq 0$. Again using the same set of arguments as used before, with Theorem 2 of Ghosh and Chakrabarti (2017) for non-zero means, (2.92) follows readily.

Step-2 Fix any i such that $\theta_{0i} = 0$. Note that,

$$\begin{aligned}
&\mathbb{E}_{\theta_{0i}}(T_{\widehat{\tau}_n}(X_i) - \boldsymbol{\theta}_{0i})^2 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \leq \mathbb{E}_{\theta_{0i}}[T^2_{\tau_n(q_n)}(X_i)] \\
&= \mathbb{E}_{\theta_{0i}}[T^2_{\tau_n(q_n)}(X_i) 1_{\{|X_i| > \sqrt{2\log\left(\frac{1}{\tau_n(q_n)}\right)}\}}] + \mathbb{E}_{\theta_{0i}}[T^2_{\tau_n(q_n)}(X_i) 1_{\{|X_i| \leq \sqrt{2\log\left(\frac{1}{\tau_n(q_n)}\right)}\}}] \\
&\leq \mathbb{E}_{\theta_{0i}}[X_i^2 1_{\{|X_i| > \sqrt{2\log\left(\frac{1}{\tau_n(q_n)}\right)}\}}] + \mathbb{E}_{\theta_{0i}}[T^2_{\tau_n(q_n)}(X_i) 1_{\{|X_i| \leq \sqrt{2\log\left(\frac{1}{\tau_n(q_n)}\right)}\}}] \\
&\lesssim e^{-\log\left(\frac{1}{\tau_n(q_n)}\right)} \sqrt{\log\left(\frac{n}{q_n}\right)} + (\tau_n(q_n))^2 \sqrt{\log\left(\frac{n}{q_n}\right)} e^{\log\left(\frac{1}{\tau_n(q_n)}\right)} \\
&\lesssim \frac{q_n}{n} \log\left(\frac{n}{q_n}\right), \qquad (2.94)
\end{aligned}$$

where we use facts similar to (15) and (43) of Ghosh and Chakrabarti (2017). As a result,

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}}(T_{\widehat{\tau}_n}(X_i) - \boldsymbol{\theta}_{0i})^2 1_{\{\widehat{\tau}_n \in [\frac{1}{n}, C\tau_n(q_n)]\}} \lesssim (n - \tilde{q}_n)\frac{q_n}{n} \log\left(\frac{n}{q_n}\right) \lesssim q_n \log n.$$

This completes the proof of (2.93) and as a consequence of this, (2.87) is also established. Combining (2.85)-(2.87) completes the proof of Theorem 2.2.7. □

2.5.3 Results related to the contraction rate of Hierarchical Bayes

Proof of Theorem 2.2.11. In order to prove Theorem 2.2.11 with the use of Lemma 2.2.9,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_0} \Pi\Big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}\Big) \\
&= \mathbb{E}_{\boldsymbol{\theta}_0}\Big[\int_{\frac{1}{n}}^{\frac{1}{\log n}} \Pi\big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}, \tau\big)\pi(\tau|\mathbf{X})d\tau\Big] \\
&= \mathbb{E}_{\boldsymbol{\theta}_0}\Big[\big(\int_{\frac{1}{n}}^{5t_n} + \int_{5t_n}^{\frac{1}{\log n}}\big)\Pi\big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}, \tau\big)\pi(\tau|\mathbf{X})d\tau\Big] \\
&\leq \mathbb{E}_{\boldsymbol{\theta}_0}\Big[\int_{\frac{1}{n}}^{5t_n} \Pi\big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}, \tau\big)\pi(\tau|\mathbf{X})d\tau\Big] + o(1) \\
&\leq \mathbb{E}_{\boldsymbol{\theta}_0}[\sup_{\frac{1}{n} \leq \tau \leq 5t_n} \Pi\big(\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0||^2 > M_n q_n \log n | \mathbf{X}, \tau\big)] + o(1). \qquad (2.95)
\end{aligned}$$

Next, applying Markov's inequality to the first term in the r.h.s. of (2.95), it is sufficient to show that

$$\mathbb{E}_{\boldsymbol{\theta}_0}\Big[\sup_{\frac{1}{n} \leq \tau \leq 5t_n} \sum_{i=1}^{n} Var(\theta_i | \mathbf{X}, \tau)\Big] \lesssim q_n \log n \qquad (2.96)$$

and

$$\mathbb{E}_{\boldsymbol{\theta}_0}\Big[\sup_{\frac{1}{n} \leq \tau \leq 5t_n} ||\boldsymbol{\theta}_0 - T_\tau(\mathbf{X})||^2\Big] \lesssim q_n \log n. \qquad (2.97)$$

Noting that the posterior distribution of θ_i given $(\boldsymbol{X}, \tau)$ depends on (X_i, τ) only, we are left to prove that

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n}\sum_{i=1}^{n} Var(\theta_i|X_i,\tau)\right] \lesssim q_n \log n. \tag{2.98}$$

In order to prove (2.98), we need to show that

$$\sum_{i:\theta_{0i}\neq 0}\mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\right] \lesssim q_n \log n \tag{2.99}$$

and

$$\sum_{i:\theta_{0i}=0}\mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\right] \lesssim q_n \log n \tag{2.100}$$

Step-1 Fix any i such that $\theta_{0i} \neq 0$. Define $r_n = \sqrt{4a\rho^2 \log n}$ where ρ is defined in the proof of Theorem 2.2.1. Since, for any fixed $x_i \in \mathbb{R}$ and $\tau > 0$, $Var(\theta_i|x_i,\tau) \leq 1 + x_i^2$,

$$\mathbb{E}_{\theta_{0i}}\left(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|\leq\sqrt{4a\rho^2\log n}\}}\right) \leq (1 + 4a\rho^2 \log n)\,. \tag{2.101}$$

Now, we want to bound $\mathbb{E}_{\theta_{0i}}\left(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|>\sqrt{4a\rho^2\log n}\}}\right)$. Again we use the fact that for any fixed $x_i \in \mathbb{R}$ and $\tau > 0$, $Var(\theta_i|x_i,\tau) \leq 1 + x_i^2$. Using this fact and the monotonicity of $E(\kappa_i^2|x_i,\tau)$ for any fixed x_i, we have as in **Step-1** of Theorem 2.2.1, for that any $\tau \in [\frac{1}{n}, 5t_n]$

$$Var(\theta_i|x_i,\tau) \leq 1 + x_i^2 E(\kappa_i^2|x_i,\tau) \leq 1 + x_i^2 E(\kappa_i^2|x_i,\frac{1}{n}) \;= \tilde{h}(x_i,\frac{1}{n})\,.$$

Hence, using the same arguments used in **Step-1** of Theorem 2.2.1,

$$\sup_{|x_i|>\sqrt{4a\rho^2\log n}} Var(\theta_i|x_i,\tau) \leq \sup_{|x_i|>\sqrt{4a\rho^2\log n}} \tilde{h}(x_i,\frac{1}{n})$$

$$\leq 1 + \sup_{|x_i|>\sqrt{4a\rho^2\log n}} \tilde{h}_1(x_i,\frac{1}{n}) + \sup_{|x_i|>\sqrt{4a\rho^2\log n}} \tilde{h}_2(x_i,\frac{1}{n}) \lesssim 1.$$

Using the above arguments,

$$\mathbb{E}_{\theta_{0i}}\left(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|>r_n\}}\right) \lesssim 1\,. \tag{2.102}$$

Combining (2.101) and (2.102), we obtain,

$$\mathbb{E}_{\theta_{0i}}\left(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\right) \lesssim \log n\,.$$

Since all of the above arguments hold uniformly for any i such that $\theta_{0i} \neq 0$, as $n \to \infty$,

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}}\Big(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\Big) \lesssim \tilde{q}_n \log n\,. \tag{2.103}$$

Step-2 Fix any i such that $\theta_{0i} = 0$. Define $s_n = \sqrt{4a\log\left(\frac{1}{t_n}\right)}$. We again decompose $\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\Big]$ as

$$\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\Big] = \mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|\leq s_n\}}\Big]$$

$$+\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|> s_n\}}\Big]\,. \tag{2.104}$$

Since for any fixed $x_i \in \mathbb{R}$ and $\tau > 0$, $Var(\theta_i|x_i,\tau) \leq E(1-\kappa_i|x_i,\tau) + J(x_i,\tau)$ and $E(1-\kappa_i|x_i,\tau)$ is non-decreasing in τ and using Lemma 2 and Lemma A.2 of Ghosh and Chakrabarti (2017), for $a \in [0.5,1)$, for any fixed $x_i \in \mathbb{R}$ and $\tau > 0$,

$$\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|\leq\sqrt{4a\log\left(\frac{1}{t_n}\right)}\}}\Big] \leq t_n^{2a}\int_0^{\sqrt{4a\log\left(\frac{1}{t_n}\right)}} e^{\frac{x^2}{2}}\phi(x)dx$$

$$\lesssim \frac{q_n}{n}\log\left(\frac{n}{q_n}\right). \tag{2.105}$$

Using the fact that for any $\tau > 0$, $Var(\theta_i|x_i,\tau) \leq 1 + x_i^2$ and the identity $x^2\phi(x) = \phi(x) - \frac{d}{dx}[x\phi(x)]$, we obtain,

$$\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)1_{\{|X_i|>\sqrt{4a\log\left(\frac{1}{t_n}\right)}\}}\Big] \lesssim \frac{q_n}{n}\log\left(\frac{n}{q_n}\right)\,. \tag{2.106}$$

On combining (2.104)-(2.106), for sufficiently large n,

$$\mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\Big] \lesssim \frac{q_n}{n}\log\left(\frac{n}{q_n}\right)\,. \tag{2.107}$$

Note that all these preceding arguments hold uniformly in i such that $\theta_{0i} = 0$. Hence, for any $a \in [0.5,1)$ as $n \to \infty$,

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}}\Big(\sup_{\frac{1}{n}\leq\tau\leq 5t_n} Var(\theta_i|X_i,\tau)\Big) \lesssim (n-\tilde{q}_n)\frac{q_n}{n}\log\left(\frac{n}{q_n}\right) \lesssim q_n\log\left(\frac{n}{q_n}\right)\,. \tag{2.108}$$

The second inequality follows due to the fact that $\tilde{q}_n \leq q_n$ and $q_n = o(n)$ as $n \to \infty$. This completes the proof for (2.98). Moving towards proving (2.97), here, we need to prove that

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}}\Big[\sup_{\frac{1}{n}\leq\tau\leq 5t_n} (T_\tau(X_i)-\theta_{0i})^2\Big] \lesssim q_n\log n \tag{2.109}$$

and

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-\theta_{0i})^2\right] \lesssim q_n \log n. \tag{2.110}$$

Step-1 Fix any i such that $\theta_{0i} \neq 0$. First, note that,

$$(T_\tau(X_i)-\theta_{0i})^2 \leq 2[(T_\tau(X_i)-X_i)^2+(X_i-\theta_{0i})^2]. \tag{2.111}$$

Observe that, the second term in the r.h.s. of (2.111) is independent of choice of τ. Also, using the fact, for any $x_i \in \mathbb{R}, |T_\tau(x_i)-x_i| \leq x_i$, we have

$$(T_\tau(X_i)-X_i)^2 1_{\{|X_i|\leq\rho\zeta_\tau\}} \lesssim \zeta_\tau^2.$$

Also, using Lemma 3 of Ghosh and Chakrabarti (2017), there exists a non-negative real-valued function h such that $|T_\tau(x_i)-x_i| \leq h(x_i,\tau)$ for all $x_i \in \mathbb{R}$ satisfying for any $\rho > c > 2$,

$$\lim_{\tau\to 0}\sup_{|x_i|>\rho\zeta_\tau} h(x_i,\tau) = 0.$$

Combining these two arguments, we conclude, as $\tau \to 0$,

$$\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-X_i)^2 \lesssim \sup_{\frac{1}{n}\leq\tau\leq 5t_n}\zeta_\tau^2(1+o(1)) \lesssim \log n. \tag{2.112}$$

Using (2.111), (2.112) and noting that $\mathbb{E}_{\theta_{0i}}(X_i-\theta_{0i})^2 = 1$, we obtain

$$\mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-\theta_{0i})^2\right] \lesssim \log n.$$

Since all of the above arguments hold uniformly for any i such that $\theta_{0i} \neq 0$, as $n \to \infty$,

$$\sum_{i:\theta_{0i}\neq 0} \mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-\theta_{0i})^2\right] \lesssim q_n \log n.$$

Step-2 Fix any i such that $\theta_{0i} = 0$. In this case,

$$\mathbb{E}_{\theta_{0i}}\left[\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-\theta_{0i})^2\right] \leq \mathbb{E}_{\theta_{0i}}[T_{t_n}^2(X_i)] \lesssim \frac{q_n}{n}\log\left(\frac{n}{q_n}\right),$$

where the inequality in the last line follows due to the use of an argument similar to that of (2.94). Note that all these preceding arguments hold uniformly in i such that $\theta_{0i} = 0$. Hence, for any $a \in [0.5, 1)$ as $n \to \infty$,

$$\sum_{i:\theta_{0i}=0} \mathbb{E}_{\theta_{0i}}\left(\sup_{\frac{1}{n}\leq\tau\leq 5t_n}(T_\tau(X_i)-\theta_{0i})^2\right) \lesssim (n-\tilde{q}_n)\frac{q_n}{n}\log\left(\frac{n}{q_n}\right) \lesssim q_n \log\left(\frac{n}{q_n}\right). \tag{2.113}$$

The second inequality follows due to the fact that $\tilde{q}_n \leq q_n$ and $q_n = o(n)$ as $n \to \infty$. This completes the proof for (2.97). As a result, Theorem 2.2.11 is also established using (2.95)-

(2.97). □

2.5.4 Results related to the ABOS of the full Bayes procedure

Proof of Theorem 2.3.1. First, note that

$$E(1-\kappa_i|\mathbf{X}) = \int_{\frac{1}{n}}^{\alpha_n} E(1-\kappa_i|\mathbf{X},\tau)\pi_{2n}(\tau|\mathbf{X})d\tau$$
$$= \int_{\frac{1}{n}}^{\alpha_n} E(1-\kappa_i|X_i,\tau)\pi_{2n}(\tau|\mathbf{X})d\tau \leq E(1-\kappa_i|X_i,\alpha_n)\cdot \tag{2.114}$$

For proving the inequality above we first use the fact that given any $x_i \in \mathbb{R}$, $E(1-\kappa_i|x_i,\tau)$ is non-decreasing in τ and next we use (C4). Now, using Theorem 4 of Ghosh et al. (2016), we have for $a \in (0,1)$, as $n \to \infty$

$$E(1-\kappa_i|X_i,\alpha_n) \leq \frac{KM}{a(1-a)} e^{\frac{X_i^2}{2}} \alpha_n^{2a}(1+o(1)). \tag{2.115}$$

Here $o(1)$ depends only on n such that $\lim_{n\to\infty} o(1) = 0$ and is independent of i. Two constants K and M are defined in (2.3) and Assumption 1, respectively. Hence, we have the following :

$$t_{1i}^{\text{FB}} = P_{H_{0i}}(E(1-\kappa_i|\mathbf{X}) > \frac{1}{2})$$
$$\leq P_{H_{0i}}\left(\frac{X_i^2}{2} > 2a\log\left(\frac{1}{\alpha_n}\right) + \log\left(\frac{a(1-a)}{2A_0KM}\right) - \log(1+o(1))\right)$$
$$= P_{H_{0i}}\left(|X_i| > \sqrt{4a\log\left(\frac{1}{\alpha_n}\right)}\right)(1+o(1))\,.$$

Note that, as $\alpha_n < 1$ for all $n \geq 1$, $4a\log\left(\frac{1}{\alpha_n}\right) > 0$ for all $n \geq 1$. Next, using the fact that under $H_{0i}, X_i \overset{iid}{\sim} \mathcal{N}(0,1)$ with $1-\Phi(t) < \frac{\phi(t)}{t}$ for $t > 0$, we get

$$t_{1i}^{\text{FB}} \leq 2\frac{\phi(\sqrt{4a\log\left(\frac{1}{\alpha_n}\right)})}{\sqrt{4a\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1)) = \frac{1}{\sqrt{\pi a}}\frac{\alpha_n^{2a}}{\sqrt{\log\left(\frac{1}{\alpha_n^2}\right)}}(1+o(1))\,. \tag{2.116}$$

Here $o(1)$ depends only on n such that $\lim_{n\to\infty} o(1) = 0$. This completes the proof of Theorem 2.3.1. □

Proof of Theorem 2.3.2. In order to provide an upper bound on the probability of type-II error induced by the decision rule (2.24), $t_{2i}^{\text{FB}} = P_{H_{1i}}(E(\kappa_i|\mathbf{X}) > \frac{1}{2})$, we first note that,

$$E(\kappa_i|\mathbf{X}) = \int_{\frac{1}{n}}^{\alpha_n} E(\kappa_i|\mathbf{X},\tau)\pi_{2n}(\tau|\mathbf{X})d\tau$$
$$= \int_{\frac{1}{n}}^{\alpha_n} E(\kappa_i|X_i,\tau)\pi_{2n}(\tau|\mathbf{X})d\tau \leq E(\kappa_i|X_i,\frac{1}{n}), \tag{2.117}$$

where inequality in the last line follows due to the fact that given any $x_i \in \mathbb{R}$, $E(\kappa_i|x_i,\tau)$ is non-increasing in τ. Fix any $\rho > 2$. Then we choose $\eta \in (0,1)$ and $\delta \in (0,1)$ such that this ρ is strictly larger than $\frac{2}{\eta(1-\delta)}$. Using arguments similar to Lemma 3 of Ghosh and Chakrabarti (2017), $E(\kappa_i|x_i,\tau)$ can be bounded above by a real-valued function $g(x_i,\tau,\eta,\delta)$, depending on η and δ such that

$$\lim_{n\to\infty} \sup_{|x_i|>\sqrt{\rho\log(n^{2a})}} g(x_i,\frac{1}{n},\eta,\delta) = 0\,, \tag{2.118}$$

where ρ,η,δ are as above. Precisely $g(x_i,\tau,\eta,\delta) = g_1(x_i,\tau) + g_2(x_i,\tau,\eta,\delta)$ where $g_1(x_i,\tau) = C(a,\eta,L)[x_i^2 \int_0^{\frac{x_i^2}{1+t_0}} e^{-\frac{u}{2}} u^{a-\frac{1}{2}} du]^{-1}$ and $g_2(x_i,\tau,\eta,\delta) = \tau^{-2a} c_0 H(a,\eta,\delta) e^{-\frac{\eta(1-\delta)x_i^2}{2}}$ where $C(a,\eta,L)$ and $H(a,\eta,\delta)$ are two constants independent of i and n and c_0 is as defined in Assumption 1. It is evident from (2.118) that for all sufficiently large n,

$$\sup_{|x_i|>\sqrt{\rho\log(n^{2a})}} g(x_i,\frac{1}{n},\eta,\delta) \le \frac{1}{2}.$$

Now define B_n and C_n as $B_n = \{g(X_i,\frac{1}{n},\eta,\delta) > \frac{1}{2}\}$ and $C_n = \{|X_i| > \sqrt{\rho\log(n^{2a})}\}$. Hence

$$\begin{aligned} t_{2i}^{\mathrm{FB}} &= P_{H_{1i}}(E(\kappa_i|\mathbf{X}) > \frac{1}{2}) \\ &\le P_{H_{1i}}(B_n) \le P_{H_{1i}}(B_n \cap C_n) + P_{H_{1i}}(C_n^c), \end{aligned} \tag{2.119}$$

for any fixed $(\eta,\delta) \in (0,1)\times(0,1)$ and $\rho > \frac{2}{\eta(1-\delta)}$. Recall that, from the definition of $g_1(\cdot)$ and $g_2(\cdot)$, it is clear that $g(\cdot)$ is monotonically decreasing in $|x_i|$ for $x_i \neq 0$. Next using the definition of B_n and C_n and with the use of (2.118), we immediately have,

$$P_{H_{1i}}(B_n \cap C_n) = 0 \tag{2.120}$$

for all sufficiently large n, depending on ρ,η,δ. Further, note that under $H_{1i}, X_i \overset{iid}{\sim} \mathcal{N}(0,1+\psi^2)$ and hence

$$\begin{aligned} P_{H_{1i}}(C_n^c) &= P_{H_{1i}}(|X_i| \le \sqrt{\rho\log(n^{2a})}) \\ &= P_{H_{1i}}\left(\frac{|X_i|}{\sqrt{1+\psi^2}} \le \sqrt{a\rho}\sqrt{\frac{2\log n}{1+\psi^2}}\right) = \left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1)), \end{aligned} \tag{2.121}$$

where the last equality follows from Assumption 2 and due to $\frac{\log\left(\frac{1}{p_n}\right)}{\log n} \to C_1$. Here the term $o(1)$ is independent of i and $\lim_{n\to\infty} o(1) = 0$. From (2.119)-(2.121) the desired result follows. □

Proof of Theorem 2.3.4. The Bayes risk corresponding to the decision rule (2.24), denoted $R_{\mathrm{OG}}^{\mathrm{FB}}$

is of the form

$$R_{\mathrm{OG}}^{\mathrm{FB}} = \sum_{i=1}^{n}[(1-p)t_{1i}^{\mathrm{FB}} + pt_{2i}^{\mathrm{FB}}] = p\sum_{i=1}^{n}\left[\left(\frac{1-p}{p}\right)t_{1i}^{\mathrm{FB}} + t_{2i}^{\mathrm{FB}}\right]. \tag{2.122}$$

Now, for $a \in [0.5, 1)$, using (2.116) we have, $t_{1i}^{\mathrm{FB}} \lesssim \frac{\alpha_n}{\sqrt{\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1))$. Hence, for $p_n \to 0$ as $n \to \infty$ such that $\underline{\lim}\, np_n > 0$, along with the choice of α_n as $\log\left(\frac{1}{\alpha_n}\right) = \log n - \frac{1}{2}\log\log n + c_n$, where $c_n = o(\log\log n)$ as $n \to \infty$, we have, $\overline{\lim}\, \frac{1}{np_n} < \infty$ and $\overline{\lim}\, \frac{n\alpha_n}{\sqrt{\log\left(\frac{1}{\alpha_n}\right)}} = 0$. These imply

$$\overline{\lim}\, \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-p}{p}\right)t_{1i}^{\mathrm{FB}} = 0\,. \tag{2.123}$$

Next, note that, from Theorem 2.3.2, for each i, the upper bound of t_{2i}^{FB} is independent of i. Therefore, we obtain,for any $\rho > 2$,

$$\sum_{i=1}^{n} t_{2i}^{\mathrm{FB}} \le n\left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1))\,. \tag{2.124}$$

As a result, using (2.122)-(2.124), for any $\rho > 2$ and any $C_1 \in (0, 1]$,

$$R_{\mathrm{OG}}^{\mathrm{FB}} \le np\left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right](1+o(1))\,. \tag{2.125}$$

This proves the first assertion of the theorem. We now prove the final assertion of the theorem for the case $C_1 = 1$. Since, by definition, $\frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}} \ge 1$, using (2.21) and (2.125), for any $\rho > 2$, and any $C_1 \in (0, 1]$,

$$1 \le \frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}} \le \frac{\left[2\Phi\left(\sqrt{\frac{a\rho C}{C_1}}\right) - 1\right]}{\left[2\Phi(\sqrt{C}) - 1\right]}(1+o(1)), \tag{2.126}$$

where $o(1)$ term is independent of i and satisfies that $\lim_{n\to\infty} o(1) = 0$. For $C_1 = 1$, taking limit inferior and limit superior in (2.126), we have, for any $\rho > 2$,

$$1 \le \underline{\lim}\, \frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}} \le \overline{\lim}\, \frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}} \le \frac{\left[2\Phi\left(\sqrt{a\rho C}\right) - 1\right]}{\left[2\Phi(\sqrt{C}) - 1\right]}. \tag{2.127}$$

Note that the decision rule does not depend on how $\rho > 2$ is chosen. Hence, the ratio $\frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}}$ is also independent of the choice of ρ for any $n \ge 1$. As a result, limit inferior and limit superior in (2.127) are also free of any particular value of ρ. Taking infimum of the r.h.s. of (2.127) over all possible choices of $\rho > 2$ and finally using continuity of $\Phi(\cdot)$, we conclude that, for the case

$C_1 = 1$ with $a = 0.5$,

$$\lim_{n\to\infty} \frac{R_{\mathrm{OG}}^{\mathrm{FB}}}{R_{\mathrm{Opt}}^{\mathrm{BO}}} = 1.$$

□

Chapter 3

Sharp Asymptotic Minimaxity for One-Group Priors in Sparse Normal Means Problem

3.1 Introduction

Multiple hypothesis testing is a prominent area of research in statistics due to its wide range of applications as discussed before. Suppose we have an observation vector $\mathbf{X} \in \mathbb{R}^n$, $\mathbf{X} = (X_1, X_2, \cdots, X_n)$, where each observation X_i is expressed as,

$$X_i = \theta_i + \epsilon_i, i = 1, 2, \cdots, n, \tag{3.1}$$

where the parameters $\theta_1, \theta_2, \cdots, \theta_n$ are unknown mean parameters and ϵ_i's are the random errors present in the model. Generally, ϵ_i's are assumed to be independent $\mathcal{N}(0, \sigma^2)$ random variables. In the literature of multiple testing, usually σ^2 is assumed to be known and hence without loss of generality, a typical choice of σ^2 is $\sigma^2 = 1$. This is the famous *normal means model* or *Gaussian sequence model*. In this chapter we are interested in problem of testing simultaneously $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ for $i = 1, \ldots, n$. Our goal in this chapter is to investigate the decision theoretic optimality of multiple testing rules using one-group priors, where the benchmark of optimality is taken to be the minimax risk with respect to two natural loss functions. This study is different from the corresponding study of optimality of multiple testing in Chapter 2, where the Bayes risk is the criterion of interest.

A multiple testing procedure is a measurable function $\boldsymbol{\psi}$ of the data as follows: $\boldsymbol{\psi} : \mathbf{X} \in \mathbb{R}^n \mapsto (\psi_i(\mathbf{X}))_{i=1,2,\cdots,n} \in \{0, 1\}^n$ where $\psi_i \equiv \psi_i(\mathbf{X}) = 1$ implies rejecting the i^{th} null hypothesis, H_{0i}.

As mentioned earlier, the primary objective in a multiple testing problem is to form a decision rule which controls some measure of overall type I error. The most popular measure one of this kind is the False Discovery Rate (FDR, defined below) introduced by Benjamini and Hochberg

(1995). Given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ and any testing procedure ψ, the FDR of $\boldsymbol{\psi}$ at $\boldsymbol{\theta}$ is defined as

$$FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{i:\theta_i=0} \psi_i}{max(\sum_{i=1}^n \psi_i, 1)}\right]. \tag{3.2}$$

Given that a testing rule controls the FDR at some desired level, the next obvious question is whether the same rule can provide any control over type II error, which can ensure power behaviour of the procedure. This probability of type II error in a multiple testing procedure can be measured by the False Negative Rate (FNR). For a testing rule $\boldsymbol{\psi}$, FNR at $\boldsymbol{\theta}$ is as defined in Arias-Castro and Chen (2017), given by

$$FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{i:\theta_i\neq 0}(1-\psi_i)}{max(\sum_{i=1}^n 1_{\theta_i\neq 0}, 1)}\right]. \tag{3.3}$$

In this context, a question in which researchers are often interested is the optimality of the testing procedure in terms of controlling the sum of type I and type II errors that is attainable by any multiple testing rule. In case of a single testing problem, this question has been addressed in depth by Ingster and Suslina (2012) and Giné and Nickl (2021) involving one test. The same question has also been studied in recent times in the context of multiple hypothesis testing. The multiple testing risk at $\boldsymbol{\theta}$ for a testing procedure $\boldsymbol{\psi}$ is defined as

$$R(\boldsymbol{\theta}, \boldsymbol{\psi}) = FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) + FNR(\boldsymbol{\theta}, \boldsymbol{\psi}). \tag{3.4}$$

Given this risk, as pointed out before, it is of interest to find its optimum value in terms of minimax criterion, defined as $R(\boldsymbol{\Theta}) = \inf_{\boldsymbol{\psi}} \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \boldsymbol{\psi})$ and actual procedures achieving this optimum value. Note that the most naive testing rule $\psi_i = 0$ for all i implies the corresponding minimax risk is exactly 1. Hence, it is necessary to include non-trivial testing procedures in our class of procedures such that the resultant $R(\boldsymbol{\Theta})$ is strictly less than 1. It is also interesting to consider the minimax risk corresponding to the classification (or Hamming) loss $L(\boldsymbol{\theta}, \boldsymbol{\psi})$, where the misclassification loss is defined as

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i=1}^n [\mathbf{1}\{\theta_i = 0\}\mathbf{1}\{\psi_i \neq 0\} + \mathbf{1}\{\theta_i \neq 0\}\mathbf{1}\{\psi_i = 0\}]. \tag{3.5}$$

Given the loss $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ as defined in (3.5), the corresponding minimax risk is defined as $R(\boldsymbol{\Theta}) = \inf_{\boldsymbol{\psi}} \sup_{\boldsymbol{\theta}\in\Theta} EL(\boldsymbol{\theta}, \boldsymbol{\psi})$. The relevant literature in such contexts has already been discussed in some detail in Subsections 1.1.4 and 1.2.2. Some technical details of references related to our work, which are already available in the literature, are presented in Section 3.2. Now we briefly highlight the contents of this chapter along with our contributions in this chapter.

In Section 3.3, the results based on one-group priors are stated. At first, we establish that when the level of sparsity is known, for the one-group priors of the form (1.7) satisfying (1.10), the global parameter τ can be chosen based on the proportion of non-null means such that the type II error probability of our decision rule (3.15) converges to 1 as n tends to infinity for $a > \frac{1}{2}$. This implies our proposed decision rule can not attain the minimax risk when $a > \frac{1}{2}$. Hence,

throughout the remaining part of the chapter, we are interested when $a = \frac{1}{2}$, i.e., *horseshoe-type* priors. Section 3.3 consists of two subsections. Subsection 3.3.1 is related to the risk obtained for the classification loss, even if the level of sparsity is known or unknown. When the level of sparsity is known, choosing the global parameter τ based on the proportion of the non-null means, we have established that our decision rule based on a broad class of one-group priors attains the minimax risk, asymptotically. Next, we prove that even if the level of sparsity is unknown, our modified decision rules, either using an empirical Bayes approach or using a full Bayes approach, also achieve the same minimax risk. These results are reported in Theorems 3.3.2, 3.3.4 and 3.3.6, respectively. On the other hand, results related to the expected loss, defined as the sum of FDR and FNR, are discussed in Subsection 3.3.2. These results establish that the previously stated decision rules also attain the minimax risk corresponding to the sum of FDR and FNR, irrespective of whether the level of sparsity is known or unknown. These results are stated in Theorems 3.3.7, 3.3.9 and 3.3.11, respectively.

Section 3.4 contains an overall discussion of our work along with some future problems related to it. Proofs corresponding to our results are provided in Section 3.5.

This chapter is based on Paul et al. (2025).

3.2 Existing results on minimax risk

The most recent and widely studied work on the minimax risk for the Gaussian sequence model is due to Abraham et al. (2024). At the initial stage of their work, under the Gaussianity assumption, they introduced the beta-min condition which is designed to separate the alternative hypothesis from the null, and is defined as follows: given $\tilde{a} \in \mathbb{R}$, set

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}(\tilde{a}, q_n) = \{\boldsymbol{\theta} \in \ell_0[q_n] : |\theta_i| \geq \tilde{a} \text{ for } i \in S_{\boldsymbol{\theta}}, |S_{\boldsymbol{\theta}}| = q_n\}, \tag{3.6}$$

where $S_{\boldsymbol{\theta}} = \{i : \theta_i \neq 0\}$. Motivated by the previous works of Arias-Castro and Chen (2017) and Rabinovich et al. (2020), they chose $\tilde{a}$ defined in (3.6) to be close to the universal threshold, $\sqrt{2\log\left(\frac{n}{q_n}\right)}$, and considered the set:

$$\boldsymbol{\Theta}_b = \boldsymbol{\Theta}(\tilde{a}_b, q_n), \tilde{a}_b = \sqrt{2\log\left(\frac{n}{q_n}\right)} + b, \tag{3.7}$$

with $\boldsymbol{\Theta}(\tilde{a}, q_n)$ is as same as (3.6). They established that for any $b \in \mathbb{R}$, the minimax risk over $\boldsymbol{\Theta}_b$ is $1 - \Phi(b) + o(1)$, i.e.

$$\inf_{\boldsymbol{\psi}} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} R(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1). \tag{3.8}$$

This result continues to hold even if $b = b_n \to \infty$ and $b = b_n \to -\infty$. As a byproduct of this result, they obtained that the oracle thresholding rule, $\psi_i = \mathbf{1}\{|X_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}\}, i = 1, \ldots, n$ is asymptotically minimax.

Since the oracle thresholding rule depends on the number of non-zero means, q_n, which is unknown quite often, they proved that both the BH procedure of Benjamini and Hochberg (1995) with appropriately chosen $\alpha = \alpha_n \to 0$ as $n \to \infty$ and ℓ-value based procedure using the two-group spike-and-slab priors with a proper choice of slab distribution results in attaining the minimax risk with b defined in (3.8) either being finite or $b = b_n$ tending to ∞ even if $q_n \lesssim n^c$, for some $0 < c < 1$. Now, we briefly mention the l-value based procedure. For a detailed discussion, see Section S-8 in the supplementary material of Abraham et al. (2024) and the references therein. Recall from Subsection 1.1.2 that in a sparse situation, in the Bayesian paradigm, usually θ_i is modelled by a two-group spike-and-slab prior, as given in (1.2). In this problem, Abraham et al. (2024) assumed the slab density to be quasi-Cauchy, such that the convolution of $\phi * f$ (f being the density of F given in (1.2)), say $\tilde{f}$ is such that

$$\tilde{f}(x) = \frac{1}{\sqrt{2\pi}} x^{-2}\left(1 - \exp\left(-\frac{x^2}{2}\right)\right), x \in \mathbb{R}.$$

Now, given any $t \in (0,1)$, they considered a multiple testing rule which depends on the posterior inclusion probability of θ_i, and is of the form,

$$\psi^{\ell}(\mathbf{X}) = (\mathbf{1}\{\ell_{i,p}(\mathbf{X}) < t\})_{i=1,2,\cdots,n},$$

where

$$\ell_{i,p}(\mathbf{X}) = \Pi_p(\theta_i = 0|\mathbf{X}) = \frac{(1-p)\phi(X_i)}{(1-p)\phi(X_i) + p\tilde{f}(X_i)}.$$

Since p is unknown, they used the Marginal Maximum Likelihood Estimate (MMLE) of p, which is defined as the maximizer of the likelihood (or equivalently log-likelihood) function of $\mathbf{X}$ given p and is of the form $\hat{p} = argmax_{p\in[\frac{1}{n},1]} L(p)$, $L(p)$ being the log-likelihood function. Hence, Abraham et al. (2024) considered the decision rule of the form $\psi^{\hat{\ell}}(\mathbf{X})$, where $\psi^{\hat{\ell}}(\mathbf{X}) = \psi^{\ell}(\mathbf{X})$ evaluated at $p = \hat{p}$.

Next, they generalized the problem of obtaining the minimax risk by relaxing assumptions both based on signal strength and the noise. For weakening the assumption based on noise, they assumed that the observations are generated from a continuous distribution on $\mathbb{R}$ such that the CDF of the observations is $\tilde{L}-$Lipschitz for some constant $\tilde{L}$. Some assumptions on the expected number of nulls were also required. On the other hand, in order to study more general situations based on the signal strength, they considered a more general set of signals containing different signal strengths. For any given $\tilde{\mathbf{a}} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{q_n})$, with $\tilde{a}_j > 0$ for all $j = 1, 2, \ldots, q_n$, let us consider the set

$$\mathbf{\Theta}(\tilde{\mathbf{a}}, q_n) = \{\boldsymbol{\theta} \in \ell_0[q_n] : \exists i_1, i_2, \ldots, i_{q_n} \text{ all distinct}, |\theta_{i_j}| \geq \tilde{a}_j \text{ for } j = 1, 2, \ldots, q_n\}, \tag{3.9}$$

To study the risk, next they introduced $\Lambda_n(\tilde{\mathbf{a}}) = \frac{1}{q_n}\sum_{j=1}^{q_n} F_{\tilde{a}_j}(a_n^*)$ where $a_n^* = (\zeta \log\left(\frac{n}{q_n}\right))^{\frac{1}{\zeta}}$ for some $\zeta > 1$. Under the assumptions mentioned above, for location models, the minimax risk is

obtained as

$$\inf_{\psi} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(\tilde{\mathbf{a}},q_n)} R(\boldsymbol{\theta},\boldsymbol{\psi}) = \Lambda_n(\tilde{\mathbf{a}}) + o(1), \tag{3.10}$$

and for scale models, the minimax risk is of the form

$$\inf_{\psi} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(\tilde{\mathbf{a}},q_n)} R(\boldsymbol{\theta},\boldsymbol{\psi}) = 2\Lambda_n(\tilde{\mathbf{a}}) - 1 + o(1), \tag{3.11}$$

Note that for the normal means model with varying signal strength assumption as given in (3.9), we have $\Lambda_n(\tilde{\mathbf{a}}) = \frac{1}{q_n}\sum_{j=1}^{q_n}[1-\Phi(\tilde{a}_j - a_n^*)]$, where $a_n^* = \sqrt{2\log\left(\frac{n}{q_n}\right)}$. Similar to the signal strength as given in (3.7), in this case, too, they were able to establish that both the BH procedure (with a proper choice of α converging to zero) and the ℓ-value based procedure achieve the risk bound, i.e.

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(\tilde{\mathbf{a}},q_n)} R(\boldsymbol{\theta},\boldsymbol{\psi}) = \Lambda_n(\tilde{\mathbf{a}}) + o(1), \tag{3.12}$$

even if the level of sparsity is unknown.

Abraham et al. (2024) also studied the minimax risk corresponding to the classification (or Hamming) loss $L(\boldsymbol{\theta},\boldsymbol{\psi})$ and obtained the minimax risk corresponding to the classification loss, which is of the form

$$\frac{1}{q_n}\inf_{\psi}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_b} \mathbb{E}L(\boldsymbol{\theta},\boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.13}$$

where $\boldsymbol{\Theta}_b$ is already defined in (3.7) and b may be either finite or $b = b_n \to \pm\infty$. This result also holds for the BH and $\ell-$value-based procedures under the assumption of polynomial sparsity stated before.

As already discussed in Subsection 1.2.2, the only work based on one-group priors related to minimax risk in the normal means model is due to Salomond (2017). Under the assumption that observations are generated from a normal distribution as given in (1.1), they considered modeling each θ_i by a continuous prior instead of a two-group prior with the hierarchical formulation as

$$\theta_i|\sigma_i^2 \overset{ind}{\sim} \mathcal{N}(0,\sigma_i^2), \sigma_i^2 \overset{ind}{\sim} \pi(\sigma_i^2), \tag{3.14}$$

where $\pi(\cdot)$ denotes a density on the positive real numbers. Here σ_i^2 plays the dual role of capturing the overall sparsity of the model along with detecting the real signals. It is clear from (3.14) that all one-group global-local priors stated in (1.7) satisfying (1.10) can be included in this rich class of priors. This class of priors was previously studied by van der Pas et al. (2016), where they were interested in obtaining the optimal posterior contraction rate of the parameter vector around the truth. In this work, they were interested in testing $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously. Motivated by the existing works of Carvalho et al. (2009), Datta and Ghosh (2013), Ghosh et al. (2016) and Ghosh and Chakrabarti (2017), they proposed a multiple

testing rule $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$, with $\psi_i = \mathbf{1}(\mathbb{E}(\varphi_i|\mathbf{X} = \mathbf{x}) > \alpha)$ for some fixed threshold α, where $\varphi_i = \frac{\sigma_i^2}{1+\sigma_i^2}$. They remarked that for $\alpha = \frac{1}{2}$, their proposed testing rule becomes equivalent to that of Ghosh and Chakrabarti (2017). Motivated by work of Arias-Castro and Chen (2017), they also considered the problem of studying the minimax bound of a risk which is defined as the sum of FDR and FNR. Based on the conditions proposed by van der Pas et al. (2016) along with some additional assumptions on the level of sparsity, they obtained an upper bound of the minimax risk. Though they derived that the minimax risk based on their chosen class of priors converges to zero as the number of hypotheses n diverges to ∞, there are some gaps present in their work, as described in Subsection 1.2.2. This implies that whether a one-group prior can attain the minimax risk for the Gaussian location model was not answered in depth. In a nutshell, the study of minimax risk based on one-group priors needs to be revisited. In the next Section, we have made a modest attempt to provide positive answers to the questions raised in the last part of this Section based on one-group priors.

3.3 Results based on one-group priors

Motivated by the works of Datta and Ghosh (2013), Ghosh et al. (2016) and Ghosh and Chakrabarti (2017), in order to study the minimax risk of different loss functions based on one-group priors for the normal means problem, we also consider the decision rule $\boldsymbol{\psi} = (\psi_1, \psi_2, \cdots, \psi_n)$, where for $i = 1, 2, \cdots, n$,

$$\psi_i \equiv \psi_i(X_i) = \mathbf{1}\{\mathbb{E}(1 - \kappa_i|X_i, \tau) > \frac{1}{2}\}. \tag{3.15}$$

Based on the decision rule (3.15), we want to investigate whether any global-local prior of the form (1.7) satisfying (1.10) can attain the minimax risk either based on Hamming loss or based on the sum of FDR and FNR, irrespective of the level of sparsity being known or not. For the theoretical study of the decision rule (3.15), we assume the slowly varying function $L(\cdot)$ defined in (1.10) satisfies **Assumption 1.**

It is evident from earlier results of Datta and Ghosh (2013), van der Pas et al. (2014), Ghosh et al. (2016), Ghosh and Chakrabarti (2017) that the global shrinkage parameter τ plays an important role in capturing the underlying sparsity of the model. However, the sparsity level is usually unknown. Since the decision rule (3.15) also contains τ, it is of equal importance to propose a modified version of it when the level of sparsity is unknown. Using the empirical Bayes estimate of τ, proposed by van der Pas et al. (2014), the decision rule modifies to, for $i = 1, 2, \cdots, n$,

$$\psi_i \equiv \psi_i^{\text{EB}}(\mathbf{X}) = \mathbf{1}\{\mathbb{E}(1 - \kappa_i|X_i, \widehat{\tau}) > \frac{1}{2}\}, \tag{3.16}$$

where $\widehat{\tau}$ is defined in (2.4). An alternative approach is to model τ by an absolutely continuous prior $\pi(\tau)$ on it, known as the full Bayes approach. Motivated by Section 2.3 of Chapter 2, here, we consider a general class of priors $\pi(\cdot)$ satisfying (C4).

In this procedure, the decision rule is modified as, for $i = 1, 2, \cdots, n$,

$$\psi_i \equiv \psi_i^{\text{FB}}(\mathbf{X}) = \mathbf{1}\{\mathbb{E}(1-\kappa_i|\mathbf{X}) > \frac{1}{2}\}, \tag{3.17}$$

where the full Bayes posterior mean of $1-\kappa_i$ is defined as,

$$\mathbb{E}(1-\kappa_i|\mathbf{X}) = \int_{\frac{1}{n}}^{\alpha_n} \mathbb{E}(1-\kappa_i|\mathbf{X},\tau)\pi(\tau|\mathbf{X})d\tau = \int_{\frac{1}{n}}^{\alpha_n} \mathbb{E}(1-\kappa_i|X_i,\tau)\pi(\tau|\mathbf{X})d\tau.$$

We will study the optimality of the decision rules (3.15), (3.16), and (3.17) depending on the level of sparsity to be known or not, for both classification loss and the loss measured as the sum of FDR and FNR. Results of the minimax risk related to the Hamming or classification loss using one-group priors are discussed in Subsection 3.3.1. On the other hand, Subsection 3.3.2 contains results based on minimax risk corresponding to the loss defined as the sum of FDR and FNR using one-group shrinkage priors.

We end this Section with a Proposition based on a lower bound on the probability of type II error corresponding to the i^{th} hypothesis. This shows that why $a > \frac{1}{2}$ in (1.10) can not be used as a good choice for the broad class of one-group priors considered here.

Proposition 3.3.1. *Suppose $X_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1), i = 1, 2, \cdots, n$. We want to test hypotheses $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously based on the decision rule (3.15) using one-group priors. Assume that $\tau \to 0$ as $n \to \infty$ such that $\frac{n\tau}{q_n} \to C \in (0, \infty)$. Also assume that slowly varying function $L(\cdot)$ satisfies Assumption 1. Then for $a > \frac{1}{2}$ in (1.10), for $i = 1, \ldots, q_n$ and $b \in \mathbb{R}$, we have,*

$$\sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i} \to 1, \text{ as } n \to \infty, \tag{3.18}$$

where t_{2i} denotes the type II error corresponding to i^{th} hypothesis.

Observe that for Theorems 3.3.2, 3.3.4 and 3.3.6, the loss function $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ for which the minimax risk has to be obtained, depends on t_{1i} and t_{2i}, the type I and type II errors corresponding to the i^{th} hypothesis, respectively. Later, it will be clear from the proofs of these two Theorems that out of type I and type II errors, type II error is the dominating one. Hence, the lower bound on the type II error over the subset $\boldsymbol{\Theta}_b$ for $a > \frac{1}{2}$ implies the overall risk also exceeds the minimax risk, $1-\Phi(b)+o(1)$, in this case. On the other hand, when one is interested in studying the minimax risk based on the loss, defined as the sum of FDR and FNR, then also, the proofs of Theorems 3.3.7, 3.3.9 and 3.3.11 imply that those mainly depend on the FNR, which is as same as the type II error. Hence, the conclusion from Proposition 3.3.1 goes through here, too. As a consequence, Proposition 3.3.1 acts as a guideline for us regarding the choice of a given in (1.10) in the subsequent subsections in order to obtain the minimax risk.

3.3.1 Results related to the classification loss

Motivated by the result of Proposition 3.3.1, from this subsection, We are interested in the special case when $a = \frac{1}{2}$ in (1.10). This subclass of priors is named as *horseshoe-type* priors. Note that it is a rich subclass which contains the three parameter beta normal mixtures with $\alpha = 0.5$ and $\beta > 0$ (e.g. horseshoe, Strawderman-Berger), the generalized double Pareto priors with $\alpha = 1$ (e.g. standard double Pareto), and the inverse–gamma priors with $\alpha = 0.5$, etc.

As stated earlier, we are interested in studying the optimality of our decision rule (3.15) based on horseshoe-type priors in the sense that whether (3.15) can attain the minimax risk over the set $\boldsymbol{\Theta}_b$ or not. At first, we assume the level of sparsity is known and use the global parameter τ as a tuning one. We propose a condition on τ, depending on the level of sparsity. The first result of our work is as follows:

Theorem 3.3.2. *Suppose $X_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1), i = 1, 2, \cdots, n$. We want to test hypotheses $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously based on the decision rule (3.15) using one-group priors. Assume that $\tau \to 0$ as $n \to \infty$ such that $\frac{n\tau}{q_n} \to C \in (0, \infty)$. Also assume that slowly varying function $L(\cdot)$ satisfies Assumption 1. Then for horseshoe-type priors (i.e. for $a = \frac{1}{2}$), we have*

$$\frac{1}{q_n} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.19}$$

where $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ is same as defined in (3.5) and $\boldsymbol{\Theta}_b$ is defined in (3.7). This result holds for any $b \in \mathbb{R}$ or $b = b_n \to \infty$.

Remark 3.3.3. *The importance of Theorem 3.3.2 lies in proposing a decision rule based on one-group priors which attains the minimax risk corresponding to the classification or Hamming loss. To the best of our knowledge, it is the first result in the literature of global-local priors in this context. In order to establish this result, the only condition one needs on τ is that τ is of the same order as $\frac{q_n}{n}$, i.e., the proportion of non-zero means, asymptotically. The same condition was previously used by Ghosh et al. (2016) while they obtained the optimality of the same testing rule with respect to the Bayes risk for the normal means model assuming the observations are generated from a two-group spike and slab distribution. Note that, the result is established via providing bounds on the type I and type II errors. For type I error, Theorem 6 of Ghosh et al. (2016) comes in handy. However, the calculation for the type II error involves the use of some technical observations and many non-trivial arguments. In a nutshell, this result illustrates that, like the BH procedure of Benjamini and Hochberg (1995) and the ℓ-value approach based on two-group spike-and-slab priors, a decision rule (3.15) based on a broad class of global-local prior can also attain the minimax risk for the normal means model under the assumption that the sparsity level is known.*

The main drawback of Theorem 3.3.2 is that it assumes the sparsity of the underlying model to be known, which holds seldom. This motivates us to consider the adaptive decision rule (3.16) and study the same in terms of the minimax risk. Now consider the following Theorem.

Theorem 3.3.4. *Consider the setup of Theorem 3.3.2. Suppose we want to test hypotheses $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously based on the decision rule (3.16) using one-*

group priors, where $\widehat{\tau}$ is the same as (2.4). Assume that the number of non-zero means q_n satisfies $q_n \propto (\log n)^{\delta_2}$ for some $\delta_2 > 0$. Also assume that slowly varying function $L(\cdot)$ satisfies Assumption 1. Then for horseshoe-type priors (i.e. for $a = \frac{1}{2}$), we have

$$\frac{1}{q_n} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.20}$$

where $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ and Θ_b are same as defined in Theorem 3.3.2. This result holds for any $b \in \mathbb{R}$ or $b = b_n \to \infty$.

Remark 3.3.5. *The key importance of Theorem 3.3.4 lies in providing a positive answer to the question that the minimax risk based on classification loss can still be attained using the decision rule, even if the level of sparsity is unknown. Though the two results (Theorem 3.3.2 and 3.3.4) attain the same risk, but the calculations are quite different. Note that, due to the introduction of $\widehat{\tau}$ in place of τ as stated in (3.16), the decision rule corresponding to rejecting i^{th} null hypothesis depends on the entire dataset $\mathbf{X}$, not on X_i only, and also has a complicated form. Hence, in order to bypass the calculations based on $\mathbb{E}(1 - \kappa_i | X_i, \widehat{\tau})$ directly, we try to incorporate some salient features of $\mathbb{E}(1 - \kappa_i | x_i, \tau)$ for any fixed $x_i \in \mathbb{R}$. Observe that, for any fixed $x_i \in \mathbb{R}$, $\mathbb{E}(1 - \kappa_i | x_i, \tau)$ is non-decreasing in τ. Hence, for calculations corresponding to the type I error, first, we truncate $\widehat{\tau}$ in a range such that the monotonicity of τ can be used. In that range, after using monotonicity, one can employ the same arguments used in Theorem 1 for controlling type I error in that part. On the complementary part of the range, we need to provide nontrivial arguments by exploiting the structure of $\widehat{\tau}$.*

Theorem 3.3.6. *Consider the setup of Theorem 3.3.2. Suppose we want to test hypotheses $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously based on the decision rule (3.17) using one-group priors, where the prior on τ satisfies (C4) with $\alpha_n \propto \frac{(\log n)^{\delta_3}}{n}$ for some $0 < \delta_3 < \frac{1}{2}$. Assume that the number of non-zero means q_n satisfies $q_n \propto (\log n)^{\delta_2}$ for some $\delta_2 > 0$. Also assume that slowly varying function $L(\cdot)$ satisfies Assumption 1. Then for horseshoe-type priors (i.e. for $a = \frac{1}{2}$), we have*

$$\frac{1}{q_n} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.21}$$

where $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ and Θ_b are same as defined in Theorem 3.3.2. This result holds for any $b \in \mathbb{R}$ or $b = b_n \to \infty$.

3.3.2 Results related to minimax risk for FDR+FNR

As alluded to earlier, in this subsection, too, we confine our interest in studying whether the decision rules based on horseshoe-type priors (3.15), (3.16) and (3.17) can achieve the minimax risk, asymptotically, irrespective of the level of sparsity to be known or not. Similar to the previous subsection, here, too, at first we assume the level of sparsity to be known and try to model τ accordingly. Our result provides an affirmative conclusion that, similar to the BH and ℓ-value procedures, as stated by Abraham et al. (2024), the decision rule (3.15) can also attain

the minimax risk assuming the proportion of non-zero means to be known. The result is stated formally below.

Theorem 3.3.7. *Suppose* $X_i \overset{ind}{\sim} \mathcal{N}(\theta_i, 1), i = 1, 2, \cdots, n$. *We want to test hypotheses* $H_{0i} : \theta_i = 0$ *vs* $H_{1i} : \theta_i \neq 0$ *simultaneously based on the decision rule* (3.15) *using one-group priors. Assume that* $\tau \to 0$ *as* $n \to \infty$ *such that* $\frac{n\tau}{q_n} \to C \in (0, \infty)$. *Also assume that slowly varying function* $L(\cdot)$ *satisfies Assumption 1. Then for horseshoe-type priors (i.e. for* $a = \frac{1}{2}$*), we have*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} R(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.22}$$

where Θ_b *is same as defined in Theorem 3.3.2 and* $R(\boldsymbol{\theta}, \boldsymbol{\psi})$ *is defined in* (3.4). *This result continues to hold for any* $b \in \mathbb{R}$ *or* $b = b_n \to \infty$.

Remark 3.3.8. *Similar to the results of the previous subsection, this result also justifies the selection of the decision rule* (3.15) *based on one-group priors since it attains the minimax risk for the loss function defined as the sum of FDR and FNR. As stated earlier, the only work in the context of obtaining an upper bound to the minimax risk based on this loss induced by one-group priors was considered by Salomond (2017). However, their choice of the minimum value of the non-zero means exceeds the universal threshold. Not only this, they did not provide any answer regarding whether their proposed decision rule can attain the minimax risk. This result provides a positive answer to the question for the horseshoe-type priors corresponding to the minimax risk when the supremum of* $\boldsymbol{\theta}$ *is considered over a set that contains the universal threshold. An advantage of this result is that the upper bound can be made arbitrarily small by taking* $b = b_n \to \infty$. *The proof is done by providing bounds on FDR and FNR. The truncation used in bounding FDR is inspired by their approach. However, we have to do delicate calculations after that. We need to use Hoeffding's inequality to bound some quantity related to FDR. The technique used in providing a bound of FNR is quite nontrivial from that of Salomond (2017). In that context, bound on type II error, as derived in previous results, comes in useful. Our proof also demonstrates that some calculations used in Salomond (2017) can be done using a simpler approach, too.*

Given that (3.15) achieves the minimax risk, it is natural to ask whether the adaptive decision rule (3.16), too, can provide the same result, when the number of non-zero means is unknown. The following result answers the question in an affirmative sense.

Theorem 3.3.9. *Consider the setup of Theorem 3.3.2. Suppose we want to test hypotheses* $H_{0i} : \theta_i = 0$ *vs* $H_{1i} : \theta_i \neq 0$ *simultaneously based on the decision rule* (3.16) *using one-group priors, where* $\widehat{\tau}$ *is the same as* (2.4). *Assume that the number of non-zero means* q_n *satisfies* $q_n \propto (\log n)^{\delta_2}$ *for some* $\delta_2 > 0$. *Also assume that slowly varying function* $L(\cdot)$ *satisfies Assumption 1. Then for horseshoe-type priors (i.e. for* $a = \frac{1}{2}$*), we have*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} R(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.23}$$

where Θ_b *is same as defined in Theorem 3.3.2 and* $R(\boldsymbol{\theta}, \boldsymbol{\psi})$ *is defined in* (3.4). *This result holds for any* $b \in \mathbb{R}$ *or* $b = b_n \to \infty$.

Remark 3.3.10. *This result elucidates that, like the BH procedure of Benjamini and Hochberg (1995) and the ℓ-value approach based on two-group spike-and-slab priors, a decision rule (3.16) based on a broad class of global-local prior can also attain the minimax risk for the normal means model, even if the sparsity level is unknown. This type of result was earlier considered by Salomond (2017) for a broader class of priors, which contains our chosen class of priors. They proved that the upper bound of the minimax risk for their proposed decision rule converges to 0 as $n \to \infty$. However, their work did not provide any answer to the question of whether their decision rule can attain the minimax risk or not. On the contrary, our result does provide a positive answer to that question. As a byproduct of our result, the minimax risk can be made arbitrarily small by choosing $b = b_n \to \infty$, even for an unknown sparse situation. In this way, their result becomes a special case of ours under our chosen class of priors under study.*

Theorem 3.3.11. *Consider the setup of Theorem 3.3.2. Suppose we want to test hypotheses $H_{0i} : \theta_i = 0$ vs $H_{1i} : \theta_i \neq 0$ simultaneously based on the decision rule (3.17) using one-group priors, where the prior on τ satisfies (C4) with $\alpha_n \propto \frac{(\log n)^{\delta_3}}{n}$ for some $0 < \delta_3 < \frac{1}{2}$. Assume that the number of non-zero means q_n satisfies $q_n \propto (\log n)^{\delta_2}$ for some $\delta_2 > 0$. Also assume that slowly varying function $L(\cdot)$ satisfies Assumption 1. Then for horseshoe-type priors (i.e. for $a = \frac{1}{2}$), we have*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_b} R(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1 - \Phi(b) + o(1), \tag{3.24}$$

where Θ_b is same as defined in Theorem 3.3.2 and $R(\boldsymbol{\theta}, \boldsymbol{\psi})$ is defined in (3.4). This result holds for any $b \in \mathbb{R}$ or $b = b_n \to \infty$.

3.4 Concluding remarks and scope for future work

In the context of multiple hypothesis testing, it is of interest to control the sum of FDR+FNR or the misclassification risk. There has been interest in finding procedures that control these risks at the optimum level with respect to the famous minimax criterion. For the normal means model, Abraham et al. (2024) obtained the minimax risk for both these losses and established that the BH procedure of Benjamini and Hochberg (1995) and the ℓ-value based procedure using spike and slab prior attain the risk, even if the sparsity pattern is unknown. This raises the question whether any decision rule based on continuous one-group priors can attain the minimax risk. In this work, we provide a positive answer to this question by considering a broad class of one-group priors both when the global shrinkage parameter is used as a tuning one or it is treated in an empirical Bayes or a full Bayes approach. To the best of our knowledge, ours is the first study of minimax optimality in the context of using one-group priors. Our results demonstrate the crucial fact that a carefully chosen decision rule based on appropriate class of one-group priors can also attain the minimax risk, even if the level of sparsity is unknown.

Though our results show that our chosen general class of continuous priors can be considered as an alternative to the 'gold standard' spike-and-slab priors for sparse problem, there are still some questions that need to be addressed from a decision theoretic viewpoint. First, when the

level of sparsity is unknown, a more natural estimate of τ is the MMLE considered by van der Pas et al. (2017a), defined as the maximizer of the marginal likelihood function of $\mathbf{X}$ given τ. So, it begs the question whether the decision rule (3.16) using the MMLE of τ can still attain the minimax risk for the two loss functions considered here. Hence, one might be interested in studying the optimality of such a procedure in this context.

Another possible extension of our work is studying the class of priors proposed by van der Pas et al. (2016). In this context, first, it is needed to figure out whether any multiple testing rule can attain the minimax risk, asymptotically, under the beta-min condition of Abraham et al. (2024), when the level of sparsity is assumed to be known. The question regarding the optimality of a testing procedure for an unknown level of sparsity is also open till now. All of these can be considered as some future problems of interest.

3.5 Proofs

Proof of Proposition 3.3.1. Let us fix any $i \geq 1$. We consider the cases $a \in (\frac{1}{2}, 1)$ and $a \geq 1$, separately.

Case 1: Let us first consider the case when $a \in (\frac{1}{2}, 1)$. Using Theorem 4 of Ghosh et al. (2016), and the fact that the slowly varying function $L(\cdot)$ is bounded above by some constant M (> 0), we have

$$E(1 - \kappa_i \big| X_i, \tau) \leq K_1 e^{\frac{X_i^2}{2}} \tau^{2a} (1 + o(1)), \tag{3.25}$$

where the $o(1)$ term above is independent of both the index i and the data point X_i, but depends on τ in such a way that $\lim_{\tau \to 0} o(1) = 0$. Now, applying the definition of type II error coupled with (3.25), the type II error rate for the i^{th} hypothesis testing problem is given by

$$\begin{aligned}
t_{2i} &= \mathbb{P}_{\theta_i \neq 0} \left(\mathbb{E}(1 - \kappa_i | X_i, \tau) \leq \frac{1}{2} \right) \\
&\geq \mathbb{P}_{\theta_i \neq 0} \left(K_1 \tau^{2a} \exp\left(\frac{X_i^2}{2} \right) (1 + o(1)) \leq \frac{1}{2} \right) \\
&= \mathbb{P}_{\theta_i \neq 0} \left(X_i^2 \leq 4a \log\left(\frac{1}{\tau} \right) (1 + o(1)) \right) \\
&= \Phi\left(\sqrt{4a \log\left(\frac{1}{\tau} \right)} (1 + o(1) - \theta_i \right) - \Phi\left(-\sqrt{4a \log\left(\frac{1}{\tau} \right)} (1 + o(1) - \theta_i \right).
\end{aligned} \tag{3.26}$$

Note that, to provide a lower bound to the type II error rate on the set $\mathbf{\Theta}_b$, it would be enough to consider only one candidate of θ_i from that set and show that the lower bound of (3.26) for that choice of θ_i tends to 1 as $n \to \infty$. Let us fix any $b \in \mathbb{R}$ and choose $\theta_i = \sqrt{2 \log\left(\frac{n}{q_n} \right)}$.

Then under the assumption that $\frac{n\tau}{q_n} \to C \in (0,\infty)$ as $n \to \infty$, we obtain

$$\log\left(\frac{1}{\tau}\right) = \log\left(\frac{n}{q_n}\right) + \log\left(\frac{1}{C}\right) + \log(1+o(1)) = \log\left(\frac{n}{q_n}\right)(1+o(1)),$$

where the $o(1)$ term above is independent of both the index i and the data point X_i, but depends on τ in such a way that $\lim_{\tau\to 0} o(1) = 0$. This implies that for all sufficiently large n, we have

$$\begin{aligned}
&\sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i} \\
&\geq \Phi\left((\sqrt{4a}-\sqrt{2})\log\left(\frac{n}{q_n}\right)\right)(1+o(1)) - \Phi\left((-\sqrt{4a}-\sqrt{2})\log\left(\frac{n}{q_n}\right)\right)(1+o(1)) \\
&= \Phi\left((\sqrt{4a}-\sqrt{2})\log\left(\frac{n}{q_n}\right)\right)(1+o(1)) \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

In the above chain of inequalities, observe that for any fixed $a \in (\frac{1}{2}, 1)$,
$\Phi\left((\sqrt{4a}-\sqrt{2})\log\left(\frac{n}{q_n}\right)\right)(1+o(1)) \to 1$ as $n \to \infty$ and $\Phi\left((-\sqrt{4a}-\sqrt{2})\log\left(\frac{n}{q_n}\right)\right)(1+o(1)) \to 0$ as $n \to \infty$.

Case 2: Next, we consider the case where $a \geq 1$. First, we derive an upper bound to $E(1-\kappa_i|X_i,\tau)$ for any fixed $\tau \in (0,1)$ and $X_i \in \mathbb{R}$.

Observe that, the posterior distribution of κ_i given X_i and τ is,

$$\pi(\kappa_i|X_i,\tau) \propto \kappa_i^{a-\frac{1}{2}}(1-\kappa_i)^{-a-1} L\left(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\right)\exp\left\{\frac{(1-\kappa)X_i^2}{2}\right\}, \text{ for } 0<\kappa_i<1.$$

Using the transformation $t = \frac{1}{\tau^2}(\frac{1}{\kappa}-1)$, $E(1-\kappa_i|X_i,\tau)$ becomes

$$E(1-\kappa_i|X_i,\tau) = \frac{\tau^2 \int_0^\infty (1+t\tau^2)^{-\frac{3}{2}} t^{-a} L(t)\exp\left(\frac{X_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}\right)dt}{\int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t)\exp\left(\frac{X_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}\right)dt}.$$

Note that,

$$\begin{aligned}
\int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t)\exp\left(\frac{X_i^2}{2}\cdot\frac{t\tau^2}{1+t\tau^2}\right)dt &\geq \int_0^\infty (1+t\tau^2)^{-\frac{1}{2}} t^{-a-1} L(t)dt \\
&= K^{-1}(1+o(1)),
\end{aligned}$$

where $o(1)$ depends only on τ and tends to zero as $\tau \to 0$. The equality in the last line follows due to the Dominated Convergence Theorem. Hence

$$E(1-\kappa_i|X_i,\tau) \leq K(A_1+A_2)(1+o(1)), \tag{3.27}$$

where

$$A_1 = \int_0^1 \frac{t\tau^2}{1+t\tau^2} \cdot \frac{1}{\sqrt{1+t\tau^2}} t^{-a-1} L(t) \exp\left(\frac{X_i^2}{2} \cdot \frac{t\tau^2}{1+t\tau^2}\right) dt.$$

and $A_2 = \int_1^\infty \frac{t\tau^2}{1+t\tau^2} \cdot \frac{1}{\sqrt{1+t\tau^2}} t^{-a-1} L(t) \exp\left(\frac{X_i^2}{2} \cdot \frac{t\tau^2}{1+t\tau^2}\right) dt$. Since, for any $t \leq 1$ and $\tau \leq 1$, $\frac{t\tau^2}{1+t\tau^2} \leq \frac{1}{2}$, using the fact $\int_0^\infty t^{-a-1} L(t) dt = K^{-1}$,

$$A_1 \leq K^{-1} \tau^2 \exp\left(\frac{X_i^2}{4}\right). \tag{3.28}$$

For providing an upper bound on A_2, we divide our calculations depending on $a = 1$ and $a > 1$. Next, note that, $1 + t\tau^2 \geq t\tau^2$ for any $t > 0$ and $t\tau^2 \geq \sqrt{t}$ if and only if $t \geq \frac{1}{\tau^4}$. As a result, for $a = 1$, we have

$$\int_1^{\frac{1}{\tau^4}} \frac{t\tau^2}{1+t\tau^2} \cdot t^{-2} L(t) dt \leq M\tau^2 \int_1^{\frac{1}{\tau^4}} t^{-1} dt = M\tau^2 \log\left(\frac{1}{\tau^4}\right), \tag{3.29}$$

and

$$\int_{\frac{1}{\tau^4}}^\infty \frac{t\tau^2}{1+t\tau^2} \cdot t^{-2} L(t) dt \leq M\tau^2 \int_{\frac{1}{\tau^4}}^\infty t^{-\frac{3}{2}} dt = 2M\tau^4. \tag{3.30}$$

Hence, combining (3.29) and (3.30), for $a = 1$, we obtain

$$A_2 \leq 2M \exp\left(\frac{X_i^2}{2}\right) \tau^2 [\log\left(\frac{1}{\tau^4}\right) + \tau^2] = 8M \exp\left(\frac{X_i^2}{2}\right) \tau^2 \log\left(\frac{1}{\tau}\right)(1 + o(1)). \tag{3.31}$$

On the other hand, for $a > 1$,

$$\int_1^\infty \frac{t\tau^2}{1+t\tau^2} \cdot t^{-a-1} L(t) dt \leq M\tau^2 \int_1^\infty t^{-a} dt = \left(\frac{M}{(a-1)}\right) \tau^2.$$

As a result, for $a > 1$,

$$A_2 \leq \left(\frac{M}{(a-1)}\right) \tau^2 \exp\left(\frac{X_i^2}{2}\right). \tag{3.32}$$

Therefore, combining (3.31) and (3.32), for $a \geq 1$, we have

$$A_2 \leq K_1 \exp\left(\frac{X_i^2}{2}\right) \tau^2 \log\left(\frac{1}{\tau}\right)(1 + o(1)), \tag{3.33}$$

where K_1 is a constant depending on M and a. Now, combining (3.27), (3.28) and (3.33), we obtain, for $a \geq 1$,

$$E(1 - \kappa_i | X_i, \tau) \leq K_2 \exp\left(\frac{X_i^2}{2}\right) \tau^2 \log\left(\frac{1}{\tau}\right)(1 + o(1)), \tag{3.34}$$

where K_2 is a constant depending on K^{-1} and K_1. Next, again using the same technique used

when $a \in (\frac{1}{2}, 1)$ with (3.34), a lower bound of t_{2i} is obtained as,

$$\begin{aligned}
t_{2i} &= \mathbb{P}_{\theta_i \neq 0}(\mathbb{E}(\kappa_i | X_i, \tau) > \frac{1}{2}) \\
&\geq \mathbb{P}_{\theta_i \neq 0}(K_2 \exp\left(\frac{X_i^2}{2}\right)\tau^2 \log\left(\frac{1}{\tau}\right)(1 + o(1)) \leq \frac{1}{2}) \\
&= \mathbb{P}_{\theta_i \neq 0}(X_i^2 \leq 4\log\left(\frac{1}{\tau}\right) - 4\log\log\left(\frac{1}{\tau}\right) - o(1) - 2\log 2) \\
&= \mathbb{P}_{\theta_i \neq 0}(X_i^2 \leq 4\log\left(\frac{1}{\tau}\right)(1 + o(1))) \\
&= \Phi(\sqrt{4\log\left(\frac{1}{\tau}\right)}(1 + o(1) - \theta_i) - \Phi(-\sqrt{4\log\left(\frac{1}{\tau}\right)}(1 + o(1) - \theta_i). \qquad (3.35)
\end{aligned}$$

Next, again using the arguments used in the previous case, we have, for all sufficiently large n,

$$\begin{aligned}
\sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b} t_{2i} &\geq \Phi((\sqrt{4} - \sqrt{2})\log\left(\frac{n}{q_n}\right))(1 + o(1)) - \Phi((-\sqrt{4} - \sqrt{2})\log\left(\frac{n}{q_n}\right))(1 + o(1)) \\
&= \Phi((\sqrt{4} - \sqrt{2})\log\left(\frac{n}{q_n}\right))(1 + o(1)) \\
&\to 1, \text{ as } n \to \infty.
\end{aligned}$$

□

Proof of Theorem 3.3.2. The desired lower bound to the left-hand side of (3.19) follows from Theorem 7 of Abraham et al. (2024). Hence, we only need to obtain an upper bound for it. Using arguments similar to Theorem 7 of Abraham et al. (2024), note that

$$\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i:\theta_i = 0} t_{1i} + \sum_{i:\theta_i \neq 0} t_{2i}, \qquad (3.36)$$

where t_{1i} and t_{2i} denote the type I and type II error rates corresponding to the i^{th} hypothesis testing problem, respectively. Hence, it is enough to show that

$$\frac{1}{q_n} \sum_{i:\theta_i = 0} t_{1i} = o(1), \qquad (3.37)$$

and

$$\frac{1}{q_n} \sum_{i:\theta_i \neq 0} \sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b} t_{2i} \leq 1 - \Phi(b) + o(1). \qquad (3.38)$$

First, we prove (3.37). Using the architecture of the proof of Theorem 6 of Ghosh et al. (2016)

with $a = \frac{1}{2}$, we have

$$t_{1i} \leq \tilde{K}_1 \frac{\tau L\left(\frac{1}{\tau^2}\right)}{\sqrt{\log\left(\frac{1}{\tau^2}\right)}}(1 + o(1)), \tag{3.39}$$

where $\tilde{K}_1$ is a positive constant that is independent of τ and i, and the $o(1)$ term is such that $\lim_{n\to\infty} o(1) = 0$ and is independent of any i. Hence, using (3.38) and under the assumption Assumption 1 on the slowly varying function $L(\cdot)$, we get

$$\begin{aligned}\frac{1}{q_n}\sum_{i:\theta_i=0} t_{1i} &\leq K_1 M \frac{(n-q_n)}{q_n}\frac{\tau}{\sqrt{\log\left(\frac{1}{\tau^2}\right)}}(1+o(1)) \\ &\leq K_1 M \frac{n\tau}{q_n}\frac{1}{\sqrt{\log\left(\frac{1}{\tau^2}\right)}}(1+o(1)) \\ &\to 0, \text{ as } n \to \infty.\end{aligned}$$

In the last line of the above chain of inequalities, we used the facts $\frac{n\tau}{q_n} \to C \in (0,\infty)$ as $n \to \infty$, $\tau \to 0$ as $n \to \infty$, and $\log\left(\frac{1}{\tau^2}\right) \to \infty$ as $\tau \to 0$. This completes the proof of (3.37).

Our next goal is to provide an upper bound to t_{2i} that holds uniformly for each $i = 1, \ldots, n$. Towards that, we first use Lemma 3 of Ghosh and Chakrabarti (2017) that asserts that, for each fixed $x \in \mathbb{R}$ and any fixed $0 < \tau < 1$, the posterior shrinkage coefficient $\mathbb{E}(\kappa|x,\tau)$ can be bounded above by a non-negative, continuously decreasing, and measurable function $g(x,\tau,\eta,\delta)$ given by

$$g(x,\tau,\eta,\delta) = g_1(x,\tau,\eta,\delta) + g_2(x,\tau,\eta,\delta),$$

where

$$g_1(x,\tau,\eta,\delta) = C_1\left[x^2\int_0^{\frac{x^2}{1+t_0}} \exp\left(-\frac{u}{2}\right)du\right]^{-1} = C_1\left[x^2\left\{1-\exp\left(-\frac{x^2}{2(1+t_0)}\right)\right\}^{-1}\right],$$

and

$$g_2(x,\tau,\eta,\delta) = C_2\tau^{-1}\exp\left(-\frac{\eta(1-\delta)}{2}x^2\right),$$

for any given $\eta, \delta \in (0,1)$. Here, C_1 and C_2 are two finite positive constants that do not depend on both x and τ (but they do depend on η and δ), and t_0 was previously defined in Assumption 1. In addition, the above nonnegative function $g(x,\tau,\eta,\delta)$ satisfies the following:

For any $\rho > \frac{2}{\eta(1-\delta)}$,

$$\lim_{\tau\to 0}\sup_{|x|>\sqrt{\rho\log\left(\frac{1}{\tau}\right)}} g(x,\tau,\eta,\delta) = 0.$$

Let us fix any $\eta \in (0,1)$ and $\delta \in (0,1)$. Choose $\rho = \frac{2}{\eta(1-\delta)}[1+\tilde{g}(\tau)]$ where $\tilde{g}(\tau) > 0$ and tends to zero as $\tau \to 0$ such that $\tau^{\tilde{g}(\tau)} \to 0$ as $\tau \to 0$. One possible choice of $\tilde{g}(\tau) = (\log\left(\frac{1}{\tau}\right))^{-\delta_1}$ for some $0 < \delta_1 < 1$. Later, we may choose a more specific choice of $\tilde{g}(\tau)$, as required.

Next, we define two events B_n and C_n as $B_n = \{g(X_i,\tau,\eta,\delta) > \frac{1}{2}\}$ and $C_n = \{|X_i| > \sqrt{\rho\log\left(\frac{1}{\tau}\right)}\}$ with ρ as defined before. Then, using the preceding observations, the type II error rate for the

i^{th} hypothesis testing problem is given by

$$\begin{aligned} t_{2i} &= P_{\theta_i \neq 0}(E(\kappa_i | X_i, \tau) > \frac{1}{2}) \\ &\leq P_{\theta_i \neq 0}(B_n) \\ &\leq P_{\theta_i \neq 0}(B_n \cap C_n) + P_{\theta_i}(C_n^c). \end{aligned} \tag{3.40}$$

Observe that, for any $\theta_i \neq 0$,

$$\begin{aligned} & P_{\theta_i \neq 0}(B_n \cap C_n) \\ & P_{\theta_i \neq 0}\left(g(X_i, \tau, \eta, \delta) > \frac{1}{2}, |X_i| > \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right) \\ & \leq P_{\theta_i \neq 0}\left(g_1(X_i, \tau, \eta, \delta) > \frac{1}{4}, |X_i| > \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right) + \\ & P_{\theta_i \neq 0}\left(g_2(X_i, \tau, \eta, \delta) > \frac{1}{4}, |X_i| > \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right). \end{aligned} \tag{3.41}$$

Since, for given τ, η, δ, the function $g_1(x, \tau, \eta, \delta)$ steadily decreases in x^2, for any $|x| > \sqrt{\rho \log(\frac{1}{\tau})}$, we have

$$g_1(x, \tau, \eta, \delta) \geq \frac{C_1}{\rho}\left[\log\left(\frac{1}{\tau}\right)\left\{1 - \tau^{\frac{\rho}{2(1+t_0)}}\right\}\right]^{-1} \to 0 \text{ as } \tau \to 0.$$

Since $\tau \to 0$ as $n \to \infty$, we have, $g_1(X_i, \tau, \eta, \delta) \leq \frac{1}{4}$ with probability 1 over the set C_n for all sufficiently large n, and this would be true for any fixed $\eta \in (0,1)$ and $\delta \in (0,1)$, and any $\theta_i \neq 0$. Therefore,

$$P_{\theta_i \neq 0}\left(g_1(X_i, \tau, \eta, \delta) > \frac{1}{4} \mid |X_i| > \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right) = 0, \text{ for all sufficiently large } n. \tag{3.42}$$

Again, we observe that, for given τ, η, δ, the function $g_2(x, \tau, \eta, \delta)$ steadily decreases in x^2. Therefore, using a similar argument, it can be shown easily that,

$$P_{\theta_i \neq 0}\left(g_2(X_i, \tau, \eta, \delta) > \frac{1}{4} \mid |X_i| > \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right) = 0, \text{ for all sufficiently large } n. \tag{3.43}$$

Combining (3.41), (3.42) and (3.43), it therefore follows that

$$P_{\theta_i \neq 0}(B_n \cap C_n) = 0, \text{ for all sufficiently large } n. \tag{3.44}$$

(3.40) and (3.44) together imply that, for all sufficiently large n, we have

$$\begin{aligned} t_{2i} &\leq P_{\theta_i \neq 0}\left(|X_i| \leq \sqrt{\rho \log\left(\frac{1}{\tau}\right)} \right) \\ &= \Phi(T_n - \theta_i) - \Phi(-T_n - \theta_i), \end{aligned} \tag{3.45}$$

where $T_n = \sqrt{\rho \log\left(\frac{1}{\tau}\right)}$. Next, we divide our calculations depending on whether θ_i is positive or not.

Case: Let us assume $\theta_i > 0$, that is, $\theta_i > \sqrt{2\log\left(\frac{n}{q_n}\right)} + b$. Then, using the lower bound on θ_i coupled with (3.45), the type II error rate t_{2i} can be bounded above as

$$\begin{aligned} t_{2i} &\leq \Phi(T_n - \theta_i) \\ &\leq \Phi\left(\sqrt{\rho \log\left(\frac{1}{\tau}\right)} - \sqrt{2\log\left(\frac{n}{q_n}\right)} - b\right), \end{aligned}$$

where the choice of ρ is same as discussed before. Observe that the decision rule does not depend on how $\eta \in (0,1), \delta \in (0,1)$ and $\rho > \frac{2}{\eta(1-\delta)}$ is chosen. Hence, taking infimum over all such ρ's and subsequently over all possible choices of $(\eta, \delta) \in (0,1) \times (0,1)$, and finally using the continuity of $\Phi(\cdot)$, we have

$$t_{2i} \leq \Phi\left(\sqrt{2(1+g(\tau))\log\left(\frac{1}{\tau}\right)} - \sqrt{2\log\left(\frac{n}{q_n}\right)} - b\right). \tag{3.46}$$

Next, under the assumption $\frac{n\tau}{q_n} \to C \in (0,\infty)$, we obtain

$$\log\left(\frac{1}{\tau}\right) = \log\left(\frac{n}{q_n}\right) + \log\left(\frac{1}{C}\right) + \log(1+o(1)) \tag{3.47}$$

and

$$g(\tau)\log\left(\frac{1}{\tau}\right) = \left(\log\left(\frac{n}{q_n}\right)\right)^{1-\delta_1}(1+o(1)). \tag{3.48}$$

Combining all these observations, we have

$$t_{2i} \leq \Phi(\sqrt{a_n + d_n} - \sqrt{a_n} - b), \tag{3.49}$$

where $a_n = 2\log\left(\frac{n}{q_n}\right)$ and $d_n = 2[\log\left(\frac{1}{C}\right) + \log(1+o(1))] + 2\left(\log\left(\frac{n}{q_n}\right)\right)^{1-\delta_1}(1+o(1))$. Now, note that

$$\begin{aligned} &\Phi(\sqrt{a_n+d_n} - \sqrt{a_n} - b) - \Phi(-b) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-b}^{\sqrt{a_n+d_n}-\sqrt{a_n}-b} \exp\left(-\frac{x^2}{2}\right) dx \\ &\leq \sqrt{a_n+d_n} - \sqrt{a_n} \\ &= \frac{(\sqrt{a_n+d_n} - \sqrt{a_n})(\sqrt{a_n+d_n} + \sqrt{a_n})}{\sqrt{a_n+d_n} + \sqrt{a_n}} \\ &= \frac{d_n}{\sqrt{a_n+d_n} + \sqrt{a_n}} \\ &\leq \frac{d_n}{2\sqrt{a_n}}. \end{aligned} \tag{3.50}$$

Note that for $d_n = (\log\left(\frac{n}{q_n}\right))^{1-\delta_1}(1+o(1))$, $\frac{1}{2} < \delta_1 < 1$, we have $d_n = o(\sqrt{a_n})$, as $n \to \infty$. As a result, we obtain

$$\begin{aligned} t_{2i} &\leq \Phi(\sqrt{a_n + d_n} - \sqrt{a_n} - b) \\ &= \Phi(-b) + \left[\Phi(\sqrt{a_n + d_n} - \sqrt{a_n} - b) - \Phi(-b)\right] \\ &= \Phi(-b) + o(1). \end{aligned} \tag{3.51}$$

Note that, this $o(1)$ is independent of i.

Case 2: Let us assume that $\theta_i < 0$, that is, $\theta_i < -\sqrt{2\log\left(\frac{n}{q_n}\right)} - b$. Then, using the lower bound on $-\theta_i$ the upper bound of t_{2i} obtained in (3.45) modifies to

$$\begin{aligned} t_{2i} &\leq 1 - \Phi(-T_n - \theta_i) \\ &\leq 1 - \Phi\left(-\sqrt{\rho\log\left(\frac{1}{\tau}\right)} + \sqrt{2\log\left(\frac{n}{q_n}\right)} + b\right). \end{aligned} \tag{3.52}$$

Again using arguments similar to those of (3.46)-(3.49), in this case, an upper bound on the probability of type II error is obtained as

$$t_{2i} \leq 1 - \Phi(-\sqrt{a_n + d_n} + \sqrt{a_n} + b), \tag{3.53}$$

where a_n and d_n are same as before. Next, again using the same arguments as used in (3.50) helps to establish that

$$\Phi(-\sqrt{a_n + d_n} + \sqrt{a_n} + b) = \Phi(b) + o(1), \tag{3.54}$$

where $o(1)$ is independent of i. Now, combining (3.52)-(3.54),

$$t_{2i} \leq 1 - \Phi(b) + o(1). \tag{3.55}$$

Finally (3.51) and (3.55) imply, for each i with $|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b$,

$$t_{2i} \leq 1 - \Phi(b) + o(1). \tag{3.56}$$

This also ensures the upper bound of the left-hand side of (3.38) is of the order of $1-\Phi(b)+o(1)$ and completes the proof of Theorem 3.3.2. □

Proof of Theorem 3.3.4. Let t_{1i}^{EB} and t_{2i}^{EB} denote the probabilities of the type I and type II errors for the empirical Bayes decision rule corresponding to the i^{th} hypothesis, respectively. Then the overall risk associated with the empirical Bayes decision rule is given by

$$\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i:\theta_i=0} t_{1i}^{EB} + \sum_{i:\theta_i\neq 0} t_{2i}^{EB}, \tag{3.57}$$

Similar to Theorem 3.3.2, here too, we need to obtain an upper bound of the left hand side of (3.20). Hence, it suffices to prove that

$$\sum_{i:\theta_i=0} t_{1i}^{EB} = o(q_n) \tag{3.58}$$

and

$$\frac{1}{q_n} \sum_{i:\theta_i \neq 0} \sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i}^{EB} \leq 1 - \Phi(b) + o(1). \tag{3.59}$$

For any $\alpha_n > 0$, we have

$$\begin{aligned} t_{1i}^{EB} &= \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}) \\ &= \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, \widehat{\tau} > 2\alpha_n) + \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, \widehat{\tau} \leq 2\alpha_n) \\ &= u_{1n} + u_{2n}, \text{ say,} \end{aligned} \tag{3.60}$$

where $u_{1n} = \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, \widehat{\tau} > 2\alpha_n)$ and $u_{2n} = \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, \widehat{\tau} \leq 2\alpha_n)$. Next, we observe that, for any given $x \in \mathbb{R}$, $\mathbb{E}(1-\kappa|x,\tau)$ is non-decreasing in τ. Hence, using (3.39), we have

$$\begin{aligned} u_{2n} &= \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}, \widehat{\tau} \leq 2\alpha_n) \\ &= \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2} \mid \widehat{\tau} \leq 2\alpha_n)\mathbb{P}_{\theta_i=0}(\widehat{\tau} \leq 2\alpha_n) \\ &\leq \mathbb{P}_{\theta_i=0}(\mathbb{E}(1-\kappa_i|X_i,2\alpha_n) > \frac{1}{2}) \\ &\leq K_1 M \frac{2\alpha_n}{\sqrt{\log\left(\frac{1}{4\alpha_n^2}\right)}}(1+o(1)). \end{aligned} \tag{3.61}$$

Let us define, $\widehat{\tau}_1 = \frac{1}{n}$ and $\widehat{\tau}_2 = \frac{1}{c_2 n}\sum_{i=1}^n \mathbf{1}(|X_i| > \sqrt{c_1 \log n})$, with $c_1 \geq 2$ and $c_2 \geq 1$. Then, $\widehat{\tau} = \max\{\widehat{\tau}_1, \widehat{\tau}_2\}$. Based on this observation, and using the definition of u_{1n}, we obtain

$$\begin{aligned} u_{1n} &\leq \mathbb{P}_{\theta_i=0}(\widehat{\tau} > 2\alpha_n) \\ &\leq \mathbb{P}_{\theta_i=0}(\widehat{\tau}_1 > 2\alpha_n) + \mathbb{P}_{\theta_i=0}(\widehat{\tau}_2 > 2\alpha_n). \end{aligned} \tag{3.62}$$

Let us choose $\alpha_n > 0$ such that $\widehat{\tau}_1 = \dfrac{1}{n} \leq 2\alpha_n$ for all sufficiently large n, so that $\mathbb{P}_{\theta_i=0}(\widehat{\tau}_1 > 2\alpha_n) = 0$, provided n is sufficiently large. Hence, (3.62) implies, for all sufficiently large n

$$\begin{aligned} u_{1n} &\leq \mathbb{P}_{\theta_i=0}(\widehat{\tau}_2 > 2\alpha_n) \\ &\leq \mathbb{P}_{\theta_i=0}(\widehat{\tau}_3 > \alpha_n) + \mathbb{P}_{\theta_i=0}(\widehat{\tau}_4 > \alpha_n), \end{aligned} \tag{3.63}$$

where $\widehat{\tau}_3 = \frac{1}{c_2 n} \sum_{i:\theta_i \neq 0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n})$ and $\widehat{\tau}_4 = \frac{1}{c_2 n} \sum_{i:\theta_i = 0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n})$. Observe that,

$$\widehat{\tau}_3 \leq \frac{1}{c_2} \frac{q_n}{n}.$$

Let us choose $\alpha_n > 0$ such that $\alpha_n \geq \frac{1}{c_2} \frac{q_n}{n}$ for all sufficiently large n, whence $\mathbb{P}_{\theta_i=0}(\widehat{\tau}_3 > \alpha_n) = 0$.

Therefore, for all sufficiently large n, we obtain

$$\begin{aligned}
u_{1n} &\leq \mathbb{P}_{\theta_i=0}(\widehat{\tau}_4 > \alpha_n) \\
&= \mathbb{P}_{\theta_i=0}\left(\frac{1}{n - q_n} \sum_{i:\theta_i=0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n}) > c_2 \frac{n\alpha_n}{n - q_n} \right) \\
&\leq \mathbb{P}_{\theta_i=0}\left(\frac{1}{n - q_n} \sum_{i:\theta_i=0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n}) > \alpha_n \right) \\
&= \mathbb{P}_{\theta_i=0}\left(\frac{1}{n - q_n} \sum_{i:\theta_i=0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n}) > p_n + \epsilon_n \right),
\end{aligned} \tag{3.64}$$

where $p_n = 2\mathbb{P}_{\theta_i=0}\left(|X_i| > \sqrt{c_1 \log n}\right)$ and $\epsilon_n = \alpha_n - p_n$.
Using Mill's ratio, we obtain,

$$p_n = 2\mathbb{P}_{\theta_i=0}\left(X_i > \sqrt{c_1 \log n}\right) \propto \frac{n^{-c_1/2}}{\sqrt{\log n}}.$$

Again, by our choice of α_n, we have

$$\epsilon_n \geq \frac{1}{c_2} \frac{q_n}{n} \left(1 - \frac{\widetilde{C}}{\sqrt{\log n}} \cdot \frac{n^{1-c_1/2}}{q_n} \right) > 0, \tag{3.65}$$

for all sufficiently large n.
Using the preceding observations and applying Hoeffding's Inequality, we obtain

$$\begin{aligned}
u_{1n} &\leq \mathbb{P}_{\theta_i=0}\left(\frac{1}{n - q_n} \sum_{i:\theta_i=0} \mathbf{1}(|X_i| > \sqrt{c_1 \log n}) > p_n + \epsilon_n \right), \\
&\leq e^{-(n-q_n) D(p_n + \epsilon_n || p_n)} \\
&= e^{-(n-q_n) D(\alpha_n || p_n)},
\end{aligned} \tag{3.66}$$

where $D(p||q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right)$ denotes the Kullback-Leibler divergence between two Bernoulli distributed random variables with parameters p and q, respectively.
Now, using the facts $\alpha_n \geq \frac{1}{c_2} \frac{q_n}{n}$, and $p_n \propto \frac{n^{-c_1/2}}{\sqrt{\log n}}$, we obtain

$$
\begin{aligned}
D\left(\alpha_n||p_n\right) &= \alpha_n \log\left(\frac{\alpha_n}{p_n}\right) + (1-\alpha_n)\log\left(\frac{1-\alpha_n}{1-p_n}\right) \\
&= \alpha_n \log\left(\frac{\alpha_n}{p_n}\right) - (1-\alpha_n)\left(\frac{\alpha_n - p_n}{1-p_n}\right)(1+o(1)) \\
&\gtrsim \frac{q_n}{n}\log(n)(1+o(1)),
\end{aligned} \tag{3.67}
$$

whence

$$
u_{1n} \leq e^{-(n-q_n)D(\alpha_n||p_n)} \lesssim e^{-(n-q_n)\frac{q_n}{n}\log(n)(1+o(1))} \lesssim e^{-q_n\log(n)(1+o(1))}. \tag{3.68}
$$

Hence, combining (3.60), (3.61) and (3.68), we have

$$
\frac{1}{q_n}\sum_{\theta_i=0} t_{1i}^{EB} \lesssim \frac{n\alpha_n}{q_n}\frac{1}{\sqrt{\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1)) + \frac{n}{q_n}e^{-q_n\log n(1+o(1))}. \tag{3.69}
$$

Note that the choice of α_n satisfies $\frac{n\alpha_n}{q_n} \to C_3$ for some $0 < C_3 < \infty$. Hence, the first term in the right-hand side of (3.69) goes to zero as $n \to \infty$. On the other hand, since q_n satisfies $q_n \propto (\log n)^{\delta_2}$ for some $\delta_2 > 0$, the second term too tends to zero as $n \to \infty$. This implies (3.58) holds.

Now we move towards establishing (3.59). Recall that, by definition, for any given $x \in \mathbb{R}$, $\mathbb{E}(\kappa|x,\tau)$ is non-increasing in τ. Since $\widehat{\tau} \geq \frac{1}{n}$, hence,

$$
\begin{aligned}
t_{2i}^{EB} &= \mathbb{P}_{\theta_i\neq 0}(\mathbb{E}(\kappa_i|X_i,\widehat{\tau}) > \frac{1}{2}) \\
&\leq \mathbb{P}_{\theta_i\neq 0}(\mathbb{E}(\kappa_i|X_i,\frac{1}{n}) > \frac{1}{2})
\end{aligned} \tag{3.70}
$$

Hence, using arguments similar to (3.40)-(3.46) with $\tau = \frac{1}{n}$, when $\theta_i > 0$ for $i \in S_{\boldsymbol{\theta}}$, we have,

$$
t_{2i}^{\mathrm{EB}} \leq \Phi(\sqrt{2(1+(\log n)^{-\delta_1})\log n} - \sqrt{2\log\left(\frac{n}{q_n}\right)} - b). \tag{3.71}
$$

Hence, in order to obtain the desired upper bound on t_{2i}^{EB}, it is sufficient to show that $\sqrt{(1+(\log n)^{-\delta_1})\log n} - \sqrt{\log\left(\frac{n}{q_n}\right)} \to 0$ as $n \to \infty$. Note that, for $q_n \propto (\log n)^{\delta_2}$,

$$
\sqrt{(1+(\log n)^{-\delta_1})\log n} - \sqrt{\log\left(\frac{n}{q_n}\right)} = \sqrt{(1+(\log n)^{-\delta_1})\log n} - \sqrt{\log n - \delta_2 \log\log n} > 0,
$$

for all n. Hence, the upper bound converging to zero is enough. Observe that

$$\begin{aligned}\sqrt{(1+(\log n)^{-\delta_1})\log n} - \sqrt{\log\left(\frac{n}{q_n}\right)} &= \sqrt{(1+(\log n)^{-\delta_1})\log n} - \sqrt{\log n - \delta_2 \log\log n} \\ &= \frac{(\log n)^{(1-\delta_1)} + \delta_2 \log\log n}{\sqrt{(1+(\log n)^{-\delta_1})\log n} + \sqrt{\log n - \delta_2 \log\log n}} \\ &\leq (\log n)^{(\frac{1}{2}-\delta_1)} + \delta_2 \frac{\log\log n}{\sqrt{\log n}}(1+o(1)) \\ &\to 0, \text{ as } n \to \infty, \end{aligned} \tag{3.72}$$

since $\frac{1}{2} < \delta_1 < 1$. Combining (3.71) and (3.72) implies the desired upper bound on t_{2i}^{EB} is established for $\theta_i > 0$ when $i \in S_{\boldsymbol{\theta}}$. The proof when $\theta_i < 0$ holds using similar arguments. This completes the proof of Theorem 3.3.4. □

Proof of Theorem 3.3.6. Similar to Theorem 3.3.4, here too, we need to obtain an upper bound of the left-hand side of (3.21). Hence, it suffices to prove that

$$\sum_{i:\theta_i=0} t_{1i}^{FB} = o(q_n) \tag{3.73}$$

and

$$\frac{1}{q_n} \sum_{i:\theta_i \neq 0} \sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i}^{FB} \leq 1 - \Phi(b) + o(1), \tag{3.74}$$

where t_{1i}^{FB} and t_{2i}^{FB} are the type I and type II errors for the full Bayes approach corresponding to i^{th} hypothesis, respectively. Note that,

$$\begin{aligned} E(1-\kappa_i|\mathbf{X}) &= \int_{\frac{1}{n}}^{\alpha_n} E(1-\kappa_i|\mathbf{X},\tau)\pi(\tau|\mathbf{X})d\tau \\ &= \int_{\frac{1}{n}}^{\alpha_n} E(1-\kappa_i|X_i,\tau)\pi(\tau|\mathbf{X})d\tau \leq E(1-\kappa_i|X_i,\alpha_n)\cdot \end{aligned} \tag{3.75}$$

For proving the inequality above, we first use the fact that given any $X_i \in \mathbb{R}$, $E(1-\kappa_i|X_i,\tau)$ is non-decreasing in τ, and next we use (C4). Now, using Theorem 4 of Ghosh et al. (2016), we have for $a = \frac{1}{2}$, as $n \to \infty$

$$E(1-\kappa_i|X_i,\alpha_n) \leq K_1 e^{\frac{X_i^2}{2}} \alpha_n (1+o(1)). \tag{3.76}$$

Here $o(1)$ depends only on n such that $\lim_{n\to\infty} o(1) = 0$ and is independent of i and K_1 is a

constant depending on M. Hence, we have the following :

$$\begin{aligned} t_{1i}^{\mathrm{FB}} &= P_{H_{0i}}(E(1-\kappa_i|\mathbf{X}) > \frac{1}{2}) \\ &\leq P_{H_{0i}}\left(\frac{X_i^2}{2} > \log\left(\frac{1}{\alpha_n}\right) - \log K_1 - \log(1+o(1))\right) \\ &= P_{H_{0i}}\left(|X_i| > \sqrt{2\log\left(\frac{1}{\alpha_n}\right)}\right)(1+o(1)). \end{aligned}$$

Note that, as $\alpha_n < 1$ for all $n \geq 1$, $2\log\left(\frac{1}{\alpha_n}\right) > 0$ for all $n \geq 1$. Next, using the fact that under $H_{0i}, X_i \overset{iid}{\sim} \mathcal{N}(0,1)$ with $1-\Phi(t) < \frac{\phi(t)}{t}$ for $t > 0$, we get

$$t_{1i}^{\mathrm{FB}} \leq 2\frac{\phi(\sqrt{2\log\left(\frac{1}{\alpha_n}\right)})}{\sqrt{2\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1)) = \sqrt{\frac{1}{\pi}}\frac{\alpha_n}{\sqrt{\log\left(\frac{1}{\alpha_n^2}\right)}}(1+o(1)), \tag{3.77}$$

where $o(1)$ depends only on n such that $\lim_{n\to\infty} o(1) = 0$. This implies, for sufficiently large n

$$\begin{aligned} \frac{1}{q_n}\sum_{i:\theta_i=0} t_{1i}^{\mathrm{FB}} &\lesssim \frac{n\alpha_n}{q_n\sqrt{\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1)) \\ &= \frac{(\log n)^{\delta_3}}{(\log n)^{\delta_2}\sqrt{\log n - \delta_2 \log\log n}}(1+o(1)) \\ &= (\log n)^{\delta_3 - \frac{1}{2} - \delta_2}(1+o(1)) \to 0, \text{ as } n \to \infty, \end{aligned} \tag{3.78}$$

since $\delta_3 < \frac{1}{2}$ and $\delta_2 > 0$. This implies (3.73) holds. Next, in order to establish (3.74) holds, it is enough to show that, for each $i = 1, \cdots, q_n$, $\sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i}^{\mathrm{FB}} \leq 1 - \Phi(b) + o(1)$, where $o(1)$ is independent of i.

In order to provide an upper bound on the probability of type-II error induced by the decision rule (3.17), $t_{2i}^{\mathrm{FB}} = P_{H_{1i}}(E(\kappa_i|\mathbf{X}) > \frac{1}{2})$, we first note that,

$$\begin{aligned} E(\kappa_i|\mathbf{X}) &= \int_{\frac{1}{n}}^{\alpha_n} E(\kappa_i|\mathbf{X},\tau)\pi(\tau|\mathbf{X})d\tau \\ &= \int_{\frac{1}{n}}^{\alpha_n} E(\kappa_i|X_i,\tau)\pi(\tau|\mathbf{X})d\tau \leq E(\kappa_i|X_i,\frac{1}{n}), \end{aligned} \tag{3.79}$$

where the inequality in the last line follows due to the fact that given any $X_i \in \mathbb{R}$, $E(\kappa_i|X_i,\tau)$ is non-increasing in τ. Hence, following the arguments of (3.71)-(3.72) with $\tau = \frac{1}{n}$, when $\theta_i > 0$ for $i = 1, \cdots, q_n$, we have,

$$\sup_{|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)}+b} t_{2i}^{\mathrm{FB}} \leq 1 - \Phi(b) + o(1), \tag{3.80}$$

where the $o(1)$ term is independent of any i. This completes the proof of Theorem 3.3.6. □

Proof of Theorem 3.3.7. The lower bound to the left hand side of (3.22) follows from Theorem 1 of Abraham et al. (2024). Hence, it is enough to establish that the upper bound to the left hand side of (3.22) is of the order $1 - \Phi(b) + o(1)$.
Using the definition of $FDR(\boldsymbol{\theta}, \boldsymbol{\psi})$, we have for any $\lambda \in (0, 1)$,

$$\begin{aligned} FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n \psi_i, 1)}\right] \\ &= \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n \psi_i, 1)}\mathbf{1}_{\sum_{i=1}^n \psi_i > \lambda q_n}\right] + \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n \psi_i, 1)}\mathbf{1}_{\sum_{i=1}^n \psi_i \leq \lambda q_n}\right] \\ &= v_{1n} + v_{2n}, \text{ say.} \end{aligned} \tag{3.81}$$

Observe that, for the set $\mathbf{1}_{\sum_{i=1}^n \psi_i > \lambda q_n}$, $max(\sum_{i=1}^n \psi_i, 1) > \lambda q_n$. Hence v_{1n} can be bounded as

$$\begin{aligned} v_{1n} &\leq \frac{\sum_{i:\theta_i=0} t_{1i}}{\lambda q_n} = \frac{(n - q_n)t_1}{\lambda q_n} \\ &\leq \frac{M}{\lambda}\frac{n\tau}{q_n}\frac{1}{\sqrt{\log\left(\frac{1}{\tau^2}\right)}}(1 + o(1)) = o(1), \text{ as } n \to \infty. \end{aligned} \tag{3.82}$$

The equality in the second line holds since by definition the type I error for i^{th} hypothesis t_{1i} is the same for all i, and denoted as t_1. The second inequality follows due to the use of (3.39) and the assumption on $L(\cdot)$. Now, note that

$$\{\sum_{i=1}^n \psi_i \leq \lambda q_n\} \subseteq \{\sum_{i:\theta_i \neq 0} \psi_i \leq \lambda q_n\}. \tag{3.83}$$

This implies

$$v_{2n} \leq \mathbb{P}_\theta(\sum_{i:\theta_i \neq 0} \psi_i \leq \lambda q_n). \tag{3.84}$$

Since, ψ_i's are independent Bernoulli random variables, hence, we use Hoeffding's inequality. Choose $0 < \lambda < \Phi(b)$. As a result, for sufficiently large n, t defined in Hoeffding's inequality is obtained as

$$\begin{aligned} t &= q_n - \sum_{i:\theta_i \neq 0} t_{2i} - \lambda q_n \\ &= (1 - \lambda)q_n - \sum_{i:\theta_i \neq 0} t_{2i} \\ &\geq [\Phi(b) - \lambda + o(1)]q_n. \end{aligned} \tag{3.85}$$

Note that, the choice of λ implies $t > 0$ for sufficiently large n. In order to obtain (3.85), we use a uniform upper bound on $t_{2i}, i : \theta_i \neq 0$ as derived in (3.56). Using (3.85), we obtain, for all

$|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b,\ i \in S_{\boldsymbol{\theta}}$

$$\mathbb{P}_\theta\Big(\sum_{i:\theta_i\neq 0}\psi_i \leq \lambda q_n\Big) \leq \exp\left(-\frac{2t^2}{q_n}\right)$$
$$\leq \exp\big(-2[\Phi(b) - \lambda + o(1)]^2 q_n\big). \tag{3.86}$$

Next, combining (3.84)-(3.86), for sufficiently large n

$$\sup_{\theta\in\Theta_b} v_{2n} = o(1),\ \text{ as } n \to \infty. \tag{3.87}$$

Now, (3.81), (3.82) and (3.87) combine to prove that

$$\sup_{\theta\in\Theta_b} FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) = o(1),\ \text{ as } n \to \infty. \tag{3.88}$$

On the other hand, for $FNR(\boldsymbol{\theta}, \boldsymbol{\psi})$, from definition, we obtain

$$FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{\mathbb{E}_\theta(\sum_{i:\theta_i\neq 0}(1-\psi_i))}{q_n}$$
$$= \frac{1}{q_n}\sum_{i:\theta_i\neq 0} t_{2i}. \tag{3.89}$$

Recall that, in (3.56), we have already derived that, for each $|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b,\ i \in S_{\boldsymbol{\theta}}$,

$$t_{2i} \leq 1 - \Phi(b) + o(1),$$

where the $o(1)$ term is independent of any i. This combining with (3.89) implies for all sufficiently large n

$$\sup_{\theta\in\Theta_b} FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) \leq 1 - \Phi(b) + o(1). \tag{3.90}$$

Finally, (3.88) and (3.90) imply the upper bound to the left hand side of (3.22) is of the order $1 - \Phi(b) + o(1)$ and completes the proof. □

Proof of Theorem 3.3.9. Here we follow the basic architecture of the proof of Theorem 3.3.7. Hence, we only need to show that the upper bound to the left-hand side of (3.23) is of the order $1 - \Phi(b) + o(1)$.
Next, using the argument of (3.81), we have

$$v_{3n} = \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n \psi_i, 1)}\mathbf{1}_{\sum_{i=1}^n \psi_i > \lambda q_n}\right],$$

and

$$v_{4n} = \mathbb{E}_\theta \left[\frac{\sum_{i:\theta_i=0} \psi_i}{max(\sum_{i=1}^n \psi_i, 1)} \mathbf{1}_{\sum_{i=1}^n \psi_i \leq \lambda q_n} \right],$$

where ψ_i is of the form (3.16). Hence, we obtain,

$$\begin{aligned} v_{3n} &\leq \frac{\sum_{i:\theta_i=0} t_{1i}^{EB}}{\lambda q_n} \\ &\lesssim \frac{n\alpha_n}{q_n} \frac{1}{\sqrt{\log\left(\frac{1}{\alpha_n}\right)}} (1 + o(1)) + \frac{n}{q_n} e^{-q_n \log n(1+o(1))}. \end{aligned} \tag{3.91}$$

Here inequality in the second line holds due to (3.69). Note that the choice of α_n satisfies $\frac{n\alpha_n}{q_n} \to C_3$ for some $0 < C_3 < \infty$. Hence, the first term to the right hand side of (3.91) goes to zero as $n \to \infty$. On the other hand, since q_n satisfies $q_n \propto \log n^{\delta_2}$ for some $\delta_2 > 0$, the second term too tends to zero as $n \to \infty$. This implies

$$v_{3n} = o(1), \text{ as } n \to \infty. \tag{3.92}$$

Again employing (3.83) and (3.84), we obtain for sufficiently large n

$$\begin{aligned} t &= q_n - \sum_{i:\theta_i \neq 0} t_{2i}^{EB} - \lambda q_n \\ &\geq [\Phi(b) - \lambda + o(1)] q_n, \end{aligned} \tag{3.93}$$

where the upper bound on t_{2i}^{EB} as obtained in (3.71) is used. This lower bound with Hoeffding's inequality confirms, for sufficiently large n

$$\sup_{\theta \in \Theta_b} v_{4n} = o(1), \text{ as } n \to \infty. \tag{3.94}$$

Combining (3.92) and (3.94) ensures

$$\sup_{\theta \in \Theta_b} FDR(\boldsymbol{\theta}, \boldsymbol{\psi}) = o(1), \text{ as } n \to \infty. \tag{3.95}$$

For $FNR(\boldsymbol{\theta}, \boldsymbol{\psi})$, from definition, we obtain

$$\begin{aligned} FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \frac{\mathbb{E}_\theta(\sum_{i:\theta_i \neq 0}(1 - \psi_i))}{q_n} \\ &= \frac{1}{q_n} \sum_{i:\theta_i \neq 0} t_{2i}^{EB}. \end{aligned} \tag{3.96}$$

Recall that, in (3.71), we have already derived that, for each $|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b$, $i \in S_{\boldsymbol{\theta}}$,

$$t_{2i}^{EB} \leq 1 - \Phi(b) + o(1),$$

where the $o(1)$ term is independent of any i. This combining with (3.96) implies for all sufficiently large n

$$\sup_{\theta\in\Theta_b} FNR(\boldsymbol{\theta},\boldsymbol{\psi}) \leq 1-\Phi(b)+o(1). \tag{3.97}$$

Finally, (3.95) and (3.97) imply the upper bound to the left hand side of (3.23) is of the order $1-\Phi(b)+o(1)$ and completes the proof. □

Proof of Theorem 3.3.11. Here also we follow the basic architecture of the proof of Theorem 3.3.7. Hence, in this case, too, it is sufficient to show that the upper bound to the left-hand side of (3.24) is of the order $1-\Phi(b)+o(1)$.
Note that, using the argument of (3.81) and (3.78), we have

$$v_{5n} = \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n\psi_i,1)}\mathbf{1}_{\sum_{i=1}^n\psi_i>\lambda q_n}\right],$$

and

$$v_{6n} = \mathbb{E}_\theta\left[\frac{\sum_{i:\theta_i=0}\psi_i}{max(\sum_{i=1}^n\psi_i,1)}\mathbf{1}_{\sum_{i=1}^n\psi_i\leq\lambda q_n}\right],$$

where ψ_i is of the form (3.17). As a result, we obtain,

$$\begin{aligned} v_{5n} &\leq \frac{\sum_{i:\theta_i=0} t_{1i}^{FB}}{\lambda q_n} \\ &\lesssim \frac{n\alpha_n}{q_n\sqrt{\log\left(\frac{1}{\alpha_n}\right)}}(1+o(1)) \\ &= o(1),\ \text{as } n\to\infty. \end{aligned} \tag{3.98}$$

Again employing arguments of (3.83) and (3.84), we obtain for sufficiently large n

$$\begin{aligned} t &= q_n - \sum_{i:\theta_i\neq 0} t_{2i}^{FB} - \lambda q_n \\ &\geq [\Phi(b)-\lambda+o(1)]q_n, \end{aligned} \tag{3.99}$$

where the upper bound on t_{2i}^{FB} as obtained in Theorem 3.3.6 is used. This lower bound with Hoeffding's inequality confirms, for sufficiently large n

$$\sup_{\theta\in\Theta_b} v_{6n} = o(1),\ \text{as } n\to\infty. \tag{3.100}$$

Combining (3.98) and (3.100) ensures

$$\sup_{\theta\in\Theta_b} FDR(\boldsymbol{\theta},\boldsymbol{\psi}) = o(1),\ \text{as } n\to\infty. \tag{3.101}$$

For $FNR(\boldsymbol{\theta}, \boldsymbol{\psi})$, from definition, we obtain

$$\begin{aligned} FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) &= \frac{\mathbb{E}_\theta(\sum_{i:\theta_i \neq 0}(1 - \psi_i))}{q_n} \\ &= \frac{1}{q_n} \sum_{i:\theta_i \neq 0} t_{2i}^{FB}. \end{aligned} \tag{3.102}$$

Recall that, in Theorem 3.3.6, we have already derived that, for each $|\theta_i| \geq \sqrt{2\log\left(\frac{n}{q_n}\right)} + b$, $i \in S_{\boldsymbol{\theta}}$,

$$t_{2i}^{FB} \leq 1 - \Phi(b) + o(1),$$

where the $o(1)$ term is independent of any i. This combining with (3.102) implies for all sufficiently large n

$$\sup_{\theta \in \Theta_b} FNR(\boldsymbol{\theta}, \boldsymbol{\psi}) \leq 1 - \Phi(b) + o(1). \tag{3.103}$$

Finally, (3.101) and (3.103) imply the upper bound to the left hand side of (3.24) is of the order $1 - \Phi(b) + o(1)$ and completes the proof. □

Chapter 4

Consistent group selection using global-local shrinkage priors in sparse normal linear regression

4.1 Introduction

In this chapter, our focus is on group selection and estimation of group regression coefficients within a linear regression framework. The importance of the problem of selection of groups as a whole has been highlighted in Chapter 1 by giving examples of specific situations where group selection, as opposed to selection of individual members in a group, would be meaningful. We examine a linear regression model containing p covariates, which are divided into G groups of potential regressors or predictors as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}. \tag{4.1}$$

Here, $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ design matrix, $\mathbf{X}_g$ is an $n \times m_g$ design matrix corresponding to the g^{th} group, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, $\boldsymbol{\beta}_g$ is an $m_g \times 1$ vector of unknown regression coefficients for the g^{th} group where $g \in \{1, \ldots, G\}$ and $\sum_{g=1}^{G} m_g = p$. We further assume that the vector of unobserved residuals $\boldsymbol{\epsilon}$ has a $N(\mathbf{0}, \sigma^2 I_n)$ distribution. Our interest is in a sparse asymptotic setting where $G \equiv G_n$ increases at the same rate as the sample size n and the number of *active* groups $G_{\mathcal{A}_n}$ grows at a slower rate than G_n, an active group being one with a non-zero true group regression coefficient vector. In this chapter, we propose and study some methods for selecting and estimating the active group coefficients using a broad class of one-group hierarchical priors on the grouped regression coefficients. Interestingly, these priors may be looked upon as *global-local* mixtures (in the grouped regression problem) of the popular g-prior of Zellner (1986).

Variable selection for grouped and ungrouped regression models has been reviewed in some

detail in several sections of Chapter 1 covering both frequentist and Bayesian literature in this domain. Our work in this chapter is a study of optimality of grouped variable selection and estimation based on a new procedure introduced for grouped regression, namely the Half-thresholding (HT) rule inspired by a similar rule in Tang et al. (2018). The motivation for this work comes from the critical study of some key references including Xu and Ghosh (2015), Xu et al. (2016), Yang and Narisetty (2020) and Boss et al. (2023) which were helpful to identify certain questions which we could possibly attempt to address. Subsection 1.2.3 comprises that discussion. For the sake of completeness, we now describe in precise technical terms the spike-and-slab priors considered by Xu and Ghosh (2015) and Yang and Narisetty (2020) in the group problem and their main findings. Xu and Ghosh (2015) considered the following hierarchical formulation:

$$\begin{aligned}
&\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n), \\
&\boldsymbol{\beta}_g \overset{ind}{\sim} \pi_0 \delta_{\{\mathbf{0}\}}(\boldsymbol{\beta}_g) + (1-\pi_0)\mathcal{N}_{m_g}(\mathbf{0}, \sigma^2\tau_g^2 I_{m_g}), \text{ for } g = 1, \cdots, G, \\
&\tau_g^2 \overset{ind}{\sim} \text{Gamma}(\frac{m_g+1}{2}, \frac{\lambda^2}{2}), g = 1, 2, \cdots, G, \\
&\sigma^2 \sim \text{Inverse Gamma}(\alpha, \gamma).
\end{aligned} \tag{4.2}$$

The prior in (4.2) is known as the Bayesian group LASSO with spike and slab priors (BGL-SS). Under the assumption of a block-orthogonal design matrix, they showed that the marginal posterior median estimator of $\boldsymbol{\beta_g}$ is zero when the norm of the corresponding block least squares estimator is less than a certain threshold. Accordingly, they declare a group to be active or inactive based on whether the posterior median of $\boldsymbol{\beta}_g$ is non-null or null. They were also able to prove that the posterior median estimator has the oracle property for group variable selection and estimation under orthogonal designs. For the same problem, Yang and Narisetty (2020) modified the BGL-SS by introducing a binary latent variable $(Z_g, g = 1, \cdots, G)$ for each group (to indicate the activeness of the group or not) and proposing spike and slab priors on group coefficients depending on the latent variables. Though this prior was quite similar to BGL-SS, their approach was different in two important ways compared to Xu and Ghosh (2015). Firstly, the slab part of the prior was made data dependent, inspired by the theory of Narisetty and He (2014) indicating better shrinkage through such a choice. Secondly, the choice of group corresponds to the vector of latent variables with highest posterior probability, rather than any rule involving marginal posterior distribution of individual group coefficients. They established variable selection consistency of their proposed decision rule.

It is needless to reiterate the obvious computational difficulties of two-groups mixture model in sparse group regression contexts. A fortiori, as commented by Yang and Narisetty (2020), for Gibbs sampling to work, $\boldsymbol{\beta}$ and $\mathbf{Z} = (Z_1, \cdots, Z_G)$ need to be clubbed together. So the need for the development of one-group priors in grouped regression from this point is paramount. We also mention once more that although the works of Xu et al. (2016) and Boss et al. (2023) developed one-group priors in this context, they are not focused on the problems of group selection and estimation and study of their optimality; questions we are really interested in.

In this chapter, we propose and study a decision rule based on a very broad class of one-group

shrinkage priors with the following hierarchical formulation

$$\boldsymbol{\beta}_g \mid \lambda_g^2, \sigma^2, \tau^2 \sim \mathcal{N}_{m_g}(\mathbf{0}, \lambda_g^2 \sigma^2 \tau^2 (\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1}), \quad \pi(\lambda_g^2) \propto (\lambda_g^2)^{-a-1} L(\lambda_g^2), \tag{4.3}$$

where a and $L(\cdot)$ are defined in (1.10) and $L(\cdot)$ satisfies Assumption 1 for $a \geq \frac{1}{2}$. For our theoretical analyses, σ^2 is assumed to be known. For the purpose of simulations, we model σ^2 as $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, the Jeffreys prior. The global parameter τ is either treated as a tuning parameter or in an empirical Bayes/ full Bayes ways, depending on whether the number of active groups is known or unknown. For deriving the theoretical properties of our rule, we restrict ourselves to a block orthogonal design. Propositions 4.2.2 and 4.2.3 of Subsection 4.2.2, derived under the block orthogonality assumption, motivate us to form our decision rule. In this chapter, we study this rule, which is based on the ratio of ℓ_2 norms of the posterior mean and the least square estimate, and call this *Half-thresholding* (HT) rule irrespective of whether the design matrix is block orthogonal or not. Clearly, this is a generalization of the corresponding rule of Tang et al. (2018) for an ungrouped problem. We also propose a corresponding *Half-thresholding* estimator of the active groups. Our main findings in this chapter have been broadly indicated in Subsection 1.2.3 and are described in details in appropriate sections of the chapter.

The remainder of the chapter is organized as follows. In Section 4.2, the hierarchical form of the modified class of global-local priors with polynomial tail and the proposed half-thresholding rule are described. Section 4.3 presents the main theoretical results of the chapter. Sections 4.5 and 4.6 deal with the simulation results and real data application, respectively. An overall discussion with some future problems can be found in Section 4.7. The chapter ends in Section 4.8, where the proofs of all theoretical results are provided.

This chapter is based on Paul et al. (2026).

4.2 Prior Selection and the Half-Thresholding Rule

Consider the linear model (4.1). Let $m = (m_1, \cdots, m_G)$ be the number of individual variables within each group, and $p = \sum_{g=1}^{G} m_g$ be the total number of variables under consideration. Let us assume that the design matrix for the g^{th} group, denoted $\mathbf{X}_g$, is of full rank, for $g = 1, 2, \cdots, G$.

4.2.1 Hierarchical form

As stated in the introduction, in this article, we consider the following Bayesian hierarchical structure given by

$$\begin{aligned} \mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n), \\ \boldsymbol{\beta}_g \mid \lambda_g^2, \sigma^2, \tau^2 &\sim \mathcal{N}_{m_g}(\mathbf{0}, \lambda_g^2 \sigma^2 \tau^2 (\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1}), \text{ independently for } g = 1, 2, \cdots, G, \\ \lambda_g^2 &\sim \pi(\lambda_g^2), \text{ independently for } g = 1, 2, \cdots, G, \text{ and} \\ (\tau, \sigma^2) &\sim \pi(\tau, \sigma^2). \end{aligned} \tag{4.4}$$

(4.4) is a s a slightly modified version of the hierarchical form proposed by Tang et al. (2018). The motivation for this modification is discussed in detail in the remark presented at the end of this subsection. In (4.1), λ_g denotes the local shrinkage parameter for the g^{th} group, and τ denotes the global shrinkage parameter. Here, $\pi(\cdot)$ denotes a non-degenerate prior distribution used to model the global shrinkage component τ, and the error variance σ^2. Polson and Scott (2010) suggested that in sparse problems, prior distributions on local and global shrinkage parameters should have the following properties:

1. The prior on the local shrinkage parameter should have thick tails to accommodate the non-null coefficients, and
2. The prior on the global shrinkage parameter should have substantial mass near the origin to account for sparsity.

Motivated by this and the previous works of Ghosh et al. (2016), and Ghosh and Chakrabarti (2017), we assume throughout this article that the prior distribution of λ_g^2 will be of the form

$$\pi(\lambda_g^2) \propto (\lambda_g^2)^{-a-1} L(\lambda_g^2). \tag{4.5}$$

In (4.5) above, a is a positive real number, and $L : (0,\infty) \to (0,\infty)$ is a measurable non-constant slowly varying function in Karamata's sense (see Bingham et al. (1987)), that is, $\frac{L(\alpha x)}{L(x)} \to 1$ as $x \to \infty$, for any $\alpha > 0$. Priors of the form given in (4.5) are naturally heavy-tailed. For an orthogonal design $\mathbf{X}$, the class of one-group shrinkage priors given in (4.4) satisfying (4.5) covers a broad array of heavy-tailed global-local shrinkage prior distributions such as the t-prior due to Tipping (2001), the negative exponential gamma prior due to Griffin and Brown (2005), the Horseshoe prior of Carvalho et al. (2009), the three-parameter beta normal mixtures of Armagan et al. (2011), the generalized double Pareto priors due to Armagan et al. (2013a), the inverse gamma priors, just to name a few. See, for instance, Ghosh et al. (2016) and Ghosh and Chakrabarti (2017), in this context.

The *global shrinkage* parameter τ is either treated as a tuning parameter τ_n based on the proportion of non-zero means or is treated in an empirical Bayes or fully Bayesian way depending on whether the proportion of active groups is known or not. For the theoretical development of this paper, we assume that the error variance term σ^2 is fixed in (4.4). See Castillo et al. (2015), Rigollet and Tsybakov (2012) for similar treatment of σ^2. On the other hand, for simulation, we employ Jeffry's prior to model the unknown σ^2. We further assume the for $a \geq \frac{1}{2}$, the slowly varying function $L(\cdot)$ defined in (4.5) satisfies Assumption 1.

Remark 4.2.1. *Several comments are in order regarding our proposed hierarchical formulation* (4.4). *Tang et al. (2018) considered the following hierarchical formulation based on the class of global-local priors in regression problem,*

$$\begin{aligned} \mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n), \\ \boldsymbol{\beta}_j \mid \lambda_j^2, \sigma^2, \tau^2 &\overset{ind}{\sim} \mathcal{N}(0, \lambda_j^2\sigma^2\tau^2), \ \textit{for } j = 1, 2, \cdots, p, \\ \lambda_j^2 &\overset{ind}{\sim} \pi(\lambda_j^2) = K(\lambda_j^2)^{-a-1} L(\lambda_j^2), \end{aligned} \tag{4.6}$$

where K and $L(\cdot)$ are same as discussed before. Hence, in case of a group selection problem, a straightforward extension in the hierarchical formulation would be to model each group coefficient $\boldsymbol{\beta}_g$ as

$$\boldsymbol{\beta}_g|\lambda_g^2, \sigma^2, \tau^2 \overset{ind}{\sim} \mathcal{N}_{m_g}(\mathbf{0}, \lambda_g^2\sigma^2\tau^2 I_{m_g})$$

keeping the prior on the local shrinkage coefficient same as before. However, due to this formulation mentioned in (4.6), Tang et al. (2018) were able to theoretically study the decision rule only when the design matrix was orthogonal.

By adding the term $(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1}$ in the formulation (4.4) above, we can motivate and theoretically study a simple thresholding rule to select a group, which does not require the orthogonality assumption of Tang et al. (2018). We can study our rule theoretically for the more general block orthogonal structure. Secondly, it is interesting to note that our proposed class of priors can be considered as a generalization of the g-prior of Zellner (1986) and has a hierarchical form similar to the block hyper-g shrinkage prior proposed by Som et al. (2014). However, the approach of Som et al. (2014) does not model global and local coefficients separately, nor do they provide any decision rule regarding the selection of a group. Finally, for hierarchical modeling, Gelman (2006) commented that the prior variance of the mean parameter should be on the same scale as that of the sufficient statistic. Under a block orthogonal design matrix, the sufficient statistic $\widehat{\boldsymbol{\beta}}_g \sim \mathcal{N}_{m_g}(\boldsymbol{\beta}_g, \sigma^2(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1})$. As a result, from (4.4), it is clear that the variance of $\widehat{\boldsymbol{\beta}}_g$ remains on the same scale as the prior distribution of $\boldsymbol{\beta}_g$ in line with the suggestion of Gelman (2006).

4.2.2 The Half-Thresholding (HT) Rule

In this subsection, we propose a rule for deciding whether a group is active or not. The rule is motivated by two key observations, which are stated as two propositions below. Proofs of these two are presented in the supplementary material.

Before stating them, we note that for a block orthogonal design matrix, $\mathbf{X}$ within the hierarchical framework of (4.4), the posterior mean of $\boldsymbol{\beta}_g$ conditioned on $(\lambda_g, \tau_n, \sigma^2, \mathcal{D})$ is given by

$$E(\boldsymbol{\beta}_g \mid \lambda_g, \tau_n, \sigma^2, \mathcal{D}) = (1 - \kappa_g)\widehat{\boldsymbol{\beta}}_g \,,$$

where $\kappa_g = 1/(1 + \lambda_g^2\tau^2)$ and $\widehat{\boldsymbol{\beta}}_g$ is the least square estimate of $\boldsymbol{\beta}_g$. Therefore, by Fubini's Theorem, it follows that the posterior mean of $\boldsymbol{\beta}_g$, denoted as $\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}$ is of the form

$$\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} = E(\boldsymbol{\beta}_g|\mathcal{D}) = (E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}))\widehat{\boldsymbol{\beta}}_g \,. \tag{4.7}$$

Thus $E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$ is the factor by which the usual estimator is shrunk in the Bayesian formulation. Under the assumption that the design matrix is block orthogonal, the propositions are about the behavior of the shrinkage factor under the null and the alternatives.

Proposition 4.2.2. *Suppose that the g^{th} group is inactive, that is, $\boldsymbol{\beta}_g^0 = \mathbf{0}$. Assume that, the group size m_g satisfies $s = \sup_{n\geq 1} \max\{m_g : g = 1, \cdots, G_n\} < \infty$. Then, $E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \xrightarrow{P} 0$ as $n \to \infty$ if and only if $\tau_n \to 0$ as $n \to \infty$.*

Proposition 4.2.3. *Suppose that the g^{th} group is active, that is, $\boldsymbol{\beta}_g^0 \neq \mathbf{0}$. Define $\mathbf{Q}_{n,g} = \frac{\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g}{n}$ and $g = 1, 2, \cdots, G$. Consider the following assumptions:*

(A1) for any active group g, there exists some global constant $C_1 > 0$ such that $e_{min}\mathbf{Q}_{n,g} > C_1$.

(A2) for any active group $g, \min_j |\beta_{gj}^0| > m_n$ with $m_n \propto n^{-b}$ and $0 \leq b < \frac{1}{2}$,

(A3) the group size m_g satisfies $s = \sup_{n\geq 1} \max\{m_g : g = 1, \cdots, G_n\} < \infty$,

(A4) the global parameter $\tau_n \to 0$ as $n \to \infty$ such that $\log\left(\frac{1}{\tau_n}\right) \lesssim \log G_n$,

Under the assumptions (A1)-(A4), $E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \xrightarrow{P} 1$ as $n \to \infty$.

Propositions 4.2.2 and 4.2.3 leads to the decision rule:

For each $g = 1, \cdots, G$,

$$\text{the } g^{th}\text{group is } \begin{array}{l} \text{considered active if } E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > 0.5, \text{ and} \\ \text{is inactive otherwise.} \end{array} \tag{4.8}$$

Remark 4.2.4. *Proposition 4.2.2 indicates that if the g^{th} group is inactive, our proposed thresholding rule detects the same, and the only condition required to establish this is that $\tau_n = o(1)$ as $n \to \infty$.*

Remark 4.2.5. *Proposition 4.2.3 implies that if $\tau_n \to 0$ not too fast, our thresholding rule successfully identifies an active group (asymptotically) provided the design matrix satisfies a simple condition. However, the condition $\log\left(\frac{1}{\tau_n}\right) \lesssim \log G_n$ allows to choose τ_n from a wide range of values. The importance of other conditions is discussed in detail in Remark 4.3.3.*

Since our proposed decision rule declares a group to be active or not depending on whether $E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$ exceeds half or not, the rule is called *Half-Thresholding (HT) rule*. Note that using (4.7), the decision rule (4.8) can be alternatively stated as:

$$\text{the } g^{th}\text{group is considered active if } \frac{||\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}||_2}{||\widehat{\boldsymbol{\beta}}_g||_2} > 0.5, \text{ for } g = 1, 2, \cdots, G. \tag{4.9}$$

For unit group size, i.e., $m_g = 1, g = 1, 2, \cdots, G$, this is exactly the variable selection rule of Tang et al. (2018). This way, our proposed thresholding rule is a generalization to that of Tang et al. (2018), although their motivation for proposing this rule was different. We define our half-thresholding(HT) estimator of $\boldsymbol{\beta}_g$ corresponding to the variable selection rule as

$$\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} I\big\{E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > 0.5\big\}, \tag{4.10}$$

where $\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}$ denotes the posterior mean of the unknown group coefficient $\boldsymbol{\beta}_g$ corresponding to the g^{th} group.

Note that our proposed decision rule (4.8) uses the global shrinkage parameter τ as a tuning parameter chosen depending on the sample size n(and the $G_{\mathcal{A}_n}$). This gives rise to the question of the treatment of τ in fact when $G_{\mathcal{A}_n}$ is unknown, which happens very often. A natural data-adaptive solution to this problem would be the use of some empirical Bayes estimate(s) of τ by learning through the data. For the recovery of a sparse normal means vector using the horseshoe prior, van der Pas et al. (2014) proposed an empirical Bayes estimator of τ given by

$$\widehat{\tau} = \max\left\{\frac{1}{n}, \frac{1}{c_2 n}\sum_{i=1}^{n} 1\left(\frac{|y_i|}{\sigma} > \sqrt{c_1 \log n}\right)\right\}, \tag{4.11}$$

where c_1 and c_2 are two positive constants with $c_1 \geq 2$ and $c_2 \geq 1$. Motivated by this, we consider the following empirical Bayes estimate of τ given by

$$\widehat{\tau}^{\mathrm{EB}} = \max\left\{\frac{1}{G_n}, \frac{1}{c_2 G_n}\sum_{g=1}^{G_n} 1\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n\right)\right\}, \tag{4.12}$$

where G_n denotes the total number of groups that varies with n and satisfies $G_n \leq n$. From the above definition, it readily follows that $\widehat{\tau}^{\mathrm{EB}}$ always lies between $\frac{1}{G_n}$ and 1. Since a lower bound of this estimator is $\frac{1}{G_n}$, it cannot collapse to zero. Collapsing of the estimator to zero is a major concern in the context of the use of such empirical Bayes procedures as mentioned by several authors, such as Carvalho et al. (2009), Scott and Berger (2010), Bogdan et al. (2008) and Datta and Ghosh (2013). It may be noted that when $\mathbf{X} = \mathbf{I}_p$ and $m_g = 1$, for $g = 1, 2, \cdots, G$, (4.12) boils down to (4.11).

Let $E(1 - \kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D})$ denote the posterior shrinkage weight corresponding to the g^{th} group evaluated at $\tau = \widehat{\tau}^{\mathrm{EB}}$. Using this empirical Bayes estimate, our proposed data-adaptive decision rule is given by

$$\text{The } g^{th}\text{group is considered active if } E(1 - \kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > 0.5, \text{ for } g = 1, 2, \cdots, G, \tag{4.13}$$

and the corresponding empirical Bayes half-thresholding(HT) estimator of $\boldsymbol{\beta}_g$, denoted $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{g,EB}$, is given by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{g,EB} = \widehat{\boldsymbol{\beta}}^{\mathrm{PM}}_{g} I\big\{E(1 - \kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > 0.5\big\}. \tag{4.14}$$

An alternative approach to the above empirical Bayes procedure is to assign a non-degenerate joint prior density to (τ, σ). In line with the recommendation of Polson and Scott (2010), for a fully Bayesian approach, we assume $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$ and $\pi(\tau)$ is the restriction of half Cauchy prior on a suitably chosen small interval near zero (details and motivations of which are available in Section 4.3.2). The full Bayesian half-thresholding (HT) decision rule is given by

$$\text{The } g^{th}\text{group is considered active if } E(1 - \kappa_g \mid \mathcal{D}) > 0.5, \text{ for } g = 1, 2, \cdots, G, \tag{4.15}$$

and the corresponding full Bayes half-thresholding(HT) estimator of $\boldsymbol{\beta}_g$, denoted $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{g,FB}$, is given by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{g,FB} = \widehat{\boldsymbol{\beta}}^{\mathrm{PM}}_{g} I\big\{E(1 - \kappa_g \mid \mathcal{D}) > 0.5\big\}. \tag{4.16}$$

For the implementation of these decision rules, one needs to sample from the posterior distribution of relevant parameters. The Gibbs sampling algorithm for these is explained in Section 4.5.1.

4.3 Main Theoretical results

In this section, under the assumption that the design matrix is block orthogonal, we present our theoretical results concerning asymptotic properties of estimation of the group coefficients and variable selection using the proposed Half-thresholding (HT) rule. Following the works of Fan and Li (2001) and Zou (2006), our aim here is to establish that the proposed half-thresholding methods defined in (4.8), (4.13) and (4.15) attain the oracle properties(defined below) asymptotically as the number of observations n grows to infinity. Let $\mathcal{A} = \{g : \boldsymbol{\beta}_g^0 \neq \mathbf{0}\}$ and $\mathcal{A}_n = \{g : \widehat{\boldsymbol{\beta}}_g^{\text{HT}} \neq \mathbf{0}\}$ denote respectively the set of true active groups and the groups declared active by our half-thresholding rule. The aforesaid articles defined a procedure δ to be an oracle if the resultant procedure can identify the true model asymptotically and the estimator corresponding to that procedure $\widehat{\boldsymbol{\beta}}(\delta)$ can achieve the optimal rate of estimation simultaneously. The exact forms of these expressions in our context will be discussed later. As mentioned before, for studying the oracle properties of the thresholding rules (4.8), (4.13), and (4.15), we treat the global shrinkage component τ either as a tuning parameter to be chosen appropriately when the number of active groups is assumed to be known or it is replaced either by an empirical Bayes estimator as given in (4.12) or is modeled by a non-degenerate prior in case the number of active groups is unknown. In both cases, however, the error variance term σ^2 is assumed to be known and does not vary with n.

Since our proposed half-thresholding (HT) rules crucially hinge upon the posterior shrinkage coefficients, for the sake of completeness, we describe below the posterior distribution of κ_g given by

$$\pi(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \propto \kappa_g^{(a+\frac{m_g}{2}-1)}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right) \exp\left(-\kappa_g \cdot \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right), 0 < \kappa_g < 1, \tag{4.17}$$

where $\mathbf{Q}_{n,g} = \frac{\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g}{n}$ and $g = 1, 2, \cdots, G$. Note that, since the error variance σ^2 is assumed to be known, the above posterior distribution of κ_g depends only on τ_n and the data $\mathcal{D}$ We repeatedly make careful exploitation of this last observation to establish the oracle properties of the half-thresholding rules proposed in this paper. On the other hand, when the global shrinkage parameter τ is replaced with an empirical Bayes estimator $\widehat{\tau}^{EB}$ or a prior is assigned to it, the posterior distribution of κ_g depends on the entire dataset $\mathcal{D}$ which makes the theoretical derivations significantly different, and technically more challenging.

4.3.1 Oracle properties of the HT procedure when τ is known

In this sub-section, we treat the *global shrinkage* parameter τ as a tuning parameter. Propositions 4.2.2 and 4.2.3 stated in Subsection 4.2.2 indicate that the half-thresholding rule of the form (4.8) correctly identifies the individual groups as active or inactive. Theorem 4.3.1 below ensures the same for the overall group selection problem when the sample size n grows to infinity. Hence the proposed half-thresholding rule defined in (4.8) enjoys model selection consistency under the assumption that the design matrix is block orthogonal. Proof of this result can be found in Section 4.8.

Theorem 4.3.1 (Variable Selection Consistency)**.** *Consider the hierarchical framework of* (4.4) *where* $\pi(\lambda_g^2)$ *is as in* (4.5) *with* $a \geq \frac{1}{2}$, $L(\cdot)$ *in* (4.5) *satisfies Assumption 1 and the half-thresholding (HT) rule* (4.8) *based on these. Let* $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \neq \mathbf{0}\}$ *and* $\widehat{\mathcal{A}}_n = \{g : \widehat{\boldsymbol{\beta}}_g^{HT} \neq \mathbf{0}\}$ *denote respectively the set of truly active groups, and the set of groups declared active by the half-thresholding rule* (4.8)*. Let* $\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$, *for* $g = 1, \cdots, G$ *and* $r_n = \frac{G_{\mathcal{A}_n}}{G_n}$. *Assume further that assumptions (A1)-(A3) are satisfied. Then we have the following.*
(i) Suppose, $G_n\tau_n[\log\left(\frac{1}{\tau_n}\right)]^{\frac{s}{2}-1} \to 0$ *as* $n \to \infty$. *Then the half-thresholding rule* (4.8) *results in variable selection consistency, i.e., we have,*

$$\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n = \mathcal{A}_n) = 1 \text{ as } n \to \infty.$$

(ii)Suppose,

(B1) the total number of active groups $|\mathcal{A}_n| = G_{\mathcal{A}_n}$ *is known and satisfies* $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some known* $0 \leq \epsilon_1 \leq \epsilon_2 < 1$, *and*

(B2) the global parameter $\tau_n \to 0$ *as* $n \to \infty$ *such that* $r_n^{\frac{1+\delta_2}{1-\epsilon_2}} \lesssim \tau_n \lesssim r_n^{\frac{1+\delta_1}{1-\epsilon_1}}$ *for some* $\delta_2 > \delta_1 > \frac{\epsilon_2-\epsilon_1}{1-\epsilon_2}$.

Then, the half-thresholding rule (4.8) *results in variable selection consistency, i.e.,*

$$\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n = \mathcal{A}_n) = 1 \text{ as } n \to \infty.$$

Remark 4.3.2. *Observe that, asserting* $\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n = \mathcal{A}_n) = 1$ *as* $n \to \infty$ *is equivalent to saying*

$$\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n \neq \mathcal{A}_n) = 0 \text{ as } n \to \infty$$

which is what we establish to prove Theorem 4.3.1. This is proved by showing

$$\sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) < \frac{1}{2}) \to 0, \text{ as } n \to \infty, \tag{4.18}$$

and

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \to 0, \text{ as } n \to \infty. \tag{4.19}$$

Thus, not only do both the probabilities of type-I error and type-II error tend to 0 *as* $n \to \infty$ *individually, but their corresponding sums also tend to* 0 *as* $n \to \infty$.

Remark 4.3.3. *Condition (A1) is very natural in variable selection problems. Johnson and Rossell (2012) and Armagan et al. (2013b) assumed the same condition on the eigenvalues of* $(\mathbf{X}^{\mathrm{T}}\mathbf{X})/n$ *while studying the posterior contraction rates in high-dimensional regression problems. Under (B1), a condition similar to (B2) of Theorem 4.3.1 was considered in Tang et al. (2018). But the main difference lies in the assumption regarding the design matrix* $\mathbf{X}$ *and the cardinality of* $|\mathcal{A}_n|$ *while proving the result. In their work, Tang et al. (2018) assumed the corresponding design matrix* $\mathbf{X}$ *to be orthogonal (i.e.* $\mathbf{X}^{\mathrm{T}}\mathbf{X} = n\mathbf{I}_p$*) and the number of active variables is independent of* n*, which restricts the applicability of their result. For an orthogonal design, assumption (A1) is trivially satisfied. Further, assuming* $|\mathcal{A}_n|$ *being fixed (that is, independent of* n*) implies that assumption (A2) is not required at all. To deal with a more general scenario, we have allowed the total number of active groups to vary with* n*. Therefore, in several aspects, Theorem 4.3.1 of the present article is a generalization of their work. Condition (A2) has been used by Zhao and Yu (2006) while studying the sign consistency of LASSO. Recently, Zhang and Xiang (2016) and Wang and Tian (2019) assumed this condition in the context of selection consistency of adaptive group LASSO in high-dimensional linear models. The assumption on the finiteness of the group size, (A3) is also frequently used in group selection problems. See the works of Xu and Ghosh (2015) and Yang and Narisetty (2020) in this context. Since, we assume* $G_n \leq n$*, (A3) indicates that the number of groups* G_n *can be assumed of the order* n*. Note that (B1) and (B2) suggest a wide range of choices for* τ_n *to achieve selection consistency. A possible choice for* τ_n *is* $\tau_n \asymp G_n^{-1-\delta}, \delta > 0$*. In the next paragraph, we mention very briefly the key steps for proving Theorem 4.3.1.*

To establish (4.18) and (4.19), we first need to obtain some concentration inequalities based on the posterior distribution of κ_g and the posterior mean of $1-\kappa_g, g = 1, 2, \cdots, G$. These are provided as Lemmas 4.8.1-4.8.4 given in Section 4.8. Next, we are required to find upper bounds for the tail probabilities of non-central and central χ^2 random variables. For appropriate non-central χ^2, assumptions (A1)-(B1), and Mill's ratio come in handy. On the other hand, tail bounds for central χ^2 random variables, results due to Gabcke (2015) help us complete the proof of our result.

The following theorem, namely, Theorem 4.3.4, establishes the fact that the half-thresholding rule in (4.8) achieves optimal estimation rate under mild conditions. Proof of this result can be found in Section 4.8.

Theorem 4.3.4. *Consider the hierarchical framework of* (4.4) *satisfying* (4.5) *with* $a \geq \frac{1}{2}$ *and the half-thresholding (HT) rule* (4.8) *based on this. Let* $\boldsymbol{\beta}^0_{\mathcal{A}_n} = \{\boldsymbol{\beta}^0_g : g \in \mathcal{A}_n\}$ *and* $\widehat{\boldsymbol{\beta}}^{HT}_{\mathcal{A}_n} = \{\widehat{\boldsymbol{\beta}}^{HT}_g : g \in \mathcal{A}_n\}$*. Assume further that assumptions (A1)-(A3) are satisfied. Then we have the following.*

Consider the following assumptions:

(C1) There exist global constants $C_1 > 0$*,* $C_2 > 0$ *with* $0 < C_1 \leq C_2 < \infty$ *such that* $0 < C_1 \leq \frac{1}{n} e_{min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \leq \frac{1}{n} e_{max}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \leq C_2 < \infty$.

(C2) *For all* $g \in \mathcal{A}_n, \min_j |\beta^0_{gj}| > C_3$, *for some global constant* $C_3 > 0$.

(C3) *The global parameter* $\tau_n \to 0$ *such that* $G_n\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right) \to \infty$ *and* $\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right) = o(\frac{1}{|\mathcal{A}_n|})$ *as* $n \to \infty$.

Suppose further that $L(\cdot)$ *in* (4.5) *satisfies Assumption 1. Then, under assumptions* (C1)−(C3), *the resultant estimator corresponding to* (4.8) *enjoys optimal estimation rate, i.e., for any vector* $\boldsymbol{\alpha}$ *with* $||\boldsymbol{\alpha}|| = 1$ *and* $\boldsymbol{\Sigma}_{\mathcal{A}_n} = \mathbf{X}^{\mathrm{T}}_{\mathcal{A}_n}\mathbf{X}_{\mathcal{A}_n}$, *we have*

$$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}^{HT}_{\mathcal{A}_n} - \boldsymbol{\beta}^0_{\mathcal{A}_n}) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } n \to \infty.$$

Remark 4.3.5. *First, note that (C1) is a slightly stronger assumption than (A1) since it assumes an upper bound on the largest eigenvalue of* $\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n}$. *This is a common assumption in the high-dimensional variable selection literature. See, e.g., assumption (A1) of Zou and Zhang (2009), who assumed exactly the same condition while establishing the oracle property of their proposed adaptive elastic net estimator. Similarly, condition (C2) is also a stronger version of (A2) as* $b = 0$ *in (A2) corresponds to (C2). This assumption is also needed in our argument to establish asymptotic normality. It is interesting to note that, (C2) is weaker than that of (A6) of Zou and Zhang (2009) where an upper bound on the rate of growth of* $|\beta^0_j|$ *was also assumed for proving asymptotic normality of adaptive elastic net estimator. Also, observe that (B1), (B2) (used in Theorem 4.3.1) and (C3) provide some choices of* τ *for achieving both selection consistency and optimal estimation rate. One such choice of* τ *is* $\tau \asymp G_n^{-1-\delta}, 0 < \delta \leq 1$. *This choice is used as an estimator in our simulation study. It also provides an idea about the range of the prior distribution of* τ *in the case of a full Bayes procedure when one models the situation using a non-degenerate prior on* τ. *This is discussed in detail in Section 4.3.2.*

The proof of Theorem 4.3.4 involves establishing two key facts, namely,

$$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n} - \boldsymbol{\beta}^0_{\mathcal{A}_n}) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } n \to \infty, \tag{4.20}$$

and

$$\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}) \xrightarrow{P} 0 \text{ as } n \to \infty, \tag{4.21}$$

where $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n} = \{\widehat{\boldsymbol{\beta}}_g : g \in \mathcal{A}_n\}$. Finally, a simple application of Slutsky's theorem results in the proof of Theorem 4.3.4. (4.20) holds due to the results of the linear model followed by the block orthogonality of the design matrix. On the other hand, to establish (4.21) one needs to use the Cauchy-Schwarz inequality along with assumptions (C1)-(C3). In this context, some of the arguments used in Lemma 3 of Ghosh and Chakrabarti (2017) come in handy to obtain the asymptotic optimality of our proposed HT estimator.

Remark 4.3.6. *To establish asymptotic normality, we have assumed a condition on the eigenvalues of the design matrix, as given in (C1). However, a particular choice of the design matrix corresponding to the* g^{th} *group trivializes the assumption and provides the following statement immediately. The proof follows similarly and is hence omitted.*

Consider the hierarchical framework of (4.4) satisfying (4.5), and the half-thresholding (HT) rule (4.8) based on this along with an orthogonal design matrix, i.e. $\mathbf{X}^{\mathrm{T}}\mathbf{X} = nI_p$. *Let* $\boldsymbol{\beta}^0_{\mathcal{A}_n} = \{\boldsymbol{\beta}^0_g : g \in \mathcal{A}_n\}$ *and* $\widehat{\boldsymbol{\beta}}^{HT}_{\mathcal{A}_n} = \{\widehat{\boldsymbol{\beta}}^{HT}_g : g \in \mathcal{A}_n\}$. *Assume* $|\mathcal{A}_n|$ *is fixed and (A3) is satisfied along with* $L(\cdot)$ *defined in (4.5) satisfies Assumption 1. Then for all* $g \in \mathcal{A}_n$, *we have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{HT}_g - \boldsymbol{\beta}^0_g\right) \xrightarrow{d} \mathcal{N}_{m_g}(\mathbf{0}, \sigma^2 I_{m_g}) \text{ as } n \to \infty.$$

The above argument implies that, when the group size reduces to unity, our result shows that the asymptotic distribution of the half-thresholding estimator proposed by Tang et al. (2018) also achieves the optimal estimation rate.

Theorem 4.3.1 along with Theorem 4.3.4 show that our proposed half-thresholding rule (4.8) is an oracle when the global shrinkage parameter is treated as a tuning one.

4.3.2 Oracle properties of the HT procedure for the empirical Bayes and full Bayes approaches

From Theorems 4.3.1 and 4.3.4 of the previous subsection, it is evident that the choice of τ plays a crucial role in the optimality of our variable selection rule and the estimate. It was shown that an appropriate choice of the global shrinkage parameter τ based on the proportion of active groups ensures such oracle properties. But as mentioned before, this proportion may not be known a priori. In such situations, we treat the global shrinkage parameter in empirical/full Bayes ways. In the next three subsections, we discuss the properties of the empirical Bayes versions of the HT rule as in (4.13) and (4.23) and its full Bayes version as in (4.15).

Oracle properties of the HT procedure using empirical Bayes approach

As described earlier, motivated by the work of van der Pas et al. (2014), we consider an empirical Bayes estimate of τ of the form (4.12). Theorem 4.3.7 and 4.3.8 together establish the significant fact that the rule (4.13) and the corresponding estimate enjoy variable selection consistency and optimal estimation rate respectively. To the best of our knowledge, this is the first result of this kind using global-local shrinkage priors in sparse high-dimensional regression problems using the empirical Bayesian method. Proof of this result can be found in Section 4.8.

Theorem 4.3.7. *Consider the hierarchical framework of (4.4) where* $\pi(\lambda_g^2)$ *satisfies (4.5) with* $a > \frac{1}{2}$, *and the empirical Bayes half-thresholding (HT) rule (4.13) based on this. Let* $\mathcal{A}_n = \{g : \boldsymbol{\beta}^0_g \neq \mathbf{0}\}$ *and* $\widehat{\mathcal{A}}_n = \{g : \widehat{\boldsymbol{\beta}}^{HT}_{g,EB} \neq \mathbf{0}\}$ *denote respectively the set of active groups, and the set of groups declared active by the half-thresholding rule (4.13). Recall,* $\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$, *for* $g = 1, \cdots, G$. *Assume that* $|\mathcal{A}_n| = G_{\mathcal{A}_n}$ *is unknown and tends to infinity as* $n \to \infty$. *Suppose that* $L(\cdot)$ *in (4.5) satisfies Assumption 1 and assumptions (A1)-(A3) hold. Also, assume that, for* $a \geq 1$, $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some* $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$ *and when* $\frac{1}{2} < a < 1$, $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$

for some $0 < \epsilon_1 < \epsilon_2 < 1 - \frac{1}{2a}$, *then we have*

$$\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n^{(EB)} = \mathcal{A}_n) = 1 \text{ as } n \to \infty.$$

Observe that the decision rule (4.13) corresponding to an individual group g depends on the whole dataset $\mathcal{D}$ and as such the rules for different g's are dependent. Proofs of these results exploit certain ideas of van der Pas et al. (2014) and Ghosh and Chakrabarti (2017), together with some non-trivial concentration inequalities involving the central and non-central χ^2 distributions to achieve the desired upper bounds to both types of error probabilities.

Regarding assumptions (A1)-(A3), see Remark 4.3.3 above. Our other assumption is on the total (unknown) number of active groups and our result on variable selection consistency holds for different broad sparsity regimes depending on the value of a in the prior on the local shrinkage coefficients.

Now we investigate the asymptotic estimation rate of our proposed empirical Bayesian half-thresholding estimate (4.14). We want to know whether the asymptotic distribution of the linear combination of $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n,EB}$ is exactly the same as that of $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n}$. Theorem 4.3.8 below provides an affirmative answer. Proof of this result can be found in Section 4.8.

Theorem 4.3.8. *Consider the hierarchical framework of* (4.4) *satisfying* (4.5) *with* $a \geq \frac{1}{2}$, *and the empirical Bayesian half-thresholding (HT) estimator* (4.14) *based on this. Assume that* $|\mathcal{A}_n| = G_{\mathcal{A}_n}$ *satisfies* $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some* $0 \leq \epsilon_1 < \epsilon_2 < \frac{1}{2}$. *Also assume that (C1) and (C2) of Theorem 4.3.4 and that* $L(\cdot)$ *in* (4.5) *satisfies Assumption 1. Then for any vector* $\boldsymbol{\alpha}$ *with* $||\boldsymbol{\alpha}|| = 1$, *we have*

$$\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathcal{A}_n}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}^{HT}_{\mathcal{A}_n,EB} - \boldsymbol{\beta}^0_{\mathcal{A}_n}) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } n \to \infty.$$

As an immediate consequence of Theorem 4.3.8, we have the following corollary. The proof of this follows using the same set of arguments as used in Theorem 4.3.8 and is hence skipped.

Corollary 4.3.9. *Consider the situation of Theorem 4.3.8 along with an orthogonal design matrix, that is,* $\mathbf{X}^{\mathrm{T}}\mathbf{X} = nI_p$. *Assume* $|\mathcal{A}_n|$ *is fixed and (A3) is satisfied and that* $L(\cdot)$ *satisfies Assumption 1. Then for all* $g \in \mathcal{A}_n$, *we have*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^{HT}_{g,EB} - \boldsymbol{\beta}^0_g\right) \xrightarrow{d} \mathcal{N}_{m_g}(\mathbf{0}, \sigma^2 I_{m_g}) \text{ as } n \to \infty.$$

HT method based on modified Empirical Bayes Estimator of τ

It may be noted that the variable selection consistency of the empirical Bayes version of our variable selection rule using the estimator (4.12) of τ was only proved when $a > \frac{1}{2}$ in the prior (4.5) for the local shrinkage parameter. However, a similar result for the case $a = \frac{1}{2}$, which, for instance, corresponds to the horseshoe prior, could not be theoretically established using the same technique. Our simulation results are very good even for $a = \frac{1}{2}$ and are indicative of the fact even in this case. So in search for an empirical Bayes estimator of τ that can be shown

to have an oracle property for all $a \geq \frac{1}{2}$, we need to dig a little deeper into the basic intuition and also the technical aspects of the proof for the $a > \frac{1}{2}$ case. The empirical Bayes estimator (4.12) used by us was motivated by a similar estimator of van der Pas et al. (2014) as in (4.11). It may be noted that the term $\frac{1}{c_2 G_n}\sum_{g=1}^{G_n} 1\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_g\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n\right)$ may intuitively be thought of as an estimator (or an estimated lower bound) of $\frac{G_{\mathcal{A}_n}}{G_n}$, the proportion of active groups. Our proof reveals that we need to ensure $G_n \tau_n^{2a}[\log(\frac{1}{\tau})]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$, where s denotes the maximum group sizes. For the $a > \frac{1}{2}$ case, $\tau_n = \frac{G_{\mathcal{A}_n}}{G_n}$ satisfies the condition, and it is quite intuitive that the empirical Bayes version of the HT procedure using $\widehat{\tau}$ can be shown to have variable selection consistency. Intuitively, for $a = \frac{1}{2}$, by choosing $\tau_n = \frac{G_{\mathcal{A}_n}}{G_n}$ the above condition is not satisfied, but the choice $\tau_n = (\frac{G_{\mathcal{A}_n}}{G_n})^{1+\delta}$ for any $\delta > 0$ works. This gives us the clue that if τ is estimated using a statistic which can be thought of as an estimator (or at least an estimated lower bound) of $(\frac{G_{\mathcal{A}_n}}{G_n})^{1+\delta}$ for some $\delta > 0$, we might have the desired result. Based on this, we consider a modified version of our early estimate as follows

$$(\widehat{\tau}^{\mathrm{EB}})^{\frac{1}{2}} = \max\left\{\frac{1}{G_n}, \frac{1}{c_2 G_n}\sum_{g=1}^{G_n} 1\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_g\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n\right)\right\}. \tag{4.22}$$

Hence, our proposed modified data-adaptive decision rule is given by

$$\text{The } g^{th}\text{group is considered active if } E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > 0.5, \text{ for } g = 1, 2, \cdots, G, \tag{4.23}$$

where $\widehat{\tau}^{\mathrm{EB}}$ is defined in (4.22). Our next theorem shows that, indeed, the empirical Bayes version of the HT procedure with the above estimate (4.22) of τ enjoys variable selection consistency. Proof of this result can be found in Section 4.8.

Theorem 4.3.10. *Consider the hierarchical framework of (4.4) satisfying (4.5) with $a \geq \frac{1}{2}$, and the half-thresholding (HT) rule (4.23) based on an empirical Bayes estimate $\widehat{\tau}^{EB}$ given in (4.22). Let $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \neq \mathbf{0}\}$ and $\widehat{\mathcal{A}}_n^{(EBmod)} = \{g : \widehat{\boldsymbol{\beta}}_{g,EB}^{HT} \neq \mathbf{0}\}$ denote respectively the set of truly active groups, and the set of groups declared active by the half-thresholding rule (4.23). Suppose that $L(\cdot)$ in (4.5) satisfies Assumption 1 and assumptions (A1)-(A3) hold. Also, assume that, for $a \geq 0.5$, $G_{\mathcal{A}_n} = O(G_n^{\epsilon})$ with $0 < \epsilon < \frac{1}{2}$, then we have*

$$\lim_{n\to\infty} P(\widehat{\mathcal{A}}_n^{(EBmod)} = \mathcal{A}_n) = 1 \text{ as } n \to \infty.$$

Remark 4.3.11. *The decision rule (4.23) based on the modified empirical Bayes estimator of τ given in (4.22) not only achieves the variable selection consistency but also the corresponding estimator attains the optimal estimation rate, that is, the statement of Theorem 4.3.8 still holds just by replacing the definition of $\widehat{\tau}^{EB}$ given in (4.12) with (4.22) in (4.13). Hence, this also implies that our proposed decision rule (4.23) with $\widehat{\tau}^{EB}$ defined in (4.22) is oracle.*

Oracle properties of the HT procedure using full Bayes approach

We now motivate and present the full Bayes approach as an alternative solution to the empirical Bayes approach mentioned in the previous subsection when $G_{\mathcal{A}_n}$ is unknown. As already proved in Theorem 4.3.1, using τ as a tuning parameter, our proposed decision rule (4.8) results in consistency in variable selection. On the other hand, Theorem 4.3.7 shows that the data-adaptive decision rule (4.13) using an empirical Bayes estimate of τ still can figure out all the true active groups present in the model. The next obvious question is whether similar results still hold in a full Bayes treatment. To answer this question for the full Bayes approach, we assume a nondegenerate prior on τ and assume the following condition on the range of prior density of τ.
D1: The prior $\Pi(\cdot)$ satisfies $\int_{\gamma_{1n}}^{\gamma_{2n}} \pi(\tau)d\tau = 1$ where γ_{1n} and γ_{2n} are positive sequences satisfying $\frac{\gamma_{2n}}{\gamma_{1n}} \to \infty$ as $n \to \infty$, $G_n\gamma_{2n}[\log\left(\frac{1}{\gamma_{2n}}\right)]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$, $\log\left(\frac{1}{\gamma_{1n}}\right) \asymp \log(G_n)$, $G_n\sqrt{\gamma_{1n}}\log\left(\frac{1}{\gamma_{1n}}\right) \to \infty$ as $n \to \infty$ and $\sqrt{\gamma_{1n}}\log\left(\frac{1}{\gamma_{1n}}\right) = o(\frac{1}{|\mathcal{A}_n|})$ as $n \to \infty$.

The motivation behind D1 comes from Theorems 4.3.1 and 4.3.4. Note that, both (B1) and (B2) provide some asymptotic order of τ to achieve selection consistency when τ is used as a tuning parameter. This observation motivates us to study whether the decision rule (4.15) can produce selection consistency of the underlying model when a non-degenerate prior on τ satisfying such conditions is considered. Similar to Theorem 4.3.8, we are also interested in studying the asymptotic distribution of the posterior mean of the group coefficient under the full Bayes approach. Our next Theorem affirmatively answers these questions. Proof of this result can be found in Section 4.8.

Theorem 4.3.12. *Consider the hierarchical framework of* (4.4)*, and the half-thresholding (HT) rule* (4.15) *where* $\pi(\lambda_g^2)$ *satisfies* (4.5) *with* $a \geq \frac{1}{2}$*,* $L(\cdot)$ *in* (4.5) *satisfies Assumption 1 and* τ *is assumed to have a non-degenerate prior distribution satisfying D1 above. Let* $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \neq \mathbf{0}\}$ *and* $\widehat{\mathcal{A}}_n = \{g : \widehat{\boldsymbol{\beta}}_{g,FB}^{HT} \neq \mathbf{0}\}$ *denote respectively the set of active groups, and the set of groups declared active by the half-thresholding rule* (4.15)*. Define,* $\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$*, for* $g = 1, \cdots, G$*. Then, under the assumptions (C1) and (C2) and that* $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some* $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$*, with* $0 < \epsilon < \frac{1}{2}$*, the decision rule* (4.15) *is an oracle.*

This result ensures that the decision rule (4.15) based on any non-degenerate proper prior on τ defined in our proposed support as given in D1 can be used as an alternative solution to the empirical Bayes approaches to provide similar results both in terms of selection consistency and optimal estimation rate. It is noteworthy that if one is interested in establishing selection consistency only, one can establish that using slightly weaker assumptions (A1) and (A2) instead of (C1) and (C2) and may assume that γ_{1n} satisfies $\log\left(\frac{1}{\gamma_{1n}}\right) \asymp \log(G_n)$ only in place of satisfying both $\log\left(\frac{1}{\gamma_{1n}}\right) \asymp \log(G_n)$ and (C3). However, we need to have stronger assumptions slightly stronger assumptions to prove selection consistency and optimal estimation rate, simultaneously. Tang et al. (2018) also studied their proposed half-thresholding rule using a non-degenerate prior on τ supported on some suitable range based on the sample size n. However, the main drawback of their approach was the assumption that the number of active variables is fixed, a condition that is rarely satisfied in high-dimensional situations. On the other hand, the optimality of our

rule (4.15) is proved without that assumption. Hence, when the group size reduces to unity for all groups, Theorem 4.3.12 confirms that our proposed rule is still an oracle without any strong assumption on the growth of the active variables. To our knowledge, this is the first result of this kind in the full Bayes approach in the literature of global-local priors when the number of active groups grows with increasing sample size n. In this way, in a sparse high-dimensional group selection problem, Theorem 4.3.12 establishes the fact that a carefully chosen broad class of global-local priors can provide optimal results even if the level of sparsity is unknown.

Remark 4.3.13. *If the support of τ is chosen such that it is concentrated in a small region covering the optimal range of values (as observed when it is used as a tuning parameter), the posterior of τ is also expected to take values near the optimal range with high probability. As a result, using arguments similar to those of Theorems 4.3.1 and 4.3.4, the oracle property of the full Bayes approach can be established. This observation is crucial for proving the optimality results for the full Bayes approach when the sparsity level is unknown. This raises the question if one can obtain similar optimality results by employing a non-informative prior on τ, as was originally proposed for the normal means model for defining the Horseshoe prior (Carvalho et al. (2009), Polson and Scott (2010)). However, establishing such results seems quite challenging, and would require us to come up with non-trivial arguments regarding the concentration of the posterior distribution of τ near the optimal value. Inspired by the intuitive appeal of the prior taken by us, we wanted to formally compare this prior with a non-informative prior in a simulation study to see if it can provide any statistical justification for the use of our prior on τ in addition to its obvious technical appeal. In our simulation study, we consider two approaches for the full Bayes procedure, one using the half-Cauchy prior on a truncated range and the other using the same half-Cauchy prior on the whole range. We observe that better results (in terms of misclassification probability and False Positive Rate) are obtained for the truncated version than the untruncated one. The results are summarized in Tables 1-9. These results make a strong case for the use of our prior in a real problem of interest, in addition to its technical attractiveness.*

4.4 Property of the Half-Thresholding (HT) estimator

Along with the variable selection consistency, one may also be interested in studying the accuracy of the half-thresholding(HT) estimator with respect to an appropriate loss function. Theorem 4.4.1, stated below, indicates that in an extremely sparse situation with no signals or in the situation where the true group coefficients are allowed to grow only up to a polynomial rate in n, our HT estimator has a clear edge over the traditional Bayes estimator. Proof of this result can be found in Section 4.8.

Theorem 4.4.1. *Consider the hierarchical framework of (4.4) where $\pi(\lambda_g^2)$ is as in (4.5) with $a \geq \frac{1}{2}$, $L(\cdot)$ in (4.5) satisfies Assumption 1 and the half-thresholding (HT) rule (4.8) based on these. Let $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \neq \mathbf{0}\}$ and $\widehat{\mathcal{A}}_n = \{g : \widehat{\boldsymbol{\beta}}_g^{HT} \neq \mathbf{0}\}$ denote respectively the set of truly active groups, and the set of groups declared active by the half-thresholding rule (4.8). Let*

$\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$, *for* $g = 1, \cdots, G$. *Then, we have the following.*
(i) In a sparse situation, under the assumption (A3), if all groups are inactive, our proposed HT estimator of $\boldsymbol{\beta}_g$ *is more accurate than that of the posterior mean with respect to squared error loss, i.e., for all sufficiently large* n,

$$\sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{HT} - \boldsymbol{\beta}_g^0\right\|_2^2 < \sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{PM} - \boldsymbol{\beta}_g^0\right\|_2^2. \tag{4.24}$$

(ii) Along with (A2)-(A4) and (C1), we further assume that for any active group $g \in \mathcal{A}_n$, $\max_j |\beta_{gj}^0| \le n^{C_3}$, *for some* $0 < C_3 < \infty$. *Then also our proposed HT estimator of* $\boldsymbol{\beta}_g$ *is more accurate than that of the posterior mean with respect to squared error loss, i.e., for all sufficiently large* n,

$$\sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{HT} - \boldsymbol{\beta}_g^0\right\|_2^2 < \sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{PM} - \boldsymbol{\beta}_g^0\right\|_2^2. \tag{4.25}$$

We end this subsection with the following remark.

Remark 4.4.2. *Theorem 4.4.1 establishes that the HT estimator is more accurate than the traditional Bayes estimator if the conditions (A2)-(A4) and (C1) hold. Note that (A4) implies that the level of sparsity is known and hence the global parameter* τ_n *is used as a tuning one. Hence, a follow-up question is whether the same result holds when the level of sparsity is unknown, when either empirical Bayes or a full Bayes treatment on* τ *is employed. The next two corollaries provide a positive answer to this question. Proofs of these follow from arguments used in Theorem 4.4.1 and Theorems 4.3.7 and 4.3.12, and hence are skipped.*

Corollary 4.4.3. *Consider the hierarchical framework of (4.4) where* $\pi(\lambda_g^2)$ *is as in (4.5) with* $a \ge \frac{1}{2}$ *and the half-thresholding (HT) rule (4.13) based on these. Let* $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \ne \mathbf{0}\}$ *and* $\widehat{\mathcal{A}}_n^{(EB)} = \{g : \widehat{\boldsymbol{\beta}}_{g,EB}^{HT} \ne \mathbf{0}\}$ *denote respectively the set of truly active groups, and the set of groups declared active by the half-thresholding rule (4.13). Let* $\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$, *for* $g = 1, \cdots, G$. *Along with (A2)-(A4) and (C1), we further assume that* $|\mathcal{A}_n| = G_{\mathcal{A}_n}$ *is unknown and* $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some unknown* $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$, *with any active group* $g \in \mathcal{A}_n$ *satisfies* $\max_j |\beta_{gj}^0| \le n^{C_3}$, *for some* $0 < C_3 < \infty$. *Then also our proposed empirical Bayes HT estimator of* $\boldsymbol{\beta}_g$ *is more accurate than that of the posterior mean with respect to squared error loss, i.e., for all sufficiently large* n,

$$\sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_{g,EB}^{HT} - \boldsymbol{\beta}_g^0\right\|_2^2 < \sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{PM} - \boldsymbol{\beta}_g^0\right\|_2^2.$$

Corollary 4.4.4. *Consider the hierarchical framework of (4.4) where* $\pi(\lambda_g^2)$ *is as in (4.5) with* $a \ge \frac{1}{2}$ *and the half-thresholding (HT) rule (4.15) based on these and* τ *is assumed to have a non-degenerate prior distribution satisfying (D1). Let* $\mathcal{A}_n = \{g : \boldsymbol{\beta}_g^0 \ne \mathbf{0}\}$ *and* $\widehat{\mathcal{A}}_n^{(FB)} = \{g : \widehat{\boldsymbol{\beta}}_{g,FB}^{HT} \ne \mathbf{0}\}$ *denote respectively the set of truly active groups, and the set of groups declared active by the half-thresholding rule (4.15). Let* $\mathbf{Q}_{n,g} = \mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g/n$, *for* $g = 1, \cdots, G$. *Along with*

(A2), (A3) and (C1), we further assume that $|\mathcal{A}_n| = G_{\mathcal{A}_n}$ *is unknown and* $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ *for some unknown* $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$*, with any active group* $g \in \mathcal{A}_n$*, satisfies* $\max_j |\beta_{gj}^0| \leq n^{C_3}$*, for some* $0 < C_3 < \infty$*. Then also our proposed full Bayes HT estimator of* $\boldsymbol{\beta}_g$ *is more accurate than that of the posterior mean with respect to squared error loss, i.e., for all sufficiently large* n*,*

$$\sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0} \left\| \widehat{\boldsymbol{\beta}}_{g,FB}^{HT} - \boldsymbol{\beta}_g^0 \right\|_2^2 < \sum_{g=1}^{G_n} E_{\boldsymbol{\beta}_g^0} \left\| \widehat{\boldsymbol{\beta}}_g^{PM} - \boldsymbol{\beta}_g^0 \right\|_2^2.$$

4.5 Simulations

In this section, we report the performance of our proposed rules in a detailed simulation study and compare that with other existing methods. We simulate data from the following true model:

$$\mathbf{y} = \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n). \tag{4.26}$$

The construction of the design matrix $\mathbf{X}_g$ is discussed separately for $n > p$ and $p > n$ below. The group coefficients $\boldsymbol{\beta}_g$ are either null or non-null. Several choices of null and non-null coefficients are considered in different simulation schemes. Different scenarios based on sample size (small, $n = 50$, moderate, $n = 200$ and large, $n = 500$), number of covariates (p), number of groups, and sparsity levels are considered. We also consider different choices of regression coefficients.

Each group regression coefficient $\boldsymbol{\beta}_g$ is modeled by a global-local shrinkage prior as given in (4.4). A standard half-Cauchy prior is used for the local shrinkage coefficient corresponding to each group, i.e. $\lambda_g \overset{iid}{\sim} C^+(0,1)$. Note that this choice of prior is included in the hierarchical formulation (4.4) satisfying (4.5) for $a = \frac{1}{2}$. Together with (4.4), we also consider the traditional modeling of $\boldsymbol{\beta}_g$ as $\boldsymbol{\beta}_g | \lambda_g^2, \sigma^2, \tau^2 \overset{ind}{\sim} \mathcal{N}_{m_g}(\mathbf{0}, \lambda_g^2 \sigma^2 \tau^2 I_{m_g})$. These two formulations are named Modified Group Horseshoe and Normal Group Horseshoe, respectively. When knowledge of the proportion of active groups is available, τ is used as a tuning parameter. Theorem 4.3.1 provides one choice of τ as $\tau_n^2 = (\frac{G_{\mathcal{A}_n}}{G_n})^{2+\delta}$ for any $\delta > 0$. Here we use $\tau_n^2 = (\frac{G_{\mathcal{A}_n}}{G_n})^{2.1}$ i.e. $\delta = 0.1$. Simulations with other choices of δ have also been conducted. When this knowledge is not available, we also consider empirical Bayes and full Bayes treatments of τ. In case of the empirical Bayes procedure, we take $c_1 = 2, c_2 = 1$ and σ equal to 1 in the definitions of $\widehat{\tau}^{\text{EB}}$, given in (4.12) and (4.22). The resulting estimators are named Modified group horseshoe EB1 (based on (4.12)) and Modified group horseshoe EB2 (based on (4.22)), respectively. Like the tuning parameter, the empirical Bayes estimate of τ also involves two constants c_1 and c_2. To study which choice of (c_1, c_2) should be used, additional simulation results involving different choices of (c_1, c_2) are added. For the full Bayes procedure, too, we consider two approaches, one by using the restriction of a standard half-Cauchy prior on τ ($\tau \sim C^+(0,1)$) on $[G^{-1.1}, G^{-1.1} \log G]$, and the other being the same half-Cauchy prior on τ on its full range. The resulting procedures are termed as Modified group horseshoe FBtruncated and Modified group horseshoe FBfull, respectively. Along with the horseshoe prior, we also consider Three parameter beta normal

(TPBN) prior using τ_n as the tuning parameter and other two hyperparameters (α_1, α_2) as $(\alpha_1, \alpha_2) = (0.5, 1)$ and $(\alpha_1, \alpha_2) = (1, 0.5)$, respectively. The resultant procedures are termed as Modified TPBN (α_1, α_2). Note that, since the simulation situations considered here are beyond the block-orthogonality assumption in the design matrix $\mathbf{X}$, we consider a group to be active or inactive using the decision rule (4.9). Also, note that the posterior mean of the group coefficient involves a choice of τ, and hence different procedures mentioned above regarding the choice of τ play a crucial role. In the case of a block-diagonal design matrix, decision rule (4.9) simplifies to (4.8). Similarly, we also use (4.13) and (4.15) for the empirical Bayes and full Bayes versions specifically for a block-diagonal design matrix. We are also going to repeat the simulation performances of the Group Spike and Diffusing prior of Yang and Narisetty (2020) (hereby named as GSD-SSS) on the group regression coefficients and the estimates computed from shotgun stochastic search algorithm(SSS) and the Bayesian Group LASSO with Spike and Slab prior (BGL-SS) due to Xu and Ghosh (2015) for comparing the performance of our rules based on one-group shrinkage prior. for the same purpose, we will also consider Group LASSO of Yuan and Lin (2006) group MCP introduced by Breheny and Huang (2009) and group SCAD due to Wang et al. (2007). The misclassification probability (MP), the false positive rate (FPR), and the true positive rate (TPR) corresponding to each of the aforementioned procedures will be compared.
All simulation situations considered here can be broadly classified into two cases, namely $n > p$ and $p > n$.

Case-1: First we consider the cases where $n > p$. For each group, each row of the design matrix $\mathbf{X}_g$ is generated from a multivariate normal distribution such that the components have zero mean and unit variance and are correlated with pairwise correlation ρ. Two values of ρ are chosen, namely 0 and 0.5, which indicate that the predictors within a group are uncorrelated and moderately correlated, respectively. In the following examples, different choices of (n, p) along with different signal strengths and different sparsity levels are to be considered.

- Example 1. We start the simulation study with a small sample size. Here, we consider a situation when sample size $n = 50$ and $p = 20$ covariates are grouped in 10 groups containing 2 covariates each. Regarding the group coefficients, we consider two situations based on the strength of the active coefficients. When the strength is weak, we let $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0.2}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$. On the other hand, for strong signal, we assume $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$. In both situations, $\mathbf{0}$ and $\mathbf{0.2}, \mathbf{1}$ are vectors of length 2, with all elements 0 or 0.2 and 1, respectively. Since the group size is 2, each row of $\mathbf{X}_g$ is generated from a $\mathcal{N}_2(0, 0, 1, 1, \rho)$ where $\mathcal{N}_2(\cdot)$ denotes a bivariate normal distribution.

- Example 2. Next, we consider the moderate sample size case. We take $n = 200$ and assume that $p = 40$ covariates are grouped in 10 groups containing 4 covariates each. We assume only the first group is active. In this case too, we consider two situations based on signal strength. Group coefficients for these two cases are, $\boldsymbol{\beta} = ((0.1, 0.2, 0.3, 0.4), \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{0})$ for weak signal strength, and $\boldsymbol{\beta} = ((1.1, 1.2, 1.3, 1.4), \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{0})$ for strong signal strength.

Here $\mathbf{0}$ is a null vector of length 4. The data generating scheme is similar to Example 1 except for necessary dimension changes.

- Example 3. For this example, we consider the case when the group sizes are different. Let us consider the scenario when $n = 200$ and $p = 50$ predictors are grouped in 16 groups with group sizes 4,3,3,2,2,2,2,2,2,4,4,4,4,2,5 and 5 respectively. Let $\boldsymbol{\beta} = ((0.1, 0.2, 0.3, 0.4), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, (0, 0.4, 0, 0), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$ for weak signal strength and $\boldsymbol{\beta} = ((1, 2, 3, 4), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, (1, 1.1, 1.2, 1.3), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$ for strong signal. In both of the cases, the first two $\mathbf{0}$ denote null vectors of length 3, the remaining are of length 2 except the last two which are of length 5. In this case, also, the predictors are generated in the same way as in Example 1 except for necessary dimension changes.

- Example 4. Next, consider a situation when the sample size is large. Suppose the sample size $n = 500$ and $p = 100$ covariates are grouped into 25 groups containing 4 covariates each. We assume that only one group is active and the group coefficients are, $\boldsymbol{\beta} = ((0.1, 0.2, 0.3, 0.4), \mathbf{0}, \cdots, \mathbf{0}, \mathbf{0})$ when the signal strength is weak and $\boldsymbol{\beta} = ((1.1, 1.2, 1.3, 1.4), \mathbf{0}, \cdots, \mathbf{0}, \mathbf{0})$ when the signal strength is strong. Here $\mathbf{0}$ is a null vector of length 4. The data generated scheme is similar to Example 1 except for necessary dimension changes.

For the next example, we consider the design matrices to be block-diagonal. Suppose $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_G})$, which is generated as previously mentioned. From $\mathbf{X}$ a block-diagonal matrix $\mathbf{Z} = (\mathbf{Z_1}, \mathbf{Z_2}, \cdots, \mathbf{Z_G})$ is obtained as

$$\begin{aligned}
\mathbf{Z_1} &= \mathbf{X_1} \\
\mathbf{Z_2} &= (\mathbf{I_n} - P_{\mathbf{Z_1}})\mathbf{X_2} \\
\mathbf{Z_3} &= (\mathbf{I_n} - P_{\mathbf{Z_1}} - P_{\mathbf{Z_2}})\mathbf{X_3} \\
&\cdots\cdots \\
&\cdots\cdots \\
\mathbf{Z_G} &= (\mathbf{I_n} - P_{\mathbf{Z_1}} - P_{\mathbf{Z_2}} - \cdots - P_{\mathbf{Z_{G-1}}})\mathbf{X_G},
\end{aligned} \tag{4.27}$$

where $\mathbf{I_n}$ denotes the identity matrix of order $n \times n$ and $P_{\mathbf{Z_g}} = \mathbf{Z_g}(\mathbf{Z}_g^{\mathrm{T}}\mathbf{Z}_g)^{-1}\mathbf{Z}_g^{\mathrm{T}}$ denotes the projection matrix on the column space of $\mathbf{Z}_g$. Now, consider the situation where the data is simulated from the following true model: $\mathbf{y} = \sum_{g=1}^{G} \mathbf{Z}_g\boldsymbol{\beta}_g + \boldsymbol{\epsilon}$, where $\mathbf{Z}_g$'s are generated from (4.27) and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$.

- Example 5. We revisit Example 1 where the design matrix now becomes $\mathbf{Z}$ instead of $\mathbf{X}$, whose columns are generated using (4.27). Here, sample size $n = 50$ and $p = 20$ covariates are grouped in 10 groups containing 2 covariates each. Let $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0.2}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$ where $\mathbf{0}$ and $\mathbf{0.2}$ are vectors of length 2, with all elements 0 or 0.2, respectively.

- Example 6. Now we revisit example 2 where the design matrix now becomes $\mathbf{Z}$ instead of $\mathbf{X}$, whose columns are generated using (4.27). Here, $n = 200$ and $p = 40$ covariates are grouped into 10 groups containing 4 covariates each. We assume that only the first

group is active and the coefficients are $\boldsymbol{\beta} = ((1.1, 1.2, 1.3, 1.4), \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{0})$ where $\mathbf{0}$ is a null vector of length 4. The data generated scheme is similar to Example 3 except for necessary dimension changes.

Case-2: Next we are interested in situations when $p > n$. We define the jth predictor in group g as $X_{gj} = Z_{gj} + Z_g$, where Z_g and Z_{gj} are independent standard normal variates. Thus, predictors within a group are correlated with a pairwise correlation of 0.5 while the predictors in different groups are independent.

- Example 7. Here, we consider a case where $n = 50$ and $p = 100$. 100 predictors are grouped into 25 groups of 4 covariates each. Let $\boldsymbol{\beta} = (\mathbf{0}, \cdots, \mathbf{0}, \mathbf{0.4}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0})$ when signal strength is weak, and $\boldsymbol{\beta} = (\mathbf{0}, \cdots, \mathbf{0}, \mathbf{1.2}, \mathbf{0}, \cdots, \mathbf{0})$ for strong signal strength. Here the $\mathbf{0}$ denotes a null vector of length 4.

- Example 8. Consider an example of a large p small n problem with $n = 50$ and $p = 100$. 100 predictors are grouped into 20 groups of 5 covariates each. For weak signal strength, the group coefficients are assumed to be, $\boldsymbol{\beta} = (\mathbf{0}, \cdots, \mathbf{0}, \mathbf{0.4}, \mathbf{0.5}, (0.65, 0.60, 0.55, 0.50, 0.45), \mathbf{0}, \cdots, \mathbf{0})$. On the other hand, when the signal strength is strong, we assume $\boldsymbol{\beta} = (\mathbf{1}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{1.2}, \mathbf{1.4}, \mathbf{0}, \cdots, \mathbf{0})$. In each case $\mathbf{0}$ denotes a null vector of length 5.

- Example 9. This is another example for large p small n problem with $n = 50$ and $p = 100$. Unlike previous situations, for the same combination of (n, p), we consider two situations where the level of sparsity in the second situation is twice the former one. Here 100 predictors are grouped into 25 groups of 4 covariates each. First, we assume 4 groups to be active out of 25 groups. Let $\boldsymbol{\beta} = (\mathbf{0}, \cdots, \mathbf{0}, \mathbf{1.8}, \mathbf{0.5}, (0.65, 0.60, 0.55, 0.50), \mathbf{0}, \cdots, \mathbf{0}, \mathbf{2.5})$ where the $\mathbf{0}$ denote null vector of length 4. Next, we assume 8 groups to be active out of 25 groups. Let $\boldsymbol{\beta} = (\mathbf{1.8}, \mathbf{1.5}, \mathbf{0}, \cdots, \mathbf{0}, \mathbf{1.8}, \mathbf{1.5}, (0.65, 0.60, 0.55, 0.50), \mathbf{0}, \cdots, \mathbf{0}, \mathbf{2.5}, (0.4, 0.45, 0.50, 0.55), \mathbf{2.5})$.

- From Theorem 4.3.1, the choice of $\tau_n^2 = (\frac{G_{\mathcal{A}_n}}{G_n})^{2+\delta}$, for any $\delta > 0$ ensures model selection consistency. A question that arises here is how to choose that δ to optimize the performance. The same goes for the choice of (c_1, c_2), also, for the empirical Bayes estimates of τ, defined in the main paper. To provide answers to these questions, we have done some simulation studies corresponding to Examples 1 and 2 of the main document with choice of τ_n obtained from $\tau_n^2 = (\frac{G_{\mathcal{A}_n}}{G_n})^{2+\delta}$, with $\delta = 0.1, 0.2, \cdots, 0.5$. The results are stated in Tables 10 and 11, respectively. On the other hand, for the empirical Bayes estimate, in order to obtain a proper choice of (c_1, c_2), we have repeated Examples 1 and 2 of the main document by keeping $c_1 = 2$ and varying c_2 over $\{1, \ldots, 5\}$ and keeping $c_2 = 1$ and varying c_1 over $\{2, 3, 4, 5\}$. The corresponding results are summarized in Tables 12-15.

- In order to study the importance of the choice of τ based on the proportion of active groups, we revisited Examples 8 and 9, where along with the choices of τ previously mentioned, we consider other choices of τ which depend on *only* G_n. Here, we use five choices of τ_n, $\tau_n = \tau_{1n} = \left(\frac{G_{\mathcal{A}_n}}{G_n}\right)^{1.05}$, $\tau_n = \tau_{2n} = \frac{1}{G_n^2}$, and $\tau_n = \tau_{3n} = G_n^{-1.4}$, $\tau_n = \tau_{4n} = G_n^{-1.6}$,

and $\tau_n = \tau_{5n} = \widehat{\tau}^{\mathrm{EB(Mod)}}$, where $\widehat{\tau}^{\mathrm{EB(Mod)}}$ is the modified version of the empirical Bayes estimate of τ defined in (4.22). We have also considered the full Bayes procedure based on half-Cauchy prior on a truncated range. The results are presented in Tables 16 and 17.

4.5.1 Gibbs Sampling

Within the hierarchical form (4.4) and using the prior distributions on τ and σ^2 stated before, the Gibbs samples are drawn from the full conditional distributions as follows:

(1) Sampling from the Posterior Distribution of $\boldsymbol{\beta}_g$:

Since, the full posterior distribution of $\boldsymbol{\beta}$ given $(\boldsymbol{\lambda}^2, \sigma^2, \tau^2, \mathcal{D})$ is

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\lambda}^2, \sigma^2, \tau^2, \mathcal{D}) \propto \exp\left[-\frac{1}{2\sigma^2} \left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{y} + \sum_{g=1}^{G} \frac{\boldsymbol{\beta}_g^{\mathrm{T}} \mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g \boldsymbol{\beta}_g}{\lambda_g^2 \tau^2} \right) \right],$$

so, we obtain for $g = 1, 2, \cdots, G$,

$$\boldsymbol{\beta}_g \mid (\boldsymbol{\beta}_{-g}, \boldsymbol{\lambda}^2, \sigma^2, \tau^2, \mathcal{D}) \sim \mathcal{N}_{m_g}(\boldsymbol{\mu}_g, \sigma^2 \boldsymbol{\Sigma}_g),$$

with $\boldsymbol{\mu}_g = (1 + \frac{1}{\lambda_g^2 \tau^2})^{-1} (\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1} (\mathbf{X}_g^{\mathrm{T}} \mathbf{y} - \sum_{g'(\neq g)=1}^{G} \mathbf{X}_g^{\mathrm{T}} \mathbf{X}_{-g} \boldsymbol{\beta}_{-g})$ and
$\boldsymbol{\Sigma}_g = (1 + \frac{1}{\lambda_g^2 \tau^2})^{-1} (\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1} = (1 - \kappa_g)(\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1}$.
With the additional assumption on the block-orthogonality of the design matrix $\mathbf{X}$, we have for $g = 1, 2, \cdots, G$,

$$\boldsymbol{\beta}_g \mid (\boldsymbol{\lambda}^2, \sigma^2, \tau^2, \mathcal{D}) \overset{ind}{\sim} \mathcal{N}_{m_g}(\boldsymbol{\mu}_g, \sigma^2 \boldsymbol{\Sigma}_g),$$

with $\boldsymbol{\mu}_g = (1 - \kappa_g) \widehat{\boldsymbol{\beta}}_g$ and $\boldsymbol{\Sigma}_g = (1 - \kappa_g)(\mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g)^{-1}$.

(2) Sampling from the Posterior Distribution of σ^2:

The full posterior distribution of σ^2 conditioned on $(\boldsymbol{\beta}, \boldsymbol{\lambda}^2, \tau^2, \mathcal{D})$ is given by

$$\pi(\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\lambda}^2, \tau^2, \mathcal{D}) \propto (\sigma^2)^{-(\frac{n}{2} + \sum_{g=1}^{G} \frac{m_g}{2} + 1)} \times \exp\left[-\frac{1}{\sigma^2} \left\{ \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} + \sum_{g=1}^{G} \frac{\boldsymbol{\beta}_g^{\mathrm{T}} \mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g \boldsymbol{\beta}_g}{2\lambda_g^2 \tau^2} \right\} \right].$$

Hence,

$$\sigma^2 \mid (\boldsymbol{\beta}, \boldsymbol{\lambda}^2, \tau^2, \mathcal{D}) \sim \text{Inverse Gamma} \left(\frac{n}{2} + \sum_{g=1}^{G} \frac{m_g}{2}, \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} + \sum_{g=1}^{G} \frac{\boldsymbol{\beta}_g^{\mathrm{T}} \mathbf{X}_g^{\mathrm{T}} \mathbf{X}_g \boldsymbol{\beta}_g}{2\lambda_g^2 \tau^2} \right).$$

(3) Sampling from the Posterior Distribution of λ_g^2:

Observe that, for each $g = 1, 2, \cdots, G$,

$$\pi(\lambda_g^2 \mid \boldsymbol{\beta}_g, \sigma^2, \tau^2, \mathcal{D}) \propto (\lambda_g^2)^{-\frac{(m_g+1)}{2}}(1+\lambda_g^2)^{-1} \times \exp\left[-\frac{1}{\lambda_g^2} \cdot \frac{\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g\boldsymbol{\beta}_g}{2\tau^2\sigma^2}\right]$$

Using the Slice-sampling approach of Damien et al. (1999), posterior sampling is done in two steps:

1. Given λ_g^2, sample u_g from the Uniform distribution supported over the interval $(0, 1+\lambda_g^2)$.
2. For given $(\boldsymbol{\beta}_g, \sigma^2, \tau^2, \mathcal{D})$, sample λ_g^2 from an inverse-gamma distribution with parameters $\frac{(m_g-1)}{2}$ and $\frac{\boldsymbol{\beta}_g^{\mathrm{T}}\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g\boldsymbol{\beta}_g}{2\tau^2\sigma^2}$, truncated over the interval $(0, \frac{1-u_g}{u_g})$.

(4) Sampling from the Posterior Distribution of τ:

The posterior distribution of τ given $\boldsymbol{\Lambda}, \mathbf{y}, \mathbf{X}$ is of the form :

$$\pi(\tau \mid \boldsymbol{\Lambda}, \mathbf{y}, \mathbf{X}) \propto \frac{1}{1+\tau^2}\left|\mathbf{M}_\tau\right|^{-\frac{1}{2}}(\mathbf{y}^{\mathrm{T}}\mathbf{M}_\tau^{-1}\mathbf{y})^{-\frac{n}{2}},$$

where $\mathbf{M}_\tau = \mathbf{I}_n + \tau^2\mathbf{X}\mathbf{W}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{W} = diag(\lambda_1^2(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_1)^{-1}, \cdots, \lambda_G^2(\mathbf{X}_G^{\mathrm{T}}\mathbf{X}_G)^{-1})$. Following the approach of Johndrow et al. (2020), samples are drawn from the above posterior distribution of τ using the Metropolis-Hastings algorithm as follows:

$$\text{proposal density: } \log(\zeta^*) \sim \mathcal{N}(\log(\zeta), s), \text{accept } \zeta \text{ w.p.} \frac{\pi(\zeta^* \mid \boldsymbol{\Lambda}, \mathbf{y}, \mathbf{X})\pi(\zeta^*)}{\pi(\zeta \mid \boldsymbol{\Lambda}, \mathbf{y}, \mathbf{X})\pi(\zeta)}$$

with $\zeta = \tau^{-2}$ and $s = \sum_{g=1}^{G} \mathbf{1}(\zeta_{\max}^{-1}\lambda_g^2 > \delta)$ where $\zeta_{\max} = \max\{\zeta, \zeta^*\}$ and $\delta = o(\frac{1}{nG})$. Here we chose $\delta = \frac{1}{n^2G}$, according to their recommendation.

4.5.2 Simulation Output

Simulation Results

4.5.3 Interpretation of simulation results

Here are our main observations from tables.

- With an increase in the magnitude of correlation among the covariates within a group, MP and FPR decrease. This indicates that when covariates form a group with a nonsignificant

Table 1: Mean True/False Positive Rate based on 100 replications(Example 1)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0162	0.0125	0.95	0.0123	0.0114	0.98
Normal GH	0.0181	0.0145	0.95	0.0123	0.0114	0.98
GSD-SSS	0.0554	0.0115	0.55	0.0499	0.0111	0.60
Modified GH(EB1)	0.0215	0.0172	0.94	0.0183	0.0148	0.95
Modified GH(EB2)	0.0175	0.0128	0.94	0.0149	0.0122	0.96
Normal GH(EB)	0.0222	0.0180	0.94	0.0206	0.0162	0.94
Modified GH(FBtruncated)	0.0160	0.0122	0.95	0.0118	0.0109	0.98
Modified GH(FBfull)	0.0237	0.0185	0.93	0.0227	0.0174	0.93
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0174	0.0127	0.94	0.0147	0.0119	0.96
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0186	0.0129	0.93	0.0156	0.0118	0.95
Normal GH(FB)	0.0276	0.0164	0.94	0.0178	0.0132	0.94
BGL-SS	0.0799	0.0222	0.40	0.0994	0.0666	0.60
Group LASSO	0.2321	0.2501	0.93	0.2312	0.2501	0.94
Group SCAD	0.2790	0.2176	0.17	0.1658	0.1143	0.37
Group MCP	0.1310	0.0722	0.35	0.0651	0.0211	0.54
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0109	0.0122	1.00	0.0094	0.0105	1.00
Normal GH	0.0129	0.0144	1.00	0.0115	0.0128	1.00
GSD-SSS	0.0103	0.0114	1.00	0.0094	0.0104	1.00
Modified GH(EB1)	0.0112	0.0124	1.00	0.0107	0.0119	1.00
Modified GH(EB2)	0.0112	0.0124	1.00	0.0107	0.0119	1.00
Normal GH(EB)	0.0131	0.0145	1.00	0.0120	0.0133	1.00
Modified GH(FBtruncated)	0.0106	0.0118	1.00	0.0092	0.0102	1.00
Modified GH(FBfull)	0.0139	0.0154	1.00	0.0132	0.0147	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0112	0.0124	1.00	0.0095	0.0106	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0111	0.0123	1.00	0.0097	0.0108	1.00
Normal GH(FB)	0.0127	0.0141	1.00	0.0109	0.0121	1.00
BGL-SS	0.0179	0.0199	1.00	0.0149	0.0165	1.00
Group LASSO	0.1998	0.2221	1.00	0.1858	0.2065	1.00
Group SCAD	0.0651	0.0722	1.00	0.0552	0.0613	1.00
Group MCP	0.0450	0.0499	1.00	0.0308	0.0343	1.00

Table 2: Mean True/False Positive Rate based on 100 replications(Example 2)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0065	0.005	0.98	0.00306	0.0034	1.000
Normal GH	0.0121	0.009	0.96	0.00531	0.0059	1.000
GSD-SSS	0.0235	0.005	0.81	0.00712	0.0038	0.963
Modified GH(EB1)	0.0108	0.006	0.95	0.00306	0.0034	1.000
Modified GH(EB2)	0.0084	0.006	0.97	0.00306	0.0034	1.000
Usual GH(EB)	0.0113	0.008	0.95	0.00531	0.0059	1.000
Modified GH(FBtruncated)	0.0071	0.0045	0.97	0.00279	0.0031	1.000
Modified GH(FBfull)	0.0129	0.011	0.97	0.00468	0.0052	1.000
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0094	0.006	0.96	0.00432	0.0048	1.000
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0084	0.006	0.97	0.00405	0.0045	1.000
Usual GH(FB)	0.0121	0.009	0.96	0.0055	0.0061	1.000
BGL-SS	0.0081	0.006	0.98	0.0049	0.0055	1.000
Group LASSO	0.3037	0.333	0.96	0.1998	0.2222	1.000
Group SCAD	0.0351	0.022	0.85	0.0170	0.0189	1.000
Group MCP	0.0311	0.011	0.80	0.0092	0.0102	1.000
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.00405	0.0045	1.00	0.00252	0.0028	1.00
Usual GH	0.00441	0.0049	1.00	0.00261	0.0029	1.00
GSD-SSS	0.00432	0.0048	1.00	0.00279	0.0031	1.00
Modified GH(EB1)	0.00522	0.0058	1.00	0.00261	0.0029	1.00
Modified GH(EB2)	0.00459	0.0051	1.00	0.00252	0.0028	1.00
Usual GH(EB)	0.00648	0.0072	1.00	0.00297	0.0033	1.00
Modified GH(FBtruncated)	0.00423	0.0047	1.00	0.00234	0.0026	1.00
Modified GH(FBfull)	0.00486	0.0054	1.00	0.00342	0.0038	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.00468	0.0052	1.00	0.00333	0.0037	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.00441	0.0049	1.00	0.00315	0.0035	1.00
Usual GH(FB)	0.00477	0.0053	1.00	0.00279	0.0031	1.00
BGL-SS	0.00499	0.0055	1.00	0.00297	0.0033	1.00
Group LASSO	0.19981	0.2221	1.00	0.19981	0.2221	1.00
Group SCAD	0.00477	0.0053	1.00	0.00383	0.0042	1.00
Group MCP	0.00444	0.0049	1.00	0.00352	0.0039	1.00

amount of correlation among themselves, the right decision within a group for a particular regressor also influences the same for the remaining individuals forming the group.

- Wang and Leng (2008) felt that, the group LASSO also may have the drawback of inconsistency in variable selection, like the usual LASSO. Xu and Ghosh (2015) proved this property in their paper. This is also reflected in our simulation setting, as in all cases, irrespective of the value of ρ, the LASSO group tends to select more variables and produces a higher FPR than the remaining methods.

- These tables also suggest that the Modified Group Horseshoe has slightly lower MP and

Table 3: Mean True/False Positive Rate based on 100 replications(Example 3)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0092	0.0034	0.95	0.0051	0.0029	0.98
Normal GH	0.0118	0.0041	0.95	0.0082	0.0036	0.96
GSD-SSS	0.0265	0.0046	0.82	0.0188	0.0029	0.87
Modified GH(EB1)	0.0120	0.0038	0.93	0.0117	0.0034	0.93
Modified GH(EB2)	0.0118	0.0035	0.92	0.0115	0.0031	0.93
Normal GH(EB)	0.0127	0.0045	0.93	0.0122	0.0039	0.93
Modified GH(FBtruncated)	0.0116	0.0033	0.93	0.0126	0.0030	0.92
Modified GH(FBfull)	0.0123	0.0041	0.93	0.0159	0.0039	0.90
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0096	0.0038	0.95	0.0063	0.0029	0.97
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0081	0.0035	0.96	0.0065	0.0031	0.97
Normal GH(FB)	0.0125	0.0043	0.93	0.0144	0.0036	0.91
BGL-SS	0.0133	0.0038	0.92	0.0104	0.0034	0.94
Group LASSO	0.1675	0.1812	0.94	0.1574	0.1691	0.94
Group SCAD	0.0313	0.0215	0.91	0.0228	0.0175	0.94
Group MCP	0.0188	0.0143	0.95	0.0135	0.0112	0.97
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.00113	0.0013	1.00	0.00105	0.0012	1.00
Normal GH	0.00131	0.0015	1.00	0.00131	0.0015	1.00
GSD-SSS	0.00113	0.0013	1.00	0.00113	0.0013	1.00
Modified GH(EB1)	0.00123	0.0014	1.00	0.00131	0.0015	1.00
Modified GH(EB2)	0.00113	0.0013	1.00	0.00105	0.0012	1.00
Normal GH(EB)	0.00123	0.0014	1.00	0.00140	0.0016	1.00
Modified GH(FBtruncated)	0.00122	0.0014	1.00	0.00088	0.0010	1.00
Modified GH(FBfull)	0.00192	0.0022	1.00	0.00157	0.0018	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.00131	0.0015	1.00	0.00123	0.0014	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.00123	0.0014	1.00	0.00123	0.0014	1.00
Normal GH(FB)	0.00158	0.0018	1.00	0.00114	0.0013	1.00
BGL-SS	0.00625	0.0072	1.00	0.00313	0.0036	1.00
Group LASSO	0.14717	0.1682	1.00	0.13475	0.1540	1.00
Group SCAD	0.05440	0.0622	1.00	0.02761	0.0315	1.00
Group MCP	0.001841	0.0021	1.00	0.00161	0.0018	1.00

FPR compared to the normal one, irrespective of the choice of ρ in all cases. When the signal strength (i.e. coefficient of the active groups) is weak, Modified Group Horseshoe produces much better results than its two-groups counterparts GSD-SSS and BGL-SS and the frequentist methods- group SCAD, group LASSO and group MCP in terms of MP, FPR, and TPR in examples 1-3, where the sample sizes are small ($n = 50$) or moderate ($n = 200$).

- Examples 1 and 2 clarify that when signal strength is strong, regardless of sample size n, the performances of BGL-SS, GSD-SSS, group SCAD and group MCP are comparable

Table 4: Mean True/False Positive Rate based on 100 replications(Example 4)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0029	0.0031	1.00	0.0023	0.0024	1.00
Normal GH	0.0029	0.0031	1.00	0.0023	0.0024	1.00
GSD-SSS	0.0029	0.0031	1.00	0.0023	0.0024	1.00
Modified GH(EB1)	0.0035	0.0036	1.00	0.0029	0.0031	1.00
Modified GH(EB2)	0.0033	0.0034	1.00	0.0027	0.0028	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0033	0.0034	1.00	0.0023	0.0024	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0033	0.0034	1.00	0.0024	0.0025	1.00
Normal GH(EB)	0.0035	0.0036	1.00	0.0029	0.0031	1.00
Modified GH(FBtruncated)	0.0029	0.0031	1.00	0.0023	0.0024	1.00
Modified GH(FBfull)	0.0034	0.0035	1.00	0.0027	0.0028	1.00
Normal GH(FB)	0.0031	0.0032	1.00	0.0024	0.0025	1.00
BGL-SS	0.0032	0.0033	1.00	0.0024	0.0025	1.00
Group LASSO	0.065	0.0681	1.00	0.0522	0.0544	1.00
Group SCAD	0.0028	0.0029	1.00	0.0023	0.0024	1.00
Group MCP	0.0029	0.0031	1.00	0.0027	0.0028	1.00
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0027	0.0028	1.00	0.0023	0.0024	1.00
Normal GH	0.0027	0.0028	1.00	0.0023	0.0024	1.00
GSD-SSS	0.0027	0.0028	1.00	0.0023	0.0024	1.00
Modified GH(EB1)	0.0025	0.0026	1.00	0.0029	0.0031	1.00
Modified GH(EB2)	0.0025	0.0026	1.00	0.0029	0.0031	1.00
Normal GH(EB)	0.0025	0.0026	1.00	0.0029	0.0031	1.00
Modified GH(FBtruncated)	0.0025	0.0026	1.00	0.0023	0.0024	1.00
Modified GH(FBfull)	0.0032	0.0033	1.00	0.0031	0.0032	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0029	0.0031	1.00	0.0023	0.0024	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0028	0.0029	1.00	0.0023	0.0024	1.00
Normal GH(FB)	0.0031	0.0032	1.00	0.0028	0.0029	1.00
BGL-SS	0.0024	0.0025	1.00	0.0020	0.0021	1.00
Group LASSO	0.0570	0.0594	1.00	0.0522	0.0544	1.00
Group SCAD	0.0028	0.0029	1.00	0.0022	0.0023	1.00
Group MCP	0.0026	0.0027	1.00	0.0022	0.0023	1.00

with those of our decision rule.

- Example 3 is different from all the remaining examples, as the group size is different in this case. Although the group size is not used in the prior distribution of the group regression coefficients in any of these methods, our half-thresholding rule successfully captures the truth and hence produces better results than those of GSD-SSS, BGL-SS, group SCAD and group MCP.

- Examples 3-6 clarify that when signal strength is strong, regardless of sample size n, the performances of BGL-SS and GSD-SSS are comparable with those of our decision rule.

Table 5: Mean True/False Positive Rate based on 100 replications(Example 5)

	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.0157	0.0119	0.95	0.0123	0.0114	0.98
Normal GH	0.0174	0.0114	0.95	0.0123	0.0114	0.98
GSD-SSS	0.0523	0.0114	0.58	0.0499	0.0111	0.60
Modified GH(EB1)	0.0218	0.0164	0.93	0.0183	0.0148	0.95
Modified GH(EB2)	0.0183	0.0125	0.93	0.0177	0.0119	0.93
Normal GH(EB)	0.0230	0.0178	0.93	0.0206	0.0162	0.94
Modified GH(FBtruncated)	0.0169	0.0121	0.94	0.0150	0.0111	0.95
Modified GH(FBfull)	0.0229	0.0133	0.89	0.0226	0.0162	0.92
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0162	0.0125	0.95	0.0146	0.0118	0.96
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0167	0.0119	0.94	0.0154	0.0116	0.95
Normal GH(FB)	0.0197	0.0141	0.93	0.0178	0.0132	0.94
BGL-SS	0.0751	0.0188	0.42	0.0601	0.0222	0.61
Group LASSO	0.3064	0.3331	0.94	0.2311	0.2501	0.94
Group SCAD	0.2574	0.1982	0.21	0.1553	0.1081	0.42
Group MCP	0.0876	0.0318	0.41	0.0586	0.0195	0.59

Table 6: Mean True/False Positive Rate based on 100 replications(Example 6)

	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH	0.00351	0.0039	1.00	0.00252	0.0028	1.00
Normal GH	0.00369	0.0041	1.00	0.00261	0.0029	1.00
GSD-SSS	0.00369	0.0041	1.00	0.00279	0.0031	1.00
Modified GH(EB1)	0.00396	0.0044	1.00	0.00261	0.0029	1.00
Modified GH(EB2)	0.00369	0.0041	1.00	0.00261	0.0029	1.00
Normal GH(EB)	0.00405	0.0045	1.00	0.00297	0.0033	1.00
Modified GH(FBtruncated)	0.00396	0.0044	1.00	0.00279	0.0031	1.00
Modified GH(FBfull)	0.00468	0.0052	1.00	0.00342	0.0038	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.00387	0.0043	1.00	0.00288	0.0032	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.00378	0.0042	1.00	0.00279	0.0031	1.00
Normal GH(FB)	0.00468	0.0052	1.00	0.00279	0.0031	1.00
BGL-SS	0.00401	0.0044	1.00	0.00203	0.0022	1.00
Group LASSO	0.19981	0.2221	1.00	0.19981	0.2221	1.00
Group SCAD	0.00387	0.0043	1.00	0.00351	0.0039	1.00
Group MCP	0.00351	0.0039	1.00	0.00243	0.0027	1.00

- Example 4 shows that when the signal strength is weak, other competing procedures can produce similar results to that obtained by ours, if the sample size is large ($n = 500$).

- In example 5, the design is taken to be block-diagonal. Since the data generation scheme is similar to that of Example 1, we expect that our method would provide better results than that of Yang and Narisetty (2020) and Xu and Ghosh (2015) when the signal strength is weak and would produce comparable results when the signal strength is strong. Table 5 confirms this.

- In example 6, we consider a case where the design matrix is block-diagonal. Since the data

Table 7: Mean True/False Positive Rate based on 100 replications(Example 7)

Small group coefficients			
Prior	MP	FPR	TPR
Modified GH	0.0115	0.0113	0.986
Normal GH	0.0118	0.0113	0.978
GSD-SSS	0.0427	0.0349	0.769
Modified GH(EB1)	0.0137	0.0133	0.975
Modified GH(EB2)	0.0136	0.0134	0.981
Normal GH(EB)	0.0165	0.0157	0.964
Modified GH(FBtruncated)	0.0118	0.0117	0.986
Modified GH(FBfull)	0.0137	0.0134	0.979
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0121	0.0119	0.982
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0122	0.0121	0.985
Normal GH(FB)	0.0118	0.0113	0.977
BGL-SS	0.0121	0.0104	0.754
Group LASSO	0.1780	0.1831	0.950
Group SCAD	0.1011	0.1041	0.901
Group MCP	0.0882	0.0887	0.931
Large group coefficients			
Prior	MP	FPR	TPR
Modified GH	0.0108	0.0113	1.00
Normal GH	0.0142	0.0148	1.00
GSD-SSS	0.0109	0.0114	1.00
Modified GH(EB1)	0.0127	0.0132	1.00
Modified GH(EB2)	0.0124	0.0129	1.00
Normal GH(EB)	0.0143	0.0149	1.00
Modified GH(FBtruncated)	0.0108	0.0113	1.00
Modified GH(FBfull)	0.0118	0.0123	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0108	0.0112	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0106	0.0110	1.00
Normal GH(FB)	0.0117	0.0122	1.00
BGL-SS	0.0112	0.0116	1.00
Group LASSO	0.1201	0.1252	1.00
Group SCAD	0.0361	0.0375	1.00
Group MCP	0.0121	0.0125	1.00

generation scheme is similar to that of Example 5, from the previous result, we expected that our method would provide better results than those of Yang and Narisetty (2020) and Xu and Ghosh (2015) when the signal strength is weak and would produce comparable results when the signal strength is strong. Table 6 confirms this.

- Examples 7-9 show that the rule (4.9) will also work even if $p > n$. In this case, too, MP and TPR corresponding to our decision rule are much better than those of GSD-SSS and better than BGL-SS for weak signal strength and produce comparable results for strong signal strength.

- Example 9 deals with situations when there is a mixture of weak and strong signal strengths. In these cases also, our proposed method outperforms GSD-SSS, BGL-SS and

Table 8: Mean True/False Positive Rate based on 100 replications(Example 8)

Small group coefficients			
Prior	MP	FPR	TPR
Modified GH	0.0137	0.0140	0.988
Normal GH	0.0142	0.0139	0.984
GSD-SSS	0.0445	0.0138	0.782
Modified GH(EB1)	0.0149	0.0141	0.980
Modified GH(EB2)	0.0144	0.0141	0.984
Normal GH(EB)	0.0158	0.0141	0.974
Modified GH(FBtruncated)	0.0135	0.0141	0.990
Modified GH(FBfull)	0.0138	0.0145	0.990
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0146	0.0143	0.984
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0145	0.0142	0.986
Normal GH(FB)	0.0137	0.0140	0.988
BGL-SS	0.0134	0.0144	0.992
Group LASSO	0.1866	0.2151	0.975
Group SCAD	0.0151	0.0176	1.000
Group MCP	0.0151	0.0176	1.000
Large group coefficients			
Prior	MP	FPR	TPR
Modified GH	0.0111	0.0131	1.00
Normal GH	0.0127	0.0149	1.00
GSD-SSS	0.0109	0.0129	1.00
Modified GH(EB1)	0.0121	0.0142	1.00
Modified GH(EB2)	0.0115	0.0135	1.00
Normal GH(EB)	0.0129	0.0152	1.00
Modified GH(FBtruncated)	0.0123	0.0145	1.00
Modified GH(FBfull)	0.0126	0.0148	1.00
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0112	0.0132	1.00
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0113	0.0133	1.00
Normal GH(FB)	0.0118	0.0139	1.00
BGL-SS	0.0121	0.0142	1.00
Group LASSO	0.1133	0.1331	1.00
Group SCAD	0.0301	0.0353	1.00
Group MCP	0.0112	0.0133	1.00

yields better results than that of group LASSO, group MCP and group SCAD.

- In general, it would not be an exaggeration to state that our decision rules provide excellent results in terms of group selection even if the design matrix is not block-diagonal. This is true when τ is treated as a tuning parameter and when it is treated as empirical or full Bayesian ways.

- The results in Tables 10 and 11 show that with an increase in the value of δ, the misclassification probability (MP) and the False Positive Rate (FPR) increase and the True Positive Rate (TPR) decreases, irrespective of the signal strength and magnitude of correlation among the covariates within a group. However, this change is more significant when the signal strength is small. Hence, based on simulation studies, we propose to choose δ to be

Table 9: Mean True/False Positive Rate based on 100 replications(Example 9)

Number of active groups=4			
Prior	MP	FPR	TPR
Modified GH	0.0175	0.0204	0.998
Normal GH	0.0169	0.0194	0.996
GSD-SSS	0.0326	0.0183	0.892
Modified GH(EB1)	0.0205	0.0232	0.994
Modified GH(EB2)	0.0164	0.0191	0.995
Normal GH(EB)	0.0203	0.0231	0.994
Modified GH(FBtruncated)	0.0155	0.0177	0.996
Modified GH(FBfull)	0.0167	0.0191	0.996
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0188	0.0212	0.994
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0180	0.0209	0.997
Normal GH(FB)	0.0139	0.0181	0.995
BGL-SS	0.0179	0.0199	0.992
Group LASSO	0.1506	0.1764	0.985
Group SCAD	0.0199	0.0214	0.988
Group MCP	0.0288	0.0272	0.963
Number of active groups=8			
Prior	MP	FPR	TPR
Modified GH	0.0102	0.0121	0.994
Normal GH	0.0109	0.0119	0.991
GSD-SSS	0.0558	0.0118	0.851
Modified GH(EB1)	0.0118	0.0121	0.989
Modified GH(EB2)	0.0116	0.0119	0.989
Normal GH(EB)	0.0131	0.0122	0.985
Modified GH(FBtruncated)	0.0109	0.0113	0.990
Modified GH(FBfull)	0.0128	0.0122	0.986
Modified TPBN($\alpha_1 = 0.5, \alpha_2 = 1$)	0.0113	0.0124	0.991
Modified TPBN($\alpha_1 = 1, \alpha_2 = 0.5$)	0.0101	0.0120	0.994
Normal GH(FB)	0.0118	0.0117	0.988
BGL-SS	0.0127	0.0121	0.986
Group LASSO	0.1406	0.1978	0.981
Group SCAD	0.0200	0.0177	0.975
Group MCP	0.0148	0.0159	0.987

small and positive, that is, we propose the choice $\delta = 0.1$.

- Tables 12-15 show that, keeping $c_1 = 2$ and varying c_2 over $\{1, \ldots, 5\}$ results in an increase in MP and FPR and a decrease in TPR for both small and large signals. The same result is also obtained with $c_2 = 1$ and $c_1 \in \{2, 3, 4, 5\}$. Based on these results, we recommend to choose $(c_1, c_2) = (2, 1)$.

- Tables 16 and 17 demonstrate that when the proportion of active groups is known, MP and FPR are at their least when τ_n is used as a tuning parameter based on this proportion. This implies that when the knowledge on the proportion is available, it is better to use that knowledge. When the level of sparsity is unknown, the tables demonstrate the following. One gets better results in these cases when one either uses the empirical Bayes method

Table 10: Mean True/False Positive Rate using different choices of δ based on 100 replications(Example 1)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH($\delta = 0.1$)	0.0162	0.0125	0.95	0.0123	0.0114	0.98
Modified GH($\delta = 0.2$)	0.0163	0.0126	0.95	0.0126	0.0118	0.98
Modified GH($\delta = 0.3$)	0.0164	0.0127	0.95	0.0138	0.0120	0.97
Modified GH($\delta = 0.4$)	0.0174	0.0126	0.94	0.0139	0.0121	0.97
Modified GH($\delta = 0.5$)	0.0174	0.0127	0.94	0.0141	0.0123	0.97
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH($\delta = 0.1$)	0.0109	0.0122	1.00	0.0094	0.0105	1.00
Modified GH($\delta = 0.2$)	0.0111	0.0123	1.00	0.0095	0.0106	1.00
Modified GH($\delta = 0.3$)	0.0112	0.0125	1.00	0.0095	0.0106	1.00
Modified GH($\delta = 0.4$)	0.0114	0.0127	1.00	0.0096	0.0107	1.00
Modified GH($\delta = 0.5$)	0.0114	0.0127	1.00	0.0097	0.0108	1.00

Table 11: Mean True/False Positive Rate using different choices of δ based on 100 replications(Example 2)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH ($\delta = 0.1$)	0.00659	0.0051	0.98	0.00306	0.0034	1.000
Modified GH($\delta = 0.2$)	0.00686	0.0054	0.98	0.00324	0.0036	1.000
Modified GH($\delta = 0.3$)	0.00731	0.0059	0.98	0.00333	0.0037	1.000
Modified GH($\delta = 0.4$)	0.00749	0.0061	0.98	0.00342	0.0038	1.000
Modified GH($\delta = 0.5$)	0.00758	0.0062	0.98	0.00351	0.0039	1.000
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH($\delta = 0.1$)	0.00405	0.0045	1.00	0.00261	0.0029	1.00
Modified GH($\delta = 0.2$)	0.00405	0.0045	1.00	0.00261	0.0029	1.00
Modified GH($\delta = 0.3$)	0.00414	0.0046	1.00	0.00261	0.0029	1.00
Modified GH($\delta = 0.4$)	0.00414	0.0046	1.00	0.00261	0.0029	1.00
Modified GH($\delta = 0.5$)	0.00423	0.0047	1.00	0.00270	0.0030	1.00

(that estimates τ using quantities which are influenced by the level of sparsity) or the full Bayes method using a prior on τ. The ad-hoc choices of τ_n using only G_n are clear under-performers in such cases.

Table 12: Mean True/False Positive Rate using $c_2 = 1$ based on 100 replications(Example 1)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_1 = 2$)	0.0172	0.0128	0.94	0.0149	0.0122	0.96
Modified GH(EB2,$c_1 = 3$)	0.0178	0.0131	0.94	0.0152	0.0125	0.96
Modified GH(EB2,$c_1 = 4$)	0.0260	0.0167	0.89	0.0258	0.0165	0.89
Modified GH(EB2,$c_1 = 5$)	0.0295	0.0172	0.86	0.0274	0.0171	0.88
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_1 = 2$)	0.0112	0.0124	1.00	0.0107	0.0119	1.00
Modified GH(EB2,$c_1 = 3$)	0.0114	0.0127	1.00	0.0110	0.0122	1.00
Modified GH(EB2,$c_1 = 4$)	0.0116	0.0129	1.00	0.0113	0.0126	1.00
Modified GH(EB2,$c_1 = 5$)	0.0119	0.0132	1.00	0.0115	0.0128	1.00

Table 13: Mean True/False Positive Rate using $c_2 = 1$ based on 100 replications(Example 2)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_1 = 2$)	0.0084	0.006	0.97	0.00306	0.0034	1.00
Modified GH(EB2,$c_1 = 3$)	0.0123	0.007	0.94	0.00315	0.0035	1.00
Modified GH(EB2,$c_1 = 4$)	0.0160	0.010	0.93	0.00342	0.0038	1.00
Modified GH(EB2,$c_1 = 5$)	0.0187	0.013	0.93	0.00369	0.0041	1.00
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_1 = 2$)	0.00459	0.0051	1.00	0.00252	0.0028	1.00
Modified GH(EB2,$c_1 = 3$)	0.00468	0.0052	1.00	0.00261	0.0029	1.00
Modified GH(EB2,$c_1 = 4$)	0.00504	0.0056	1.00	0.00288	0.0032	1.00
Modified GH(EB2,$c_1 = 5$)	0.00522	0.0058	1.00	0.00306	0.0034	1.00

4.6 Real data analysis

In this section, we compare the performance of our variable selection and estimation rules with some existing methods when applied to a real dataset. Here we consider two datasets, both of them are available in R.

Diabetes dataset. This dataset was used initially by Efron (2004), which is available in R package *care*. This dataset was previously studied by Tang et al. (2018). It contains ten baseline variables (predictors): age, sex, body mass index (BMI), average blood pressure (BP), six blood serum measurements (TC, LDL, HDL, TCH, LTH, GLU), and a quantitative measure of disease progression one year after baseline (response) for 442 diabetes patients. The baseline variables are standardized to have zero mean and unit l_2 norm. The response variable is centered to have zero mean. The different methods discussed in previous simulations were applied and the results are presented in Table 18. It was observed that five variables (Gender, BMI, BP,

Table 14: Mean True/False Positive Rate using $c_1 = 2$ based on 100 replications(Example 1)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_2 = 1$)	0.0172	0.0128	0.94	0.0149	0.0122	0.96
Modified GH(EB2,$c_2 = 2$)	0.0176	0.0129	0.94	0.0161	0.0123	0.95
Modified GH(EB2,$c_2 = 3$)	0.0188	0.0131	0.93	0.0163	0.0126	0.95
Modified GH(EB2,$c_2 = 4$)	0.0189	0.0132	0.93	0.0175	0.0128	0.94
Modified GH(EB2,$c_2 = 5$)	0.0201	0.0134	0.92	0.0178	0.0131	0.94
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_2 = 1$)	0.0112	0.0124	1.00	0.0107	0.0119	1.00
Modified GH(EB2,$c_2 = 2$)	0.0113	0.0126	1.00	0.0110	0.0122	1.00
Modified GH(EB2,$c_2 = 3$)	0.0115	0.0128	1.00	0.0113	0.0125	1.00
Modified GH(EB2,$c_2 = 4$)	0.0118	0.0131	1.00	0.0116	0.0129	1.00
Modified GH(EB2,$c_2 = 5$)	0.0119	0.0132	1.00	0.0117	0.0130	1.00

Table 15: Mean True/False Positive Rate using $c_1 = 2$ based on 100 replications(Example 2)

Small group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_2 = 1$)	0.0084	0.006	0.97	0.00306	0.0034	1.00
Modified GH(EB2,$c_2 = 2$)	0.0103	0.007	0.96	0.00315	0.0035	1.00
Modified GH(EB2,$c_2 = 3$)	0.0121	0.009	0.96	0.00324	0.0036	1.00
Modified GH(EB2,$c_2 = 4$)	0.0140	0.010	0.95	0.00342	0.0038	1.00
Modified GH(EB2,$c_2 = 5$)	0.0140	0.010	0.95	0.00351	0.0039	1.00
Large group coefficients						
	$\rho = 0$			$\rho = 0.5$		
Prior	MP	FPR	TPR	MP	FPR	TPR
Modified GH(EB2,$c_2 = 1$)	0.00459	0.0051	1.00	0.00252	0.0028	1.00
Modified GH(EB2,$c_2 = 2$)	0.00468	0.0052	1.00	0.00261	0.0029	1.00
Modified GH(EB2,$c_2 = 3$)	0.00486	0.0054	1.00	0.00279	0.0031	1.00
Modified GH(EB2,$c_2 = 4$)	0.00495	0.0055	1.00	0.00306	0.0034	1.00
Modified GH(EB2,$c_2 = 5$)	0.00495	0.0055	1.00	0.00306	0.0034	1.00

HDL, LTH) are selected by all the methods. These five variables are the first ones that enter the regression equation in Efron (2004). Among all the methods, BGL-SS selects the highest number of variables. The methods based on frequentist approaches also select LDL, but not HDL, On the other hand, all Bayesian approaches select HDL, in addition to five common variables, but not choose TCH.

Next, we are interested in the prediction problem. For this, the 442 observations in the dataset are divided into a training set and a test set. The training set has 280 observations and the test set has 162 observations, as same as Tang et al. (2018). The methods used in the diabetes example are applied for the training data and the mean squared prediction error (MSPE) is estimated based on the test data for each method. The results are presented in Table

Table 16: Mean True/False Positive Rate using different choices of τ based on 100 replications(Example 8)

Small group coefficients			
Prior	MP	FPR	TPR
Modified GH(τ_{1n})	0.0137	0.0140	0.988
Modified GH(τ_{2n})	0.0151	0.0148	0.983
Modified GH(τ_{3n})	0.0149	0.0144	0.982
Modified GH(τ_{4n})	0.0151	0.0146	0.982
Modified GH(τ_{5n})	0.0144	0.0141	0.984
Modified GH(FBtruncated)	0.0127	0.0135	0.992
Large group coefficients			
Prior	MP	FPR	TPR
Modified GH(τ_{1n})	0.0111	0.0131	1.00
Modified GH(τ_{2n})	0.0126	0.0148	1.00
Modified GH(τ_{3n})	0.0123	0.0145	1.00
Modified GH(τ_{4n})	0.0126	0.0148	1.00
Modified GH(τ_{5n})	0.0115	0.0135	1.00
Modified GH(FBtruncated)	0.0112	0.0132	1.00

Table 17: Mean True/False Positive Rate using different choices of τ based on 100 replications(Example 9)

Number of active groups=4			
Prior	MP	FPR	TPR
Modified GH(τ_{1n})	0.0175	0.0204	0.998
Modified GH(τ_{2n})	0.0186	0.0214	0.996
Modified GH(τ_{3n})	0.0185	0.0211	0.998
Modified GH(τ_{4n})	0.0184	0.0209	0.995
Modified GH(τ_{5n})	0.0164	0.0191	0.995
Modified GH(FBtruncated)	0.0154	0.0176	0.996
Number of active groups=8			
Prior	MP	FPR	TPR
Modified GH (τ_{1n})	0.0102	0.0121	0.994
Modified GH(τ_{2n})	0.0133	0.0139	0.988
Modified GH(τ_{3n})	0.0132	0.0138	0.988
Modified GH(τ_{4n})	0.0127	0.0136	0.989
Modified GH(τ_{5n})	0.0116	0.0119	0.989
Modified GH(FBtruncated)	0.0106	0.0109	0.990

19. It ensures that in terms of estimated MSPE, our method yields results comparable to those of the existing methods in this literature.

Birth weight data. We consider the birth weight dataset from Hosmer and Lemeshow (1989) with the group methods, which is available in the R package *grpreg*. This dataset was previously analyzed by Yuan and Lin (2006). The birth weight dataset records the birth weights of 189 babies and 16 predictors concerning the mother. These 16 covariates are divided into 8 groups named as mother's age in years, mother's weight in pounds at the last menstrual period, mother's race, smoking status during pregnancy, number of previous premature labours, history

Table 18: Performance of different methods in Diabetes dataset

Method	Age	Gender	Bmi	Bp	TC	LDL	HDL	TCH	LTH	GLU
Group SCAD	0	1	1	1	1	1	0	1	1	0
Group MCP	0	1	1	1	1	1	0	1	1	0
Group LASSO	0	1	1	1	1	1	0	1	1	0
BGL-SS	0	1	1	1	1	0	1	1	1	1
GSD-SS	0	1	1	1	1	0	1	0	1	0
Modified GH(EB)	0	1	1	1	1	0	1	0	1	0
Modified GH(FB)	0	1	1	1	1	0	1	0	1	0

Table 19: Mean squared prediction error corresponding to different methods for Diabetes dataset

Method	MSPE
Group SCAD	0.484159
Group MCP	0.486948
Group LASSO	0.486289
BGL-SS	0.489933
GSD-SS	0.484794
Modified GH(EB)	0.484990
Modified GH(FB)	0.484291

of hypertension presence of uterine irritability, and number of physician visits during the first trimester. The data were collected at Baystate Medical Center, Springfield, Massachusetts, during 1986. For the prediction problem, the 189 observations in the dataset are divided into a training and a test part. The training part has 126 observations and the test part has 63 observations. The methods used in the diabetes example are applied to the training data and the mean squared prediction error (MSPE) is calculated based on the test data for each method. The results are presented in Table 20. The results indicate that in terms of estimated MSPE, our method yields significantly better result than those of the existing ones. We have observed that, for moderate birth-weights, our method has much better performance than the remaining ones. For the remaining cases, it produces comparable results.

Table 20: Mean squared prediction error corresponding to different methods for Birh weight dataset

Method	MSPE($\times 10^{-2}$)
Group SCAD	9.668395
Group MCP	9.607762
Group LASSO	9.565219
BGL-SS	9.564448
GSD-SS	9.208861
Modified GH(EB)	8.564448
Modified GH(FB)	8.516276

4.7 Concluding remarks and scope for future work

In this chapter, we studied the problem of selecting relevant groups of predictors/regressors in a sparse high-dimensional regression model, assuming that the potential regressors are inherently grouped. For the selection of groups, we considered a half-thresholding rule based on a broad class of global-local shrinkage priors with polynomial tails. We also studied an estimator of group regression coefficients based on this rule.

We extended the idea of one-group global-local shrinkage priors to group regression problems to facilitate group selection by factoring in the similarity and dependence within a group and heterogeneity across groups through the prior distribution of the group coefficients. Based on such priors, we have proposed a half-thresholding (HT) rule that enjoys oracle property irrespective of whether the level of sparsity is known or unknown. Our results also establish that the corresponding HT estimator is more accurate than the usual posterior mean under squared error loss for most interesting sparse situations. In our simulation studies, we have compared different versions of our proposed decision rule (corresponding to the treatment of τ as a tuning parameter or in empirical/full Bayesian ways) to some well-known group selection methods in the literature. We have demonstrated that all our proposed decision rules outperform these methods when the number of observations is small or moderate. Even when the number of observations is large our method has a distinct advantage over that of Yang and Narisetty (2020) when the signal strength is low. When the number of observations is large and the signal strength is high, the methods due to Yang and Narisetty (2020) and Xu and Ghosh (2015) are comparable to ours. Finally, when our methods are applied to real data, they produce results similar to their two-groups counterparts in terms of mean squared prediction error. Therefore, our proposed HT rule can be a viable alternative to the methods available in the literature based on spike and slab priors when dealing with sparse situations.

Throughout this work, our focus is only on selecting and estimating relevant groups as a whole. However in many situations, the selection of both the relevant groups and individual variables within a chosen group might be desirable. One such example is presented by Huang et al. (2012). In genome-wide association studies, genetic variations in the same gene form a natural group. However, a genetic variation related to the disease does not necessarily mean that all other variations in the same gene are also associated with the disease. Hence, in this situation, the selection of important genes both at the group and individual levels becomes important. Development and study of optimality of bi-level selection rules is a problem of keen interest for the future. The literature on this topic is at best at a nascent stage as far as we know.

Another less visited area in this literature is the situation where the possible groups overlap. As mentioned by Huang et al. (2012), in genomic data analysis involving both genes and pathways, many important genes belong to more than one pathway. Hence, in this context, it is challenging to select important variables without all the groups that contain them. Some of the works in this direction are due to Jacob et al. (2009), Liu and Ye (2010), Zhao et al. (2009), etc. Selection of variables in an efficient way in such cases requires more attention and may be a problem of serious research in future.

4.8 Proofs

4.8.1 Proofs of Lemmas, Propositions and Theorems

4.8.2 Some important Lemmas

We first state and prove Lemmas 4.8.1 to 4.8.4 which are crucial to proving the main theorems of our work.

Lemma 4.8.1. *Let L be a nonnegative, measurable, slowly varying function defined over an interval unbounded to the right. Then the following results hold.*

(1) L^α *is slowly varying for all* $\alpha \in \mathbb{R}$.

(2) $\frac{\log L(x)}{\log x} \to 0$ *as* $x \to \infty$.

(3) For every $\alpha > 0$, $x^{-\alpha}L(x) \to 0$ *and* $x^\alpha L(x) \to \infty$ *as* $x \to \infty$.

(4) For $\alpha < -1$, $-\frac{\int_x^\infty t^\alpha L(t)dt}{x^{\alpha+1}L(x)} \to \frac{1}{\alpha+1}$ *as* $x \to \infty$.

(5) There exists a global constant $A_0 > 0$ *such that, for any* $\alpha > -1$, $\frac{\int_{A_0}^x t^\alpha L(t)dt}{x^{\alpha+1}L(x)} \to \frac{1}{\alpha+1}$ *as* $x \to \infty$.

Proof. See Bingham et al. (1987). □

Lemma 4.8.2. *Let* $L : (0,\infty) \to (0,\infty)$ *be a measurable and integrable function such that for fixed* $a > 0$, $\int_0^\infty t^{-a-1}L(t)dt = K^{-1}$, *with* $K \in (0,\infty)$. *Assume* $\tau_n \to 0$ *as* $n \to \infty$. *Then*

$$\int_0^1 u^{a+\frac{m_g}{2}-1}(1-u)^{-a-1}L\left(\frac{1}{\tau_n^2}(\frac{1}{u}-1)\right)du = K^{-1}(\tau_n^2)^{-a}(1+o(1))\,,$$

where the $o(1)$ *term is such that* $\lim_{n\to\infty} o(1) = 0$.

Proof. The proof follows using the same set of arguments used to establish Lemma 5 of Ghosh et al. (2016). □

Lemma 4.8.3. *Consider the hierarchical framework of (4.4) where the local shrinkage parameters are modeled with the class of priors given by (4.5). Suppose* $\tau_n \to 0$ *as* $n \to \infty$. *Then for any given* $a \in (0,1)$, *there exists* $A_0 \geq 1$ *such that*

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq \frac{A_0 K}{a(1-a)}(\tau_n^2)^a L\left(\frac{1}{\tau_n^2}\right)\exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)(1+o(1)).$$

Assume that the slowly varying function $L(\cdot)$ satisfies Assumption 1 for some $a \geq 1$. Then

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq \frac{KM}{a}\tau_n \exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)(1+o(1)).$$

The terms $o(1)$ in both inequalities above tend to zero as $n \to \infty$.

Proof. The proof for the case $a \in (0,1)$ follows using the same set of arguments employed by Ghosh et al. (2016) to establish Theorem 4 of their paper.

Let us now consider the case $a \geq 1$. First note that

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) = \frac{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)\exp\left\{(1-\kappa_g)\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right\}d\kappa_g}{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)\exp\left\{(1-\kappa_g)\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right\}d\kappa_g}. \tag{4.28}$$

Using the transformation $s = \frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)$ in the integrals above, we obtain

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) = \tau_n^2 \frac{\int_0^\infty (1+s\tau_n^2)^{-\frac{m_g}{2}-1} s^{-a} L(s)\exp\left(\frac{s\tau_n^2}{1+s\tau^2}\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)ds}{\int_0^\infty (1+s\tau_n^2)^{-\frac{m_g}{2}} s^{-a-1} L(s)\exp\left(\frac{s\tau_n^2}{1+s\tau^2}\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)ds}. \tag{4.29}$$

Note that

$$\begin{aligned}\int_0^\infty (1+s\tau_n^2)^{-\frac{m_g}{2}} s^{-a-1} L(s)\exp\left(\frac{s\tau_n^2}{1+s\tau_n^2}\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)ds &\geq \int_0^\infty (1+s\tau_n^2)^{-\frac{m_g}{2}} s^{-a-1} L(s) ds \\ &= K^{-1}(1+o(1)),\end{aligned} \tag{4.30}$$

where the last equality above follows from the Dominated Convergence Theorem. Combining (4.29) and (4.30), we obtain

$$\begin{aligned}&E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq K\tau_n^2 \int_0^\infty (1+s\tau_n^2)^{-\frac{m_g}{2}-1} s^{-a} L(s)\exp\left(\frac{s\tau_n^2}{1+s\tau_n^2}\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)ds(1+o(1)) \\ &= K\tau_n^2\left(\int_0^1 + \int_1^{\frac{1}{\tau_n}} + \int_{\frac{1}{\tau_n}}^\infty\right)(1+s\tau_n^2)^{-\frac{m_g}{2}-1} s^{-a} L(s)\exp\left(\frac{s\tau_n^2}{1+s\tau_n^2}\cdot\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)ds(1+o(1)) \\ &= K(A_{1,\tau_n} + A_{2,\tau_n} + A_{3,\tau_n})(1+o(1)), \text{ say.}\end{aligned} \tag{4.31}$$

Observe that for $s \in (0,1)$ and $\tau_n \in (0,1)$, $\frac{s\tau_n^2}{1+s\tau_n^2} \leq \frac{1}{2}$. Also, $\int_0^\infty s^{-a-1}L(s)dt = K^{-1}$.

Therefore it follows that

$$A_{1,\tau_n} \leq K^{-1}\tau_n^2 \exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{4\sigma^2}\right). \tag{4.32}$$

Likewise, for $s \in [1, \frac{1}{\tau_n})$ and $\tau_n \in (0,1)$, using the above arguments, we obtain

$$A_{2,\tau_n} \leq K^{-1}\tau_n \exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{4\sigma^2}\right). \tag{4.33}$$

Finally, using (4) of Lemma 4.8.1, we have

$$A_{3,\tau_n} \leq \exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)\int_{\frac{1}{\tau}}^{\infty} s^{-a-1}L(s)ds \leq \frac{\tau_n}{a}M\exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)(1+o(1)). \tag{4.34}$$

Combining (4.31)-(4.34), the desired result follows. □

Lemma 4.8.4. *Consider the framework of Lemma 4.8.3. Then under Assumption 1, for any arbitrary constants $\eta \in (0,1), q \in (0,1)$ and any fixed $\tau > 0$,*

$$P(\kappa_g > \eta|\tau,\sigma^2,\mathcal{D}) \leq \frac{(a+\frac{m_g}{2})(1-\eta q)^a}{\tau^{2a}(\eta q)^{a+\frac{m_g}{2}}c_0}\exp\left(-\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g\eta(1-q)}{2\sigma^2}\right).$$

Proof. The proof follows using a similar set of arguments used by Ghosh et al. (2016) to establish Theorem 5 in their paper. □

4.8.3 Proofs of Propositions

Proof of Proposition 4.2.2: First, we prove the if part. Towards that, first, we consider the case when $a \in (0,1)$. Using Lemma 4.8.3, we obtain

$$E(1-\kappa_g \mid \tau_n,\sigma^2,\mathcal{D}) \leq \frac{A_0 K}{a(1-a)}(\tau_n^2)^a L(\frac{1}{\tau_n^2})\exp\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g}\widehat{\boldsymbol{\beta}}_g}{2\sigma^2}\right)(1+o(1)). \tag{4.35}$$

When $\tau_n \to 0$ as $n \to \infty$, using Part (3) of Lemma 4.8.1,

$$\lim_{n\to\infty}(\tau_n^2)^a L(\frac{1}{\tau_n^2}) = \lim_{n\to\infty}(\frac{1}{\tau_n^2})^{-a}L(\frac{1}{\tau_n^2}) = 0. \tag{4.36}$$

Under a block-orthogonal design, from the standard theory of linear regression, the distribution of the ordinary least square estimator $\widehat{\boldsymbol{\beta}}_g$ is given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0\right) \sim \mathcal{N}_{m_g}(\mathbf{0}, \sigma^2\mathbf{Q}_{n,g}^{-1}).$$

Clearly, if $\boldsymbol{\beta}_g^0 = \mathbf{0}$, $\sqrt{n}\widehat{\boldsymbol{\beta}}_g \sim \mathcal{N}_{m_g}(\mathbf{0}, \sigma^2 \mathbf{Q}_{n,g}^{-1})$. Therefore, $\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} \sim \chi^2_{m_g}$, whence

$$\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} = O_p(1), \text{ for all } n. \tag{4.37}$$

Combining (4.35) - (4.37), and using Slutsky's Theorem, it readily follows that

$$E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Next, we consider the case when case $a \geq 1$. Observe that the upper bound to $E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$ is similar to the upper bound when $a \in (0,1)$. Hence, the proof follows using the same set of arguments as in the case $a \in (0,1)$.

Now, we consider the 'Only if' part. Here, we will show that if $\tau_n \nrightarrow 0$ as $n \to \infty$, $P(E(1 - \kappa_g|\tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \nrightarrow 0$ as $n \to \infty$.

Since, $\tau_n \nrightarrow 0$ as $n \to \infty$, implies, there exists at least one subsequence of τ_n, say τ_{n_k}, such that as $k \to \infty$, $\tau_{n_k} > \epsilon$ for some $\epsilon > 0$. Hence, the remaining calculations will be based on that subsequence of τ_n. Note that

$$\begin{aligned} E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) &= \int_0^{\frac{1}{4}} \kappa_g \pi(\kappa_g|\tau_n, \sigma^2, \mathcal{D}) d\kappa_g + \int_{\frac{1}{4}}^1 \kappa_g \pi(\kappa_g|\tau_n, \sigma^2, \mathcal{D}) d\kappa_g \\ &\leq \frac{1}{4} + P(\kappa_g > \frac{1}{4}|\tau_n, \sigma^2, \mathcal{D}). \end{aligned} \tag{4.38}$$

Therefore, we have,

$$P\left(E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) < \frac{1}{2}\right) \geq P\left(P(\kappa_g > \frac{1}{4}|\tau_n, \sigma^2, \mathcal{D}) < \frac{1}{4}\right). \tag{4.39}$$

Now, substituting $\eta = \frac{1}{4}$ in Lemma 4.8.4, some simple algebra yields, for any fixed $\tau > 0$,

$$\begin{aligned} P\left(P(\kappa_g > \frac{1}{4}|\tau_n, \sigma^2, \mathcal{D}) < \frac{1}{4}\right) &\geq P\left(\frac{(a+\frac{1}{2})(1-\frac{1}{4}q)^a}{{\tau_n}^{2a}(\eta q)^{a+\frac{1}{2}} c_0} \exp\left(-\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g} \widehat{\boldsymbol{\beta}}_g (1-q)}{8\sigma^2}\right) < \frac{1}{4}\right) \\ &= P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > \frac{8}{(1-q)}[2a \log\left(\frac{1}{\tau}\right) + K_2]\right), \end{aligned} \tag{4.40}$$

where K_2 is a constant depending on a, q, c_0. Note that, for sufficiently large k, $\log\left(\frac{1}{\tau_{n_k}}\right) < \log\left(\frac{1}{\epsilon}\right)$, where the ϵ is mentioned previously. This ensures a lower bound on (4.40), as $k \to \infty$

$$\begin{aligned} P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > \frac{8}{(1-q)}[2a \log\left(\frac{1}{\tau_{n_k}}\right) + K_2]\right) &\geq P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > \frac{8}{(1-q)}[2a \log\left(\frac{1}{\epsilon}\right) + K_2]\right) \\ &\nrightarrow 0 \text{ as } k \to \infty. \end{aligned} \tag{4.41}$$

Combining (4.38)-(4.41) completes the proof of the Only if part and the proof of Proposition 4.2.2 is completed.

□

Proof of Proposition 4.2.3: It would be enough to show that $E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \xrightarrow{P} 0$ as $n \to \infty$ when $\boldsymbol{\beta}_g^0 \neq \mathbf{0}$.

Let us fix $\epsilon_0 > 0$. Then

$$\begin{aligned} E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) &= \int_0^{\frac{\epsilon_0}{2}} \kappa_g \pi(\kappa_g|\tau_n, \sigma^2, \mathcal{D}) d\kappa_g + \int_{\frac{\epsilon_0}{2}}^1 \kappa_g \pi(\kappa_g|\tau_n, \sigma^2, \mathcal{D}) d\kappa_g \\ &\leq \frac{\epsilon_0}{2} + P(\kappa_g > \frac{\epsilon_0}{2}|\tau_n, \sigma^2, \mathcal{D}) \,. \end{aligned} \tag{4.42}$$

Therefore, for a given $\epsilon_0 > 0$,

$$P\left(E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \epsilon_0\right) \leq P\left(P(\kappa_g > \frac{\epsilon_0}{2}|\tau_n, \sigma^2, \mathcal{D}) > \frac{\epsilon_0}{2}\right). \tag{4.43}$$

Now, substituting $\eta = \frac{\epsilon_0}{2}$ in Lemma 4.8.4, some simple algebra yields

$$\begin{aligned} P(E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \epsilon_0) &\leq P\left(\frac{(a+\frac{m_g}{2})(1-\eta q)^a}{{\tau_n}^{2a}(\eta q)^{a+\frac{m_g}{2}} c_0} \exp\left(-\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} Q_{n,g} \widehat{\boldsymbol{\beta}}_g \eta(1-q)}{2\sigma^2}\right) > \frac{\epsilon_0}{2}\right) \\ &= P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right), \end{aligned} \tag{4.44}$$

where $d_n = \frac{4}{\epsilon_0(1-q)}\left[d' + a \cdot \log\left(\frac{1}{\tau_n^2}\right)\right]$, d' being a constant is independent of n. Observe now that using (A1), we have

$$\begin{aligned} &P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) \\ &\leq \sum_{j=1}^{m_g} P\left(-\frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} \leq \frac{\sqrt{n}(\widehat{\beta}_{gj} - \beta_{gj}^0)}{\sigma\sqrt{\zeta_j}} \leq \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}}\right), \end{aligned} \tag{4.45}$$

where ζ_j is the j^{th} diagonal element of $\mathbf{Q}_{n,g}^{-1}$. Now the cases $\min_j \beta_{gj}^0 > 0$ and $\min_j \beta_{gj}^0 < 0$ are dealt separately as follows.

Case-(I): Assume $\min_j \beta_{gj}^0 > 0$. Using (4.45), we have

$$\begin{aligned} P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) &\leq \sum_{j=1}^{m_g} \left[1 - \Phi\left(\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}}\right)\right] \\ &\leq \sum_{j=1}^{m_g} \frac{\phi\left(\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}}\right)}{\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}}}, \end{aligned}$$

where the inequality in the last line follows using the fact that $1-\Phi(t) \leq \frac{\phi(t)}{t}$, for any $t>0$. Under assumption (A2), we have

$$\min_j \beta_{gj}^0 > m_n \text{ for all } g \in \mathcal{A}_n \text{ with } m_n \propto n^{-b} \text{ and } 0 \leq b < \frac{1}{2}. \tag{4.46}$$

Next we use the fact $d_n \asymp \log\left(\frac{1}{\tau_n}\right)$ and the assumption (A4). Hence, we obtain $\frac{\sqrt{d_n}}{\sqrt{n}\beta_{gj}^0} \to 0$ as $n \to \infty$.

Therefore, we have

$$\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} = \frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}}(1+o(1)), \tag{4.47}$$

where $o(1)$ term tends to zero as $n \to \infty$. Since, $\frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} = o\left(\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}}\right)$, we have, for for any $\epsilon > 0$,

$$\frac{\sqrt{nd_n}\beta_{gj}^0}{\sigma\zeta_j\sqrt{C_1}} < \epsilon \frac{n{\beta_{gj}^0}^2}{\sigma^2\zeta_j} \tag{4.48}$$

for sufficiently large n. Let $e_1, e_2, \cdots, e_{m_g}$ be the eigenvalues of $\mathbf{Q}_{n,g}$. Then $\sum_{j=1}^{m_g} \frac{1}{e_j} = \sum_{j=1}^{m_g} \zeta_j$. Next note that, under (A1), $e_j > C_1$ for all $j = 1, 2, \cdots, m_g$. This implies, ζ_j is bounded above for all $j = 1, 2, \cdots, m_g$, i.e. for all $j = 1, 2, \cdots, m_g$, $\zeta_j < C_3$ for some $0 < C_3 < \infty$. Combining (4.45) - (4.48) and using (A3) with the above observation, it follows

$$P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) \leq \frac{\sigma s\sqrt{C_3}}{\sqrt{n}m_n} \exp\left\{-(\frac{1}{2}-\epsilon)\frac{nm_n^2}{2\sigma^2 C_3}\right\}. \tag{4.49}$$

Using (4.46) and choosing $\epsilon \in (0, \frac{1}{2})$ yields,

$$P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) = o(1), \text{ as } n \to \infty. \tag{4.50}$$

Case-(II): Assume $\min_j \beta_{gj}^0 < 0$. This implies some of $\beta_{gj}^0 < 0$'s are negative and the remaining are positive. Let us assume that $\beta_{gj}^0 < 0$ for all $j = 1, \cdots, \tilde{m}_g$, and $\beta_{gj}^0 > 0$ for all $j = \tilde{m}_g + 1, \cdots, m_g$ for some $\tilde{m}_g$. Hence again using (4.45) we have

$$\begin{aligned}
&P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) \\
&\leq \sum_{j=1}^{m_g} P\left(-\frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} \leq \frac{\sqrt{n}(\widehat{\beta}_{gj} - \beta_{gj}^0)}{\sigma\sqrt{\zeta_j}} \leq \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}}\right) \\
&\leq \sum_{j=1}^{\tilde{m}_g} \Big[P(Z > a_{1n,j}) - P(Z > a_{2n,j})\Big] + \sum_{j=\tilde{m}_g+1}^{m_g} \frac{\phi\left(\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}}\right)}{\frac{\sqrt{n}\beta_{gj}^0}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}}}, \text{ say}
\end{aligned} \tag{4.51}$$

where $a_{1n,j} = -\frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta^0_{gj}}{\sigma\sqrt{\zeta_j}}$ and $a_{2n,j} = \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} - \frac{\sqrt{n}\beta^0_{gj}}{\sigma\sqrt{\zeta_j}}$, $j = 1, 2, \cdots, \tilde{m}_g$ and $Z \sim \mathcal{N}(0,1)$. For providing an upper bound to the first term on the right hand side of (4.51), we use arguments used in (4.46)-(4.49). It is well known

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \left(\frac{1}{z} - \frac{1}{z^3}\right) \leq P(Z > z) \leq \frac{1}{\sqrt{2\pi} z} e^{-\frac{z^2}{2}}.$$

Using this it follows from (4.51) that

$$\sum_{j=1}^{\tilde{m}_g} \left[P(Z > a_{1n,j}) - P(Z > a_{2n,j}) \right] \leq \sum_{j=1}^{\tilde{m}_g} \frac{1}{\sqrt{2\pi}} \left[\frac{1}{a_{1n,j}} \exp\left(-\frac{a^2_{1n,j}}{2}\right) + \left(\frac{1}{a_{2n,j}} + \frac{1}{a^3_{2n,j}}\right) \exp\left(-\frac{a^2_{2n,j}}{2}\right) \right]$$
$$\leq \frac{3}{\sqrt{2\pi}} \sum_{j=1}^{\tilde{m}_g} \left[\frac{1}{a_{1n,j}} \exp\left(-\frac{a^2_{1n,j}}{2}\right) \right], \tag{4.52}$$

where inequality in the last line holds due to the use of $a_{1n,j} \leq a_{2n,j}, j = 1, 2, \cdots, \tilde{m}_g$. Note that using (A2) and the assumption (A4), we have $-\frac{\sqrt{d_n}}{\sqrt{n}\beta^0_{gj}} \to 0$ as $n \to \infty$.

Therefore, we have

$$a_{1n,j} = -\frac{\sqrt{n}\beta^0_{gj}}{\sigma\sqrt{\zeta_j}} - \frac{1}{\sqrt{C_1}} \cdot \sqrt{\frac{d_n}{\zeta_j}} = -\frac{\sqrt{n}\beta^0_{gj}}{\sigma\sqrt{\zeta_j}} (1 + o(1)), \tag{4.53}$$

where $o(1)$ term tends to zero as $n \to \infty$. Hence, using arguments similar to that of (4.48) and (4.49), we obtain

$$P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n \right) = o(1), \text{ as } n \to \infty. \tag{4.54}$$

Hence combining (4.44), (4.50) and (4.54), the proof of Proposition 4.2.3 is complete.

□

4.8.4 Proofs of Theorems

Proof of Theorem 4.3.1: First, we observe that

$$P(\widehat{\mathcal{A}}_n \neq \mathcal{A}_n) \leq \sum_{g \in \mathcal{A}_n} P(E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) < \frac{1}{2}) + \sum_{g \notin \mathcal{A}_n} P(E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}). \tag{4.55}$$

To prove this result, it suffices to show

$$\sum_{g \in \mathcal{A}_n} P(E(1 - \kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) < \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.56}$$

and

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.57}$$

both when $\frac{1}{2} \leq a < 1$, and $a \geq 1$.

Proof of (4.56): Fix an arbitrary $\epsilon_0 > 0$. Now, using the arguments employed in the proof of proposition 4.2.3 and applying Lemma 4.8.4 with $\eta = \frac{\epsilon_0}{2}$, we have,

$$\sum_{g \in \mathcal{A}_n} P(E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \leq \sum_{g \in \mathcal{A}_n} P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right). \tag{4.58}$$

Since $G_{\mathcal{A}_n} \leq n$, it follows from (4.49) that

$$\sum_{g \in \mathcal{A}_n} P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) \leq \frac{\sigma s \sqrt{C_3}\sqrt{n}}{m_n} \exp\left\{-(\frac{1}{2} - \epsilon)\frac{n m_n^2}{2\sigma^2 C_3}\right\}(1 + o(1)) \to 0, \text{ as } n \to \infty,$$

whence again using (4.46),

$$\sum_{g \in \mathcal{A}_n} P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} < d_n\right) = o(1), \text{ as } n \to \infty. \tag{4.59}$$

Hence, (4.58) coupled with (4.59) completes the proof of (4.56).

Proof of (4.57):

Case (I): First consider the case when $a \in [\frac{1}{2}, 1)$. Using Lemma 4.8.3 and our previous arguments, it follows that for all $g \notin \mathcal{A}_n$,

$$P\left(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}\right) \leq P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > M_n\right)(1 + o(1)), \tag{4.60}$$

where $M_n = 2\log\left(\frac{C_4}{(\tau_n^2)^a L(\frac{1}{\tau_n^2})}\right)$ and C_4 is a global constant that is independent of n. In (4.60), the $o(1)$ is such that it is independent of any specific group g, and $\lim_{n\to\infty} o(1) = 0$. Now observe that for all $g \notin \mathcal{A}_n$, $\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} \sim \chi^2_{m_g}$. We consider the two cases $m_g = 1$, and $m_g > 1$ separately. For $m_g = 1$, we use

$$P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > M_n\right) = P(|Z| > \sqrt{M_n}) = \sqrt{\frac{2}{\pi}} e^{-\frac{M_n}{2}} M_n^{-\frac{1}{2}}(1 + o(1)),$$

where Z denotes a standard normal random variable and the last line follows due to Mill's ratio. On the other hand, for $m_g \geq 2$, we first observe that a $\chi^2_{m_g}$ distribution can equivalently be regarded as a Gamma$(\frac{m_g}{2}, \frac{1}{2})$ distribution, with shape parameter $\frac{m_g}{2}$, and scale parameter $\frac{1}{2}$.

Then we have

$$P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > M_n\right) = \frac{1}{2^{\frac{m_g}{2}}\Gamma(\frac{m_g}{2})}\int_{M_n}^{\infty} e^{-\frac{u}{2}}u^{\frac{m_g}{2}-1}du = \frac{1}{\Gamma(\frac{m_g}{2})}\int_{M_n/2}^{\infty} e^{-u}u^{\frac{m_g}{2}-1}du, \quad (4.61)$$

where $\Gamma(r) = \int_0^{\infty} e^{-u}u^{r-1}du$ denotes the gamma function evaluated at $r > 0$. Now, we state below a result due to Gabcke (2015) that is instrumental in completing the remainder of this proof. This is presented as Lemma 4.8.5 below.

Lemma 4.8.5. *When $r \geq 1$ and $c > r + 1$,*

$$e^{-c}c^{r-1} \leq \int_c^{\infty} e^{-u}u^{r-1}du \leq re^{-c}c^{r-1},$$

that is, for sufficiently large $c > 0$,

$$\int_c^{\infty} e^{-u}u^{r-1}du \lesssim re^{-c}c^{r-1}.$$

Thus, using Lemma 4.8.5 coupled with the (4.61) and the fact that $M_n \to \infty$ as $n \to \infty$, we have, for all sufficiently large n, for all $g \notin \mathcal{A}_n$,

$$e^{-\frac{M_n}{2}}M_n^{\frac{m_g}{2}-1} \leq P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > M_n\right) \lesssim e^{-\frac{M_n}{2}}M_n^{\frac{s}{2}-1}, \quad (4.62)$$

where $s = \sup_{n\geq 1}\max_{g\in\{1,2,\cdots,G_n\}} m_g$, is finite. Using this observation, and combining (4.60)-(4.62), we have,

$$\sum_{g\notin\mathcal{A}_n} P\left(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}\right) \lesssim G_n(\tau_n^2)^a L(\frac{1}{\tau_n^2})\left[\log\left(\frac{1}{(\tau_n^2)^a L(\frac{1}{\tau_n^2})}\right)\right]^{\frac{s}{2}-1}. \quad (4.63)$$

Hence, for $a \in [\frac{1}{2}, 1)$ using Assumption 1 on $L(\cdot)$, the term of the right-hand side of (4.63) converges to 0, as $n \to \infty$ if $G_n\tau_n[\log\left(\frac{1}{\tau_n}\right)]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$. Note that (B1) and (B2) imply $G_n\tau_n\left[\log\left(\frac{1}{\tau_n}\right)\right]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$. This completes the proof of (4.57) when $\frac{1}{2} \leq a < 1$.

Case (II): Now we consider the situation $a \geq 1$. Using similar arguments employed to prove **Case (I)**, one can easily verify that there exists a constant C_5 independent of n, such that $M_n = 2\log\left(\frac{C_5}{\tau_n}\right)$ and

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \lesssim G_n\tau_n\left[\log\left(\frac{1}{\tau_n}\right)\right]^{\frac{s}{2}-1}. \quad (4.64)$$

Again observe (B1) and (B2) imply $G_n\tau_n\left[\log\left(\frac{1}{\tau_n}\right)\right]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$, for $a \geq 1$. Hence under the assumption of (B1) and (B2), the right hand side of (4.64) goes to 0 as $n \to \infty$, for each fixed $a \geq 1$, which establishes (4.57). This completes the proof of Theorem 4.3.1.

□

Proof of Theorem 4.3.4: Define $T = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \boldsymbol{\beta}^0_{\mathcal{A}_n})$. Then $T = T_1 + T_2$, where $T_1 = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n} - \boldsymbol{\beta}^0_{\mathcal{A}_n})$ and $T_2 = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n})$. Now, it boils down to show that,

$$T_1 \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } n \to \infty, \tag{4.65}$$

and

$$T_2 \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{4.66}$$

First, we prove (4.65). Note that, due to the block-diagonal property of the design matrix $\mathbf{X}$, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n} = \boldsymbol{\Sigma}^{-1}_{\mathcal{A}_n}\mathbf{X}^{\mathrm{T}}_{\mathcal{A}_n}\mathbf{y}$ and using the standard theory of linear models, it readily follows that for $||\boldsymbol{\alpha}|| = 1$, $T_1 \sim \mathcal{N}(0, \sigma^2)$.

For the time being, let us assume (4.66) to be true. Then, combining (4.65) and (4.66), coupled with Slutsky's Theorem, the desired asymptotic normality of $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n}$ follows.

We now turn our focus on establishing (4.66) above. Towards that, using Cauchy-Schwarz inequality, we have

$$\begin{aligned} T_2^2 &\leq (\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathcal{A}_n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}) \\ &\leq C_2 n(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n})^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A}_n} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}) = C_2 \sum_{g \in \mathcal{A}_n} \left\| \sqrt{n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_g - \widehat{\boldsymbol{\beta}}_g) \right\|_2^2. \end{aligned} \tag{4.67}$$

The second inequality in (4.67) follows due to the assumption on the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ as given in (C1) and using the fact that $e_{max}(\mathbf{X}^{\mathrm{T}}_{\mathcal{A}_n}\mathbf{X}_{\mathcal{A}_n}) \leq e_{max}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$. Next, observe that using the form of the posterior mean $\widehat{\boldsymbol{\beta}}^{\mathrm{PM}}_g$ as given by (4.7) coupled with the definition of the half-thresholding estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_g$ given by (4.10), one may rewrite the difference $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_g - \widehat{\boldsymbol{\beta}}_g)$ as

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_g - \widehat{\boldsymbol{\beta}}_g) &= \sqrt{n}\left[E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) > 0.5\right\} - 1\right]\widehat{\boldsymbol{\beta}}_g \\ &= -\sqrt{n}\widehat{\boldsymbol{\beta}}_g E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) - \sqrt{n}\widehat{\boldsymbol{\beta}}_g E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq 0.5\right\}. \end{aligned} \tag{4.68}$$

Note that

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq 0.5 \text{ if and only if } E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \geq 0.5. \tag{4.69}$$

Thus,

$$0 \leq E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq 0.5\right\} \leq E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}), \tag{4.70}$$

whence

$$\|\sqrt{n}\widehat{\boldsymbol{\beta}}_g E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq 0.5\right\}\| \leq \|\sqrt{n}\widehat{\boldsymbol{\beta}}_g E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})\|. \tag{4.71}$$

Now to establish (4.66), let us first define the following random variables:

$$W_{n,g} = \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2}, \text{ and } U_{n,g} = W_{n,g}E(\kappa_g^2 \mid \tau_n, \sigma^2, \mathcal{D}).$$

Combining (4.67) - (4.71) together with the triangle inequality for the ℓ_2 norm, we obtain

$$\begin{aligned} T_2^2 &\leq 4C_2 \sum_{g\in\mathcal{A}_n} n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g E(\kappa_g^2 \mid \tau_n, \sigma^2, \mathcal{D}) \\ &\leq \frac{4C_2\sigma^2}{C_1} \sum_{g\in\mathcal{A}_n} \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} E(\kappa_g^2 \mid \tau_n, \sigma^2, \mathcal{D}) = \frac{4C_2\sigma^2}{C_1} \sum_{g\in\mathcal{A}_n} U_{ng}. \end{aligned} \tag{4.72}$$

The second inequality above follows using (C1) and the fact that for all g $\in \mathcal{A}, e_{min}(\frac{\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g}{n}) \geq e_{min}(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n})$. Thus to prove (4.66), it is enough to show that

$$\sum_{g\in\mathcal{A}_n} U_{ng} \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{4.73}$$

Using the above definitions coupled with (4.17), we obtain

$$\begin{aligned} U_{n,g} &= W_{n,g} \frac{\int_0^1 \kappa_g^2 \cdot \kappa_g^{(a+\frac{m_g}{2}-1)}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right) \exp\left(-\kappa_g \cdot \frac{W_{n,g}}{2}\right) d\kappa_g}{\int_0^1 \kappa_g^{(a+\frac{m_g}{2}-1)}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right) \exp\left(-\kappa_g \cdot \frac{W_{n,g}}{2}\right) d\kappa_g} \\ &= J(W_{n,g}, \tau), \text{ say}. \end{aligned} \tag{4.74}$$

Next, we follow the arguments used in Lemma 3 of Ghosh and Chakrabarti (2017) to find an upper bound to $J(W_{n,g}, \tau)$. First note that, given any $c > 2$, we can find $\eta, q \in (0,1)$ such that $c = \frac{2}{\eta(1-q)}$. Following Ghosh and Chakrabarti (2017) there exists a non-negative measurable function $h(W_{n,g}, \tau) = h_1(W_{n,g}, \tau) + h_2(W_{n,g}, \tau)$, where $h_1(W_{n,g}, \tau) = C_*[W_{n,g}\int_0^{\frac{W_{n,g}}{1+t_0}} e^{-\frac{u}{2}} u^{m_g+a-1} du]^{-1}$ and $h_2(W_{n,g}, \tau) = C^* W_{n,g}\tau^{-2a} e^{-\frac{\eta(1-q)}{2}W_{n,g}}$ where t_0 is as in Assumption 1 and C_* and C^* are two constants which depends on $a, \eta, q, L(\cdot)$ and satisfies:
for any $W_{n,g}$,

$$J(W_{n,g}, \tau) \leq h(W_{n,g}, \tau), \tag{4.75}$$

and we also have for any $\rho > c$,

$$\lim_{\tau_n \to 0} \sup_{W_{n,g} > 2a\rho \frac{1}{\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right)}} h(W_{n,g}, \tau_n) = 0. \tag{4.76}$$

Let $\epsilon > 0$ be given. Then

$$P\Big(\sum_{g\in\mathcal{A}_n} U_{ng} > \epsilon\Big) \leq \sum_{g\in\mathcal{A}_n} P(U_{ng} > \frac{\epsilon}{|\mathcal{A}_n|}).$$

Let us fix some $c > 2$ and any $\rho > c$. Let B_n and C_n denote the events $\{U_{n,g} > \frac{\epsilon}{|\mathcal{A}_n|}\}$ and

$\{W_{n,g} > 2a\rho\frac{1}{\sqrt{\tau}\log\left(\frac{1}{\tau}\right)}\}$, respectively. Then,

$$\sum_{g\in\mathcal{A}_n} P(U_{n,g} > \frac{\epsilon}{|\mathcal{A}_n|}) = \sum_{g\in\mathcal{A}_n} P(B_n) = \sum_{g\in\mathcal{A}_n} P(B_n \cap C_n) + \sum_{g\in\mathcal{A}_n} P(B_n \cap C_n^c)$$
$$\leq \sum_{g\in\mathcal{A}_n} P(B_n \cap C_n) + \sum_{g\in\mathcal{A}_n} P(C_n^c). \tag{4.77}$$

Using (4.74) and (4.75) along with the form of $h_k(W_{n,g},\tau), k = 1, 2$, it follows that for sufficiently large n, under the assumption (C3),

$$\sup_{W_{n,g}>2a\rho\frac{1}{\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right)}} h_1(W_{n,g},\tau_n) \leq \sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right) = o(\frac{1}{|\mathcal{A}_n|}). \tag{4.78}$$

Also note that

$$\sup_{W_{n,g}>2a\rho\frac{1}{\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right)}} h_2(W_{n,g},\tau) \leq \frac{\tau_n^{-2a-\frac{1}{2}}}{\log\left(\frac{1}{\tau_n}\right)} e^{-\frac{a\rho\eta(1-q)}{\sqrt{\tau_n}\log\left(\frac{1}{\tau_n}\right)}} = o(\frac{1}{|\mathcal{A}_n|}). \tag{4.79}$$

Observe that, with the definition of B_n, and using (4.78) and (4.79) we have, for sufficiently large n,

$$\text{for all } g \in \mathcal{A}_n, P(B_n \cap C_n) = 0. \tag{4.80}$$

This implies the first term in the right-hand side of (4.77) goes to zero and we are left with the second term only.

For the second term, under the assumption of (C2) and (C3), using similar set of arguments used in (4.45)-(4.59), we can show that

$$\lim_{n\to\infty} \sum_{g\in\mathcal{A}_n} P(C_n^c) = 0. \tag{4.81}$$

Since $\epsilon > 0$ is arbitrary, combining (4.74)-(4.81) ensures (4.73) holds. This completes the proof of Theorem 4.3.4. □

Proof of Theorem 4.3.7: As noted before,

$$P(\widehat{\mathcal{A}}_n^{(\mathrm{EB})} \neq \mathcal{A}_n) \leq \sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) < \frac{1}{2}) + \sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}). \tag{4.82}$$

To prove (4.82), it suffices to show

$$\sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) < \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.83}$$

and

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.84}$$

both when $0.5 < a < 1$, and $a \geq 1$.
Note that, for fixed $\mathcal{D} = \{\mathbf{y}\}$ and σ^2, $E(\kappa_g \mid \tau, \sigma^2, \mathcal{D})$ is a non-increasing function of τ. Moreover, $\widehat{\tau}^{\text{EB}} \geq \gamma_n$, where $\gamma_n = \frac{1}{G_n}$, for $n \geq 1$. Combining these two facts, we obtain

$$\sum_{g \in \mathcal{A}_n} P(E(\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) \leq \sum_{g \in \mathcal{A}_n} P(E(\kappa_g \mid \gamma_n, \sigma^2, \mathcal{D}) > \frac{1}{2}). \tag{4.85}$$

Observe that, the sequence $\gamma_n = G_n^{-1}$ for $n \geq 1$ satisfies $\log\left(\frac{1}{\tau_n}\right) \asymp \log(G_n)$. Therefore, under assumptions (A1)-(A3) of Theorem 4.3.1, using exactly the same set of arguments employed in proving Theorem 4.3.1 (when τ was a tuning parameter), one has

$$\lim_{n \to \infty} \sum_{g \in \mathcal{A}_n} P(E(\kappa_g \mid \gamma_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) = 0. \tag{4.86}$$

Therefore, (4.85) and (4.86) together yield

$$\lim_{n \to \infty} \sum_{g \in \mathcal{A}_n} P(E(\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) = 0,$$

which completes the proof of (4.83). Now, we are left to prove (4.84). Note that, for any $\alpha_n > 0$,

$$\begin{aligned} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) = P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ \widehat{\tau}^{\text{EB}} \leq 2\alpha_n) + \\ P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ \widehat{\tau}^{\text{EB}} > 2\alpha_n). \end{aligned} \tag{4.87}$$

To complete the proof, we now appropriately choose $\{\alpha_n\}_{n \geq 1} > 0$ such that $\alpha_n \to 0$ as $n \to \infty$ so that both

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\text{EB}} > 2\alpha_n) = o(1), \text{ as } n \to \infty \tag{4.88}$$

and

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\text{EB}} \leq 2\alpha_n) = o(1), \text{ as } n \to \infty. \tag{4.89}$$

For studying (4.88), we define $\widehat{\tau}_1 = \frac{1}{G_n}$ and $\widehat{\tau}_2 = \frac{1}{c_2 G_n} \sum_{g=1}^{G_n} 1\left\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n\right\}$, where $c_1 \geq 2$, and $c_2 \geq 1$. Clearly, $\widehat{\tau}^{\text{EB}} = \max\left\{\widehat{\tau}_1, \widehat{\tau}_2\right\}$. Therefore, we have,

$$\begin{aligned} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ \widehat{\tau}^{\text{EB}} > 2\alpha_n) &\leq P(\widehat{\tau}^{\text{EB}} > 2\alpha_n) \\ &\leq P(\widehat{\tau}_1 > 2\alpha_n) + P(\widehat{\tau}_2 > 2\alpha_n). \end{aligned} \tag{4.90}$$

We note that if $\alpha_n > 0$ is such that

$$\frac{1}{G_n} \leq 2\alpha_n, \text{ for all sufficiently large } n, \tag{4.91}$$

whence

$$P(\widehat{\tau}_1 > 2\alpha_n) = 0, \text{ for all sufficiently large } n. \tag{4.92}$$

Thus, (4.90) coupled with (4.92) yields

$$P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \epsilon_0, \ \widehat{\tau}^{\text{EB}} > 2\alpha_n) \leq P(\widehat{\tau}_2 > 2\alpha_n), \tag{4.93}$$

for all sufficiently large n for α_n satisfying (4.91). Let us define $\widehat{\tau}_3 = \frac{1}{c_2 G_n} \sum_{g \in \mathcal{A}_n} 1\left\{ \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n \right\}$, and $\widehat{\tau}_4 = \frac{1}{c_2 G_n} \sum_{g \notin \mathcal{A}_n} 1\left\{ \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n \right\}$, so that $\widehat{\tau}_2 = \widehat{\tau}_3 + \widehat{\tau}_4$. Clearly,

$$P(\widehat{\tau}_2 > 2\alpha_n) \leq P(\widehat{\tau}_3 > \alpha_n) + P(\widehat{\tau}_4 > \alpha_n). \tag{4.94}$$

Observe that

$$\widehat{\tau}_3 = \frac{1}{c_2 G_n} \sum_{g \in \mathcal{A}_n} 1\left\{ \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1 \log G_n \right\} \leq \frac{1}{c_2 G_n} \sum_{g \in \mathcal{A}_n} 1 = \frac{G_{\mathcal{A}_n}}{c_2 G_n}. \tag{4.95}$$

We now observe that if $\alpha_n > 0$ is chosen such that

$$\frac{G_{\mathcal{A}_n}}{c_2 G_n} \leq \alpha_n, \text{ for all sufficiently large } n, \tag{4.96}$$

then

$$P(\widehat{\tau}_3 > \alpha_n) = 0, \text{ for all sufficiently large } n. \tag{4.97}$$

Note that condition (4.96) implies condition (4.91) and therefore (4.92) is automatically satisfied for this choice of α_n. For bounding the second term in the right-hand side of (4.94), we consider the generalized version of Chernoff-Hoeffding bound for independent but non i.i.d. sequence of random variables, which is stated in the next lemma.

Lemma 4.8.6. *Let $Z_1, Z_2, \cdots, Z_m$ be m independent $0-1$ random variables with $\mathbb{E}(Z_i) = p_i$, $i = 1, 2, \cdots, m$. Let $Z = \sum_{i=1}^m Z_i$, $\mu = \mathbb{E}(Z) = \sum_{i=1}^m p_i$ and $p = \frac{\mu}{m}$. Then*

$$\mathbb{P}(Z \geq \mu + \lambda) \leq \exp\left\{ -\{mH_p(p + \frac{\lambda}{m})\} \right\}, \quad \textit{for} \ \ 0 < \lambda < m - \mu,$$

and

$$\mathbb{P}(Z \leq \mu - \lambda) \leq \exp\left\{ -\{mH_{1-p}(1 - p + \frac{\lambda}{m})\} \right\}, \quad \textit{for} \ \ 0 < \lambda < \mu,$$

where $H_p(x) = x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right)$ *is the relative entropy of* x *w.r.t.* p.

Note that

$$\begin{aligned}
P(\widehat{\tau}_2 > 2\alpha_n) &\leq P(\widehat{\tau}_4 > \alpha_n) \\
&= P\Bigg[\sum_{g\notin\mathcal{A}_n} 1\Bigg\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\Bigg\} > c_2\alpha_n G_n\Bigg] \\
&= P\Bigg[\frac{1}{(G_n - G_{\mathcal{A}_n})}\sum_{g\notin\mathcal{A}_n} 1\Bigg\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\Bigg\} > c_2\frac{\alpha_n G_n}{(G_n - G_{\mathcal{A}_n})}\Bigg] \\
&\leq P\Bigg[\frac{1}{(G_n - G_{\mathcal{A}_n})}\sum_{g\notin\mathcal{A}_n} 1\Bigg\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\Bigg\} \geq \alpha_n\Bigg] \\
&= P\Big[S_n - \mathbb{E}(S_n) \geq \alpha_n(G_n - G_{\mathcal{A}_n}) - \mathbb{E}(S_n)\Big] \quad \text{say,} \qquad (4.98)
\end{aligned}$$

where the second inequality holds due to $c_2 G_n > G_n - G_{\mathcal{A}_n}$ and $S_n = \sum_{g\notin\mathcal{A}_n} 1\Big\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\Big\}$. To find an upper bound for (4.98), we use Lemma 4.8.6 by taking $m = G_n - G_{\mathcal{A}_n}$, $Z_i = 1\Big\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\Big\}$, $\mathbb{E}(Z_i) = p_i = P\left(\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\right) \leq m_g G_n^{-\frac{c_1}{2}}(\log G_n)^{\frac{m_g}{2}-1}$, $\mu \leq sG_n^{-\frac{c_1}{2}+1}(\log G_n)^{\frac{s}{2}-1}$, and $p = \frac{\mu}{G_n - G_{\mathcal{A}_n}} \leq sG_n^{-\frac{c_1}{2}}(\log G_n)^{\frac{s}{2}-1}$, where s denotes the maximum of the group size. For $\lambda = \alpha_n(G_n - G_{\mathcal{A}_n}) - \mu$, we have $0 < \lambda < m - \mu$. Also, we have $p + \frac{\lambda}{G_n - G_{\mathcal{A}_n}} = \alpha_n$. Hence,

$$H_p(p + \frac{\lambda}{G_n - G_{\mathcal{A}_n}}) = \alpha_n\log\left(\frac{\alpha_n}{p}\right) + (1-\alpha_n)\log\left(\frac{1-\alpha_n}{1-p}\right). \qquad (4.99)$$

Also, note that, since $c_1 \geq 2$, $\frac{p}{\alpha_n} \to 0$ as $n \to \infty$ under the assumption that $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$.

Recall that, $\frac{\log\left(\frac{1}{1-y}\right)}{y} \to 1$ as $y \to 0$. Hence, with $y = \frac{p-\alpha_n}{1-\alpha_n}$, the second term in the right hand side of (4.99) is of the form

$$\log\left(\frac{1-\alpha_n}{1-p}\right) = \frac{p-\alpha_n}{1-\alpha_n}(1 + o(1)), \qquad (4.100)$$

where $o(1)$ depends only on n such that $\lim_{n\to\infty} o(1) = 0$. Hence, using (4.99) and (4.100), an lower bound of $H_p(\alpha_n)$ is given by

$$\begin{aligned}
H_p(\alpha_n) &= \alpha_n\log\left(\frac{\alpha_n}{p}\right) + (1-\alpha_n)\cdot\frac{(p-\alpha_n)}{(1-\alpha_n)}(1+o(1)) \\
&= \alpha_n\log\left(\frac{\alpha_n}{p}\right)(1+o(1)) \gtrsim \alpha_n(1+o(1)). \qquad (4.101)
\end{aligned}$$

With the use of Lemma 4.8.6 and (4.101) with the assumption $G_{\mathcal{A}_n} = o(G_n)$, the upper bound

of (4.98) is obtained as

$$P(\widehat{\tau}_2 > 2\alpha_n) \leq P\left[\frac{1}{(G_n - G_{\mathcal{A}_n})}\sum_{g\notin\mathcal{A}_n} 1\left\{\frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2} > c_1\log G_n\right\} \geq \alpha_n\right]$$
$$\leq e^{-\alpha_n G_n(1+o(1))} \leq e^{-\frac{G_{\mathcal{A}_n}}{c_2}(1+o(1))}, \tag{4.102}$$

where inequality in the last line holds due to (4.96). By choosing $\alpha_n = c_2\frac{G_{\mathcal{A}_n}}{G_n}$, we immediately see that $\alpha_n \to 0$ and satisfies (4.91) and (4.96). Thus, with the above choice of α_n, and combining (4.93) and (4.102), we obtain

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ \widehat{\tau}^{\mathrm{EB}} > 2\alpha_n) \leq G_n e^{-\frac{G_{\mathcal{A}_n}}{c_2}(1+o(1))}.$$

Since, $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$, we can conclude that

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} > 2\alpha_n) = o(1),\ \text{as } n\to\infty. \tag{4.103}$$

We now proceed to prove that

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} \leq 2\alpha_n) = o(1),\ \text{as } n\to\infty. \tag{4.104}$$

Case-1 $a \geq 1$. Now using the monotonicity of the shrinkage coefficient and then by Markov's inequality, the term in the left hand side of (4.104) can be bounded as

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} \leq 2\alpha_n)$$
$$\leq \sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \leq 2\sum_{g\notin\mathcal{A}_n} E\left(E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D})\right), \tag{4.105}$$

where the outer expectation is w.r.t. to $W_{n,g} = \frac{n\widehat{\boldsymbol{\beta}}_g^{\mathrm{T}}\mathbf{Q}_{n,g}\widehat{\boldsymbol{\beta}}_g}{\sigma^2}$. Since, by definition, $E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D}) \leq 1$, so, we have

$$E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D}) \leq 1_{\{W_{n,g} > 2c_1\log G_n\}} + E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D})1_{\{W_{n,g}\leq 2c_1\log G_n\}}. \tag{4.106}$$

Next, with the use of arguments similar to Lemma 1 of Paul and Chakrabarti (2025), we provide an upper bound on $E(1-\kappa_g \mid \tau, \sigma^2, \mathcal{D})$ for any $\tau \in (0,1)$.

$$E(1-\kappa_g \mid \tau, \sigma^2, \mathcal{D}) \leq \left(\tau^2 e^{\frac{W_{n,g}}{4}} + K\int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot\frac{1}{(1+t\tau^2)^{\frac{m_g}{2}}}t^{-a-1}L(t)e^{\frac{t\tau^2}{1+t\tau^2}\cdot\frac{W_{n,g}}{2}}dt\right)(1+o(1)), \tag{4.107}$$

where the term $o(1)$ depends only on τ and $\lim_{\tau\to 0} o(1) = 0$ and $\int_0^\infty t^{-a-1}L(t)dt = K^{-1}$.

Now, using (4.106) and (4.107), we obtain, for any $\tau \in (0,1)$

$$E\left(E(1-\kappa_g \mid \tau, \sigma^2, \mathcal{D})\right) \leq P(W_{n,g} > 2c_1 \log G_n) + E\left[\left(\tau^2 e^{\frac{W_{n,g}}{4}} + K\int_1^\infty \frac{t^{-a}\tau^2}{(1+t\tau^2)^{\frac{m_g}{2}+1}} L(t) e^{\frac{t\tau^2}{1+t\tau^2}\cdot\frac{W_{n,g}}{2}} dt\right)(1+o(1))1_{\{W_{n,g}\leq 2c_1\log G_n\}}\right] \tag{4.108}$$

Since, for all $g \notin \mathcal{A}_n$, $W_{n,g} \sim \chi^2_{m_g}$, so, by Lemma 4.8.5,

$$P(W_{n,g} > 2c_1 \log G_n) \lesssim G_n^{-c_1}(\log G_n)^{\frac{s}{2}-1}. \tag{4.109}$$

Next, for the second term in the right-hand side of (4.108), noting that the term $(1+o(1))$ is independent of any particular g,

$$\begin{aligned} E\left[\tau^2 e^{\frac{W_{n,g}}{4}}(1+o(1))1_{\{W_{n,g}\leq 2c_1\log G_n\}}\right] &= \tau^2 \frac{1}{2^{\frac{m_g}{2}}\Gamma(\frac{m_g}{2})}\int_0^{2c_1\log G_n} e^{\frac{u}{4}} e^{-\frac{u}{2}} u^{\frac{m_g}{2}-1} du(1+o(1)) \\ &= \tau^2 \frac{1}{\Gamma(\frac{m_g}{2})}\int_0^{c_1\log G_n} e^{-\frac{u}{2}} u^{\frac{m_g}{2}-1} du(1+o(1)) \\ &\lesssim \tau^2(1+o(1)). \end{aligned} \tag{4.110}$$

Finally, for the third term in the right-hand side of (4.108), note that

$$\begin{aligned} &\int_0^{2c_1\log G_n}\int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot\frac{1}{(1+t\tau^2)^{\frac{m_g}{2}}} t^{-a-1}L(t) e^{\frac{t\tau^2}{1+t\tau^2}\cdot\frac{u}{2}} e^{-\frac{u}{2}} u^{\frac{m_g}{2}-1} dt du \\ &= \int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot\frac{1}{(1+t\tau^2)^{\frac{m_g}{2}}} t^{-a-1}L(t)\left(\int_0^{2c_1\log G_n} e^{-\frac{u}{2(1+t\tau^2)}} u^{\frac{m_g}{2}-1} du\right) dt \\ &= \int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot t^{-a-1}L(t)\left(\int_0^{\frac{2c_1\log G_n}{1+t\tau^2}} e^{-\frac{z}{2}} z^{\frac{m_g}{2}-1} dz\right) dt. \end{aligned} \tag{4.111}$$

Now, we provide an upper bound on (4.111) separately for $a = 1$ and $a > 1$.
For $a > 1$, using the boundedness of $L(t)$, it is easy to show that,

$$\int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot t^{-a-1}L(t)\left(\int_0^{\frac{2c_1\log G_n}{1+t\tau^2}} e^{-\frac{z}{2}} z^{\frac{m_g}{2}-1} dz\right) dt \lesssim \tau^2. \tag{4.112}$$

For $a = 1$, we have the following

$$\begin{aligned} &\int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot t^{-a-1}L(t)\left(\int_0^{\frac{2c_1\log G_n}{1+t\tau^2}} e^{-\frac{z}{2}} z^{\frac{m_g}{2}-1} dz\right) dt \\ &\leq \int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot t^{-a-1}L(t)\left(\int_0^{\frac{2c_1\log G_n}{t\tau^2}} e^{-\frac{z}{2}} z^{\frac{m_g}{2}-1} dz\right) dt \leq C_7\int_1^\infty \frac{t\tau^2}{1+t\tau^2}\cdot t^{-a-1}L(t)dt, \end{aligned}$$

where C_7 is a global constant independent of n. Next, note that, $1+t\tau^2 \geq \sqrt{t}$ if and only if

$t \geq \frac{1}{\tau^4}$. As a result, we have

$$\int_1^{\frac{1}{\tau^4}} \frac{t\tau^2}{1+t\tau^2} \cdot t^{-2} L(t) dt \leq M\tau^2 \int_1^{\frac{1}{\tau^4}} t^{-1} dt = M\tau^2 \log\left(\frac{1}{\tau^4}\right), \tag{4.113}$$

and

$$\int_{\frac{1}{\tau^4}}^{\infty} \frac{t\tau^2}{1+t\tau^2} \cdot t^{-2} L(t) dt \leq M\tau^2 \int_{\frac{1}{\tau^4}}^{\infty} t^{-\frac{3}{2}} dt = 2M\tau^4. \tag{4.114}$$

Hence, for $a = 1$, the upper bound on (4.111) is obtained as

$$\int_1^{\infty} \frac{t\tau^2}{1+t\tau^2} \cdot t^{-a-1} L(t) \left(\int_0^{\frac{2c_1 \log G_n}{1+t\tau^2}} e^{-\frac{z}{2}} z^{\frac{m_g}{2}-1} dz \right) dt \leq 2M\tau^2 [\log\left(\frac{1}{\tau^4}\right) + \tau^2]. \tag{4.115}$$

As a consequence, for $a \geq 1$, the upper bound of the third term in the right-hand side of (4.108) is of the form

$$\begin{aligned} &E\left[K \int_1^{\infty} \frac{t\tau^2}{1+t\tau^2} \cdot \frac{1}{(1+t\tau^2)^{\frac{m_g}{2}}} t^{-a-1} L(t) e^{\frac{t\tau^2}{1+t\tau^2} \cdot \frac{W_{n,g}}{2}} dt (1+o(1)) 1_{\{W_{n,g} \leq c_1 \log G_n\}} \right] \\ &\lesssim \tau^2 [\log\left(\frac{1}{\tau^2}\right) + \tau^2](1+o(1)). \end{aligned} \tag{4.116}$$

Finally, substituting the upper bounds obtained in (4.109), (4.110) and (4.116) in (4.108) with $\tau = \alpha_n = \frac{c_2 G_{\mathcal{A}_n}}{G_n}$, the upper bound on (4.104) can be obtained as

$$\begin{aligned} &\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} \leq 2\alpha_n) \\ &\lesssim G_n^{-c_1+1} (\log G_n)^{\frac{s}{2}-1} + \frac{(G_{\mathcal{A}_n})^2}{G_n}(1+o(1)) + \frac{(G_{\mathcal{A}_n})^2}{G_n} \log G_n (1+o(1)). \end{aligned}$$

Hence, for $c_1 \geq 2$ and $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$, we have

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} \leq 2\alpha_n) = o(1), \text{ as } n \to \infty. \tag{4.117}$$

This completes the proof for $a \geq 1$.

Case-II $\frac{1}{2} < a < 1$. Again using the monotonicity of the shrinkage coefficient, the term in the left-hand side of (4.104) can be bounded as

$$\begin{aligned} &\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\mathrm{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\mathrm{EB}} \leq 2\alpha_n) \\ &\leq \sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid 2\alpha_n, \sigma^2, \mathcal{D}) > \frac{1}{2}) \lesssim G_n \left(\frac{G_{\mathcal{A}_n}}{G_n}\right)^{2a} \left[\log\left(\frac{G_n}{G_{\mathcal{A}_n}}\right)\right]^{\frac{s}{2}-1} (1+o(1)). \end{aligned}$$

Here inequality in the last line follows due to the use of arguments similar to the proof of

Theorem 4.3.1 where τ is assumed to be a tuning one. Note that, for $\frac{1}{2} < a < 1$, there exists $\epsilon \in (0, 1 - \frac{1}{2a})$ such that $2a(1-\epsilon) > 1$. Hence, when $\frac{1}{2} < a < 1$, for $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < 1 - \frac{1}{2a}$, we conclude

$$\sum_{g \notin \mathcal{A}_n} P(E(1 - \kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, \widehat{\tau}^{\text{EB}} \leq 2\alpha_n) = o(1), \text{ as } n \to \infty \tag{4.118}$$

and the proof is completed for $\frac{1}{2} < a < 1$. □

Proof of Theorem 4.3.8: To prove Theorem 4.3.8, we employ a similar set of arguments used in the proof of Theorem 4.3.4 when τ is used as a tuning parameter. Here, T_1 is as defined in Theorem 4.3.4 and $T_2 = \boldsymbol{\alpha}^{\text{T}} \boldsymbol{\Sigma}_{\mathcal{A}_n}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n, EB}^{\text{HT}} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n})$. As in Theorem 4.3.4, it follows that for proving $T_2 \xrightarrow{P} 0$, it suffices to show that

$$\sum_{g \in \mathcal{A}_n} U_{ng} \xrightarrow{P} 0 \text{ as } n \to \infty, \tag{4.119}$$

where

$$W_{n,g} = \frac{n \widehat{\boldsymbol{\beta}}_g^{\text{T}} \mathbf{Q}_{n,g} \widehat{\boldsymbol{\beta}}_g}{\sigma^2}, \text{ and } U_{n,g} = W_{n,g} E(\kappa_g^2 \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}).$$

Now to establish (4.119), let us first fix any $\epsilon_0 > 0$, and take $\gamma_n = \frac{1}{G_n}$, for $n \geq 1$. Note that, for fixed $\mathcal{D} = \{\mathbf{y}\}$ and σ^2, $E(\kappa_g \mid \tau, \sigma^2, \mathcal{D})$ is non-increasing in τ, and $\widehat{\tau}^{\text{EB}} \geq \gamma_n$ for all $n \geq 1$. Therefore, for $\tau_n = \gamma_n$ and using the monotonicity of τ, we only need to show that

$$\sum_{g \in \mathcal{A}_n} U_{n,g}^* = \sum_{g \in \mathcal{A}_n} W_{n,g} E(\kappa_g^2 \mid \gamma_n, \sigma^2, \mathcal{D}) \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{4.120}$$

Next, we proceed following the steps as mentioned in (4.74)-(4.77) with $\tau = \frac{1}{G_n}$. Now we define C_n as $C_n = \{W_{n,g} > 2a\rho \frac{\sqrt{G_n}}{\log G_n}\}$ and B_n as $B_n = \{U_{n,g}^* > \frac{\epsilon_0}{|\mathcal{A}_n|}\}$. Hence, to obtain the optimal estimation rate, it is enough to show that

$$\lim_{n \to \infty} \sum_{g \in \mathcal{A}_n} P(B_n) = 0. \tag{4.121}$$

Here, observe that under the assumption that $|\mathcal{A}_n| = O(G_n^{\epsilon}), 0 < \epsilon < \frac{1}{2}$, we have, for $k = 1, 2$,

$$\sup_{W_{n,g} > 2a\rho \frac{\sqrt{G_n}}{\log G_n}} h_k(W_{n,g}, \frac{1}{G_n}) = o(\frac{1}{|\mathcal{A}_n|}).$$

As a consequence of this, we have for all $g \in \mathcal{A}_n$,

$$P(B_n \cap C_n) = 0.$$

This also ensures

$$\lim_{n\to\infty} \sum_{g\in\mathcal{A}_n} P(B_n \cap C_n) = 0. \tag{4.122}$$

Next note that for $\tau = \frac{1}{G_n}$, we have $G_n\sqrt{\tau}\log\left(\frac{1}{\tau}\right) \to \infty$ as $n \to \infty$, and hence using similar set of arguments used in (4.45)-(4.59), we can show that

$$\lim_{n\to\infty} \sum_{g\in\mathcal{A}_n} P(C_n^c) = 0. \tag{4.123}$$

Finally combining (4.122) and (4.123) implies (4.121) and (4.120) holds. As a result, (4.119) is also established, and that along with the use of Slutsky's Theorem completes the proof of Theorem 4.3.8. □

Proof of Theorem 4.3.10: Here also, we will show that

$$\sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) < \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.124}$$

and

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) = o(1), \text{ as } n \to \infty, \tag{4.125}$$

both when $0.5 \leq a < 1$, and $a \geq 1$, where $\widehat{\tau}^{\text{EB}}$ is defined in (4.18).
Now, using the same technique as used in the proof of Theorem 4.3.7 and taking $\gamma_n = (\frac{1}{G_n})^2$, for $n \geq 1$, and then following steps similar to (4.85) and (4.86), we obtain

$$\lim_{n\to\infty} \sum_{g\in\mathcal{A}_n} P(E(\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}) = 0,$$

which completes the proof of (4.124). Now, we are left to prove (4.125). Again following the same steps as given in the proof of Theorem 4.3.7 along with the use of (4.102) with the use of (4.22), we obtain

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ (\widehat{\tau}^{\text{EB}})^{\frac{1}{2}} > 2\alpha_n) \leq G_n e^{-\frac{G_{\mathcal{A}_n}}{c_2}(1+o(1))}.$$

Since, $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$, we can conclude that

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, (\widehat{\tau}^{\text{EB}})^{\frac{1}{2}} > 2\alpha_n) = o(1), \text{ as } n \to \infty. \tag{4.126}$$

Next using the monotonicity of $E(1-\kappa_g \mid \tau, \sigma^2, \mathcal{D})$, it follows

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2},\ (\widehat{\tau}^{\text{EB}})^{\frac{1}{2}} \leq 2\alpha_n) \leq \sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid (2\alpha_n)^2, \sigma^2, \mathcal{D}) > \frac{1}{2}) \lesssim G_n \alpha_n^2 [\log\left(\frac{1}{\alpha_n}\right)]^{\frac{s}{2}}.$$

With the choice of $\alpha_n = \frac{c_2 G_{\mathcal{A}_n}}{G_n}$, $c_2 \geq 1$ for all $n \geq 1$, and since, $G_n^{\epsilon_1} \lesssim G_{\mathcal{A}_n} \lesssim G_n^{\epsilon_2}$ for some $0 < \epsilon_1 < \epsilon_2 < \frac{1}{2}$, we conclude that

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \widehat{\tau}^{\text{EB}}, \sigma^2, \mathcal{D}) > \frac{1}{2}, (\widehat{\tau}^{\text{EB}})^{\frac{1}{2}} \leq 2\alpha_n) = o(1),\ \text{as } n \to \infty. \tag{4.127}$$

Combining (4.126) and (4.127), we obtain (4.125) and the proof is completed. □

Proof of Theorem 4.3.12: To show that the decision rule (4.15) is an oracle, at first, we show the selection consistency part. Using similar arguments used before, again we have

$$P(\widehat{\mathcal{A}}_n^{(\text{FB})} \neq \mathcal{A}_n) \leq \sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) < \frac{1}{2}) + \sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) > \frac{1}{2}). \tag{4.128}$$

Here also, our target is to show the following:

$$\sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) < \frac{1}{2}) = o(1),\ \text{as } n \to \infty, \tag{4.129}$$

and

$$\sum_{g\notin\mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) > \frac{1}{2}) = o(1),\ \text{as } n \to \infty, \tag{4.130}$$

both when $0.5 \leq a < 1$, and $a \geq 1$.
In order to show (4.129), first note that with the use of D1,

$$E(\kappa_g \mid \sigma^2, \mathcal{D}) = \int_{\gamma_{1n}}^{\gamma_{2n}} E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})\pi(\tau_n|\mathcal{D})d\tau \leq E(\kappa_g \mid \gamma_{1n}, \sigma^2, \mathcal{D}).$$

Here, first, we use the fact that given τ_n, σ^2 and $\mathcal{D}$, the posterior mean of κ_g depends only on g^{th} group. Inequality in the last line follows since, for any fixed τ_n and σ^2, $E(\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$ is non-increasing in τ. As a result of this, we have

$$\sum_{g\in\mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) < \frac{1}{2}) \leq \sum_{g\in\mathcal{A}_n} P(E(\kappa_g \mid \gamma_{1n}, \sigma^2, \mathcal{D}) > \frac{1}{2}). \tag{4.131}$$

Next using the same steps as used in (4.58)-(4.59) with $\tau_n = \gamma_{1n}$ and noting that γ_{1n} satisfies $\log\left(\frac{1}{\tau_n}\right) \asymp \log(G_n)$, along with the use of (4.131), (4.129) is established.

In case of proving (4.130), note that

$$E(1-\kappa_g \mid \sigma^2, \mathcal{D}) = \int_{\gamma_{1n}}^{\gamma_{2n}} E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})\pi(\tau_n|\mathcal{D})d\tau$$
$$\leq E(1-\kappa_g \mid \gamma_{2n}, \sigma^2, \mathcal{D}).$$

Here, first, we use the fact that given τ_n, σ^2 and $\mathcal{D}$, the posterior mean of $1-\kappa_g$ depends only on g^{th} group. Inequality in the last line follows since, for any fixed τ_n and σ^2, $E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$ is non-decreasing in τ. As a consequence, it follows that

$$\sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \sigma^2, \mathcal{D}) > \frac{1}{2}) \leq \sum_{g \notin \mathcal{A}_n} P(E(1-\kappa_g \mid \gamma_{2n}, \sigma^2, \mathcal{D}) > \frac{1}{2})$$
$$\lesssim G_n \gamma_{2n} [\log\left(\frac{1}{\gamma_{2n}}\right)]^{\frac{s}{2}-1}(1+o(1)).$$

Inequality in the last line follows due to the use of arguments similar to those used in (4.60)-(4.64). Finally, under the assumption that $G_n\gamma_{2n}[\log\left(\frac{1}{\gamma_{2n}}\right)]^{\frac{s}{2}-1} \to 0$ as $n \to \infty$, (4.130) also holds and completes the proof related to variable selection consistency.
Next, we move forward to show the optimal estimation. Here we aim to show,

$$\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathcal{A}}^{\frac{1}{2}} (\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_{\mathcal{A},FB} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}}) \xrightarrow{P} 0 \text{ as } n \to \infty.$$

Now making use of the same arguments as used in (4.67)-(4.73) of Theorem 4.3.4, we only need to show that

$$\sum_{g \in \mathcal{A}_n} W_{n,g} E(\kappa_g^2 \mid \gamma_{1n}, \sigma^2, \mathcal{D}) \xrightarrow{P} 0 \text{ as } n \to \infty. \tag{4.132}$$

Next we follow the same steps as mentioned in (4.74)-(4.81) with $\tau = \gamma_{1n}$. Note that since γ_{1n} satisfies both (C3) and $\log\left(\frac{1}{\gamma_{1n}}\right) \asymp \log(G_n)$, it is immediate that (4.132) holds and as a result, $T_2 \xrightarrow{P} 0$ as $n \to \infty$. Finally, the use of Slutsky's Theorem completes the proof of Theorem 4.3.12. □

Proof of Theorem 4.4.1. The proof of (4.24) follows as a byproduct of the proof of (4.25). Hence, we now focus on proving (4.25). From the definition of the mean squared error of an estimator, one can write the mean squared error of the half-thresholding estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{HT}}_g$ as

$$E_{\boldsymbol{\beta}^0}\left\|\widehat{\boldsymbol{\beta}}^{\mathrm{HT}} - \boldsymbol{\beta}^0\right\|_2^2 = \sum_{g \in \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\right\|_2^2 + \sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\right\|_2^2$$
$$= \sum_{g \in \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\right\|_2^2 + \sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2, \tag{4.133}$$

where the last line follows from the fact $\boldsymbol{\beta}_g^0 = \mathbf{0}$ for each $g \notin \mathcal{A}_n$. Similarly, the mean squared

error of the posterior mean $\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}$ can be obtained as

$$E_{\boldsymbol{\beta}^0}\left\|\widehat{\boldsymbol{\beta}}^{\mathrm{PM}}-\boldsymbol{\beta}^0\right\|_2^2=\sum_{g\in\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}-\boldsymbol{\beta}_g^0\right\|_2^2+\sum_{g\notin\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2. \tag{4.134}$$

Hence, to establish part (ii) of Theorem 4.4.1, it would be enough to show that, for all sufficiently large n,

$$\sum_{g\notin\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2=\sum_{g\notin\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2+e_{1n},\ \text{with } e_{1n}<0, \tag{4.135}$$

and

$$\sum_{g\in\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}-\boldsymbol{\beta}_g^0\right\|_2^2=\sum_{g\in\mathcal{A}_n}E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}-\boldsymbol{\beta}_g^0\right\|_2^2+e_{2n},\ \text{where } e_{2n}=o(e_{1n}). \tag{4.136}$$

Proof of (4.135): From the definition of the half-thresholding estimate, $\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}$ given in (4.10), it follows that,

$$\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2=\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\big\{E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\le 0.5\big\}. \tag{4.137}$$

Hence, for any g, $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2\le\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2$. As a result, for obtaining (4.135), it is enough to establish that for any $g\notin\mathcal{A}_n$, $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2>0$ with some positive probability. Therefore, we are interested in exploring the behavior of $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\big\{E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\le 0.5\big\}$ for $g\notin\mathcal{A}_n$. Observe that, from Proposition 4.2.2, it follows that

$$E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\xrightarrow{P}0\ \text{as } n\to\infty,\ \text{uniformly in } g\notin\mathcal{A}_n.$$

This implies that, for any $g\notin\mathcal{A}_n$

$$I\big\{E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\le 0.5\big\}=1\ \text{on the set } B_{1n}\ \text{and}\ \lim_{n\to\infty}P_{\boldsymbol{\beta}_g^0=0}(B_{1n})=1, \tag{4.138}$$

where $B_{1n}=\big\{E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\le 0.5\big\}$. Again, for any $g\notin\mathcal{A}_n$, using the definition of $\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}$,

$$\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2>0\ \text{on the set } B_{2n}\ \text{and}\ P_{\boldsymbol{\beta}_g^0=0}(B_{2n})=1,\ \text{for all } n, \tag{4.139}$$

where $B_{2n}=\left\{\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2>0\right\}$. Hence, combining (4.137)- (4.139) along with the definitions of B_{1n} and B_{2n}, we have $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\big\{E(1-\kappa_g\mid\tau_n,\sigma^2,\mathcal{D})\le 0.5\big\}>0$ on the set $B_{1n}\cap B_{2n}$, with

$$P_{\boldsymbol{\beta}_g^0=0}(B_{1n}\cap B_{2n})\ge P_{\boldsymbol{\beta}_g^0=0}(B_{1n})+P_{\boldsymbol{\beta}_g^0=0}(B_{2n})-1\to 1\ \text{as } n\to\infty. \tag{4.140}$$

Hence, for all sufficiently large n, uniformly in $g\notin\mathcal{A}_n$, $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2-\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\right\|_2^2>0$ on the set $B_{1n}\cap B_{2n}$

with $P_{\boldsymbol{\beta}_g^0=0}(B_{1n} \cap B_{2n}) \to 1$. This along with (4.137) implies, $\sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0=0}\Big[\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2 I\big\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5\big\}\Big] > 0$ for sufficiently large n. Let us define,

$$
\begin{aligned}
e_{1n} &= \sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\big\|_2^2 - \sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2 \\
&= -\sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0=0}\Big[\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2 I\big\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5\big\}\Big].
\end{aligned}
$$

Then, using our previous arguments, it follows that $e_{1n} < 0$ for sufficiently large n. Therefore, for all sufficiently large n, we have

$$
\sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}\big\|_2^2 = \sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2 + e_{1n}, \text{ with } e_{1n} < 0,
$$

which completes the proof of (4.135). It is important to observe that, when all groups are inactive, $\mathcal{A}_n = \emptyset$. In this case, the first terms in both (4.133) and (4.134) will disappear. Consequently, (4.24) trivially follows from here. This concludes the proof of part (i) of the present theorem. Now we prove (4.136).

Proof of (4.136): Note that, for any $g \in \mathcal{A}_n$

$$
\begin{aligned}
E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 &= E_{\boldsymbol{\beta}_g^0}\Big[\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\}\Big] + E_{\boldsymbol{\beta}_g^0}\Big[\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} \neq \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\}\Big] \\
&= E_{\boldsymbol{\beta}_g^0}\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \boldsymbol{\beta}_g^0\big\|_2^2 + E_{\boldsymbol{\beta}_g^0}\Big[\Big(\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 - \big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \boldsymbol{\beta}_g^0\big\|_2^2\Big) 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} \neq \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\}\Big].
\end{aligned} \tag{4.141}
$$

Next, observe that, using the definition of $\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}$, we have $1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} \neq \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\} = 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}$. Using this fact, the second term in the right hand side of (4.141) is of the form

$$
\begin{aligned}
&E_{\boldsymbol{\beta}_g^0}\Big[\Big(\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 - \big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \boldsymbol{\beta}_g^0\big\|_2^2\Big) 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} \neq \widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\}\Big] \\
&= E_{\boldsymbol{\beta}_g^0}\Big[\Big(\big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\big\|_2^2 - \big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \boldsymbol{\beta}_g^0\big\|_2^2\Big) 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}\Big] \\
&= E_{\boldsymbol{\beta}_g^0}\Big[\Big(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2\Big) 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}\Big].
\end{aligned} \tag{4.142}
$$

Hence, for proving (4.136), first we need to obtain an upper bound of $\sum_{g \in \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\Big[\Big(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \big\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\big\|_2^2\Big) 1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}$.

Note that,

$$E_{\boldsymbol{\beta}_g^0}\left(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right)^2 \leq 2\left[E_{\boldsymbol{\beta}_g^0}\left(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right)^2 + E_{\boldsymbol{\beta}_g^0}\left(\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right)^2\right]$$
$$\leq 8(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\boldsymbol{\beta}_g^0 E_{\boldsymbol{\beta}_g^0}\left(\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right) + 2E_{\boldsymbol{\beta}_g^0}\left(\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right)^2$$
$$\leq 8(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\boldsymbol{\beta}_g^0 E_{\boldsymbol{\beta}_g^0}\left(\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2\right) + 2E_{\boldsymbol{\beta}_g^0}\left(\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2\right)^2. \tag{4.143}$$

In the above chain of inequalities, the first inequality holds because of the fact $(a-b)^2 \leq 2(a^2+b^2)$, for any two real numbers a and b. The second inequality here follows from the Cauchy-Schwarz inequality, while the last one follows from the definition of $\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}$, and the fact $\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 \leq \left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2$, for any arbitrary group g.
Since, $\widehat{\boldsymbol{\beta}}_g \sim \mathcal{N}_{m_g}(\boldsymbol{\beta}_g, \sigma^2(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1})$, using standard theory of multivariate normal distributions, it follows that

$$E\left(\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2\right) = \|\boldsymbol{\beta}_g^0\|_2^2 + \sigma^2 Tr((\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1})$$
$$= \|\boldsymbol{\beta}_g^0\|_2^2 + \frac{\sigma^2}{n}\sum_{j=1}^{m_g}\eta_j \leq \|\boldsymbol{\beta}_g^0\|_2^2 + \frac{\sigma^2 C_3 s}{n}, \tag{4.144}$$

where η_j is the j^{th} diagonal element of $n(\mathbf{X}_g^{\mathrm{T}}\mathbf{X}_g)^{-1}$. Note that, (C1) implies η_j is bounded for all $j = 1, \ldots, m_g$. Also, observe that, $E((\sum_{j=1}^{m_g}\widehat{\beta}_{gj})^2)^2 \leq m_g \sum_{j=1}^{m_g} E((\widehat{\beta}_{gj})^4)$. As a result,

$$E((\widehat{\beta}_{gj})^4) = (\beta_{gj}^0)^4 + 6\sigma^2\eta_j\frac{(\beta_{gj}^0)^2}{n} + 3\frac{\sigma^4}{n^2}(\eta_j)^2 \leq (\beta_{gj}^0)^4 + 6\frac{\sigma^2 C_3}{n}(\beta_{gj}^0)^2 + 3\frac{\sigma^4 C_3}{n^2}. \tag{4.145}$$

Combining (4.143)-(4.145), we obtain, for all $g \in \mathcal{A}_n$

$$E_{\boldsymbol{\beta}_g^0}\left(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right)^2 \leq C_5 \max_{j=1,\ldots,m_g}(\beta_{gj}^0)^4, \tag{4.146}$$

for some arbitrary constant C_5, independent of both g and j. On the other hand, using the definition of $\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}}$ followed by (4.44) and (4.49), we have

$$E_{\boldsymbol{\beta}_g^0}[1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}] = P_{\boldsymbol{\beta}_g^0}(E(1-\kappa_g|\tau,\sigma^2,\mathcal{D}) \leq \frac{1}{2})$$
$$\leq \frac{C_7}{\sqrt{n}m_n}\exp\left(-C_8 n m_n^2\right)(1+o(1)), \tag{4.147}$$

where $o(1)$ is independent of any $g \in \mathcal{A}_n$ and C_7 and C_8 are some constants independent of any $g \in \mathcal{A}_n$. Hence, using (4.146), (4.147) and by Cauchy-Schwarz inequality along with (A2), for

all sufficiently large n,

$$\begin{aligned}
&\sum_{g\in\mathcal{A}_n} E_{\boldsymbol{\beta}_g^0}\left[\left(2(\boldsymbol{\beta}_g^0)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2\right)1\{\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} = \mathbf{0}\}\right] \\
&\lesssim \sum_{g\in\mathcal{A}_n} \max_{j=1,\ldots,m_g} (\beta_{gj}^0)^2 n^{\frac{b}{2}-\frac{1}{4}} \exp\left(-0.5C_8 n^{1-2b}\right)(1+o(1)) \\
&\lesssim n^{\tilde{C}_1}\exp\left(\tilde{C}_2 \times n^{\tilde{C}_3}\right)(1+o(1)),
\end{aligned} \tag{4.148}$$

for some $\tilde{C}_i > 0$, $i = 1, 2, 3$. Here the last line holds since for all $g \in \mathcal{A}_n, \max_{j=1,\ldots,m_g} \beta_{gj}^0 \le n^{C_3}$, for some $C_3 > 0$ and $|\mathcal{A}_n| \le n$. Let us define $e_{2n} = E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{HT}} - \boldsymbol{\beta}_g^0\right\|_2^2 - E_{\boldsymbol{\beta}_g^0}\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}} - \boldsymbol{\beta}_g^0\right\|_2^2$. Next using (4.141)-(4.148), we have, for all sufficiently large n, $e_{2n} \lesssim n^{\tilde{C}_1}\exp\left(\tilde{C}_2 \times n^{\tilde{C}_3}\right)(1+o(1))$. Hence, we are left to prove that, $e_{2n} = o(e_{1n})$, as $n \to \infty$. For this, we need an upper bound on e_{1n}.

Towards this, first observe that,

$$\begin{aligned}
&E_{\boldsymbol{\beta}_g^0=0}\left[\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5\right\}\right] \\
&\ge E_{\boldsymbol{\beta}_g^0=0}\left[\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5\right\} I\left\{\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2 \ge \frac{1}{n^2}\right\}\right] \\
&\ge \frac{1}{n^2} E_{\boldsymbol{\beta}_g^0=0}\left[(E(1-\kappa_g|\tau_n, \sigma^2, \mathcal{D}))^2 I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5 \cap \left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2 \ge \frac{1}{n^2}\right\}\right] \\
&\ge n^{-(4a+4)} E_{\boldsymbol{\beta}_g^0=0}\left[I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \le 0.5 \cap \left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2 \ge \frac{1}{n^2}\right\} I\left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \ge \frac{1}{n^{(2a+1)}}\right\}\right] \\
&\ge \frac{1}{n^{(4a+4)}} P_{\boldsymbol{\beta}_g^0=0}(B_{1n} \cap B_{3n} \cap B_{4n}) \ge \frac{1}{n^{(4a+4)}}[P_{\boldsymbol{\beta}_g^0=0}(B_{1n}) + P_{\boldsymbol{\beta}_g^0=0}(B_{3n}) + P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) - 2],
\end{aligned} \tag{4.149}$$

where B_{1n} is the same as defined earlier, $B_{3n} = \left\{\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2 \ge \frac{1}{n^2}\right\}$ and $B_{4n} = \left\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \ge \frac{1}{n^{(2a+1)}}\right\}$. Here in the chain of inequalities, the second one holds due to the definition of the posterior mean along with the lower bound on the least square estimate. The fourth one follows due to Markov's inequality on the event B_{4n}. The last one follows from Boole's inequality. Recall that, Proposition 4.2.2 implies, $\lim_{n\to\infty} P_{\boldsymbol{\beta}_g^0=0}(B_{1n}) = 1$. Hence, our next target is to show that, $\lim_{n\to\infty} P_{\boldsymbol{\beta}_g^0=0}(B_{3n}) = 1$ and $\lim_{n\to\infty} P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) = 1$. Towards that, note, under (C1), we have

$$\frac{W_{n,g}}{C_2} \le \frac{n\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2}{\sigma^2} \le \frac{W_{n,g}}{C_1}.$$

Hence, for $W_{n,g} \sim \chi^2_{m_g}$,

$$P_{\boldsymbol{\beta}_g^0=0}(B_{3n}) = P_{\boldsymbol{\beta}_g^0=0}\left(\frac{n\left\|\widehat{\boldsymbol{\beta}}_g\right\|_2^2}{\sigma^2} > \frac{1}{n\sigma^2}\right) \geq P_{\boldsymbol{\beta}_g^0=0}\left(\frac{W_{n,g}}{C_2} > \frac{1}{n\sigma^2}\right)$$
$$\to 1 \text{ as } n \to \infty. \tag{4.150}$$

Hence, it is only left to show that, $\lim_{n\to\infty} P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) = 1$. For this, we need to obtain a lower bound on $E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D})$. We consider two cases $a \in [\frac{1}{2}, 1)$ and $a \geq 1$, separately.

Case 1: $a \in [\frac{1}{2}, 1)$**.** Using the definition of the posterior distribution of κ_g given $\tau, \sigma^2, \mathcal{D}$, we have

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) = \frac{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)\exp\left\{-\kappa_g \cdot \frac{W_{n,g}}{2}\right\}d\kappa_g}{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)\exp\left\{-\kappa_g \cdot \frac{W_{n,g}}{2}\right\}d\kappa_g}$$
$$\geq \exp\left(-\frac{W_{n,g}}{2}\right)\frac{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)d\kappa_g}{\int_0^1 \kappa_g^{a+\frac{m_g}{2}-1}(1-\kappa_g)^{-a-1} L\left(\frac{1}{\tau_n^2}(\frac{1}{\kappa_g}-1)\right)d\kappa_g}, \tag{4.151}$$

where the inequality holds using the monotonicity of the function $\kappa \mapsto \exp\big(-\kappa x^2/2\big), 0 < \kappa < 1$. Next, using the transformation $t = \frac{1}{\tau_n^2}(\frac{1}{\kappa_g} - 1)$ in the integrals above, we obtain

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \geq \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \frac{\int_0^\infty (1+t\tau_n^2)^{-\frac{m_g}{2}-1} t^{-a} L(t)dt}{\int_0^\infty (1+t\tau_n^2)^{-\frac{m_g}{2}} t^{-a-1} L(t)dt}$$
$$= K\exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_0^\infty (1+t\tau_n^2)^{-\frac{m_g}{2}-1} t^{-a} L(t)dt(1+o(1))$$
$$\geq K\exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_{t_0}^{\frac{t_0}{\tau_n^2}} (1+t\tau_n^2)^{-\frac{m_g}{2}-1} t^{-a} L(t)dt(1+o(1))$$
$$\geq K\exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 (1+t_0)^{-\frac{m_g}{2}-1} \int_{t_0}^{\frac{t_0}{\tau_n^2}} t^{-a} L(t)dt(1+o(1))$$
$$\geq \frac{K}{(1-a)}\exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 (1+t_0)^{-\frac{s}{2}-1} L(\frac{t_0}{\tau_n^2})(\frac{t_0}{\tau_n^2})^{(1-a)}(1+o(1))$$
$$\geq K_1\exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^{2a}(1+o(1)), \tag{4.152}$$

where K_1 is a constant depending on t_0, a, s. Here, the equality in the second line holds due to using the Dominated Convergence Theorem on the denominator. The inequality in the fourth line holds because of the monotonicity of the function $t \mapsto (1+t)^{-\frac{m_g}{2}-1}$ in the interval $[t_0, \frac{t_0}{\tau_n^2}]$. (A2) followed by (5) of Lemma 4.8.1 ensures the inequality in the fifth line. The final one holds due to Assumption 1 on $L(\cdot)$.

Case 2: $a \geq 1$**.** In this case, too, following arguments same as those of (4.151) and (4.152), we

have

$$\begin{aligned} E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) &\geq K \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_0^{\infty} (1+t\tau_n^2)^{-\frac{m_g}{2}-1} t^{-a} L(t)dt(1+o(1)) \\ &\geq K \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_1^{\frac{1}{\tau_n^2}} (1+t\tau_n^2)^{-\frac{m_g}{2}-1} t^{-a} L(t)dt(1+o(1)) \\ &\geq 2^{-\frac{m_g}{2}-1} K \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_1^{\frac{1}{\tau_n^2}} t^{-a} L(t)dt(1+o(1)) \\ &\geq 2^{-\frac{s}{2}-1} K c_0 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \int_1^{\frac{1}{\tau_n^2}} t^{-a} dt(1+o(1)). \end{aligned} \tag{4.153}$$

Since, for $a = 1$, $\int_1^{\frac{1}{\tau_n^2}} t^{-a}dt = \log\left(\frac{1}{\tau_n^2}\right)$ and for $a > 1$, $\int_1^{\frac{1}{\tau_n^2}} t^{-a}dt = \frac{1}{(a-1)}[1-\tau_n^{2(a-1)}]$, using (4.153), we obtain

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \geq \begin{cases} K_2 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \log\left(\frac{1}{\tau_n^2}\right)(1+o(1)), a = 1, \\ K_3 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 (1+o(1)), a > 1, \end{cases} \tag{4.154}$$

where K_2 and K_3 are constants depending on t_0, a, s. Combining (4.152) and (4.154), for $a \geq \frac{1}{2}$,

$$E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \geq \begin{cases} K_1 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^{2a}(1+o(1)), 0 < a < 1, \\ K_2 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 \log\left(\frac{1}{\tau_n^2}\right)(1+o(1)), a = 1, \\ K_3 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^2 (1+o(1)), a > 1. \end{cases} \tag{4.155}$$

Hence, we need to show $\lim_{n\to\infty} P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) = 1$, for $a \in [\frac{1}{2}, 1)$ and $a \geq 1$. When $\frac{1}{2} \leq a < 1$,

$$\begin{aligned} P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) &= P_{\boldsymbol{\beta}_g^0=0}\left(E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \geq \frac{1}{n^{(2a+1)}}\right) \\ &\geq P_{\boldsymbol{\beta}_g^0=0}\left(K_1 \exp\left(-\frac{W_{n,g}}{2}\right)\tau_n^{2a}(1+o(1)) \geq \frac{1}{n^{(2a+1)}}\right) \\ &= P_{\boldsymbol{\beta}_g^0=0}\left(\log K_1 - \frac{W_{n,g}}{2} - 2a\log\left(\frac{1}{\tau_n}\right) + o(1) \geq -(2a+1)\log n\right) \\ &= P_{\boldsymbol{\beta}_g^0=0}\left(\frac{W_{n,g}}{2} \leq (2a+1)\log n - 2a\log\left(\frac{1}{\tau_n}\right) + o(1) + \log K_1\right). \end{aligned} \tag{4.156}$$

Observe that, under (A4), for sufficiently large n,

$$(2a+1)\log n - 2a\log\left(\frac{1}{\tau_n}\right) \geq (2a+1)\log n - 2a\log n = \log n. \tag{4.157}$$

Therefore, combining (4.156) and (4.157), for $a \in (0,1)$,

$$P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) \geq P_{\boldsymbol{\beta}_g^0=0}\Big(W_{n,g} \leq 2\log n + 2\log K_1 + o(1)\Big) \to 1, \text{ as } n \to \infty. \tag{4.158}$$

For $a \geq 1$, using the lower bound obtained in (4.155) and following similar arguments of (4.156) and (4.157), again we can show that,

$$P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) \to 1, \text{ as } n \to \infty. \tag{4.159}$$

Combining (4.158) and (4.159), for $a > 0$, we obtain

$$P_{\boldsymbol{\beta}_g^0=0}(B_{4n}) \to 1, \text{ as } n \to \infty. \tag{4.160}$$

Combining (4.149), (4.150) and (4.160), we have, for sufficiently large n,

$$e_{1n} = -\sum_{g \notin \mathcal{A}_n} E_{\boldsymbol{\beta}_g^0=0}\left[\left\|\widehat{\boldsymbol{\beta}}_g^{\mathrm{PM}}\right\|_2^2 I\big\{E(1-\kappa_g \mid \tau_n, \sigma^2, \mathcal{D}) \leq 0.5\big\}\right] \lesssim -n^{-(4a+3)}(1+o(1)).$$

This ensures $e_{2n} = o(e_{1n})$, as $n \to \infty$. As a result, (4.136) is established. The proof of Theorem 4.4.1 is completed using (4.135) and (4.136). □

Chapter 5

Asymptotic Bayes Optimality for Sparse Count Data

5.1 Introduction

In this era of enormous supply of high-dimensional data, multiple hypothesis testing has been a growing area of research in statistics. Asymptotic theory of multiple testing in the context of normal models has received serious attention as described earlier in this thesis. But the normality assumption is not valid in many situations and there is a need to develop methods and build corresponding theories of statistical inference for such data. Our work in this chapter is a modest step in that direction. In this chapter, our interest is in modelling and inference on high-dimensional "quasi-sparse" count data. We often come across count data where most of the counts are near zero with few of them being really large or moderate. This would possibly indicate that the rates of occurrences corresponding to most observations are very small while for the rest they are possibly moderate or large. Such data are referred to as quasi-sparse count data. In this situation, detecting the large signals out of a pool of noises/small signals is the problem of interest. One such situation, inspired by Datta and Dunson (2016), involving modelling of count data was described in Subsection 1.2.4. This paper is the main motivation of our work in this chapter. Datta and Dunson (2016) consider an observation vector $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$, where each Y_i is modeled as, $Y_i \sim Poi(\theta_i)$. Here $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ is the parameter of interest. Recall from Subsection 1.2.4 that Datta and Dunson (2016) considered the following hierarchical formulation (for their theoretical analysis):

$$\theta_i|\lambda_i, \tau \overset{ind}{\sim} Ga(\alpha, \lambda_i^2\tau^2), \lambda_i^2 \overset{ind}{\sim} \pi_1(\lambda_i^2), \tag{5.1}$$

where $\pi_1(\cdot)$ is the density for λ_i^2. They either used τ as fixed or let $\tau \to 0$ as $n \to \infty$ for proving their results. Integrating out θ_i and defining $\kappa_i = 1/(1+\lambda_i^2\tau^2)$, the marginal distribution of Y_i given κ_i is obtained as

$$f(Y_i|\kappa_i) \propto (1-\kappa_i)^{Y_i}\kappa_i^{\alpha},$$

i.e., the marginal distribution of each Y_i follows a negative binomial distribution with size α and probability of success $1-\kappa_i$. They assumed Gauss Hypergeometric (GH) prior on κ_i as in (1.30) (with $a_1 = a_2 = \frac{1}{2}$), which was originally introduced by Armero and Bayarri (1994) in the context of a specific queuing problem.

In the Bayesian paradigm, the usual approach to model a sparse vector $\boldsymbol{\theta}$ with mostly small values of coordinates is through a two-group prior like a spike-and-slab prior with a distribution concentrated around 0 with (high) probability $1-p$ and a heavy-tailed distribution for capturing the large values with probability p. Datta and Dunson (2016) in fact modeled each θ_i as a scale mixture of two Gamma distributions in the following way:

$$\theta_i \overset{iid}{\sim} (1-p)Ga(\alpha,\beta) + pGa(\alpha,\beta+\delta), i = 1,2,\cdots,n, \tag{5.2}$$

where the choices of α, β and δ are as described in Subsection 1.2.4. As a result, the marginal distribution of Y_i is obtained as a mixture of two negative binomial distributions having the form:

$$Y_i \overset{iid}{\sim} (1-p)NB(\alpha, \frac{1}{\beta+1}) + pNB(\alpha, \frac{1}{\beta+\delta+1}), i = 1,2,\cdots,n. \tag{5.3}$$

As indicated before, by introducing a set of latent variables $\nu_i, i = 1,2,\cdots,n$, such that $\nu_i = 0$ indicates $H_{0i} : \theta_i \sim Ga(\alpha,\beta)$ and $\nu_i = 1$ indicates $H_{1i} : \theta_i \sim Ga(\alpha,\beta+\delta)$ with $P(\nu_i = 1) = p$ and $P(\nu_i = 0) = 1-p$, the marginal prior on θ_i becomes of the form (5.2). In a simulation study, Datta and Dunson (2016) showed that the performance of their proposed decision rule (described in Subsection 1.2.4) using GH prior with $\gamma \neq 1$ is better than that of the horseshoe prior ($\gamma = 1$) for simultaneous testing of H_{0i} vs H_{1i} for $i = 1,2,\cdots,n$, in terms of the misclassification probability when the data is generated from a two-group mixture. They also showed that average mean squared error of the estimate of θ_i's based on the GH prior with $\gamma \neq 1$ is smaller than that using the horseshoe estimator ($\gamma = 1$). They also stated an upper bound on the type I error of their decision rule when θ_i's are truly generated using (5.2). However expression for the type II error and study of the optimal Bayes risk of multiple testing rules within any decision theoretic framework were missing in their work.

Inspired by the results of Carvalho et al. (2009), Carvalho et al. (2010), Ghosh et al. (2016) and Datta and Dunson (2016), we have considered in this chapter a broad class of global-local shrinkage priors of the form (5.1) with

$$\pi_1(\lambda_i^2) = K(\lambda_i^2)^{-a-1}L(\lambda_i^2), \tag{5.4}$$

where $K, a \geq 1, L(\cdot)$ are the same as defined in (1.10) of Subsection 1.1.3. Similar to Ghosh et al. (2016), we are interested in studying whether any decision rule based on such one-group priors can attain the optimal Bayes risk, asymptotically, with respect to additive symmetric $0-1$ loss, when the data is truly generated from a two-group mixture.

We now briefly highlight the main contents and our contributions in this chapter. We start with our theoretical results. In Section 5.2, under the additive symmetric $0-1$ loss function, we have obtained an expression for the Bayes risk (in Theorem 5.4.1) corresponding to the Bayes oracle rule induced by the two-group prior given in (5.2) under some assumptions on the model

parameters p, β and δ. In Section 5.3, as an alternative to the two-group approach, we propose a multiple testing rule based on our chosen class of one-group priors to be used for simultaneous testing when the true θ_i's are generated from a two-group model. The rule is similar to that in Datta and Dunson (2016) and is defined in (5.15). Corresponding to our class of priors, we provide some posterior concentration inequalities for κ_i in Subsection 5.4.2 (see Theorems 5.4.3-5.4.6). The conclusions about concentration and shrinkage properties obtained in these results are similar to those obtained by Ghosh et al. (2016) for the sparse normal means model and Datta and Dunson (2016) based on Gauss Hypergeometric prior for quasi-sparse count data. Based on the concentration inequalities, in Subsection 5.4.3, we obtain non-trivial upper bounds for both type I and type II error probabilities of this rule. These are reported in Theorems 5.4.10 and 5.4.12 respectively. Although Datta and Dunson (2016) claimed an upper bound for the type I error corresponding to the Gauss Hypergeometric prior, a careful inspection reveals that there is a soft spot in the argument. Using the line of proof of Theorem 5.4.10, it is possible that their argument can be made rigorous and also the statement of their result can be made more precise, but we have not formally pursued it in this thesis. Finally, when the theoretical proportion of non-nulls (p) is known, we prove that the Bayes risk corresponding to the decision rule (5.15) based on our proposed class of priors attains the optimal Bayes risk, up to some multiplicative constant. This is the first main contribution of this work and is stated in Theorem 5.4.1. When the theoretical proportion of non-nulls is unknown, based on an empirical Bayes estimate $\widehat{\tau}$ of τ as proposed by Yano et al. (2021), in Subsection 5.4.4, we provide bounds on the probabilities of type I and type II errors of our rule where τ is replaced by $\widehat{\tau}$. These are presented in Theorems 5.4.14 and 5.4.15 respectively. We are able to show that the Bayes risk of this empirical Bayes decision rule also attains the optimal Bayes risk, up to some multiplicative constant. This is formally stated in Theorem 5.4.2. These results are the first of their kind in the literature.

We have presented the results of a simulation study in Section 5.5 and a real data analysis in Section 5.6. Our results lend strong support to our theoretical findings. In Section 5.7, we present an overall discussion along with some possible extensions of this work. Section 5.8 contains the proofs of all theorems and other theoretical results.

This chapter is based on Paul and Chakrabarti (2024).

5.2 The Two-group prior and the Bayes Oracle

To capture the sparsity of the $\boldsymbol{\theta}$ in the context of our problem, the popular approach would be to use a two-component mixture prior where with probability p, θ_i is degenerate at zero and with probability $(1-p)$ it is generated either from a distribution degenerate at some non-zero value or a Gamma distribution. See Datta and Dunson (2016) and the references therein in this context. But as commented by them, such mixture modelling may cause computational instability. Inspired by the work of Datta and Dunson (2016), we model each θ_i as a sample from a scale mixture of two gamma distributions of the form (5.2). As a result of this, the marginal distribution of the observations is a mixture of two negative binomial distributions as given in (5.3).

Now we consider the problem of multiple hypothesis testing of $H_{0i} : \nu_i = 0$ against $H_{1i} : \nu_i = 1$, for $i = 1, 2, \cdots, n$ within a Bayesian decision theoretic framework. We consider the usual $0-1$ loss function for individual tests and assume that the overall loss of a multiple testing procedure is the sum of losses corresponding to individual tests. In this way, our approach is based on an additive loss function and is similar to that of Bogdan et al. (2011) for the normal means model. Additive losses were earlier considered by Lehmann (1957) in the context of testing.

Let t_{1i} and t_{2i} denote the probabilities of type I and type II error of the i^{th} test respectively, and are defined as

$$t_{1i} = P_{H_{0i}}(H_{0i} \text{ is rejected})$$

and

$$t_{2i} = P_{H_{1i}}(H_{0i} \text{ is accepted}).$$

Then the Bayes risk R of a multiple testing procedure is obtained as

$$R = \sum_{i=1}^{n} [(1-p)t_{1i} + pt_{2i}]. \tag{5.5}$$

Recalling Bogdan et al. (2011), it is easy to see that a multiple testing rule minimizing the Bayes risk R is the one that applies the simple Bayes classifier for each individual test. So the optimal rule, for each i, rejects H_{0i} in favor of H_{1i} if

$$\frac{f_{H_{1i}}(Y_i)}{f_{H_{0i}}(Y_i)} > \frac{1-p}{p}, \tag{5.6}$$

where $f_{H_{1i}}$ and $f_{H_{0i}}$ are the marginal densities of Y_i under the alternative and null hypotheses respectively. Using (5.3), the optimal testing rule given in (5.6) is simplified as

$$\text{reject } H_{0i} \text{ if} \quad Y_i > C, \tag{5.7}$$

where

$$C = C_{p,\alpha,\beta,\delta} = \frac{\log\left(\frac{1-p}{p}\right) + \alpha \log\left(\frac{\beta+\delta+1}{\beta+1}\right)}{\log\left[\frac{(\beta+\delta)(\beta+1)}{\beta(\beta+\delta+1)}\right]}. \tag{5.8}$$

Due to the presence of p, α, β and δ, the decision rule (5.7) is termed as *Bayes Oracle*. Its performance can not be equalled by any testing rule based on finite samples if these parameters are unknown. Let us denote the Bayes risk of the Bayes Oracle as $R^{\text{BO}}_{\text{Opt}}$.

Note that, under the mixture model (5.3), the marginal distributions of Y_i under both null and alternative hypotheses are independent of the choice of i. The threshold of the Bayes Oracle, given in (5.8), corresponding to a two-group prior is also independent of i. Therefore, we drop i from t_{1i} and t_{2i} and rename those as t_1^{BO} and t_2^{BO}, respectively. Therefore, we can write

$$R^{\text{BO}}_{\text{Opt}} = n[pt_1^{\text{BO}} + (1-p)t_2^{\text{BO}}]. \tag{5.9}$$

As in Bogdan et al. (2011), we want to study the behaviour of $R^{\text{BO}}_{\text{Opt}}$ under a suitable asymptotic setting. Since our basic model is non-normal, we cannot simply mimic the setting of Bogdan

et al. (2011) as in their Assumption (A) described formally in Chapter 2. Before writing out the asymptotic setting, we want to motivate the logic behind it. Note that here we want to model quasi-sparse count data with a mixture of $Ga(\alpha, \beta)$ and $Ga(\alpha, \beta + \delta)$ prior. In order to ensure that under the null the θ_i's are concentrated near zero, we let $\beta \to 0$ and keep α fixed at a small value. In order to accommodate possible large values under the alternative, we further assume that δ is fixed value bounded away from zero, which ensures that the prior mean and the variance are larger by a factor of magnitude under the alternative compared to the null. Finally, sparsity dictates that we assume $p \to 0$ as $n \to \infty$. Calculations of Theorem 5.4.1 reveal that under these assumptions, the error bounds corresponding to the Bayes oracle rule (5.7) (with C as in (5.8)) are of the form,

$$t_1^{\text{BO}} \leq \mathbb{P}_{H_{0i}}\left(Y_i > \frac{\log\left(\frac{1}{p}\right) + \alpha \log(1+\delta)}{\log\left(\frac{1}{\beta}\right)}(1 + o(1))\right)$$

and

$$t_2^{\text{BO}} = P_{H_{1i}}\left(Y_i \leq \frac{\log\left(\frac{1}{p}\right) + \alpha \log(1+\delta)}{\log\left(\frac{1}{\beta}\right)}(1 + o(1))\right),$$

where the $o(1)$ term is non-random, independent of index i and goes to 0 as $n \to \infty$. If we further assume that $\beta \propto p^{C_1}$ for some $C_1 > 1$, then the type II error, t_2^{BO} and eventually the power of the Bayes Oracle rule is asymptotically strictly between 0 and 1. In this chapter, we are interested in these situations, since otherwise, even the Bayes Oracle rule cannot guarantee non-trivial inference. We now state our asymptotic setting in Assumption 3 below. Our asymptotic analysis of the Bayes Oracle (5.7) and our proposed rule are done under Assumption 3. This is discussed in more details later in this chapter.

Assumption 3. $p \to 0$, α and δ are fixed and bounded away from both 0 and infinity, $\beta \propto p^{C_1}$ for some $C_1 > 1$.

Under Assumption 3, we have,

$$t_2^{\text{BO}} = (\beta + \delta + 1)^{-\alpha} \text{ and } t_1^{\text{BO}} = o(p) \text{ as } n \to \infty.$$

and $t_1^{\text{BO}} = o(p)$ as $n \to \infty$. Combining these two, we have

$$R_{\text{Opt}}^{\text{BO}} = np(\delta + 1)^{-\alpha}(1 + o_1(1)), \tag{5.10}$$

where $o_1(1)$ is non-random, independent of index i and goes to 0 as $n \to \infty$. Detailed calculations establishing the above facts on the error probabilities are available in the proof of Theorem 5.4.1 in Section 5.8. Our next target is to find expressions/upper bounds for t_{1i} and t_{2i} corresponding to our decision rules (5.15) and (5.17) based on priors (5.1) satisfying (5.4). This will help to study the optimality of our decision rules. In the next subsection, we motivate our decision rules by highlighting a connection between the posterior means of θ_i under two-group priors and our class of priors.

5.3 Multiple testing rules using one-group priors

In this section, we propose simultaneous hypothesis testing rules based on a broad class of global-local shrinkage priors to be applied in the context of quasi-sparse Poisson count data described before. We want to investigate in the following sections how they perform vis-a-vis the optimal rule, namely the Bayes Oracle when the means are actually generated from a two-group prior. We first briefly recall that Datta and Dunson (2016) had also applied a decision rule in this context using a subclass of the Gauss Hypergeometric (GH) priors described earlier. But they did not study its optimality properties in a decision theoretic framework. Our main interest lies in such a study using one-group priors in the context of count data. Towards that we consider a broad class of global-local shrinkage priors satisfying (5.1) and (5.4) described in Subsection 5.1. Incidentally, it may be noted that the a large spectrum of the Three Parameter Beta Normal (TPBN) priors (with $\alpha > 0$ and $\beta \geq 1$) and Generalized Double Pareto priors (with $\alpha \geq 2$ and $\beta > 0$) are contained in our general class. The forms of these two classes of priors are discussed in equations (12) and (13) of Ghosh et al. (2016). The GH priors defined in Subsection 1.2.4 contain the TPBN family. But Datta and Dunson (2016) only considered a subclass of that is mentioned before.

We now introduce our decision rule based on our chosen class of priors. It may first be observed that the posterior distribution of θ_i's given $(Y_1, \cdots, Y_n, \tau)$ are independent and the posterior distribution of θ_i depends only on Y_i, κ_i and τ. First note that using such priors, the posterior distribution of θ_i given κ_i, τ and Y_i is of the form

$$\theta_i | Y_i, \kappa_i, \tau \sim \text{Ga}(Y_i + \alpha, 1 - \kappa_i)\,. \tag{5.11}$$

By Fubini's theorem, the posterior mean of θ_i is therefore given by

$$E(\theta_i | Y_i, \tau) = (1 - E(\kappa_i | Y_i, \tau))(Y_i + \alpha)\,. \tag{5.12}$$

Under a two-group mixture prior, the posterior distribution of θ_i's given $\mathbf{Y} = (Y_1, \cdots, Y_n)$ are independent and, in fact, the posterior distribution of θ_i only depends on Y_i. When θ_i is generated from the two-group prior (5.2), one has

$$\mathbb{E}(\theta_i | Y_i) = [(1 - w_i)\frac{\beta}{\beta + 1} + w_i \frac{\beta + \delta}{\beta + \delta + 1}](Y_i + \alpha) = w_i^*(Y_i + \alpha)\,, \tag{5.13}$$

where $w_i = \mathbb{P}(H_{1i} | Y_i)$ denotes the posterior probability of H_{1i} being true and w_i^* is the observation-specific shrinkage weight. Comparing (5.12) and (5.13), it is apparent that the posterior shrinkage weight $1 - \mathbb{E}(\kappa_i | Y_i, \tau)$ mimics the role of w_i^* when one uses one-group prior instead of a two-group prior in a sparse situation.

Under a two-group framework, the optimal multiple testing rule for an additive $0 - 1$ loss, is given by:

$$\text{Reject}\, H_{0i} \,\text{if} \quad w_i > \frac{1}{2}\,, i = 1, 2, \cdots, n.$$

When $\beta \to 0$ and δ is fixed as $n \to \infty$, the first term in w_i^* goes to zero while the second goes to $\frac{\delta}{\delta+1}$ and thus $w_i^* \to \frac{\delta}{\delta+1} w_i$. As a result, the optimal testing rule is approximately:

$$\text{Reject } H_{0i} \text{ if } \quad \frac{\delta+1}{\delta} w_i^* > \frac{1}{2}, i = 1, 2, \cdots, n. \tag{5.14}$$

Using this and the previous observation (connecting $1 - \mathbb{E}(\kappa_i|Y_i, \tau)$ and w_i^*), a natural testing rule for simultaneous hypothesis testing using our class of priors can be taken as:

$$\text{Reject } H_{0i} \text{ if } \quad 1 - \mathbb{E}(\kappa_i|Y_i, \tau) > \frac{\delta}{2(\delta+1)}, i = 1, 2, \cdots, n. \tag{5.15}$$

It may be recalled that the decision rule in Datta and Dunson (2016) based on GH prior is similar and based on the same intuitive observation. Our decision rule is inspired by their work.

Our main theoretical result (Theorem 5.4.1) shows excellent performance of (5.15) in terms of Bayes risk (with respect to additive $0-1$ loss), if, among other things, τ is chosen proportional to p if p is known. When p is unknown, this strategy will not work. In this case, one may expect to get good results if τ is replaced by some $\widehat{\tau}$ that is of the order of the proportion of non-zero means using observed data. Keeping this in mind and inspired by the estimate $\max\{1, \sum_{i=1}^n \mathbf{1}\{Y_i \geq 1\}\}$ of the number of non-null effects in a Poisson count data considered by Yano et al. (2021), we propose to employ an empirical Bayes estimate $\widehat{\tau}$ of τ given by

$$\widehat{\tau} = \max\left\{\frac{1}{n}, \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y_i \geq 1\}\right\}. \tag{5.16}$$

We then apply the decision rule (5.15) by replacing τ with $\widehat{\tau}$. Thus the new rule for our multiple testing problem is obtained as:

$$\text{Reject } H_{0i} \text{ if } \quad 1 - E(\kappa_i|Y_i, \widehat{\tau}) > \frac{1}{2}, i = 1, 2, \cdots, n, \tag{5.17}$$

where $1 - E(\kappa_i|Y_i, \widehat{\tau})$ is the posterior shrinkage weight $1 - E(\kappa_i|Y_i, \tau)$ evaluated at $\tau = \widehat{\tau}$. The definition of $\widehat{\tau}$ ensures that it never collapses to zero, avoiding a major concern for empirical Bayes estimates in the literature. See in this context Carvalho et al. (2009), Scott and Berger (2010), Bogdan et al. (2008) and Datta and Ghosh (2013).

As mentioned before, our main interest is in the study of asymptotic optimality of rules (5.15) and (5.17) when applied in a two-group setting. Since both (5.15) and (5.17) depend on the posterior mean of κ_i, we first need the form of the posterior distribution of κ_i given Y_i and τ. It may be noted that using (5.1) and (5.4), the posterior distribution of κ_i's are independently distributed given $(Y_1, Y_2, \cdots, Y_n, \tau)$. The posterior distribution of κ_i depends only on Y_i given τ and has the following form :

$$\pi(\kappa_i|Y_i, \tau) \propto \kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1} L\left(\frac{1}{\tau^2}\left(\frac{1}{\kappa_i} - 1\right)\right), 0 < \kappa_i < 1. \tag{5.18}$$

In the next section, using (5.18), we provide some bounds on probabilities of type I and type II errors. They are later used to study the optimality property of the decision rule (5.15) for

quasi-sparse count data generated from a two-group model. However, the posterior distribution of κ_i given Y_i and τ evaluated at $\tau = \widehat{\tau}$ does not have an explicit form like (5.18) since the empirical Bayes estimate of τ depends on the entire dataset $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$. We need to devise new techniques for the study of optimality of (5.17) to take care of this issue. The technical details of this will be made clear later in this chapter.

5.4 Theoretical Results

This section is divided into several subsections, the first of which describes the main theoretical findings of this work. The second subsection is related to some concentration inequalities which are essential for proving our main results in this paper. The third one uses the results obtained in the former subsection to provide bounds on probabilities of type I and type II errors of decision rule (5.15). In the last subsection, we derive the bounds on probabilities of type I and type II errors of the empirical Bayes decision rule (5.17).

5.4.1 Results on Asymptotic Bayes Risk under sparsity using one-group priors

In this subsection, we present and discuss our main optimality results (in terms of Bayes risk with respect to additive $0-1$ loss) for our proposed simultaneous testing rules when the true means are generated from a two-component mixture prior (5.2). Our optimality results are derived under the sparse asymptotic regime described in Assumption 3 in Section 5.2. Our first result is described in Theorem 5.4.1 below, proof of which is provided in Section 5.8.

Theorem 5.4.1. *Let $Y_i \sim Poi(\theta_i)$ independently for $i = 1, 2, \cdots, n$ and suppose each θ_i is generated from (5.2). Suppose we want to test $H_{0i} : \nu_i = 0$ against $H_{1i} : \nu_i = 1$, for $i = 1, 2, \cdots, n$ using decision rule (5.15) induced by the class of priors (5.1) satisfying (5.4), where $L(\cdot)$ satisfies (A1) and (A2) defined in Section 5.3 with $a > 1$. Also assume that p, β and δ of the two-group model satisfy Assumption 3. Further assume that $\tau \to 0$ as $n \to \infty$. Then the Bayes risk (with respect to additive $0-1$ loss) of the multiple testing rule (5.15), denoted R_{OG}, satisfies,*

$$1 \leq \liminf_{n\to\infty} \frac{R_{OG}}{R^{BO}_{Opt}} \leq \limsup_{n\to\infty} \frac{R_{OG}}{R^{BO}_{Opt}} \leq (\delta+1)^{\alpha} \mathbb{P}\left(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)} \right). \tag{5.19}$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$.

Theorem 5.4.1 shows that the Bayes risk (R_{OG}) of the decision rule (5.15) based on one-group priors asymptotically remains within a multiplicative factor of the Bayes risk ($R^{\mathrm{BO}}_{\mathrm{Opt}}$) of the optimal decision rule in the two-group setting. In particular, we may say,

$$R_{\mathrm{OG}} = O(R^{\mathrm{BO}}_{\mathrm{Opt}}), \text{ as } n \to \infty.$$

In practice, p would generally not be known. We suspect that if in (5.15), τ is replaced by a $\widehat{\tau}$ that is a reasonable estimate of p, then we might be able to mimic the optimality result in the previous theorem. Our next theorem confirms this intuition, where we make use of the estimator (5.16) in replacing τ by $\widehat{\tau}$ in (5.15). The resulting empirical Bayesian multiple testing rule (5.17) is shown to possess similar optimality property (in Theorem 5.4.2 below). We also need a very mild additional assumption that $p \propto n^{-\epsilon}$ for some $0 < \epsilon < 1$ which covers most cases of theoretical and practical interest. We emphasize that ϵ need not be known for our result to go through. We only need that $\epsilon \in (0, 1)$. Proof of this result is provided in Section 5.8.

Theorem 5.4.2. *Let $Y_i \sim Poi(\theta_i)$ independently for $i = 1, 2, \cdots, n$ and suppose each θ_i is generated from (5.2). Suppose we want to test $H_{0i} : \nu_i = 0$ against $H_{1i} : \nu_i = 1$, for $i = 1, 2, \cdots, n$ using decision rule (5.17) induced by the class of priors (5.1) and (5.4), where $L(\cdot)$ satisfies (A1) and (A2) with $a > 1$. Also assume that p, β and δ of the two-group model satisfy Assumption 3 with $p \propto n^{-\epsilon}$ for some $0 < \epsilon < 1$. Then the Bayes risk (with respect to additive $0-1$ loss) of the multiple testing rule (5.17), denoted R_{OG}^{EB}, satisfies,*

$$1 \leq \liminf_{n\to\infty} \frac{R_{OG}^{EB}}{R_{Opt}^{BO}} \leq \limsup_{n\to\infty} \frac{R_{OG}^{EB}}{R_{Opt}^{BO}} \leq (\delta+1)^{\alpha} \mathbb{P}\left(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)} \right), \tag{5.20}$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$.

Theorems 5.4.1 and 5.4.2 are the first results of their kind in the literature on quasi-sparse count data. As far as we know, ours is the first formal study of asymptotic optimality of multiple testing in this context. Ghosh et al. (2016) proved a similar result for the same class of priors using an empirical Bayes estimate of τ given by van der Pas et al. (2014) for the normal means model. Our result confirms similar phenomenon even if the marginal distribution of observations is Negative Binomial.

5.4.2 Posterior Concentration inequalities

In this subsection, we describe some concentration and moment inequalities for the posterior distribution of the shrinkage coefficient κ_i's. These are essential for obtaining upper bounds on probabilities of type I and type II errors for decision rules (5.15) and (5.17). These upper bounds are discussed in the next subsections.

Before stating our results here, first we motivate how these may be useful for deriving the bounds for type I and type II errors. Let t_{1i} and t_{2i} denote the probabilities of type I and type II errors, respectively, corresponding to the i^{th} test. Recall by definition, $t_{1i} = \mathbb{P}_{H_{0i}}(E(1-\kappa_i|Y_i, \tau) > \frac{\delta}{2(\delta+1)})$ and $t_{2i} = \mathbb{P}_{H_{1i}}(E(\kappa_i|Y_i, \tau) \geq \frac{\delta+2}{2(\delta+1)})$. Now, with (5.18) being the posterior distribution of κ_i given Y_i and τ, it seems that finding exact asymptotic order of t_{1i} and t_{2i} is not feasible as the density involves an unknown L. Hence, Datta and Ghosh (2013) and Ghosh et al. (2016) (for normal means model) and Datta and Dunson (2016) (for count data) tried to derive some non-trivial asymptotic bounds for the same, although Datta and Dunson (2016) only derived this for t_{1i}. Following a similar strategy as in Ghosh et al. (2016), we will try to obtain bounds

on $E(1-\kappa_i|Y_i,\tau)$ and $E(\kappa_i|Y_i,\tau)$ and apply them to produce bounds on t_{1i} and t_{2i}.
Towards that, we first need the following concentration inequality on the posterior distribution of κ_i. Proof of this result is given in Section 5.8.

Theorem 5.4.3. *Suppose $Y_i \sim Poi(\theta_i)$ independently for $i = 1, 2, \cdots, n$. Consider the one-group prior given in (5.1) satisfying (5.4), where $L(\cdot)$ satisfies (A1). Then for any fixed $\epsilon \in (0,1)$, $a > 0$ and any $\tau > 0$,*

$$\mathbb{P}(\kappa_i < \epsilon|Y_i,\tau) \leq \frac{a}{c_0(a+\alpha)}(K_0^{-a}-K_1^{-a})^{-1}(\tau^2)^{a-Y_i}\left(\frac{\epsilon}{1-\epsilon}\right)^{a+\alpha} L\left(\frac{1}{\tau^2}\right)(1+K_1\tau^2)^{(Y_i+\alpha)}(1+o(1))\,,$$

where the $o(1)$ term is non-random, independent of index i and depends only on τ such that $\lim_{\tau\to 0} o(1) = 0$, and K_0 and K_1 are as in Lemma 5.8.1.

It may be noted that this result is true for any $a > 0$ in the definition of $\pi_1(\lambda_i^2)$. Next we describe an upper bound on the posterior mean of the shrinkage coefficient $(1-\kappa_i)$. Later, this bound will be used for proving an upper bound on t_{1i}. proof of which is provided in Section 5.8.

Theorem 5.4.4. *Consider the setup of Theorem 5.4.3 with the global-local prior of the form (5.1) satisfying (5.4) where $L(\cdot)$ satisfies Assumption 1. Then, for any fixed $\tau > 0$ and for any $Y_i \in [0, a-1)$ and $a > 1$,*

$$\mathbb{E}(1-\kappa_i|Y_i,\tau) \leq \frac{a}{c_0}(K_0^{-a}-K_1^{-a})^{-1}\tau^2[K^{-1}+\frac{M}{(a-Y_i-1)}](1+K_1\tau^2)^{(Y_i+\alpha)}\,,$$

where K_0 and K_1 are as in Lemma 5.8.1.

Remark 5.4.5. *For the normal means model, Ghosh et al. (2016) derived a result on the posterior distribution of κ_i similar to Theorem 5.4.3 of this work. However, their result is based on the Dominated Convergence Theorem. In contrast, ours uses a lower bound on the normalizing quantity of the posterior distribution of κ_i given Y_i and τ, as given in Lemma 5.8.1 and the properties of slowly varying function, as given in Lemma 4.8.1. They also obtained a result on the posterior mean of $1-\kappa_i$ similar to Theorem 5.4.4. Under Assumption 1 on $L(\cdot)$, they provided an upper bound on $\mathbb{E}(1-\kappa_i|Y_i,\tau)$ by using the definition, $\mathbb{E}(1-\kappa_i|Y_i,\tau) = \int_0^1 \mathbb{P}(\kappa_i < \epsilon|Y_i,\tau)d\epsilon$. The same technique can not be used in our situation as bounding $\int_0^1(\frac{\epsilon}{1-\epsilon})^{a+\alpha}d\epsilon$ by $\frac{1}{1-(a+\alpha)}$ is possible only when $0 < a+\alpha < 1$. However, for any $a > 1$, $a+\alpha > 1$ always. As a result, we need to use a completely different argument to obtain a non-trivial upper bound on the posterior expectation of the shrinkage coefficient $1-\kappa_i$ in Theorem 5.4.4.*

After having a non-trivial bound on $\mathbb{E}(1-\kappa_i|Y_i,\tau)$, our next step is to search for a similar bound on $\mathbb{E}(\kappa_i|Y_i,\tau)$ for the purpose of handling the type II error. Since, for any $\eta \in (0,1)$, $\mathbb{E}(\kappa_i|Y_i,\tau)$ can be expressed as the sum of $\mathbb{E}(\kappa_i\mathbf{1}\{\kappa_i > \eta\}|Y_i,\tau)$ and $\mathbb{E}(\kappa_i\mathbf{1}\{\kappa_i \leq \eta\}|Y_i,\tau)$, we need to find an upper bound for $\mathbb{P}(\kappa_i > \eta|Y_i,\tau)$ to bound $\mathbb{E}(\kappa_i\mathbf{1}\{\kappa_i > \eta\}|Y_i,\tau)$. The second term in the sum can be handled easily. The next theorem provides a concentration inequality in this context. Proof of this result is given in Section 5.8.

Theorem 5.4.6. *Under the set up of Theorem 5.4.3, for any fixed $\eta \in (0,1)$ and $\delta_1 \in (0,1)$ and for all sufficiently small $\tau(>0)$,*

$$\mathbb{P}(\kappa_i > \eta | Y_i, \tau) \leq \frac{(a+\alpha)}{Kc_0}\tau^{-2a}\left(\frac{1-\eta}{1-\eta\delta_1}\right)^{Y_i}(\eta\delta_1)^{-(a+\alpha)}.$$

It may be noted that this result is true by taking any $a > 0$.

Results obtained in Theorems 5.4.3-5.4.6 reveal some interesting characteristics of the posterior distribution of κ_i based on our choice of priors. These results also lead to the following three corollaries, whose proofs are immediate and hence omitted.

We first have a corollary of Theorem 5.4.3.

Corollary 5.4.7. *Under the assumptions of Theorem 5.4.3, $\mathbb{P}(\kappa_i \geq \epsilon | Y_i, \tau) \to 1$ as $\tau \to 0$ for any fixed $\epsilon \in (0,1)$ whenever $0 \leq Y_i < a$.*

Datta and Dunson (2016) also obtained a similar result for the Gauss hypergeometric prior under quasi-sparse count data. Thus for any fixed $0 \leq Y_i < a$, for our choice of one-group global-local shrinkage priors, the posterior distribution of κ_i concentrates near 1 with high probability for small values of τ.

We next have a corollary to Theorem 5.4.4.

Corollary 5.4.8. *Under the assumptions of Theorem 5.4.4, $\mathbb{E}(1-\kappa_i | Y_i, \tau) \to 0$ as $\tau \to 0$ uniformly in $Y_i \in [0, a-1)$.*

This corollary ensures that in the case of global-local priors, for small values of τ, the noise observations will be shrunk towards the origin.

Theorem 5.4.6 yields the following corollary.

Corollary 5.4.9. *Under the assumptions of Theorem 5.4.6, $\mathbb{P}(\kappa_i > \eta | Y_i, \tau) \to 0$ as $Y_i \to \infty$ for any sufficiently small $\tau > 0$.*

This corollary implies that for our choice of one-group global-local shrinkage priors, the posterior distribution of κ_i concentrates near 0 with high probability for large values of Y_i if τ is small enough.

5.4.3 Type I and Type II error bounds for testing the rule (5.15)

As mentioned earlier, this subsection provides some asymptotic bounds on the probabilities of both types of errors for the decision rule (5.15) under the assumption that p is known. Theorem 5.4.10 gives a non-trivial upper bound on the probability of type I error (t_{1i}) for the decision rule (5.15). On the other hand, Theorem 5.4.12 provides an upper bound on the probability of type II error (t_{2i}). We may recall that Datta and Dunson (2016) did not study the decision theoretic optimality of their decision rule (based on GH prior) which has the same structure as (5.15). They only proved that the probability of type I error goes to zero asymptotically. However, their

argument has a soft spot and the upper bound for type I error also needs modification. We feel that proof of our result (Theorem 5.4.10) on the upper bound of type I error may be helpful in getting rid of these lacunae but we have not pursued this in this thesis. Proofs of these results are provided in Section 5.8.

Theorem 5.4.10. *Let $Y_i \sim Poi(\theta_i)$ independently for $i = 1, 2, \cdots, n$ and suppose each θ_i is generated from (5.2). Suppose we want to test $H_{0i} : \nu_i = 0$ against $H_{1i} : \nu_i = 1$, for $i = 1, 2, \cdots, n$ using decision rule (5.15) induced by the class of priors (5.1) satisfying (5.4), where $L(\cdot)$ satisfies (A1) and (A2) with $a > 1$. Also assume that p, β, and δ of the two-group model satisfy Assumption 3. Further, assume that $\tau \to 0$ as $n \to \infty$. Then as $n \to \infty$, the probability of type I error (t_{1i}) of the decision rule (5.15), satisfies*

$$t_{1i} \equiv t_1 \leq \frac{2\alpha\beta}{a}.$$

Remark 5.4.11. *Under Assumption 3, not only does t_1 go to zero as $n \to \infty$, but its rate of convergence towards zero is faster than that of p. This fact will be used in obtaining an asymptotic expression for the Bayes risk R_{OG}, reported in Theorem 5.4.1.*

Theorem 5.4.12. *Consider the set-up of Theorem 5.4.10. Then as $n \to \infty$, the probability of type II error (t_{2i}) of the decision rule (5.15), denoted t_2, satisfies*

$$t_{2i} \equiv t_2 \leq \mathbb{P}\left(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)}\right)(1 + o(1)),$$

where the $o(1)$ term is non-random, independent of index i and depends on a, α and τ such that $\lim_{\tau \to 0} o(1) = 0$ and $Y \sim NB(\alpha, \frac{1}{\delta+1})$.

Remark 5.4.13. *This is the first result of its kind in the literature on one-group priors based on quasi-sparse count data. It is important to observe that the techniques used in this result are completely different from those of Theorem 7 of Ghosh et al. (2016), which is also related to an upper bound of the type II error for the sparse normal means model. Their result is based on the usage of the upper bound on $\mathbb{P}(\kappa_i > \eta | Y_i, \tau)$. In contrast, ours uses the definition of the type II error, followed by a careful division on the range of Y_i and lastly the Dominated Convergence Theorem.*

5.4.4 Type I and Type II error bounds corresponding to an empirical Bayes procedure

In this subsection, we present asymptotic bounds on the probabilities of type I and type II errors of the individual decision rules (5.17) corresponding to the empirical Bayes approach discussed before. These are reported in Theorems 5.4.14 and 5.4.15 respectively. Approaches involved in the proofs of these theorems are significantly different from those employed for proving Theorem 5.4.10 and Theorem 5.4.12. Note that, when τ is used as a tuning parameter, the decision rule (5.15) corresponding to i^{th} test depends on i^{th} observation Y_i only. On the other hand, $\widehat{\tau}$ depends

on (see (5.16)) the entire dataset $(Y_1, Y_2, \cdots, Y_n)$. Obviously, this necessitates serious changes in the approach. Towards that, we introduce a cut-off value that is asymptotically of the order p. Next, we divide the range of $\widehat{\tau}$ into two mutually disjoint parts depending on whether $\widehat{\tau}$ exceeds the cut-off or not. For one part, noting that for any fixed $y \geq 0$, $\mathbb{E}(1-\kappa|y,\tau)$ is non-decreasing in τ, we can use results already reported for the case when τ is used as a tuning parameter. For the other part, the form of $\widehat{\tau}$ along with the independence of Y_i's come into play for obtaining asymptotic bounds on the quantities of interest. We now present Theorem 5.4.14.

Theorem 5.4.14. *Let $Y_i \sim Poi(\theta_i)$ independently for $i = 1, 2, \cdots, n$ and suppose each θ_i is generated from (5.2). Suppose we want to test $H_{0i} : \nu_i = 0$ against $H_{1i} : \nu_i = 1$, for $i = 1, 2, \cdots, n$ using decision rule (5.17) induced by the class of priors (5.1) satisfying (5.4), where $L(\cdot)$ satisfies (A1) and (A2) with $a > 1$. Also assume that p, β and δ of the two-group model satisfy Assumption 3 with $p \propto n^{-\epsilon}$, for some $0 < \epsilon < 1$. Then as $n \to \infty$, the probability of type I error of the decision rule (5.17), denoted t_{1i}^{EB}, satisfies*

$$t_{1i}^{EB} \leq \frac{2\alpha\beta}{a} + \alpha\beta + e^{-(2\log 2 - 1)(1-(\beta+\delta+1)^{-\alpha})np(1+o(1))},$$

where the $o(1)$ term is non-random, independent of index i and tends to zero as $n \to \infty$.

Theorem 5.4.15. *Consider the set-up of Theorem 5.4.14. Then as $n \to \infty$, the probability of type II error of the decision rule (5.17), denoted t_{2i}^{EB} satisfies*

$$t_{2i}^{EB} \leq \mathbb{P}\left(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)}\right)(1+o(1)),$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$ and the $o(1)$ term is non-random, independent of index i and tends to zero as $n \to \infty$.

Remark 5.4.16. *While proving Theorem 5.4.14 and Theorem 5.4.15, we have used some arguments similar to those of Ghosh et al. (2016). But only these arguments are not enough to establish these results. We need to use Theorem 5.4.10 and Theorem 5.4.12 too to complete the proofs in this subsection. Our work shows that with some non-trivial modifications, the techniques used by Ghosh et al. (2016) can be useful even if the underlying model is not normal.*

5.5 Simulation Results

In this section, we are going to compare through simulations the performance of our decision rules (5.15) and (5.17) vis-a-vis the Bayes Oracle (5.7), when the data is truly generated from a two-group model and the data is modelled using a member of TPBN families as the prior on the θ_i's. We will also use the Gauss Hypergeometric (GH) prior of Datta and Dunson (2016). The main reason to include GH prior is that this prior is not fully included in our chosen class of priors, and hence any optimality property of it does not follow theoretically using our results. Hence, it is of interest to also study the simulation performance of their decision rule. Our

Table 21: Average misclassification probabilities based on 1000 replications

Sparsity level	Two-group prior	TPBN-tun	TPBN-EB	Gauss Hypergeometric
0.01	0.032	0.032	0.032	0.032
0.02	0.048	0.048	0.048	0.048
0.03	0.056	0.056	0.058	0.058
0.04	0.061	0.063	0.064	0.062
0.05	0.065	0.068	0.069	0.067
0.06	0.076	0.081	0.084	0.082
0.07	0.084	0.090	0.093	0.090
0.08	0.089	0.096	0.099	0.095
0.09	0.093	0.101	0.104	0.100
0.10	0.095	0.104	0.107	0.102
0.11	0.102	0.112	0.116	0.111
0.12	0.105	0.116	0.121	0.115
0.13	0.109	0.122	0.130	0.120
0.14	0.113	0.130	0.139	0.128
0.15	0.115	0.134	0.142	0.134
0.16	0.119	0.139	0.147	0.138
0.17	0.125	0.147	0.155	0.149
0.18	0.125	0.150	0.160	0.151
0.19	0.128	0.155	0.168	0.155
0.20	0.131	0.161	0.172	0.162

comparison will be based on the average value of the proportion of misclassified hypotheses in the simulation, which is used as an estimate of the risk with respect to additive $0-1$ loss.

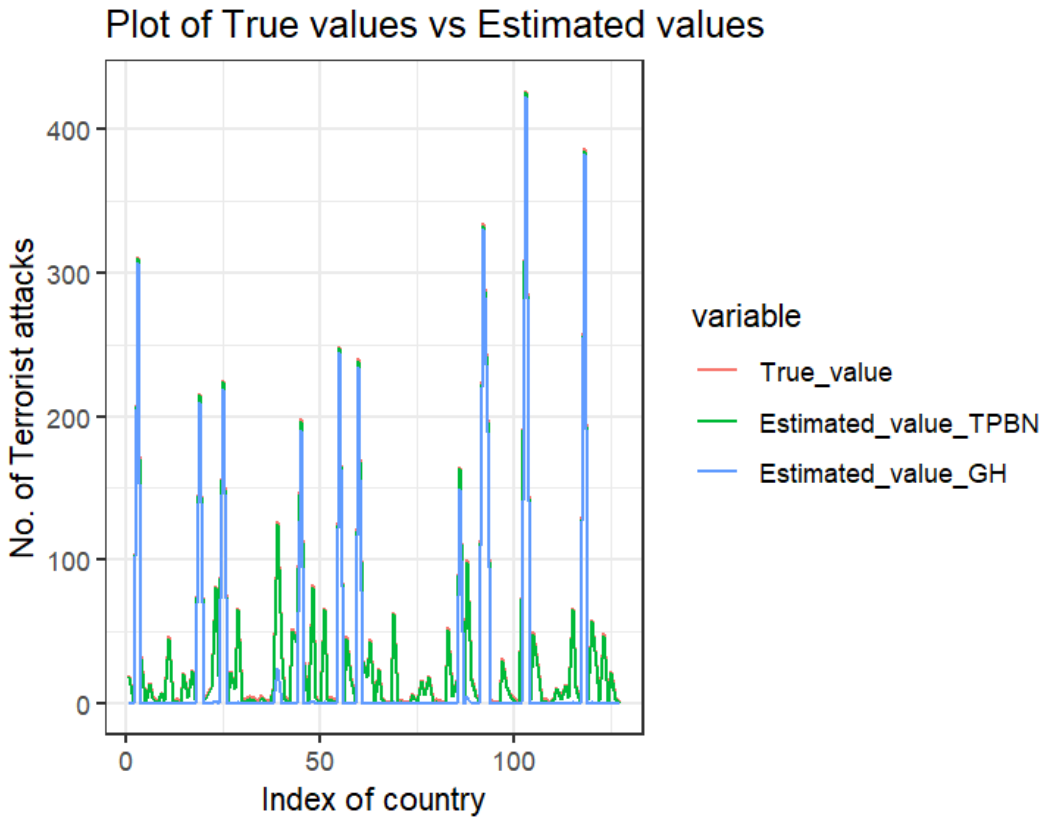
The data generating scheme is as follows. For any fixed level of sparsity $p \in (0,1)$, we generate $n = 200$ independent observations $Y_1, \cdots, Y_n$, from a two-group model (5.3). Motivated by Datta and Dunson (2016), we take $\alpha = 0.5, \beta = 1, \delta = 10$. All the rules under study are applied on the generated dataset of size 200. This procedure is repeated 1000 times and the average number of misclassifications is reported. For the simulation study, we use three parameter beta normal (TPBN) prior with both hyperparameters being 1.5 and choose the global parameter τ of the same order as p. We also use the same GH prior as in the study of Datta and Dunson (2016). We also use the empirical Bayes estimate of τ, given in (5.16). The results are presented in Table 21, where $p \in \{0.01, 0.02, \cdots, 0.2\}$. In the table the column TPBN-tun refers to the results obtained for rule (5.15), while TPBN-EB refers to those for rule (5.17).

When the level of sparsity is small, the misclassification probabilities of decision rules based on our chosen one-group prior are quite close to those of the Oracle rule, which aligns well with the results established in this chapter. The results corresponding to the GH prior are also similar to those of ours, which indicates that the theoretical optimality property is possibly also true for the testing rule of Datta and Dunson (2016).

5.6 Real data analysis

We also apply our modelling in the Global Terrorism Dataset for the years 1970-2020, which consists of the number of terrorist attacks (Y_i) in different countries over the previously mentioned period. This dataset is available on the website Global Terrorism Database on request. The Figure 5.1 contains the number of terrorist attacks reported in different countries (names blinded) over the years, along with corresponding estimates using our approach and that of the GH prior-based approach. The estimates are the (empirical Bayes) posterior expected means of the rates of the terrorist attacks for different countries. We have used the empirical Bayes version of the posterior mean here as the fraction of "nulls" is unknown here and used $\widehat{\tau} = \max\{\frac{1}{n}, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq 100\}\}$. The change in the definition of $\widehat{\tau}$ from (5.17) is done due to the fact that we found it is reasonable to consider an observation coming from "non-null" group if the average number of attacks in one year is at least 2. It is quite clear from the figure that our prior provides better estimates than those of the GH prior when the number of terrorist attacks is moderate. On the other hand, even if one is interested in considering the countries

Figure 5.1: Performance of our method on real dataset



which are worst hit, then also, our approach provides a somewhat more accurate estimate of the true number of attacks than that of Datta and Dunson (2016). The counts for the top 10 worst hit countries along with the estimates obtained by us and that using Datta and Dunson (2016) are illustrated in Table 22.

Finally, in order to understand the role of estimate of τ based on the dataset, we consider

Table 22: Performance of our method on real dataset for worst hit countries

Country no.	No. of Terrorist Attacts	Estimate using GH	Estimate using TPBN
Country 1	426	424	424
Country 2	386	384	385
Country 3	334	331	332
Country 4	311	307	309
Country 5	249	244	248
Country 6	240	235	239
Country 7	216	210	215
Country 8	225	220	224
Country 9	198	191	197
Country 10	198	191	196

Table 23: Misclassification probabilities with the choice of empirical Bayes estimate of τ

Non-null	Estimate of τ	MP
$\#\{Y_i \geq 100\}$	$\max\{\frac{1}{n}, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq 50\}\}$	0.09448
$\#\{Y_i \geq 100\}$	$\max\{\frac{1}{n}, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq 100\}\}$	0.00787
$\#\{Y_i \geq 100\}$	$\max\{\frac{1}{n}, \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq 200\}\}$	0.02362

different empirical Bayes estimates of τ. We consider countries with total number of attacks at least 100 as belonging to non-null group. Here the difference in estimates lies in using different thresholds to categorize an observation as a signal or not. We choose three values of the thresholds, namely $50, 100$ and 200, namely, and compute the misclassification probabilities. The results are provided in Table 23. It is clear from the table that the choice of τ plays a pivotal role in controlling the misclassification probability. The misclassification probability, when the sample estimate of τ is of order p, is significantly less compared to other choices of τ. This provides a clear picture of the role of the global parameter τ, when in a multiple testing problem involving count data, the mean parameter is modeled by one-group prior.

5.7 Concluding remarks and scope for future work

In this chapter, we study modelling of high-dimensional count data containing a lot of values near zero, namely "quasi-sparse" count data. Assuming that observations are generated from a Poisson distribution with mean θ_i and modelling θ_i by (5.2), at first, we obtained an expression for the optimal Bayes risk under an additive $0-1$ loss function. This was obtained under suitable assumptions on the model parameters. Next, our focus was to study the optimality property of a decision rule based on a class of one-group priors when the true data is generated from a two-group model. In this context, we have been able to establish that irrespective of whether the level of sparsity is known or not, our decision rules using global-local priors attain the optimal Bayes risk, up to some multiplicative constant. To the best of our knowledge, these results are the first of their kind in the literature on sparse count data.

Throughout our calculations, we either treat τ as a tuning parameter or use an empirical Bayes estimate of it. A question that arises naturally is whether the same optimality property holds for a similar decision rule when one uses a non-degenerate prior on τ. Questions can also be asked regarding the choice of the prior on τ or the range on which it has to be defined. Though we have not discussed these here, we hope techniques used in Section 2.3 of Chapter 2 of this thesis might be helpful for this purpose also.

Hamura et al. (2022) also studied a class of global-local priors for modeling quasi-sparse count data. They proposed two priors, namely, inverse gamma (IG) prior and extremely heavy-tailed (EH) prior, and proved that both of them possess the *tail robustness* property introduced by Carvalho et al. (2010). However, they kept the global parameter fixed and did not study the asymptotic optimality in terms of the Bayes risk of the decision rule induced by their chosen class of priors. We expect that by letting the global parameter go to zero as the sample size increases and using arguments similar to ours, one may be able to establish results similar to Theorem 5.4.1 and Theorem 5.4.2 of this work. We also expect that our techniques will be handy in the case of the Gauss Hypergeometric prior used by Datta and Dunson (2016) for providing a non-trivial bound on the probability of type II error and the corresponding Bayes risk of their decision rule, irrespective of the underlying sparsity to be known or unknown. It may be recalled that the simulation performance of the GH prior is very encouraging in terms of Bayes risk.

As mentioned by Hamura et al. (2022), one possible drawback of the Poisson modelling is that for the Poisson distribution, the mean and the variance are the same, a situation which does not always hold, specially in sparse count data. One possible remedy is to use the Negative Binomial distribution to model the situation and try to provide answer to the same questions discussed here. However, instead of considering a particular distribution, one may consider a general class of distributions and study such optimality for that class. We have already obtained some results in this context when observations are generated from a subclass of one parameter exponential family distributions containing Normal, Poisson and negative binomial as special cases. Interestingly, these three also belong to the *Natural Exponential Family with Quadratic Variance function*, or in short, NEF-QVF, proposed by Morris (1982). However, optimality results in terms of Bayes risk are still not established for the class of Morris (1982). We will be actively pursuing this in the near future.

5.8 Proofs

Lemma 5.8.1. *Let $L : (0,\infty) \to (0,\infty)$ be a measurable function satisfying Assumption 1 and a be any positive real number. Suppose $\tau \to 0$ as $n \to \infty$. Then for any $y = 0, 1, 2, \cdots$, $\alpha > 0$, and fixed $K_1 > K_0 > \max\{1, t_0\}$,*

$$\int_0^1 u^{a+\alpha-1}(1-u)^{y-a-1} L\left(\frac{1}{\tau^2}(\frac{1}{u}-1)\right) du \geq \frac{c_0(K_0^{-a}-K_1^{-a})}{a}(\tau^2)^{y-a}(1+K_1\tau^2)^{-(y+\alpha)},$$

where t_0 is as in Assumption 1 of Section 2.2.

Proof. Let $I = \int_0^1 u^{a+\alpha-1}(1-u)^{y-a-1}L\left(\frac{1}{\tau^2}(\frac{1}{u}-1)\right)du$. Then with the change of variable $t = \frac{1}{\tau^2}(\frac{1}{u}-1)$, we have

$$I = (\tau^2)^{y-a}\int_0^\infty (1+t\tau^2)^{-(y+\alpha)}t^{y-a-1}L(t)dt\,. \tag{5.21}$$

By Assumption 1 on $L(\cdot)$, $\exists K_0(>1)$ such that $L(t) \geq c_0$ if $t \geq K_0 \geq t_0$. Hence, we obtain

$$\begin{aligned}\int_0^\infty (1+t\tau^2)^{-(y+\alpha)}t^{y-a-1}L(t)dt &\geq \int_{K_0}^\infty (1+t\tau^2)^{-(y+\alpha)}t^{y-a-1}L(t)dt\\ &\geq c_0\int_{K_0}^\infty (1+t\tau^2)^{-(y+\alpha)}t^{-a-1}dt\\ &\geq c_0\int_{K_0}^{K_1} (1+t\tau^2)^{-(y+\alpha)}t^{-a-1}dt\\ &\geq \frac{c_0(K_0^{-a}-K_1^{-a})}{a}(1+K_1\tau^2)^{-(y+\alpha)}. \end{aligned}\tag{5.22}$$

Finally, the result follows from (5.21) and (5.22). □

Proof of Theorem 5.4.3. Fix any $\epsilon \in (0,1)$. Then

$$\mathbb{P}(\kappa_i < \epsilon|Y_i,\tau) = \frac{\int_0^\epsilon \kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1}L\left(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\right)d\kappa_i}{\int_0^1 \kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1}L\left(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\right)d\kappa_i}\,.$$

Now using the change of variable $t = \frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)$ to the numerator and denominator of the previous quantity and applying (5.22) of Lemma 5.8.1 for the denominator, we have

$$\mathbb{P}(\kappa_i < \epsilon|Y_i,\tau) \leq \frac{a}{c_0}(K_0^{-a}-K_1^{-a})^{-1}(1+K_1\tau^2)^{(Y_i+\alpha)}\int_{\frac{1}{\tau^2}(\frac{1}{\epsilon}-1)}^\infty (1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt\,. \tag{5.23}$$

Also, note that

$$\begin{aligned}\int_{\frac{1}{\tau^2}(\frac{1}{\epsilon}-1)}^\infty (1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt &= (\tau^2)^{-Y_i-\alpha}\int_{\frac{1}{\tau^2}(\frac{1}{\epsilon}-1)}^\infty \left(\frac{t\tau^2}{1+t\tau^2}\right)^{Y_i+\alpha} t^{-\alpha-a-1}L(t)dt\\ &\leq (\tau^2)^{-Y_i-\alpha}\int_{\frac{1}{\tau^2}(\frac{1}{\epsilon}-1)}^\infty t^{-\alpha-a-1}L(t)dt\,. \end{aligned}\tag{5.24}$$

Next using Lemma 4.8.1 and the definition of $L(\cdot)$, we obtain

$$\int_{\frac{1}{\tau^2}(\frac{1}{\epsilon}-1)}^\infty t^{-\alpha-a-1}L(t)dt = \frac{1}{(a+\alpha)}(\tau^2)^{a+\alpha}\left(\frac{\epsilon}{1-\epsilon}\right)^{a+\alpha}L\left(\frac{1}{\tau^2}\right)(1+o(1))\,. \tag{5.25}$$

Using (5.23)-(5.25) the proof of Theorem 5.4.3 is obtained easily. □

Proof of Theorem 5.4.4. Using the definition,

$$\mathbb{E}(1-\kappa_i|Y_i,\tau)=\frac{\int_0^1\kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a}L\Big(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\Big)d\kappa_i}{\int_0^1\kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1}L\Big(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\Big)d\kappa_i}.$$

Again using the change of variable $t=\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)$ in the numerator and denominator of the previous quantity and applying (5.22) of Lemma 5.8.1 for the denominator, we have

$$\mathbb{E}(1-\kappa_i|Y_i,\tau)\leq\frac{a}{c_0}(K_0^{-a}-K_1^{-a})^{-1}\tau^2(1+K_1\tau^2)^{(Y_i+\alpha)}\int_0^\infty(1+t\tau^2)^{-(Y_i+\alpha+1)}t^{Y_i-a}L(t)dt\,. \quad (5.26)$$

Next, we divide the range of integration into two parts, namely in $t\in(0,1)$ and $t\geq 1$. We observe that

$$\tau^2\int_0^1(1+t\tau^2)^{-(Y_i+\alpha+1)}t^{Y_i-a}L(t)dt\leq\tau^2K^{-1}\,. \quad (5.27)$$

Next using Assumption 1 on $L(\cdot)$, it is easy to see that for any $Y_i\in[0,a-1)$,

$$\tau^2\int_1^\infty(1+t\tau^2)^{-(Y_i+\alpha+1)}t^{Y_i-a}L(t)dt\leq\frac{\tau^2M}{(a-Y_i-1)}\,. \quad (5.28)$$

Finally, Theorem 5.4.4 follows from (5.26)-(5.28). □

Proof of Theorem 5.4.6. Fix any $\eta\in(0,1)$ and $\delta_1\in(0,1)$. Now using the definition,

$$\mathbb{P}(\kappa_i>\eta|Y_i,\tau)=\frac{\int_\eta^1\kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1}L\Big(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\Big)d\kappa_i}{\int_0^1\kappa_i^{a+\alpha-1}(1-\kappa_i)^{Y_i-a-1}L\Big(\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)\Big)d\kappa_i}.$$

Then using the change of variable $t=\frac{1}{\tau^2}(\frac{1}{\kappa_i}-1)$ to both the numerator and denominator of the right-hand side of the above equation, we obtain, for any $\delta_1\in(0,1)$,

$$\begin{aligned}\mathbb{P}(\kappa_i>\eta|Y_i,\tau)&=\frac{\int_0^{\frac{1}{\tau^2}(\frac{1}{\eta}-1)}(1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt}{\int_0^\infty(1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt}\\&\leq\frac{\int_0^{\frac{1}{\tau^2}(\frac{1}{\eta}-1)}(1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt}{\int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta_1}-1)}^\infty(1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt}\,. \quad (5.29)\end{aligned}$$

Note that the numerator of (5.29) can be further bounded as

$$\begin{aligned}\int_0^{\frac{1}{\tau^2}(\frac{1}{\eta}-1)}(1+t\tau^2)^{-(Y_i+\alpha)}t^{Y_i-a-1}L(t)dt&=(\tau^2)^{-Y_i}\int_0^{\frac{1}{\tau^2}(\frac{1}{\eta}-1)}(\frac{t\tau^2}{1+t\tau^2})^{Y_i}(1+t\tau^2)^{-\alpha}t^{-a-1}L(t)dt\\&\leq(\tau^2)^{-Y_i}(1-\eta)^{Y_i}\int_0^\infty t^{-a-1}L(t)dt\\&=K^{-1}\Big(\frac{1-\eta}{\tau^2}\Big)^{Y_i}. \quad (5.30)\end{aligned}$$

Here the inequality occurs due to the fact that $t\tau^2/(1+t\tau^2)$ is increasing in t for any fixed $\tau > 0$ whenever $t \in (0, \frac{1}{\tau^2}(\frac{1}{\eta} - 1))$. Next we use $\int_0^\infty t^{-a-1}L(t)dt = K^{-1}$.
On the other hand, for the denominator, we have

$$\begin{aligned}
&\int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta_1}-1)}^{\infty} (1+t\tau^2)^{-(Y_i+\alpha)} t^{Y_i-a-1} L(t)dt \\
&= (\tau^2)^{-Y_i-\alpha} \int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta_1}-1)}^{\infty} \left(\frac{t\tau^2}{1+t\tau^2}\right)^{Y_i+\alpha} t^{-a-\alpha-1} L(t)dt \\
&\geq (\tau^2)^{-Y_i-\alpha}(1-\eta\delta_1)^{Y_i+\alpha} \int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta_1}-1)}^{\infty} t^{-a-\alpha-1} L(t)dt \\
&\geq c_0(\tau^2)^{-Y_i-\alpha}(1-\eta\delta_1)^{Y_i+\alpha} \int_{\frac{1}{\tau^2}(\frac{1}{\eta\delta_1}-1)}^{\infty} t^{-a-\alpha-1} dt \\
&= \frac{c_0}{(a+\alpha)} \left(\frac{1}{\tau^2}\right)^{Y_i-a} (1-\eta\delta_1)^{Y_i-a} (\eta\delta_1)^{a+\alpha}. \qquad (5.31)
\end{aligned}$$

In the chain of inequalities, the first one holds due to the fact that $t\tau^2/(1+t\tau^2)$ is increasing in t for any fixed $\tau > 0$ whenever $t \geq \frac{1}{\tau^2}(\frac{1}{\eta\delta_1} - 1)$. The second inequality follows from Assumption 1 on $L(\cdot)$ by noting that for any fixed $\eta \in (0,1)$ and $\delta_1 \in (0,1)$, τ can be made small enough so that $\frac{1}{\tau^2}(\frac{1}{\eta\delta_1} - 1) \geq t_0$. Finally, (5.30) and (5.31) provide the upper bound on (5.29) and complete the proof of Theorem 5.4.6. □

Proof of Theorem 5.4.10. In order to calculate type I error, note that, the posterior distribution of κ_i given Y_i and τ has the same from for all $i = 1, 2, \cdots, n$. Also, under H_{0i}, the marginal distribution of Y_i is the same for all $i = 1, 2, \cdots, n$. These two facts indicate that the probability of type I error of i^{th} decision rule, denoted as t_{1i} is independent of i. We further then have,

$$\begin{aligned}
t_{1i} \equiv t_1 &= \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i,\tau) > \frac{\delta}{2(\delta+1)}) \\
&\leq \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i,\tau) > \frac{\delta}{2(\delta+1)}, Y_i \leq \frac{a}{2}) + \mathbb{P}_{H_{0i}}(Y_i > \frac{a}{2}). \qquad (5.32)
\end{aligned}$$

Now concentrate on the first term in (5.32). In this case, using Theorem 5.4.4, as $\tau \to 0$, this probability can be further bounded by

$$\begin{aligned}
&\mathbb{P}_{H_{0i}}\left(M_1\tau^2[K^{-1} + \frac{M}{\frac{a}{2}-1}](1+K_1\tau^2)^{(Y_i+\alpha)} > \frac{\delta}{2(\delta+1)}, Y_i \leq \frac{a}{2}\right) \\
&\leq \mathbb{P}\left(M_2\tau^2(1+K_1\tau^2)^{(\frac{a}{2}+\alpha)} > \frac{\delta}{2(\delta+1)}\right) \\
&= 0, \text{for all sufficiently large } n, \qquad (5.33)
\end{aligned}$$

where M_1 and M_2 are constants depending only on c_0, a, α, K_0 and K_1 and are independent of Y_i. Since, under H_{0i}, $Y_i \overset{iid}{\sim} NB(\alpha, \frac{1}{\beta+1})$, for the second term of (5.32), using the Markov's inequality, we have

$$\mathbb{P}_{H_{0i}}(Y_i > \frac{a}{2}) \leq \frac{2\alpha\beta}{a}. \qquad (5.34)$$

The proof is completed using (5.32)-(5.34). □

Proof of Theorem 5.4.12. Note that under H_{1i}, the distribution of Y_i is the same for all i, $i = 1, 2, \cdots, n$. Also, note that, the form of the posterior distribution of κ_i given Y_i and τ is same for all $i = 1, 2, \cdots, n$. As a consequence of this, the probability of type II error of i^{th} decision rule is denoted as t_2.
Hence the upper bound on t_2 is of the form

$$\begin{aligned} t_2 &= \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\tau) \geq \frac{\delta+2}{2(\delta+1)}) \\ &\leq \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\tau) \geq \frac{\delta+2}{2(\delta+1)}, Y_i > a) + \mathbb{P}_{H_{1i}}(Y_i \leq a). \end{aligned} \tag{5.35}$$

First note that

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}_{H_{1i}}(Y_i \leq a) &= \lim_{n\to\infty} \sum_{y=0}^{[a]} \binom{y+\alpha-1}{y} (1 - \frac{1}{\beta+\delta+1})^y (\beta+\delta+1)^{-\alpha} \\ &= \sum_{y=0}^{[a]} \binom{y+\alpha-1}{y} (1 - \frac{1}{\delta+1})^y (\delta+1)^{-\alpha} = \mathbb{P}(Y \leq a), \end{aligned} \tag{5.36}$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$ and $[x]$ denotes the greatest integer less than or equal to x. Note that

$$\lim_{n\to\infty} \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\tau) \geq \frac{\delta+2}{2(\delta+1)}, Y_i > a) = \lim_{n\to\infty} \mathbb{E}_{H_{1i}}(Z_n),$$

where $Z_n = 1\{\mathbb{E}(\kappa_i|Y_i,\tau) \geq \frac{\delta+2}{2(\delta+1)}, Y_i > a\}$. Therefore

$$\begin{aligned} \mathbb{E}_{H_{1i}}(Z_n) &= \int_\Omega 1\{\mathbb{E}(\kappa|y,\tau) \geq \frac{\delta+2}{2(\delta+1)}, y > a\} f_n(y) d\mu \\ &= \int_\Omega g_n(y) d\mu, \text{say}, \end{aligned}$$

where $f_n(y)$ denotes the probability mass function of Y_i at y under H_{1i} and is of the form

$$f_n(y) = \binom{y+\alpha-1}{y} (1 - \frac{1}{\beta+\delta+1})^y (\beta+\delta+1)^{-\alpha}, y = 0, 1, 2, \cdots$$

and μ is the counting measure on $\Omega = \{0, 1, 2, \cdots\}$. Here,

$$E(\kappa|y,\tau) = \frac{\int_0^1 \kappa^{a+\alpha}(1-\kappa)^{y-a-1} L\left(\frac{1}{\tau^2}(\frac{1}{\kappa}-1)\right) d\kappa}{\int_0^1 \kappa^{a+\alpha-1}(1-\kappa)^{y-a-1} L\left(\frac{1}{\tau^2}(\frac{1}{\kappa}-1)\right) d\kappa}. \tag{5.37}$$

We now show that $f_n(y) \leq h(y)$ for all $y \in \Omega$, for some $h(y)$ such that it's distribution is independent of n and $\int_\Omega h(y) d\mu < \infty$. Since, $\beta \to 0$ as $n \to \infty$, for sufficiently large n, we

assume $\beta \leq 1$. As a result,

$$\begin{aligned} f_n(y) &\leq \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+2})^y(\delta+1)^{-\alpha} \\ &= \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha}\frac{(1-\frac{1}{\delta+2})^y}{(1-\frac{1}{\delta+1})^y} \\ &= \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha}(1+\frac{1}{\delta^2+2\delta})^y. \end{aligned}$$

Hence, if we define $h(y) = \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha}(1+\frac{1}{\delta^2+2\delta})^y$, then

$$\sum_{y=0}^{\infty} h(y) = \mathbb{E}[(1+\frac{1}{\delta^2+2\delta})^Y] = \mathbb{E}[e^{Y\log\left(1+\frac{1}{\delta^2+2\delta}\right)}],$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$. Recall that if $X \sim NB(r,q)$, then the moment generating function (MGF) of t exists if $t < -\log(1-q)$. Here $q = \frac{1}{\delta+1}$. Therefore $\mathbb{E}[e^{Y\log\left(1+\frac{1}{\delta^2+2\delta}\right)}]$ exists if $\log\left(1+\frac{1}{\delta^2+2\delta}\right) < \log\left(1+\frac{1}{\delta}\right)$, i.e., for any $\delta > 0$. Which implies, $f_n(y) \leq h(y) = \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha}(1+\frac{1}{\delta^2+2\delta})^y$ and $\int_\Omega h(y)d\mu < \infty$. Also, note that, for all $y \in \Omega, \lim_{n\to\infty} f_n(y) = f(y) = \binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha}$. Define, $m(y) = \lim_{\tau\to 0}\mathbb{E}(\kappa|y,\tau)$. Hence, for all $y \in \Omega$,

$$\lim_{n\to\infty} g_n(y) = 1\{m(y) \geq \frac{\delta+2}{2(\delta+1)}, y > a\}\binom{y+\alpha-1}{y}(1-\frac{1}{\delta+1})^y(\delta+1)^{-\alpha} = g(y), \text{say.}$$

Therefore, by the Dominated Convergence Theorem,

$$\int_\Omega g_n(y)d\mu \to \int_\Omega g(y)d\mu \text{ as } n \to \infty.$$

Under Assumption 1, for $y > a$ using (5.37), we have

$$m(y) = \frac{Beta(a+\alpha+1, y-a)}{Beta(a+\alpha, y-a)} = \frac{a+\alpha}{y+\alpha}.$$

Hence, we obtain

$$\begin{aligned} \lim_{n\to\infty}\mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\tau) \geq \frac{\delta+2}{2(\delta+1)}, Y_i > a) &= \mathbb{P}(m(Y) \geq \frac{\delta+2}{2(\delta+1)}, Y > a) \\ &= \mathbb{P}(\frac{a+\alpha}{Y+\alpha} \geq \frac{\delta+2}{2(\delta+1)}, Y > a) \\ &= \mathbb{P}\left(a < Y \leq 2a+\alpha-\frac{2(a+\alpha)}{(\delta+2)}\right), \end{aligned} \tag{5.38}$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$. Finally the proof of Theorem 5.4.12 is obtained by combining (5.35), (5.36) and (5.38).

□

Proof of Theorem 5.4.14. First, let us find an expression for $\gamma_n = P(Y_i \geq 1)$. We have,

$$\begin{aligned}\gamma_n &= (1-p)\mathbb{P}_{H_{0i}}(Y_i \geq 1) + p\mathbb{P}_{H_{1i}}(Y_i \geq 1) \\ &= p[\mathbb{P}_{H_{1i}}(Y_i \geq 1) + \frac{(1-p)}{p}\mathbb{P}_{H_{0i}}(Y_i \geq 1)]\end{aligned}$$

Since, $Y_i \overset{iid}{\sim} NB(\alpha, \frac{1}{\beta+1})$, under H_{0i}, using Markov's inequality and Assumption 3, we have, $\frac{1}{p}\mathbb{P}_{H_{0i}}(Y_i \geq 1) \to 0$ as $n \to \infty$. Also, note that, under H_{1i},

$$\mathbb{P}(Y_i \geq 1) = (1 - (\beta + \delta + 1)^{-\alpha}).$$

Combining these two, we obtain, as $n \to \infty$,

$$\gamma_n = (1 - (\beta + \delta + 1)^{-\alpha})p(1 + o(1)). \tag{5.39}$$

Now, using the definition of type I error of the i^{th} decision rule in (5.17), we have

$$\begin{aligned}t_{1i}^{\text{EB}} &= \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}) \\ &= \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}) + \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, Y_i > \frac{a}{2}).\end{aligned} \tag{5.40}$$

For the second term of (5.40), using Markov's inequality, we have

$$\mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, Y_i > \frac{a}{2}) \leq \mathbb{P}_{H_{0i}}(Y_i > \frac{a}{2}) \leq \frac{2\alpha\beta}{a}. \tag{5.41}$$

For the first term in (5.40), we obtain

$$\begin{aligned}&\mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}) \\ &= \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}, \widehat{\tau} \leq 2\gamma_n) \\ &+ \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}, \widehat{\tau} > 2\gamma_n).\end{aligned} \tag{5.42}$$

Now, using the form of γ_n, we note that

$$\begin{aligned}&\mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, \widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}, \widehat{\tau} \leq 2\gamma_n) \\ &\leq \mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i, 2\gamma_n) > \frac{\delta}{2(\delta+1)}, 0 \leq Y_i \leq \frac{a}{2}) \\ &= 0, \text{ for all sufficiently large } n.\end{aligned} \tag{5.43}$$

The inequality above holds due to the fact that for any fixed $y \geq 0$, $\mathbb{E}(1-\kappa|y, \tau)$ is non-decreasing in τ. The last assertion follows from (5.33).

Now let us concentrate on the second term of (5.42). Define $\widehat{\tau}_1 = \frac{1}{n}$ and $\widehat{\tau}_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \geq 1\}$.

Hence, $\widehat{\tau} = \max\{\widehat{\tau}_1, \widehat{\tau}_2\}$. Also, note that $\gamma_n \sim (1-(\beta+\delta+1)^{-\alpha})p$ and $p \propto n^{-\epsilon}, 0 < \epsilon < 1$ implies $\frac{1}{n} < 2\gamma_n$ for all sufficiently large n. Next, we use the fact that $\{\widehat{\tau} > 2\gamma_n\}$ implies that either of the two events $\{\widehat{\tau}_1 > 2\gamma_n\}$ or $\{\widehat{\tau}_2 > 2\gamma_n\}$ will happen. Using these facts for all sufficiently large n, we have

$$\begin{aligned}
&\mathbb{P}_{H_{0i}}(\mathbb{E}(1-\kappa_i|Y_i,\widehat{\tau}) > \frac{\delta}{2(\delta+1)}, 0 \le Y_i \le \frac{a}{2}, \widehat{\tau} > 2\gamma_n) \le \mathbb{P}_{H_{0i}}(\widehat{\tau} > 2\gamma_n) \\
&\le \mathbb{P}_{H_{0i}}(\widehat{\tau}_1 > 2\gamma_n) + \mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n) \\
&= \mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n) \\
&\le \mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n, Y_i < 1) + \mathbb{P}_{H_{0i}}(Y_i \ge 1) \\
&\le \mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n, Y_i < 1) + \alpha\beta\,. \qquad (5.44)
\end{aligned}$$

Now we are only left with only the first term of (5.44). Note that

$$\{\widehat{\tau}_2 > 2\gamma_n, Y_i < 1\} \subseteq \{\frac{1}{n}\sum_{\substack{j=1\\(j\neq i)}}^{n} \mathbf{1}\{Y_j \ge 1\} > 2\gamma_n\}.$$

As a result, using the Chernoff-Hoeffding inequality, we have

$$\begin{aligned}
\mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n, Y_i < 1) &\le \mathbb{P}_{H_{0i}}(\frac{1}{n}\sum_{\substack{j=1\\(j\neq i)}}^{n} \mathbf{1}\{Y_j \ge 1\} > 2\gamma_n) \\
&\le \mathbb{P}(\frac{1}{n-1}\sum_{\substack{j=1\\(j\neq i)}}^{n} \mathbf{1}\{Y_j \ge 1\} \ge 2\gamma_n) \\
&\le e^{-(n-1)D(2\gamma_n||\gamma_n)}\,, \qquad (5.45)
\end{aligned}$$

where $D(x||y)$ is defined as, $D(x||y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$. Next using these facts that $\log\left(\frac{1}{1-x}\right) \sim x$ as $x \to 0$ and $\gamma_n \to 0$ as $n \to \infty$ and also using the calculations of Theorem 10 of Ghosh et al. (2016), we have

$$D(2\gamma_n||\gamma_n) = (2\log 2 - 1)\gamma_n(1+o(1)) = (2\log 2 - 1)(1-(\beta+\delta+1)^{-\alpha})p(1+o(1))\,. \qquad (5.46)$$

Here $o(1)$ is non-random and tends to 0 as $n \to \infty$. Hence using (5.45) and (5.46), we obtain

$$\mathbb{P}_{H_{0i}}(\widehat{\tau}_2 > 2\gamma_n, Y_i < 1) \le e^{-(2\log 2-1)(1-(\beta+\delta+1)^{-\alpha})np(1+o(1))}\,. \qquad (5.47)$$

Combining the above arguments the proof of Theorem 5.4.14 follows. □

Proof of Theorem 5.4.15. Let us fix any $C_3 \in (0,1)$. Now, using the definition of type II error

of the i^{th} decision rule in (5.17), we have

$$\begin{aligned} t_{2i}^{\text{EB}} &= \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) \geq \frac{\delta+2}{2(\delta+1)}) \\ &= \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) \geq \frac{\delta+2}{2(\delta+1)}, \widehat{\tau} < C_3\gamma_n) + \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) \geq \frac{\delta+2}{2(\delta+1)}, \widehat{\tau} \geq C_3\gamma_n)\,, \end{aligned} \tag{5.48}$$

where γ_n is same as defined in the previous theorem. Let us first bound the second term in the r.h.s. of (5.48). Using the fact that for any fixed y, $\mathbb{E}(\kappa|y,\tau)$ is decreasing in τ, towards that, we first note that

$$\{\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) \geq \frac{\delta+2}{2(\delta+1)}, \widehat{\tau} \geq C_3\gamma_n\} \subseteq \{\mathbb{E}(\kappa_i|Y_i,C_3\gamma_n) \geq \frac{\delta+2}{2(\delta+1)}\}\,. \tag{5.49}$$

Now applying the same set of arguments used in the proof of Theorem 5.4.12 and noting under H_{1i}, $Y_i \overset{iid}{\sim} NB(\alpha, \frac{1}{\beta+\delta+1})$, we obtain, as $n \to \infty$,

$$\mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,C_3\gamma_n) \geq \frac{\delta+2}{2(\delta+1)}) \leq \mathbb{P}\bigg(Y \leq 2a+\alpha-\frac{2(a+\alpha)}{(\delta+2)}\bigg), \tag{5.50}$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$. Now our aim is to show that the first term in (5.48) goes to 0 as $n \to \infty$. Note that, $\widehat{\tau}_2 \geq \frac{1}{n}\sum_{\substack{j=1\\(j\neq i)}}^{n} \mathbf{1}\{Y_j \geq 1\}$. Using this observation and noting that due to independence, the distribution of remaining Y_j's do not depend on that of Y_i, we have

$$\begin{aligned} \mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) \geq \frac{1}{2}, \widehat{\tau} < C_3\gamma_n) &\leq \mathbb{P}_{H_{1i}}(\widehat{\tau} < C_3\gamma_n) \\ &\leq \mathbb{P}_{H_{1i}}(\widehat{\tau}_2 < C_3\gamma_n) \\ &\leq \mathbb{P}_{H_{1i}}(\frac{1}{n}\sum_{\substack{j=1\\(j\neq i)}}^{n} \mathbf{1}\{Y_j \geq 1\} \leq C_3\gamma_n) \\ &= \mathbb{P}(-\frac{1}{n-1}(\sum_{\substack{j=1\\(j\neq i)}}^{n}(\mathbf{1}\{Y_j \geq 1\} - \gamma_n)) \geq (1-\frac{n}{n-1}C_3)\gamma_n)\,, \end{aligned} \tag{5.51}$$

where probability in the last statement is calculated using the marginal distribution (5.3) of $Y_j, j \neq i$. Since, $1-\frac{n}{n-1}C_3 \to 1-C_3$ as $n \to \infty$ and $C_3 \in (0,1)$, so, $1-\frac{n}{n-1}C_3 > 0$ for all

sufficiently large n. Finally using Markov's inequality,

$$\begin{aligned}
&\mathbb{P}(-\frac{1}{n-1}(\sum_{\substack{j=1\\(j\neq i)}}^{n}(\mathbf{1}\{Y_i \geq 1\} - \gamma_n)) \geq (1 - \frac{n}{n-1}C_3)\gamma_n) \\
&\leq \mathbb{P}\bigg(|\frac{1}{n-1}(\sum_{\substack{j=1\\(j\neq i)}}^{n}(\mathbf{1}\{Y_i \geq 1\} - \gamma_n))| \geq (1 - \frac{n}{n-1}C_3)\gamma_n\bigg) \\
&\leq \frac{\gamma_n(1-\gamma_n)}{(n-1)(1-\frac{n}{n-1}C_3)^2\gamma_n^2} \\
&= \frac{(1-\gamma_n)}{(1-C_3)^2 n\gamma_n}(1+o(1)) = o(1)\,. \qquad (5.52)
\end{aligned}$$

At the final step, we use $\gamma_n \sim (1-(\beta+\delta+1)^{-\alpha})p$ and $p \propto n^{-\epsilon}, 0 < \epsilon < 1$. Combining (5.51) and (5.52), we obtain, as $n \to \infty$,

$$\mathbb{P}_{H_{1i}}(\mathbb{E}(\kappa_i|Y_i,\widehat{\tau}) > \frac{\delta+2}{2(\delta+1)}, \widehat{\tau} < C_3\gamma_n) = o(1)\,.$$

Hence using (5.50), as $n \to \infty$, we have

$$t_{2i}^{\mathrm{EB}} \leq \mathbb{P}\bigg(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)}\bigg)(1+o(1))\,, \qquad (5.53)$$

since the r.h.s. of (5.50) is bounded away from zero under Assumption 3. This completes the proof of Theorem 5.4.15. □

Proof of Theorem 5.4.1. First, we study the probabilities of type I error (t_1^{BO}) and type II error (t_2^{BO}) of the decision rule when each θ_i is modeled by a two-group prior of the form (5.2). Since, under H_{0i}, $Y_i \overset{iid}{\sim} NB(\alpha, \frac{1}{\beta+1})$, using Markov's inequality and Assumption 3, t_1^{BO} can be bounded as

$$\begin{aligned}
t_1^{\mathrm{BO}} &= \mathbb{P}_{H_{0i}}(Y_i > C_{p,\alpha,\beta,\delta}) \\
&= \mathbb{P}_{H_{0i}}\bigg(Y_i > \frac{\log\left(\frac{1}{p}\right) + \alpha\log(1+\delta)}{\log\left(\frac{1}{\beta}\right)}(1+o(1))\bigg) \\
&= \mathbb{P}_{H_{0i}}(Y_i > \frac{1}{C_1}(1+o(1))) \\
&\leq \alpha\beta C_1(1+o(1))\,. \qquad (5.54)
\end{aligned}$$

Here, inequality in (5.54) holds due to Markov's inequality. Since under the assumption $\beta \propto p^{C_1}$, for some $C_1 > 1$, we have, $\frac{t_1^{\mathrm{BO}}}{p} \to 0$ as $n \to \infty$. Again using Assumption 3, we have

$$\begin{aligned}
t_2^{\mathrm{BO}} &= \mathbb{P}_{H_{1i}}(Y_i \leq C_{p,\alpha,\beta,\delta}) \\
&= \mathbb{P}_{H_{1i}}\bigg(Y_i \leq \frac{1}{C_1}(1+o(1))\bigg). \qquad (5.55)
\end{aligned}$$

Since under $H_{1i}, Y_i \overset{iid}{\sim} NB(\alpha, \frac{1}{\beta+\delta+1})$, using the probability mass function (p.m.f.) of a negative binomial random variable, we have for all sufficiently large n,

$$t_2^{\text{BO}} = \mathbb{P}_{H_{1i}}(Y_i = 0) = (\beta + \delta + 1)^{-\alpha}.$$

Hence under Assumption 3, we have, for all sufficiently large n,

$$t_2^{\text{BO}} = (\delta + 1)^{-\alpha}(1 + o(1)). \tag{5.56}$$

As a consequence of this, the asymptotic expression for the optimal Bayes risk based on the decision rule (5.7) with (5.8), denoted $R_{\text{Opt}}^{\text{BO}}$, can be written as

$$\begin{aligned} R_{\text{Opt}}^{\text{BO}} &= np[\frac{(1-p)}{p} t_1^{\text{BO}} + t_2^{\text{BO}}] \\ &= npt_2^{\text{BO}}(1 + o_1(1)), \end{aligned} \tag{5.57}$$

where the term $o_1(1)$ is non-random, independent of index i and tends to zero as $n \to \infty$. Now note that under Assumption 3, $t_1/p \to 0$ as $n \to \infty$ where t_1 is the probability of type I error corresponding to i^{th} decision rule of the form (5.15) based on our class of one-group priors, as obtained in Theorem 5.4.10. Using this and Theorem 5.4.12, the asymptotic expression for the Bayes risk based on the decision rule, denoted by R_{OG} can be written as

$$R_{\text{OG}} = npt_2(1 + o_2(1)), \tag{5.58}$$

by observing that t_2 is also bounded away from zero as in (5.56) using the same argument. Here also, the term $o_2(1)$ is non-random, independent of index i and tends to zero as $n \to \infty$. Next, we take the ratio of (5.58) to (5.57). The lower bound is obtained using the fact that $\frac{R_{\text{OG}}}{R_{\text{Opt}}^{\text{BO}}} \geq 1$. For the upper bound, observe that, the lower bound on t_2^{BO} is bounded away from zero and using Theorem 5.4.12, we have

$$\limsup_{n\to\infty} \frac{R_{\text{OG}}}{R_{\text{Opt}}^{\text{BO}}} \leq (\delta + 1)^{\alpha} \mathbb{P}\left(Y \leq 2a + \alpha - \frac{2(a+\alpha)}{(\delta+2)}\right),$$

where $Y \sim NB(\alpha, \frac{1}{\delta+1})$. □

Proof of Theorem 5.4.2. The asymptotic expression for the Bayes risk based on the decision rule (5.17), denoted by $R_{\text{OG}}^{\text{EB}}$, is of the form

$$R_{\text{OG}}^{\text{EB}} = \sum_{i=1}^{n} [(1-p)t_{1i}^{\text{EB}} + pt_{2i}^{\text{EB}}] = p \sum_{i=1}^{n} \left[\frac{1-p}{p} t_{1i}^{\text{EB}} + t_{2i}^{\text{EB}}\right]. \tag{5.59}$$

Note that, we already proved in Theorem 5.4.1 that $t_2^{\text{BO}} = (\delta + \beta + 1)^{-\alpha}$ is bounded away from 0, for all sufficiently large n. Hence, using Theorem 5.4.15 along with Assumption 3, the proof

of our desired result is obtained if we establish that

$$\sum_{i=1}^{n} \frac{1-p}{p} t_{1i}^{\mathrm{EB}} = o(n) \,\mathrm{as}\, n \to \infty\,. \tag{5.60}$$

To prove (5.60), we use Theorem 5.4.14 where, it is derived that the upper bound of t_{1i}^{EB} is the same for each $i = 1, 2, \cdots, n$. Further using the fact that $1 - p \leq 1$, we obtain,

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1-p}{p} t_{1i}^{\mathrm{EB}} \leq \frac{2\alpha\beta}{ap} + \frac{\alpha\beta}{p} + \frac{1}{p} e^{-(2\log 2 - 1)(1-(\beta+\delta+1)^{-\alpha})np(1+o(1))}\,. \tag{5.61}$$

By Assumption 3, the first two terms in the right-hand side of (5.61) go to 0 as $n \to \infty$. For the third term, note that, for $p \propto n^{-\epsilon}, 0 < \epsilon < 1$, $np \to \infty$ as $n \to \infty$ and $\log\left(\frac{1}{p}\right) = o(np)$. As a result, the third term too goes to 0 as $n \to \infty$ and this proves (5.60). This completes the proof of Theorem 5.4.2. □

Bibliography

Abraham, K., Castillo, I., and Roquain, É. (2024). Sharp multiple testing boundary for sparse sequences. *The Annals of Statistics*, 52(4):1564–1591.

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653.

Abramovich, F., Grinshtein, V., and Pensky, M. (2007). On optimality of bayesian testimation in the normal means problem. *The Annals of Statistics*, 35(5):2261–2286.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine learning*, 73:243–272.

Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001.

Armagan, A., Clyde, M., and Dunson, D. (2011). Generalized beta mixtures of gaussians. *Advances in neural information processing systems*, 24.

Armagan, A., Dunson, D. B., and Lee, J. (2013a). Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018.

Armero, C. and Bayarri, M. (1994). Prior assessments for prediction in queues. *Journal of the Royal Statistical Society: Series D (the Statistician)*, 43(1):139–153.

Banerjee, S., Castillo, I., and Ghosal, S. (2021). Bayesian inference in high-dimensional models. *arXiv preprint arXiv:2101.04491*.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (1997). Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577.

Belitser, E. and Nurushev, N. (2020). Needles and straw in a haystack: Robust confidence for possibly sparse sequences. *Bernoulli*, 26(1):191–225.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012). Bayesian shrinkage. *arXiv preprint arXiv:1212.6088*.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Cambridge University Press.

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–1579.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the benjamini-hochberg procedure with some bayesian rules for multiple testing. *arXiv preprint arXiv:0805.2479*.

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). Group inverse-gamma gamma shrinkage for sparse linear models with block-correlated regressors. *Bayesian Analysis*, 1(1).

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, pages 1685–1704.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Casella, G., Ghosh, M., Gill, J., and Kyung, M. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, 5(2):369–411.

Castillo, I., Schmidt-Hieber, J., and Van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

Castillo, I. and Szabó, B. (2020). Spike and slab empirical bayes sparse credible sets. *Bernoulli*, 26(1):127–158.

Castillo, I. and Van Der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069–2101.

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344.

Datta, J. and Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4):971–983.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):41–67.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1):1–22.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Gabcke, W. (2015). Derivation of the riemann-siegel formula with explicit estimates of its remainders. *http://dx.doi.org/10.53846/goediss-5113*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531.

Ghosh, P. and Chakrabarti, A. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis*, 12(4):1133–1161.

Ghosh, P., Tang, X., Ghosh, M., and Chakrabarti, A. (2016). Asymptotic properties of bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Analysis*, 11(3):753–796.

Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, Technical report, University of Warwick.

Griffin, J. E. and Brown, P. J. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, page 12(1):135–159.

Guo, W., He, L., and Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics*, 42(3):1070–1101.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Hamura, Y., Irie, K., and Sugasawa, S. (2022). On global-local shrinkage priors for count data. *Bayesian Analysis*, 17(2):545–564.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, 106(496):1383–1393.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hosmer, D. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4).

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684.

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). Scalable approximate mcmc algorithms for the horseshoe prior. *Journal of Machine Learning Research*, pages 21(73):1–61.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical bayes selection of wavelet thresholds. *The Annals of Statistics*, 33(4):1700–1752.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Lehmann, E. L. (1957). A theory of some multiple decision problems, i. *The Annals of Mathematical Statistics*, pages 1–25.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian analysis*, 5(1):151–170.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Liu, J. and Ye, J. (2010). Fast overlapping group lasso. *arXiv preprint arXiv:1009.0306*.

Martin, R., Mess, R., and Walker, S. G. (2017). Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206.

Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g-prior. *Annals of Statistics*, 39(5):2740–2765.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, pages 65–80.

Mukhopadhyay, M., Samanta, T., and Chakrabarti, A. (2015). On consistency and optimality of bayesian variable selection based on g-prior in normal linear regression models. *Annals of the Institute of Statistical Mathematics*, 67(5):963–997.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Papaspiliopoulos, O. and Rossell, D. (2017). Bayesian block-diagonal variable selection and model averaging. *Biometrika*, page 104(2):343–359.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Paul, S. and Chakrabarti, A. (2024). Asymptotic bayes optiamlity for sparse count data. *arXiv preprint arXiv:2401.05693*.

Paul, S. and Chakrabarti, A. (2025). Posterior contraction rate and asymptotic bayes optimality for one group global–local shrinkage priors in sparse normal means problem. *Annals of the Institute of Statistical Mathematics*, 77:787–819.

Paul, S., Ghosh, P., and Chakrabarti, A. (2025). Sharp asymptotic minimaxity for one-group priors in sparse normal means problem. *arXiv preprint arXiv:2505.16428*.

Paul, S., Ghosh, P., and Chakrabarti, A. (2026). Consistent group selection using global-local prior in high dimensional setup. *Annals of the Institute of Statistical Mathematics*, 78::399–433.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311.

Rabinovich, M., Ramdas, A., Jordan, M. I., and Wainwright, M. J. (2020). Optimal rates and trade-offs in multiple testing. *Statistica Sinica*, 30(2):741–762.

Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E., and Roth, V. (2009). The bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 881–888.

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horse- shoe prior. *arXiv preprint arXiv:1810.09004*.

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Romano, J. P. and Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, 34(4):1850–1873.

Roy, R. and Bhandari, S. K. (2024). Asymptotic bayes' optimality under sparsity for exchangeable dependent multivariate normal test statistics. *Statistics & Probability Letters*, 207:110030.

Salomond, J.-B. (2017). Risk quantification for the thresholding rule for multiple testing using gaussian scale mixtures. *arXiv preprint arXiv:1711.08705*.

Sarkar, S. K. (2007). Stepup procedures controlling generalized fwer and generalized fdr. *The Annals of Statistics*, 35(6):2405–2420.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 135–168.

Sarkar, S. K. and Guo, W. (2009). On a generalized false discovery rate. *The Annals of Statistics*, 37(3):1545–1565.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of bayesian multiple testing. *Journal of statistical planning and inference*, 136(7):2144–2162.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Som, A., Hans, C. M., and MacEachern, S. N. (2014). Block hyper-g priors in bayesian regression. *arXiv preprint arXiv:1406.6419*.

Song, Q. (2020). Bayesian shrinkage towards sharp minimaxity. *Electronic Journal of Statistics*, page 14(2).

Song, Q. and Liang, F. (2023). Nearly optimal bayesian shrinkage for high- dimensional regression. *Science China Mathematics*.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The annals of statistics*, 31(6):2013–2035.

Sun, W. and Tony Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):393–424.

Tang, X., Xu, X., Ghosh, M., and Ghosh, P. (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80(2):215–246.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic journal of statistics*, 10(1):976–1000.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11(2):3196–3225.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274.

van der Pas, S. L., Kleijn, B. J., and Van Der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wang, M. and Tian, G.-L. (2019). Adaptive group lasso for high-dimensional generalized linear models. *Statistical Papers*, 60:1469–1486.

Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4):1369.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *Australasian Joint Conference on Artificial Intelligence*, pages 229–240. Springer.

Yang, X. and Narisetty, N. N. (2020). Consistent group selection with bayesian high dimensional modeling. *Bayesian Analysis*, 15(3):909–935.

Yano, K., Kaneko, R., and Komaki, F. (2021). Minimax predictive density for sparse count data. *Bernoulli*.

Yuan, M. and Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

Zhang, C. and Xiang, Y. (2016). On the oracle property of adaptive group lasso in high-dimensional linear models. *Statistical Papers*, 57:249–265.

Zhang, C.-H. (2005). General empirical bayes wavelet methods and exactly adaptive minimax estimation. *The Annals of Statistics*, 33(1):54–100.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733.