

GENERALISED BOOTSTRAP TECHNIQUES

SNIGDHANSU BHUSAN CHATTERJEE

Thesis submitted to the Indian Statistical Institute
in partial fulfilment of the requirements for the award
of the degree of Doctor of Philosophy.

THEORETICAL STATISTICS AND MATHEMATICS UNIT
INDIAN STATISTICAL INSTITUTE
203 B. T. ROAD
CALCUTTA 700 035, INDIA

1999

Acknowledgements

This work was done under the supervision of Professor Arup Bose. I would like to record my gratitude to him for his constant encouragement, guidance and inspiration. He has been a friend and guide to me for the past few years, and I profited immensely in his company. I am grateful to him for the time he spent on discussing our research and scrutinizing it critically and with painstaking care; and also for allowing our joint works to be included in this thesis.

I shared an office with Mr. Saunak Mishra for the better part of the last four years. I am thankful to him for his companionship and encouragement. He was at receiving end of many a proof when they were at the 'brainwave' stage, and as a computer scientist must have found them quite unusual. He, along with Dr. Biman Chakraborty, Mr. Ujjwal Bhattacharya and others shared the task of maintaining and updating the softwares in the computers of the Stat-Math Unit, and I am grateful to all of them.

This thesis is written in L^AT_EX, partly on a Linux machine. The codes for simulating data and statistical analysis were sometimes compiled using gcc. Some data analysis, and the graphical displays were prepared using R. All these tools are free software. I wish to use this opportunity to acknowledge my indebtedness to everyone who has been a part of the free software movement.

I would like to thank Professor H. L. Koul and Professor A. Ivanov for reading drafts of some of the material of this thesis and commenting on them. Their comments have been most helpful.

The Institute's library staff, inmates and staff of R. S. Hostel and everyone in Stat-Math Unit have been co-operative and friendly, which helped a lot. Stat-Math Unit provided a great atmosphere for research, which made everyday worth looking forward to.

This research is supported by a grant of CSIR, India. I am thankful to them.

Contents

Acknowledgements	i
1 INTRODUCTION TO RESAMPLING TECHNIQUES	1
1.1 Introduction	1
1.2 The jackknife	2
1.3 The bootstrap and generalised bootstrap	3
1.4 An overview of Chapters 2-8	10
2 GENERALISED BOOTSTRAP TECHNIQUES	15
2.1 Introduction	15
2.2 The Uncorrelated Weights Bootstrap	16
2.3 Examples of uncorrelated weights bootstrap techniques	19
2.4 The correlated weights bootstrap	20
2.5 Weighted jackknives as generalised bootstrap	22
2.6 The generalised residual bootstrap	24
3 VARIANCE ESTIMATION IN HIGH DIMENSIONAL LINEAR REGRESSIONS: FIRST ORDER RESULTS	27
3.1 Introduction	27
3.2 Assumptions and notations	28
3.2.1 The model	28
3.2.2 The resampling scheme	30
3.3 Main Results	32
3.4 Some simulation results	34
3.5 Proofs	37
4 COMPARISON OF BOOTSTRAP AND JACKKNIFE VARIANCE ESTIMATORS IN LINEAR REGRESSIONS: SECOND ORDER RESULTS	45
4.1 Introduction	45
4.2 Notations, assumptions and preliminaries	46
4.3 Second order results for $p = 1$	49

4.4	E-class bootstrap: second order results for $p > 1$	53
4.5	R class bootstrap: second order results for $p > 1$	54
4.6	Results from simulation experiments	59
4.7	Proofs of theorems: $p = 1$ case	67
4.8	Proofs for general p	84
5	ESTIMATION OF DISTRIBUTION OF LEAST SQUARES ESTIMATOR IN HIGH DIMENSIONAL LINEAR REGRESSIONS: FIRST ORDER RE- SULTS AND DIMENSION ASYMPTOTICS	92
5.1	Introduction	92
5.2	Assumptions and main result	93
5.2.1	Conditions on <i>UBS</i> weights	93
5.2.2	Assumptions on the model	94
5.2.3	Distributional consistency result	94
5.3	Proofs	96
6	ESTIMATION OF DISTRIBUTION OF LEAST SQUARES ESTIMATOR IN LINEAR REGRESSIONS: SECOND ORDER ACCURACY RESULTS	108
6.1	Introduction	108
6.2	Notations and assumptions	109
6.3	Main results	112
6.4	Proofs	115
7	ASYMPTOTICS AND GENERALISED BOOTSTRAP FOR SOLUTIONS OF MARTINGALE ESTIMATING EQUATIONS	119
7.1	Introduction	119
7.2	Technical conditions	122
7.2.1	Conditions on <i>UBS</i> weights	122
7.2.2	Assumptions for subsection 7.3.1	123
7.2.3	Assumptions for subsections 7.3.2 and 7.3.3	124
7.2.4	Assumptions for subsection 7.3.4	125
7.3	Main Results	127
7.3.1	Asymptotics and bootstrap for $p = 1$	127
7.3.2	Asymptotic representation of variance estimator	129
7.3.3	Higher order accuracy of <i>UBS</i>	131
7.3.4	Dimension asymptotics	133
7.4	Examples of <i>UBS</i> resampling techniques	135
7.5	Examples of martingale estimating equations	138
7.5.1	Finite dimensional parameter problems	138

7.5.2	Semiparametric inference problems	143
7.6	Computational aspects and simulation results	146
7.7	Proofs	149
8	GENERALISED BOOTSTRAP FOR MINIMUM CONTRAST (M_m) ESTIMATORS	159
8.1	Introduction	159
8.2	Preliminaries	160
8.3	Bootstrap representation results for $m = 1$	162
8.4	Results for the general m case	166
8.5	Computational examples	169
	References	174

Chapter 1

INTRODUCTION TO RESAMPLING TECHNIQUES

1.1 Introduction

A typical problem in statistics is as follows: there is some observable data $\mathbf{X}_n = (X_1, \dots, X_n)$, and a parameter of interest θ which is related in such a way to the distribution of $\mathbf{X}_n$ that meaningful conclusions about θ can be drawn based on $\mathbf{X}_n$. Sometimes data $\mathbf{X}_n$ is observed keeping the objective parameter θ in mind, at other times the parameter appears while trying to 'model' the observed data.

Once the data is observed and the parameter fixed, the questions that have to be addressed are as follows:

- (I) How to estimate θ from the data $\mathbf{X}_n$?
- (II) Given an estimator $T_n = T(\mathbf{X}_n)$ of θ , how good is the estimator, by some standard of 'goodness'?
- (III) Given an estimator $T_n = T(\mathbf{X}_n)$ of θ , how to obtain confidence sets, test hypothesis and settle other such inferential questions about θ ?

Question (I) on estimating θ from $\mathbf{X}_n$ is fundamental to statistics and even now new ideas about estimation come up as new problems come across the statistical horizon. But the other questions, (II) and (III) are equally important and interesting. A traditional way to answer (II) is to report the variance or mean squared error of T_n , which itself often has to be estimated from the data. The last question may be answered once an estimate of the probability distribution of T_n is obtained. Most often, computation of the variance of T_n or its entire distribution are intractable problems. The classical approach is to use the theory of weak convergence, and approximate the distribution of T_n by its weak limit as $n \rightarrow \infty$. A major step is to obtain a result like $a_n(T_n - \theta) \Rightarrow N(0, \tau^2)$, for some $\tau^2 > 0$ and some sequence $a_n \rightarrow \infty$. Often τ^2 is unknown, and one has to use some estimate of it based on $\mathbf{X}_n$.

Example 1.1.1 Suppose X_i 's are independently and identically distributed (hereafter *i.i.d*) according to some distribution F on the real line with mean θ and unknown variance ξ^2 and

the parameter of interest is the mean θ . Suppose the estimator $T_n = \sum X_i/n$ is used. Using the independence and identical nature of X_i 's, one obtains ξ^2/n as the variance of T_n , and $n^{1/2}(T_n - \theta) \Rightarrow N(0, \xi^2)$. An estimator of ξ^2 is given by $\sum (X_i - T_n)^2/n$.

Example 1.1.2 Suppose (x_i, y_i) , $i = 1, \dots, n$ are observed pairs of real values, and the model is $y_i = \beta x_i + e_i$, $i = 1, \dots, n$. Here the parameter of interest is β and it is assumed that e_i 's are i.i.d. with mean 0 and variance ξ^2 . The 'least squared error' estimator for β is given by $T_n = (\sum x_i^2)^{-1} \sum x_i y_i$, and the variance of T_n is $(\sum x_i^2)^{-1} \xi^2$. With standard assumptions, one has that $(\sum x_i^2)^{1/2}(T_n - \beta) \Rightarrow N(0, \xi^2)$. Defining $r_i = y_i - T_n x_i$, and an estimate of ξ^2 is given by $\sum r_i^2/n - (\sum r_i/n)^2$.

In the above examples the variance or the limiting distribution was easily estimated because the form of the statistic T_n was simple. However, consider the estimator of β obtained by minimising $\sum |y_i - \beta x_i|$ in Example 1.1.2. There is no easily derived or computed variance estimate, or even asymptotic variance estimate available for that statistic. Estimators like the above occur quite naturally, and variance of such estimators have to be estimated. This is not so simple outside the realm of cases like Examples 1.1.1 or 1.1.2. Efron and Le Page (1992) say 'much of statistical history concerns finding approximations to ..(the variance).. for more general estimators ... and avoiding estimators that do not have such formulas.'

A significant breakthrough was achieved by Tukey in 1958, when he proposed a new method for estimating the variance of T_n . This is the jackknife, described below.

1.2 The jackknife

Following the initial ideas of Quenouille, Tukey proposed the following method for computing the variance of any statistic $T_n = T(\mathbf{X}_n)$.

Consider the data set $\mathbf{X}_{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ obtained by deleting the i^{th} data point X_i from the originally observed data set. Let $T_{(i)} = T(\mathbf{X}_{(i)})$, $i = 1, \dots, n$. The jackknife estimate of variance is given by

$$V_J = (n-1)n^{-1} \sum (T_{(i)} - \bar{T})^2,$$

where $\bar{T} = \sum T_{(i)}/n$. If T_n is the sample mean as in Example 1.1.1, then this formula matches with the traditional unbiased variance estimator. It was observed that the jackknife is generally a trustworthy variance estimator (Miller (1964)), except when applied to non-smooth estimators like the median where it performed poorly.

Estimating the entire distribution of T_n is a more difficult problem. With asymptotic normality, a confidence interval of θ may be given by $T_n \pm z_{\alpha/2} V^{1/2}$, where z_p is a standard normal percentile and V is a variance estimate. For studentised statistic the normal percentile is replaced by a percentile from a t distribution of appropriate degrees of freedom. As Efron and

Le Page (1992) remark, with studentised statistics the jackknife was a major disappointment in the sense that the correct degrees of freedom to be associated to V_J remained elusive. Although we have described the jackknife performed by deleting one observation from the data set, this idea can be extended to the case where d observations are deleted at a time, $1 \leq d < n$ thus yielding the delete- d jackknife. Often the upper bound on d is lower than n , for example for U -statistics with kernel of order m , d can be at most $n - m$. Incidentally, Wu (1990) established that a jackknife based on deleting d observations at a time where d/n does not converge to either zero or one, may be used for distributional approximation.

This state of affairs persisted till 1979, when statistical thinking was revolutionised by the introduction of the bootstrap, by Efron (1979). The next section discusses the bootstrap.

1.3 The bootstrap and generalised bootstrap

Efron (1979) introduced the bootstrap in an attempt to understand the jackknife better. Let us discuss the bootstrap in the framework of Example 1.1.1. In order to estimate $\theta = \int x dF(x) = T(F)$, the statistic $T(F_n) = \int x dF_n(x)$ was used, where F_n is the empirical distribution $F_n(x) = n^{-1} \sum I_{\{X_i \leq x\}}$. Then $n^{1/2}(T(F_n) - T(F))$ was studied, its variance and limiting distribution calculated. Efron suggested that in order to approximate the distribution of $\eta_n = a_n(T(F_n) - T(F))$, we may use the statistic $\eta_n^* = a_n(T(F_n^*) - T(F_n))$, where F_n^* is an empirical distribution from F_n . Conditional on the data, the distribution function F_n is completely known, and F_n^* may be obtained by simple random sampling with replacement from the data. Then conditional on the data, the distribution of η_n^* is a random variable that may be a close approximation of the distribution of η_n .

The motivation behind this novel idea is that F_n is a close approximation of F , and F_n^* relates to F_n much the same way in which F_n relates to F , so η_n^* is a data-driven imitation of η_n . This idea is beautifully explained in details in Efron and Tibshirani (1993) and Hall (1992). With this motivation in the background, it may be understood that the function F_n need not be the empirical distribution or F_n^* need not be generated the way described above; any close approximation G_n of F , and some 'empirical-from- G_n ' can replace F_n and F_n^* in the definition of η_n^* . While Tukey's idea of estimating variance using jackknife was a shift from routine statistical theorizing, the bootstrap marked the beginning of the now popular computational approach to statistics.

The important aspect of both the jackknife and the bootstrap is the repeated use of the original sample, by different processes of drawing sample-from-the-sample, or *resampling*. Thus these techniques, and their generalisations we discuss hereafter, are broadly called resampling techniques.

Two important papers came soon after the publication of Efron's ideas about the bootstrap. Bickel and Freedman (1981) showed that consistency was obtainable under very general

circumstances for many statistical functionals, and Singh (1981) established that the bootstrap approximation is often better than the classical asymptotic normal approximation, in the sense that the bootstrap distribution captures the second order term in the Edgeworth expansion of the original distribution. It is interesting to note that recently bootstrap theory and its generalisations have been used for the problem of estimation also, as in Efron and Tibshirani (1998). Thus bootstrap is a breakthrough in three aspects:

- It is a computation based technique of obtaining the distribution of a large class of statistics.
- It often produces an estimate that is a better approximation than the asymptotic normal approximation.
- New estimation techniques can be derived from it.

Over the past two decades a significant part of statistical research concentrated on broadening the class of problems on which bootstrap may be applied, and establishing that by using bootstrap, better approximations are obtainable than what is obtained using more traditional methods. Excellent review of the bootstrap and the jackknife is available in the books by Efron and Tibshirani (1993), Hall (1992), Shao and Tu (1995), Davison and Hinkley (1997), Politis, Romano and Wolf (1999) and in many other references.

A couple of examples are in order at this point to illustrate the way the bootstrap and jackknife are applied, and to introduce the concept of generalised bootstrap.

Example 1.3.1 Let $X_1, \dots, X_n$ be an *i.i.d.* sample from some probability distribution function F , and $\theta = Eh(X_1, X_2, \dots, X_m)$ be the parameter of interest. Assume that the function $h(\cdot)$ is symmetric in its arguments. The estimator for θ is the U -statistics

$$U_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m}).$$

Let us first describe the jackknife for U_n . Consider a subset of $n - d$ observations from $X_1, \dots, X_n$, say $X_{j_1}, \dots, X_{j_{n-d}}$. Consider the U -statistics formed from these observations only. Since there are $\binom{n}{d}$ ways to select the subset of observations, there are as many such U -statistics. The jackknife statistic is determined from these $\binom{n}{d}$ U -statistics.

The bootstrap scheme of Efron (1979), when applied to U -statistics, is as follows: from the data $X_1, \dots, X_n$, draw a simple random sample with replacement, say $X_1^*, \dots, X_n^*$. Now the bootstrap statistic is defined by

$$U_n^* = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} h(X_{i_1}^*, \dots, X_{i_m}^*).$$

Conditional on the data $X_1, \dots, X_n$, the distribution of the statistic U_n^* typically closely approximates that of U_n . The variance of U_n may also be estimated by using U_n^* . Conditional on the data, the random component in U_n^* is induced only by the scheme of simple random sampling with replacement from the data, or in other words, by the scheme of drawing 'sample-from-the-sample' or resample from the data. Theoretically, this is a randomisation scheme whose distribution is completely known, and thus the distribution of U_n^* can be calculated perfectly and completely. However, since this distribution is generally difficult to obtain exactly, a very routine technique used in bootstrap is to approximate it by Monte Carlo techniques. This means that the distribution of U_n^* may be approximated by repeatedly drawing $(X_1^*, \dots, X_n^*)$ from $(X_1, \dots, X_n)$, say a large B number of times, computing U_n^* for every such resample, and finally computing the histogram formed by these B values. This Monte Carlo step is important from the computational viewpoint, since it effectively transfers a problem of mathematical complications to a computational one.

Suppose w_i is the number of times the data point X_i is drawn in the bootstrap sample. Then $(w_{n:1}, \dots, w_{n:n})$ follows a *Multinomial* $(n, 1/n, \dots, 1/n)$ distribution. A different way of writing the bootstrap U -statistic is

$$U_n^* = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m}) \prod_{j=1}^m w_{n:i_j}. \quad (1.3.1)$$

Clearly, the distribution of U_n^* conditional on the data is dependent only on $(w_{n:1}, \dots, w_{n:n})$. In case of the delete- d jackknife too, the statistic can be written in the above form, for $(w_{n:1}, \dots, w_{n:n})$ following a different distribution.

Note that we might have started by defining the resample U -statistics to be (1.3.1), for some 'resample weights' $w_{n:1}, \dots, w_{n:n}$ appropriately chosen. Then the bootstrap or the delete- d jackknives would have been obtained as special cases. By doing that, we would potentially be able to do a number of things: (a) in problems where the bootstrap (or jackknife) is known to be inconsistent, weights from some other distribution might possibly achieve consistency; (b) even where the bootstrap (or jackknife) is applicable, in order to get 'the best' resample estimate one may use this broader class and get resampling approximations that are better according to some notion of efficiency; (c) in practical problems, computations with weights other than those obtained from the bootstrap or the jackknives may be faster and easier; (d) in some problems, the algebra might be easier or more insightful while working with general weights than for particular choices implied by the usual bootstrap or the jackknife; (e) one gets a deeper insight into the technical and philosophical aspects of why does the bootstrap or jackknife work in some problems but not in others; (f) it yields unified way of studying several resampling schemes simultaneously in a given problem, including typically the jackknives and most of the variations of the bootstrap found in literature.

Efron (1992) raised six questions relating to classical asymptotics and the bootstrap, and using generalised bootstrap, partial answers to two important questions may be obtained: what kind of statistics to use for testing and building confidence sets for parameters, and what are the computationally efficient ways of resampling.

We will refer to resampling with weights $w_{n:1}, \dots, w_{n:n}$ as *generalised bootstrap*. Other names by which this general form of resampling is referred to in literature is the *weighted bootstrap* (Barbe and Bertail (1995)) since the emphasis is on the 'bootstrap weights' $w_{n:i}$'s, and the *exchangeable bootstrap* (Praestgaard and Wellner (1993)) since it is customary to assume that for every fixed n , the weights $w_{n:1}, \dots, w_{n:n}$ are exchangeable. The idea of bootstrapping with random weights probably appeared first in Rubin (1981). Bootstrapping with exchangeable weights have been treated in Efron (1982), Lo (1987), Weng (1989), Zheng and Tu (1988), Praestgaard and Wellner (1993). Other generalised bootstrap methods may be found in Boos and Monahan (1986), Lo (1991), Hardle and Marron (1991), Mammen ((1992a) and (1993)). A review can be found in Barbe and Bertail (1995). Generalised bootstrap in U -statistics has been treated in Huskova and Janssen (1993a,b) and Janssen (1994). Statistical research with generalised bootstrap has largely been dominated by an approach based on empirical processes and weighted empirical processes. Important contributions in this direction include Csorgo and Mason (1989), Gine and Zinn (1989, 1990), Koul (1992), Mason and Newton (1992), Praestgaard and Wellner (1993), Mammen (1996); see Gine (1997) for a review. In this thesis we do not consider the empirical process approach because of several reasons. In Chapter 2 we discuss these reasons in details.

The form (1.3.1) is not the only way to do generalised bootstrap in U -statistics. A very general form of (1.3.1) is given by $U_n^* = \binom{n}{m}^{-1} \sum h(X_{i_1}, \dots, X_{i_m}) w_{n:i_1 \dots i_m}$, where $w_{n:i_1 \dots i_m}$ is some triangular sequence of random variables. In particular, one may assume that

$$w_{n:i_1 \dots i_m} = m^{-1} \sum_{j=1}^m w_{n:i_j}.$$

to yield generalised bootstrap statistic

$$U_{Bn} = \binom{n}{m}^{-1} m^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} h(X_{i_1}, \dots, X_{i_m}) \sum_{j=1}^m w_{n:i_j}. \quad (1.3.2)$$

We have not encountered such weights in bootstrap literature. But (1.3.2) has interesting possibilities. It suggests new ways of 'bootstrapping' and 'jackknifing' U -statistics. For the delete-1 jackknife, consider η_i to be the vector whose i^{th} entry is zero and the rest are one. The jackknife weights are (properly scaled) η_i 's, $i = 1, \dots, n$, with $1/n$ probability each. Now notice that in using (1.3.2), the notion of 'deleting the i^{th} observation' is lost, since if i is one of $i_1, \dots, i_m$, then that $h(X_{i_1}, \dots, X_{i_m})$ is merely downweighted, but not dropped entirely from the expression of the U -statistics. It is also easy to see that (1.3.2) is computationally much

simpler, less time and space consuming than (1.3.1), thus illustrating the computational aspect of using generalised bootstrap.

Example 1.3.2 Consider the linear regression model

$$y_i = \mathbf{x}_i^T \beta + e_i, \quad i = 1, \dots, n$$

The parameter $\beta \in \mathbb{R}^p$ is unknown and suppose it is estimated by the least squares estimator $\hat{\beta}_n = (\sum \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum \mathbf{x}_i y_i$. There are two different bootstrap techniques for linear models, the so called *residual bootstrap* and the *paired bootstrap* (see Efron (1982)).

In the case of the residual bootstrap, a sample of size n is drawn with replacement from the centered residuals $r_i - \bar{r}$, say $r_1^*, \dots, r_n^*$. Here $r_i = y_i - \mathbf{x}_i^T \hat{\beta}_n$ and $\bar{r} = \sum r_i / n$. The i^{th} bootstrap data point is defined by $y_i^* = \mathbf{x}_i^T \hat{\beta}_n + r_i^*$. The bootstrap estimate of the parameter is calculated by applying least squares to this data, pretending that $\hat{\beta}_n$ is unknown.

For paired bootstrap, a sample $(\mathbf{x}_1^*, y_1^*), \dots, (\mathbf{x}_n^*, y_n^*)$ of size n is drawn with replacement from the observed data $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$. The bootstrap model is $y_i^* = \mathbf{x}_i^* \beta + e_i^*$, and the bootstrap estimator is obtained by using least squares on this.

The residual bootstrap is extensively studied by Freedman (1981, 1984), Bickel and Freedman (1983), Navidi (1989), Lahiri (1992) among others. Among its many nice properties are that the bootstrap distribution is consistent even when the dimension p of the parameter tends to infinity with the data size n , as long as $p/n \rightarrow 0$. This shows that bootstrap is consistent under high dimensionality. Other properties are that for fixed p , the bootstrap distribution estimate is second order accurate, thus it is a better approximation than asymptotic normality.

Liu and Singh (1992) showed two properties of this bootstrap for the problem of estimating the variance of $\hat{\beta}_n$. The first is that it is inconsistent if the errors e_i 's are heteroscedastic, and the second is that it is more efficient than the delete-1 jackknife or the paired bootstrap for estimating the variance if the e_i 's are homoscedastic.

In an important and significant contribution, Lahiri (1992) showed that the residual bootstrap with some modifications is second order accurate for M -estimators in multiple linear regression.

For the paired bootstrap, consistency of the distribution estimate was obtained under very general conditions in Mammen (1993), when the regressors are random and $p^4/n^3 \rightarrow 0$. Liu and Singh (1992) showed that for fixed p and non-random regressors, the paired bootstrap can consistently estimate the variance of $\hat{\beta}_n$ even when e_i 's are heteroscedastic, but as already pointed out, it is less efficient than the residual bootstrap if e_i 's are homoscedastic. It was understood from heuristic calculations and other considerations that the paired bootstrap distribution estimator is not second order accurate (Mammen (1993), Hall (1992), Hall and Mammen (1994)).

It seemed that the residual bootstrap had a very definite edge over the paired bootstrap because of its second order accuracy. Moreover, since it only involved replacement of the un-

known parameters by their estimates, and noise by the estimated noise terms, it is potentially applicable to many other statistical problems. The residual bootstrap also has less computational requirement in some problems. For example in linear regression, the matrix $(\sum \mathbf{x}_i \mathbf{x}_i^T)^{-1}$ may be used for every bootstrap sample, but for the paired bootstrap, one need to calculate $(\sum \mathbf{x}_i^* \mathbf{x}_i^*)^{-1}$ for every bootstrap sample. It is not even guaranteed that such matrix inverses always exist. As mentioned in Example 1.3.1, the Monte Carlo method of computation of the bootstrap approximations is often unavoidable, so this meant that a matrix inversion was needed for every Monte Carlo step. The bootstrap technique is two decades old, and except for the last few years the computational tools at hand were neither remarkably fast nor were they as widely available worldwide as they are today, so the issue of limiting the amount of computation was important. Also, in more general problems, there is some ambiguity about the notion of 'data pairs', and the paired bootstrap heavily depends on the definition of data pairs.

All these lead to preferring the residual bootstrap over the paired bootstrap, and it seemed that its only shortcoming was not being consistent under heteroscedasticity. This issue was addressed early by Wu (1986), Liu (1988) and Mammen (1992a,b,1993,1995) with a modification of the usual residual bootstrap, called the external or wild bootstrap. This technique is as follows:

Consider an *i.i.d.* sample $t_1, \dots, t_n$ from some distribution H with mean zero and variance one, and define $r_i^* = t_i r_i$, and $y_i^* = \mathbf{x}_i^T \hat{\beta}_n + r_i^*$. The rest is as in the case of the classical residual bootstrap. Consistency of variance estimation under heteroscedasticity (Wu (1986), Liu and Singh (1992)), and of distribution estimation under very general conditions (Mammen (1993)) was obtained; and over the years this form of bootstrap has been used in a number of problems, see <http://statlab.uni-heidelberg.de/reports> for some examples. It may be seen that the external bootstrap is also a kind of generalised bootstrap.

The various factors mentioned above resulted in the attention of the statistical community somewhat shifting away from resampling based on data pairs. The various advantages of using residual based techniques has lead to a default understanding that for any problem at hand, a bootstrap technique must mean a residual based technique. Yet it is not necessarily true that a residual based technique is simpler or better than a non-residual based technique. Even for the simple example of least squares in linear regression, using uncentered residuals would generally result in incorrect results. Modifications required for problems like that of M -estimators are often not very simple, see Lahiri (1992).

The fact remained that a number of resampling techniques used data pairs, and that it was potentially applicable to a very wide variety of problems. In a number of very diverse problems like resampling in non-stationary time series, degenerate U -statistics (see discussion by Arcones and Gine (1992a)) and a variety of other problems, bootstrap with resample size m

instead of n was found to work, where $m/n \rightarrow 0$. In the context of regression, such bootstrap uses data pairs, there being no natural notions of ‘residual bootstrap’ with only m observations instead of n , which retains the original design exactly. It can also be seen that in regression the delete- d jackknives also use the data pairs. The m out of n bootstrap and the delete- d jackknife have received separate attention, see Bickel and Sakov (1999) for the former and the works of Shao, Wu and others (referenced in Shao and Tu (1995), Shao and Chen (1998) and elsewhere) for the latter. However, the underlying unity among these resampling techniques does not seem to have received enough attention. Also, the use of residuals is not recommendable everywhere. We have already commented on the fact that the classical residual bootstrap is not consistent under heteroscedasticity. Mammen (1996) showed that in high dimensional regression models, if the estimator is obtained under wrong assumptions, the empirical process of residuals tends to be biased towards the wrong assumption. This is true even if the linear fit from the ‘wrong’ estimator is consistent, see (A5) in Mammen (1996). Loosely speaking, this reflects the phenomenon that the residuals reflect the assumptions and not the nature of the data. Since the residuals are sensitive to the assumptions, bootstrap based on residuals is also expected to exhibit such sensitivity.

Our work is mainly on a new class of generalised bootstrap techniques, which may be seen as generalisation of the resampling methods based on data pairs. In linear regression, this new class appropriately captures the idea of a generalised bootstrap based on data pairs. The proposed generalised bootstrap extends naturally to more complicated problems where a notion of a ‘data pair’ is no longer easily and uniquely defined. The central theme of this thesis is the robustness of such techniques. Thus in the subsequent chapters we consider models that are very broad based and study the new generalised bootstrap scheme on these. Resampling techniques are used either for the problem of variance estimation of some statistic, or for the problem of estimating the distribution of a statistic. We extend the known results of standard data pair based resampling schemes to this new class of generalised bootstrap techniques. We compare the new generalised bootstrap with other resampling schemes, and in this context also develop a further new generalised bootstrap, based on residuals. The proposed new techniques are discussed in details in Chapter 2. However, the new residual based generalised bootstrap is not studied in details, because of its lack of robustness.

We show that at least for some members of the new generalised ‘paired’ bootstrap, a second order accuracy of distribution approximation is obtainable after some bias correction, thereby addressing the major criticism that data pair based methods lack higher order accuracy. We establish consistency results for the new class of resampling techniques for a very wide variety of problems and models, some of which have not been considered in resampling literature before. We consider estimators obtained by solving estimating equations or minimising functionals, both smooth and non-smooth, and show that the new generalised bootstrap is typically consistent. Lastly, we discuss the computational aspects of using the proposed generalised bootstrap.

There are several data examples discussed throughout this thesis. These are aimed at establishing two points, that the new generalised bootstrap may be considered in fairly complicated problems, and there is a computational edge of some members of the new class of generalised bootstrap over other techniques.

1.4 An overview of Chapters 2-8

In Chapter 2 the new generalised ‘paired’ bootstrap technique is first introduced. We call this the uncorrelated weights bootstrap (*UBS*). Some examples of the *UBS* are discussed. It is then established that the different weighted jackknife variance estimators are of the same form as the *UBS* variance estimator, but the technical assumptions that we make on the *UBS* later on are generally not satisfied for some of these weighted jackknives. Another related technique called correlated weights bootstrap, which is a generalisation of the block bootstrap, is also discussed. Then the new generalised residual bootstrap is introduced, and some interesting aspects about it are mentioned.

In Chapters 3 to 6 we revisit the problem of least squares in linear regression. Chapters 3 and 4 are on estimating the variance of the least squares estimator, and Chapters 5 and 6 on estimating its distribution. The linear regression model is given by $y_{ni} = \mathbf{x}_{ni}^T \beta + e_{ni}$, where for each n the observations are the pairs $\{(y_{ni}, \mathbf{x}_{ni}), i = 1, \dots, n\}$, $\beta \in \mathbb{R}^p$ is the unknown parameter and $\{e_{ni}, i = 1, \dots, n\}$ are the noise terms. The regressor or design variables $\mathbf{x}_{ni}$ may or may not be random. High dimensionality is captured by allowing the parameter dimension p to tend to infinity with the data size n . The errors $\{e_{ni}, i = 1, \dots, n\}$ are random variables which can have the same variance (homoscedastic) or may have different variances (heteroscedastic). An estimate of β is obtained by minimising the sum of squared errors $\sum_{i=1}^n (y_{ni} - \mathbf{x}_{ni}^T \beta)^2$, to obtain the least squares estimator $\hat{\beta} = \hat{\beta}_n = (\sum_{i=1}^n \mathbf{x}_{ni} \mathbf{x}_{ni}^T)^{-1} \sum_{i=1}^n \mathbf{x}_{ni} y_{ni}$. The matrix inverse is assumed to exist always.

Liu and Singh (1992) showed that resampling schemes satisfy one of the following

$$n^{3/2}(V_B - V_n) = Y_n + o_P(1) \quad (1.4.1)$$

$$n^{3/2}(V_B - V_n) = Z_n + o_P(1) \quad (1.4.2)$$

where V_B denotes the resampling estimate of $V_n = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$. The definitions of Y_n and Z_n are given later. Typically, Y_n and Z_n will have limiting distributions. If all the error variances are equal then all the resampling schemes discussed give consistent estimate of the variance of $\hat{\beta}$. However, those that satisfy (1.4.1) have greater asymptotic relative efficiency compared to those that satisfy (1.4.2). On the other hand, resampling schemes that satisfy (1.4.2) are consistent estimates also under heteroscedasticity, but not those that satisfy (1.4.1). Liu and Singh (1992) thus classified resampling schemes as belonging to either *E*-class (*efficient*) or *R*-class (*robust*) according to whether they satisfy (1.4.1) or (1.4.2).

Chapter 3 is on an asymptotic representation of the *UBS* variance estimator similar to (1.4.2) under very general model assumptions, where the number of parameters can tend to infinity, regressors can be random, and the errors may be heteroscedastic and may be inter-dependent. When $p \rightarrow \infty$ the rate at which it may increase compared to data size n is an important issue. It is shown that the best possible rate for variance estimation, $p^2/n \rightarrow 0$ is attainable. Some simulation based results on the performance of different *UBS* techniques are discussed. This chapter is based on Chatterjee (1998) and Chatterjee and Bose (1999a). The asymptotic representation in Chapter 3 shows that the *UBS* belongs to *R*-class. The *R*-class representation is already known for the paired bootstrap and the delete-1 jackknife. Our result extends this in two ways: (i) the same representation holds for many other resampling techniques, notably the delete- d jackknife for general d ; and (ii) the representation holds under very general model conditions, so that the same representation is valid for many time series models, dynamic linear models and many parameter regression models.

In Chapter 4 we make a deeper analysis of (1.4.1) and (1.4.2) by obtaining higher order expansions. This leads to a finer classification of schemes within the *E*-class and *R*-class. We make a thorough study of all resampling schemes and their possible generalisations based on these higher order expansions. Based on asymptotic mean squared error comparison, it is shown that a hitherto unknown member of the class of *GBS* techniques is the best in *E*-class, and a *UBS* scheme is best in *R*-class. Some data examples are reported which show that the higher order terms can make significant difference in small and medium sized data. This chapter is based on Chatterjee and Bose (1997, 1998a,b).

In Chapter 5 we move from the problem of estimating the variance of the least squares estimator in linear regression to that of estimating its entire distribution. Consider the model $y_{ni} = \mathbf{x}_{ni}^T \beta + e_{ni}$, where for each n , the data pairs $(y_{ni}, \mathbf{x}_{ni})$ are independent, identically distributed random vectors. This model has also been used by Mammen (1993) for studying the paired bootstrap and the external bootstrap, and allows for random regressors and for the dimension to tend to infinity with data size, thus capturing high dimensionality. This model does not imply the traditional assumption that given the regressors, the errors are conditionally independent and identically distributed with mean zero. We establish the consistency of *UBS* for estimating the distribution of the least squares estimator of β . It is shown that there is a *UBS* scheme that is consistent under $p/n \rightarrow 0$, although for most *UBS* schemes $p^2/n \rightarrow 0$ is required for consistency. This improves the hitherto known 'best' rate of $p^4/n^3 \rightarrow 0$ obtained by Mammen (1993) for the paired bootstrap. This chapter is based on Chatterjee and Bose (1998c).

The *UBS* estimator is typically biased for the least squares estimator in linear regression. This bias is not negligible in second order calculations, and this has been an important reason why data pair based resampling techniques lack second order accuracy. In Chapter 6 we consider a bias corrected bootstrap statistic for a sub-class of *UBS* techniques, and establish its second

order accuracy. The second order consistency result is also obtained for external bootstrap. To obtain these results we have established and used a new Edgeworth expansion result for studentised sample average of independent (but not necessarily identically distributed) random variables. This chapter is based on Chatterjee and Bose (1999c).

The assumptions on the linear regression model in Chapter 6 allow for heteroscedasticity and random regressors. However, we feel that the model conditions can be further relaxed. Improvements are possible by relaxing moment assumptions on the noise terms as well as on the design variables, which may be random. Further relaxation is possible if the Edgeworth expansion result, can be proved under weaker assumptions. This will broaden the class of models as well as resampling techniques on which our results may hold. These are important issues that we have not pursued in this thesis.

The results of the above chapters serve as the foundation for more general problems that we consider in Chapters 7 and 8. In Chapter 7 we consider bootstrapping estimators obtained by solving $\sum \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) = 0$. A basic assumption is that $\sum \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta_0)$ is a martingale, where β_0 is the true parameter value. We also assume that the functions ϕ_{ni} are smooth in the argument β . The bootstrap is done by solving $\sum w_{ni} \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) = 0$.

First we establish that the *UBS* is consistent for estimating the distribution of the estimator if the parameter is real valued, later extending to the case when the dimension tends to infinity with data size. We aim at using minimal assumptions for these consistency results, in order to emphasise the robustness and versatility of the *UBS* techniques.

By strengthening the martingale structure to an independence structure, we obtain asymptotic representation of variance estimators as in Chapter 3. This establishes that the *UBS* yields consistent variance estimator under heteroscedasticity. It is observed that in some *M*-estimation problems in linear regression, the *UBS* variance estimator can be more efficient than the weighted jackknife of Liu and Singh, which is consistent only under homoscedasticity. Thus the *UBS* is always robust, and sometimes it is efficient also, and the classification of Liu and Singh (1992) does not hold in general *M*-estimation problems.

We also establish the second order accuracy of the studentised, bias corrected *UBS* estimator for estimating the distribution of the bias corrected, studentised statistic as in Chapter 6. Several examples of statistics obtained by solving martingale equations are discussed. We show how standard assumptions in some problems, like *M*-estimation, nonlinear regression or generalised linear models, are sufficient conditions for our technical assumptions. One important feature of the present set-up is that one may consider models with nuisance parameters, even if such parameters are infinite dimensional ones. Our treatment is essentially based on the profile of the parameter of interest, and suitable model assumptions allow us to use plug-in estimates of the nuisance parameters. An important class of problems is the estimator of the parametric part in semiparametric problems, and our results may be used here for asymptotics and bootstrap for such estimators. The problem of co-integrating time series is also discussed

in details.

The computational complexity of resampling is an important aspect. Based on some heuristic considerations, we suggest a one-step non-iterative method of obtaining the bootstrap estimator. When applicable this greatly reduces the computational burden. An example based on simulated data from a nonlinear regression problem is presented, which shows that the non-iterative technique produces a good *UBS* approximation in some problems. Parts of Chapter 7 is based on Chatterjee and Bose (1999b).

In Chapter 8 we consider minimum contrast estimators or M_m estimators obtained by minimising U -functionals that are convex in the parameter, which are not necessarily smooth. That is, the parameter is $a_* \in \mathbb{R}^d$ where $Ef(a_*, Z_1, \dots, Z_m) = \min_a Ef(a, Z_1, \dots, Z_m)$ for some real function f which is convex in a , and $Z_1, \dots, Z_m$ are i.i.d. copies of some random variable over an arbitrary probability space. For an i.i.d. sample $Z_1, \dots, Z_n$, defining

$$Q_n(a) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} f(a, Z_{i_1}, Z_{i_2}, \dots, Z_{i_m}),$$

the minimum contrast estimator is obtained as a_n where $Q(a_n) = \min_a Q_n(a)$. The important difference between the models of Chapter 7 and 8 are that in the former the criterion function is smooth in the parameter, in the latter it may be non-smooth but it is convex. Also, in Chapter 7 a general martingale structure is used, but in Chapter 8 we use independent and identically distributed random variables.

The *UBS* technique in Chapter 8 is slightly complicated since U -functionals are involved, but it is a natural extension of the ideas used in previous chapters. It is defined as the minimiser of the randomly weighted U -functional $\binom{n}{m}^{-1} \sum w_{n:i_1, i_2, \dots, i_m} f(a, X_{i_1}, X_{i_2}, \dots, X_{i_m})$ where $w_{n:i_1, i_2, \dots, i_m}$ are the 'bootstrap weights'. The classical methods of jackknifing and bootstrapping U -statistics as well as the generalised bootstrap for U -statistics studied by Huskova and Janssen (1993a,b) may be seen as special cases of the *UBS*.

In this set-up, the consistency of *UBS* distribution and variance estimators is established. We discuss some new forms of resampling for U -functionals, and it is seen that an unknown *UBS* technique has remarkable advantages in terms of computational time and memory storage space requirement. Several data examples are considered to compare the performance of the new technique and more traditional methods of resampling, which establish that the accuracies of approximation of the new and traditional methods are similar, and the new method has the edge because of its less computational requirement. We analyse a data set whose earlier statistical analysis was rather involved. We show that a simple *UBS* bootstrap hypothesis test based on multivariate medians can be done. These examples show that the new generalised bootstrap may be considered in fairly complicated problems in a simple way, and there is a computational edge of some *UBS* resampling technique over other techniques.

The problem of higher order accuracy of bootstrap for minimum contrast estimators is not addressed in this thesis, and we intend to do so in the future. This is a delicate problem. From Arcones and Gine (1992a) and some of the results of Jurekova and Sen (1996) it is apparent that the second order results are sensitive to the assumptions on the smoothness of f and the dimension of the parameter. The problem of degenerate U -statistics have also not been addressed.

Chapter 2

GENERALISED BOOTSTRAP TECHNIQUES

2.1 Introduction

In the introduction we have mentioned that generalised bootstrap may be carried out using either residuals or data pairs. In this chapter we first introduce the new class of resampling techniques that may be viewed as generalised bootstrap using data pairs. This new bootstrap called the *uncorrelated weights bootstrap (UBS)* is our major object of study in the subsequent chapters. Following a broad discussion of the scheme in Section 2.2, in Section 2.3 we indicate how some standard resampling techniques may be viewed as special cases of this scheme, which we call the *uncorrelated weights bootstrap*. A related generalised bootstrap technique, the *correlated weights bootstrap*, is discussed in Section 2.4. It is more suited for dependent data, where any underlying independent mechanism is absent. This bootstrap is not considered in this thesis.

In Section 2.5, which is adapted from Chatterjee (1998), we discuss some standard weighted jackknife techniques that are used in the context of variance estimation in linear regression. An important feature of these weighted jackknives is that the weights assigned to each data point are functions of the design variables. We show that the weighted jackknife variance estimators are of the same form as the uncorrelated weights bootstrap variance estimator. However, the technical conditions that are subsequently used on the uncorrelated weights bootstrap are generally not satisfied for some of the weighted jackknives.

Then in Section 2.6 we discuss in details different generalised bootstrap techniques using residuals in linear regression. The generalised residual bootstrap is not studied in details in this thesis, except in Chapter 4. Also, in Chapter 6, we have established second order accuracy of the external bootstrap estimator for the distribution of the least squares estimator in linear regression. The external bootstrap is a generalised residual bootstrap, and along with the classical residual bootstrap of Efron (1979), probably is the most commonly studied resampling technique. Section 2.6 is adapted from Chatterjee and Bose (1998a, 1998b).

There seems to be a diversity in the behaviour of different members of the class of generalised residual bootstrap. We discuss this issue in details in the context of estimating the variance of the least squares estimator in linear regression. The classification of resampling techniques

based on asymptotic representation introduced by Liu and Singh (1992), briefly mentioned in Chapter 1, is used to analyse this diversity. In Section 3.1 of Chapter 3 we discuss these asymptotic representations in details. The results of Chapter 3, Chapter 4 and subsection 7.3.2 of Chapter 7 are motivated from such asymptotic representations.

2.2 The Uncorrelated Weights Bootstrap

The new generalised bootstrap scheme based on data pairs is perhaps best introduced through least squares estimator in linear regression. Let $y_i = \mathbf{x}_i^T \beta + e_i$ and $\hat{\beta}_n = (\sum \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum \mathbf{x}_i y_i$.

Let $\{w_{ni}; 1 \leq i \leq n, n \geq 1\}$ be a triangular array of non-negative random variables to be used as weights. The resampling scheme is carried out by weighting each data point $(y_i, \mathbf{x}_i)$ with the random weight $\sqrt{w_{ni}}$, then computing the statistic of interest and taking expectation with respect to the random weight vector.

In this new scheme the original data is transformed, so that the i^{th} bootstrap data point is $(y_i^*, \mathbf{x}_i^*) = w_{ni}^{1/2}(y_i, \mathbf{x}_i)$. That is, in order to get the bootstrap least squares estimator of the parameter, minimise with respect to β

$$\sum_{i=1}^n (y_i^* - \beta^T \mathbf{x}_i^*)^2 = \sum_{i=1}^n w_{ni} (y_i - \beta^T \mathbf{x}_i)^2 \quad (2.2.1)$$

Notice that in the bootstrap world (2.2.1) works conditional on the originally observed data $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$, so the only source of randomness is in the bootstrap weights $\{w_{ni}\}$'s. The 'bootstrap normal equations', that is, estimating equations that are solved in order to obtain the bootstrap estimator are

$$\sum_{i=1}^n w_{ni} \mathbf{x}_i (y_i - \mathbf{x}_i^T \beta) = 0 \quad (2.2.2)$$

For more general problems, where the notion of a data pair is not necessarily present, one still can carry out the above bootstrap by formally extending either (2.2.1) or (2.2.2). Thus, very generally, for data $\{X_{ni}\}$ and parameters obtained by minimising

$$\sum_{i \in I} \Psi_{ni}(X_{ni}, \lambda),$$

the bootstrap can be done by minimising

$$\sum_{i \in I} w_{ni} \Psi_{ni}(X_{ni}, \lambda).$$

where I is some indexing set, λ is the parameter of interest and Ψ_{ni} some appropriate sequence of functions. Similarly, for parameters obtained by solving

$$\sum_{i \in I} \psi_{ni}(X_{ni}, \lambda) = 0,$$

the bootstrap estimates can be obtained by solving

$$\sum_{i \in I} w_{ni} \psi_{ni}(X_{ni}, \lambda) = 0. \quad (2.2.3)$$

Several known resampling techniques are special cases of our set-up. In Section 2.3 we discuss these schemes. It can be seen that this new bootstrap is really ‘bootstrapping equations’ rather than ‘bootstrapping data points’. The idea of bootstrapping equations probably appeared first in Freedman and Peters (1984) in the context of regression with econometric variables. A similar idea appeared later on also in Rao and Zhao (1992).

Resampling with general weights have been considered earlier. A review upto 1995 may be found in Barbe and Bertail (1995). Much of the literature in this topic is based on empirical processes and results derived therefrom; for example, Praetgaard and Wellner (1993), Mason and Newton (1992). Typically, the ‘bootstrap weights’ $\{w_{ni}\}$ that can be derived from considering classical resampling schemes are at the least exchangeable, moreover, some of them can be written as $w_{ni} = nX_{ni}/(\sum X_{ni})$ for some sequence of *i.i.d.* random variables X_{ni} . Empirical processes typically require some nice structure on the random variables studied, and exchangeability or independence of the random variables is of considerable technical importance.

Empirical process techniques are powerful and are applicable in very complicated situations. However, in our research we do not use empirical process based techniques. Some of our results are on variance estimation where exchangeability of the weights are not necessary, and using empirical processes seems difficult in such situations. Empirical process based methodology also does not appear to be suitable for higher order accuracy results, where the more classical approach of Edgeworth expansions may be used. Empirical process based methodology is typically used to obtain consistency uniformly over some parameter space. For this, the parameter space often has to satisfy certain restrictions, usually in terms of the nature of an envelope function. The conditions that are typically used for more classical methods are often more broad based. Thus the models considered in this thesis are more diverse, but the results are on pointwise consistency in place of the uniform consistency that one gets by using empirical processes. For most statistical problems the pointwise consistency is enough. For example, when considering linear regression, we generally allow random regressors, heteroscedastic errors and dimension of parameter increasing with data size. It is difficult to consider all these simultaneously in a single empirical process framework. In all the subsequent chapters we obtain some interesting properties of resampling schemes using weights that typically do not satisfy the conditions that empirical process based methodology would require.

The bootstrap weights $\{w_{ni}\}$ has to satisfy certain conditions in order to obtain various consistency results. These requirements depend on the problem at hand. The first kind of conditions are on the various lower order moments and cross moments of w_{ni} ’s. The second kind of conditions is variations of the condition that w_{ni} ’s are non-negative. For example, a condition may be a bound on the probability that a large number of w_{ni} , $i = 1, \dots, n$ are zero.

Another restriction could be to require $P[w_{ni} > 0] = 1$ for all i and n , for some constant k . The third kind of conditions relates to the interrelationship among the w_{ni} 's. For example, that $\{(w_{n1}, \dots, w_{nn}), i = 1, \dots, n\}$ are exchangeable, or that they are independent and absolutely continuous.

We look at three different kinds of problems in subsequent chapters. The first is the asymptotic behaviour of such bootstrap variance estimators through asymptotic representation of the bootstrap variance. For this, various moment conditions on the weights are needed, as also some support restrictions. Second, we look at consistency of the bootstrap for estimating the entire distribution of the parameter estimator. Here moment conditions are slightly different and relaxed, conditions on the support of the distribution is minimal, but additionally, a row-wise exchangeability of w_{ni} is required. Third is the higher order accuracy of such bootstrap distribution as estimators of the distribution of the parameter estimator. Here moment conditions and conditions on the support are more stringent, and additionally, it is assumed that w_{ni} 's are *i.i.d.* absolutely continuous with a condition on the common density. The conditions for second order accuracy of distribution may be relaxed, but we do not attempt it here.

Since certain moment conditions on the weights appear repeatedly, we may discuss them here. Let $Var(w_{ni}) = \sigma_n^2$ and assume that the quantities

$$E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

are functions of the powers $i, j, k \dots$ only, and not of the indices $a, b, c \dots$. Thus we can write

$$c_{ijk\dots} = E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

Note that if the weights are assumed to be exchangeable, then the above condition follows. But exchangeability of weights is not a necessary condition. The weights are *always* assumed to satisfy the following conditions, irrespective of the particular problems we address:

$$Ew_{ni} = 1, \quad (2.2.4)$$

$$\sigma_n^2 > 0, \quad (2.2.5)$$

$$c_{11} = O(n^{-1}), \quad (2.2.6)$$

These restrictions are on the first two moments of w_{ni} 's. In generalised bootstrap literature, one commonplace assumption is $\sum_{i=1}^n w_{ni} = n$, see Mason and Newton (1992) and Hall and Mammen (1994) for example. By a simple algebra this implies $c_{11} = -(n-1)^{-1}$, and hence (2.2.6) is satisfied.

Assumption (2.2.6) implies that the weights $\{w_{ni}\}$ are asymptotically uncorrelated. Hence we call our proposed generalised bootstrap scheme the **Uncorrelated Weights Bootstrap** (hereafter *UBS*). In the discussion of the correlated weights bootstrap later on, the importance of (2.2.6) is further elucidated.

2.3 Examples of uncorrelated weights bootstrap techniques

In this section we discuss some special cases of *UBS* resampling techniques. In each such case, we identify a random mechanism, and show how the *UBS* conditions imposed in subsequent chapters, relate to it. In all the examples we consider, the weights are exchangeable, sometimes they are independent. A more detailed discussion of these resampling schemes, along with the technical conditions they do or do not satisfy, is given in Chapter 7.

Example 2.3.1 THE CLASSICAL BOOTSTRAP Suppose $\mathbf{w}_n = (w_{n:1}, \dots, w_{n:n})$ is a random sample from *Multinomial*($n, 1/n, \dots, 1/n$). This is essentially the classical bootstrap of Efron (1979).

Example 2.3.2 THE BAYESIAN BOOTSTRAP Consider $\mathbf{w}_n$ to be a random sample from *Dirichlet*($\alpha, \dots, \alpha$) distribution. Dirichlet weights appear in generalised bootstrap context in two ways, first as an absolutely continuous version of the weights of Example 2.3.1, second as weights for Bayesian bootstrap when Dirichlet process prior is used (Rubin (1981), Lo (1991)). Absolutely continuous weights are often used for showing higher order accuracy of generalised bootstrap (Zheng and Tu (1988)).

Example 2.3.3 THE DELETE- d JACKKNIVES For some integer d in $\{1, \dots, n\}$, consider the n dimensional vectors $\eta_{n:i_1, i_2, \dots, i_d}$ defined as follows: The j^{th} coordinate of $\eta_{n:i_1, \dots, i_d}$ is 0 if j is one of $i_1, \dots, i_d$, else it is $n/(n-d)$. Thus there are $\binom{n}{d}$ such vectors. Consider w_n to be a sample from the set of η , each of the $\binom{n}{d}$ vectors having the same probability of being selected. Such weights can be identified with the delete- d jackknife weights. The idea of associating a random variable and taking expectations with respect to it for calculating jackknife estimators is not common, for a different treatment see Praestgaard and Wellner (1993).

Example 2.3.4 THE BOOTSTRAP CLONE METHOD Consider the weights w_{ni} to be *i.i.d.* random variables, satisfying certain moment and support restrictions. Absolute continuity and restrictions on the density of the weights may also be imposed.

Very often in place of *i.i.d.* random variables, weights of the form $w_{ni} = nZ_{ni}/(\sum Z_{ni})$ are used, where for every n the random variables $Z_{n1}, \dots, Z_{nn}$ are *i.i.d.* absolutely continuous. This is the 'bootstrap clone' method of Lo (1991). In (2.2.3), in most situations one may replace such w_{ni} by X_{ni} without loss. This covers the case of Dirichlet weights also, discussed in Example 2.3.2.

Example 2.3.5 THE WEIGHTED LIKELIHOOD BOOTSTRAP In case $\Psi_{ni}(\cdot)$ has a log-likelihood interpretation, the weighted likelihood bootstrap of Newton and Raftery (1994) is also a special case of *UBS*. This can be used for approximate Bayesian inference as suggested by them.

Example 2.3.6 THE m OUT OF n BOOTSTRAP Consider the m out of n bootstrap, where $m \rightarrow \infty$ and $m/n \rightarrow 0$. This is carried out by selecting m data points randomly out of the n available. If the selection is without replacement, this scheme can be identified with the delete- $(n-m)$ jackknife. If the selection is with replacement, the weights are a random sample from $Multinomial(m, 1/n, \dots, 1/n)$. The m out of n bootstrap with replacement has turned out to be a crucial tool in many problems where more traditional resampling methods fail. For example, in case of degenerate U -statistics and non-stationary autoregressive processes. Recently, Bickel and Sakov (1999) has addressed the issue of the choice of m in this bootstrap.

Example 2.3.7 (DOUBLE BOOTSTRAP OR NESTED BOOTSTRAP) The double bootstrap or nested bootstrap is a common technique that is employed for bootstrap confidence intervals of studentised statistics with unknown variances. Here the weights $(w_{n1}, \dots, w_{nn})$ follow $Multinomial(n, M_{n1}/n, \dots, M_{nn}/n)$ conditional on $M_n = (M_{n1}, \dots, M_{nn})$, which in turn follows $Multinomial(n, 1/n, \dots, 1/n)$. Thus this is a two stage sampling scheme, where the model for the second stage sampling depends on the outcome of the first stage. The behaviour of this bootstrap is similar to the classical bootstrap described in Example 2.3.1, but this is more likely to concentrate on fewer data points than the classical bootstrap.

Example 2.3.8 POLYA-EGGENBERGER BOOTSTRAP This technique is similar to the double bootstrap, except that none of the conditional probabilities for the second stage sampling is zero. Here $(w_{n1}, \dots, w_{nn})$ follow $Multinomial(n, D_{n1}/n, \dots, D_{nn}/n)$ conditional on the absolutely continuous random vector $D_n = (D_{n1}, \dots, D_{nn})$, which follows $Dirichlet_n(\alpha, \dots, \alpha)$ for some parameter $\alpha > 0$. The behaviour of this bootstrap is similar to the classical bootstrap and the double bootstrap.

Example 2.3.9 THE MULTIVARIATE HYPERGEOMETRIC BOOTSTRAP In this scheme, the weights come from the multivariate hypergeometric distribution

$$Prob[w_{n1} = k_1, \dots, w_{nn} = k_n] = \binom{nK}{n}^{-1} \binom{K}{k_1} \dots \binom{K}{k_n}; \quad \sum_{i=1}^n k_i = n, \quad 0 \leq k_i \leq K$$

where K is some integer parameter. The behaviour of this bootstrap is similar to the classical bootstrap, or the double bootstrap or Polya-Eggenberger bootstrap.

Examples 2.3.8 and 2.3.9 are so-called urn models, as Praestgaard and Wellner (1993) have pointed out. Many more such schemes may be described. See Muliere and Walker (1998) for some more examples.

2.4 The correlated weights bootstrap

Let us digress a bit and consider a different form of resampling, which is much in use for weakly dependent observations. This is the block bootstrap. There are various forms of the

block bootstrap, for example, non-overlapping blocks (Carlstein (1986)), overlapping blocks (Kunsch (1989) and Liu and Singh (1992a)), circular blocks (Politis and Romano (1992)) and stationary blocks (Politis and Romano (1994)). The different forms of block bootstrap have been widely used, see for example Lahiri (1999) and references therein. Higher order properties of such schemes are also known, see Lahiri (1996), Gotze and Kunsch (1996) and references therein. The block bootstrap is also a generalised bootstrap. We illustrate this point with the following example:

Example 2.4.1 BLOCK BOOTSTRAP Let $X_1, \dots, X_n$ be observations from a stationary sequence of weakly dependent observations. Suppose $EX_i = \mu$, and this is estimated using $T_n = \sum X_i/n$. For a block size l , the *circular block bootstrap* considers the n blocks of observations $B_i = (X_i, X_{i+1}, \dots, X_{i+l-1})$, $i = 1, 2, \dots, n$ where for $j > n$, X_j is taken to be X_{j-n} . Then k blocks are drawn with replacement from the B_i 's, where approximately $n \approx kl$, and these k blocks $B_i^*, \dots, B_k^*$ are treated as bootstrap data.

Let

$$\xi_{ij} = \begin{cases} 1 & \text{if } j = i, \dots, i+l-1, \quad i = 1, \dots, n-l+1 \\ 1 & \text{if } j = i, \dots, n, 1, 2, \dots, l - (n-i-1), \quad i = n-l, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

Let $\eta_i = (\xi_{i1}, \dots, \xi_{in})^T$, $i = 1, \dots, n$. Consider the random vector $\mathbf{w}$ such that

$$Prob[\mathbf{w} = \eta_i] = 1/n, \quad i = 1, \dots, n$$

It can be seen that the circular block bootstrap is drawing a sample of size k with replacement from the distribution of $\mathbf{w}$. Let $\mathbf{w} = (w_1, \dots, w_n)^T$. The correlation structure of the circular block bootstrap weights is:

$$Corr(w_i, w_{i+h}) = \begin{cases} 1 - nh/l(n-l) & \text{if } h < l \\ -l/(n-l) & \text{if } h \geq l \end{cases}$$

To capture the dependence in $\{X_i\}$ it is always required that $l \rightarrow \infty$ and $l/n \rightarrow 0$ with n . It can be easily seen that then any two fixed weights have a correlation tending to one. We may thus call this the *Correlated Weights Bootstrap*, as opposed to the *Uncorrelated Weights Bootstrap* discussed earlier. Correlated weights bootstrap, or their more well known counterparts the block bootstraps, are extremely useful for resampling with dependent observations, and are designed for such data. Using them for data where correlation does not play a significant role, like with independent or even uncorrelated observations, is unnecessary and inefficient technology. Throughout the present study we often discuss dependent data, like autoregressive processes, but we do not need to consider equivalents of the block bootstrap or correlated weights on it, since we exploit some underlying structure present in the model. Needless to say, for most of the problems we discuss, using correlated weights would produce consistent but generally inefficient estimators.

2.5 Weighted jackknives as generalised bootstrap

Consider the linear regression model $y_i = \mathbf{x}_i^T \beta + e_i$. Assume that $\mathbf{x}_i$'s are non-random, and e_i 's are independent with finite variances. Suppose the problem is to obtain the dispersion matrix of the least squares estimator $\hat{\beta} = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$. Let $\mathbf{X}$ be the $n \times p$ dimensional matrix whose i^{th} row is $\mathbf{x}_i^T$.

Let d be an integer in $[1, n]$. From the set $\{1, 2, \dots, n\}$ choose d integers. The delete- d parameter estimate $\hat{\beta}_{(-d)}$ is the usual least squares estimate from the data with the chosen d observations dropped. Such estimates are found for all possible $\binom{n}{d}$ choices of sets of d observations to delete. The jackknife variance estimate is an appropriately scaled version of $\sum (\hat{\beta}_{(-d)} - \hat{\beta})^2$ where the sum runs over all possible choices. A detailed study of the delete- d jackknife can be found in Shao (1987, 1988, 1989), Wu (1986, 1990), Shao and Wu (1987, 1989), Shao and Tu (1995). For example, in the usual delete-1 jackknife, introduced by Quenouille (1956), the variance estimator simplifies to

$$V_J = \frac{1}{n(n-1)} \sum_{i=1}^n (J_i - \hat{\beta})^2.$$

Here the *pseudovalue* J_i is defined as $n\hat{\beta} - (n-1)\hat{\beta}_{(-i)}$, and $\hat{\beta}_{(-i)}$ stands for the least squares estimate from the data set with i^{th} observation deleted.

In the above sum often weights are attached to the individual summands to yield the weighted jackknives, with weights possibly depending on the subset of ignored observations. We now describe the important weighted jackknife schemes. Let $\mathbf{G} = \mathbf{X}^T \mathbf{X}$ and for any i , $\delta_i = \mathbf{x}_i^T \mathbf{G}^{-1} \mathbf{x}_i$. For an $(n \times p)$ matrix $\mathbf{A}$ and $S \in \mathcal{S}_{n,d}$, $\mathbf{A}_S$ is defined to be the submatrix of $\mathbf{A}$ formed by deleting the $i_1^{\text{th}}, \dots, i_d^{\text{th}}$ rows of $\mathbf{A}$. Denote $\mathbf{X}_S^T \mathbf{X}_S$ by $\mathbf{G}_S$. We assume that $\mathbf{G}_S$ is positive definite for all choices of $S \in \mathcal{S}_{n,d}$, as long as $d + p \leq n$. Let $\hat{\beta}_S = \mathbf{G}_S^{-1} \mathbf{X}_S^T \mathbf{y}_S$. The variance estimator of Hinkley (1977) is

$$V_{JH} = n(n-p)^{-1} \sum_{i=1}^n (1 - \delta_i)^2 (\hat{\beta}_{(-i)} - \hat{\beta})(\hat{\beta}_{(-i)} - \hat{\beta})^T$$

and that of Wu (1986) is

$$V_{JW} = \left(\binom{n-p}{d-1} |\mathbf{G}| \right)^{-1} \sum_{S \in \mathcal{S}_{n,d}} |\mathbf{G}_S| (\hat{\beta}_S - \hat{\beta})(\hat{\beta}_S - \hat{\beta})^T$$

A detailed study of these weighted jackknives is made in Shao and Wu (1987). The third weighted jackknife scheme, due to Liu and Singh (1992), is defined for $p \equiv 1$ case, and does not have any simple extension to higher dimensional models. This estimator is defined as

$$V_{JLS} = \frac{L_n}{n^2(n-1)} \sum_{i=1}^n \frac{(J_i - \hat{\beta})^2}{x_i^2}$$

where $L_n = \sum_{i=1}^n x_i^2$.

The generalised bootstrap described in Section 2.2 is carried out by selecting random weights $\{w_{ni}; i = 1, \dots, n\}$ and then transforming the i^{th} data point (x_i, y_i) to $(\sqrt{w_{ni}}x_i, \sqrt{w_{ni}}y_i)$. Let the least squares estimate of parameter β be $\hat{\beta}$ using the original data and $\hat{\beta}_B$ with the transformed data. Assume that the weights satisfy $E_B(w_{ni} - 1)^2 = \sigma_n^2$, $i = 1, \dots, n$ where σ_n^2 may depend on n . The generalised bootstrap variance estimate is defined to be

$$\sigma_n^{-2} E_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T \quad (2.5.1)$$

In the next theorem we show that the jackknife variance estimators defined above are of the form (2.5.1) by defining the weight distribution appropriately.

Theorem 2.5.1 *In the context of variance estimation of the least squares estimator in linear models, the jackknife resampling schemes are special cases of generalised bootstrap.*

Proof of Theorem 2.5.1 To establish this theorem, we have to show that the jackknives are driven by a scheme of random weights.

Let $\mathcal{N}_n = \{1, 2, \dots, n\}$ and $\mathcal{S}_{n,d} = \{\text{all subsets of size } d \text{ from } \mathcal{N}_n\}$. We will identify a typical element of $\mathcal{S}_{n,d}$ by $S = \{i_1, i_2, \dots, i_d\}$. There are $\binom{n}{d}$ such elements in $\mathcal{S}_{n,d}$. S^C denotes the set $\mathcal{N}_n \setminus S$. Let $\{\xi_S, S \in \mathcal{S}_{n,d}\}$ be a collection of vectors in $\mathbb{R}^n$ defined by $\xi_S = (\xi_S(1), \xi_S(2), \dots, \xi_S(n))$, where

$$\xi_S(i) = \begin{cases} c_S & \text{if } i \in S^C \\ 0 & \text{if } i \in S \end{cases}$$

where $c_S, S \in \mathcal{S}_{n,d}$ is a set of constants. Suppose $\{p_S, S \in \mathcal{S}_{n,d}\}$ is a set of probabilities, that is $p_S \geq 0$ for all $S \in \mathcal{S}_{n,d}$ and $\sum_{S \in \mathcal{S}_{n,d}} p_S = 1$. Consider the random vector $W_n = (w_{n1}, w_{n2}, \dots, w_{nn})$ which has the following probability law

$$P[W_n = \xi_S] = p_S, \quad S \in \mathcal{S}_{n,d} \quad (2.5.2)$$

Observe that for $i \in \mathcal{N}_n$

$$w_{ni} = \begin{cases} c_S & \text{if } i \notin S \\ 0 & \text{if } i \in S \end{cases}$$

It can now be easily seen that if the i^{th} data point (x_i, y_i) is scaled by $\sqrt{w_{ni}}$, and the least squares estimate calculated for the transformed data, the resulting estimate is $\hat{\beta}_S$. For the different jackknife variance estimates, we only have to choose the constants c_S and p_S appropriately, so that the mean squared deviation from one is same but possibly depending on n . These condition can be summarised as

$$\sum_{S: i \notin S} (c_S - 1)^2 p_S = n^{-1} \sum_{i=1}^n \sum_{S: i \notin S} (c_S - 1)^2 p_S, \quad i = 1, 2, \dots, n \quad (2.5.3)$$

One way of ensuring (2.5.3) holds is to specify p_S , $S \in \mathcal{S}_{n,d}$ and then take $c_S = 1 + k p_S^{-1/2}$, $S \in \mathcal{S}_{n,d}$ for some constant $k > 0$.

For the delete- d jackknife, we take $p_S = \binom{n}{d}^{-1}$, $S \in \mathcal{S}_{n,d}$ so that the expression in (2.5.3) do not depend on S . For the weighted jackknives, we have the following relations

$$p_S \propto |\mathbf{G}_S| \text{ for Wu (1986)} \quad (2.5.4)$$

$$p_i \propto (1 - \delta_i)^2 \text{ for Hinkley (1977)} \quad (2.5.5)$$

$$p_i \propto \frac{1}{x_i^2} \text{ for Liu and Singh (1992)} \quad (2.5.6)$$

Note that for Hinkley's jackknife and Liu and Singh's jackknife the set S can only be singletons, so when $S = \{i\}$, we use the notation ξ_i and p_i in place of ξ_S and p_S respectively.

Thus we can see that with an appropriate choice of weights, the jackknife variance estimate is of the form (2.5.1). ■

Incidentally, for (2.5.4) it can be seen using the Binet-Cauchy expansion that the sum $\sum_S d(n-d)^{-1} (\binom{n-p}{d-1} |\mathbf{G}|)^{-1} |\mathbf{G}_S|$ is $\frac{n-d-p+1}{n-d}$. Wu (1986) had pointed out that the extra factor of $\frac{n-d-p+1}{n-d}$ should be absorbed in the $\hat{\beta}_S - \hat{\beta}$ for exact standardisation. Our proposed estimator (2.5.1) does that automatically. Hence the entire discussion leading to the proposed jackknife estimator in Wu (1986) can now be motivated as resampling from a generalised bootstrap viewpoint. In Chapter 4 we discuss higher order properties of these weighted jackknife schemes.

In subsequent chapters we discuss a number of technical conditions on *UBS* weights which do not hold in general for the weighted jackknife weights. Thus these schemes deserve separate attention.

2.6 The generalised residual bootstrap

Consider the linear regression problem described in Section 2.5, and let $\hat{\beta}$ be the least squares estimator. Let

$$\begin{aligned} \mathbf{I} &= \text{identity matrix of appropriate order} \\ \mathbf{J} &= n^{-1} \mathbf{1} \mathbf{1}^T \\ \hat{\beta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e} \\ \mathbf{P}_x &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \text{ the projection on the column space of } \mathbf{X} \\ \hat{\mathbf{y}} &= \mathbf{X} \hat{\beta} = \mathbf{X} \beta + \mathbf{P}_x \mathbf{e} \\ \mathbf{R} &= \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P}_x) \mathbf{e} \\ \mathbf{R}_0 &= (\mathbf{I} - \mathbf{J})(\mathbf{I} - \mathbf{P}_x) \mathbf{e} \\ \mathbf{G} &= \mathbf{X}^T \mathbf{X} \end{aligned}$$

Let $\mathbf{W}$ be an $n \times n$ matrix, whose i^{th} row is denoted by $\mathbf{w}_i^T$. Suppose $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ are *i.i.d.* samples from the distribution *Multinomial*(1, 1/ n , 1/ n , ..., 1/ n). If we define $\mathbf{y}^* = \mathbf{X} \hat{\beta} + \mathbf{W} \mathbf{R}_0$,

we get the usual classical residual bootstrap of Efron (1979). A generalisation of this is possible by taking any random $\mathbf{W}$ to be the weight matrix, subject to certain conditions. Also $\mathbf{R}$ or other residuals may be used instead of $\mathbf{R}_0$. However, we restrict attention to $\mathbf{R}$ and $\mathbf{R}_0$ in this thesis. Using other forms of residuals may lead to a bootstrap that is more robust, in the sense that model assumptions can be more relaxed. See Haslett and Hayes (1998) for some discussion on different residuals.

If the resampled data set is formed by

$$\mathbf{y}^* = \mathbf{X}\hat{\beta} + \mathbf{W}\mathbf{R}^* \quad (2.6.1)$$

where $\mathbf{W}$ is a random matrix and $\mathbf{R}^*$ is either $\mathbf{R}$ or $\mathbf{R}_0$, we call it a *Generalised Residual Bootstrap (GBS)*. In view of (2.6.1), we have $\hat{\beta}_B = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}^*$ and hence $\hat{\beta}_B - \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{R}^*$. We have already seen that the usual residual bootstrap is a particular case. If $\mathbf{R}^* = \mathbf{R}$ and $\mathbf{W}$ is diagonal with entries that are *i.i.d.* observations from a random variable with zero mean and unit variance, we have the *external or wild bootstrap* of Wu (1986). If $\mathbf{R}^* = \mathbf{R}$ and $\mathbf{W} = (\mathbf{I} - \mathbf{J})\mathbf{W}_1$, where $\mathbf{W}_1$ is diagonal with entries that are *i.i.d.* observations from a random variable with zero mean and unit variance, we obtain the *weighted bootstrap* of Liu (1988). In general, the behaviour of the *GBS* technique depends on the nature of the matrix $\mathbf{W}$. However, the choice of the residual vector is affected in an interesting way. If $\mathbf{E}_B\mathbf{W} = c\mathbf{J}$, define $\mathbf{W}_0 = (\mathbf{W} - c\mathbf{J})(\mathbf{I} - \mathbf{J})$. Using the idempotence of $\mathbf{J}$, we have

$$\begin{aligned} \mathbf{W}\mathbf{R}_0 &= (\mathbf{W} - c\mathbf{J})\mathbf{R}_0 + c\mathbf{J}\mathbf{R}_0 \\ &= (\mathbf{W} - c\mathbf{J})(\mathbf{I} - \mathbf{J})\mathbf{R} \\ &= \mathbf{W}_0\mathbf{R} \end{aligned} \quad (2.6.2)$$

Thus the choice of $\mathbf{R}^*$ to be $\mathbf{R}$ or $\mathbf{R}_0$ may be offset by the choice of the random weighting matrix.

The *GBS* variance estimator depends heavily on the nature of the weights and so the definition varies according to the resampling scheme. We now give three different variance estimators, to incorporate some of the existing and related estimators in our study.

$$V_{GB1} = \frac{1}{n\sigma_n^2} \mathbf{E}_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T \quad (2.6.3)$$

$$V_{GB2} = \mathbf{E}_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T \quad (2.6.4)$$

$$V_{GB3} = \frac{1}{\sigma_n^2} \mathbf{E}_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T \quad (2.6.5)$$

In the above σ_n^2 is the variance of the bootstrap weights. It can be deduced that V_{GB2} , and V_{GB3} are respectively of the same form as the variance estimate under residual bootstrap and external bootstrap. The variance estimator (2.6.4) is essentially a special case of (2.6.3), and consistency of this requires $n\sigma_n^2 \rightarrow 1$. However, it deserves separate attention because of two

reasons. First, because this is the variance expression for the classical residual bootstrap, for which $\sigma_n^2 = (n-1)/n^2$. Second, in Chapter 4, we obtain a result showing that the ‘best under homoscedasticity’ resampling scheme has a variance expression of this form. It can be seen that all the variance estimators are scaled versions of $E_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T$, the scaling depending on σ_n^2 . These variance estimators can be viewed as a natural extension of the variance estimate used in Barbe and Bertail (1995) for generalised bootstrapping of the sample mean.

The *GBS* variance estimators V_{GB1} , V_{GB2} and V_{GB3} behave in an interesting way. Under certain conditions V_{GB1} and V_{GB2} belongs to the *E*-class. Under some other conditions, V_{GB3} belongs to the *R*-class. These results are discussed in details in Chapter 4. The behaviour of *GBS* is largely governed by whether $\mathbf{W}$ is a diagonal matrix or not. Generally speaking, if off-diagonal entries are non-zero, the scheme is consistent only under homoscedasticity.

Chapter 3

VARIANCE ESTIMATION IN HIGH DIMENSIONAL LINEAR REGRESSIONS: FIRST ORDER RESULTS

3.1 Introduction

Consider the problem of estimating the variance of the least squares estimator in linear regression. In Liu and Singh (1992) first order approximations of the variance estimate of some resampling schemes were discussed under appropriate assumptions. They classified the resampling schemes as satisfying one of the following

$$n^{3/2}(V_B - V_n) = Y_n + o_P(1) \quad (3.1.1)$$

$$n^{3/2}(V_B - V_n) = Z_n + o_P(1) \quad (3.1.2)$$

where $V_n = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$ and V_B denotes its resampling estimate. Let $L_n = \sum_{i=1}^n x_i^2$ and $Ee_i^2 = \tau_i^2$. Also, if the errors are homoscedastic then $\tau_i^2 = \tau^2$, $i = 1, \dots, n$. If the parameter dimension $p = 1$, then $Y_n = n^{1/2} \frac{1}{L_n} \sum_{i=1}^n (e_i^2 - \tau^2)$, and $Z_n = n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2)$. The definitions of Y_n and Z_n for general p are given in Chapter 4.

Typically, Y_n and Z_n will have limiting distributions. If all the τ_i^2 's are equal then all the resampling schemes discussed give consistent estimate of the variance of $\hat{\beta}$. However, those that satisfy (3.1.1) have greater asymptotic relative efficiency compared to those that satisfy (3.1.2). On the other hand, resampling schemes that satisfy (3.1.2) give consistent estimates also under heteroscedasticity of the error terms. Schemes that satisfy (3.1.1) in general do not yield consistent estimates under heteroscedasticity. Thus those that satisfy (3.1.1) are more *efficient* under homoscedasticity, whereas those that satisfy (3.1.2) are *robust* under heteroscedasticity. Liu and Singh (1992) thus classified resampling schemes as belonging to either *E*-class or *R*-class according to whether they satisfy (3.1.1) or (3.1.2). It was shown that the residual bootstrap and the weighted jackknife due to Liu and Singh (1992) (for $p = 1$) belong to the *E*-class, the rest of the standard resampling schemes, *i.e.* the external bootstrap, paired bootstrap and the delete-1 jackknife belong to the *R*-class. The weighted jackknife of Liu and Singh (1992) does not have an extension to the multiple regression case.

In Chapter 2, we introduced the *UBS* resampling technique, special cases of which are the paired bootstrap, delete- d jackknives and the Bayesian bootstrap and many other examples. In

this chapter we establish an asymptotic representation of the *UBS* estimator of the variance, which is similar to the *R*-class representation obtained by Liu and Singh (1992). The *R*-class representation is already known for the paired bootstrap and the delete-1 jackknife. Our result extends this in two ways: (i) the same representation holds for many other resampling techniques, notably the delete- d jackknife for general d ; and (ii) the representation holds under very general model assumptions, where we allow for random regressors, heteroscedastic errors and the dimension of the parameter p to tend to infinity with data size n . Many time series models, dynamic linear models and regression models can be considered in this framework.

In Section 3.2 the different model conditions are discussed in details. Some discussion on the conditions required on *UBS* weights for this particular problem at hand is presented in subsection 3.2.2. Then in Section 3.3 the main results are stated. Some simulation experiments are reported in Section 3.4. These simulations are used for comparing different *UBS* schemes. The proofs of the theorems are in Section 3.5.

The contents of this chapter is adapted from Chatterjee and Bose (1999a).

3.2 Assumptions and notations

3.2.1 The model

The data consists of the observations $\{(\mathbf{x}_{i:n}, y_{i:n}), i = 1, \dots, n\}$, where $\mathbf{x}_{i:n}$'s are $p_n \times 1$ vectors and $y_{i:n}$'s are real valued. We assume the linear regression model given by

$$y_{i:n} = \mathbf{x}_{i:n}^T \beta_n + e_{i:n}, \quad i = 1, \dots, n \quad (3.2.1)$$

for some fixed but unknown sequence of constants β_n and some function $e_{i:n}$ which represents the 'error'. We henceforth write $p, \beta, \mathbf{x}_i, y_i, e_i$ respectively for $p_n, \beta_n, \mathbf{x}_{i:n}, y_{i:n}, e_{i:n}$. Let $\mathbf{X}$ denote the $n \times p$ matrix whose i^{th} row is formed by $\mathbf{x}_i^T$. Let $\mathbf{X}^T$ denote the transpose of $\mathbf{X}$. Also let $\mathbf{Y}$ and $\mathbf{e}$ be the n -dimensional vectors whose i^{th} entries are y_i and e_i respectively. Then the model with n observations may be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \quad (3.2.2)$$

Let $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ be the projection matrix on the column space of $\mathbf{X}$. Let $\mathbf{A}_{ij}$ denote the $(i, j)^{\text{th}}$ element of the matrix $\mathbf{A}$. Also, for any real symmetric matrix $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$, $\lambda_{\max}(\mathbf{A})$, $\lambda_i(\mathbf{A})$, respectively, denote the maximum, minimum, maximum in absolute value and i^{th} eigenvalue of $\mathbf{A}$. Throughout the rest of this chapter we use the generic c and k to denote constants, without implying they are the same, wherever they appear.

We now state the conditions which we impose on our linear model. The regressors $\mathbf{x}_i$'s may or may not be random. The first condition, on the dimension growth, is

$$p^2/n \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (3.2.3)$$

We assume the following conditions if $\mathbf{x}_i$'s are non-random:

$$\sup_{1 \leq i \leq n} \|\mathbf{x}_i\|^2 = O(p), \quad (3.2.4)$$

$$\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) > c > 0. \quad (3.2.5)$$

The assumptions on the errors are

$$Ee_i^2 = \tau_i^2 < c < \infty, \quad (3.2.6)$$

$$Ee_ie_j = 0; \quad i, j \text{ different}, \quad (3.2.7)$$

$$\sup Ee_i^2e_je_k = O(n^{-1}), \quad i, j, k \text{ different}, \quad (3.2.8)$$

$$\sup Ee_ie_je_ke_l = O(n^{-2}), \quad i, j, k, l \text{ different}, \quad (3.2.9)$$

$$\sup Ee_i^4 < \infty. \quad (3.2.10)$$

In case the $\mathbf{x}_i$'s are random, we need probabilistic versions of (3.2.4) and (3.2.5). Suppose $\mathcal{A}$ is the set on which $\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) > c > 0$. Then the conditions assumed are

$$\sup_{1 \leq i \leq n} E\|\mathbf{x}_i\|^2 = O(p), \quad (3.2.11)$$

$$P[\mathcal{A}] = 1 - O(p^2n^{-2}). \quad (3.2.12)$$

When $\mathbf{x}_i$'s are random, the assumptions on the error are

$$e_i \text{ is independent of } \mathbf{x}_i, (\mathbf{x}_j, e_j), \quad j \leq i \quad \forall i, \quad (3.2.13)$$

$$Ee_i = 0, \quad (3.2.14)$$

$$Ee_i^2 = \tau_i^2 < c < \infty, \quad (3.2.15)$$

$$\sup Ee_i^4 < \infty. \quad (3.2.16)$$

A precise definition of our models is now as follows:

Model 1: (Fixed regressors) This is (3.2.1) with non-random $\mathbf{x}_i$'s satisfying conditions (3.2.3)-(3.2.10).

Model 2: (Random regressors) This is (3.2.1) with random $\mathbf{x}_i$'s satisfying conditions (3.2.3), (3.2.11)-(3.2.16).

It may be mentioned here that our conditions on the errors allow for the e_i 's to come from several standard dependent structures, such as the autoregressive or autoregressive conditional heteroscedastic. If the e_i 's are mean zero normal random variables, not necessarily independent, then (3.2.8) and (3.2.9) follow from (3.2.7). This follows from

Lemma 3.2.1 (Wick) *If (N_1, N_2, N_3, N_4) is a normal random vector with mean zero then*

$$E(N_1N_2N_3N_4) = E(N_1N_2)E(N_3N_4) + E(N_1N_3)E(N_2N_4) + E(N_1N_4)E(N_2N_3).$$

Note that (3.2.5) or (3.2.12) ensures that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists with a large probability. We want to estimate $\mathbf{V}_n = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$. Under Model 1, this is $\mathbf{V}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{T}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$, where $\mathbf{T}$ is a diagonal matrix with i^{th} diagonal element τ_i^2 . Notice that this is a $p \times p$ matrix, and for us $p \rightarrow \infty$. Hence we try to estimate $\xi^T\mathbf{V}_n\xi$ for any $\xi \in \mathbb{R}^p$ with $\|\xi\| = 1$. Our main result is a representation for $\xi^T\mathbf{V}_{\text{UBS}}\xi$, where $\mathbf{V}_{\text{UBS}}$ is an appropriate bootstrap estimate of $\mathbf{V}_n$. This also gives an element wise bound of $\mathbf{V}_n$ and results for estimating the variance of linear combinations of the elements of β .

For some resampling schemes that we consider, condition (3.2.5) is not sufficient. For example, a paired bootstrap sample can include only the data points indexed by $\mathcal{I}_m = \{i_1, i_2, \dots, i_m\}$. Let $\mathbf{X}^*$ be the $m \times p$ matrix whose j^{th} row is $\mathbf{x}_{i_j}^T$. If $m < p$ there is no possibility that this design matrix is of full column rank. However, this case has exponentially small probability as we show later. Even if $m \geq p$, we still need nonsingularity of $\mathbf{X}^{T*}\mathbf{X}^*$. So an equivalent of (3.2.5) is needed for submatrices of the design matrix. We now state here a more general condition. Suppose m_0 is a specified integer in the range $[n/3]$ to n . Assume that

$$m^{-1}\mathbf{X}^{T*}\mathbf{X}^* > k_1\mathbf{I}, \quad k_1 > 0 \quad (3.2.17)$$

for every such choice of subset $\mathcal{I}_m$ of size m from $\{1, 2, \dots, n\}$ and for every m in $[m_0, n]$. (For two matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ we write $\mathbf{A}_1 > \mathbf{A}_2$ if and only if $\mathbf{A}_1 - \mathbf{A}_2$ is positive definite.) Note that this condition depends on m_0 ; the higher its value, the weaker is the condition. For $m_0 = n$ condition (3.2.17) is same as (3.2.5). For the model with random regressors, the corresponding condition is that the set $\mathcal{A}$, on which

$$m^{-1}\mathbf{X}^{T*}\mathbf{X}^* > k_1\mathbf{I}, \text{ for every } \mathcal{I}_m \text{ and every } m \in [m_0, n], \quad (3.2.18)$$

has a probability $1 - O(p^2/n)$. There are *UBS* resampling schemes which require only $m_0 = n$. The delete- d jackknife requires $m_0 = n - d$. However, the paired bootstrap requires as low an m_0 as possible. The more stringent assumption (3.2.17) or (3.2.18) makes resampling schemes like the paired bootstrap and the different jackknives feasible.

3.2.2 The resampling scheme

We now discuss the conditions on the bootstrap weights used for *UBS* resampling in the context of the present chapter.

Let $\{w_{ni}; 1 \leq i \leq n, n \geq 1\}$ be a triangular array of non-negative random variables to be used as weights. *We drop the suffix n from the notation of the weights.*

The weights $w_1, \dots, w_n$ used for resampling satisfy certain restrictions on the first few moments which we now state. Recall that $V(w_i) = \sigma_n^2$ and assume that the quantities

$$E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

are functions of the powers $i, j, k \dots$ only, and not of the indices $a, b, c \dots$. Thus we can write

$$c_{ijk\dots} = E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

Note that if the weights are assumed to be exchangeable, then the above condition follows. But exchangeability of weights is not a necessary condition.

Let $\mathcal{W}$ be the set on which at least m_0 of the weights are greater than some fixed constant $k_2 > 0$. The value of m_0 is the same as that of assumptions (3.2.17) or (3.2.18). The weights are assumed to satisfy certain conditions.

$$E(w_i) = 1, \quad (3.2.19)$$

$$\sigma_n^2 \rightarrow k > 0, \quad (3.2.20)$$

$$P_B[Kn \geq \sum_{i=1}^n w_i \geq kn, \quad K > k > 0] = 1, \quad (3.2.21)$$

$$P_B[\mathcal{W}] = 1 - O(p^2 n^{-1}), \quad (3.2.22)$$

$$c_{11} = O(n^{-1}), \quad (3.2.23)$$

$$\forall i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 3, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+1}) \quad (3.2.24)$$

$$\forall i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 4, \quad c_{i_1 i_2 \dots i_k} = O(\min(n^{-k+2}, 1)) \quad (3.2.25)$$

We define the bootstrap estimate of β to be

$$\hat{\beta}_B = \begin{cases} (\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_y & \text{on the set } \mathcal{W} \cap \mathcal{A} \\ \hat{\beta} & \text{otherwise} \end{cases}$$

and the bootstrap variance estimate to be

$$V_{UBS} = \sigma_n^{-2} E_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T.$$

For Model 1, take the set $\mathcal{A}$ to be the entire sample space, so that the definition of $\hat{\beta}_B$ does not depend on the model used. If y_i 's are i.i.d., the *UBS* variance estimate coincides with the generalised bootstrap variance estimate used for estimating the variance of the sample mean. See Barbe and Bertail (1995) for use of this statistic for other statistical functionals.

A slight variant of the above conditions on weights can be effected by dropping (3.2.20) and letting the common variance go to zero, so that the weights are asymptotically degenerate. With that condition, the conditions on various mixed moments can be slightly relaxed. For details about this variation refer to Chatterjee (1998) or Chapter 7. By taking into account this variation, all the examples of resampling schemes discussed in Section 2.3 of Chapter 2 may be considered for the results of the present chapter.

One significant condition on the weights is (3.2.22), which we now discuss. This condition is related to (3.2.17) and (3.2.18). If instead, we restrict the condition to require that all the weights are bounded away from zero, then we can take $m_0 = n$ in (3.2.17). For example, the weights can be taken to be *i.i.d.* unit mean random variables that are supported on some interval (k, K) , with $0 < k < K < \infty$. For the delete- d jackknives, we need (3.2.17) with $m_0 = n-d$ only. For the paired bootstrap, that is for a random sample from $Multinomial(n, 1/n, 1/n, \dots, 1/n)$, Proposition 3.3.1 of the next section shows that (3.2.22) is satisfied for $m_0 = [n/3]$, the greatest integer less than or equal to $n/3$. Note that the condition on $\mathcal{W}$ is in terms of ‘at least m_0 ’ weights being positive, so an upper bound on m_0 is really meaningful. There is a duality between model conditions and conditions on the resampling weights. We may relax certain model conditions by making the conditions on resampling weights more stringent. For example, for resampling with independent weights, we can ignore the model conditions (3.2.8) and (3.2.9).

Bootstrap schemes for linear models have been discussed in Efron (1979), Freedman (1981), and in Bickel and Freedman (1983). Hinkley (1977), Wu (1986), Shao and Wu (1987) have studied consistency of different bootstrap and jackknife schemes in heteroscedastic linear models. The bootstrap in regression models with many parameters has been considered by Bickel and Freedman (1983) and Mammen (1993), who respectively showed the consistency of the residual bootstrap and the wild bootstrap (which is a generalisation of the residual bootstrap) and paired bootstrap.

Our generalised bootstrap can be looked upon as a weighted least squares analysis with random weights, and then simple averaging over such weighted least squares estimates. With easily available weighted least squares routines this can lead to significant ease in implementing resampling in linear models. Even though we discuss estimating the variance of the ordinary least squares estimator here, our technique potentially extends to the corresponding problem for the weighted least squares.

3.3 Main Results

Let

$$(\mathbf{T}_n)_{ij} = \begin{cases} e_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Theorem 3.3.1 *Assume the conditions of Model 1. Also assume (3.2.17) for some m_0 . Then for any $\xi \in \mathbb{R}^p$ with $\|\xi\| = 1$,*

$$\begin{aligned} & n^{3/2} p^{-1} \xi^T (\mathbf{V}_{\text{UBS}} - \mathbf{V}_n) \xi \\ &= n^{3/2} p^{-1} \xi^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T [\mathbf{T}_n - \mathbf{T}] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \xi + O_P(pn^{-1/2}). \end{aligned} \quad (3.3.1)$$

In particular, *UBS* is a consistent resampling technique for the model. The distributional asymptotics for the variance estimator can essentially be developed from here, by noting that

the leading term in the variance representation is a linear combination of e_i^2 's. Note that the leading term in (3.3.1) is bounded in probability. Since this result does not depend on the particular choice of weights, these weights can be chosen according to convenience. This exact rate is not obtained under Model 2, but otherwise a very similar representation theorem holds.

Theorem 3.3.2 *Assume the conditions of Model 2. Also assume (3.2.18) for some m_0 . Then for any $\xi \in \mathbb{R}^p$ with $\|\xi\| = 1$,*

$$\begin{aligned} & n^{3/2} p^{-1} \xi^T (V_{\text{UBS}} - V_n) \xi \\ = & n^{3/2} p^{-1} \xi^T [I_{\mathcal{A}}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T - V_n] \xi + O_P(pn^{-1/2}) \end{aligned} \quad (3.3.2)$$

Notice that typically $(\hat{\beta} - \beta) = O_P((n/p)^{-1/2})$. The leading term in (3.3.2) is of the form $n^{-1/2} \sum_{i=1}^n (Z_i - EZ_i)$ for some standardized random variables Z_i (see (3.5.7) in the section on proofs for details on the random variables Z_i) plus some remaining terms which are $O_P(pn^{-1/2})$. Suppose we assume that $E[\xi^T(\hat{\beta} - \beta)]^4 < \infty$. This simultaneously ensures that the indicator on the set $\mathcal{A}$ can be ignored, and that the variance of $(n/p)^{-1/2} \xi_i^T(\hat{\beta} - \beta)$, that is $np^{-1} \xi^T V_n \xi$, is consistently estimated by $np^{-1} \xi^T V_{\text{UBS}} \xi$ in the sense that $np^{-1} \xi^T (V_{\text{UBS}} - V_n) \xi \rightarrow 0$ in probability.

If $\{w_1, w_2, \dots, w_n\}$ follows *Multinomial* $(n, 1/n, 1/n, \dots, 1/n)$, we get the paired bootstrap. The moment conditions on these weights can be easily verified by direct calculation. Through the following proposition we show that condition (3.2.22) is also satisfied with $m_0 = n/3$.

Proposition 3.3.1 *Suppose $(X_1, \dots, X_n)$ follows *Multinomial* $(n, 1/n, \dots, 1/n)$ distribution. If $\{m_n\}$ is such that $m_n/n < 1/3$, then the probability that at least m_n of the X 's are positive is greater than $1 - e^{-\alpha n}$, for some constant $\alpha > 0$.*

The proof of all the results are given in Section 3.5.

REMARK Our results show that for fixed p the *UBS* we study belongs to the *R*-class. Note that two special cases of *UBS*, the paired bootstrap and the delete-1 jackknife were already known to belong to the *R*-class.

The random regressor model has been considered in Mammen (1993) in the context of the paired bootstrap and the wild bootstrap. The model there was based on observing *i.i.d.* variables $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$, where $\mathbf{x}_i$'s were $p \times 1$ vectors with the assumed model relation being $y_i = \mathbf{x}_i^T \beta + e_i$ for some constant but unknown β . The dimension p was allowed to vary with n . In Chapter 5 we have considered the problem of estimating the distribution of the least squares estimate of a linear regression parameter, using the *UBS* resampling scheme, in the same set-up as that of Mammen (1993). Consistency of the *UBS* technique for estimating the distribution function is shown there. Since in this chapter we are estimating the variance

of the least squares estimator as opposed to the distribution function, as in Mammen (1993), the fourth moment condition (3.2.16) is imposed.

Our random regressor model is more general than Mammen (1993) in the sense that the assumption of *i.i.d.* nature of the data is modified to (3.2.13). However, in place of assuming $p^{1+\delta}/n \rightarrow 0$ for $\delta \geq 1/3$ as in Mammen (1993), or $\delta > 0$ as in Chapter 5, we need $p^2/n \rightarrow 0$. This is explained by the fact that in general the bias in the least squares estimate is of the order $p/n^{1/2}$, and so the mean squared error that we are estimating requires (3.2.3).

We now check that with regressors taken to be independently and identically distributed, as assumed in Mammen (1993), our model conditions (3.2.11)-(3.2.15) hold. Note that if $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$ are *i.i.d.* then $\{(\mathbf{x}_i, e_i), i = 1, \dots, n\}$ are also *i.i.d.* and this, in turn, implies (3.2.13)-(3.2.15). Suppose the regressors $\mathbf{x}_{i:n}$ are *i.i.d.*, with $\sup_n \sup_{\|d\|=1} E|d^T \mathbf{x}_{i:n}|^4 < \infty$ as in condition 2.2 of Theorem 1 of Mammen (1993). The condition (3.2.11) follows directly from Lemma 0 of Mammen (1993). Assume without loss that $E\mathbf{x}_i\mathbf{x}_i^T = \mathbf{I}$, and letting $\mathbf{A} = n^{-1} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T - \mathbf{I}$, we have $n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I} + \mathbf{A}$, so that

$$\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) = 1 + \lambda_{\min}(\mathbf{A}) \geq 1 - \lambda_{\max}(\mathbf{A})$$

and hence

$$\begin{aligned} P[|\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X})| \leq 1/2] &\leq P[|1 - \lambda_{\max}(\mathbf{A})| \leq 1/2] \\ &\leq P[|\lambda_{\max}(\mathbf{A})| \geq 1/2] \\ &\leq 4E(\lambda_{\max}(\mathbf{A}))^2 = O(p^{3/2}n^{-1}), \end{aligned}$$

with the last relation following from Lemma 1 of Mammen (1993). This verifies (3.2.12).

3.4 Some simulation results

To gauge how the different *UBS* schemes perform, especially when we have random regressors and/or dependent errors, we carried out a small simulation experiment. We chose the following five *UBS* schemes:

- (a) the *Multinomial*($n; 1/n, 1/n, \dots, 1/n$) bootstrap(MB);
- (b) the *Dirichlet*($n; 1/n, 1/n, \dots, 1/n$) bootstrap(DB);
- (c) the *Uniform*($1/2, 3/2$) bootstrap(UB);
- (d) the *Beta*(2, 7) bootstrap(BB1);
- (e) the *Beta*(7, 2) bootstrap(BB2).

Note that the first choice corresponds to the case of the paired bootstrap and the second one corresponds to the case of the Bayesian bootstrap with a Dirichlet process prior. For the last three choices, a sample of size n is generated from the given distribution and used for resampling. The last two schemes were considered to see how asymmetry of the generating distribution affects the performance of the resampling method. Computationally, the third choice is the easiest and fastest.

We considered three models with $p = 1$. The first model is the simple one of estimating the variance of the sample mean. This is intended to serve as a benchmark. The second model is the autoregressive (AR) model of order one with IID innovations. This is a well-known model in time series and the results obtained for this model are an indication of what is to be expected in similar models such as the AR of order $p > 1$. The third model is also the autoregressive model but here the errors, instead of being IID are assumed to have an ARCH (autoregressive conditional heteroscedastic) structure. The ARCH model is widely used in econometrics to model financial time series data, see Bera and Higgins (1993) for a review.

We now give a precise description of the three models.

Experiment 1: $\mathbf{X}_t = \beta + \epsilon_t$. In this case $\hat{\beta} = \sum_{i=1}^n \mathbf{X}_t / n$ and $V_n = n^{-1}$. For simulations we fix $\beta = 7.0$ and take the errors ϵ_t to be an i.i.d. sequence from $Normal(0, 1)$. This is the simplest example when Model 1 conditions are satisfied.

Experiment 2: $\mathbf{X}_t = \beta \mathbf{X}_{t-1} + \epsilon_t$, and $\mathbf{X}_0 = 0$. Here $\hat{\beta} = (\sum_{i=1}^n \mathbf{X}_{i-1}^2)^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_{i-1}$ and it is known that if $|\beta| < 1$ then $n^{1/2}(\hat{\beta} - \beta) \rightarrow N(0, 1 - \beta^2)$. For simulation, the innovations ϵ_t are taken to be i.i.d. $Normal(0, 1)$. The process has an explosive behaviour if $|\beta| > 1$, in which case normal limits are not obtained for the least squares estimate. With this in mind, we choose two different values of β as follows.

Experiment 2a: We take $\beta = -0.5$. Note that this is well within the safe zone.

Experiment 2b: We take $\beta = 0.9$. This value is close to the boundary and here the usual bootstrap is not expected to perform well.

Experiment 3: $\mathbf{X}_t = \beta \mathbf{X}_{t-1} + \epsilon_t$, where $\mathbf{X}_0 = 0$ and ϵ_t is an ARCH process satisfying $\epsilon_1 \sim N(0, \gamma^2)$ and $[\epsilon_t | \epsilon_s, s \leq t] \sim N(0, \gamma_t^2)$ where γ_t is in general a polynomial in $\epsilon_s, s \leq t$. We choose $\gamma_t^2 = \alpha + \beta \epsilon_{t-1}^2$. In this case, the (conditional) least squares estimate of β is given by $\hat{\beta} = (\sum_{i=1}^n \mathbf{X}_{i-1}^2)^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_{i-1}$. If $\gamma^2 = \alpha / (1 - \beta)$, then it is known that $n^{1/2}(\hat{\beta} - \beta) \rightarrow N(0, \gamma^2(1 - \beta^2))$. For our simulation we take $\alpha = 0.5$, $\beta = 0.4$ and $\gamma^2 = 5/6$. As earlier, we work with two different values of β . In **Experiment 3a** we take $\beta = -0.5$ and in **Experiment 3b** we take $\beta = 0.9$.

In each case we fix n , the size of the data, then randomly generate a data set satisfying the assumed conditions and use resampling on it. In all three experiments we first compute

the least squares estimate of β , and then resample for the variance of this estimator. Since in all experiments we consider, the least squares estimator has a limiting normal distribution, the different bootstrap variance estimates are compared through their appropriate asymptotic variances.

Results are presented in Tables 1, 2, 3, 4 and 5 respectively. We use the notation av for the asymptotic variance. The resample size was 5000 for Experiment 1 and 10000 for Experiment 2(a), 2(b), 3(a), 3(b). As the table entries show, *UBS* using *Beta*(2, 7) and *Beta*(7, 2) weights lead to slight underestimation and overestimation of the variance, possibly due to the difference in higher order terms. The results from the other resampling schemes are almost identical. *UBS* with *i.i.d.* uniform weights is recommended.

The starred figures are scaled up 1000 times.

TABLE 1: EXPERIMENT 1

n	$\hat{\beta}$		$av.*$		MB		DB		UB		BB1		BB2	
					β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*
30	6.88	61.30	6.88	61.92	7.01	80.79	6.88	61.55	6.88	51.99	6.88	82.23		
40	6.94	17.99	6.94	17.90	7.00	20.13	6.94	18.14	6.94	15.51	6.95	23.27		
100	7.01	10.22	7.01	10.24	6.99	10.93	7.01	10.46	7.01	8.99	7.01	13.21		
200	6.99	4.95	6.99	4.90	6.99	5.08	6.99	4.90	6.99	4.30	6.99	6.49		
600	6.99	1.66	6.99	1.64	6.99	1.64	6.99	1.67	6.99	1.43	6.99	2.14		
1000	6.95	1.10	6.95	1.07	6.95	1.10	6.95	1.11	6.95	0.96	6.95	1.42		
2000	7.01	0.47	7.01	0.47	7.01	0.47	7.01	0.48	7.01	0.40	7.01	0.61		

TABLE 2: EXPERIMENT 2(a)

n	$\hat{\beta}$		$av.*$		MB		DB		UB		BB1		BB2	
					β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*
30	-0.60	21.40	-0.61	16.44	-0.61	16.19	-0.60	15.32	-0.60	13.83	-0.60	20.32		
50	-0.59	13.03	-0.59	12.76	-0.59	11.90	-0.59	12.05	-0.59	10.47	-0.59	15.18		
100	-0.52	7.24	-0.53	6.31	-0.52	6.08	-0.52	6.26	-0.53	5.25	-0.53	7.96		
200	-0.45	3.99	-0.45	3.71	-0.45	3.54	-0.45	3.70	-0.45	3.16	-0.45	4.76		
500	-0.54	1.41	-0.54	1.21	-0.54	1.21	-0.54	1.21	-0.54	1.08	-0.54	1.62		
1000	-0.51	0.73	-0.51	0.70	-0.51	0.69	-0.51	0.72	-0.51	0.62	-0.51	0.91		
2000	-0.50	0.38	-0.50	0.36	-0.50	0.36	-0.50	0.36	-0.50	0.31	-0.50	0.48		

TABLE 2: EXPERIMENT 2(b)

n	$\hat{\beta}$	av.*	MB		DB		UB		BB1		BB2	
			β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*
30	0.96	2.85	0.93	2.76	0.93	2.66	0.96	3.92	0.96	3.71	0.96	5.50
50	0.92	3.15	0.91	2.66	0.91	2.51	0.92	3.22	0.92	2.79	0.92	4.26
100	0.87	2.49	0.87	2.70	0.87	2.62	0.87	2.81	0.88	2.49	0.88	3.96
200	0.81	1.68	0.81	1.94	0.81	1.87	0.81	1.90	0.81	1.64	0.81	2.51
500	0.85	0.56	0.85	0.58	0.85	0.57	0.85	0.58	0.85	0.50	0.85	0.76
1000	0.91	0.17	0.91	0.17	0.91	0.16	0.91	0.17	0.91	0.15	0.91	0.21
2000	0.91	0.09	0.91	0.09	0.91	0.09	0.91	0.09	0.91	0.08	0.91	0.11

TABLE 2: EXPERIMENT 3(a)

n	$\hat{\beta}$	av.*	MB		DB		UB		BB1		BB2	
			β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*
30	-0.57	22.39	-0.59	17.12	-0.58	15.52	-0.57	16.65	-0.57	14.19	-0.57	21.78
50	-0.68	10.84	-0.65	36.96	-0.66	27.97	-0.68	47.62	-0.68	39.14	-0.68	61.14
100	-0.68	5.43	-0.67	7.43	-0.67	7.03	-0.68	7.02	-0.68	6.02	-0.68	9.45
200	-0.49	3.80	-0.49	5.43	-0.49	5.25	-0.49	5.37	-0.49	4.81	-0.49	7.14
500	-0.54	1.41	-0.54	1.91	-0.54	1.88	-0.54	1.92	-0.54	1.64	-0.54	2.44
1000	-0.55	0.70	-0.55	2.02	-0.55	1.95	-0.55	2.03	-0.55	1.77	-0.55	2.70
2000	-0.50	0.37	-0.50	0.83	-0.50	0.83	-0.50	0.84	-0.50	0.72	-0.50	1.05

TABLE 2: EXPERIMENT 3(b)

n	$\hat{\beta}$	av.*	MB		DB		UB		BB1		BB2	
			β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*	β_B	V_B^*
30	0.93	4.42	0.91	3.42	0.91	3.35	0.93	5.13	0.93	4.51	0.93	6.76
50	0.94	2.31	0.92	2.84	0.92	2.65	0.94	4.14	0.94	3.54	0.94	5.44
100	0.95	1.05	0.94	1.28	0.94	1.25	0.95	1.68	0.95	1.50	0.95	2.19
200	0.92	0.76	0.92	0.86	0.92	0.84	0.92	0.88	0.92	0.76	0.92	1.15
500	0.91	0.33	0.91	0.38	0.91	0.38	0.91	0.39	0.91	0.35	0.91	0.51
1000	0.90	0.18	0.90	0.24	0.90	0.23	0.90	0.24	0.90	0.21	0.90	0.31
2000	0.88	0.11	0.88	0.16	0.88	0.16	0.88	0.17	0.88	0.14	0.88	0.21

3.5 Proofs

There are two important conclusions to be derived from the model conditions. For Model 1 they are

$$\max_i (\mathbf{P}_x)_{ii} = O(p/n), \quad (3.5.1)$$

$$\left\| \sum_{i=1}^n \mathbf{x}_i e_i \right\| = O_P(p^{1/2} n^{1/2}). \quad (3.5.2)$$

These are quickly verified as follows:

$$\begin{aligned} \max_i (\mathbf{P}_x)_{ii} &= \max_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \leq n^{-1} \max_i \|\mathbf{x}_i\|^2 [\lambda_{\min}(n^{-1} \mathbf{X}^T \mathbf{X})]^{-1} \\ &\leq cn^{-1} \max_i \|\mathbf{x}_i\|^2 = O(p/n), \text{ and} \end{aligned}$$

$$E \left\| \sum_{i=1}^n \mathbf{x}_i e_i \right\|^2 = E \sum_{i=1}^n e_i^2 \mathbf{x}_i^T \mathbf{x}_i = \sum_{i=1}^n \tau_i^2 \|\mathbf{x}_i\|^2 \leq c \sum_{i=1}^n \|\mathbf{x}_i\|^2 = O(pn).$$

For Model 2 the equivalent conclusions are

$$\max_i (\mathbf{P}_x)_{ii} I_A = O_P(p/n), \quad (3.5.3)$$

$$\left\| \sum_{i=1}^n \mathbf{x}_i e_i \right\| = O_P(p^{1/2} n^{1/2}). \quad (3.5.4)$$

These are also easily verified by using similar arguments as earlier.

Note that $\mathbf{X}$ is an $n \times p$ matrix of rank p . Let $\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q}^T$ be the singular value decomposition of $\mathbf{X}$. That is, $\mathbf{D}$ is an $p \times p$ diagonal matrix with positive diagonal elements, $\mathbf{P}$ is an $n \times p$ matrix such that $\mathbf{P}^T \mathbf{P} = \mathbf{I}$ and $\mathbf{Q}$ is a $p \times p$ orthogonal matrix. The spectral representation of $\mathbf{X}^T \mathbf{X}$ is given by $\mathbf{X}^T \mathbf{X} = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^T$. Let $\Lambda = n^{-1}\mathbf{D}^2$. Note that the minimum eigenvalue of the diagonal matrix Λ is bounded away from zero, so that the maximum eigenvalue of Λ^{-1} is bounded above. The notation $\Lambda^{1/2}$ is sometimes used for $n^{-1/2}\mathbf{D}$. Note that the diagonal entries of $\mathbf{D}$ are all positive.

Let $\mathbf{W}_D$ be the $n \times n$ diagonal matrix with i^{th} diagonal element w_i . Let us also use the notation $W_i = \frac{(w_i - 1)}{\sigma_n}$, and $\mathbf{W}_{(n \times n)} = \text{Diag}(w_1, w_2, \dots, w_n)$.

Since our two theorems state almost the identical result under two different models, we use an approach that proves both results simultaneously. Note that under Model 1 the set $\mathcal{A}$ is the entire sample space.

We first prove that with a high probability a condition like (3.2.5) is also true for the bootstrap design matrix.

Lemma 3.5.1 *Assume Model 1. Also assume condition (3.2.17) for an appropriate choice of $m_0 > p$. Then under the set $\mathcal{W} \cap \mathcal{A}$, all eigenvalues of the matrix $n^{-1}\mathbf{X}^T \mathbf{W}_D \mathbf{X}$ are greater than $k > 0$, where k is a constant.*

Proof of Lemma 3.5.1. Let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$ be the eigenvalues of $\mathbf{X}^T \mathbf{W}_D \mathbf{X}$. We have to show a positive lower bound for λ_1/n under $\mathcal{W} \cap \mathcal{A}$. Note that we get λ_1 by minimizing $\frac{\xi^T \mathbf{X}^T \mathbf{W}_D \mathbf{X} \xi}{\xi^T \xi}$ with respect to all vectors $\xi \in \mathbb{R}^p$. Now note that

$$\min_{\xi \in \mathbb{R}^p} \frac{\xi^T \mathbf{X}^T \mathbf{W}_D \mathbf{X} \xi}{\xi^T \xi}$$

$$\begin{aligned}
&= \min_{\xi \in \mathbb{R}^p} \min_{\eta = \mathbf{D}\xi} \frac{\eta^T \mathbf{P}^T \mathbf{W}_D \mathbf{P} \eta}{\eta^T \eta} \frac{\xi^T \mathbf{D}^2 \xi}{\xi^T \xi} \\
&\geq \min_{\xi \in \mathbb{R}^p} \min_{\eta \in \mathbb{R}^p} \frac{\eta^T \mathbf{P}^T \mathbf{W}_D \mathbf{P} \eta}{\eta^T \eta} \frac{\xi^T \mathbf{D}^2 \xi}{\xi^T \xi}
\end{aligned}$$

By condition (3.2.5), on the set $\mathcal{A}$ we have n^{-1} times the second term bounded below by a positive constant. Thus in order to complete the proof, we need to show that $\frac{\eta^T \mathbf{P}^T \mathbf{W}_D \mathbf{P} \eta}{\eta^T \eta}$ has a positive lower bound, as η varies. First observe that the model condition (3.2.17) says that on $\mathcal{A}$, for all $m >$ an appropriate m_0 if $\mathcal{I}_m = \{i_1, i_2, \dots, i_m\}$ is a subset of $\{1, 2, \dots, n\}$, and if $\mathbf{X}^*$ is the $m \times p$ matrix whose j^{th} row is $\mathbf{x}_{i_j}^T$, then $\frac{1}{m} \mathbf{X}^{*T} \mathbf{X}^* > k_1 \mathbf{I}$. If $\mathbf{P}^*$ is the submatrix of $\mathbf{P}$ corresponding to $\mathbf{X}^*$, this implies $\mathbf{P}^{*T} \mathbf{P}^* = \sum_{i \in \mathcal{I}_m} \mathbf{p}_i \mathbf{p}_i^T > k_1 \frac{m}{n} \mathbf{I} \geq \frac{k_1}{3} \mathbf{I}$.

Suppose $\mathcal{S}$ is the set of indices of weights that are greater than some fixed k_2 . Under $\mathcal{W}$, $\mathcal{S}$ has m elements, where $m \geq m_0$. Thus the same set of indices $\mathcal{S}$ is also an $\mathcal{I}_m$ for which (3.2.17) holds. Therefore under the set $\mathcal{W} \cap \mathcal{A}$,

$$\mathbf{P}^T \mathbf{W}_D \mathbf{P} = \sum_{i=1}^n w_i \mathbf{p}_i \mathbf{p}_i^T > k_2 \sum_{\{w_i \in \mathcal{S}\}} \mathbf{p}_i \mathbf{p}_i^T > \frac{k_2 k_1}{3} \mathbf{I}$$

This completes the proof. ■

Let $\mathbf{U}_B = \mathbf{P}^T \mathbf{W}_D \mathbf{P}$. Then Lemma 3.5.1 allows us to conclude that

$$\lambda_{\max}(\mathbf{U}_B^{-1}) I_{\mathcal{W} \cap \mathcal{A}} < k < \infty. \quad (3.5.5)$$

Using (3.5.5) we have a more precise statement about the nature of the matrix $\mathbf{U}_B^{-1}$.

Lemma 3.5.2 *On the set $\mathcal{W} \cap \mathcal{A}$*

$$\mathbf{U}_B^{-1} = \mathbf{I} + \sigma_n \mathbf{R}_W, \quad (3.5.6)$$

where $\lambda_{\max}(\mathbf{E}_B \mathbf{R}_W^2 I_{\mathcal{W} \cap \mathcal{A}}) = O(p^2 n^{-1})$.

Proof of Lemma 3.5.2. Since $\mathbf{U}_B = \mathbf{P}^T \mathbf{W}_D \mathbf{P}$, we have $\mathbf{U}_B = \mathbf{I} + \sigma_n \mathbf{P}^T \mathbf{W} \mathbf{P} = \mathbf{I} + \sigma_n \mathbf{R}_B$, say. Then it is easily seen that $\text{tr}(\mathbf{R}_B^2) = \text{tr}(\mathbf{W} \mathbf{P}_x \mathbf{P}_x^T \mathbf{W})$. From (3.5.1) it now follows that $\text{tr} \mathbf{E}_B(\mathbf{R}_B^2) = O(p^2/n)$, thus eventually all the eigenvalues of $\mathbf{E}_B \mathbf{R}_B$ are less than 1, and also $\lambda_{\max}(\mathbf{E}_B \mathbf{R}_B^2) = O(p^2 n^{-1})$. This means that by taking an inverse we can write $\mathbf{U}_B^{-1} = \mathbf{I} + \sigma_n \mathbf{R}_B \sum_{k \geq 0} \sigma_n^k \mathbf{R}_B^k = \mathbf{I} + \sigma_n \mathbf{R}_W$, say, then $\lambda_{\max}(\mathbf{E}_B \mathbf{R}_W^2 I_{\mathcal{W} \cap \mathcal{A}}) = O(p^2 n^{-1})$. ■

Proof of Theorem 3.3.1 Note that

$$\begin{aligned}
&\hat{\beta}_B - \hat{\beta} \\
&= [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_D \mathbf{e} - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] I_{\mathcal{W} \cap \mathcal{A}} \mathbf{e} \\
&= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{W}_D - \mathbf{I}) \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}} + [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1}] \mathbf{X}^T \mathbf{W}_D \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}}
\end{aligned}$$

$$\begin{aligned}
&= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{W}_D - \mathbf{I}) \mathbf{e} I_{W \cap \mathcal{A}} \\
&\quad + [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1}] \mathbf{X}^T (\mathbf{W}_D - \mathbf{I}) \mathbf{e} I_{W \cap \mathcal{A}} \\
&\quad + [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1}] \mathbf{X}^T \mathbf{e} I_{W \cap \mathcal{A}} \\
&= \sigma_n (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{e} I_{W \cap \mathcal{A}} + \sigma_n [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1}] \mathbf{X}^T \mathbf{W} \mathbf{e} I_{W \cap \mathcal{A}} \\
&\quad + [(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{X})^{-1}] \mathbf{X}^T \mathbf{e} I_{W \cap \mathcal{A}} \\
&= \sigma_n C_n I_{W \cap \mathcal{A}} + \sigma_n T_{1n} I_{W \cap \mathcal{A}} + T_{2n} I_{W \cap \mathcal{A}} \text{ (say)},
\end{aligned}$$

and thus

$$\sigma_n^{-1} (\hat{\beta}_B - \hat{\beta}) = C_n I_{W \cap \mathcal{A}} + T_{1n} I_{W \cap \mathcal{A}} + \sigma_n^{-1} T_{2n} I_{W \cap \mathcal{A}}.$$

Recall that $V_{\text{UBS}} = \sigma_n^{-1} E_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T$. We show that the contributing term in the representation of the bootstrap variance estimate comes from C_n , and the other terms are negligible. We need to compute $\xi^T V_{\text{UBS}} \xi$ for all $\{\xi \in \mathbb{R}^p : \|\xi\| = 1\}$. But since $\mathbf{Q}$ is an orthogonal matrix, we may as well write $\xi = \mathbf{Q}c$, and varying ξ over the unit sphere in $\mathbb{R}^p$ is equivalent to taking all $\{c \in \mathbb{R}^p : \|c\| = 1\}$. We show the following:

$$I_{\mathcal{A}} E_B \xi^T C_n C_n^T \xi = O_P(pn^{-1}), \quad (3.5.7)$$

$$I_{\mathcal{A}} E_B \xi^T C_n C_n^T \xi I_{W^c} = O_P(p^2 n^{-2}), \quad (3.5.8)$$

$$I_{\mathcal{A}} E_B \xi^T T_{1n} T_{1n}^T \xi I_W = O_P(\sigma_n^2 p^2 n^{-2}), \quad (3.5.9)$$

$$I_{\mathcal{A}} \sigma_n^{-2} E_B \xi^T T_{2n} T_{2n}^T \xi I_W = O_P(p^2 n^{-2}), \quad (3.5.10)$$

$$\sigma_n^{-1} I_{\mathcal{A}} E_B \xi^T C_n T_{2n}^T \xi I_W = O_P(p^2 n^{-2}), \quad (3.5.11)$$

$$I_{\mathcal{A}} E_B \xi^T C_n T_{1n}^T \xi I_W = O_P(p^2 n^{-2}). \quad (3.5.12)$$

The first term (3.5.7) is actually $I_{\mathcal{A}}[\xi^T(\hat{\beta} - \beta)]^2 + O_P(p^2 n^{-2})$. For Model 1 since $\mathcal{A}$ is the entire sample space, this implies

$$E_B \xi^T C_n C_n^T \xi - \xi^T V_n \xi = O_P(pn^{-3/2}) \quad (3.5.13)$$

is the leading term, with all other terms negligible. We now proceed to check (3.5.7)-(3.5.13).

PROOF OF (3.5.7) AND (3.5.13). With an elementary simplification using the singular value decomposition of $\mathbf{X}$, we have

$$E_B \xi^T C_n C_n^T \xi = E_B n^{-1} c^T \Lambda^{-1} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1} c. \quad (3.5.14)$$

Recall that $\mathbf{T}_n = \text{diag}(e_i^2)$. Now define the following:

$$\mathbf{M} = n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{T}_n \mathbf{P} \Lambda^{-1/2} c,$$

$$\mathbf{M}_1 = c_{11} n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{e} \mathbf{e}^T \mathbf{P} \Lambda^{-1/2} c,$$

$$\mathbf{M}_2 = c_{11} n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{T}_n \mathbf{P} \Lambda^{-1/2} c.$$

From (3.5.14) it can be easily seen that $E_B \xi^T C_n C_n^T \xi = \mathbf{M} + \mathbf{M}_1 - \mathbf{M}_2$. In proving the results we first note that $\mathbf{P}_x = \mathbf{P}\mathbf{P}^T$. If we define $\eta_b = \sum_{a=1}^p c_a \mathbf{P}_{ba} \lambda_a^{-1/2}$, where the a^{th} element of the vector c is c_a and the a^{th} diagonal entry in the diagonal matrix Λ is λ_a , then on $\mathcal{A}$ we have $\eta_b^2 \leq k h_{\max}$, where $h_{\max}$ is the maximum diagonal entry of $\mathbf{P}_x$, known to be $O(p/n)$ from (3.5.1), and k is some positive constant. Now note that $\mathbf{M} = n^{-1} \sum_{b=1}^n e_b^2 \eta_b^2$, $\mathbf{M}_1 = c_{11} n^{-1} (\sum_{b=1}^n e_b \eta_b)^2$, $\mathbf{M}_2 = c_{11} n^{-1} \sum_{b=1}^n e_b^2 \eta_b^2$. Hence we can conclude $I_A \mathbf{M}_1 = O_P(p^2 n^{-2})$ and $I_A \mathbf{M}_2 = O_P(p n^{-2})$ after some algebra involving (3.2.23), (3.2.5), (3.2.4), (3.5.1) and (3.5.2). Similarly, $I_A \mathbf{M}^2 = O_P(p^2 n^{-2})$, and note also that $E\mathbf{M} = \xi^T \mathbf{V}_n \xi$. This proves (3.5.7). For (3.5.13) we have under Model 1, $\mathbf{M} - \xi^T \mathbf{V}_n \xi = n^{-1} \sum_{b=1}^n (e_b^2 - \tau_b^2) \eta_b^2 = O_P(p n^{-3/2})$.

PROOF OF (3.5.8) We only need to look at $I_A E_B \mathbf{M} I_{\mathcal{W} \cap \mathcal{A}}$, since the other terms are of smaller order. Now $I_A E_B \mathbf{M} I_{\mathcal{W} \cap \mathcal{A}} = [I_A E_B \mathbf{M}^2]^{1/2} [1 - P_B(\mathcal{W})]^{1/2} = O_P(p^2 n^{-2})$ follows immediately from the proof of (3.5.7) and Assumption (3.2.22).

PROOF OF (3.5.9) Note that $T_{1n} = -\sigma_n (\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{e}$, so we have that $\xi^T T_{1n} I_{\mathcal{W} \cap \mathcal{A}} = -\sigma_n n^{-1/2} c^T \Lambda^{-1/2} \mathbf{U}_B^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P} \mathbf{P}^T \mathbf{W} \mathbf{e}$ and thus

$$\xi^T T_{1n} T_{1n}^T \xi I_{\mathcal{W} \cap \mathcal{A}} \quad (3.5.15)$$

$$= \sigma_n^2 n^{-1} \mathbf{e}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{U}_B^{-1} \Lambda^{-1/2} c c^T \Lambda^{-1/2} \mathbf{U}_B^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}} \quad (3.5.16)$$

$$\leq \sigma_n^2 n^{-1} \mathbf{e}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{U}_B^{-1} \Lambda^{-1} \mathbf{U}_B^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}} \quad (3.5.17)$$

$$\leq k \sigma_n^2 n^{-1} \mathbf{e}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{U}_B^{-2} \mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}} \quad (3.5.18)$$

$$\leq \sigma_n^2 n^{-1} \mathbf{e}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}} \quad (3.5.19)$$

$$\leq \sigma_n^2 n^{-1} \mathbf{e}^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} I_A \quad (3.5.20)$$

The steps involved in the above reduction are as follows: (3.5.17) follows from (3.5.16) since $c c^T < \mathbf{I}$; (3.5.18) follows from (3.5.17) since $\Lambda^{-1} < k \mathbf{I}$ on $\mathcal{A}$ from condition (3.2.5); (3.5.19) follows from (3.5.18) since, from Lemma 3.5.1, $\mathbf{U}_B^{-1} < k \mathbf{I}$ on the set $\mathcal{W} \cap \mathcal{A}$. The last inequality is obvious. Now routine but lengthy algebra shows that the bootstrap expectation of the expression in (3.5.20) is $O_P(\sigma_n^2 p^2 n^{-2})$.

PROOF OF (3.5.10) The proof of this part is similar to that of (3.5.9). We have $\sigma_n^{-1} T_{2n} = -(\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e}$, and hence $\sigma_n^{-1} \xi^T T_{2n} I_{\mathcal{W} \cap \mathcal{A}} = -n^{-1/2} c^T \Lambda^{-1/2} \mathbf{U}_B^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P} \mathbf{P}^T \mathbf{e} I_{\mathcal{W} \cap \mathcal{A}}$ and thus

$$\sigma_n^{-2} \xi^T T_{2n} T_{2n}^T \xi I_{\mathcal{W} \cap \mathcal{A}} \leq n^{-1} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{e} I_A \quad (3.5.21)$$

by following similar steps as in the previous part of the proof. Now again routine algebra shows that the bootstrap expectation of the expression in (3.5.21) is $O_P(p^2 n^{-2})$.

PROOF OF (3.5.11) We have to show $\sigma_n^{-1} E_B \xi^T C_n T_{2n}^T \xi I_{\mathcal{W} \cap \mathcal{A}} = O_P(p^2 n^{-2})$. Using

$$\sigma_n^{-1} \xi^T T_{2n} I_{W \cap A} = -n^{-1/2} c^T \Lambda^{-1/2} \mathbf{U}_B^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P}^T \mathbf{e} I_{W \cap A}, \text{ and}$$

$$\xi^T C_n I_{W \cap A} = n^{-1/2} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} I_{W \cap A},$$

we have

$$\begin{aligned} & \sigma_n^{-1} \xi^T C_n T_{2n}^T \xi I_{W \cap A} \\ &= n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{U}_B^{-1} \Lambda^{-1/2} c I_{W \cap A} \\ &= n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_{W \cap A} \\ & \quad + \sigma_n n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{R}_W \Lambda^{-1/2} c I_{W \cap A}, \end{aligned}$$

where Lemma 3.5.2 states the nature of the matrix $\mathbf{R}_W$. Thus we have

$$\begin{aligned} & \mathbf{E}_B \sigma_n^{-1} \xi^T C_n T_{2n}^T \xi I_{W \cap A} \\ & \leq I_A \mathbf{E}_B n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c \end{aligned} \quad (3.5.22)$$

$$+ \sigma_n n^{-1} \mathbf{E}_B c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{R}_W \Lambda^{-1/2} c I_{W \cap A}. \quad (3.5.23)$$

Consider (3.5.22) first:

$$\begin{aligned} \mathbf{E}_B n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c & \leq k n^{-1} \mathbf{E}_B \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{e} \\ & = \mathbf{M}_3 + \mathbf{M}_4 \end{aligned} \quad (3.5.24)$$

where $\mathbf{M}_3 = (1 - c_{11}) k n^{-1} \mathbf{e}^T \mathbf{P}_x \mathbf{M}_5 \mathbf{e}$, $\mathbf{M}_4 = c_{11} k n^{-1} \mathbf{e}^T \mathbf{P}_x \mathbf{e}$, $\mathbf{M}_5 = \text{diag}((\mathbf{P}_x)_{ii})$. The reduction to (3.5.24) follows from techniques similar to those used earlier, for example in case of (3.5.21). Now it is easily verified that $I_A \mathbf{M}_4$ is $O_P(p n^{-2})$. A direct computation using (3.2.8)-(3.2.10) shows that $I_A \mathbf{M}_3 = O_P(p^2 n^{-2})$.

Now consider (3.5.23):

$$\begin{aligned} & \sigma_n n^{-1} [\mathbf{E}_B c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P} \mathbf{R}_W \Lambda^{-1/2} c I_{W \cap A}] \\ & \leq \sigma_n n^{-1} [\mathbf{E}_B \|\mathbf{R}_W \Lambda^{-1/2} c I_{W \cap A}\|^2]^{1/2} [\mathbf{E}_B \|\mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_{W \cap A}\|^2]^{1/2} \end{aligned} \quad (3.5.25)$$

Now $\mathbf{E}_B \|\mathbf{R}_W \Lambda^{-1/2} c I_{W \cap A}\|^2 = c^T \Lambda^{-1/2} \mathbf{E}_B \mathbf{R}_W^2 I_{W \cap A} \Lambda^{-1/2} c$
 $\leq \lambda_{\max}(\mathbf{E}_B \mathbf{R}_W^2 I_{W \cap A}) \|\Lambda^{-1/2} c I_A\|^2 = O(p^2 n^{-1})$ and

$$\begin{aligned} & \mathbf{E}_B \|\mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_{W \cap A}\|^2 \\ &= \mathbf{E}_B \|\mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_A\|^2 \\ &= \mathbf{E}_B [\mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{e} I_A] [c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_A] \\ &\leq [\mathbf{E}_B (\mathbf{e}^T \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{e} I_A)^2 \mathbf{E}_B (c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{e} \mathbf{e}^T \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_A)^2]^{1/2} \\ &= O_P(p^{3/2} n^{-3/2}) = O_P(p^2 n^{-1}) \end{aligned}$$

The rates follow from routine calculations. This ensures that (3.5.25) is $O_P(p^2 n^{-2})$.

PROOF OF (3.5.12). We have to show $E_B \xi^T C_n T_{\ln}^T \xi I_{W \cap A} = O_P(p^2 n^{-2})$. As in the previous part of the proof, $E_B \sigma_n^{-1} \xi^T C_n T_{\ln}^T \xi I_{W \cap A}$

$$\leq \sigma_n E_B n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} e e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_A \quad (3.5.26)$$

$$+ \sigma_n^2 n^{-1} E_B c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} e e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c I_{W \cap A} \quad (3.5.27)$$

Consider (3.5.26) first:

$$\begin{aligned} & I_A E_B n^{-1} c^T \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} e e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c \\ & \leq k n^{-1} I_A E_B e^T \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} e \\ & = O_P(p^2 n^{-2}) \end{aligned}$$

by direct computation. For (3.5.27), we have the following:

$$\begin{aligned} & \|e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c\|^2 I_{W \cap A} \\ & = e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} c c^T \Lambda^{-1/2} \mathbf{R}_W^{-1} \mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{W} e I_{W \cap A} \\ & \leq k e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} \Lambda^{-1/2} \mathbf{P}^T \mathbf{W} \mathbf{P}_x \mathbf{W} e I_{W \cap A} \\ & \leq k p^2 n^{-1} e^T \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P}_x \mathbf{W} \mathbf{P} e I_A \quad (\text{by Lemma 3.5.2}) \end{aligned}$$

and this last is $O_P(p^4 n^{-2})$ from calculations as in (3.5.20). Therefore the quantity in (3.5.27) is $O_P(\sigma_n^2 p^2 n^{-2})$. This completes the proof of the theorem. ■

Proof of Proposition(3.3.1). We denote m_n by m and note that $\text{Prob}[\text{at least } m \text{ of the } X\text{'s are positive}] = 1 - \sum_{i=0}^{m-1} \text{Prob}[\text{exactly } i \text{ of the } X\text{'s are positive}]$. We can exclude the case $i = 0$, since at least one of the X 's is always positive. For $j = 1, \dots, m-1$,

$$\begin{aligned} & \text{Prob}[\text{exactly } j \text{ of the } X\text{'s are positive}] \\ & = \frac{n!}{n^n} \binom{n}{j} \sum_{a_1+a_2+\dots+a_j=n-j, a_i \geq 0} \frac{1}{(a_1+1)!(a_2+1)! \dots (a_j+1)!} \\ & \leq \frac{n! j^{(n-j)}}{n^n (n-j+1)(n-j)!} \binom{n}{j} \sum_{a_1+a_2+\dots+a_j=n-j, a_i \geq 0} \frac{(n-j)!}{(a_1)!(a_2)! \dots (a_j)!} \frac{1}{j^{n-j}} \\ & = \frac{n! n! j^{(n-j)}}{n^n (n-j+1)! j! (n-j)!}. \end{aligned}$$

The inequality in the middle comes from $(a_1+1)(a_2+1) \dots (a_j+1) \geq 1 + \sum_{i=1}^j a_i$, since all the a_i 's are non-negative, and then we use $\sum_{i=1}^j a_i = n-j$.

We can bound $\frac{n! n! j^{(n-j)}}{n^n (n-j+1)! j! (n-j)!}$ above by $n^{-1} \binom{n}{j} n^{-2j} \frac{j!}{j!}$, and using an upper bound in the Stirling's approximation of $\frac{j!}{j!}$ from Feller ((1968), page 54), we see that a further upper bound

is $n^{-1} \binom{j}{n}^{n-2j} \frac{e^j}{\sqrt{2\pi j}}$, and across $j = 1, \dots, m-1$, this is bounded above by $n^{-1} \left(\frac{m}{n}\right)^{n-2m} \frac{e^m}{\sqrt{2\pi}}$.

Now

$$\begin{aligned}
 \sum_{i=0}^{m-1} \text{Prob}[\text{exactly } i \text{ of the } X\text{'s are positive}] &\leq \sum_{i=0}^{m-1} n^{-1} \left(\frac{m}{n}\right)^{n-2m} \frac{e^m}{\sqrt{2\pi}} \\
 &\leq \left(\frac{m}{n}\right)^{n-2m} \frac{e^m}{\sqrt{2\pi}} \\
 &\leq \frac{1}{\sqrt{2\pi}} c_2^{n-2c_2n} e^{c_2n} \quad (\text{since } m \leq c_2n \leq n/2) \\
 &= \frac{1}{\sqrt{2\pi}} e^{-n[(1-2c_2)\Delta - c_2]} \quad (\text{putting } -\log(c_2) = \Delta)
 \end{aligned}$$

We only have to show that $\alpha = -(1-2c_2)\log(c_2) - c_2 > 0$, which is true if c_2 is small enough.

An elementary computation shows that $\alpha = 0$ near $c_2 = 0.34128$, and taking $c_2 = 1/3$ yields

$\alpha > 0$. ■

Chapter 4

COMPARISON OF BOOTSTRAP AND JACKKNIFE VARIANCE ESTIMATORS IN LINEAR REGRESSIONS: SECOND ORDER RESULTS

4.1 Introduction

Consider the oxygen uptake data from Rawlings, Pantula, Dickey (1998, page 124). It is a data across five variables measuring physical fitness of thirty-one persons, and the oxygen uptake by them. The authors make a regression modelling and have a hypothesis that two of the variables do not influence oxygen uptake. Consider the contrast of the difference of these two means. In order to apply standard theory, one must have an estimate of the variance of the estimated contrast, usually estimated by resampling techniques. We applied three resampling techniques on this data, all of which are consistent under unequal error variances namely the delete-1 jackknife, paired bootstrap and Wu's external bootstrap. The variance estimates are respectively 30.29, 30.85, 19.80. The difference between these estimates is too large to be ignored, and calls for a close comparison of different resampling techniques.

In this chapter we consider all the different resampling techniques that have been used for the problem of variance estimation, and compare in linear regression their higher order characteristics. We have mentioned earlier (Section 3.1) that for this problem resampling schemes classify in two groups according to the asymptotic expansion of their variance estimate. That is, they are either consistent under heteroscedasticity (hence *robust* or in *R*-class), or relatively efficient under homoscedasticity (hence *efficient* or in *E*-class). A finer distinction within either class may be made, in order to identify the scheme that is better than others. Such a distinction may be made by considering higher order terms in the asymptotic expansion for variance estimates.

In Chapter 3 we have already discussed the *UBS* variance estimator and shown that it is consistent under heteroscedasticity. Here we show that different members of this class exhibit different higher order behaviour. One important result in this chapter is that among the *R*-class resampling techniques, some *UBS* techniques outperform other resampling schemes in terms of second order efficiency. For the simple linear regression ($p = 1$) case we are actually able

to identify a best *UBS* technique. The technical conditions on the *UBS* weights used for this chapter are discussed later.

In the *E*-class, a *GBS* scheme (introduced in Section 2.6 of Chapter 2) is best, in the sense that the variance estimate has least mean squared error. Interestingly, this result does not depend on the dimension p of the parameter. This generalised bootstrap scheme has apparently not appeared in the literature earlier.

Following a discussion of the model in Section 4.2, we state our results for the simple linear regression case in Section 4.3. Sections 4.4 and 4.5 are respectively on *E*-class and *R*-class results for multiple regressions.

The optimal resampling schemes may have data dependent parameter values, and thus may be difficult to simulate. So the usefulness of considering a heuristically justified 'more or less optimal' technique has to be studied. In Section 4.6 we discuss some simulation experiments with some *R*-class resampling techniques to study these issues. It turns out that for small and medium sized data, the improvement in using a close to optimal technique over other techniques is considerable.

In order to investigate whether higher order terms actually matter in computation, we have carried out extensive simulations and analysed some real life data-sets in Section 4.6. An example from a real data set was given at the beginning of this chapter. We show simulated examples from *R*-class here. It turns out that for small and moderate sample sizes, and depending on the dimension of the problem and the signal to noise ratio, higher order terms are significantly different across resampling schemes. The *i.i.d.* *UBS* scheme that we recommend in Section 4.6 always seems to do at least as good as other resampling schemes. In addition they are computationally easier and potentially applicable to other models and problems where many of the other known schemes may not even be applicable. These issues are taken up in Chapter 7.

In Section 4.7 we prove the results of Section 4.3, and in Section 4.8 we prove the results of Sections 4.4 and 4.5.

This chapter is adapted from Chatterjee and Bose (1997, 1998a, 1998b).

4.2 Notations, assumptions and preliminaries

Suppose the regression model is given by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

where $\mathbf{y}$ is the $(n \times 1)$ observed vector of responses and $\mathbf{X}$ is the $(n \times p)$ size design matrix whose i^{th} row $\mathbf{x}_i^T$ is given by the i^{th} observation of the $(p \times 1)$ vector of explanatory variables. The parameter β is a $(p \times 1)$ vector of unknown constants and $\mathbf{e}^T = (e_1 \ e_2 \ \dots \ e_n)$ is the noise vector. Suppose $\hat{\beta}$ is the least squares estimate of β as calculated from the data. If $Ee_i^2 = \tau_i^2$

then

$$V_R = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{T} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (4.2.1)$$

where $\mathbf{T} = \text{diag}(\tau_i^2, 1 \leq i \leq n)$. Under homoscedasticity of the noise terms, *i.e.* when $\tau_i^2 = \tau^2$, $i = 1, \dots, n$, the expression for variance of $\hat{\beta}$ simplifies to

$$V_E = \tau^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (4.2.2)$$

For the rest of this chapter V_R and V_E will be as defined in (4.2.1) and (4.2.2), respectively.

Generally for multiple regression problems two common and natural assumptions on the design matrix are

$$d_i/n \in (m, M); \text{ for some } 0 < m \leq M < \infty \quad (4.2.3)$$

$$\mathbf{x}_i^T \mathbf{x}_i < K < \infty, \quad \forall i = 1, 2, \dots, n \quad (4.2.4)$$

where $d_i, 1 \leq i \leq p$ are the eigenvalues of $\mathbf{X}^T \mathbf{X}$. Assumption (4.2.4) appears in Liu and Singh (1992) and it essentially eliminates the possibility of an influential design point that can alter the asymptotics. However, in order to make the resampling schemes like the paired bootstrap and the different jackknives feasible on the model, we need slightly stronger assumptions than (4.2.3). This has been discussed in details in Chapter 3. We assume that

$$\frac{1}{m} \mathbf{X}^{T*} \mathbf{X}^* > k \mathbf{I}, \quad k > 0 \quad (4.2.5)$$

for some $m \geq n/3$, and all choices of $(m \times p)$ submatrix $\mathbf{X}^*$ that can be formed by choosing any m rows of $\mathbf{X}$. By $\mathbf{A} > \mathbf{B}$ we mean that $\mathbf{A} - \mathbf{B}$ is positive definite. As remarked by Wu (1986, page 1344), condition (4.2.5) seems necessary for most schemes that use the data-pairs for resampling. But let us emphasise that (4.2.5) and (4.2.4) are only sufficient conditions under which all known resampling schemes work. Individual resampling schemes may work under weaker conditions. It turns out that the best resampling plans do work under (4.2.3). For the noise terms we make the assumption that the e_i 's are independent with bounded eighth moment. Again this is only a sufficient condition, actual requirement for most resampling schemes being weaker than this.

For transparency, we consider first the simple linear regression. By comparing the second order terms of the variance estimates, we show that the weighted jackknife of Liu and Singh (1992) is more efficient than the residual bootstrap in the E -class. The GBS is not considered in this case, but it is considered in the multiple linear regression case.

We first show that the weighted jackknives of Hinkley (1977) and Wu (1986) also belong to the R -class. For the model with heteroscedastic errors, under the additional but natural model condition

$$2 \sum_{i=1}^n x_i^2 \sum_{i=1}^n x_i^4 \tau_i^2 > \sum_{i=1}^n x_i^4 \sum_{i=1}^n x_i^2 \tau_i^2 \quad (4.2.6)$$

a *UBS* scheme can be chosen for which the second order term of the variance estimate has least mean squared error compared to other resampling schemes in the *R*-class.

We call a resampling scheme *strongly optimal* if it has least mean squared error compared to any other resampling scheme in its class. We show that *UBS* with symmetric *i.i.d.* weights is strongly optimal in *R*-class under the given conditions in simple linear regression.

We then consider the multiple regression problem. Among the usual resampling techniques which are applicable to the multiple regression problem, only the residual bootstrap belongs to the *E*-class. However, we show that under certain conditions, *GBS* techniques also belong to the *E*-class, and this broadens the *E*-class considerably. It turns out that there is a *GBS* scheme which is strongly optimal, in the sense that the variance estimate has least mean squared error.

As mentioned above, for simple linear regression a *UBS* technique can be found which is better than any other technique in the *R*-class. Such universal dominance may not be attainable in multiple regression. A class of resampling schemes *weakly optimal* if for any fixed resampling scheme there is one member of the class whose mean squared error is less than that of the fixed scheme. However, we show that for any specified resampling technique, a class of *UBS* techniques can be chosen which is weakly optimal. This establishes a limited superiority of the *UBS* over other resampling techniques.

The behaviour of different resampling techniques discussed above is summarised in the following table.

TABLE : CLASSIFICATION OF RESAMPLING TECHNIQUES

	<i>E</i> -class	<i>R</i> -class
Simple Regression	Residual Bootstrap, Weighted Jackknife (Liu, Singh), STRONGLY OPTIMAL: <i>i.i.d. GBS</i>	External Bootstrap, Weighted Bootstrap (Liu), Weighted Jackknife (Wu), Weighted Jackknife (Hinkley), Paired Bootstrap, Delete- <i>d</i> Jackknife, $d \geq 1$ STRONGLY OPTIMAL: <i>i.i.d. UBS</i>
Multiple Regression	Residual Bootstrap, STRONGLY OPTIMAL: <i>i.i.d. GBS</i>	External Bootstrap, Weighted Bootstrap (Liu) , Weighted Jackknife (Wu), Weighted Jackknife (Hinkley), Paired Bootstrap, Delete- <i>d</i> Jackknife, $d \geq 1$ WEAKLY OPTIMAL: <i>i.i.d. UBS</i>

We now discuss the conditions on *UBS* weights relevant to this chapter. The conditions are similar to those of Chapter 3, but higher moment conditions are needed here.

Recall that $\mathcal{W}$ is the set on which at least m_0 of the weights are greater than some fixed constant $k_2 > 0$. The value of m_0 is $\geq n/3$. The weights are required to satisfy

$$E(w_i) = 1 \quad (4.2.7)$$

$$P_B[Kn \geq \sum_{i=1}^n w_i \geq kn, \quad K > k > 0] = 1 \quad (4.2.8)$$

$$P_B[\mathcal{W}] = 1 - O(p^2 n^{-1}) \quad (4.2.9)$$

$$K > \sigma_n^2 > 0 \quad (4.2.10)$$

$$c_{11} = O(n^{-1}) \quad (4.2.11)$$

$$\forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 3, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+1} \sigma_n^{-1}) \quad (4.2.12)$$

$$\forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 4, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+2}) \quad (4.2.13)$$

$$\forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 6, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+3}) \quad (4.2.14)$$

$$\forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 8, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+4}) \quad (4.2.15)$$

4.3 Second order results for $p = 1$

Throughout this section we assume $p = 1$ and one dimensional versions of (4.2.3)-(4.2.5). This is equivalent to assuming that for all i , $0 < k < |x_i| < K < \infty$. We also assume that the errors are independent with bounded sixth moments. As mentioned earlier, some of the schemes below will work under weaker restrictions. This section is for an easier understanding of the results to come for general p , and also to relate to the first order results of Liu and Singh (1992). Hence we do not consider the *GBS* scheme here but in the next section, and separately compute the results for the delete-1 jackknife, which is a special case of *UBS*.

We use the notation $L_n = \sum_{i=1}^n x_i^2$. If $E\epsilon_i^2 = \tau_i^2$, then

$$V_n = \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 \tau_i^2 \quad (4.3.1)$$

Under homoscedasticity of the noise terms, *i.e.* when $\tau_i^2 = \tau^2$, $i = 1, \dots, n$, the expression for variance of $\hat{\beta}$ simplifies to

$$v_n = \frac{1}{L_n} \tau^2 \quad (4.3.2)$$

We assume the existence of the following limits.

$$\alpha = \lim_{n \rightarrow \infty} \left(\frac{n}{L_n} \right)^{1/2}; \quad \alpha_0 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n x_i^2 \tau_i^2; \quad \alpha_1 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n x_i^4 \tau_i^2;$$

$$\alpha_2 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n x_i^6 \tau_i^2; \quad \gamma_1 = \lim_{n \rightarrow \infty} \frac{\sum x_i}{n}; \quad \gamma_2 = \lim_{n \rightarrow \infty} \frac{\sum x_i^4}{n}; \quad \gamma_3 = \lim_{n \rightarrow \infty} \frac{\sum x_i^6}{n}$$

$$\text{Let } Y_n = n^{1/2} \frac{1}{L_n} \sum_{i=1}^n (e_i^2 - \tau^2) \quad Z_n = n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2)$$

Recall that the definitions of variance estimators of resampling schemes are given in Chapter 2. Recall also that the residual bootstrap and the weighted jackknife of Liu and Singh (1992) belong to the E -class. In our first theorem we deal with the asymptotic expansions of these two resampling techniques.

Theorem 4.3.1 *Suppose the error terms have the same variance.*

(i) *(Residual Bootstrap) The variance estimate of the residual bootstrap satisfies*

$$n^{3/2}(V_{GB2} - v_n) = Y_n - \frac{1}{n^{1/2}} [\alpha^2 \left(\frac{\sum e_i}{n^{1/2}} \right)^2 - 2\alpha^3 \gamma_1 \left(\frac{\sum e_i}{n^{1/2}} \right) \left(\frac{\sum x_i e_i}{(\sum x_i^2)^{1/2}} \right) + (\alpha^2 + \alpha^4 \gamma_1^2) \left(\frac{\sum x_i e_i}{(\sum x_i^2)^{1/2}} \right)^2] + O_p(n^{-1})$$

(ii) *(Weighted Jackknife Liu and Singh (1992)) The variance estimate of Liu and Singh's weighted jackknife satisfies*

$$n^{3/2}(V_{JLS} - v_n) = Y_n - n^{-1/2} [\alpha^2 \left(\frac{\sum_{i=1}^n x_i e_i}{(\sum_{i=1}^n x_i^2)^{1/2}} \right)^2 - \alpha^2 \tau^2] + O_p(n^{-1})$$

A simple computation shows that the second order term for the residual bootstrap has a greater mean squared error than the second order term for the weighted jackknife of Liu and Singh (1992). However, the weighted jackknife can be bettered by a strongly optimal GBS technique. We prove this in Section 4.4.

We introduce some notation to present the results for the R -class. Let

$$\eta_1 = \frac{\sum x_i e_i}{(\sum x_i^2)^{1/2}}, \quad \eta_2 = \frac{\sum x_i^3 e_i}{(\sum x_i^6)^{1/2}} \quad \text{and} \quad \eta = (\eta_1 \quad \eta_2)^T$$

$$\mathbf{H} = \begin{pmatrix} \alpha^6 \gamma_2 & -\alpha^5 \gamma_3^{1/2} \\ -\alpha^5 \gamma_3^{1/2} & 0 \end{pmatrix}, \quad \mathbf{Q} = \eta^T \mathbf{H} \eta$$

Theorem 4.3.2 (i) (*Delete-1 Jackknife*) The variance estimate from the delete-1 jackknife satisfies

$$n^{3/2}(V_J - V_n) = Z_n + n^{-1/2}[\mathbf{Q} - \alpha^4\alpha_0 + 2\alpha^6\alpha_1] + O_P(n^{-1})$$

(ii) (*External bootstrap* Wu (1986) and *Weighted Bootstrap* Liu (1988)) The variance estimate from the external bootstrap and weighted bootstrap satisfies

$$n^{3/2}(V_{GB3} - V_n) = n^{3/2}(V_{WB} - V_n) = Z_n + n^{-1/2}\mathbf{Q} + O_P(n^{-1})$$

(iii) (*Weighted Jackknife* Hinkley (1977)) The variance estimate from Hinkley's weighted jackknife satisfies

$$n^{3/2}(V_{JH} - V_n) = Z_n + n^{-1/2}[\mathbf{Q} + \alpha^4\alpha_0 + O_P(n^{-1})]$$

(iv) (*Weighted Jackknife* Wu (1986)) The variance estimate from Wu's weighted jackknife satisfies

$$n^{3/2}(V_{JW} - V_n) = Z_n + n^{-1/2}[\mathbf{Q} + \alpha^6\alpha_1] + O_P(n^{-1})$$

(v) (*Uncorrelated Weights Bootstrap*) The variance estimate from the UBS under (4.2.7)-(4.2.15) satisfies

$$n^{3/2}(V_{UBS} - V_n) = Z_n + n^{-1/2}T_n + O_P(n^{-1})$$

where the second order term is given by

$$T_n = \mathbf{Q} + 3\sigma_n^2 c_{22} \alpha^8 \alpha_0 \gamma_2 - 2\sigma_n c_3 \alpha^6 \alpha_1 + n(3\sigma_n^2 c_{112} - 2\sigma_n c_{12} - c_{11}) \alpha^4 \alpha_0$$

The various mixed-moments determine the term T_n in Theorem 4.3.2 (v), and hence the exact second order value depends on the sequence of random weights used in UBS. For particular choice of weights, the above expression simplifies considerably, and we present some important special cases as a corollary.

Corollary 4.3.1 (a) The variance estimate from the paired bootstrap satisfies

$$n^{3/2}(V_{UBS} - V_n) = Z_n + n^{-1/2}[\mathbf{Q} + 3\alpha^8\alpha_0\gamma_2 - 2\alpha^6\alpha_1] + O_P(n^{-1})$$

(b) The variance estimate from the UBS with i.i.d. weights satisfies

$$n^{3/2}(V_{UBS} - V_n) = Z_n + n^{-1/2}[\mathbf{Q} + 3\sigma_n^2 \alpha^8 \alpha_0 \gamma_2 - 2\sigma_n c_3 \alpha^6 \alpha_1] + O_P(n^{-1})$$

In particular, if the weights are symmetric about their mean, then the second order term is $\mathbf{Q} + 3\sigma_n^2 \alpha^8 \alpha_0 \gamma_2$.

Observe that all the second order terms in Theorem 4.3.2 and Corollary 4.3.1 are of the form $\mathbf{Q} + c$, for some constant c depending on the model conditions and the resampling technique involved. Thus the mean squared error for such terms is $\text{Var}(\mathbf{Q}) + (\mu + c)^2$, where $E\mathbf{Q} = \mu = \alpha^6(\alpha^2\alpha_0\gamma_2 - 2\alpha_1)$. The minimum mean squared error is achieved when $\mu + c = 0$. We say that a resampling scheme is asymptotically *second order optimal* if this happens. If $\gamma_2 = \alpha = 1$, then the delete-1 jackknife is second order optimal. If $\alpha_1 = \alpha^2\alpha_0\gamma_2$, then the paired bootstrap is second order optimal. Note that the second order term for *UBS* depends on σ_n^2 and c_3 . So the performance of any *UBS* scheme can be characterised by the second and third moments of the weights. For a *UBS* resampling scheme with *i.i.d.* weights to be second order optimal, the following relation must be satisfied.

$$\alpha^2\alpha_0\gamma_2(1 + 3\sigma_n^2) = 2\alpha_1(1 + \sigma_n c_3). \quad (4.3.3)$$

Suppose the additional model condition (4.2.6) holds. If we assume that the weights come from a distribution that is not skewed, and their variance is given by

$$\sigma_n^2 = \frac{2\alpha_1 - \alpha^2\alpha_0\gamma_2}{3\alpha^2\alpha_0\gamma_2} \quad (4.3.4)$$

then the resulting *UBS* is second order efficient. The additional model condition (4.2.6) comes from requiring that $2\alpha_1 > \alpha^2\alpha_0\gamma_2$. Thus it is seen that under very general conditions, a *UBS* resampling technique exists that is superior to other techniques. Under other less general model conditions we have shown that the paired bootstrap is also optimal. The delete-1 jackknife is also optimal under certain other restrictions on the model. These are other examples of *UBS*.

Consider the problem of estimation of the common unknown mean if all the error variances are equal. Then all the x_i 's are equal to 1. Then $Y_n = Z_n = n^{-1/2} \sum (e_i^2 - \sigma_i^2)$. So in this particular model the distinction between *E*-class and *R*-class does not remain and all the resampling schemes have the same first order behaviour. The second order term for residual bootstrap, weighted bootstrap, and external bootstrap is $-n^{-1}(\sum_{i=1}^n e_i)^2$. For the delete-1 jackknife, weighted jackknife (Liu and Singh), weighted jackknife (Hinkley), weighted jackknife (Wu) and paired bootstrap, the second order term is $-n^{-1}(\sum_{i=1}^n e_i)^2 + n^{-1} \sum_{i=1}^n e_i^2$. Note that in this special case (4.2.6) is satisfied, and the variance requirement in (4.3.4) is that $\sigma_n^2 = 1/3$. Consider the two point discrete distribution supported with equal probability on $1 \pm \sqrt{\frac{2}{3}}$. An *UBS* with *i.i.d.* weights from this distribution is efficient. The different resampling schemes with second order term equal to $-n^{-1}(\sum_{i=1}^n e_i)^2 + n^{-1} \sum_{i=1}^n e_i^2$ are all efficient. This, in particular proves the superiority of the delete-1 jackknife over the residual bootstrap for the restricted model under consideration.

4.4 E-class bootstrap: second order results for $p > 1$

In this section we discuss the conditions under which a *GBS* belongs to the *E*-class. Assume that the elements W_{ij} of $\mathbf{W}_0$ satisfy

$$E W_{ij} = 0, \quad 1 \leq i, j \leq n \quad (4.4.1)$$

$$Var W_{ij} = \sigma_n^2 > 0, \quad 1 \leq i, j \leq n \quad (4.4.2)$$

$$Corr(W_{ij}, W_{ik}) = c_{11} = O(n^{-1}), \quad j \neq k, \quad \forall 1 \leq i \leq n \quad (4.4.3)$$

$$Corr(W_{ij}, W_{lk}) = 0, \quad 1 \leq i \neq l \leq n, \quad \forall j, k \quad (4.4.4)$$

As in the case of UBS resampling schemes, the most crucial conditions are (4.4.3) and (4.4.4).

For the residual bootstrap, observe that $\mathbf{W}_0 = \mathbf{W} - \mathbf{J}$. Also, rows of $\mathbf{W}$ are independent. So the above conditions are satisfied with

$$E_B \mathbf{W} = \mathbf{J} \quad (4.4.5)$$

$$\sigma_n^2 = \frac{n-1}{n^2} \quad \forall 1 \leq i, j \leq n \quad (4.4.6)$$

$$c_{11} = -\frac{1}{n-1}, \quad j \neq k, \quad \forall 1 \leq i \leq n \quad (4.4.7)$$

Theorem 4.4.1 shows the consistency of the *GBS* and establishes the second order results under the above assumptions. Recall that the definitions of variance estimators considered here are given in Chapter 2. Let us define

$$d_1 = \tau^2[p + c_{11} \sum_{i \neq j} ((\mathbf{P}_x))_{ij}], \quad d_2 = \tau^2[p + c_{11} \sum_{i \neq j} ((\mathbf{P}_x))_{ij} - (n^2 \sigma_n^2 - n)]$$

Theorem 4.4.1 *Suppose that the multiple linear regression model satisfies conditions (4.2.3) - (4.2.5) and the errors are independent with finite eighth moment. If the resampling is done by using weights satisfying the conditions (4.4.1)-(4.4.4), and the residuals $\mathbf{R}$, then the following expansions hold for the two variance estimators given in (2.6.3) and (2.6.4)*

$$V_{GB1} - V_E = \left\{ \frac{1}{n^{3/2}} \frac{\sum (e_i^2 - \tau^2)}{n^{1/2}} - \frac{d_1}{n^2} + O_P(n^{-5/2}) \right\} \{n \mathbf{G}^{-1}\}$$

$$V_{GB2} - V_E = \left\{ \frac{1}{n^{3/2}} \frac{\sum (e_i^2 - \tau^2)}{n^{1/2}} - \frac{d_2}{n^2} + O_P(n^{-5/2}) \right\} \{n \mathbf{G}^{-1}\}$$

Thus V_{GB1} is consistent, and V_{GB2} is consistent if $\lim_{n \rightarrow \infty} n \sigma_n^2 = 1$.

Theorem 4.4.1 shows the consistency of both variance estimators using the *GBS* in homoscedastic linear regression models. Note that the variance estimator V_{GB1} is very general and is valid for all choices of σ_n^2 . However V_{GB2} requires $n \sigma_n^2 \rightarrow 1$. We now state a corollary of Theorem 4.4.1 to take care of the important special case of residual bootstrap. This corollary

uses the identification of the usual residual bootstrap as a *GBS* with the variance estimator being V_{GB2} . Let

$$d_3 = \tau^2[p + 1 - n^{-1} \sum_{i \neq j} ((P_x))_{ij}], \quad d_{13} = \tau^2[2 - n^{-1} \frac{(\sum x_i)^2}{\sum x_i^2}]$$

Corollary 4.4.1 *Suppose resampling is done by using the classical residual bootstrap. Then the following expansions hold for the variance estimator in multiple and simple linear regression respectively*

$$\begin{aligned} V_{GB1} - V_E &= \left\{ \frac{1}{n^{3/2}} \frac{\sum (e_i^2 - \tau^2)}{n^{1/2}} - \frac{d_3}{n^2} + O_P(n^{-5/2}) \right\} \{n\mathbf{G}^{-1}\} \\ V_{GB2} - V_E &= \left\{ \frac{1}{n^{3/2}} \frac{\sum (e_i^2 - \tau^2)}{n^{1/2}} - \frac{d_{13}}{n^2} + O_P(n^{-5/2}) \right\} \left\{ \frac{n}{\sum x_i^2} \right\} \end{aligned}$$

The second part of Corollary 4.4.1 matches the corresponding calculations in Section 4.3. Based on the second order terms of the variance estimate, a comparison of different resampling schemes can be made. As in Section 4.3, we call a resampling scheme second order optimal if the second order term in the variance estimate of that resampling scheme is smaller in absolute value in comparison with other schemes. We now discuss a special *GBS* scheme, that is always second order optimal, in the sense that its second order term is identically zero. Suppose $\{W_{ij}, 1 \leq i, j \leq n\}$ are i.i.d. random variables with mean zero and variance $\sigma_n^2 = \frac{n+p}{n^2}$. These random variables are easily seen to satisfy (4.4.1)-(4.4.4).

Theorem 4.4.2 *Under the conditions of Theorem 4.4.1, if a *GBS* resampling is done with i.i.d. mean zero weights having variance $\sigma_n^2 = \frac{n+p}{n^2}$, and if the bootstrap variance estimator is $V_{GB2} = E_B(\hat{\beta}_B - \hat{\beta})(\hat{\beta}_B - \hat{\beta})^T$, then*

$$V_{GB2} - V_E = \left\{ \frac{1}{n^{3/2}} \frac{\sum (e_i^2 - \tau^2)}{n^{1/2}} + O_P(n^{-5/2}) \right\} \{n\mathbf{G}^{-1}\}$$

*Thus, this *GBS* technique is always second order optimal.*

As stated earlier, for homoscedastic linear regression models, the *E*-class resampling schemes are better than the *R*-class ones by a comparison of the first order terms. Within the *E*-class, Theorem 4.4.2 shows that i.i.d. weights *GBS* is the best. Apart from the efficiency consideration, let us also emphasise the significant computational advantage of using i.i.d. weights over other resampling methods.

4.5 R class bootstrap: second order results for $p > 1$

Liu and Singh (1992) showed that the paired bootstrap, the delete-1 jackknife, the external or wild bootstrap and the weighted bootstrap belong to the *R*-class. In Section 4.3 we showed

that the weighted jackknives proposed by Hinkley (1977) and Wu (1986) also belong to the R -class. From Chapter 3 it follows that all the different delete- d jackknives belong to the R -class. We have already commented earlier that the paired bootstrap and the delete- d jackknives are special cases of the UBS . Recall that the external bootstrap is a GBS technique for which $\mathbf{W}$ is diagonal with entries that are *i.i.d.* observations from a random variable with zero mean and finite variance. It has been observed earlier (Liu and Singh (1992), and Section 4.3 here) that the variance estimator for the weighted bootstrap of Liu (1988) is same as that of the external bootstrap, so we mention the result for the external bootstrap only. In this section we will study the second order properties of the external bootstrap, the UBS , and the weighted jackknives.

Let us introduce the following notation

$$\begin{aligned}\mathbf{T} &= \text{diag}(\tau_i^2, 1 \leq i \leq n) \\ \mathbf{T}_n &= \text{diag}(c_i^2, 1 \leq i \leq n) \\ \mathbf{D} &= \text{diag}(-2\tau_i^2((\mathbf{P}_x)_{ii} + \sum_{a=1}^n \tau_a^2((\mathbf{P}_x)_{ia}^2), 1 \leq i \leq n) \\ \delta_i &= \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \\ \mathbf{D}_{Wu} &= \text{diag}(\delta_i \tau_i^2, 1 \leq i \leq n) \\ \mathbf{F} &= \mathbf{X}^T (\mathbf{T}_n - \mathbf{T}) \mathbf{X} \\ \mathbf{S} &= \mathbf{X}^T \mathbf{D} \mathbf{X}\end{aligned}$$

Theorem 4.5.1 Suppose that the multiple linear regression model satisfies conditions (4.2.3) - (4.2.5) and the errors are independent with finite eighth moment. Then the following expansions hold for the variance estimate

(i) (External Bootstrap)

$$\begin{aligned}V_{GB3} - V_R &= \frac{1}{n^{3/2}} \{n\mathbf{G}^{-1}\} \{n^{-1/2}\mathbf{F}\} \{n\mathbf{G}^{-1}\} \\ &\quad + n^{-2} \{n\mathbf{G}^{-1}\} \{\mathbf{S}\} \{n\mathbf{G}^{-1}\} + O_P(n^{-5/2})\end{aligned}$$

(ii) Weighted Jackknife (Hinkley)

$$\begin{aligned}V_{JH} - V_R &= \frac{1}{n^{3/2}} \{n\mathbf{G}^{-1}\} \{n^{-1/2}\mathbf{F}\} \{n\mathbf{G}^{-1}\} \\ &\quad + n^{-2} \{n\mathbf{G}^{-1}\} \left\{ \frac{p}{(n-p)} \mathbf{X}^T \mathbf{T} \mathbf{X} + \mathbf{S} \right\} \{n\mathbf{G}^{-1}\} + O_P(n^{-5/2})\end{aligned}$$

(iii) Weighted Jackknife (Wu)

$$\begin{aligned}V_{JW} - V_R &= \frac{1}{n^{3/2}} \{n\mathbf{G}^{-1}\} \{n^{-1/2}\mathbf{F}\} \{n\mathbf{G}^{-1}\} \\ &\quad + n^{-2} \{n\mathbf{G}^{-1}\} \{ \mathbf{X}^T \mathbf{D}_{Wu} \mathbf{X} + \mathbf{S} \} \{n\mathbf{G}^{-1}\} + O_P(n^{-5/2})\end{aligned}$$

(iv) UBS satisfying (4.2.7)-(4.2.15)

$$V_{\text{UBS}} - V_{\text{R}} = \frac{1}{n^{3/2}} \{n\mathbf{G}^{-1}\} \{n^{-1/2}\mathbf{F}\} \{n\mathbf{G}^{-1}\} \\ + n^{-2} \{n\mathbf{G}^{-1}\} \{\mathbf{X}^T \mathbf{M} \mathbf{X}\} \{n\mathbf{G}^{-1}\} + O_p(n^{-5/2}) \text{ where}$$

$$\mathbf{M}_{ii} = \sigma_n^2 [3c_4 \tau_i^2 ((\mathbf{P}_{\mathbf{x}})_{ii}^2) + c_{22} \sum_{a \neq i} \tau_a^2 ((\mathbf{P}_{\mathbf{x}})_{ia}^2) + \sum_{a=1}^n \tau_a^2 (\mathbf{P}_{\mathbf{x}})_{ia}^2 - 2\sigma_n c_3 \tau_i^2 (\mathbf{P}_{\mathbf{x}})_{ii}^2 \\ - 2\tau_i^2 (\mathbf{P}_{\mathbf{x}})_{ii} + 2\sigma_n^2 c_{31} \tau_i^2 \sum_{a \neq i} ((\mathbf{P}_{\mathbf{x}})_{ia}^2) \\ \mathbf{M}_{ij} = \sigma_n^2 [2c_{31} \{\tau_i^2 ((\mathbf{P}_{\mathbf{x}})_{ii} ((\mathbf{P}_{\mathbf{x}})_{ij} + \tau_i^2 ((\mathbf{P}_{\mathbf{x}})_{jj} ((\mathbf{P}_{\mathbf{x}})_{ij})_{ij}) \\ + c_{211} \sum_{a \neq i,j} \tau_a^2 ((\mathbf{P}_{\mathbf{x}})_{ia} ((\mathbf{P}_{\mathbf{x}})_{ja} + c_{211} (\tau_i^2 + \tau_j^2) \sum_{a \neq i,j} ((\mathbf{P}_{\mathbf{x}})_{ia} ((\mathbf{P}_{\mathbf{x}})_{ja} \\ + c_{22} \{\tau_j^2 ((\mathbf{P}_{\mathbf{x}})_{ii} ((\mathbf{P}_{\mathbf{x}})_{ij} + \tau_i^2 ((\mathbf{P}_{\mathbf{x}})_{jj} ((\mathbf{P}_{\mathbf{x}})_{ij})_{ij})\} \\ + c_{11} (\sum_{a=1}^n \tau_a^2 (\mathbf{P}_{\mathbf{x}})_{ia} (\mathbf{P}_{\mathbf{x}})_{ja}) \\ - c_{11} [\tau_j^2 + \tau_i^2] (\mathbf{P}_{\mathbf{x}})_{ij} + \sigma_n [c_{21} (\tau_i^2 + \tau_j^2) (\mathbf{P}_{\mathbf{x}})_{ij}]$$

The exact expression for the different delete- d jackknives and the paired bootstrap schemes can be calculated by imputing the exact values of the different moments involved. Let us compare the second order terms of the above expressions from external bootstrap, Hinkley's weighted jackknife and Wu's weighted jackknife. It can be easily seen that both these weighted jackknives are better than the external bootstrap, as seen for $p = 1$ case. But it is also of interest to see how UBS with *i.i.d.* weights performs, since that yielded the second order optimal results for $p = 1$, based on a particular choice of σ_n^2 (see Section 4.3). For general p , this involves expectation of matrices, and we have to see how close different matrices are to the $(p \times p)$ matrix whose all entries are zero. We take two criteria for comparing matrices. .

The first one is based on quadratic forms. Consider two $(n \times n)$ matrices $\mathbf{A}$ and $\mathbf{B}$. If both are positive definite, say $\mathbf{B}$ is closer to the zero matrix compared to $\mathbf{A}$ if $\mathbf{A} - \mathbf{B}$ is positive definite. If $\mathbf{A}$ is positive definite and $\mathbf{B}$ negative definite, we say $\mathbf{B}$ is closer to zero than $\mathbf{A}$ if $\mathbf{A} + \mathbf{B}$ is positive definite. Essentially, the criterion checks whether for every quadratic form

$$\text{Criterion 1} \quad |\xi^T \mathbf{A} \xi| > |\xi^T \mathbf{B} \xi| \quad (4.5.1)$$

Criterion 1 is based on a comparison of all the eigenvalues of $\mathbf{B}$ and $\mathbf{A}$. If either of the matrices has both positive and negative eigenvalues then instead of comparing every quadratic form, we can use a criterion that reflects the average behaviour of quadratic forms. The second criterion is a relaxation of (4.5.1), where we desire that the absolute value of the average of eigenvalues of $\mathbf{A}$ is greater than the corresponding quantity of $\mathbf{B}$.

Criterion 2

$$|tr(\mathbf{A})| \geq |tr(\mathbf{B})| \quad (4.5.2)$$

Here $tr(\mathbf{A})$ denotes the trace of $\mathbf{A}$. Trace induces an inner product and hence a norm on $\mathcal{M}_n$, the space of all $(n \times n)$ matrices by $\langle \mathbf{M}_1, \mathbf{M}_2 \rangle = tr(\mathbf{M}_1 \mathbf{M}_2^T)$. Note that the matrices $\mathbf{A}$ and $\mathbf{B}$ that occur as second order terms in the variance are symmetric matrices, so if we take a decomposition $\mathbf{A} = \mathbf{A}_1 \mathbf{A}_1^T$ and $\mathbf{B} = \mathbf{B}_1 \mathbf{B}_1^T$, the trace criterion compares the Euclidean norms of $\mathbf{A}_1$ and $\mathbf{B}_1$, viewed as vectors in $\mathbb{R}^{n^2}$. Apart from the above two criteria, other possible criteria for comparison are the absolute values of determinants of $\mathbf{A}$ and $\mathbf{B}$, or the maximum and minimum eigenvalues. However, in our case both $\mathbf{A}$ and $\mathbf{B}$ involve sums of matrices, and these criteria are not easily applicable.

We call a resampling scheme *second order better* than another resampling scheme if the second order term of the variance of the first one is closer to the zero matrix than that of the latter. We will explicitly mention which criterion we use for comparison. In the multiple regression we provide a range of values for σ_n^2 , depending on the maximum and minimum eigenvalues of certain matrices, for which the *i.i.d. UBS* performs better than any other specific resampling scheme. Whether there is a single choice of σ_n^2 with which *i.i.d. UBS* will perform uniformly better depends on the design. We make the additional model assumptions

$$\mathbf{S} = \mathbf{X}^T \mathbf{D} \mathbf{X} \leq 0 \quad a.s. \quad (4.5.3)$$

$$\mathbf{X}^T \mathbf{U}_2 \mathbf{X} \geq 0 \quad a.s. \quad (4.5.4)$$

where $\mathbf{U}_2 = diag(\sum_{a \neq i} \tau_a^2((\mathbf{P}_x)_{ia}^2) - 2\tau_i^2((\mathbf{P}_x)_{ii}^2))$. For the multiple regression model assumption (4.2.6) translates to (4.5.3), and the one-dimensional equivalent of (4.5.4) is a triviality. Before stating the theorem on optimality of the *i.i.d. UBS*, let us introduce some more notation. Let

$$\mathbf{U}_{ij} = \begin{cases} \sum_{a \neq i} \tau_a^2((\mathbf{P}_x)_{ia}^2) & \text{if } i = j \\ \{\tau_j^2((\mathbf{P}_x)_{ii}((\mathbf{P}_x)_{ij} + \tau_i^2((\mathbf{P}_x)_{jj}((\mathbf{P}_x)_{ij})\} & \text{if } i \neq j \end{cases}$$

$$\lambda_{\max \mathbf{A}} = \max_{\xi^T \xi = 1} \xi^T \mathbf{X}^T \mathbf{A} \mathbf{X} \xi, \quad \lambda_{\min \mathbf{A}} = \min_{\xi^T \xi = 1} \xi^T \mathbf{X}^T \mathbf{A} \mathbf{X} \xi$$

for any symmetric matrix $\mathbf{A}$. It may be observed that the variance expression for *i.i.d. UBS* as stated in Theorem 4.5.1 is

$$\begin{aligned} V_{\text{UBS}} - V_{\text{R}} &= \frac{1}{n^{3/2}} \{n\mathbf{G}^{-1}\} \{n^{-1/2}\mathbf{F}\} \{n\mathbf{G}^{-1}\} \\ &\quad + n^{-2} \{n\mathbf{G}^{-1}\} \{\sigma_n^2 \mathbf{X}^T \mathbf{U} \mathbf{X} + \mathbf{S}\} \{n\mathbf{G}^{-1}\} + O_P(n^{-5/2}) \end{aligned}$$

The following theorem has four parts, where part (i) and (ii) are based on **Criterion 1** and the other two on **Criterion 2**.

Theorem 4.5.2 Assume the conditions of Theorem 4.5.1 and the model conditions (4.5.3)-(4.5.4). Suppose two R -class resampling schemes are used to estimate the variance, one of which is a UBS with i.i.d. weights having mean one, skewness zero and finite fourth moment.

(i) Suppose the other resampling scheme has $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ as the second order term in the variance, and $\mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} > 0$. Then if

$$-\frac{\lambda_{\min}(2\mathbf{D} + \mathbf{A})}{\lambda_{\min \mathbf{U}}} \leq \sigma_n^2 \leq \frac{\lambda_{\min \mathbf{A}}}{\lambda_{\max \mathbf{U}}}$$

the UBS technique is second order better than that resampling scheme, in the sense of **Criterion 1**.

(ii) Suppose the other resampling scheme has $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ as the second order term in the variance, and $\mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} < 0$. Then if

$$\frac{\lambda_{\max \mathbf{A}}}{\lambda_{\min \mathbf{U}}} \leq \sigma_n^2 \leq -\frac{\lambda_{\max}(2\mathbf{D} + \mathbf{A})}{\lambda_{\max \mathbf{U}}}$$

the UBS technique is second order better than that resampling scheme, in the sense of **Criterion 1**.

(iii) Suppose the other resampling scheme has $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ as the second order term in the variance, and $\text{tr}(\mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X}) > 0$. Then if

$$-\frac{\text{tr}(\mathbf{X}^T (2\mathbf{D} + \mathbf{A}) \mathbf{X})}{\text{tr}(\mathbf{X}^T \mathbf{U} \mathbf{X})} \leq \sigma_n^2 \leq \frac{\text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})}{\text{tr}(\mathbf{X}^T \mathbf{U} \mathbf{X})}$$

the UBS technique is second order better than that resampling scheme, in the sense of **Criterion 2**.

(iv) Suppose the other resampling scheme has $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ as the second order term in the variance, and $\text{tr}(\mathbf{X}^T (\mathbf{D} + \mathbf{A}) \mathbf{X}) < 0$. Then if

$$\frac{\text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{X})}{\text{tr}(\mathbf{X}^T \mathbf{U} \mathbf{X})} \leq \sigma_n^2 \leq -\frac{\text{tr}(\mathbf{X}^T (2\mathbf{D} + \mathbf{A}) \mathbf{X})}{\text{tr}(\mathbf{X}^T \mathbf{U} \mathbf{X})}$$

the UBS technique is second order better than that resampling scheme, in the sense of **Criterion 2**.

The above theorem provides four intervals as range of σ_n^2 for UBS to be second order better than the other resampling scheme. In the course of the proof it follows that $\lambda_{\min \mathbf{U}} > 0$, so that the denominator in the above theorem is always positive. Also, the model assumptions makes the intervals of parts (iii) and (iv) of the theorem non-empty, and have non-empty intersection with the positive half of the real line. Additional assumptions are required to guarantee the same for the intervals when using **Criterion 1**.

When $p = 1$, in Section 4.3 we showed that the UBS is second order optimal in the sense of being second order better than any other resampling scheme. Theorem 4.5.2 shows that while

using **Criterion 2** the same result holds, under the additional but mild assumption that for the resampling that is compared with *UBS*, the trace of the matrix in the second order term is non-zero. If one uses the much stronger **Criterion 1**, this optimality holds under extra model conditions which are fairly relaxed.

4.6 Results from simulation experiments

As an illustration we consider the design matrix from the oxygen uptake data from Rawlings, Pantula, Dickey (1998, page 124). Since in our results the design matrix is fixed, we take the design matrix from this data as fixed, and vary the response variable. We fix a set of five parameter values, all between 1.79 and 2.31, and vary the noise process to generate different sets of data, then estimate the parameters using least squares. In this framework we study the distribution of the estimated variance of the contrast given by $(-0.25, -0.25, -0.25, -0.25, 1)$. Our extensive simulations with other linear combinations of parameters and different design matrices yield similar conclusions.

With noise variances fixed, the actual variance of the least squares estimate (say V) is known, and for every noise sequence and every resampling plan this is estimated. Say for a particular resampling plan the variance estimator is called V_B . Our theorem states a relation like

$$n^{3/2}(V_B - V) = V_f + R_B$$

with V_f being the lead term in the expansion. We want to check whether the residual terms R_B differ significantly across resampling techniques, and whether for reasonable *UBS* techniques the residuals are indeed small. We present the density plots of $n^{3/2}(V_B - V)$ for three different kinds of independent noise random variables with data size 15 and 31.

Figures 4.1, 4.2 and 4.3 uses data with size 15, by considering the first fifteen design points from the oxygen uptake data. Figures 4.4, 4.5 and 4.6 consider the full design matrix based on 31 observations.

The noises are generated as follows. Let $\tau_i = 2 \log(i + 2)$. For Figures 4.1 and 4.4, the i^{th} noise was generated from $Normal(0, \tau_i^2)$ distribution. For Figures 4.2 and 4.5, the i^{th} noise was generated from *Exponential* distribution with mean τ_i , which was left shifted so that the mean became zero. For Figures 4.3 and 4.6, the i^{th} noise was generated from *Gamma* distribution with parameters 10 and τ_i , which was also left shifted so that the mean became zero.

Although our theorems on optimality of *UBS* depend on choice of σ_n^2 (the variance of the weights), in order to check the performance of *i.i.d. UBS* schemes generally we ignore the variance conditions of the theorem, and simply work with weights from distributions with mean one, skewness zero and finite moments. The two *i.i.d. UBS* schemes chosen for simulations are

as follows. The first one is a two-point distribution, putting equal masses on 0.15 and 1.85, this we call *discrete UBS*. This is a slightly modified version of the optimal *UBS* for the variance of the sample mean in the one-dimensional case. The other *UBS* scheme uses weights from a mean 1, variance $1/3$ *Uniform* distribution, and is called *Uniform UBS*.

The estimated densities are denoted as follows:

1. Wu's jackknife - dashed line (Figure A, 4.1- 4.6).
2. Hinkley's jackknife - dashed line (Figure B, 4.1- 4.6).
3. Standard delete-1 jackknife - longer dashed line (Figure B, 4.1- 4.6).
4. *Uniform UBS* - dashed line (Figure C, D, 4.1- 4.6).
5. Paired bootstrap - longer dashed line (Figure C, 4.1- 4.6).
6. External bootstrap - longer dashed line (Figure D, 4.1- 4.6).
7. *Discrete UBS* - dotted line (all figures).
8. Simulated data density - solid line (all figures).

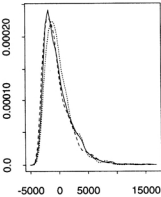
As the figures show, the *Discrete UBS*, *Uniform UBS* and *Wu's jackknife* estimate the densities remarkably well, even in small samples. *Hinkley's jackknife* also does reasonably well, although there is some underestimation of the right tail. The *paired bootstrap* performs very badly with small samples as expected, but does reasonably well even with moderate sized samples. The usual *delete-1 jackknife* and *external bootstrap* perform poorly. The external bootstrap also substantially underestimates the right tail, even for $n = 31$. We have checked with data sets of size 80 and found that the bias in delete-1 jackknife and external bootstrap persists.

The figures show that in small or moderate sample sizes the *i.i.d. UBS* is distinctly better than all resampling schemes, with Wu's jackknife being the only scheme that is comparably good. It may be mentioned in the context of the weighted jackknives that the use of the diagonal elements of P_x is very special to the least squares estimation and linear regression.

On the other hand, the *UBS* schemes can be easily set up for other types of regression problems, see Chapter 7. Given that its performance is comparable to Wu's jackknife in linear regressions, it would be worthwhile to study it in other regression problems. The performance of the weighted jackknives may improve further if instead of delete-1 jackknife, a delete- d jackknife is used, but then computational difficulties and data storage problems eliminate any advantage. In any case, the theoretical edge and the computational advantages that *UBS* has over other schemes still remains.

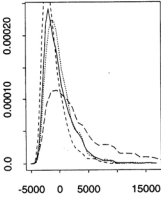
Figure 4.1: Normal distribution, data size=15

Figure 1A



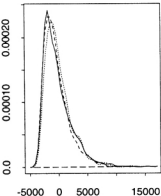
Simulated, Discrete, Wu

Figure 1B



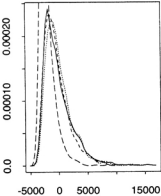
Simulated, Discrete, Hinkley, Jackknife

Figure 1C



Simulated, Discrete, Uniform, Paired

Figure 1D



Simulated, Discrete, Uniform, External

Figure 4.2: Exponential distribution, data size=15

Figure 2A

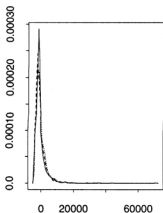
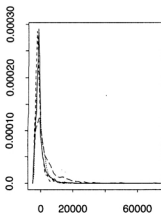


Figure 2B



Simulated, Discrete, Wu

Simulated, Discrete, Hinkley, Jackknife

Figure 2C

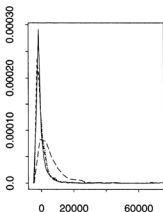
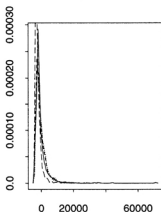


Figure 2D

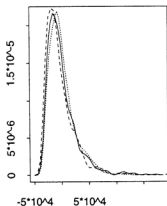


Simulated, Discrete, Uniform, Paired

Simulated, Discrete, Uniform, External

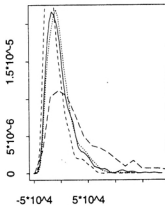
Figure 4.3: Gamma distribution, data size=15

Figure 3A



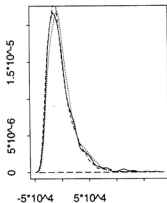
Simulated, Discrete, Wu

Figure 3B



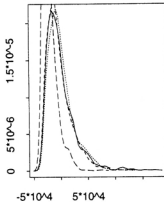
Simulated, Discrete, Hinkley, Jackknife

Figure 3C



Simulated, Discrete, Uniform, Paired

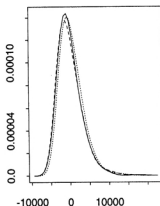
Figure 3D



Simulated, Discrete, Uniform, External

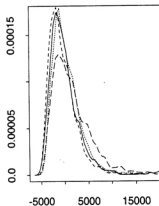
Figure 4.4: Normal distribution, data size=31

Figure 4A



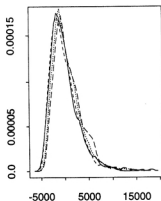
Simulated, Discrete, Wu

Figure 4B



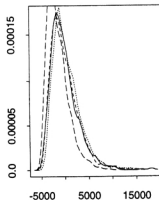
Simulated, Discrete, Hinkley, Jackknife

Figure 4C



Simulated, Discrete, Uniform, Paired

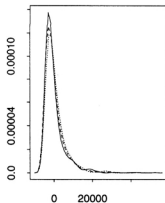
Figure 4D



Simulated, Discrete, Uniform, External

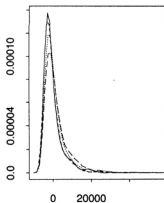
Figure 4.5: Exponential distribution, data size=31

Figure 5A



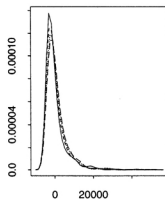
Simulated, Discrete, Wu

Figure 5B



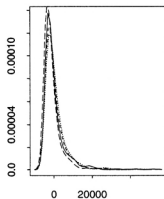
Simulated, Discrete, Hinkley, Jackknife

Figure 5C



Simulated, Discrete, Uniform, Paired

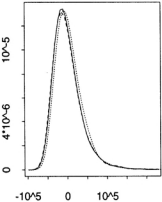
Figure 5D



Simulated, Discrete, Uniform, External

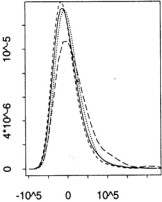
Figure 4.6: Gamma distribution, data size=31

Figure 6A



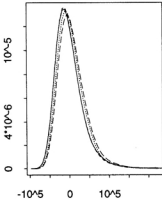
Simulated, Discrete, Wu

Figure 6B



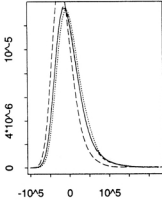
Simulated, Discrete, Hinkley, Jackknife

Figure 6C



Simulated, Discrete, Uniform, Paired

Figure 6D



Simulated, Discrete, Uniform, External

4.7 Proofs of theorems: $p = 1$ case

Lemma 4.7.1 *Under the assumption of homoscedasticity of errors, the following relations hold:*

$$(a) \quad \frac{n}{L_n} n^{-1} \sum_{i=1}^n e_i^2 \xrightarrow{P} \alpha^2 \tau^2$$

$$(b) \quad \frac{n}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 \xrightarrow{P} \alpha^2 \tau^2$$

Proof of Lemma 4.7.1 (a) We have to prove $\frac{1}{L_n} \sum_{i=1}^n e_i^2 \xrightarrow{P} \alpha^2 \tau^2$ when $e_1, \dots, e_n$ are independent with mean zero, variance τ^2 and finite sixth moment. Fix $\epsilon > 0$.

$$\begin{aligned} P[|n^{-1} \sum_{i=1}^n e_i^2 - \tau^2| > \epsilon] &\leq \epsilon^{-2} n^{-2} E\left(\sum_{i=1}^n (e_i^2 - \tau^2)\right)^2 \\ &= \epsilon^{-2} n^{-2} E\sum_{i=1}^n (e_i^2 - \tau^2)^2 \\ &\rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

(b) Here we work under same conditions as in (a). Again fix $\epsilon > 0$.

$$\begin{aligned} P\left[\left|\frac{1}{L_n} \sum_{i=1}^n x_i^2 e_i^2 - \tau^2\right| > \epsilon\right] &= P\left[\left|\frac{1}{L_n} \sum_{i=1}^n x_i^2 (e_i^2 - \tau^2)\right| > \epsilon\right] \\ &\leq \epsilon^{-2} \frac{1}{L_n^2} E\left[\sum_{i=1}^n x_i^2 (e_i^2 - \tau^2)\right]^2 \\ &\leq \epsilon^{-2} \frac{1}{L_n^2} E\sum_{i=1}^n x_i^4 (e_i^2 - \tau^2)^2 \\ &\rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

■

Lemma 4.7.2 *Under the assumption of heteroscedasticity of errors, the following relations hold:*

$$(a) \quad \frac{n}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 \xrightarrow{P} \alpha^4 \alpha_0$$

$$(b) \quad \frac{n^2}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 \xrightarrow{P} \alpha^6 \alpha_1$$

Proof of Lemma 4.7.2 (a) We have $e_1, \dots, e_n$ as independent random variables with mean zero, finite sixth moment and $V(e_i) = \tau_i^2$. Fix $\epsilon > 0$.

$$\begin{aligned}
 P[|n^{-1} \sum_{i=1}^n x_i^4 e_i^2 - \alpha_1| > 2\epsilon] &= P[|n^{-1} \sum_{i=1}^n x_i^4 (e_i^2 - \tau_i^2)| > \epsilon], \text{ for large } n \\
 &\leq \epsilon^{-2} n^{-2} E\left[\sum_{i=1}^n x_i^4 (e_i^2 - \tau_i^2)\right]^2 \\
 &= \epsilon^{-2} n^{-2} E \sum_{i=1}^n x_i^8 (e_i^2 - \tau_i^2)^2 \\
 &\rightarrow 0 \text{ as } n \rightarrow \infty
 \end{aligned}$$

(b) Here we have the same set-up as part (a). Again fix $\epsilon > 0$.

$$\begin{aligned}
 P[|n^{-1} \sum_{i=1}^n x_i^2 e_i^2 - \alpha_0| > 2\epsilon] &= P[|n^{-1} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2)| > \epsilon], \text{ for large } n \\
 &\leq \epsilon^{-2} n^{-2} E\left[\sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2)\right]^2 \\
 &= \epsilon^{-2} n^{-2} E \sum_{i=1}^n x_i^4 (e_i^2 - \tau_i^2)^2 \\
 &\rightarrow 0 \text{ as } n \rightarrow \infty
 \end{aligned}$$

Proof of Theorem 4.3.1(i) The residuals from the regression problem are denoted by

$$r_i = e_i - (\hat{\beta} - \beta)x_i, 1 \leq i \leq n$$

$$\begin{aligned}
 V_B &= \frac{1}{L_n} \left\{ \frac{1}{n} \sum_{i=1}^n r_i^2 - \bar{r}^2 \right\} \\
 &= \frac{1}{L_n} \left\{ \frac{1}{n} \sum_{i=1}^n (e_i^2 + (\hat{\beta} - \beta)^2 x_i^2 - 2(\hat{\beta} - \beta)x_i e_i) \right\} - \frac{1}{L_n} \bar{r}^2 \\
 &= \frac{1}{nL_n} \sum_{i=1}^n e_i^2 - n^{-1}(\hat{\beta} - \beta)^2 - \frac{1}{n^2 L_n} \left(\sum_{i=1}^n r_i \right)^2 \\
 &= \frac{1}{nL_n} \sum_{i=1}^n e_i^2 - n^{-1}(\hat{\beta} - \beta)^2 - \frac{1}{n^2 L_n} \left[\left(\sum_{i=1}^n e_i \right)^2 + \frac{1}{L_n} \left(\sum_{i=1}^n x_i e_i \right)^2 \left(\sum_{i=1}^n x_i \right)^2 \right. \\
 &\quad \left. - 2 \frac{1}{L_n} \left(\sum_{i=1}^n x_i e_i \right) \left(\sum_{i=1}^n e_i \right) \left(\sum_{i=1}^n x_i \right) \right]
 \end{aligned}$$

Thus we have

$$n^{3/2}(V_B - v_n) = \frac{1}{nL_n} n^{3/2} \sum_{i=1}^n (e_i - \tau_i^2) - n^{-1} n^{3/2} (\hat{\beta} - \beta)^2$$

$$\begin{aligned}
& -\frac{1}{n^2 L_n} n^{3/2} [(\sum_{i=1}^n e_i)^2 + \frac{1}{L_n^2} (\sum_{i=1}^n x_i e_i)^2 (\sum_{i=1}^n x_i)^2 \\
& - 2 \frac{1}{L_n} (\sum_{i=1}^n x_i e_i) (\sum_{i=1}^n e_i) (\sum_{i=1}^n x_i)] \\
& = Y_n - \frac{1}{n^{1/2}} (X_{1n})
\end{aligned}$$

where

$$\begin{aligned}
Y_n &= \frac{1}{L_n} n^{1/2} \sum_{i=1}^n (e_i - \tau_i^2) \\
X_{1n} &= \{n^{1/2} \frac{1}{L_n} (\sum_{i=1}^n x_i e_i)\}^2 + n \frac{1}{L_n} [(\sum_{i=1}^n n^{-1/2} e_i)^2 \\
& - 2n^{1/2} (\hat{\beta} - \beta) (\sum_{i=1}^n n^{-1/2} e_i) (\sum_{i=1}^n n^{-1} x_i) + (n^{1/2} (\hat{\beta} - \beta))^2 (\sum_{i=1}^n n^{-1} x_i)^2]
\end{aligned}$$

are all bounded in probability. This proves Theorem 4.3.1(i). ■

Proof of Theorem 4.3.1(ii)

$$V_{JLS} = \frac{L_n}{n^2(n-1)} \sum_{i=1}^n \frac{(J_i - \hat{\beta})^2}{x_i^2}$$

where from the definition of pseudo-value we have $J_i = n\hat{\beta} - (n-1)\hat{\beta}_{(-i)}$ where $\hat{\beta}_{(-i)}$ is the estimate for β with the i_{th} observation deleted. Thus

$$\begin{aligned}
J_i &= n\hat{\beta} - (n-1)\hat{\beta}_{(-i)} \\
&= \beta + n \frac{1}{L_n} \sum_{j=1}^n x_j e_j - (n-1) \frac{\sum_{j=1}^n x_j e_j - x_i e_i}{L_n - x_i^2} \\
&= \beta + n \frac{1}{L_n} \sum_{j=1}^n x_j e_j - n \frac{1}{L_n} \sum_{j=1}^n x_j e_j (1 + \frac{x_i^2}{L_n - x_i^2}) + n \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2}) \\
&\quad + \frac{1}{L_n} \sum_{j=1}^n x_j e_j (1 + \frac{x_i^2}{L_n - x_i^2}) - \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2}) \\
&= \beta - n \frac{1}{L_n} \sum_{j=1}^n x_j e_j \frac{x_i^2}{L_n - x_i^2} + n \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2}) \\
&\quad + \frac{1}{L_n} \sum_{j=1}^n x_j e_j (1 + \frac{x_i^2}{L_n - x_i^2}) - \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2}) \\
&= \hat{\beta} - n \frac{1}{L_n} \sum_{j=1}^n x_j e_j \frac{x_i^2}{L_n - x_i^2} + n \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2}) \\
&\quad + \frac{1}{L_n} \sum_{j=1}^n x_j e_j \frac{x_i^2}{L_n - x_i^2} - \frac{1}{L_n} x_i e_i (1 + \frac{x_i^2}{L_n - x_i^2})
\end{aligned}$$

Let $\frac{1}{L_n - x_i^2} = \frac{1}{L_n}(1 + \epsilon_{ni})$ so that $\epsilon_{ni} = \frac{x_i^2}{L_n - x_i^2}$ is a $O(n^{-1})$ term. Thus we have

$$\begin{aligned} J_i - \hat{\beta} &= n \frac{1}{L_n} x_i e_i - \frac{1}{L_n} x_i e_i - n \frac{1}{L_n^2} x_i^2 (1 + \epsilon_{ni}) \sum_{j=1}^n x_j e_j \\ &\quad - \frac{1}{L_n} x_i e_i \epsilon_{ni} + n \frac{1}{L_n^2} x_i^3 e_i (1 + \epsilon_{ni}) + \frac{1}{L_n^2} x_i^2 (1 + \epsilon_{ni}) \sum_{j=1}^n x_j e_j \end{aligned} \quad (4.7.1)$$

and hence

$$\begin{aligned} &\frac{J_i - \hat{\beta}}{x_i} \\ &= \frac{1}{L_n} n e_i + \frac{1}{L_n} [-e_i + n \frac{1}{L_n} x_i^2 e_i (1 + \epsilon_{ni}) - n \frac{1}{L_n} x_i (1 + \epsilon_{ni}) \sum_{j=1}^n x_j e_j - e_i \epsilon_{ni}] \\ &\quad + \frac{1}{L_n^2} x_i (1 + \epsilon_{ni}) \sum_{j=1}^n x_j e_j \\ &= \frac{1}{L_n} (n-1) e_i + n \frac{1}{L_n^2} (1 + \epsilon_{ni}) x_i^2 e_i - n \frac{1}{L_n} (1 + \epsilon_{ni}) x_i (\hat{\beta} - \beta) \\ &\quad - \frac{1}{L_n} \epsilon_{ni} e_i + \frac{1}{L_n} (1 + \epsilon_{ni}) x_i (\hat{\beta} - \beta) \end{aligned}$$

So we have

$$\begin{aligned} &\sum_{i=1}^n \frac{(J_i - \hat{\beta})^2}{x_i^2} \\ &= (n-1)^2 \frac{1}{L_n^2} \sum_{i=1}^n e_i^2 + n^2 \frac{1}{L_n^4} \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^4 e_i^2 + n^2 \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^2 \\ &\quad + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni}^2 e_i^2 + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^2 + 2 \frac{1}{L_n^3} (n-1) n \sum_{i=1}^n (1 + \epsilon_{ni}) x_i^2 e_i^2 \\ &\quad - 2 \frac{1}{L_n^2} (n-1) n (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni}) x_i e_i - 2 \frac{1}{L_n^2} (n-1) \sum_{i=1}^n \epsilon_{ni} e_i^2 \\ &\quad + 2 \frac{1}{L_n^2} (n-1) (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni}) x_i e_i - 2 \frac{1}{L_n^3} n^2 (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^3 e_i \\ &\quad - 2 \frac{1}{L_n^3} n \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i^2 e_i^2 + 2 \frac{1}{L_n^3} n (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^3 e_i \\ &\quad + 2 \frac{1}{L_n^2} n (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i e_i - 2 \frac{1}{L_n^2} n (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^2 \\ &\quad - 2 \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i e_i \\ &= (n-1)^2 \frac{1}{L_n^2} \sum_{i=1}^n e_i^2 + n^2 \frac{1}{L_n^4} \sum_{i=1}^n x_i^4 e_i^2 + n^2 \frac{1}{L_n} (\hat{\beta} - \beta)^2 + \frac{1}{L_n} (\hat{\beta} - \beta)^2 \end{aligned}$$

$$\begin{aligned}
& + 2 \frac{1}{L_n^3} (n-1)n \sum_{i=1}^n x_i^2 e_i^2 - 2 \frac{1}{L_n} (n-1)n(\hat{\beta} - \beta)^2 \\
& + 2 \frac{1}{L_n} (n-1)(\hat{\beta} - \beta)^2 - 2 \frac{1}{L_n^3} n^2 (\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i \\
& + 2 \frac{1}{L_n^3} n(\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i - 2 \frac{1}{L_n} n(\hat{\beta} - \beta)^2 + n^2 \frac{1}{L_n^4} \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^4 e_i^2 \\
& + n^2 \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^2 + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni}^2 e_i^2 + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^2 \\
& + 2 \frac{1}{L_n^3} (n-1)n \sum_{i=1}^n \epsilon_{ni} x_i^2 e_i^2 - 2 \frac{1}{L_n^2} (n-1)n(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} x_i e_i \\
& - 2 \frac{1}{L_n^2} (n-1) \sum_{i=1}^n \epsilon_{ni} e_i^2 + 2 \frac{1}{L_n^2} (n-1)(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} x_i e_i \\
& - 2 \frac{1}{L_n^3} n^2 (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^3 e_i - 2 \frac{1}{L_n^3} n \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i^2 e_i^2 \\
& + 2 \frac{1}{L_n^3} n(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^3 e_i + 2 \frac{1}{L_n^2} n(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i e_i \\
& - 2 \frac{1}{L_n^2} n(\hat{\beta} - \beta)^2 \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^2 - 2 \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i e_i \\
& = (n-1)^2 \frac{1}{L_n^2} \sum_{i=1}^n e_i^2 + n^2 \frac{1}{L_n} (\hat{\beta} - \beta)^2 + 2 \frac{1}{L_n^3} (n-1)n \sum_{i=1}^n x_i^2 e_i^2 \\
& \quad - 2 \frac{1}{L_n} (n-1)n(\hat{\beta} - \beta)^2 + O_P(n^{-1/2})
\end{aligned}$$

And hence we have

$$\begin{aligned}
V_{JLS} &= \frac{L_n}{n^2(n-1)} \sum_{i=1}^n \frac{(J_i - \hat{\beta})^2}{x_i^2} \\
&= \frac{1}{n^2 L_n} (n-1) \sum_{i=1}^n e_i^2 + \frac{1}{n-1} (\hat{\beta} - \beta)^2 + 2 \frac{1}{L_n^2} \frac{1}{n} \sum_{i=1}^n x_i^2 e_i^2 \\
&\quad - 2 \frac{1}{n} (\hat{\beta} - \beta)^2 + O_P(n^{-5/2})
\end{aligned}$$

So we have

$$\begin{aligned}
& n^{3/2} (V_{JLS} - v_n) \\
&= n^{3/2} \frac{1}{n L_n} \sum_{i=1}^n (e_i^2 - \tau_i^2) - n^{3/2} \frac{1}{n^2 L_n} \sum_{i=1}^n e_i^2 + \frac{1}{(n-1)} n^{1/2} \{n^{1/2} (\hat{\beta} - \beta)\}^2 \\
&\quad + 2n^{3/2} \frac{1}{L_n^2} \frac{1}{n} \sum_{i=1}^n x_i^2 e_i^2 - 2n^{1/2} \frac{1}{n} \{n^{1/2} (\hat{\beta} - \beta)\}^2 + O_P(n^{-1}) \\
&= Y_n + n^{-1/2} X_{3n} + O_P(n^{-1})
\end{aligned}$$

where

$$Y_n = n^{1/2} \frac{1}{L_n} \sum_{i=1}^n (e_i^2 - \tau_i^2)$$

$$X_{3n} = \{n^{1/2}(\hat{\beta} - \beta)\}^2 \left(-1 + \frac{1}{n-1}\right) + 2n \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 - \frac{1}{L_n} \sum_{i=1}^n e_i^2$$

Note that X_{3n} is bounded in probability. This proves Theorem 4.3.1(ii). ■

Proof of Theorem 4.3.2(i) We have

$$\begin{aligned} & \sum_{i=1}^n (J_i - \hat{\beta})^2 \\ = & (n-1)^2 \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + n^2 \frac{1}{L_n^4} \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^6 e_i^2 + n^2 \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^4 \\ & + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni}^2 x_i^2 e_i^2 + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^4 + 2(n-1)n \frac{1}{L_n^3} \sum_{i=1}^n (1 + \epsilon_{ni}) x_i^4 e_i^2 \\ & - 2(n-1)n \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni}) x_i^3 e_i - 2(n-1) \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni} x_i^2 e_i^2 \\ & + 2(n-1) \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni}) x_i^3 e_i - 2n^2 \frac{1}{L_n^3} (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^5 e_i \\ & - 2n \frac{1}{L_n^3} \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i^4 e_i^2 + 2n \frac{1}{L_n^3} (\hat{\beta} - \beta) \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^5 e_i \\ & + 2n \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i^3 e_i - 2n \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n (1 + \epsilon_{ni})^2 x_i^4 \\ & - 2 \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} (1 + \epsilon_{ni}) x_i^3 e_i \\ = & (n-1)^2 \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + n^2 \frac{1}{L_n^4} \sum_{i=1}^n x_i^6 e_i^2 + n^2 \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 \\ & + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 + 2(n-1)n \frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 - 2(n-1)n \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i \\ & + 2(n-1) \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i - 2n^2 \frac{1}{L_n^3} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^5 e_i \\ & + 2n \frac{1}{L_n^3} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^5 e_i - 2n \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 \\ & + n^2 \frac{1}{L_n^4} \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^6 e_i^2 + n^2 \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^4 \\ & + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni}^2 x_i^2 e_i^2 + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n \epsilon_{ni} (2 + \epsilon_{ni}) x_i^4 + 2(n-1)n \frac{1}{L_n^3} \sum_{i=1}^n \epsilon_{ni} x_i^4 e_i^2 \end{aligned}$$

$$\begin{aligned}
& -2(n-1)n\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} x_i^3 e_i - 2(n-1)\frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni} x_i^2 e_i^2 \\
& + 2(n-1)\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni} x_i^3 e_i - 2n^2\frac{1}{L_n^3}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni}(2 + \epsilon_{ni})x_i^5 e_i \\
& - 2n\frac{1}{L_n^3} \sum_{i=1}^n \epsilon_{ni}(1 + \epsilon_{ni})x_i^4 e_i^2 + 2n\frac{1}{L_n^3}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni}(2 + \epsilon_{ni})x_i^5 e_i \\
& + 2n\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni}(1 + \epsilon_{ni})x_i^3 e_i - 2n\frac{1}{L_n^2}(\hat{\beta} - \beta)^2 \sum_{i=1}^n \epsilon_{ni}(2 + \epsilon_{ni})x_i^4 \\
& - 2\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n \epsilon_{ni}(1 + \epsilon_{ni})x_i^3 e_i \\
= & (n-1)^2\frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + n^2\frac{1}{L_n^2}(\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 + 2(n-1)n\frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 \\
& - 2(n-1)n\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i + O_P(n^{-1/2})
\end{aligned}$$

Therefore

$$\begin{aligned}
& V_J \\
= & \frac{1}{n(n-1)} \sum_{i=1}^n (J_i - \hat{\beta})^2 \\
= & \frac{n-1}{n} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + \frac{n}{n-1} \frac{1}{L_n^2}(\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 + 2\frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 \\
& - 2\frac{1}{L_n^2}(\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i + O_P(n^{-5/2})
\end{aligned}$$

And hence

$$\begin{aligned}
n^{3/2}(V_J - V_n) &= n^{3/2} \left[\frac{(n-1)}{n} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 - \frac{1}{L_n^2} \sum_{i=1}^n \tau_i^2 x_i^2 \right] \\
&+ n^{3/2} \frac{n}{n(n-1)} \frac{1}{L_n^2} \{n^{1/2}(\hat{\beta} - \beta)\}^2 \sum_{i=1}^n x_i^4 + 2n^{3/2} \frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 \\
&- 2n^{3/2} \frac{1}{L_n^2} \{n^{1/2}(\hat{\beta} - \beta)\} \sum_{i=1}^n n^{-1/2} x_i^3 e_i + O_P(n^{-1}) \\
= & Z_n + n^{-1/2} T_{1n} + O_P(n^{-1})
\end{aligned}$$

where

$$\begin{aligned}
Z_n &= n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2) \\
T_{1n} &= \frac{n^3}{n-1} \frac{1}{L_n^2} \{n^{1/2}(\hat{\beta} - \beta)\}^2 \sum_{i=1}^n n^{-1} x_i^4 - 2n^2 \frac{1}{L_n^2} \{n^{1/2}(\hat{\beta} - \beta)\} \sum_{i=1}^n n^{-1/2} x_i^3 e_i
\end{aligned}$$

$$+ 2n^2 \frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 - n \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2$$

Note that Z_n and T_{1n} are bounded in probability. This proves Theorem 4.3.2(i). ■

Proof of Theorem 4.3.2(ii)

$$\begin{aligned} V_{\text{EXB}} &= V_{\text{WB}} = \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 r_i^2 \\ &= \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (y_i - \hat{\beta} x_i)^2 \\ &= \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (x_i(\beta - \hat{\beta}) + e_i)^2 \\ &= \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (x_i^2 (\hat{\beta} - \beta)^2 + e_i^2 - 2x_i e_i (\hat{\beta} - \beta)) \\ &= \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + (\hat{\beta} - \beta)^2 \frac{1}{L_n^2} \sum_{i=1}^n x_i^4 - 2(\hat{\beta} - \beta) \frac{1}{L_n^2} \sum_{i=1}^n x_i^3 e_i \end{aligned}$$

Thus we have

$$n^{3/2} (V_{\text{EXB}} - V_n) = Z_n + n^{-1/2} T_{2n} \quad (4.7.2)$$

where

$$\begin{aligned} Z_n &= n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2) \\ T_{2n} &= n \frac{1}{L_n^2} \{n^{1/2} (\hat{\beta} - \beta)\}^2 \sum_{i=1}^n x_i^4 - 2n^2 \frac{1}{L_n^2} \{n^{1/2} (\hat{\beta} - \beta)\} \sum_{i=1}^n n^{-1/2} x_i^3 e_i \end{aligned}$$

Note that Z_n and T_{2n} are bounded in probability. This proves Theorem 4.3.2(ii). ■

Proof of Theorem 4.3.2(iii)

$$V_{\text{JH}} = \frac{n}{(n-1)^3} \sum_{i=1}^n \left(\frac{L_n - x_i^2}{L_n} \right)^2 (J_i - \hat{\beta})^2$$

We have

$$\begin{aligned} J_i - \hat{\beta} &= \frac{n-1}{L_n - x_i^2} [x_i e_i - x_i \frac{1}{L_n} \sum_{j=1}^n x_j e_j] \\ &= \frac{n-1}{L_n - x_i^2} [x_i e_i - x_i (\hat{\beta} - \beta)] \end{aligned}$$

so

$$\left(\frac{L_n - x_i^2}{L_n} \right)^2 (J_i - \hat{\beta})^2 = (n-1)^2 \frac{1}{L_n^2} [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i]$$

and hence

$$n^{3/2}(V_{JH} - V_n) = Z_n + n^{-1/2}T_{3n}$$

where

$$\begin{aligned} Z_n &= n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2) \\ T_{3n} &= \frac{n^2}{n-1} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 + \frac{n^3}{n-1} \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 - 2 \frac{n^3}{n-1} \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i \end{aligned}$$

Note that Z_n and T_{3n} are bounded in probability. This proves Theorem 4.3.2(iii). ■

Proof of Theorem 4.3.2(iv)

$$\begin{aligned} &V_{JW} \\ &= \frac{1}{L_n} \sum_{i=1}^n \frac{1}{L_n - x_i^2} [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i] \\ &= \frac{1}{L_n^2} \sum_{i=1}^n [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i] \\ &\quad + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni} [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i] \\ &= \frac{1}{L_n^2} \sum_{i=1}^n [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i] + \frac{1}{L_n^2} \sum_{i=1}^n \epsilon_{ni} x_i^2 e_i^2 + O_P(n^{-5/2}) \\ &= \frac{1}{L_n^2} \sum_{i=1}^n [x_i^2 e_i^2 + x_i^4 (\hat{\beta} - \beta)^2 - 2(\hat{\beta} - \beta) x_i^3 e_i] + \frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 + O_P(n^{-5/2}) \end{aligned}$$

so that

$$n^{3/2}(V_{JW} - V_n) = Z_n + n^{-1/2}T_{4n} + O_P(n^{-1})$$

where

$$\begin{aligned} Z_n &= n^{3/2} \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 (e_i^2 - \tau_i^2) \\ n^{-2}T_{4n} &= \frac{1}{L_n^3} \sum_{i=1}^n x_i^4 e_i^2 + \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 \sum_{i=1}^n x_i^4 - 2 \frac{1}{L_n^2} (\hat{\beta} - \beta) \sum_{i=1}^n x_i^3 e_i \end{aligned}$$

Note that Z_n and $n^{-2}T_{4n}$ are bounded in probability. This proves Theorem 4.3.2(iv). ■

We now present the proof of the result for resampling with *UBS*. We use the notation W_a for the standardised weights $\frac{(w_a - 1)}{\sigma_a}$.

Lemma 4.7.3 *The following hold for resampling with UBS in linear models*

$$E_B(\sum_{i=1}^n W_i x_i^2)^4 = O(n^2) \quad (4.7.3)$$

$$E_B(\sum_{i=1}^n (w_i - 1)x_i^2)^4 = O(n^2) \quad (4.7.4)$$

$$E_B(\sum_{i=1}^n (w_i - 1)x_i e_i)^4 = O_P(n^2) \quad (4.7.5)$$

$$E_B(\sum_{i=1}^n W_i x_i e_i)^4 = O_P(n^2) \quad (4.7.6)$$

Proof of Lemma 4.7.3 Note that it is enough to prove (4.7.3) and (4.7.6). Because of (3.2.20), the other two relations follow.

We have

$$\begin{aligned} E_B(\sum_{i=1}^n W_i x_i^2)^4 &= E_B[\sum_{i=1}^n W_i^4 x_i^8 + 3 \sum_{i \neq j} W_j^2 W_i^2 x_i^4 x_j^4 + 4 \sum_{i \neq j} W_j^3 W_i x_i^6 x_j^2 \\ &\quad + 6 \sum_{i \neq j \neq k} W_i^2 W_j W_k x_i^4 x_j^2 x_k^2 + \sum_{i \neq j \neq k \neq l} W_i W_j W_k W_l x_i^2 x_j^2 x_k^2 x_l^2] \\ &= O(n^2) \end{aligned}$$

For (4.7.6), we have

$$\begin{aligned} E_B(\sum_{i=1}^n W_i x_i e_i)^4 &= E_B[\sum_{i=1}^n W_i^4 x_i^4 e_i^4 + 3 \sum_{i \neq j} W_i^2 W_j^2 x_i^2 x_j^2 e_i^2 e_j^2 + 4 \sum_{i \neq j} W_i^3 W_j x_i^3 x_j e_i^3 e_j \\ &\quad + 6 \sum_{i \neq j \neq k} W_i^2 W_j W_k x_i^2 x_j x_k e_i^2 e_j e_k \\ &\quad + \sum_{i \neq j \neq k \neq l} W_i W_j W_k W_l x_i x_j x_k x_l e_i e_j e_k e_l] \\ &= O_P(n^2) \end{aligned}$$

We also have the following result.

Lemma 4.7.4 *The following hold for resampling with UBS in linear models*

$$E_B(\sum_{i=1}^n W_i x_i^2)^6 = O(n^3) \quad (4.7.7)$$

$$E_B(\sum_{i=1}^n (w_i - 1)x_i^2)^6 = O(n^3) \quad (4.7.8)$$

$$E_B(\sum_{i=1}^n (w_i - 1)x_i e_i)^6 = O_P(n^3) \quad (4.7.9)$$

$$E_B(\sum_{i=1}^n W_i x_i e_i)^6 = O_P(n^3) \quad (4.7.10)$$

Proof of Lemma 4.7.4 Note that it is enough to prove (4.7.7) and (4.7.10). Because of (3.2.20), the other two relations follow. We state a series of relations, which in turn prove the required result. All these relations in this series can be verified by direct calculation. Use the notation $\sum_{abc\dots}$ to indicate the sum over the set of indices $a \neq b \neq c \neq \dots$. We also use e as an index, however there is no scope for confusion with the noise term.

$$\begin{aligned}
E_B \sum_{i=1}^n W_a^6 x_i^6 e_i^6 &= O_P(n) \\
E_B \sum_{i=1}^n W_a^6 x_i^{12} &= O_P(n) \\
E_B \sum_{ab} W_a^5 W_b x_a^5 e_a^5 x_b e_b &= O(n^2) \\
E_B \sum_{ab} W_a^5 W_b x_a^{10} x_b^2 &= O(n^2) \\
E_B \sum_{ab} W_a^4 W_b^2 x_a^8 x_b^4 &= O(n^2) \\
E_B \sum_{ab} W_a^4 W_b^2 x_a^4 e_a^4 x_b^2 e_b^2 &= O(n^2) \\
E_B \sum_{ab} W_a^3 W_b^3 x_a^3 e_a^3 x_b^3 e_b^3 &= O(n^2) \\
E_B \sum_{ab} W_a^3 W_b^3 x_a^6 x_b^6 &= O(n^2) \\
E_B \sum_{abc} W_a^4 W_b W_c x_a^4 e_a^4 x_b e_b x_c e_c &= O(n^2) \\
E_B \sum_{abc} W_a^4 W_b W_c x_a^8 x_b^2 x_c^2 &= O(n^2) \\
E_B \sum_{abc} W_a^3 W_b^2 W_c x_a^3 e_a^3 x_b^2 e_b^2 x_c e_c &= O(n^2) \\
E_B \sum_{abc} W_a^3 W_b^2 W_c x_a^6 x_b^4 x_c^2 &= O(n^2) \\
E_B \sum_{abc} W_a^2 W_b^2 W_c^2 x_a^2 e_a^2 x_b^2 e_b^2 x_c^2 e_c^2 &= O(n^3) \\
E_B \sum_{abc} W_a^2 W_b^2 W_c^2 x_a^4 x_b^4 x_c^4 &= O(n^3) \\
E_B \sum_{abcd} W_a^3 W_b W_c W_d x_a^3 e_a^3 x_b e_b x_c e_c x_d e_d &= O(n^2) \\
E_B \sum_{abcd} W_a^3 W_b W_c W_d x_a^6 x_b^2 x_c^2 x_d^2 &= O(n^2) \\
E_B \sum_{abcd} W_a^2 W_b^2 W_c W_d x_a^2 e_a^2 x_b^2 e_b^2 x_c e_c x_d e_d &= O(n^2) \\
E_B \sum_{abcd} W_a^2 W_b^2 W_c W_d x_a^4 x_b^4 x_c^2 x_d^2 &= O(n^2)
\end{aligned}$$

$$E_B \sum_{abcde} W_a^2 W_b W_c W_d W_e x_a^2 e_a^2 x_b e_b x_c e_c x_d e_d x_e e_e = O(n^2)$$

$$E_B \sum_{abcde} W_a^2 W_b W_c W_d W_e x_a^4 x_b^2 x_c^2 x_d^2 x_e^2 = O(n^2)$$

$$E_B \sum_{abcdef} W_a W_b W_c W_d W_e W_f x_a e_a x_b e_b x_c e_c x_d e_d x_e e_e x_f e_f = O(n^2)$$

$$E_B \sum_{abcdef} W_a W_b W_c W_d W_e W_f x_a^2 x_b^2 x_c^2 x_d^2 x_e^2 x_f^2 = O(n^2)$$

Proof of Theorem 4.3.2(v) We have

$$\begin{aligned} \hat{\beta}_B - \beta &= \frac{\sum_{i=1,n} w_i y_i x_i}{\sum_{i=1,n} w_i x_i^2} - \frac{\sum_{i=1,n} y_i x_i}{\sum_{i=1,n} x_i^2} \\ &= \frac{1}{L_n} \left[\sum_{i=1}^n (w_i - 1) x_i e_i \right] + \sum_{i=1}^n w_i x_i e_i \left[\frac{1}{\sum_{i=1}^n w_i x_i^2} - \frac{1}{\sum_{i=1}^n x_i^2} \right] \\ &= C_n + D_n \end{aligned}$$

where

$$\begin{aligned} C_n &= \frac{1}{L_n} \left[\sum_{i=1}^n (w_i - 1) x_i e_i \right] \\ D_n &= \sum_{i=1}^n w_i x_i e_i \left[\frac{1}{\sum_{i=1}^n w_i x_i^2} - \frac{1}{\sum_{i=1}^n x_i^2} \right] \end{aligned}$$

Define the following

$$\begin{aligned} T_{1n} &= \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]} \\ T_{2n} &= \frac{[\sum_{i=1}^n x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]} \\ T_{3n} &= \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]}{[\sum_{i=1}^n x_i^2]^2} \\ T_{4n} &= \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^2}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^2} \\ T_{5n} &= (\hat{\beta} - \beta) \frac{\sum_{i=1}^n (w_i - 1) x_i^2}{\sum_{i=1}^n x_i^2} \\ T_{6n} &= (\hat{\beta} - \beta) \frac{[\sum_{i=1}^n (w_i - 1) x_i^2]^2}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]} \\ T_{41n} &= \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^2}{[\sum_{i=1}^n x_i^2]^3} \\ T_{42n} &= \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^3} \end{aligned}$$

$$\begin{aligned}
T_{61n} &= (\hat{\beta} - \beta) \frac{[\sum_{i=1}^n (w_i - 1)x_i^2]^2}{[\sum_{i=1}^n x_i^2]^2} \\
T_{62n} &= (\hat{\beta} - \beta) \frac{[\sum_{i=1}^n (w_i - 1)x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2][\sum_{i=1}^n x_i^2]}
\end{aligned}$$

Using the above terms, we have the following relations

$$\begin{aligned}
D_n &= -T_{1n} - T_{2n} \\
T_{1n} &= T_{3n} - T_{4n} \\
T_{2n} &= T_{5n} - T_{6n} \\
T_{4n} &= T_{41n} - T_{42n} \\
T_{6n} &= T_{61n} - T_{62n}
\end{aligned}$$

We show the terms involving

$$T_{4n}^2, T_{6n}^2, T_{4n}T_{6n}, T_{4n}T_{5n}, T_{4n}T_{3n}, T_{5n}T_{6n}, T_{3n}T_{6n}, T_{62n}C_n, T_{42n}C_n$$

are negligible, and explicitly calculate the rest of the terms, i.e.

$$C_n^2, T_{3n}^2, T_{5n}^2, T_{3n}C_n, T_{5n}C_n, T_{3n}T_{5n}, T_{61n}C_n, T_{41n}C_n$$

$$\begin{aligned}
&\sigma_n^{-2} E_B T_{4n}^2 \\
&= E_B \frac{[\sum_{i=1}^n W_i x_i e_i]^2 [\sum_{i=1}^n (w_i - 1)x_i^2]^4}{[\sum_{i=1}^n w_i x_i^2]^2 [\sum_{i=1}^n x_i^2]^4} \\
&\leq \frac{1}{n^2 k} \frac{1}{L_n^4} E_B \left[\sum_{i=1}^n W_i x_i e_i \right]^2 \left[\sum_{i=1}^n (w_i - 1)x_i^2 \right]^4 \\
&\leq \frac{1}{n^2 k} \frac{1}{L_n^4} [E_B (\sum_{i=1}^n W_i x_i e_i)^6]^{1/3} [E_B (\sum_{i=1}^n (w_i - 1)x_i^2)^6]^{2/3} \\
&= O_P(n^{-3})
\end{aligned}$$

$$\begin{aligned}
\sigma_n^{-2} E_B T_{6n}^2 &= (\hat{\beta} - \beta)^2 \frac{1}{L_n^2} E_B \sigma_n^2 \frac{[\sum_{i=1}^n W_i x_i^2]^4}{[\sum_{i=1}^n w_i x_i^2]^2} \\
&\leq \sigma_n^2 (\hat{\beta} - \beta)^2 \frac{1}{n^2 k} \frac{1}{L_n^2} E_B \left[\sum_{i=1}^n W_i x_i^2 \right]^4 \\
&= O_P(n^{-3})
\end{aligned}$$

From the above two inequalities, it can be seen that $\sigma_n^{-2} |E_B T_{4n} T_{6n}|$ is also $O_P(n^{-3})$.

$$\begin{aligned}
& |\sigma_n^{-2} E_B T_{4n} T_{5n}| \\
&= |\sigma_n^{-2} E_B (\hat{\beta} - \beta) \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^3}}| \\
&= \sigma_n^2 \frac{1}{L_n^3} |(\hat{\beta} - \beta) E_B \frac{[\sum_{i=1}^n W_i x_i e_i] [\sum_{i=1}^n W_i x_i^2]^3}{\sum_{i=1}^n w_i x_i^2}}| \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| E_B \left| \sum_{i=1}^n W_i x_i e_i \right| \left| \sum_{i=1}^n W_i x_i^2 \right|^3 \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| [E_B (\sum_{i=1}^n W_i x_i e_i)^4]^{3/4} [E_B (\sum_{i=1}^n W_i x_i^2)^4]^{1/4} \\
&= O_P(n^{-5/2})
\end{aligned}$$

$$\begin{aligned}
|\sigma_n^{-2} E_B T_{4n} T_{3n}| &= |\sigma_n^{-2} E_B \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i]^2 [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^4}}| \\
&\leq \sigma_n^3 \frac{1}{nk} \frac{1}{L_n^4} |E_B (\sum_{i=1}^n W_i x_i e_i)^2| \left| \sum_{i=1}^n W_i x_i^2 \right|^3 \\
&\leq \sigma_n^3 \frac{1}{nk} \frac{1}{L_n^4} [E_B (\sum_{i=1}^n W_i x_i e_i)^4]^{1/2} [E_B (\sum_{i=1}^n W_i x_i^2)^6]^{1/2} \\
&= O_P(n^{-5/2})
\end{aligned}$$

$$\begin{aligned}
|\sigma_n^{-2} E_B T_{6n} T_{5n}| &= \sigma_n^{-2} (\hat{\beta} - \beta)^2 \frac{1}{L_n^2} |E_B \frac{[\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2]}}| \\
&\leq \sigma_n \frac{1}{nk} \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 |E_B (\sum_{i=1}^n W_i x_i^2)^3| \\
&\leq \sigma_n \frac{1}{nk} \frac{1}{L_n^2} (\hat{\beta} - \beta)^2 [E_B (\sum_{i=1}^n W_i x_i^2)^6]^{1/2} \\
&= O_P(n^{-5/2})
\end{aligned}$$

$$\begin{aligned}
& |\sigma_n^{-2} E_B T_{3n} T_{6n}| \\
&= |\sigma_n^{-2} (\hat{\beta} - \beta) E_B \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^3}}| \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| E_B \left| \sum_{i=1}^n W_i x_i e_i \right| \left| \sum_{i=1}^n W_i x_i^2 \right|^3 \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| [E_B (\sum_{i=1}^n W_i x_i e_i)^4]^{1/4} [E_B (\sum_{i=1}^n W_i x_i^2)^4]^{3/4} \\
&= O_P(n^{-5/2})
\end{aligned}$$

$$\begin{aligned}
& |\sigma_n^{-2} \mathbb{E}_B T_{62n} C_n| \\
&= \sigma_n^{-2} |(\hat{\beta} - \beta) \mathbb{E}_B \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i] [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^3}}| \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| |\mathbb{E}_B [\sum_{i=1}^n W_i x_i e_i] [\sum_{i=1}^n W_i x_i^2]^3| \\
&\leq \sigma_n^2 \frac{1}{nk} \frac{1}{L_n^3} |(\hat{\beta} - \beta)| [\mathbb{E}_B (\sum_{i=1}^n W_i x_i e_i)^4]^{1/4} [\mathbb{E}_B (\sum_{i=1}^n W_i x_i^2)^4]^{3/4} \\
&= O_P(n^{-5/2})
\end{aligned}$$

$$\begin{aligned}
& |\sigma_n^{-2} \mathbb{E}_B T_{42n} C_n| \\
&= \sigma_n^{-2} |\mathbb{E}_B \frac{[\sum_{i=1}^n (w_i - 1) x_i e_i]^2 [\sum_{i=1}^n (w_i - 1) x_i^2]^3}{[\sum_{i=1}^n w_i x_i^2] [\sum_{i=1}^n x_i^2]^4}}| \\
&\leq \sigma_n^3 \frac{1}{nk} \frac{1}{L_n^4} |\mathbb{E}_B [\sum_{i=1}^n W_i x_i e_i]^2 [\sum_{i=1}^n W_i x_i^2]^3| \\
&\leq \sigma_n^3 \frac{1}{nk} \frac{1}{L_n^4} [\mathbb{E}_B (\sum_{i=1}^n W_i x_i e_i)^4]^{1/2} [\mathbb{E}_B (\sum_{i=1}^n W_i x_i^2)^6]^{1/2} \\
&= O_P(n^{-5/2})
\end{aligned}$$

We will now consider terms which are not negligible upto second order approximation.

$$\begin{aligned}
\mathbb{E}_B C_n^2 &= \frac{1}{L_n^2} \mathbb{E}_B [\sum_{i=1}^n (w_i - 1) x_i e_i]^2 \\
&= \frac{1}{L_n^2} \mathbb{E}_B \sum_{i=1}^n (w_i - 1)^2 x_i^2 e_i^2 + \frac{1}{L_n^2} \sum_{i \neq j} \mathbb{E}_B (w_i - 1) x_i e_i (w_j - 1) x_j e_j \\
&= \sigma_n^2 \frac{1}{L_n^2} [\sum_{i=1}^n x_i^2 e_i^2 + c_n \sum_{i \neq j} x_i x_j e_i e_j] \\
&= \sigma_n^2 \frac{1}{L_n^2} [\sum_{i=1}^n x_i^2 e_i^2] + \sigma_n^2 \frac{1}{L_n^2} c_n [\sum_{i \neq j} x_i x_j e_i e_j]
\end{aligned}$$

$$\begin{aligned}
& \sigma_n^{-2} \mathbb{E}_B T_{3n}^2 \\
&= \sigma_n^2 \frac{1}{L_n^4} \mathbb{E}_B [\sum_{i=1}^n W_i x_i e_i]^2 [\sum_{j=1}^n W_j x_j^2]^2 \\
&= \sigma_n^2 \frac{1}{L_n^4} \mathbb{E}_B [\sum_{i=1}^n W_i^2 x_i^2 e_i^2 + \sum_{i \neq j} W_i W_j x_i e_i x_j e_j] [\sum_{i=1}^n W_i^2 x_i^4 + \sum_{i \neq j} W_i W_j x_i^2 x_j^2]
\end{aligned}$$

$$\begin{aligned}
&= \sigma_n^2 \frac{1}{L_n^4} E_B \left[\sum_{i=1}^n W_i^4 x_i^6 e_i^2 + \sum_{i \neq j} W_i^2 W_j^2 x_i^2 e_i x_j^4 + \sum_{i=1}^n \sum_{k \neq l} W_k W_l W_i^2 x_i^2 e_i^2 x_k^2 x_l^2 \right. \\
&\quad \left. + \sum_{k=1}^n \sum_{i \neq j} W_i W_j W_k^2 x_i x_j x_k^4 e_i e_j + \sum_{i \neq j} \sum_{k \neq l} W_i W_j W_k W_l x_i x_j x_k^2 x_l^2 e_i e_j \right] \\
&= \sigma_n^2 \frac{1}{L_n^4} E_B \left[\sum_{i=1}^n W_i^4 x_i^6 e_i^2 + \sum_{i \neq j} W_i^2 W_j^2 x_i^2 e_i x_j^4 + 2 \sum_{i \neq j} W_i^3 W_j x_i^4 x_j^2 e_i^2 \right. \\
&\quad \left. + \sum_{i=1}^n \sum_{k \neq i} \sum_{l \neq i, k} W_k W_l W_i^2 x_i^2 x_k^2 x_l^2 e_i^2 + \sum_{k=1}^n \sum_{i \neq j} W_i W_j W_k^2 x_i x_j x_k^4 e_i e_j \right. \\
&\quad \left. + \sum_{i \neq j} \sum_{k \neq l} W_i W_j W_k W_l x_i x_j x_k^2 x_l^2 e_i e_j \right]
\end{aligned}$$

$$\begin{aligned}
\sigma_n^{-2} E_B T_{5n}^2 &= (\hat{\beta} - \beta)^2 \frac{1}{L_n^2} E_B \left(\sum_{i=1}^n W_i x_i^2 \right)^2 \\
&= (\hat{\beta} - \beta)^2 \frac{1}{L_n^2} \left(\sum_{i=1}^n x_i^4 + c_n \sum_{i \neq j} x_i^2 x_j^2 \right)
\end{aligned}$$

$$\begin{aligned}
&\sigma_n^{-2} E_B C_n T_{3n} \\
&= \sigma_n \frac{1}{L_n^3} E_B \left(\sum_{i=1}^n W_i x_i e_i \right)^2 \left(\sum_{i=1}^n W_i x_i^2 \right) \\
&= \sigma_n \frac{1}{L_n^3} E_B \left[\sum_{i=1}^n W_i^2 x_i^2 e_i^2 + \sum_{i \neq j} W_i W_j x_i x_j e_i e_j \right] \left[\sum_{i=1}^n W_i x_i^2 \right] \\
&= \sigma_n \frac{1}{L_n^3} E_B \left[\sum_{i=1}^n W_i^3 x_i^4 e_i^2 + \sum_{i \neq k} W_i^2 W_k x_i^2 x_k^2 e_i^2 \right. \\
&\quad \left. + 2 \sum_{i \neq j} W_i^2 W_j x_i^3 x_j e_i e_j + \sum_{i \neq j} \sum_{k \neq j} \sum_{k \neq i} W_i W_j W_k x_i x_j x_k^2 e_i e_j \right]
\end{aligned}$$

$$\begin{aligned}
\sigma_n^{-2} E_B C_n T_{5n} &= (\hat{\beta} - \beta) \frac{1}{L_n^2} E_B \left[\sum_{i=1}^n W_i x_i e_i \right] \left[\sum_{i=1}^n W_i x_i^2 \right] \\
&= (\hat{\beta} - \beta) \frac{1}{L_n^2} \left[\sum_{i=1}^n x_i^3 e_i + c_n \sum_{i \neq j} x_i x_j^2 e_i \right]
\end{aligned}$$

$$\begin{aligned}
&\sigma_n^{-2} E_B T_{3n} T_{5n} \\
&= (\hat{\beta} - \beta) \frac{1}{L_n^3} E_B \left[\sum_{i=1}^n W_i x_i e_i \right] \left[\sum_{i=1}^n W_i x_i^2 \right]^2 \\
&= (\hat{\beta} - \beta) \frac{1}{L_n^3} E_B \left[\sum_{i=1}^n W_i x_i e_i \right] \left[\sum_{i=1}^n W_i^2 x_i^4 + \sum_{i \neq j} W_i W_j x_i^2 x_j^2 \right]
\end{aligned}$$

$$\begin{aligned}
&= \sigma_n(\hat{\beta} - \beta) \frac{1}{L_n^3} E_B \left[\sum_{i=1}^n W_i^3 x_i^5 e_i + \sum_{i \neq k} W_i^2 W_k x_i^4 x_k e_i \right. \\
&\quad \left. + 2 \sum_{i \neq j} W_i^2 W_j x_i^3 x_j^2 e_i + \sum_{i \neq j} \sum_{k \neq j, k \neq i} W_i W_j W_k x_i^2 x_j^2 x_k e_k \right]
\end{aligned}$$

Note that $\sigma_n^{-2} E_B C_n T_{61n}$ is identical with above. Also, $\sigma_n^{-2} E_B C_n T_{41n}$ is same as $\sigma_n^{-2} E_B T_{3n}^2$. So that the second order term T_n is given by

$$\begin{aligned}
&n^{-2} T_{5n} \\
&= \frac{1}{L_n^2} c_{11} \sum_{i \neq j} x_i x_j e_i e_j + 3\sigma_n^2 \frac{1}{L_n^4} \left[c_4 \sum_{i=1}^n x_i^6 e_i^2 \right. \\
&\quad + c_{22} \sum_{i \neq j} x_i^2 e_i^2 x_j^4 + c_{112} \sum_{i=1}^n \sum_{k \neq l} x_i^2 x_k^2 x_l^2 e_i^2 + 2c_{31} \sum_{i \neq j} x_i^4 x_j^2 e_j^2 \\
&\quad + c_{112} \sum_{k=1}^n \sum_{i \neq j} x_i x_j x_k^4 e_i e_j + c_{1111} \sum_{i \neq j} \sum_{k \neq l} x_i x_j x_k^2 x_l^2 e_i e_j \Big] \\
&\quad + \frac{1}{L_n^4} \left(\sum_{i=1}^n x_i e_i \right)^2 \left[\sum_{i=1}^n x_i^4 + c_{11} \sum_{i \neq j} x_i^2 x_j^2 \right] \\
&\quad - 2\sigma_n \frac{1}{L_n^3} \left[c_3 \sum_{i=1}^n x_i^4 e_i^2 + c_{12} \sum_{i \neq j} x_i^2 x_j^2 e_i^2 \right. \\
&\quad + 2c_{12} \sum_{i \neq j} x_i^3 x_j e_i e_j + c_{111} \sum_{i \neq j} \sum_{k \neq j, k \neq i} x_i x_j x_k^2 e_i e_j \Big] \\
&\quad - 2 \frac{1}{L_n^3} \left(\sum_{i=1}^n x_i e_i \right) \left[\sum_{i=1}^n x_i^3 e_i + c_{11} \sum_{i \neq j} x_i x_j^2 e_i \right] \\
&\quad + 4\sigma_n \frac{1}{L_n^4} \left(\sum_{i=1}^n x_i e_i \right) \left[c_3 \sum_{i=1}^n x_i^5 e_i + c_{12} \sum_{i \neq k} x_i^4 x_k e_i \right. \\
&\quad \left. + 2c_{12} \sum_{i \neq j} x_i^3 x_j^2 e_i + c_{111} \sum_{i \neq j} \sum_{k \neq j, k \neq i} x_i^2 x_j^2 x_k e_k \right] \\
&= \frac{1}{L_n^2} c_{11} \sum_{i \neq j} x_i x_j e_i e_j + 3\sigma_n^2 \frac{1}{L_n^4} \left[c_{22} \sum_{i \neq j} x_i^2 e_i^2 x_j^4 + c_{112} \sum_{i=1}^n \sum_{k \neq l} x_i^2 x_k^2 x_l^2 e_i^2 \right. \\
&\quad + c_4 \sum_{i=1}^n x_i^6 e_i^2 + 2c_{31} \sum_{i \neq j} x_i^4 x_j^2 e_j^2 \Big] + \frac{1}{L_n^4} \left(\sum_{i=1}^n x_i e_i \right)^2 \left[\sum_{i=1}^n x_i^4 + c_{11} \sum_{i \neq j} x_i^2 x_j^2 \right] \\
&\quad - 2\sigma_n \frac{1}{L_n^3} \left[c_3 \sum_{i=1}^n x_i^4 e_i^2 + c_{12} \sum_{i \neq j} x_i^2 x_j^2 e_i^2 \right] \\
&\quad - 2 \frac{1}{L_n^3} \left(\sum_{i=1}^n x_i e_i \right) \left[\sum_{i=1}^n x_i^3 e_i + c_{11} \sum_{i \neq j} x_i x_j^2 e_i \right] \\
&\quad + 4\sigma_n \frac{1}{L_n^4} \left(\sum_{i=1}^n x_i e_i \right) \left[c_3 \sum_{i=1}^n x_i^5 e_i + O_P(n^{-3}) \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{L_n^4} \left(\sum_{i=1}^n x_i e_i \right)^2 \sum_{i=1}^n x_i^4 - 2 \frac{1}{L_n^3} \left(\sum_{i=1}^n x_i e_i \right) \sum_{i=1}^n x_i^3 e_i \\
&\quad + 3 \sigma_n^2 \frac{1}{L_n^4} \left[c_{22} \sum_{i \neq j} x_i^2 e_i^2 x_j^4 + c_4 \sum_{i=1}^n x_i^6 e_i^2 + 2 c_{31} \sum_{i \neq j} x_i^4 x_j^2 e_j^2 \right] - 2 \sigma_n \frac{1}{L_n^3} c_3 \sum_{i=1}^n x_i^4 e_i^2 \\
&\quad + \frac{1}{L_n^2} \sum_{i=1}^n x_i^2 e_i^2 (3 \sigma_n^2 c_{112} - 2 \sigma_n c_{12} - c_{11}) + O_P(n^{-3})
\end{aligned}$$

Note that Z_n and T_{3n} are bounded in probability. This proves Theorem 4.3.2(v). ■

4.8 Proofs for general p

Proof of Theorem 4.4.1

From (2.6.3) we have

$$\begin{aligned}
V_{GB1} &= \frac{1}{n \sigma_B^2} E_B (\hat{\beta}_B - \hat{\beta}) (\hat{\beta}_B - \hat{\beta})^T \\
&= \frac{1}{n \sigma_B^2} E_B \mathbf{G}^{-1} \mathbf{X}^T \mathbf{W}_0 \mathbf{R} \mathbf{R}^T \mathbf{W}_0 \mathbf{X} \mathbf{G}^{-1} \\
&= \frac{1}{n \sigma_B^2} E_B \mathbf{G}^{-1} \mathbf{X}^T \mathbf{C}_B \mathbf{X} \mathbf{G}^{-1} \quad \text{say}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{C}_B &= \mathbf{W}_0 \mathbf{e} \mathbf{e}^T \mathbf{W}_0^T - \mathbf{W}_0 \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W}_0^T - \mathbf{W}_0 \mathbf{P}_x \mathbf{e} \mathbf{e}^T \mathbf{W}_0^T + \mathbf{W}_0 \mathbf{P}_x \mathbf{e} \mathbf{e}^T \mathbf{P}_x \mathbf{W}_0^T \\
&= \mathbf{C}_1 - \mathbf{C}_2 - \mathbf{C}_2^T + \mathbf{C}_3
\end{aligned}$$

By routine calculation of the bootstrap expectation of the $(i, j)^{th}$ element of the matrices involved, we have the following:

$$\begin{aligned}
E_B C_{1ij} &= \begin{cases} \sigma_B^2 [\sum_{k=1}^n e_k^2 + c_{11} \sum_{k \neq l} e_k e_l] & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \\
E_B C_{2ij} &= \begin{cases} \sigma_B^2 [\sum_{1 \leq a, c \leq n} ((\mathbf{P}_x))_{ac} e_c e_a \\ + c_{11} \sum_{1 \leq c \leq n} \sum_{a \neq b} ((\mathbf{P}_x))_{bc} e_a e_c] & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \\
E_B C_{3ij} &= \begin{cases} \sigma_B^2 [\sum_{a, b, d=1}^n ((\mathbf{P}_x))_{ab} ((\mathbf{P}_x))_{ad} e_b e_d \\ + c_{11} \sum_{b, d=1}^n \sum_{a \neq c} ((\mathbf{P}_x))_{ab} ((\mathbf{P}_x))_{cd} e_b e_d] & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Note that all the off-diagonal terms in the above calculations are zero, and all the diagonal terms are identical. Thus we have $E_B \mathbf{C}_B = \sigma_B^2 \rho \mathbf{I}$, where

$$\rho = \left[\sum_{k=1}^n e_k^2 + c_{11} \sum_{k \neq l} e_k e_l \right] - 2 \left[\sum_{1 \leq a, c \leq n} ((\mathbf{P}_x))_{ac} e_c e_a + c_{11} \sum_{1 \leq c \leq n} \sum_{a \neq b} ((\mathbf{P}_x))_{bc} e_a e_c \right]$$

$$+[\sum_{a,b,d=1}^n ((\mathbf{P}_x))_{ab}((\mathbf{P}_x))_{ad}e_b e_d + c_{11} \sum_{b,d=1}^n \sum_{a \neq c} ((\mathbf{P}_x))_{ab}((\mathbf{P}_x))_{cd}e_b e_d] \quad (4.8.1)$$

Hence the variance expression is

$$\begin{aligned} V_{GB1} &= t_n^2 \mathbf{G}^{-1} + \frac{g}{n} \mathbf{G}^{-1} \text{ where} \\ g &= c_{11} \mathbf{e}^T [n\mathbf{J} - \mathbf{I}] \mathbf{e} + nc_{11} \mathbf{e}^T \mathbf{P}_x \mathbf{J} \mathbf{P}_x \mathbf{e} - 2nc_{11} \mathbf{e}^T \mathbf{J} \mathbf{P}_x \mathbf{e} - (1 - 2c_{11}) \mathbf{e}^T \mathbf{P}_x \mathbf{e} \end{aligned}$$

The first order term of the variance expression is from the first term of $E_B C_{1ij}$ and the rest of the terms are collected in g . It can be easily seen that after ignoring terms that are of a smaller order the expectation of g is d_1 , and $g - d_1 = O_P(n^{-1/2})$.

For the other variance estimator the calculations are as earlier. In this case we have that $V_{GB2} = E_B \mathbf{G}^{-1} \mathbf{X}^T \mathbf{C}_B \mathbf{X} \mathbf{G}^{-1}$, where $\mathbf{C}_B$ is as earlier, and we have $E_B \mathbf{C}_B = \sigma_B^2 \rho \mathbf{I}$, where ρ is given in (4.8.1). Hence

$$\begin{aligned} V_{GB2} &= t_n^2 \mathbf{G}^{-1} + \frac{r}{n} \mathbf{G}^{-1} \text{ where} \\ r &= (n\sigma_B^2 - 1) \mathbf{e}^T \mathbf{e} + n\sigma_B^2 g \end{aligned}$$

where g is as earlier. It can be easily seen that after ignoring terms that are of a smaller order the expectation of r is d_2 , and $d - d_2 = O_P(n^{-1/2})$. ■

Proof of Theorem 4.5.1

We do not show all the steps in this proof, since they follow the arguments used in the proofs given earlier.

(i) *External bootstrap* Here we use the same notations as that of *GBS*. This time the weight matrix is $\mathbf{W} = \text{diag}(w_1, w_2, \dots, w_n)$ with

$$\begin{aligned} E_B w_i &= 0 \\ \text{Var} w_i &= \sigma_B^2 \\ \text{Corr}(w_i, w_j) &= 0 \end{aligned}$$

Following the notations of *GBS* we have

$$\begin{aligned} E_B C_{1ij}^2 &= \begin{cases} \sigma_B^2 e_i^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \\ E_B C_{2ij}^2 &= \begin{cases} \sigma_B^2 \sum_{1 \leq c \leq n} ((\mathbf{P}_x))_{ic} e_c e_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \\ E_B C_{3ij}^2 &= \begin{cases} \sigma_B^2 \sum_{b,d=1}^n ((\mathbf{P}_x))_{ib} ((\mathbf{P}_x))_{id} e_b e_d & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus we have

$$\begin{aligned} &V_{GB3} - V_R \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{T}_n - \mathbf{T}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{R}_n \end{aligned}$$

where easy calculations show that $\mathbf{R}_n = O_P(n^{-5/2})$ ■

(ii)- (iii) *Weighted jackknives*

For this part of the proof we use the notation $\mathbf{G}_{(-i)} = \sum_{j \neq i} \mathbf{x}_j \mathbf{x}_j^T$, so that we have for the least squares estimate on the data with the i^{th} observation deleted

$$\begin{aligned}\hat{\beta}_{(-i)} &= \mathbf{G}_{(-i)}^{-1} \sum_{j \neq i} \mathbf{x}_j y_j \\ &= \mathbf{G}_{(-i)}^{-1} [\mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i] \\ &= \mathbf{G}_{(-i)}^{-1} \mathbf{G} \hat{\beta} - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i \mathbf{x}_i^T \beta - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i e_i\end{aligned}$$

and thus

$$\begin{aligned}\hat{\beta}_{(-i)} - \hat{\beta} &= (\mathbf{G}_{(-i)}^{-1} \mathbf{G} - \mathbf{I}) \hat{\beta} - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i \mathbf{x}_i^T \beta - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i e_i \\ &= \mathbf{G}_{(-i)}^{-1} (\mathbf{G} - \mathbf{G}_{(-i)}) \hat{\beta} - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i \mathbf{x}_i^T \beta - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i e_i \\ &= \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i \mathbf{x}_i^T (\hat{\beta} - \beta) - \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i e_i \\ &= -\mathbf{G}_{(-i)}^{-1} [\mathbf{x}_i e_i - \mathbf{x}_i \mathbf{x}_i^T (\hat{\beta} - \beta)] \\ &= -\mathbf{G}_{(-i)}^{-1} [a_i - b_i] \text{ say}\end{aligned}$$

$$\begin{aligned}(a_i - b_i)(a_i - b_i)^T &= a_i a_i^T - a_i b_i^T - b_i a_i^T + b_i b_i^T \\ &= e_i^2 \mathbf{x}_i \mathbf{x}_i^T - 2e_i \mathbf{x}_i^T (\hat{\beta} - \beta) \mathbf{x}_i \mathbf{x}_i^T + [\mathbf{x}_i^T (\hat{\beta} - \beta)]^2 \mathbf{x}_i \mathbf{x}_i^T \\ &= [e_i - \mathbf{x}_i^T (\hat{\beta} - \beta)]^2 \mathbf{x}_i \mathbf{x}_i^T \\ &= r_i^2 \mathbf{x}_i \mathbf{x}_i^T\end{aligned}$$

Since

$$\mathbf{G}_{(-i)}^{-1} = \mathbf{G}^{-1} + \frac{\mathbf{G}^{-1} \mathbf{x}_i \mathbf{x}_i^T \mathbf{G}^{-1}}{1 - \delta_i}$$

we have

$$(1 - \delta_i)^2 \mathbf{G}_{(-i)}^{-1} \mathbf{x}_i \mathbf{x}_i^T \mathbf{G}_{(-i)}^{-1} = \mathbf{G} \mathbf{x}_i \mathbf{x}_i^T \mathbf{G} \quad (4.8.2)$$

so that for Hinkley's jackknife $V_{JH} = \frac{n}{n-p} \mathbf{G}^{-1} \mathbf{X}^T \mathbf{D}_H \mathbf{X} \mathbf{G}^{-1}$ where $\mathbf{D}_H = \text{diag}(r_i^2, 1 \leq i \leq n)$.

Now

$$\begin{aligned}\mathbf{D}_H &= \text{diag}((e_i^2 - 2e_i \mathbf{x}_i^T (\hat{\beta} - \beta) + [\mathbf{x}_i^T (\hat{\beta} - \beta)]^2)) \\ &= \mathbf{T}_{ni} - 2\mathbf{D}_{2i} + \mathbf{D}_{3i} \text{ where}\end{aligned}$$

$$\mathbf{T}_n = \text{diag}(e_i^2, 1 \leq i \leq n)$$

$$\mathbf{D}_2 = \text{diag}(e_i \mathbf{x}_i^T (\hat{\beta} - \beta), 1 \leq i \leq n)$$

$$\mathbf{D}_3 = \text{diag}([\mathbf{x}_i^T (\hat{\beta} - \beta)]^2, 1 \leq i \leq n)$$

Note that $E[-2\mathbf{D}_2 + \mathbf{D}_3] = \mathbf{D}$ and therefore the result follows for Hinkley's jackknife.

For Wu's jackknife, using (4.8.2) we have

$$(1 - \delta_i)\mathbf{G}_{(-i)}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{G}_{(-i)}^{-1} = (1 + \delta_i + \delta_i^2(1 - \delta_i)^{-1})\mathbf{G}\mathbf{x}_i\mathbf{x}_i^T\mathbf{G}$$

so that $\mathbf{V}_{\text{JW}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_4\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$ where

$$\begin{aligned}\mathbf{D}_4 &= \text{diag}((1 + \delta_i + \delta_i^2(1 - \delta_i)^{-1})\tau_i^2, \quad 1 \leq i \leq n) \\ &= \text{diag}(1 + \delta_i + \delta_i^2(1 - \delta_i)^{-1})[\mathbf{T}_{ni} - 2\mathbf{D}_{2i} + \mathbf{D}_{3i}], \quad 1 \leq i \leq n) \\ &= \text{diag}(\mathbf{T}_{ni} + [\delta_i\mathbf{T}_{ni} - 2\mathbf{D}_{2i} + \mathbf{D}_{3i}] + \delta_i^2(1 - \delta_i)^{-1}\mathbf{T}_{ni} \\ &\quad + [\delta_i + \delta_i^2(1 - \delta_i)^{-1}] [-2\mathbf{D}_{2i} + \mathbf{D}_{3i}], \quad 1 \leq i \leq n)\end{aligned}$$

Our assumptions imply that $\delta_i = O(n^{-1})$, so that

$$\begin{aligned}\mathbf{V}_{\text{JW}} - \mathbf{V}_R &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{T}_n - \mathbf{T})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &\quad + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{D}_W + \mathbf{D})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + O_P(n^{-5/2})\end{aligned}$$

where

$$\mathbf{D}_W = \text{diag}(\delta_i\tau_i^2, \quad 1 \leq i \leq n)$$

therefore the result follows for Wu's jackknife.

(iv) *UBS*

$$\begin{aligned}\hat{\beta}_B - \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{e}I_A + [(\mathbf{X}^T\mathbf{W}_D\mathbf{X})^{-1} - (\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T\mathbf{W}\mathbf{e}I_A \\ &= C_n I_A + D_n I_A \quad (\text{say})\end{aligned}$$

and

$$\begin{aligned}D_n &= T_{1n} + T_{2n} \\ T_{1n} &= T_{11n} - T_{12n} \\ T_{2n} &= T_{21n} - T_{22n} \\ T_{12n} &= T_{121n} + T_{122n} \\ T_{22n} &= T_{221n} + T_{222n} \quad \text{where}\end{aligned}$$

$$\begin{aligned}T_{1n} &= [(\mathbf{X}^T\mathbf{W}_D\mathbf{X})^{-1} - (\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{e} \\ T_{2n} &= [(\mathbf{X}^T\mathbf{W}_D\mathbf{X})^{-1} - (\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{X}^T\mathbf{e} \\ T_{11n} &= (\mathbf{X}^T\mathbf{X})^{-1}[\mathbf{X}^T(\mathbf{I} - \mathbf{W})\mathbf{X}](\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{W} - \mathbf{I})\mathbf{e} \\ T_{12n} &= (\mathbf{X}^T\mathbf{W}_D\mathbf{X})^{-1}[\mathbf{X}^T(\mathbf{I} - \mathbf{W})\mathbf{P}_X(\mathbf{I} - \mathbf{W})\mathbf{P}_X(\mathbf{W} - \mathbf{I})\mathbf{e}\end{aligned}$$

$$\begin{aligned}
T_{21n} &= (\mathbf{X}^T \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{X}] (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e} \\
T_{22n} &= (\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{X}] (\mathbf{X}^T \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{X}] (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e} \\
T_{121n} &= (\mathbf{X}^T \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{W} - \mathbf{I}) \mathbf{e} \\
T_{122n} &= (\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{W} - \mathbf{I}) \mathbf{e} \\
T_{221n} &= (\mathbf{X}^T \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x \mathbf{e} \\
T_{222n} &= (\mathbf{X}^T \mathbf{W}_D \mathbf{X})^{-1} [\mathbf{X}^T (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x (\mathbf{I} - \mathbf{W}) \mathbf{P}_x \mathbf{e}
\end{aligned}$$

As in the proof of the $p = 1$ case, we show that the term from $C_n C_n^T$ is the leading term in the variance expression. The terms that contribute in the second order are $T_{11n} T_{11n}^T$, $T_{21n} T_{21n}^T$, $T_{11n} C_n^T$, $T_{121n} C_n^T$, $T_{21n} C_n^T$, $T_{221n} C_n^T$ and the transposes of the last four matrices. The rest of the terms are negligible. We first show the negligible cases.

$$\begin{aligned}
& \sigma_n^{-2} \mathbb{E}_B (T_{12n})_i^2 \\
&= \sigma_n \mathbb{E}_B \left[\sum_{beh=1}^n \sum_{acdfg=1}^p (\mathbf{X}^T \mathbf{W}_D \mathbf{X})_{ia}^{-1} x_{ba} W_b x_{bc} (\mathbf{X}^T \mathbf{X})_{cd}^{-1} x_{ed} W_e x_{ef} (\mathbf{X}^T \mathbf{X})_{fg}^{-1} x_{hg} W_h e_h \right]^2 \\
&\leq \sigma_n n^{-6} \sum_{acdfg=1}^p [\mathbb{E}_B (\sum_{1 \leq b \leq n} x_{ba} W_b x_{bc})^6 \mathbb{E}_B (\sum_{1 \leq e \leq n} x_{ed} W_e x_{ef})^6 \mathbb{E}_B (\sum_{1 \leq h \leq n} x_{hg} W_h e_h)^6] \\
&= O_P(\sigma_n n^{-3})
\end{aligned}$$

$$\begin{aligned}
& \sigma_n^{-2} \mathbb{E}_B (T_{22n})_i^2 \\
&= \mathbb{E}_B \left[\sum_{beh=1}^n \sum_{acdfg=1}^p (\mathbf{X}^T \mathbf{W}_D \mathbf{X})_{ia}^{-1} x_{ba} W_b x_{bc} (\mathbf{X}^T \mathbf{X})_{cd}^{-1} x_{ed} W_e x_{ef} (\mathbf{X}^T \mathbf{X})_{fg}^{-1} x_{hg} e_h \right]^2 \\
&\leq n^{-6} \sum_{acdfg=1}^p [\mathbb{E}_B (\sum_{beh=1}^n x_{ba} W_b x_{bc} x_{ed} W_e x_{ef} x_{hg} e_h)]^2 \\
&\leq n^{-5} \sum_{acdfg=1}^p [\mathbb{E}_B (\sum_{1 \leq b \leq n} x_{ba} W_b x_{bc})^4 \mathbb{E}_B (\sum_{1 \leq e \leq n} x_{ed} W_e x_{ef})^4] \\
&= O_P(n^{-3})
\end{aligned}$$

We now show that $\sigma_n^{-2} \mathbb{E}_B T_{11n}^2 I_{AC}$ and $\sigma_n^{-2} \mathbb{E}_B T_{21n}^2 I_{AC}$ do not contribute to the second order term.

$$\begin{aligned}
& \sigma_n^{-2} \mathbb{E}_B T_{11n}^2 I_{AC} \\
&\leq \sigma_n^{-2} [\mathbb{E}_B T_{11n}^4 \mathbb{E}_B I_{AC}]^{1/2} \\
&= O_P(\sigma_n \mathbb{E}_B (I_{AC})^{1/2} n^{-2}) \\
&= O_P(n^{-2})
\end{aligned}$$

We get the above relation because $\sigma_n^{-2} \mathbb{E}_B T_{11n}^4 = O_P(n^{-4})$. The calculation for the quantity $\sigma_n^{-2} \mathbb{E}_B T_{21n}^2 I_{AC}$ is similar.

$$\begin{aligned}
& |\sigma_n^{-2} E_B T_{122ni} C_{nj}| \\
&= |\sigma_n^3 E_B [\sum_{b e f g s=1}^n \sum_{a c d r=1}^p (X^T W_D X)_{ia}^{-1} x_{ba} W_b x_{bc} \\
&\quad (X^T X)_{cd}^{-1} x_{cd} W_c (P_x)_{ef} W_f (P_x)_{fg} W_g e_g (X^T X)_{jr}^{-1} x_{sr} W_s e_s | \\
&\leq O_P(\sigma_n^3 n^{-5/2})
\end{aligned}$$

The calculation for $\sigma_n^{-2} E_B T_{222ni} C_{nj}$ is similar. It can also be easily shown that the following terms $\sigma_n^{-2} E_B T_{11ni} T_{12nj}$, $\sigma_n^{-2} E_B T_{11ni} T_{21nj}$, $\sigma_n^{-2} E_B T_{11ni} T_{22nj}$, $\sigma_n^{-2} E_B T_{21ni} T_{12nj}$, $\sigma_n^{-2} E_B T_{21ni} T_{22nj}$, $\sigma_n^{-2} E_B T_{12ni} T_{22nj}$ do not contribute to the second order term, using Cauchy-Schwartz inequality. The rest of the terms are contributing, and the following calculations are similar to the techniques used earlier. We have the following results:

$$\sigma_n^{-2} E_B T_{11n} T_{11n}^T = \sigma_n^2 (X^T X)^{-1} X^T C_1 X (X^T X)^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$\begin{aligned}
C_{1ii} &= c_4 e_i^2 ((P_x))_{ii}^2 + c_{22} \sum_{a \neq i} e_a^2 ((P_x))_{ia}^2 \\
C_{1ij} &= c_{31} ((P_x))_{ij} [e_i^2 ((P_x))_{ii} + e_j^2 ((P_x))_{jj}] + c_{211} \sum_{a \neq i, j} e_a^2 ((P_x))_{ia} ((P_x))_{ja} \\
&\quad + c_{22} e_i e_j [((P_x))_{ii} ((P_x))_{jj} + ((P_x))_{ij}^2] + c_{211} [\sum_{a \neq i, j} \{ e_i e_a ((P_x))_{ii} ((P_x))_{ja} \\
&\quad + e_j e_a ((P_x))_{ij} ((P_x))_{ja} + e_i e_a ((P_x))_{ia} ((P_x))_{ij} \\
&\quad + e_j e_a ((P_x))_{jj} ((P_x))_{ia} \}] + c_{1111} \sum_{a \neq i, j} \sum_{b \neq i, j} e_b e_a ((P_x))_{ib} ((P_x))_{ja}
\end{aligned}$$

$$\sigma_n^{-2} E_B T_{21n} T_{21n}^T = (X^T X)^{-1} X^T C_2 X (X^T X)^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$C_{2ij} = \begin{cases} (\sum_{a=1}^n e_a (P_x)_{ia})^2 & \text{if } i = j \\ c_{11} (\sum_{a=1}^n \sum_{1 \leq b \leq n} e_a e_b (P_x)_{ia} (P_x)_{jb}) & \text{if } i \neq j \end{cases}$$

$$\sigma_n^{-2} E_B T_{11n} C_n^T = -\sigma_n (X^T X)^{-1} X^T C_3 X (X^T X)^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$C_{3ij} = \begin{cases} c_3 e_i^2 (P_x)_{ii}^2 & \text{if } i = j \\ c_{21} [e_i e_j (P_x)_{ii} + e_j^2 (P_x)_{ij}] & \text{if } i \neq j \end{cases}$$

$$\sigma_n^{-2} E_B T_{21n} C_n^T = -(X^T X)^{-1} X^T C_4 X (X^T X)^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$C_{4ij} = \begin{cases} e_i \sum_{a=1}^n e_a (P_x)_{ia} & \text{if } i = j \\ c_{11} e_j \sum_{a=1}^n e_a (P_x)_{ia} & \text{if } i \neq j \end{cases}$$

$$\sigma_n^{-2} E_B T_{121n} C_n^T = \sigma_n^2 (X^T X)^{-1} X^T C_5 X (X^T X)^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$C_{5ii} = c_4 e_i^2 ((P_x))_{ii}^2 + c_{31} e_j^2 \sum_{a \neq i} ((P_x))_{ia}^2 + c_{31} e_i \sum_{a \neq i} ((P_x))_{ii} ((P_x))_{ia}$$

$$\begin{aligned} C_{5ij} = & c_{31}e_ie_j((\mathbf{P}_x))_{ii}^2 + c_{31}e_j^2((\mathbf{P}_x))_{ij}((\mathbf{P}_x))_{jj} + c_{211}e_j \sum_{a \neq i, j} e_a((\mathbf{P}_x))_{aa}((\mathbf{P}_x))_{ia} \\ & + c_{22}e_j^2((\mathbf{P}_x))_{ii}((\mathbf{P}_x))_{ij} + c_{22}e_ie_j((\mathbf{P}_x))_{ij}^2 + c_{211}e_j^2 \sum_{a \neq i, j} ((\mathbf{P}_x))_{ja}((\mathbf{P}_x))_{ia} \end{aligned}$$

$$\sigma_n^{-2} E_B C_n C_n^T = -(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}_6 \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + O_P(n^{-5/2}) \text{ where}$$

$$C_{6ij} = \begin{cases} e_i^2 & \text{if } i = j \\ c_{11}e_ie_j & \text{if } i \neq j \end{cases}$$

Thus we have

$$\begin{aligned} V_{UBS} - V_R = & (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{D}_1 - \mathbf{T}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ & + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{M} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} + O_P(n^{-5/2}) \end{aligned}$$

where

$$\begin{aligned} \mathbf{M}_{ii} = & \sigma_n^2 [3c_4e_i^2((\mathbf{P}_x))_{ii}^2 + c_{22} \sum_{a \neq i} e_a^2((\mathbf{P}_x))_{ia}^2] + \left(\sum_{a=1}^n e_a(\mathbf{P}_x)_{ia}^2 - 2\sigma_n c_3 e_i^2(\mathbf{P}_x)_{ii}^2 \right. \\ & \left. - 2e_i \sum_{a=1}^n e_a(\mathbf{P}_x)_{ia} + 2\sigma_n^2 c_{31} e_i^2 \sum_{a \neq i} ((\mathbf{P}_x))_{ia}^2 \right) \\ \mathbf{M}_{ij} = & \sigma_n^2 [2c_{31} \{e_i^2((\mathbf{P}_x))_{ii}((\mathbf{P}_x))_{ij} + e_j^2((\mathbf{P}_x))_{jj}((\mathbf{P}_x))_{ij}\} \\ & + c_{211} \sum_{a \neq i, j} e_a^2((\mathbf{P}_x))_{ia}((\mathbf{P}_x))_{ja} + c_{211}(e_i^2 + e_j^2) \sum_{a \neq i, j} ((\mathbf{P}_x))_{ia}((\mathbf{P}_x))_{ja} \\ & + c_{22} \{e_j^2((\mathbf{P}_x))_{ii}((\mathbf{P}_x))_{ij} + e_i^2((\mathbf{P}_x))_{jj}((\mathbf{P}_x))_{ij}\} \\ & + c_{11}e_ie_j + c_{11} \left(\sum_{a=1}^n \sum_{1 \leq b \leq n} e_a e_b (\mathbf{P}_x)_{ia} (\mathbf{P}_x)_{jb} \right) \\ & - c_{11} [e_j \sum_{a=1}^n e_a (\mathbf{P}_x)_{ia} + e_i \sum_{a=1}^n e_a (\mathbf{P}_x)_{ja}] \\ & + \sigma_n [c_{21}(e_i^2 + e_j^2)(\mathbf{P}_x)_{ij} + c_{21}e_ie_j(\mathbf{P}_x)_{ii} + \mathbf{P}_x)_{jj}]] \end{aligned}$$

This establishes the result for UBS . ■

Proof of Theorem 4.5.2 We will first show that $\mathbf{X}^T \mathbf{U} \mathbf{X} \geq 0$. Let

$$\begin{aligned} \mathbf{D}_1 &= \text{diag}(((\mathbf{P}_x))_{11}, ((\mathbf{P}_x))_{22}, \dots, ((\mathbf{P}_x))_{nn}) \\ \mathbf{D}_2 &= \text{diag}(\tau_1^2, \tau_1^2, \dots, \tau_1^2) \\ \mathbf{U}_1 &= \mathbf{D}_1 \mathbf{P}_x \mathbf{D}_2 + \mathbf{D}_2 \mathbf{P}_x \mathbf{D}_1 \\ \mathbf{U}_2 &= \text{diag} \left(\sum_{a \neq i} \tau_a^2 ((\mathbf{P}_x))_{ia}^2 - 2\tau_i^2 ((\mathbf{P}_x))_{ii}^2 \right) \end{aligned}$$

Note that $\mathbf{U} = \sigma_n^2(\mathbf{U}_1 + \mathbf{U}_2)$. It is trivially seen that $\mathbf{X}^T \mathbf{U}_1 \mathbf{X}$ is positive definite, and (4.5.4) implies that $\mathbf{X}^T \mathbf{U}_2 \mathbf{X}$ is also positive definite.

(i) It is given that $\mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X} > 0$, and we require for $\xi \in \mathbb{R}^p$

$$\xi^T \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X}\xi \geq \xi^T \mathbf{X}^T(\mathbf{D} + \sigma_n^2 \mathbf{U})\mathbf{X}\xi \geq -\xi^T \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X}\xi$$

Note that we can restrict ξ to be in $U = \{\xi \in \mathbb{R}^p : \xi^T \xi = 1\}$. Then one of the requirement is

$$\begin{aligned} \xi^T \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X}\xi &\geq \xi^T \mathbf{X}^T(\mathbf{D} + \sigma_n^2 \mathbf{U})\mathbf{X}\xi \quad \forall \xi \in U \\ \Leftrightarrow \xi^T \mathbf{X}^T \mathbf{A} \mathbf{X} \xi &\geq \sigma_n^2 \xi^T \mathbf{X}^T \mathbf{U} \mathbf{X} \xi \quad \forall \xi \in U \\ \Leftrightarrow \xi^T \mathbf{X}^T \mathbf{A} \mathbf{X} \xi &\geq \sigma_n^2 \sup_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{U} \mathbf{X} \eta \quad \forall \xi \in U \\ \Leftrightarrow \inf_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{A} \mathbf{X} \eta &\geq \sigma_n^2 \sup_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{U} \mathbf{X} \eta \end{aligned}$$

and the other requirement is

$$\begin{aligned} \xi^T \mathbf{X}^T(\mathbf{D} + \sigma_n^2 \mathbf{U})\mathbf{X}\xi &\geq -\xi^T \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X}\xi \quad \forall \xi \in U \\ \Leftrightarrow \xi^T \mathbf{X}^T(2\mathbf{D} + \mathbf{A})\mathbf{X}\xi &\geq \sigma_n^2 - \xi^T \mathbf{X}^T \mathbf{U} \mathbf{X} \xi \quad \forall \xi \in U \\ \Leftrightarrow \xi^T \mathbf{X}^T \mathbf{A} \mathbf{X} \xi &\geq \sigma_n^2 - \inf_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{U} \mathbf{X} \eta \quad \forall \xi \in U \\ \Leftrightarrow \inf_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{A} \mathbf{X} \eta &\geq \sigma_n^2 \inf_{\eta \in U} \eta^T \mathbf{X}^T \mathbf{U} \mathbf{X} \eta \end{aligned}$$

This proves (i). The proof for (ii) is similar.

(iii) Given that $\text{Tr} \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X} > 0$, we require

$$\text{Tr} \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X} \geq \text{Tr} \mathbf{X}^T(\mathbf{D} + \sigma_n^2 \mathbf{U})\mathbf{X} \geq -\text{Tr} \mathbf{X}^T(\mathbf{D} + \mathbf{A})\mathbf{X}$$

We use linearity of trace to reduce the requirements to

$$\frac{-\text{Tr} \mathbf{X}^T(2\mathbf{D} + \mathbf{A})\mathbf{X}}{\text{Tr} \mathbf{X}^T \mathbf{U} \mathbf{X}} \leq \sigma_n^2 \leq \frac{\text{Tr} \mathbf{X}^T \mathbf{A} \mathbf{X}}{\text{Tr} \mathbf{X}^T \mathbf{U} \mathbf{X}}$$

This proves (iii). The proof for (iv) is similar. ■

Chapter 5

ESTIMATION OF DISTRIBUTION OF LEAST SQUARES ESTIMATOR IN HIGH DIMENSIONAL LINEAR REGRESSIONS: FIRST ORDER RESULTS AND DIMENSION ASYMPTOTICS

5.1 Introduction

In Chapters 3 and 4 we have discussed the problem of estimating the variance of the least squares estimator in linear regression. In this chapter we discuss the problem of estimating the entire distribution of the least squares estimator in linear regression.

Consider the linear model

$$y_{i:n} = \mathbf{x}_{i:n}^T \beta_n + e_{i:n}, \quad i = 1, \dots, n \quad (5.1.1)$$

where $y_{i:n}$ are n observations, $\mathbf{x}_{i:n}$ are the observed values of random design vectors in $\mathbb{R}^p$, β_n is a p dimensional unknown but constant parameter vector and $e_{i:n}$ are unobservable noise terms. We allow the regressors to be random, and the dimension p to tend to infinity with n , the data size. The model is same as the one considered in Mammen (1993). In many applications, models are used where the dimension p is not small compared to n . Then an asymptotic approach where p is fixed is misleading since the high dimensionality of the model is lost asymptotically. Asymptotics where p may increase with n has been considered in Huber (1981), Haberman (1977a,b), Shorack (1982), Bickel and Freedman (1983), Portnoy (1984,1985,1988), Mammen (1989) and Sauermann (1989) and others in different contexts. Mammen (1989) has appropriately named this kind of asymptotic study *dimension asymptotics*.

Let $\hat{\beta}$ be the least squares estimator of β . For a vector $c = (c_1 \dots c_p)^T$, consider the problem of estimating the distribution of $n^{1/2} c^T (\hat{\beta} - \beta)$ as $n \rightarrow \infty$. Mammen (1993) has shown that under certain conditions the paired bootstrap and wild bootstrap methods provide consistent estimators. In this chapter we extend the paired bootstrap based consistency result of Mammen (1993) to the case of UBS resampling. The conditions on the random weights used for resampling are given in the next section.

A major consideration in dimension asymptotics is to obtain the highest possible rate of growth of p with respect to n . The paired bootstrap typically requires $p^4/n^3 \rightarrow 0$ (Mammen

(1993)). Let $\{w_{i:n}\}$ be the sequence of non-negative, exchangeable weights used for *UBS*, and let $\sigma_n^2 = \text{Var}(w_{i:n})$. Our major result in this chapter is to show that by letting $\sigma_n^2 \rightarrow 0$, we can have best dimension asymptotic results. That is, distributional consistency results hold for generalised bootstrap even under the stringent condition $p/n \rightarrow 0$ when σ_n^2 tend to zero at an appropriate rate. The rate $p/n \rightarrow 0$ is the best achievable when bootstrapping from residuals, as proved by Bickel and Freedman (1983). We also study consistency of resampling with random weights whose variance satisfies a positive lower bound, for example the delete- d jackknives with $d/n \rightarrow c \in (0, 1)$ and the paired bootstrap. For these techniques, we show consistency when $p^2/n \rightarrow 0$. The consistency of m out of n bootstrap is also established.

The contents of this chapter is adapted from Chatterjee and Bose (1998c).

5.2 Assumptions and main result

5.2.1 Conditions on *UBS* weights

We will henceforth drop the suffix n from the notations, thus $y_i = y_{i:n}$, $\mathbf{x}_i = \mathbf{x}_{i:n}$, $\beta = \beta_n$ and $e_i = e_{i:n}$ from now on. We will also have occasion to use the usual matrix notation for the above, thus $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)^T$, $\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n)^T$ and $\mathbf{e} = (e_1 \ e_2 \ \dots \ e_n)^T$.

The conditions on *UBS* weights to be imposed for this chapter are discussed below. Let $\mathcal{W}$ be the set on which at least m_0 of the weights are greater than some fixed constant $k_2 > 0$. The value of m_0 is $\geq n/3$.

$$E(w_i) = 1 \quad (5.2.1)$$

$$0 < \sigma_n^2 = o(np^{-5/2}) \quad (5.2.2)$$

$$P_B[\mathcal{W}] = 1 - O(pn^{-1}) \quad (5.2.3)$$

$$c_{11} = O(n^{-1}) \quad (5.2.4)$$

$$\forall \ i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 3, \quad c_{i_1 i_2 \dots i_k} = O(n^{-k+1}) \quad (5.2.5)$$

$$\forall \ i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 4, \quad c_{i_1 i_2 \dots i_k} = O(\min(n^{-k+2}, 1)) \quad (5.2.6)$$

The bootstrap estimator is defined as

$$\hat{\beta}_B = (\sum w_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum w_i \mathbf{x}_i y_i I_{\mathcal{W}} + \hat{\beta}(1 - I_{\mathcal{W}})$$

Condition (5.2.3) ensures that $\hat{\beta}_B$ is the first term on the right hand side of the above relation a large number of times. Thus $\hat{\beta}_B$ is realised by ordinary least squares on the randomly weighted data $\{\sqrt{w_i}(\mathbf{x}_i, y_i), \ i = 1, \dots, n\}$, with a correction for the cases when a substantial number of weights are zero.

5.2.2 Assumptions on the model

We assume

$$(\mathbf{x}_i, y_i) \quad \text{are i.i.d.} \quad (5.2.7)$$

$$E y_i^2 < \infty, \quad i = 1, \dots, n \quad (5.2.8)$$

$$E \|\mathbf{x}_i\|^2 < \infty \quad i = 1, \dots, n \quad (5.2.9)$$

We also assume

$$\beta \text{ minimises } b \text{ in } E(y_i - \mathbf{x}_i^T b)^2 \quad (5.2.10)$$

However a careful study of the results of this chapter and Mammen (1993) reveals that the assumption about identical nature of their distribution may be dropped in favour of some simple moment conditions. However, we retain the *i.i.d.* assumption for simplicity.

We assume that $E\mathbf{x}_i\mathbf{x}_i^T$ is nonsingular, so that β is uniquely defined, and the additional assumption $E(y_i - \mathbf{x}_i^T\beta)^2 > 0$ ensures that the following are always fulfilled after a standardisation:

$$E\mathbf{x}_i\mathbf{x}_i^T = \mathbf{I}, \quad i = 1, \dots, n \quad (5.2.11)$$

$$0 < \inf_n E e_i^2 \leq \sup E e_i^2 < \infty \quad i = 1, \dots, n \quad (5.2.12)$$

As Mammen (1993) points out, the model conditions imply $E\mathbf{x}_i e_i = 0$, $\forall i$, but they do not imply the traditional homoscedastic assumption that given $\mathbf{X}$ the e_i are conditionally *i.i.d.* with mean zero. Also, the parameter β actually defines the linear model 'nearest' to the data, which is often a useful thing to consider even when linear models do not hold exactly.

5.2.3 Distributional consistency result

We denote $\mathcal{L}(\cdot | \mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n)$ by $\mathcal{L}^B$. The sup-norm distance between distributions is denoted by d_∞ .

Theorem 5.2.1 *Choose $c \in \mathbb{R}^p$ with $\|c\| = 1$. Consider a linear model (5.1.1) satisfying (5.2.7) - (5.2.12). For a fixed δ , set K to be the smallest integer greater than or equal to $2/\delta$ and assume that*

$$p^{1+\delta}/n \rightarrow 0 \quad (5.2.13)$$

$$\sup_n \sup_{\|d\|=1} E |d^T \mathbf{x}_i|^{4K} (1 + e_i^2) < \infty \quad (5.2.14)$$

$$E |(c^T \mathbf{x}_i) e_i|^2 > 0 \quad (5.2.15)$$

$$E |(c^T \mathbf{x}_i) e_i|^{2+\gamma} < \infty \text{ for some } \gamma > 0. \quad (5.2.16)$$

Consider resampling with weights which satisfy $c_{22} \rightarrow 1$ in addition to conditions given in subsection 5.2.1. Then

$$d_{\infty}(\mathcal{L}^B(n^{1/2}\sigma_n^{-1}c^T(\hat{\beta}_B - \hat{\beta})), \mathcal{L}(n^{1/2}c^T(\hat{\beta} - \beta))) \rightarrow 0 \text{ in probability} \quad (5.2.17)$$

holds if any of the following holds:

$$\delta \geq 1 \quad \text{and} \quad \sigma_n^2 > k > 0 \quad \forall n \quad (5.2.18)$$

$$1 > \delta \geq 1/3 \quad \text{and} \quad \sigma_n^2 = O(n^{-1}) \quad \forall n \quad (5.2.19)$$

$$1 > \delta \geq 0 \quad \text{and} \quad \sigma_n^4 p^5 n^{-2} \rightarrow 0 \quad (5.2.20)$$

REMARK 1. (a) The condition (5.2.20) ensures that $\delta = 0$ (i.e. $p/n \rightarrow 0$) is enough if $\sigma_n^2 = o(n^{-3/2})$. (b) For the m out of n bootstrap with replacement, with $m \rightarrow \infty$ and $m/n \rightarrow 0$ and $\sigma_n^2 = n/m$, consistency is achieved if either $p^2/n \rightarrow 0$ or if $\delta < 1$ and $p^5/m^2 \rightarrow 0$. Note that this restricts the choice of m to require that $nm^{-2(1+\delta)/5} \rightarrow 0$. This is a criterion for choice of m . Bickel and Sakov (1999) have also suggested some choices of m for the m out of n bootstrap in other contexts.

REMARK 2. Theorem 5.2.1 implies that the paired bootstrap is consistent if $p^2/n \rightarrow 0$, which is weaker than the rate $p^4/n^3 \rightarrow 0$ obtained by Mammen (1993). This is because in our general framework for resampling, we do not use the special properties of the paired bootstrap. Mammen (1993) has also shown that if $E[e_i|\mathbf{x}_i] = 0$ and $p/n \rightarrow 0$, then the paired bootstrap is consistent. For our results, we do not need any condition on conditional expectation of errors.

REMARK 3. We have replaced the Lindeberg condition of Mammen (1993) with a Lyapunov-type condition in (5.2.16). This is for convenience in the bootstrap distributional asymptotics. Conditions (5.2.13) and (5.2.14) are such that weakening of one leads to strengthening of the other. Mammen(1993) had conjectured, following Yin, Bai and Krishnaiah (1988), that for the design where elements of $\mathbf{x}_i$ are *i.i.d.*, finite fourth moments of the elements of $\mathbf{x}_i$ would suffice.

REMARK 4. Two important points about the results, that have been pointed out by Mammen (1993). The first is that if $E(e_i|\mathbf{x}_i) \neq 0$, then the bias of $n^{1/2}c^T(\hat{\beta} - \beta)$ is of the order $p/n^{1/2}$, which may tend to ∞ if $\delta < 1$. Thus for the mean zero normal approximation to work requires $\delta \geq 1$. The other one is that the conditional law $\mathcal{L}(n^{1/2}c^T(\hat{\beta} - \beta)|\mathbf{x}_1, \dots, \mathbf{x}_n)$ is also consistently estimated by *UBS*. This is because, as will be seen later, both the conditional and unconditional laws converge to the same limit $N(0, E(c^T \mathbf{x}_1)^2 e_1^2)$.

REMARK 5. In Chapters 3 and 4, *UBS* with asymptotically degenerate weights ($\sigma_n^2 \rightarrow 0$) was also considered. Consistency and asymptotic representation of variance estimator for such weights were obtained there. An important example of asymptotically degenerate weights is the delete-1 jackknife, for which the fourth central moment of the weights is $O(n)$. However, the degenerate weights that are considered here have bounded fourth moment, and this rules out the delete-1 jackknife in the present context. It can also be shown that distributional

consistency results do not hold if the fourth moment is unbounded. Note that Wu (1990) has shown that the delete- d jackknife with $d/n \rightarrow 0$ (this implies that c_4 is unbounded) is generally inconsistent for the distribution.

REMARK 6. It may be observed that the conditions on weights are similar to those used in resampling theory based on empirical processes. However, there are significant differences between the general empirical process based approach that is widely used and our approach to resampling in the present chapter.

Suppose, for some statistic, P, P_n, P_{Bn} are the actual law, the sample (empirical) law and the bootstrap law, respectively. The major thrust of empirical process literature is usually to prove both

$$\begin{aligned} n^{1/2}(P_n - P) &\Rightarrow G_P \\ n^{1/2}(P_{Bn} - P_n) &\Rightarrow G_P \end{aligned}$$

either almost surely or in probability, where G_P is a P -Brownian bridge. This implies usual bootstrap consistency, that is, $F_{Bn}(t) - F_n(t) \rightarrow 0$ uniformly over t , where F denotes the distribution function. For bootstrapping with random weights, usual assumptions on the weights are that they are either independent or exchangeable, with some restrictions on their lower order moments. Let $W_{i:n}$ be the weights used for resampling. Then, along with other conditions, Praestgaard and Wellner (1993) assume that for some $c > 0$,

$$n^{-1} \sum_{i=1}^n (W_{i:n} - 1)^2 \xrightarrow{P} c \quad (5.2.21)$$

See also Mason and Newton (1992), where an equivalent of (5.2.21) is given in W_{III} . The assumption (5.2.21) is only slightly weaker than $\sigma_n^2 = \text{Var}(W_{i:n}) \rightarrow c^2 > 0$. Relaxing the positive lower bound condition on σ_n^2 implies that weights may be considered that are asymptotically degenerate. It is evident from (5.2.20) that degenerate weights yield the best dimension asymptotic results.

REMARK 7. Theorem 5.2.1 establishes consistency of the UBS distribution estimator but does not give any information on its higher order behaviour. In Chapter 6 we study the higher order properties of UBS distribution estimator for certain class of weights.

5.3 Proofs

In order to prove the theorem, we need some results which we state now. Define $A = \mathbf{I} - n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ and $A^* = \mathbf{I} - n^{-1} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T}$. Also let $\lambda_{amaz}(A)$ the maximum of the absolute values of the eigenvalues of A .

Lemma 5.3.1 *Under the conditions of Theorem 5.2.1*

$$E_B \lambda_{amaz}(A^{*4}) = O_P(\max\{p^{2+\delta} n^{-2}, \sigma_n^4 p^3 n^{-2}\}) \quad (5.3.1)$$

5.3. PROOFS

Lemma 5.3.2 *Under the assumptions of Theorem 5.2.1, on the set where both the quantities $\lambda_{\max}(A) < 1$ and $\lambda_{\max}(A^*) < 1$, we have*

$$n^{1/2} \sigma_n^{-1} c^T (\hat{\beta}_B - \hat{\beta}) = n^{-1/2} \sum_{i=1}^n W_i c^T \mathbf{x}_i e_i + r_{nB} + o_P(1) \quad (5.3.2)$$

where $P_B[|r_{nB}| > \epsilon] = O_P(\sigma_n^4 p^4 n^{-2})$ for any $\epsilon > 0$.

The proof of the above results are given later. Conditions (5.2.13) and (5.2.14) are required for the proof of the above two lemma. We also require Theorem 5.3.1, which is Lemma 4.6 of Praetgaard and Wellner (1993). A variation of Theorem 5.3.1 may be found in Arenal-Gutierrez and Matran (1996). Their Theorem 3.3 may be used to get more general results under some circumstances. However, in this chapter as well as in Chapters 7 and 8 we use Theorem 5.3.1 instead of Theorem 3.3 of Arenal-Gutierrez and Matran (1996) since the technical conditions of Theorem 5.3.1 are generally easier to verify.

The m out of n bootstrap is one important resampling scheme where the set-up of Arenal-Gutierrez and Matran (1996) is directly applicable but that of Praetgaard and Wellner (1993) is not. Another important difference is that the Arenal-Gutierrez and Matran (1996) results are on almost sure convergence. A third important difference is in the method of proof. The proof of Theorem 5.3.1 uses Hajek's central limit theorem, which is a key technique in generalised bootstrap result. See the discussion of Mason and Newton (1992) and Praetgaard and Wellner (1993) on this. Arenal-Gutierrez and Matran (1996) use an independent technique of proof.

Theorem 5.3.1 (Praetgaard and Wellner (1993)) *Let $\{m\}$ be a sequence of natural numbers, let $\{a_{mj}\}$ be a triangular array of constants, and let B_{mj} , $j = 1, \dots, m$, $m \in \{m\}$ be a triangular array of row-exchangeable random variables such that*

$$m^{-1} \sum_{j=1}^m (a_{mj} - \bar{a}_m)^2 \rightarrow \tau^2 > 0, \quad m^{-1} \max_{j=1, \dots, m} (a_{mj} - \bar{a}_m)^2 \rightarrow 0$$

$$m^{-1} \sum_{j=1}^m (B_{mj} - \bar{B}_m)^2 \xrightarrow{P} c^2 > 0, \quad \lim_{K \rightarrow \infty} \limsup_{m \rightarrow \infty} E(B_{mj} - \bar{B}_m)^2 I_{\{|B_{mj} - \bar{B}_m| > K\}} = 0$$

Then

$$\frac{1}{\sqrt{m}} \sum_{j=1}^m (a_{mj} B_{mj} - \bar{a}_m \bar{B}_m) \Rightarrow N(0, c^2 \tau^2) \quad (5.3.3)$$

Proof of Theorem 5.2.1 One part of the proof involves the arguments to show that $\mathcal{L}(n^{1/2} c^T (\hat{\beta} - \beta)) \rightarrow N(0, E(c^T \mathbf{x}_1)^2 e_1^2)$, and this is identical with Mammen (1993), so we will not repeat it here. Note that $P[\lambda_{\max}(A) > 1] = o(1)$ and $P_B[\lambda_{\max}(A^*) > 1] = o_P(1)$. This is shown later. Thus by using Lemma 5.3.2 it is sufficient to show that $\mathcal{L}^B(n^{1/2} \sigma_n^{-1} c^T (\hat{\beta}_B - \hat{\beta})) \rightarrow N(0, E(c^T \mathbf{x}_1)^2 e_1^2)$ in probability. Note that for resampling with UBS weights, we have

$\mathbf{x}_i^* = \sqrt{w_i} \mathbf{x}_i$, $y_i^* = \sqrt{w_i} y_i$, $\mathbf{e}_i^* = \sqrt{w_i} \mathbf{e}_i$. The matrix $\mathbf{W}$ is the diagonal matrix with i^{th} diagonal entry w_i , and $W_i = \sigma_n^{-1}(w_i - 1)$. Now

$$\begin{aligned} & \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x] \\ &= \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, I_W = 1] + \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, I_W = 0] \\ &\leq \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, I_W = 1] + \mathbb{P}_B[I_W = 0] \\ &= \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, I_W = 1] + o_P(1) \end{aligned}$$

For the rest of this chapter it is implicitly assumed that calculations are carried out on the set W , although we drop the indicator of this set to simplify notations. Note that

$$\begin{aligned} \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x] &= \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, \lambda_{\max}(A^*) < 1] \\ &\quad + \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, \lambda_{\max}(A^*) \geq 1] \\ &= \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, \lambda_{\max}(A^*) < 1] + o_P(1) \end{aligned}$$

since

$$\begin{aligned} \mathbb{P}_B[n^{1/2} c^T (\hat{\beta}_B - \hat{\beta}) \leq x, \lambda_{\max}(A^*) \geq 1] &\leq \mathbb{P}_B[\lambda_{\max}(A^*) \geq 1] \\ &\leq \mathbb{E}_B[\lambda_{\max}(A^{*2})] \\ &= o_P(1) \text{ by Lemma 5.3.1} \end{aligned}$$

But on the set $\lambda_{\max}(A^*) < 1$, from Lemma 5.3.2, (5.3.2) holds. To complete the proof of the theorem we need to establish that

$$\mathcal{L}_B(n^{-1/2} \sum_{i=1}^n W_i c^T \mathbf{x}_i \mathbf{e}_i) \rightarrow N(0, E(c^T \mathbf{x}_1)^2 e_1^2) \quad (5.3.4)$$

in probability. For this part of the proof, let $\xi_i = c^T \mathbf{x}_i \mathbf{e}_i$. Note that for the bootstrap distribution, which is conditional on the data, ξ_i 's are constants.

We now apply Theorem 5.3.1. In our set-up, conditional on the data, $m = n$, $a_{mi} = \xi_i$ and $B_{mi} = W_i$. The first two conditions of the theorem are satisfied in probability from (5.2.15) and (5.2.16), with $\tau^2 = E\xi_1^2$. Also, since ξ_i and W_i are mean zero random variables, the mean adjustments in (5.3.3) may be ignored. Note that variance of W_i is identically one, so the third condition of the theorem is satisfied. From Lemma 4.7 and Lemma 3.1 of Praetgaard and Wellner (1993), the last condition of the above theorem is satisfied because of (5.2.1) and (5.2.6). This completes the proof of the theorem. $\blacksquare$

PROOF OF LEMMA 5.3.1

$$\text{Let } A_B = A^* - A = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - n^{-1} \sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T} = -\sigma_n n^{-1} \sum_{i=1}^n W_i \mathbf{x}_i \mathbf{x}_i^T$$

If $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ respectively denote the maximum and minimum eigenvalue, then it can be easily seen

$$\lambda_{\max}(A^*) \leq \lambda_{\max}(A) + \lambda_{\max}(A_B), \quad \lambda_{\min}(A^*) \geq \lambda_{\min}(A) + \lambda_{\min}(A_B)$$

Hence we have $\lambda_{\max}(A^{*4}) \leq 8[\lambda_{\max}(A^4) + \lambda_{\max}(A_B^4)]$. From Lemma 1 of Mammen (1993),

$$\lambda_{\max}(A^4) = O_P(p^{2+\delta}n^{-2}) \quad (5.3.5)$$

Thus we only have to establish the rates for A_B . We use the notation $\sum^{\neq}$ to denote sum over indices that are different from one another. Now

$$\begin{aligned} \text{tr}(A_B^4) &= \sigma_n^4 n^{-4} \left[\sum_{i=1}^n W_i^4 \|\mathbf{x}_i\|^8 + 4 \sum_{ij}^{\neq} W_i^3 W_j (\mathbf{x}_i \mathbf{x}_j^T)^2 \|\mathbf{x}_j\|^4 \right. \\ &\quad + \sum_{ij}^{\neq} W_i^2 W_j^2 [2 \|\mathbf{x}_i\|^2 \|\mathbf{x}_j\|^2 (\mathbf{x}_i^T \mathbf{x}_j)^2 + (\mathbf{x}_i^T \mathbf{x}_j)^4] \\ &\quad + \sum_{ijk}^{\neq} W_i^2 W_j W_k [4 \|\mathbf{x}_i\|^2 (\mathbf{x}_i^T \mathbf{x}_j)(\mathbf{x}_i^T \mathbf{x}_k)(\mathbf{x}_j^T \mathbf{x}_k) + 2(\mathbf{x}_i^T \mathbf{x}_j)^2 (\mathbf{x}_i^T \mathbf{x}_k)^2 (\mathbf{x}_k \mathbf{x}_k^T)] \\ &\quad \left. + \sum_{ijkl}^{\neq} W_i W_j W_k W_l (\mathbf{x}_i^T \mathbf{x}_j)(\mathbf{x}_j^T \mathbf{x}_k)(\mathbf{x}_k^T \mathbf{x}_l)(\mathbf{x}_l^T \mathbf{x}_i) \right] \end{aligned}$$

where $k_1, \dots, k_4$ are constants. We only have to show the rate for the above expression. To this end we calculate the bootstrap expectation of $\text{tr}(A_B^4)$.

$$\begin{aligned} \mathbb{E}_B \text{tr}(A_B^4) &= \sigma_n^4 n^{-4} \left[\sum_{i=1}^n c_4 \|\mathbf{x}_i\|^8 + 4 \sum_{ij}^{\neq} c_{31} (\mathbf{x}_i \mathbf{x}_j^T)^2 \|\mathbf{x}_j\|^4 \right. \\ &\quad + \sum_{ij}^{\neq} c_{22} [2 \|\mathbf{x}_i\|^2 \|\mathbf{x}_j\|^2 (\mathbf{x}_i^T \mathbf{x}_j)^2 + (\mathbf{x}_i^T \mathbf{x}_j)^4] \\ &\quad + \sum_{ijk}^{\neq} c_{211} [4 \|\mathbf{x}_i\|^2 (\mathbf{x}_i^T \mathbf{x}_j)(\mathbf{x}_i^T \mathbf{x}_k)(\mathbf{x}_j^T \mathbf{x}_k) + 2(\mathbf{x}_i^T \mathbf{x}_j)^2 (\mathbf{x}_i^T \mathbf{x}_k)^2 (\mathbf{x}_k \mathbf{x}_k^T)] \\ &\quad \left. + \sum_{ijkl}^{\neq} c_{1111} (\mathbf{x}_i^T \mathbf{x}_j)(\mathbf{x}_j^T \mathbf{x}_k)(\mathbf{x}_k^T \mathbf{x}_l)(\mathbf{x}_l^T \mathbf{x}_i) \right] \\ &= O_P(\sigma_n^4 n^{-2} p^3) \end{aligned}$$

The last statement follows from using the moment conditions on the weights used for UBS and (5.2.14). Thus we have

$$\mathbb{E}_B \lambda_{\max}(A_B^4) = O_P(\sigma_n^4 p^3 n^{-2}) \quad (5.3.6)$$

Putting together the results of (5.3.5) and (5.3.6) the lemma is established. ■

In order to prove Lemma 5.3.2 we need two more results which we now state and prove. For any i , define $A_i = \sum_{j \neq i} \mathbf{x}_j \mathbf{x}_j^T - nI$. Observe that $A = n^{-1}A_i + n^{-1}\mathbf{x}_i \mathbf{x}_i^T$, and A_i and $\mathbf{x}_i$ are independent. We employ this fact repeatedly in proofs that come later, so it is useful to discuss the rate for $\lambda_{\max}(A_i)$.

Lemma 5.3.3 *Under the conditions of Theorem 5.2.1*

$$\sup_i E \lambda_{\max} \left(\frac{A_i}{n} \right)^{2K} = O(p^{K+K\delta/2} n^{-K})$$

PROOF OF LEMMA 5.3.3 Note that for any i , $\frac{A_i}{n} = A - n^{-1}\mathbf{x}_i \mathbf{x}_i^T$. We have

$$\lambda_{\max} \left(\frac{A_i}{n} \right)^{2K} \leq 2^{2K-1} [\lambda_{\max}(A^{2K}) + \lambda_{\max}(n^{-1}\mathbf{x}_i \mathbf{x}_i^T)^{2K}]$$

From Lemma 1 of Mammen (1993), we have $E \lambda_{\max}(A^{2K}) = O(p^{K+K\delta/2} n^{-K})$. Further $n^{-1}\mathbf{x}_i \mathbf{x}_i^T$ is a rank 1 matrix, with its only non-zero eigenvalue being $n^{-1} \|\mathbf{x}_i\|^2$. Now using Lemma 0 of Mammen (1993), we have $E \|\mathbf{x}_i\|^{4K} p^{-2K}$ is uniformly bounded over i , so that $E \lambda_{\max}(n^{-1}\mathbf{x}_i \mathbf{x}_i^T)^{2K}$ is $O(p^{2K} n^{-2K})$ uniformly over i . This concludes the proof. ■

Now define

$$\begin{aligned} \hat{\beta}^K &= \beta + n^{-1} \sum_{h=1}^{K-1} A^h \sum_{i=1}^n \mathbf{x}_i e_i \\ \hat{\beta}_B^K &= \beta + n^{-1} \sum_{h=1}^{K-1} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* \end{aligned}$$

Lemma 5.3.4 *Under the assumptions of Theorem 5.2.1, on the set where both the quantities $\lambda_{\max}(A) < 1$ and $\lambda_{\max}(A^*) < 1$, we have*

$$n^{1/2} \sigma_n^{-1} c^T (\hat{\beta}_B^K - \hat{\beta}^K) = r_{nB}$$

where $P_B[|r_{nB}| > \epsilon] \xrightarrow{P} 0$ for any $\epsilon > 0$.

PROOF OF LEMMA 5.3.4 Note that

$$\begin{aligned} & n^{1/2} \sigma_n^{-1} c^T (\hat{\beta}_B^K - \hat{\beta}^K) \\ &= n^{-1/2} \sigma_n^{-1} c^T \left(\sum_{h=1}^{K-1} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* - \sum_{h=1}^{K-1} A^h \sum_{i=1}^n \mathbf{x}_i e_i \right) \\ &= n^{-1/2} \sigma_n^{-1} c^T \sum_{h=1}^{K-1} [(A^{*h} - A^h) \sum_{i=1}^n w_i \mathbf{x}_i e_i + A^h \sum_{i=1}^n (w_i - 1) \mathbf{x}_i e_i] \end{aligned} \quad (5.3.7)$$

The sum over h is a finite sum with K being a constant, so it suffices to consider the individual terms in the summations and establish the following three results for every h :

$$\mathbb{E}_B[n^{-1/2}\sigma_n^{-1}c^TA^h\sum_{i=1}^n(w_i-1)\mathbf{x}_i e_i]^2 = o_P(1) \quad (5.3.8)$$

$$\mathbb{E}_B[n^{-1/2}\sigma_n^{-1}c^T(A^{*h}-A^h)\sum_{i=1}^n(w_i-1)\mathbf{x}_i e_i] = o_P(1) \quad (5.3.9)$$

$$\mathbb{E}_B[n^{-1/2}\sigma_n^{-1}c^T(A^{*h}-A^h)\sum_{i=1}^n\mathbf{x}_i e_i] = o_P(1) \quad (5.3.10)$$

Proof of (5.3.8)

$$\begin{aligned} & \mathbb{E}_B[n^{-1/2}\sigma_n^{-1}c^TA^h\sum_{i=1}^n(w_i-1)\mathbf{x}_i e_i]^2 \\ &= n^{-1}\mathbb{E}_B[\sum_{i=1}^n W_i(c^TA^h\mathbf{x}_i e_i)^2] \\ &= n^{-1}\sum_{i=1}^n (c^TA^h\mathbf{x}_i)^2 e_i^2 + n^{-1}c_{11}\sum_{ij}^{\neq} (c^TA^h\mathbf{x}_i e_i)(c^TA^h\mathbf{x}_j e_j) \\ &= n^{-1}(1-c_{11})\sum_{i=1}^n (c^TA^h\mathbf{x}_i)^2 e_i^2 + n^{-1}c_{11}[\sum_{i=1}^n (c^TA^h\mathbf{x}_i e_i)^2] \end{aligned} \quad (5.3.11)$$

From Lemma 2 of Mammen (1993) we know that $n^{-1}[\sum_{i=1}^n (c^TA^h\mathbf{x}_i e_i)^2] = O_P(p)$, and since $c_{11} = O(n^{-1})$, the second expression in (5.3.11) is $O_P(p/n) = o_P(1)$. For the other quantity in (5.3.11), recall that

$$A = n^{-1}\sum_{i=1}^n (\mathbf{x}_i \mathbf{x}_i^T - \mathbf{I}) = n^{-1}A_i + n^{-1}\mathbf{x}_i \mathbf{x}_i^T, \quad (5.3.12)$$

Thus we have

$$\begin{aligned} & n^{-1}\sum_{i=1}^n (c^TA^h\mathbf{x}_i)^2 e_i^2 \\ &= n^{-3}\sum_{i=1}^n [c^T(A_i + \mathbf{x}_i \mathbf{x}_i^T)A^{h-1}\mathbf{x}_i]^2 e_i^2 \\ &\leq 2n^{-3}\sum_{i=1}^n [(c^TA_i A^{h-1}\mathbf{x}_i)^2 e_i^2 + (c^T\mathbf{x}_i \mathbf{x}_i^T A^{h-1}\mathbf{x}_i)^2 e_i^2] \end{aligned} \quad (5.3.13)$$

Consider the second term in the above expression. We have

$$n^{-3}\sum_{i=1}^n [c^T\mathbf{x}_i \mathbf{x}_i^T A^{h-1}\mathbf{x}_i]^2 e_i^2 \leq n^{-3}\lambda_{\max}(A^{2(h-1)})\sum_{i=1}^n [c^T\mathbf{x}_i]^2 \|\mathbf{x}_i\|^4 e_i^2 \quad (5.3.14)$$

Now by an easy use of (5.2.14) and Lemma 0 of Mammen (1993),

$$E[c^T \mathbf{x}_i]^2 \|\mathbf{x}_i\|^4 e_i^2 \leq [E(c^T \mathbf{x}_i)^4 e_i^2]^{1/2} [E\|\mathbf{x}_i\|^8 e_i^2]^{1/2} = O(p^2)$$

By using this and Lemma 5.3.1, at (5.3.14) we have $O_P((\frac{p^{1/2+\delta/4}}{n^{1/2}})^{2(h-1)} p^2 n^{-2}) = o_P(1)$. We are left with the first term in (5.3.13). We use the same technique as above, that is, writing one of the A as in (5.3.12) and thus reducing the power of A . The above scheme is used repeatedly for the rest of the calculations. For example, in the next step, for the first term in (5.3.13), we have

$$\begin{aligned} n^{-3} \sum_{i=1}^n [c^T A_i A^{h-1} \mathbf{x}_i]^2 e_i^2 &= n^{-1} \sum_{i=1}^n [c^T \frac{A_i}{n} A^{h-1} \mathbf{x}_i]^2 e_i^2 \\ &= n^{-3} \sum_{i=1}^n [c^T \frac{A_i}{n} (A_i + \mathbf{x}_i \mathbf{x}_i^T) A^{h-2} \mathbf{x}_i]^2 e_i^2 \\ &\leq 2n^{-3} \sum_{i=1}^n [(c^T \frac{A_i}{n} A_i A^{h-2} \mathbf{x}_i)^2 e_i^2 + (c^T \frac{A_i}{n} \mathbf{x}_i \mathbf{x}_i^T A^{h-2} \mathbf{x}_i)^2 e_i^2] \end{aligned}$$

Now since A_i and $\mathbf{x}_i$ are independent, and $\lambda_{\max}(n^{-1} A_i) = O_P(p^{1/2+\delta/4} n^{-1/2})$, the second term is obtained to be $o_P(1)$, using similar calculations as in (5.3.14), based on (5.2.14) and Lemma 0. The term left is $n^{-1} \sum_{i=1}^n (c^T (\frac{A_i}{n})^2 A^{h-2} \mathbf{x}_i)^2 e_i^2$. Repeating this process, ultimately we have to show $n^{-1} \sum_{i=1}^n (c^T (\frac{A_i}{n})^h \mathbf{x}_i)^2 e_i^2 = o_P(1)$, which follows easily using (5.2.14). This concludes the proof of (5.3.8).

For (5.3.9) and (5.3.10) we show the results for $h = 1$ and $h \geq 2$ separately. The result for $h = 1$ follows from direct calculation, and for higher h we employ the technique of reducing the power of A that we used in the above proof.

Proof of (5.3.9) for $h = 1$. We have

$$\begin{aligned} &n^{-1/2} \sigma_n^{-1} c^T [(A^{*h} - A^h) \sum_{i=1}^n (w_i - 1) \mathbf{x}_i e_i] \\ &= n^{-1/2} c^T A_B \sum_{i=1}^n W_i \mathbf{x}_i e_i \\ &= n^{-1/2} \sum_{i=1}^n W_i [c^T A_B \mathbf{x}_i] e_i \\ &= -n^{-1/2} n^{-1} \sigma_n \sum_{i=1}^n \sum_{j=1}^n W_i W_j [c^T \mathbf{x}_j \mathbf{x}_j^T \mathbf{x}_i] e_i \end{aligned}$$

A direct calculation shows that we have $E_B[n^{-1/2} n^{-1} \sigma_n \sum_{i=1}^n \sum_{j=1}^n W_i W_j [c^T \mathbf{x}_j \mathbf{x}_j^T \mathbf{x}_i] e_i]^2$ is $O_P(\sigma_n^2 p^2 n^{-1})$. This is $o_P(1)$ under any one of the three alternative conditions of the theorem.

Proof of (5.3.10) for $h = 1$. As earlier

$$\begin{aligned} n^{-1/2} \sigma_n^{-1} c^T (A^{*h} - A^h) \sum_{i=1}^n \mathbf{x}_i e_i &= n^{-1/2} n^{-1} \sigma_n^{-1} c^T (A_B) \sum_{i=1}^n \mathbf{x}_i e_i \\ &= -n^{-1/2} n^{-1} \sum_{j=1}^n W_j c^T \mathbf{x}_j \mathbf{x}_j^T \left(\sum_{i=1}^n \mathbf{x}_i e_i \right) \end{aligned} \quad (5.3.15)$$

Again a direct calculation shows that

$$\mathbb{E}_B [n^{-1/2} n^{-1} \sum_{j=1}^n W_j c^T \mathbf{x}_j \mathbf{x}_j^T \left(\sum_{i=1}^n \mathbf{x}_i e_i \right)]^2$$

is $O_P(p/n)$.

Thus (5.3.9) and (5.3.10) are proved for $h = 1$, completing the proof of the Lemma when (5.2.18) holds since $\delta \geq 1$ implies $K \leq 2$. We now concentrate on (5.3.9) and (5.3.10) under (5.2.19) or (5.2.20) and $h > 1$.

Proof of (5.3.9) for $h > 1$. Since $A^* = A + A_B$, we have

$$\begin{aligned} &n^{-1/2} \sigma_n^{-1} c^T A^{*h} \sum_{i=1}^n (w_i - 1) \mathbf{x}_i e_i \\ &= n^{-1/2} c^T A^{*h} \sum_{i=1}^n W_i \mathbf{x}_i e_i \\ &= n^{-1/2} c^T (A + A_B) A^{*(h-1)} \sum_{i=1}^n W_i \mathbf{x}_i e_i \\ &= n^{-1/2} c^T A A^{*(h-1)} \sum_{i=1}^n W_i \mathbf{x}_i e_i + n^{-1/2} c^T A_B A^{*(h-1)} \sum_{i=1}^n W_i \mathbf{x}_i e_i \end{aligned} \quad (5.3.16)$$

Repeating this process h times, we have

$$\begin{aligned} &n^{-1/2} \sigma_n^{-1} c^T A^{*h} \sum_{i=1}^n (w_i - 1) \mathbf{x}_i e_i \\ &= \sum_{j=0}^{h-1} c^T A^j A_B A^{*(h-j-1)} n^{-1/2} \sum_{i=1}^n W_i \mathbf{x}_i e_i + c^T A^h n^{-1/2} \sum_{i=1}^n W_i \mathbf{x}_i e_i \end{aligned} \quad (5.3.17)$$

The last term in (5.3.17) cancels out with the identical term left from (5.3.9). For the rest of the terms in (5.3.17), for any $j = 0, \dots, h-1$ we have

$$\begin{aligned} &\mathbb{E}_B |n^{-1/2} c^T A^j A_B A^{*(h-j-1)} \sum_{i=1}^n W_i \mathbf{x}_i e_i| \\ &\leq [\mathbb{E}_B \|c^T A^j A_B\|^2]^{1/2} [\mathbb{E}_B \|A^{*(h-1)} n^{-1/2} \sum_{i=1}^n W_i \mathbf{x}_i e_i\|^2]^{1/2} \end{aligned}$$

$$\begin{aligned}
&\leq \lambda_{\max}(A^j)[\mathbb{E}_B \lambda_{\max}(A_B^2)]^{1/2}[\mathbb{E}_B \lambda_{\max}(A^{*2(h-j-1)})||n^{-1/2} \sum_{i=1}^n W_i \mathbf{x}_i e_i||^2]^{1/2} \\
&\leq [\mathbb{E}_B \lambda_{\max}(A_B^2)]^{1/2}[\mathbb{E}_B ||n^{-1/2} \sum_{i=1}^n W_i \mathbf{x}_i e_i||^2]^{1/2} \\
&= O_P(\sigma_n p^{3/4} n^{-1/2} p^{1/2}) \\
&= O_P(\sigma_n p^{5/4} n^{-1/2})
\end{aligned}$$

With the second set of conditions in Theorem 5.2.1, that is, $p^4/n^3 \rightarrow 0$ and $\sigma_n^2 = O(n^{-1})$ the above is $o_P(1)$. With the third set of conditions also the same result holds.

Proof of (5.3.10) for $h > 1$. The calculations for this part is slightly more elaborate but in spirit similar to what was used for (5.3.9). Note that the expression in (5.3.10) can be expanded by using $A^* = A + A_B$, so every term in the expansion has A and A_B in various combinations. The terms without any A_B factors cancel out. The terms that involve two or more A_B matrices can be shown to be $o_P(1)$ easily. This leaves all the terms that have a single A_B matrix in them which are also $o_P(1)$. Thus

$$\begin{aligned}
&n^{-1/2} \sigma_n^{-1} c^T A^{*h} \sum_{i=1}^n \mathbf{x}_i e_i \\
&= n^{-1/2} \sigma_n^{-1} c^T A^{*h} \sum_{i=1}^n \mathbf{x}_i e_i \\
&= n^{-1/2} \sigma_n^{-1} c^T (A + A_B) A^{*(h-1)} \sum_{i=1}^n \mathbf{x}_i e_i \\
&= n^{-1/2} \sigma_n^{-1} c^T A A^{*(h-1)} \sum_{i=1}^n \mathbf{x}_i e_i + n^{-1/2} \sigma_n^{-1} c^T A_B A^{*(h-1)} \sum_{i=1}^n \mathbf{x}_i e_i \quad (5.3.18)
\end{aligned}$$

We first deal with the second term in (5.3.18). This is split into two terms, thus

$$\begin{aligned}
&n^{-1/2} \sigma_n^{-1} c^T A_B A^{*(h-1)} \sum_{i=1}^n \mathbf{x}_i e_i \\
&= n^{-1/2} \sigma_n^{-1} c^T A_B A A^{*(h-2)} \sum_{i=1}^n \mathbf{x}_i e_i + n^{-1/2} \sigma_n^{-1} c^T A_B^2 A^{*(h-2)} \sum_{i=1}^n \mathbf{x}_i e_i \quad (5.3.19)
\end{aligned}$$

First take the second term in (5.3.19).

$$\begin{aligned}
&\mathbb{E}_B |n^{-1/2} \sigma_n^{-1} c^T A_B^2 A^{*(h-2)} \sum_{i=1}^n \mathbf{x}_i e_i| \\
&\leq [\sigma_n^{-2} \mathbb{E}_B \lambda_{\max}(A_B^4)]^{1/2} [\mathbb{E}_B \lambda_{\max}(A^{*2(h-2)}) ||n^{-1/2} \sum_{i=1}^n \mathbf{x}_i e_i||^2]^{1/2} \\
&= O_P(\sigma_n^{-1} \sigma_n^2 p^2 n^{-1}) \\
&= o_P(1)
\end{aligned}$$

Proceeding in this way, the terms from (5.3.18) and (5.3.19) that have two or more A_B terms in them can be shown to be small, so we are eventually left to show

$$\mathbb{E}_B[n^{-1/2}\sigma_n^{-1}c^TA^aA_BA^{(h-a-1)}\sum_{i=1}^n\mathbf{x}_i\mathbf{e}_i]^2 = o_P(1) \quad (5.3.20)$$

for $a = 0, \dots, h-1$. For convenience, set $\eta_1 = A^a c$ and $\eta_2 = n^{-1/2}A^{(h-a-1)}\sum_{i=1}^n\mathbf{x}_i\mathbf{e}_i$. Thus (5.3.20) reduces to showing $\sigma_n^{-1}\eta_1^TA_B\eta_2 = o_P(1)$.

$$\begin{aligned} \mathbb{E}_B[\sigma_n^{-1}\eta_1^TA_B\eta_2]^2 &= n^{-2}\mathbb{E}_B[\sum_{i=1}^n W_i\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 \\ &= n^{-2}(1 - c_{11})\sum_{i=1}^n[\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 + n^{-2}c_{11}[\sum_{i=1}^n\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 \end{aligned}$$

Consider the second term.

$$\begin{aligned} &c_{11}[n^{-1}\sum_{i=1}^n\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 \\ &= c_{11}[\eta_1^T(n^{-1}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T - \mathbf{I})\eta_2 + \eta_1^T\eta_2]^2 \\ &= c_{11}[\eta_1^TA\eta_2 + \eta_1^T\eta_2]^2 \\ &\leq 2c_{11}[(c^TA^h n^{-1/2}\sum_{i=1}^n\mathbf{x}_i\mathbf{e}_i)^2 + (c^TA^{h-1}n^{-1/2}\sum_{i=1}^n\mathbf{x}_i\mathbf{e}_i)^2] \\ &= O_P(p/n) \end{aligned}$$

The last step follows from Lemma 2 of Mammen (1993) where it has been established that $(c^TA^h n^{-1/2}\sum_{i=1}^n\mathbf{x}_i\mathbf{e}_i)^2 = O_P(p)$, for $h = 1, \dots, K-1$, and by using $c_{11} = O(n^{-1})$. We now show $n^{-2}\sum_{i=1}^n[\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 = o_P(1)$. For this term, we have $n^{-2}\sum_{i=1}^n[\eta_1^T\mathbf{x}_i\mathbf{x}_i^T\eta_2]^2 = n^{-2}\sum_{i=1}^n[\eta_1^T\mathbf{x}_i]^2[\mathbf{x}_i^T\eta_2]^2$. We will show the following

$$\begin{aligned} (i) \quad &E[\eta_1^T\mathbf{x}_i]^2 = O(1) \\ (ii) \quad &E[\eta_2^T\mathbf{x}_i]^2 = O(p) \end{aligned}$$

and the proof will be complete. Consider (i) first. This is true if $a = 0$, so we look at $a \geq 1$ case. Then

$$\begin{aligned} E[c^TA^a\mathbf{x}_i]^2 &= n^{-2}E[c^T(\mathbf{x}_i\mathbf{x}_i^T + A_i)A^{a-1}\mathbf{x}_i]^2 \\ &\leq 2n^{-2}E[c^T\mathbf{x}_i\mathbf{x}_i^TA^{a-1}\mathbf{x}_i]^2 + 2n^{-2}E[c^TA_iA^{a-1}\mathbf{x}_i]^2 \end{aligned}$$

For the first term,

$$\begin{aligned} n^{-2}E[c^T\mathbf{x}_i\mathbf{x}_i^TA^{a-1}\mathbf{x}_i]^2 &= n^{-2}[E(c^T\mathbf{x}_i)^4]^{1/2}[E(\mathbf{x}_i^TA^{a-1}\mathbf{x}_i)^4]^{1/2} \\ &= O_P(p^2n^{-2}\lambda_{amax}A^{2(a-1)}) \\ &= o_P(1) \end{aligned}$$

The term $E[c^T \frac{A_i}{n} A^{a-1} \mathbf{x}_i]^2$ is treated by breaking up A again if $a-1 \geq 1$, and using the independence of $\frac{A_i}{n}$ and $\mathbf{x}_i$. This proves (i). For (ii), write $b = h - a - 1$ and since the relation easily holds for $b = 0$, we look at $b \geq 1$. Note that

$$[\mathbf{x}_i^T A^b n^{-1/2} \sum_{j=1}^n \mathbf{x}_j e_j]^2 \leq 2[\mathbf{x}_i^T A^b n^{-1/2} \mathbf{x}_i e_i]^2 + 2[\mathbf{x}_i^T A^b n^{-1/2} \sum_{j \neq i} \mathbf{x}_j e_j]^2$$

The first term is easily seen to be $O_P(p^2 n^{-1}) = o_P(p)$. We use the notation $\eta_i = n^{-1/2} \sum_{j \neq i} \mathbf{x}_j e_j$ for the remaining part of the proof. Then the second term of the above expression is $E[\mathbf{x}_i^T A^b \eta_i]^2$. Using independence of $\mathbf{x}_i$ and η_i this is seen to be $O(p)$ if $b = 0$, so take $b \geq 1$. Then

$$E[\mathbf{x}_i^T A^b \eta_i]^2 \leq 2n^{-2} E[\mathbf{x}_i^T A^{b-1} \mathbf{x}_i]^2 [\mathbf{x}_i^T \eta_i]^2 + 2E[\mathbf{x}_i^T A^{b-1} \frac{A_i}{n} \eta_i]^2$$

Using a Cauchy-Schwartz inequality and independence of $\mathbf{x}_i$ and η_i the first term is seen to be $o(p)$ and the second term is treated by techniques similar to ones used earlier. This completes the proof. ■

PROOF OF LEMMA 5.3.2 Note that the bootstrap and the original estimators are given by $\hat{\beta}_B = (\sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T})^{-1} \sum_{i=1}^n \mathbf{x}_i^* y_i^*$ and $\hat{\beta} = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$. Since $E_B \lambda_{amaz}(A^*) = o_P(1)$, we can write

$$(\sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T})^{-1} = n^{-1}(\mathbf{I} - A^*)^{-1} = n^{-1}[\mathbf{I} + A^* + \dots] \text{ and hence}$$

$$\begin{aligned} \hat{\beta}_B &= (\sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T})^{-1} \sum_{i=1}^n \mathbf{x}_i^* y_i^* = \beta + (\sum_{i=1}^n \mathbf{x}_i^* \mathbf{x}_i^{*T})^{-1} \sum_{i=1}^n \mathbf{x}_i^* e_i^* \\ &= \beta + n^{-1} \sum_{h \geq 0} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* \end{aligned}$$

$$\begin{aligned} \text{Similarly, } \hat{\beta} &= (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i = \beta + (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^n \mathbf{x}_i e_i \\ &= \beta + n^{-1} \sum_{h \geq 0} A^h \sum_{i=1}^n \mathbf{x}_i e_i \end{aligned}$$

Thus

$$\begin{aligned} & n^{1/2} \sigma_n^{-1} (\hat{\beta}_B - \hat{\beta}) \\ &= n^{-1/2} \sum_{h \geq 0} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* - n^{-1/2} \sum_{h \geq 0} A^h \sum_{i=1}^n \mathbf{x}_i e_i \\ &= n^{-1/2} \sum_{i=1}^n W_i c^T \mathbf{x}_i e_i + n^{-1/2} \sum_{h=1}^{K-1} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* - n^{-1/2} \sum_{h=1}^{K-1} A^h \sum_{i=1}^n \mathbf{x}_i e_i \\ &\quad + n^{-1/2} \sum_{h \geq K} A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* - n^{-1/2} \sum_{h \geq K} A^h \sum_{i=1}^n \mathbf{x}_i e_i \end{aligned}$$

From Lemma 5.3.4, it follows that

$$n^{-1/2} \sum_{h=1}^{K-1} c^T A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* - n^{-1/2} \sum_{h=1}^{K-1} c^T A^h \sum_{i=1}^n \mathbf{x}_i e_i = r_{1nB}$$

where $P_B[|r_{1nB}| > \epsilon] = o_P(1)$ for every $\epsilon > 0$. Note that in Lemma 3 of Mammen (1993), it has been shown that $n^{-1/2} \sum_{h \geq K} c^T A^h \sum_{i=1}^n \mathbf{x}_i e_i = o_P(1)$. Now let $n^{-1/2} \sum_{h \geq K} c^T A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^* = r_{2nB}$, then we now show $P_B[|r_{2nB}| > \epsilon] = o_P(1)$ to complete the proof. We use the technique of Lemma 3 of Mammen (1993), thus for almost all data sequences

$$|n^{-1/2} \sum_{h \geq K} c^T A^{*h} \sum_{i=1}^n \mathbf{x}_i^* e_i^*| \leq \lambda_{\max}(A^{*K}) \frac{1}{1 - \lambda_{\max}(A^*)} \|n^{-1/2} \sum_{i=1}^n \mathbf{x}_i^* e_i^*\|$$

$$\begin{aligned} \text{So } P_B[|r_{2nB}| > \epsilon] &\leq \epsilon^{-2} E_B r_{2nB}^2 \\ &\leq \epsilon^{-2} [E_B \lambda_{\max}(A^{*2K})] [E_B \|n^{-1/2} \sum_{i=1}^n \mathbf{x}_i^* e_i^*\|^2] \end{aligned} \quad (5.3.21)$$

$$\leq \epsilon^{-2} [E_B \lambda_{\max}(A^{*4})] [E_B \|n^{-1/2} \sum_{i=1}^n \mathbf{x}_i^* e_i^*\|^2] \quad \text{if } \delta < 2 \quad (5.3.22)$$

At (5.3.21) we can separate the cases where $\delta \geq 2$ and $\delta < 2$. Notice that when we have $E_B \lambda_{\max}(A^*) = O_P(p^{1/2+\delta/4} n^{-1/2})$, then (5.3.21) is $o_P(1)$, the calculations being identical with that of Lemma 3 of Mammen (1993). If $\delta \geq 2$, then $K = 1$ and from Lemma 5.3.1, we have $E_B \lambda_{\max}(A^*) = O_P(p^{1/2+\delta/4} n^{-1/2})$. Hence we need to look at $\delta < 2$ case under the condition $E_B \lambda_{\max}(A^{*4}) = O_P(\sigma_n^4 p^3 n^{-2})$. For this case, we use (5.3.22) and get $P_B[|r_{2nB}| > \epsilon] = O_P(\sigma_n^4 n^{-2} p^4)$. The last relation is $o_P(1)$ under any one of the three sets of possible conditions on σ_n^2 and δ given in the theorem. This completes the proof. $\blacksquare$

Chapter 6

ESTIMATION OF DISTRIBUTION OF LEAST SQUARES ESTIMATOR IN LINEAR REGRESSIONS: SECOND ORDER ACCURACY RESULTS

6.1 Introduction

In Chapter 5 we discussed the consistency of the *UBS* estimator for the distribution of the least squares parameter estimator in linear regression. In early pathbreaking works, Singh (1981) and Bickel and Freedman (1981) have established that the bootstrap approximation of distribution is often better than more traditional, asymptotic normality based approximations for functions of sample means of *i.i.d.* variables. It turns out that the bootstrap histogram often captures the second order term in the Edgeworth expansion of the distribution of parameter estimator, which the normal approximation may not. Following their works, a major research interest in bootstrap theory has been to understand the second order behaviour of the bootstrap distribution in a variety of problems.

Consider the linear regression model

$$y_i = \mathbf{x}_i^T \beta + e_i \quad (6.1.1)$$

where the explanatory variables $\mathbf{x}_i$'s are vectors in $\mathbb{R}^p$, and the e_i 's are independent noise terms. The parameter β is an unknown constant in $\mathbb{R}^p$ which is estimated from the data. A common estimating technique is to use least squares, which yields the estimator $\hat{\beta}_n = (\sum \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum \mathbf{x}_i y_i$. Consider the problem of estimating the distribution of $\hat{\beta}_n$. When $\{e_i\}$ are *i.i.d.*, it is known that residual based weighted bootstrap techniques are second order accurate (Zheng and Tu (1988)). Other related results are by Navidi (1989), Qumsiyeh (1990), Lahiri (1992). These results are based on Edgeworth expansions for the distribution of the statistic and its bootstrap version.

For all the papers mentioned above, it was assumed that the design vectors $\mathbf{x}_i$'s are non-random, and that the errors are *i.i.d.* with mean zero and finite variance. However, a number of important and useful models not represented in this framework are : (a) the errors are not necessarily *i.i.d.*, in particular, they are heteroscedastic; (b) the regressors are random; (c)

the resampling scheme uses the data pairs $(\mathbf{x}_i, y_i)$ instead of residuals, as in paired bootstrap or the jackknives.

It is generally understood that methods like the paired bootstrap or delete- d jackknives would not have higher order accuracy (Hall (1992), Mammen (1993)). In data pair resampling methods such as the paired bootstrap, the design changes with every bootstrap sample and consequently there is a bias in the bootstrap estimator that shows up in second order calculations. This is partially responsible for the lack of second order accuracy. A natural estimator to consider under these circumstances would be a bias corrected version of the original bootstrap estimator.

In this chapter, we consider the problem of estimating the distribution of the studentised least squares estimator in linear regression in random and fixed regressor models. Note that since the errors are heteroscedastic, appropriate studentisation of $\hat{\beta}_n$ is an important issue, which we discuss in details later on. For the model with non-random regressors, the errors are allowed to be heteroscedastic. For the random regressor case, we restrict our attention to the simple linear regression ($p = 1$) case. Extensions to more general ($p \geq 1$) models is feasible, but the algebra is messy. The technical conditions on the models are given later.

Our major result here is to establish the second order accuracy of the distribution of a bias corrected *UBS* estimator for studentised $\hat{\beta}_n$. We restrict our attention only to a sub-class of *UBS* schemes derived from *i.i.d.* weights for which the Edgeworth expansion is derivable. It is plausible that with appropriate conditions on the bootstrap weights, some other bias adjusted *UBS* procedures would also be second order accurate. We also establish the second order accuracy of the external bootstrap distribution estimator in the present problem. The results are based on Edgeworth expansions for studentised sums of independent (but not necessarily identically distributed) random variables, which may be of independent interest. It must be mentioned that the conditions imposed in this chapter on the model and the *UBS* weights are quite strong. They can perhaps be relaxed by using the techniques of Babu and Bai (1992, 1993, 1996) or Babu and Singh (1983, 1984). However, we have not pursued this aspect.

This chapter is adapted from Chatterjee and Bose (1999c).

6.2 Notations and assumptions

We first state a general condition that is used in all the results of this chapter.

Let $\{X_n\}$ be a sequence of independent, absolutely continuous random variables defined on the real line. Let the density of X_n be $f_n(\cdot)$. Define the function $\gamma_n(x) = \min\{f_n(x), f_n(-x)\}$, and let $\Gamma_n(x) = n^{-1} \sum_{i=1}^n \gamma_i(x)$. We need an assumption on the behaviour of $\Gamma_n(x)$. This condition will be used to derive Edgeworth expansions of studentised sample sums of $\{X_i\}$. A Cramer-type condition follows from the assumption below.

Assumption A Assume for all large n , $\exists$ a real valued function $g(x) \geq 0$ such that $\mathcal{A} = \{x : \Gamma_n(x) \geq g(x)\}$ has positive measure, and $0 < \int_{\mathcal{A}} g(x) dx < \infty$.

We now introduce some notations. For symmetric matrices A and B , $A > B$ means that $A - B$ is positive definite. For a vector c , $\|c\|$ denotes its Euclidean norm. Let $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{e} = (e_1, \dots, e_n)^T$, $\mathbf{X} = (\mathbf{x}_1 : \dots : \mathbf{x}_n)^T$. The projection operator on the column space of $\mathbf{X}$ is $\mathbf{P}_x = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. We also let $\kappa = \max_{i=1, \dots, n} \{(\mathbf{P}_x)_{ii}\}$ where $(\mathbf{P}_x)_{ii}$ denotes the i^{th} diagonal element of $\mathbf{P}_x$. Identity matrices will be denoted by $\mathbf{I}$, their order being apparent from the contexts. Let $\tau_i^2 = E e_i^2$ and $D_n = n^{-1} \sum \mathbf{x}_i \mathbf{x}_i^T$. The notations k, K , with or without suffix, are generic for constants.

For the model with non-random regressors, the conditions are

$$D_n > k_1 \mathbf{I} \text{ for some } n \quad (6.2.1)$$

$$\|\mathbf{x}_i\| < K_2 \text{ for all } i \quad (6.2.2)$$

$$\kappa = O(n^{-1}) \quad (6.2.3)$$

$$E e_i = 0 \text{ for all } i \quad (6.2.4)$$

$$\tau_i^2 > k_3 > 0 \text{ for all } i \quad (6.2.5)$$

$$E|e_i|^L < K_4 < \infty \text{ for all } i \quad (6.2.6)$$

Let a_n be the number of non-zero x_i 's. Two important consequences of the above assumptions are

$$\begin{aligned} \liminf_n n^{-1} \sum x_i^2 \tau_i^2 &> C_0 > 0 \\ a_n/n &> k_1/K_2 \end{aligned}$$

Since the terms with $x_i = 0$ do not appear in T_n , we may assume $a_n = n$ without loss of generality.

For the random regressors model, the conditions are

$$E x_1^{16} < \infty \quad (6.2.7)$$

$$E(x_1)^2 > k_5 > 0 \quad (6.2.8)$$

$$E(x_1 e_1)^2 > k_6 > 0 \quad (6.2.9)$$

$$E x_1^3 e_1 = 0 \quad (6.2.10)$$

$$E(x_1 e_1)^{12} < \infty \quad (6.2.11)$$

Assume, without loss of generality, that the x_i values are all non-zero.

For both the models, assume that the random variables $\{x_i e_i\}$ are absolutely continuous satisfying Assumption A.

Thus precisely, the two models considered are:

Model A : (non-random regressors) $\{y_i\}$'s satisfy (6.1.1) where the conditions (6.2.1)-(6.2.6) hold for some L , and $\{x_i e_i\}$ are absolutely continuous satisfying Assumption A.

Model B : (random regressors) The pairs (x_i, e_i) are *i.i.d.*, and $\{y_i\}$'s satisfy (6.1.1) where (6.2.7)-(6.2.11) hold, and $\{x_i e_i\}$ are absolutely continuous satisfying Assumption A.

In either model consider the least squares estimator $\hat{\beta}_n = D_n^{-1} n^{-1} \sum x_i y_i$. From our assumptions on either model, it follows that this matrix is nonsingular. For any p -dimensional fixed vector c with $\|c\| = 1$, let $T_{n0} = n^{1/2} c^T D_n (\hat{\beta}_n - \beta)$. Using the notation $x_i = c^T x_i$, $T_{n0} = n^{-1/2} \sum x_i e_i$. Let $\{r_i = y_i - x_i^T \hat{\beta}_n\}$ be the usual residuals. Define $S_n^2 = n^{-1} \sum x_i^2 r_i^2$, which is an estimate of $V(T_{n0}) = n^{-1} \sum x_i^2 \tau_i^2$. Thus $T_n = S_n^{-1} T_{n0}$, is a studentised version of the centered $\hat{\beta}_n$. Let $F_n(x) = \text{Prob}[T_n \leq x]$ be the distribution function of T_n . Our goal is to make second order accurate approximation of F_n . The resampling schemes we discuss are the uncorrelated weights bootstrap (UBS) and the external bootstrap of Wu (1986). The following additional model condition is required for bootstrap.

For any constant $C > 1$, let $p_i(C) = \text{Prob}[x_i^2 e_i^2 \in [C^{-1}, C]]$.

Assumption B For some $c \in (0, 1)$, there exists a constant $C > 1$ such that $n^{-1} \sum p_i(C) \geq c$ for all large n .

THE EXTERNAL BOOTSTRAP: (Wu (1986)) Consider a triangular sequence of random variables, whose n^{th} row consists of a set of *i.i.d.* random variables $U_{1:n}, \dots, U_{n:n}$, where $\{U_{i:n} : i \geq 1\}$ is independent from $\{(x_i, e_i) : i \geq 1\}$. For simplicity we henceforth use the notation U_i for $U_{i:n}$. Define $y_i^* = x_i^T \hat{\beta}_n + U_i r_i$ for $i = 1, \dots, n$. The external bootstrap estimator is given by $\hat{\beta}_{En} = D_n^{-1} n^{-1} \sum x_i y_i^*$. The bootstrap equivalent of T_{n0} is $T_{En0} = n^{1/2} c^T D_n (\hat{\beta}_{En} - \hat{\beta}_n) = n^{-1/2} \sum U_i x_i r_i$. Thus the studentised bootstrap statistic is $T_{En} = (n^{-1} \sum U_i^2 x_i^2 r_i^2)^{-1/2} T_{En0}$ and let F_{En} be its (conditional) distribution function given $\{r_i\}$ and $\{x_i\}$.

The following moment conditions are assumed on $U_{i:n}$:

$$EU_{1:n} = 0 \quad (6.2.12)$$

$$EU_{1:n}^2 = 1 \quad (6.2.13)$$

$$EU_{1:n}^3 \rightarrow 0 \text{ as } n \rightarrow \infty \quad (6.2.14)$$

$$\sup_n EU_{1:n}^6 < \infty \quad (6.2.15)$$

THE UNCORRELATED WEIGHTS BOOTSTRAP (UBS): Consider a triangular sequence of random variables, whose n^{th} row consists of a set of *i.i.d.* random variables $w_{1:n}, \dots, w_{n:n}$, where $\{w_{i:n} : i \geq 1\}$ is independent of $\{(x_i, e_i) : i \geq 1\}$. For simplicity we henceforth use w_i for $w_{i:n}$. Consider the transformed data $(x_i^*, y_i^*) = w_i^{1/2} (x_i, y_i)$ as the bootstrap data. Use ordinary least squares on this transformed data to get the UBS estimator $\hat{\beta}_{Bn} = D_{Bn}^{-1} n^{-1} \sum w_i x_i y_i$,

where $D_{Bn} = n^{-1} \sum w_i \mathbf{x}_i \mathbf{x}_i^T$. The matrix D_{Bn} is always positive definite whenever (6.2.1) and (6.2.16) below holds. The condition (6.2.16) is in general not necessary for implementing *UBS* resampling, and is indeed not satisfied by many schemes like the paired bootstrap or the delete- d jackknives. However, since we assume that w_i 's are *i.i.d.* such schemes are anyway not being considered here, and (6.2.16) is a minor additional restriction.

Since the design variables change with every bootstrap sample, it naturally implies that conditional on the data, the bootstrap expectation of $\hat{\beta}_{Bn}$ is not $\hat{\beta}_n$. The bias factor does not affect first order consistency results (see Chapter 5); but it is not negligible in the second order term. Let $E(w_{1:n} - 1)^2 = \sigma_n^2$ and $W_i = \sigma_n^{-1}(w_i - 1)$, and the matrices $\mathbf{W}_0$ and $\mathbf{W}$ denote the diagonal matrices whose i^{th} diagonal entries are respectively w_i and W_i . Algebra shows that the difference $\hat{\beta}_n - E_B \hat{\beta}_{Bn} = \sigma_n E_B(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_0 \mathbf{e}$, and the quantity inside the expectation may be estimated by $B_{Bn} = \sigma_n (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_0 \mathbf{r}$, where the vector of residuals is denoted by $\mathbf{r} = (\mathbf{I} - \mathbf{P}_x) \mathbf{e}$. The bias corrected *UBS* statistic that is equivalent to T_{n0} is $T_{Bn0} = n^{1/2} c^T D_n \sigma_n^{-1} (\hat{\beta}_{Bn} - \hat{\beta}_n + B_{Bn}) = n^{-1/2} \sum W_i x_i r_i$. Thus the studentised bootstrap statistic is $T_{Bn} = (n^{-1} \sum W_i^2 x_i^2 r_i^2)^{-1/2} T_{Bn0}$ and let F_{Bn} be its (conditional) distribution function given $\{r_i\}$.

Note that T_{Bn0} has the same form as T_{En0} with W_i 's in place of U_i 's. This means that the results for the two generalised bootstrap resampling are similar.

The following conditions are assumed on the *UBS* weights:

$$0 < k < w_{1:n} \text{ with probability } 1 \quad (6.2.16)$$

$$E w_{1:n} = 1 \quad (6.2.17)$$

$$0 < E(w_{1:n} - 1)^2 = \sigma_n^2 < \infty \quad (6.2.18)$$

$$E W_{1:n}^3 \rightarrow 1 \text{ as } n \rightarrow \infty \quad (6.2.19)$$

$$\sup_n E W_{1:n}^6 < \infty \quad (6.2.20)$$

In addition to the conditions stated above, the following condition on the density of U_i (in case of external bootstrap) or W_i (in case of *UBS*) is always assumed to be satisfied:

Assumption C The distribution of U_1 (W_1) has a density $f(x)$ such that for some positive numbers x_0 and b , we have $f(x) > b$ for all $x \in (-x_0, x_0)$.

Recall that in both the resampling schemes we use *i.i.d.* absolutely continuous weights.

6.3 Main results

Note that T_n is a studentised sample mean from independent random variables which are not necessarily identically distributed. Thus in order to approximate F_n we first establish

an Edgeworth expansion result for studentised statistic from independent but not necessarily identically distributed random variables. This result does not seem to exist in the literature in the form in which we need it.

We now state some moment conditions on $\{X_i\}$. Some of these may be relaxed as in Hall (1987).

$$EX_i = 0 \text{ for all } i \quad (6.3.1)$$

$$EX_i^2 = \tau_i^2 < K_1 < \infty \quad (6.3.2)$$

$$\sum_{i \in \mathcal{R}} EX_i^4 = O(n) \quad (6.3.3)$$

$$\sum_{i \in \mathcal{R}} EX_i^6 = o(n^2) \quad (6.3.4)$$

Under the above set of assumptions, we consider Edgeworth expansion for the statistic $t_n = (\sum X_i^2)^{-1/2} \sum X_i$. The standard notations $\Phi(\cdot)$ and $\phi(\cdot)$ are used for the standard normal distribution function and density function respectively.

Theorem 6.3.1 *For random variables $\{X_n\}$ satisfying Assumption A and (6.3.1)-(6.3.4),*

$$\sup_{y \in \mathcal{R}} |P[(\sum X_i^2)^{-1/2} \sum X_i \leq y] - \Phi(y) - n^{-1/2} p(y) \phi(y)| = o(n^{-1/2}) \quad (6.3.5)$$

for some polynomial $p(y)$ of degree 3 whose coefficients are continuous functions of the averages of the first three moments of X_i , $i = 1, \dots, n$.

REMARK 1. Recall that Hall (1987) established Edgeworth expansion for the studentised sample mean of *i.i.d.* random variables with finite third moments and non-singular distributions. The non-singularity of the distribution was used to establish a Cramer condition type result (see Lemma 2.2 of Hall (1987)), using a result on the density of the absolutely continuous part of the random variable (see Lemma 2.1 of Hall (1987)). For our set-up, we have assumed for convenience that the random variables are all absolutely continuous. The proof of Theorem 6.3.1 is along the same lines as that of Hall (1987), and is sketched in Section 6.4.

REMARK 2. The above theorem a partial generalisation of the main results of three important papers. Bai and Zhao (1986) considered nonstudentised sample averages from independent random variables, our result is for the studentised case. Hall (1987) established Edgeworth expansion for studentised sample average of *i.i.d.* random variables under minimal conditions, we extend it to the independent case, but do not bother about minimal conditions. By using standard techniques, our result can be extended to functions of sample averages of independent random variables, thereby extending the result of Bai and Rao (1991). Our result may also be viewed as a variant of Theorem 20.6 of Bhattacharya and Ranga Rao (1976), where individual random variables are required to satisfy the Cramer condition. We derive a 'conditional Cramer condition', as in Lemma 2.2. of Hall (1987) in the present set-up, and for

this we need assumption A. However, by following a different approach based on Edgeworth expansions for strongly non-lattice random variables, assumptions about absolute continuity of X_n 's can possibly be replaced by strong non-latticeness of X_i 's.

For weighted sums of i.i.d. random variables, a sufficient condition for ensuring assumption A is given below. The proof is easy and is omitted.

Proposition 6.3.1 *Suppose $\{\xi_i\}$ are i.i.d. mean zero absolutely continuous random variables with density $f(x)$ such that for some constants $x_0 > 0$ and $b > 0$, $f(x) > b > 0$ for all $x \in (-x_0, x_0)$. Let $\{a_i\}$ be a sequence of non-random constants such that for large n , and for some $\alpha > 0$; at least $[an]$ of $\{|a_i|, i = 1, \dots, n\}$ lie in $[a^{-1}, a]$ for some $a > 1$. Then $\{X_i = a_i \xi_i\}$ satisfies Assumption A.*

We now apply Theorem 6.3.1 to establish Edgeworth expansion of F_n , under Models A and B. Notice that T_n is not of the same form as t_n , which is the statistic of Theorem 6.3.1, since the errors e_i have been estimated by the residuals r_i while studentising. The various moment conditions in Models A and B ensure that using the residuals instead of the actual unobservable errors does not alter the second order behaviour (see Lemma 6.4.3). This is an important step in the studentisation of $\hat{\beta}_n$.

Theorem 6.3.2 (A) (Edgeworth expansion for T_n in Model A) *Assume the conditions of Model A with $L = 8$. Then*

$$\sup_x |F_n(x) - \Phi(x) - n^{-1/2} p(x) \phi(x)| = o(n^{-1/2})$$

for some third degree polynomial $p(x)$ whose coefficients are continuous functions of averages of moments of $x_i e_i$, $i = 1, \dots, n$ upto the third order.

(B) (Edgeworth expansion for T_n in Model B) *Assume the conditions of Model B. Then the conclusion given in (A) above holds.*

The following theorems state the second order accuracy of the external bootstrap and UBS for F_n in Model A and B.

Theorem 6.3.3 (a) *Assume the conditions of Model A with $L = 12$ and assumption B. Then*

$$\sup_x |F_{En}(x) - F_n(x)| = o(n^{-1/2}) \text{ almost surely,}$$

(b) *Assume the conditions of Model B. Then the same conclusion holds as in (a).*

Theorem 6.3.4 (a) *Assume the conditions of Model A with $L = 12$ and assumption B. Then*

$$\sup_x |F_{Bn}(x) - F_n(x)| = o(n^{-1/2}) \text{ almost surely,}$$

(b) *Assume the conditions of Model B. Then the same conclusion holds as in (a).*

REMARK 1: The moment conditions for either the external bootstrap or *UBS* are satisfied in many situations. For example, take V to follow $\text{beta}(4, \gamma_n)$, and let $EV = \mu_1$ and $E(V - \mu_1)^2 = \mu_2$. Now define $U_{1:n} = \mu_2^{-1/2}(V - \mu_1)$. As long as $\gamma_n \rightarrow \infty$ the moment conditions (6.2.12)-(6.2.15) are satisfied. This is precisely the property that Zheng and Tu (1988) found useful for the *Dirichlet*(4, ..., 4) weights. If V follows *Gamma*(4, γ) distribution, and U_1 is defined as centered and standardised V then again (6.2.12)-(6.2.15) holds. For $\gamma < 1/4$, the same weights can also be used in *UBS*, by a location shift. A mixture of Uniform distributions can also be used to get random variables with required moment conditions, and such distributions are typically easy to sample from.

REMARK 2: In its full generality, the *UBS* technique includes as special cases, the paired bootstrap, delete- d jackknives, Bayesian bootstrap, weighted likelihood bootstrap and the m out of n bootstrap. But since in this chapter we consider the *UBS* weights to be *i.i.d.* and absolutely continuous, the jackknives and paired bootstrap are not covered here. However, it is interesting to note that condition (6.2.19) is violated for jackknife weights, so a second order accuracy of the jackknife histogram is typically not to be expected. On the other hand, for the paired bootstrap the weights used follow *Multinomial*($n, 1/n, \dots, 1/n$) distribution, for which all the moment conditions are satisfied, so it is possible that the bias corrected paired bootstrap estimator is second order accurate.

6.4 Proofs

SKETCH OF PROOF OF THEOREM 6.3.1 Theorem 6.3.1 is proved by closely imitating the arguments of Hall (1987). Hence we give only a sketch of the proof. In Hall (1987) the X_j 's are *i.i.d.* but in our case it is not so. Let $Y_j = X_j - E(X_j|X_j)$, $p_j = P[X_j > 0|X_j]$, $\beta_{2j} = E[Y_j^2|X_j]$ and $\psi_j(t) = E[\exp(itY_j)|X_j]$, where all the above expectations are conditional on $|X_j|$, as is obvious from the notation. The arguments of Hall (1987) may be followed, by substituting appropriate averages of moments (or conditional moments) in place of the moments (or conditional moments) from the single distribution considered by Hall (1987). Note that the remarks of Hall (1987, page 922, after (2.4)) do not hold in case X_i 's are not *i.i.d.*. However, Assumption A allows us to overcome these issues. With a notation similar to that of Hall (1987), we may then define $\mu_2 = n^{-1} \sum EX_i^2$ and $\nu_2 = n^{-1} \sum E\beta_{2j}$ among other things, and follow his arguments. One additional technical condition that is required for the present set-up is that both μ_2 and ν_2 must have a positive lower bound. This follows from Assumption A. We also need a parallel to Lemma 2.2 of Hall (1987), which is stated below. We omit its proof, which is easy.

Lemma 6.4.1 *If X_i 's are absolutely continuous with densities satisfying assumption A, then*

$$\sup_{|t|>\epsilon} n^{-1} \sum E|\psi_j(t)| < 1$$

for all $\epsilon > 0$ for all large n .

Then, by following arguments similar to that of Hall (1987, pages 926-930), we obtain

$$\sup_{y \in \mathbb{R}} |P[t_n \leq y] - \Phi(y) - n^{-1/2} q(y) \phi(y)| = o(n^{-1/2})$$

for some third degree polynomial $q(y)$ whose coefficients depend on moments upto third order of X_i 's and $(2p_i - 1)|X_i|$'s. From Theorem 20.6 of Bhattacharya and Ranga Rao (1976) it follows that $q(y)$ must be as stated in (6.3.5), with coefficients depending only on the averages of moments upto third order of the random vectors $(X_i, n^{-1/2}(X_i^2 - \tau_i^2))^T$, $i = 1, \dots, n$, since if two second order Edgeworth expansions exist for the same quantity they may differ only by $o(n^{-1/2})$. However, now we can use the special structure of our statistic to see that the coefficients of $q(y)$ must be functions of $n^{-1} \sum EX_i^2$ and $n^{-1} \sum EX_i^3$, since $EX_i = 0$ for all i . An easy cross check is accomplished by observing that this is indeed the case when X_i 's are *i.i.d* random variables. This completes the sketch of the proof of Theorem 6.3.1. ■

We first introduce some convenient notation before going into the proof of Theorem 6.3.2. Define $S_{n1}^2 = n^{-1} \sum x_i^2 e_i^2$ and $T_{n1} = S_{n1}^{-1} T_{n0}$. Let the distribution function of T_{n1} be F_{n1} . We define $\epsilon_n = n^{-1/2}(\log \log n)^{-1}$.

Lemma 6.4.2 For random variables R_n and C_n ,

$$\text{If } P[|R_n| > \epsilon_n] = O(a_n) \text{ and } \sup_x \sup_{|y_n| \leq \epsilon_n} |P[C_n \leq x] - P[C_n \leq x - y]| = O(b_n)$$

$$\text{Then } \sup_x |P[C_n + R_n \leq x] - P[C_n \leq x]| = O(a_n + b_n)$$

The proof of this lemma is routine and is omitted.

Lemma 6.4.3 Assume the model conditions (6.2.1) - (6.2.6) with $L = 8$. Then we have that $T_n = T_{n1} + R_{n1}$ where $P[|R_{n1}| > \epsilon_n] = o(n^{-1/2})$.

PROOF OF LEMMA 6.4.3 Let $h_{ij} = \mathbf{x}_i^T D_n^{-1} \mathbf{x}_j$, $\delta_i = n^{-1} \sum_{j=1}^n h_{ij} e_j$, then $r_i = e_i + \delta_i$. Then we have $n^{-1} \sum x_i^2 r_i^2 = n^{-1} \sum x_i^2 e_i^2 + \delta$, say. With a lot of algebra, one can show that $E\delta^4 = O(n^{-4})$ using the assumed moment conditions. We now have $T_{n1} = T_{n2} + R_{n1}$ where $R_{n1} = S_{n1}^{-1} S_{n2}^{-1} (S_{n1} + S_{n2})^{-1} T_{n0} \delta$. Let us define the sets

$$\begin{aligned} \mathcal{A}_{n0} &= \{|\delta| < 4^{-1} C_0\} \\ \mathcal{A}_{n1} &= \{n^{-1} \sum x_i^2 r_i^2 > 2^{-1} n^{-1} \sum x_i^2 e_i^2\} \\ \mathcal{A}_{n2} &= \{n^{-1} \sum x_i^2 e_i^2 > 2^{-1} n^{-1} \sum x_i^2 r_i^2\} \end{aligned}$$

Using the moment bound on δ , we have $1 - P[\mathcal{A}_{n0}] = P[|\delta| > 4^{-1}C_0] = O(n^{-1})$. Using this fact, we now show that the probabilities of the sets $\mathcal{A}_{n1}$ and $\mathcal{A}_{n2}$ are $1 - O(n^{-1})$.

$$\begin{aligned}
1 - P[\mathcal{A}_{n2}] &= P[n^{-1} \sum x_i^2 e_i^2 < 2^{-1} n^{-1} \sum x_i^2 \tau_i^2] \\
&\leq P[|n^{-1} \sum x_i^2 (e_i^2 - \tau_i^2)| > 2^{-1} C_0] \\
&\leq 4C_0^{-2} E[n^{-1} \sum x_i^2 (e_i^2 - \tau_i^2)]^2 = O(n^{-1}) \text{ and} \\
1 - P[\mathcal{A}_{n1}] &= P[n^{-1} \sum x_i^2 r_i^2 < 2^{-1} n^{-1} \sum x_i^2 \tau_i^2] \\
&\leq P[|\delta| > 4^{-1} C_0] + P[n^{-1} |\sum x_i^2 (e_i^2 - \tau_i^2)| > 4^{-1} C_0] = O(n^{-1})
\end{aligned}$$

It now follows that the set $\mathcal{A} = \mathcal{A}_{n1} \cap \mathcal{A}_{n2}$ has a probability $1 - O(n^{-1})$. Thus we have

$$\begin{aligned}
P[R_{n1} > \epsilon_n] &= P[|S_n^{-1} S_{n1}^{-1} (S_n + S_{n1})^{-1} T_{n0} \delta| > \epsilon_n] \\
&= P[|S_n^{-1} S_{n1}^{-1} (S_n + S_{n1})^{-1} T_{n0} \delta| > \epsilon_n, \mathcal{A}] + O(n^{-1}) \\
&\leq P[|T_{n0} \delta| > k\epsilon_n] + O(n^{-1}) \\
&\leq k\epsilon_n^{-2} [ET_{n0}^4 E\delta^4]^{1/2} + O(n^{-1}) \\
&= o(n^{-1/2})
\end{aligned}$$

using the fact that $E\delta^4 = O(n^{-4})$. This completes the proof. $\blacksquare$

PROOF OF THEOREM 6.3.2 (a) From Lemma 6.4.3, we observe that T_n can be approximated by T_{n1} with an error R_{n1} for which $P[|R_{n1}| > \epsilon_n] = o(n^{-1/2})$. Now observe that the assumptions of Theorem 6.3.1 are satisfied by the random variables $\{x_i e_i\}$, so that F_{n1} has an Edgeworth expansion of the form (6.3.5). Hence the following is satisfied:

$$\sup_x \sup_{|y| \leq \epsilon_n} |F_{n1}(x) - F_{n1}(x - y)| = o(n^{-1/2})$$

Thus, the conditions of Lemma 6.4.2 are satisfied with $C_n = T_{n1}$ and $R_n = R_{n1}$, so that we have

$$\sup_x |F_n(x) - F_{n1}(x)| = o(n^{-1/2})$$

Now use Theorem 6.3.1 again, to get the result.

The proof of Theorem 6.3.2 (b) is similar. An equivalent of Lemma 6.4.3 is needed for this case. This may be proved by routine moment computations. $\blacksquare$

PROOF OF THEOREM 6.3.3 (a) Suppose we establish

$$\sup_x |F_{En}(x) - \Phi(x) - n^{-1/2} \hat{p}(x) \phi(x)| = o(n^{-1/2}) \quad (6.4.1)$$

where the polynomial $\hat{p}(x)$ is of the same form as that of $p(x)$, with the coefficients being functions of the averages of sample moments upto third order of $x_i e_i$, $i = 1, 2, \dots, n$. Note that

our moment assumptions imply that the difference between the sample moments and their expectations tend to zero almost surely. We use Theorem 6.3.1 for the Edgeworth expansion of F_{E_n} . The moment conditions of that theorem are satisfied easily, and only Assumption A needs to be verified. Consider the set $\mathcal{A}_{n3}$ on which at least $[cn/3]$ of the $x_i^2 r_i^2$ lie in $[(2C)^{-1}, 2C]$ where c and C are as in Assumption B. Clearly, over this set, using Proposition 6.3.1 Assumption A can be verified. One other part is to show that the complement of $\mathcal{A}_{n3}$ cannot occur infinitely often. This is proved as follows: For every n , because of Assumption B on the model, we have that with a probability $1 - O(n^{-2})$, at least $[cn/3]$ of the $x_i^2 e_i^2$ lie in $[C^{-1}, C]$, where c and C are as in assumption B. From this, it can be shown that the probability that at least $[cn/3]$ of the $x_i^2 r_i^2$ lie in $[(2C)^{-1}, 2C]$ is $1 - O(n^{-2})$. Now use Borel-Cantelli lemma to establish almost sure convergence.

The proofs of part (b) of this theorem and Theorem 6.3.4 are similar. ■

Chapter 7

ASYMPTOTICS AND GENERALISED BOOTSTRAP FOR SOLUTIONS OF MARTINGALE ESTIMATING EQUATIONS

7.1 Introduction

Stepping ahead from the case of least squares estimator in linear regression that we considered in the previous chapters we consider estimating equations in very general terms in this chapter. The goal of this chapter is to obtain results similar to the previous chapters, for estimators obtained by solving martingale equations

$$\sum_{i=1}^n \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) = 0 \quad (7.1.1)$$

where

- (i) The functions $\{\phi_{ni}(\cdot), i = 1, \dots, n, n \geq 1\}$ are known functions taking values in $\mathbb{R}^p$.
- (ii) β is a parameter taking values in $\mathcal{B} \subset \mathbb{R}^p$ for some integer p , and its unknown true value is denoted by β_0 . The values of p and β_0 may depend on n .
- (iii) $\{y_{ni}, i = 1, \dots, n, n \geq 1\}$ is an observed triangular sequence of random variables.
- (iv) $\{\mathbf{x}_{ni}, i = 1, \dots, n, n \geq 1\}$ is an observable sequence which may or may not be random, taking values in $\mathcal{X} \subset \mathbb{R}^k$ for $i = 1, \dots, n$. These stand for 'covariates' in a typical statistical problem.

We assume that at the true value β_0 of the parameter β , the functions ϕ_{ni} form a triangular array of martingale differences, and hence we call (7.1.1) martingale estimating equations. Typically $\{y_{ni}\}$ is modelled in terms of some function of the covariates $\{x_{ni}\}$ and parameter β_0 , along with an additive error term $\{e_{ni}\}$. Generally, $\{e_{ni}, i = 1, \dots, n, n \geq 1\}$ is an unobservable sequence of random variables. The joint distribution of $(e_{ni}, \mathbf{x}_{ni})$ may depend on β_0 . The function involving $\mathbf{x}_{ni}$ and β_0 represents the signal term, and e_{ni} is the noise.

The above framework contains as special cases the estimation problems of (i) M -estimators of location, scale, and regression; (ii) non-linear regression; (iii) generalised linear models; (iv)

certain kinds of semiparametric regression and density estimation problems; (v) auto-regression; (vi) co-integration; (vii) local estimating equations; (viii) likelihood based inference problems; and many other examples.

We define our resample estimator as the solution of

$$\sum_{i=1}^n w_{ni} \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) = 0 \quad (7.1.2)$$

where $\{w_{ni}, i = 1, \dots, n, n \geq 1\}$ is a triangular sequence of random variables, independent of $\{(e_{ni}, \mathbf{x}_{ni})\}$. These are the so called 'bootstrap weights'. We discuss the conditions on w_{ni} in Section 7.2. Sometimes estimators are obtained by the following minimisation problem:

$$\min_{\beta} \sum \Phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) \quad (7.1.3)$$

for some array of functions $\Phi_{ni}(\cdot)$. Define the *UBS* estimator in this case as the solution to the following problem:

$$\min_{\beta} \sum w_{ni} \Phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \beta) \quad (7.1.4)$$

In case Φ_{ni} 's are smooth, solutions to these can be obtained using (7.1.1) and (7.1.2) respectively. Thus minimisation problems with smooth criterion functions may also be treated as part of our set-up. The case of non-smooth Φ_{ni} is treated in Chapter 8 in the special framework of M_m estimators. Early ideas relating to bootstrapping with (7.1.2) or (7.1.4) may be found in Freedman and Peters (1984) and Rao and Zhao (1992).

The results of this chapter require diverse technical conditions. We put all these conditions together in Section 7.2. Our different theorems require different conditions both on the model as well as on the *UBS* weights.

In Section 7.3 we state the main results of this chapter. There are four major results. The asymptotic normality of the estimator; the first order consistency of the *UBS* distribution estimator; second order consistency of the distribution of studentised, bias corrected *UBS* statistic for the distribution of the studentised, bias corrected estimator and, asymptotic representation of *UBS* variance estimator.

First we consider the case of a real valued parameter for greater transparency. In subsection 7.3.1 we obtain asymptotic representation of the estimator as well as its bootstrap version. This is used to establish asymptotic normality of the estimator and consistency of the *UBS* for estimating the distribution. We aim at establishing our results under minimal conditions on the functions ϕ_{ni} . However, the conditions on ϕ_{ni} may be further relaxed, and this is indicated in a remark in that subsection.

Then in subsection 7.3.2 we consider the special case of independent ϕ_{ni} , $i = 1, \dots, n$. We establish an asymptotic representation of the *UBS* variance estimator, along the lines of Chapter 3 and Liu and Singh (1992). This asymptotic representation shows that the *UBS* remains

consistent under heteroscedasticity for a very broad class of estimators and models, including M -estimators in linear regression, generalised linear models and non-linear regressions. It is known from Bose and Kushary (1996) that the classification of common resampling schemes as *efficient* or *robust* along the lines of Liu and Singh (1992) does not hold generally for M -estimators in linear regression. Our result establishes that UBS estimators are always robust, sometimes they are also efficient.

The estimator obtained by solving estimating equations will be typically biased for the parameter. In subsection 7.3.3 we retain the assumption of independent ϕ_{ni} and consider a bias corrected, studentised estimator and show that a subclass of UBS yields a second order accurate approximation to its distribution. This higher order accuracy result may be used for obtaining higher order accurate confidence sets for the parameter or testing hypothesis relating to it. This result is an extension of Theorem 6.3.4 of Chapter 6 to the case of estimating equations.

In subsection 7.3.4 we consider the case of a general p dimensional parameter and allow p to tend to infinity with the data size n . Asymptotic normality and consistency of UBS results along the lines of subsection 7.3.1 are obtained. The UBS consistency result of this subsection is an extension of Theorem 5.2.1 of Chapter 5 to the case of martingale estimating equations.

In Section 7.4 we discuss several resampling schemes that are covered in our framework. This is a more detailed discussion on the resampling techniques which were briefly mentioned in Chapter 2. We discuss how the conditions on UBS weights are variously satisfied.

Examples are abundant in statistics where estimators are obtained by solving martingale equations. In Section 7.5 we consider several such examples. In subsection 7.5.1 we discuss several examples in details, including non-linear regression, M -estimation in linear models, generalised linear models and autoregressive processes. In subsection 7.5.2 we consider a slightly more general model framework, that of estimators obtained by solving 'approximate' martingale estimating equations. Here $\{\phi_{ni} = m_{ni} + r_{ni}\}$, where $\{m_{ni}\}$ is a martingale difference array, and $\{r_{ni}\}$ is an array of 'small' perturbation terms. Important examples of this set-up is that of estimating the parametric component in semiparametric problems, after using a data-driven plug-in for the non-parametric component. We discuss how our methodology extends to this slightly more general framework. Four important examples are discussed.

In Section 7.6 we discuss some computational aspects. Solving (7.1.1) or (7.1.2) would typically require iterative techniques and the bootstrap Monte Carlo further implies that (7.1.2) has to be solved a large number of times. Based on some heuristics, we suggest a one-step non-iterative method of solving (7.1.2), once (7.1.1) is solved. When applicable, this greatly reduces the computational burden. An example based on simulated data from a nonlinear regression problem is presented, which shows that the non-iterative technique produces a good UBS approximation in some problems at least. The proofs of all the theorems are given in Section 7.7. Some of the material of this chapter are adapted from Chatterjee and Bose (1999b).

7.2 Technical conditions

We first state the conditions on *UBS* weights, and then state the model conditions used for the different results of this chapter. We write $\phi_{ni}(e_{ni}, \mathbf{x}_{ni}, \lambda) = \phi_{ni}(\lambda)$, and assume that for all $1 \leq i \leq n$ and all $n \geq 1$, the functions $\phi_{ni}(\lambda)$ are differentiable twice almost everywhere. Two conventions are used for the conditions listed below: Any condition stated for a random function is assumed to hold almost surely unless otherwise stated; and the usage of the term ‘for all λ ’ signifies that the stated condition holds for all values λ in the interior of the parameter space, which is always assumed to be an open set. Also, throughout this chapter, k and K , with or without suffix, are used as generic for constants.

7.2.1 Conditions on *UBS* weights

Let $\{w_{i,n}; 1 \leq i \leq n, n \geq 1\}$ be a triangular array of non-negative random variables to be used as weights. We drop the suffix n from the notation of the weights. The notations P_B, E_B will respectively denote bootstrap probability and expectation, conditional on the data. Recall that $V(w_i) = \sigma_n^2$ and assume that the quantities

$$E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

are functions of the powers $i, j, k, \dots$ only, and not of the indices $a, b, c, \dots$. Thus we can write

$$c_{ijk\dots} = E\left(\frac{w_a - 1}{\sigma_n}\right)^i \left(\frac{w_b - 1}{\sigma_n}\right)^j \left(\frac{w_c - 1}{\sigma_n}\right)^k \dots$$

Note that if the weights are assumed to be exchangeable, then this is true.

Let $\mathcal{W}$ be the set on which at least m_0 of the weights are greater than some fixed constant $k_2 > 0$. The value of m_0 may be related to the model assumptions, and will be specified later. We now state some conditions on the weights. Various combinations of these conditions are used in proving our different results on *UBS* bootstrap throughout this chapter. The conditions are stated keeping in view that later we allow $p \rightarrow \infty$ in some results.

$$E w_i = 1 \tag{7.2.1}$$

$$0 < \sigma_n^2 = o(\min a_n^2 p^{-3}, n) \tag{7.2.2}$$

$$c_{11} = O(n^{-1}) \tag{7.2.3}$$

$$c_{22} \rightarrow 1 \tag{7.2.4}$$

$$c_4 < \infty \tag{7.2.5}$$

$$P_B[\mathcal{W}] = 1 - O(p^2 n^{-1}) \tag{7.2.6}$$

$$0 < k < \sigma_n^2 < K < \infty \tag{7.2.7}$$

$$c_{i_1 i_2 \dots i_k} = O(n^{-k+1}) \quad \forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 3, \tag{7.2.8}$$

$$c_{i_1 i_2 \dots i_k} = O(\min(n^{-k+2}, 1)) \quad \forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 4. \quad (7.2.9)$$

$$0 \leftarrow \sigma_n^2 < K < \infty \quad (7.2.10)$$

$$c_{i_1 i_2 \dots i_k} = O(n^{-k+1} \sigma_n^{-1}) \quad \forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 3, \quad (7.2.11)$$

$$c_{i_1 i_2 \dots i_k} = O(n^{-k+2}) \quad \forall \quad i_1, i_2, \dots, i_k \text{ satisfying } \sum_{j=1}^k i_j = 4. \quad (7.2.12)$$

$$P_B[w_i > k > 0] = 1 \quad (7.2.13)$$

$$c_3 \rightarrow 1 \quad (7.2.14)$$

$$c_8 < \infty \quad (7.2.15)$$

7.2.2 Assumptions for subsection 7.3.1

We adopt the notations $\phi_{(k+1)ni}(\lambda) = \frac{\partial}{\partial \lambda} \phi_{kni}(\lambda)$, for $k = 0, 1, 2, \dots$, where $\phi_{0ni}(\lambda) \equiv \phi_{ni}(\lambda)$; $\phi_{kni}(\beta_0) = \phi_{kni}$ for $k = 0, 1, 2, \dots$. Let $S_{nj} = \sum_{i=1}^j \phi_{ni}$.

For the sequence $\{a_n\}$ described later on in (7.2.17), define

$$\begin{aligned} s_n^2 &= \sum_{i=1}^n E \phi_{ni}^2 / a_n^2 \\ \gamma_n &= [\sum_{i=1}^n \phi_{ni}^2]^{-1/2} \sum_{i=1}^n \phi_{1ni}. \end{aligned}$$

Assume that for each $n \geq 1$, there is a sequence of σ -fields $\mathcal{F}_{n1} \subset \dots \mathcal{F}_{nn}$, such that $\{S_{nj}, \mathcal{F}_{nj}, j = 1, \dots, n\}$ is a martingale sequence. The following are a set of conditions on this martingale:

(A0)

$$E \phi_{ni} = 0 \text{ for all } 1 \leq i \leq n, \quad n \geq 1. \quad (7.2.16)$$

$$\exists \text{ a sequence } \{a_n \rightarrow \infty\} \text{ such that } K_1 > \sum_{i=1}^n E \phi_{ni}^2 / a_n^2 > k_1 > 0 \quad (7.2.17)$$

(A1)

$$0 < k_2 < \sum_{i=1}^n E \phi_{1ni} / a_n^2 < K_2 \quad (7.2.18)$$

$$0 < \sum_{i=1}^n \phi_{1ni}(\lambda) / a_n^2 \text{ for all } \lambda \quad (7.2.19)$$

(A2) There exist M_{2ni} , possibly random and depending on β_0 , and a sequence $\{c_n\}$ such that

$$\sup_{|\lambda - \beta_0| < \delta_0} |\phi_{2ni}(\lambda)| \leq M_{2ni} \text{ for some fixed } \delta_0 > 0. \quad (7.2.20)$$

$$E\left(\sum_{i=1}^n M_{2ni}\right)^2 = O(c_n^2) \text{ where } a_n^{-4} c_n^2 = O(1). \quad (7.2.21)$$

(A3) For some sequences $\{b_n\}$ and $\{d_n\}$,

$$E[a_n^{-1} \sum (\phi_{1ni} - E\phi_{1ni})]^2 = O(b_n^2) \text{ where } a_n^{-2} b_n^2 = o(1). \quad (7.2.22)$$

$$E[a_n^{-1} \sum (\phi_{ni}^2 - E\phi_{ni}^2)]^2 = O(d_n^2) \text{ where } a_n^{-2} d_n^2 = o(1). \quad (7.2.23)$$

(A4) (Asymptotic normality assumptions)

The triangular sequence $X_{ni} = (\sum_{i=1}^n E\phi_{ni}^2)^{-1/2} \phi_{ni}$ satisfies

$$\max_i |X_{ni}| \xrightarrow{P} 0, \quad \sum_{i=1}^n X_{ni}^2 \xrightarrow{P} 1, \quad \sup_n E(\max_i X_{ni}^2) < \infty \quad (7.2.24)$$

$$\mathcal{F}_{n,i} \subset \mathcal{F}_{n+1,i} \text{ for } i = 1, \dots, n, n \geq 1 \quad (7.2.25)$$

7.2.3 Assumptions for subsections 7.3.2 and 7.3.3

In both subsections 7.3.2 and 7.3.3, we assume that ϕ_{ni} , $i = 1, \dots, n$ are *independent*. Assume that

$$\phi_{ni}(\lambda + t) = \phi_{ni}(\lambda) + \phi_{1ni}(\lambda)t + 2^{-1}\phi_{2ni}(\lambda)t^2 + R_{ni}(t, \lambda)t^2 \quad (7.2.26)$$

where $|R_{ni}(t, \lambda)| < k|t|^\alpha$ for each λ for some $0 < \alpha \leq 1$.

Thus for this restricted set-up we can use simplified expressions for some of the assumptions stated earlier. Following is the list of assumptions in this framework:

$$E\phi_{ni} = 0 \text{ for all } 1 \leq i \leq n, n \geq 1. \quad (7.2.27)$$

$$\sum_{i=1}^n E|\phi_{ni}|^L = O(n) \quad (7.2.28)$$

$$\sum_{i=1}^n E|\phi_{1ni}|^L = O(n) \quad (7.2.29)$$

$$\sum_{i=1}^n E|\phi_{2ni}|^L = O(n) \quad (7.2.30)$$

$$n^{-1} \sum_{i=1}^n \phi_{1ni} > k_1 > 0, \quad (7.2.31)$$

$$E\phi_{ni}^2 > k_1 > 0, \text{ for all } \lambda \quad (7.2.32)$$

$$\phi_{1ni}(\lambda) > k_2 > 0, \text{ for all } \lambda \quad (7.2.33)$$

In the above, the constant L is to be specified in the results. For some resampling schemes that we consider, condition (7.2.31) is not sufficient. Recall the discussion about model conditions in Chapter 3. Suppose m_0 is a specified integer in the range $[n/3]$ to n . For any integer m in $\{m_0, \dots, n\}$ consider the subset $\mathcal{I}_m = \{i_1, i_2, \dots, i_m\}$ of $\{1, 2, \dots, n\}$. We assume

$$m^{-1} \sum_{i \in \mathcal{I}_m} \phi_{1ni} > k_1 > 0 \quad (7.2.34)$$

for every such choice of subset $\mathcal{I}_m$ of size m from $\{1, 2, \dots, n\}$ and for every m in $[m_0, n]$. The more stringent assumption (7.2.34) is required to make resampling schemes like the paired bootstrap and the different jackknives feasible.

Note that (7.2.27)-(7.2.30) covers the conditions listed in (A0) and (A2)-(A3). The conditions (7.2.34) reflect the condition (A1).

7.2.4 Assumptions for subsection 7.3.4

The model conditions here are extensions of the conditions (A0)-(A4) for the p dimensional case, along with a condition on the growth rate of p with respect to n . The following notations will be used: $\|c\|$ is the Euclidean norm of a vector c , A^T is the transpose of the matrix A , $\text{tr}(A)$ is the trace of A , $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are respectively the maximum and minimum eigenvalue of A , and for symmetric matrices A and B , $A > B$ means $A - B$ is positive definite.

Note that each ϕ_{ni} is actually a vector valued function, thus

$$\phi_{ni}(\lambda) = (\phi_{ni1}(\lambda), \dots, \phi_{nip}(\lambda))^T$$

For each $\phi_{nia}(\lambda)$, we assume that the following Taylor expansion holds:

$$\phi_{nia}(\lambda + t) = \phi_{nia}(\lambda) + \phi_{nia}^T(\lambda)t + 2^{-1}t^T H_{2nia}(\lambda_1)t \quad (7.2.35)$$

for $\lambda_1 = \lambda + ct$ for some $0 < c < 1$. We consider two separate possibilities: the dimension p is a fixed positive integer; or $p \rightarrow \infty$ as $n \rightarrow \infty$. As earlier, we often denote $\phi_{ni}(\beta_0)$, $\phi_{1nia}(\beta_0)$ etc. by ϕ_{ni} , ϕ_{1nia} etc.

Let $S_{nj} = \sum_{i=1}^j \phi_{ni}$. Assume that for every n , there is a sequence of σ -fields $\mathcal{F}_{n1} \subset \dots \mathcal{F}_{nn}$, such that $\{S_{nj}, \mathcal{F}_{nj}, j = 1, \dots, n\}$ is a martingale sequence.

(B0) Assume that

$$E\phi_{ni} = 0 \quad (7.2.36)$$

$$\exists \text{ a sequence } \{a_n \rightarrow \infty\} \text{ such that } K > \left[\sup_{\|c\|=1} \sum_{i=1}^n E(c^T \phi_{ni})^2 \right] / a_n^2 > k > 0 \quad (7.2.37)$$

(B1) Assume that

$$\sum_{i=1}^n \sum_{a=1}^p E \|\phi_{1nia}\|^2 = O(a_n^2 p^2) \quad (7.2.38)$$

$$\sum_{i=1}^n \sum_{a=1}^p E \|\phi_{1nia} - E\phi_{1nia}\|^2 = O(a_n^2 p^2) \quad (7.2.39)$$

(B2) For the symmetric matrix H_{2nia} in (7.2.35), for some $\delta_0 > 0$ there exists a symmetric matrix M_{2nia} such that

$$\text{for all } \{t : \|t\| \leq \delta_0\} \quad H_{2nia}(\beta_0 + t) < M_{2nia} \quad (7.2.40)$$

$$\sum_{i=1}^n \sum_{a=1}^p E \lambda_{\max}^2(M_{2nia}) = O(a_n^4 p n^{-1}) \quad (7.2.41)$$

(B3) Let $\phi_{1ni}(\lambda)$ be the $(p \times p)$ matrix, whose a^{th} row is given by $\phi_{1nia}^T(\lambda)$, for $a = 1, \dots, p$. Let $\Gamma_{1n}(\lambda) = a_n^{-2} \sum_{i=1}^n \phi_{1ni}(\lambda)$. Assume that Γ_{1n} is nonsingular. Let $G_{1n} = a_n^{-2} \sum_{i=1}^n E \phi_{1ni}$. Assume

$$0 < k_2 < \lambda_{\min}(2^{-1}(G_{1n} + G_{1n}^T)) \quad (7.2.42)$$

$$\sup_{\|c\|=1} \|\Gamma_{1n}^{-1} c\| = O(p) \quad (7.2.43)$$

(B4) (strengthening of (B3)) Recall that $\Gamma_{1n}(\lambda) = a_n^{-2} \sum_{i=1}^n \phi_{1ni}(\lambda)$. Consider all $\lambda \in \{\lambda : \|\lambda - \beta_0\| < \delta_0\}$, where without loss of generality δ_0 is same as in (B2). Assume

$$0 < k_2 < \lambda_{\min}(2^{-1}(\Gamma_{1n}(\lambda) + \Gamma_{1n}^T(\lambda))) \quad (7.2.44)$$

$$\sup_{\|c\|=1} \|\Gamma_{1n}^{-1}(\lambda) c\| = O(p) \quad (7.2.45)$$

(B5) Consider any fixed vector $c \in \mathbb{R}^p$ with $\|c\| = 1$. Let

$$\begin{aligned} s_n^2 &= p^{-2} \left[\left(\sum_{i=1}^n E \phi_{1ni} \right)^{-1} c \right]^T \left[\sum_{i=1}^n E \phi_{ni} \phi_{ni}^T \right] \left[\left(\sum_{i=1}^n E \phi_{1ni} \right)^{-1} c \right] \\ X_{ni} &= -s_n^{-1} \left[\left(\sum_{i=1}^n E \phi_{1ni} \right)^{-1} c \right]^T \phi_{ni} \end{aligned}$$

Then X_{ni} satisfies

$$\max_i |X_{ni}| \xrightarrow{p} 0, \quad \sum_{i=1}^n X_{ni}^2 \xrightarrow{p} 1, \quad \sup_n E(\max_i X_{ni}^2) < \infty \quad (7.2.46)$$

$$\mathcal{F}_{n,i} \subset \mathcal{F}_{n+1,i} \quad \text{for } i = 1, \dots, n, n \geq 1 \quad (7.2.47)$$

7.3 Main Results

The estimates of the parameters are obtained by solving (7.1.1). We first establish consistency of the estimator and an asymptotic representation, in Theorem 7.3.1 for one dimensional parameter and in Theorem 7.3.5 for general p dimensional parameter, allowing p to tend to infinity with n . Asymptotic normality of the estimator follows as a corollary. The *UBS* estimator also admits an asymptotic linearised representation, which is obtained in Theorems 7.3.2 and 7.3.6 for $p = 1$ and general p respectively. The consistency of the *UBS* distribution estimator follows from this.

Assuming that ϕ_{ni} , $i = 1, \dots, n$ are independent, we obtain two more results. In Theorem 7.3.3 an asymptotic representation of the *UBS* variance estimator is obtained. After bias correction and studentisation of the original estimator as well as its *UBS* equivalent, a second order accuracy result is obtained in Theorem 7.3.4. For these two results we consider one dimensional parameter only for technical convenience, and the results can be easily extended to higher dimensional parameter space. Further details and remarks on the results are given in the different subsections.

The technical conditions and the assumptions of the theorems are all given in Section 7.2. We do not refer to this fact in the statements of the theorems.

7.3.1 Asymptotics and bootstrap for $p = 1$

The first result is on consistency and asymptotic representation of the estimator obtained by solving (7.1.1).

Theorem 7.3.1 *Under (A0)-(A3) exists a sequence $\{\hat{\beta}_n\}$ of solutions of (7.1.1) such that*

$$a_n(\hat{\beta}_n - \beta_0) = O_P(1) \text{ and} \quad (7.3.1)$$

$$\gamma_n(\hat{\beta}_n - \beta_0) = -\left(\sum_{i=1}^n E\phi_{ni}^2\right)^{-1/2} \sum_{i=1}^n \phi_{ni} + r_n \quad (7.3.2)$$

where $r_n = o_P(1)$. If further, $c_n = O(n)$, $b_n = O(1)$ and $d_n = O(1)$, then $r_n = O_P(a_n^{-1})$.

Henceforth we work with that sequence of estimates $\{\hat{\beta}_n\}$ which satisfies Theorem 7.3.1. Assumption (A4) ensures that $\sum_i X_{ni} = a_n^{-1} s_n^{-1} \phi_{ni} \xrightarrow{D} N(0, 1)$ by Theorem 3.2 of Hall and Heyde (1980). The usual case of ' $n^{1/2}$ -consistency' is when $a_n = n^{1/2}$. This may be seen to hold in many cases satisfying (A0)-(A4). The following corollary is immediate.

Corollary 7.3.1 *Assume (A0)-(A4). Then*

$$\gamma_n(\hat{\beta}_n - \beta_0) \xrightarrow{D} N(0, 1) \quad (7.3.3)$$

Let

$$\begin{aligned} F_n(x) &= \text{Prob}[\gamma_n(\hat{\beta}_n - \beta_0) \leq x] \\ \hat{g}_n &= a_n^{-2} \sum_{i=1}^n \phi_{ni}^2(\hat{\beta}_n) \\ \hat{\gamma}_{1n} &= a_n^{-2} \sum_{i=1}^n \phi_{1ni}(\hat{\beta}_n) \\ \gamma_{nB} &= \sigma_n^{-1} [\sum_{i=1}^n \phi_{ni}^2(\hat{\beta}_n)]^{-1/2} \sum_{i=1}^n \phi_{1ni}(\hat{\beta}_n) \end{aligned}$$

For any bootstrap random variable T_{Bn} , the notation $O_{PB}(\xi_n)$ (and $o_{PB}(\xi_n)$) will mean that for any fixed $\epsilon > 0$, the bootstrap probability conditional on the data $P_B[\xi_n^{-1}|T_{Bn}| > \epsilon]$ is respectively $O_P(1)$ (and $o_P(1)$).

The bootstrap estimator is obtained by solving (7.1.2). The next theorem is on its asymptotic representation, which is similar to (7.3.2).

Theorem 7.3.2 *Assume the conditions (A0)-(A3) and the bootstrap weights satisfy conditions (7.2.1) - (7.2.3). Then there exists a sequence $\{\hat{\beta}_B\}$ of solutions of (7.1.2) such that*

(a) *for fixed $\epsilon, \delta > 0$, $\exists K > 0$ and integer n_0 such that for all $n \geq n_0$*

$$\text{Prob}[P_B(\sigma_n^{-1} a_n |\hat{\beta}_B - \hat{\beta}_n| \leq K) < 1 - \epsilon] < \delta \quad (7.3.4)$$

(b)

$$\gamma_{nB}(\hat{\beta}_B - \hat{\beta}_n) = -(\sum_{i=1}^n E\phi_{ni}^2)^{-1/2} \sum_{i=1}^n W_i \phi_{ni} + r_{nB} \quad (7.3.5)$$

where $r_{nB} = o_{PB}(1)$. Further, if $c_n = O(n)$, $b_n = O(1)$ and $d_n = O(1)$, then $r_{nB} = O_{PB}(a_n^{-1})$.

Let

$$F_{Bn}(x) = P_B[\gamma_{nB}(\hat{\beta}_B - \hat{\beta}_n) \leq x].$$

The consistency of the UBS distribution estimator F_{Bn} for F_n is an immediate consequence of the above two theorems and Theorem 5.3.1. This is stated in the following corollary.

Corollary 7.3.2 *Let $\{w_{ni}, i = 1, \dots, n\}$ be exchangeable for every n . Assume (A0)-(A4) and conditions (7.2.1)-(7.2.5) on the bootstrap weights. Then*

$$\sup_x |F_{Bn}(x) - F_n(x)| \rightarrow 0 \text{ in probability} \quad (7.3.6)$$

REMARK 1 Assumption (A2) is less restrictive than the Lipschitz condition (C.1) of Lahiri (1992). We compare our assumptions and results in details with that of Lahiri (1992) in subsection 7.3.3 where we consider higher order accuracy of UBS.

REMARK 2 Existence of second derivative ϕ_{2ni} of ϕ_{ni} and related assumptions may be replaced by the assumption that ϕ_{ni} admits the Taylor series approximation of the form $\phi_{ni}(a+t) = \phi_{ni}(a) + t\phi_{1ni}(a) + o(|t|)$ as $|t| \rightarrow 0$. Note that the functions ϕ_{ni} are random, so the remainder term has a probabilistic interpretation. The assumption needed in that case would be similar to assumption (h1) of Koul (1996).

REMARK 3 The assumption (7.2.19) that $a_n^{-2} \sum \phi_{1ni}(\lambda) > 0$ for any λ may be replaced by requiring only $a_n^{-2} \sum \phi_{1ni}(\beta_0) > 0$. Then the proof of Theorem 7.3.1 remains unchanged, but the proof of Theorem 7.3.2 has to be modified, and the class of *UBS* techniques slightly restricted.

REMARK 4 An important property of the bootstrap weights we need is (7.2.3). Many resampling schemes like the classical bootstrap or the delete- d jackknife satisfies $\sum_{i=1}^n w_{ni} = C_n$ for some sequence of constants C_n . After appropriate scaling to ensure $E_B w_{ni} = 1$, for such schemes it is easily seen that $c_{11} = -(n-1)^{-1}$. Without (7.2.3), the calculations involved in the proof of Theorem 7.3.2 may still be carried out, but the rate conditions would change. The various block bootstrap techniques can be identified in the set-up of the bootstrap weights where (7.2.3) is violated and instead, $c_{11} \rightarrow 1$ typically. Also typically, for such a framework the rate of convergence of various bootstrap quantities are slower. Block bootstrap techniques are powerful tools used for resampling in case of weak dependence. However, in our framework dependencies are of a special kind owing to our ‘martingale’ assumptions, and block bootstrap techniques are not required here. Hence we retain (7.2.3) throughout.

7.3.2 Asymptotic representation of variance estimator

In subsection 7.3.1, it was established that the *UBS* resampling scheme is consistent for approximating the distribution of the parameter estimate. However, since the asymptotic distribution of $\hat{\beta}_n$ is normal, one may simply wish to estimate the asymptotic variance of $\hat{\beta}_n$ using bootstrap. A variance estimate is also required often for confidence intervals (or sets), and hypothesis testing. In this subsection we consider the *UBS* variance estimator and obtain its asymptotic representation. For this subsection and next, we restrict our attention to the case ϕ_{ni} , $i = 1, \dots, n$ are *independent*. Without the assumption of independence, the result of this subsection would be much weaker, and that of the next subsection would require entirely different treatment.

For this subsection, we assume that the *UBS* weights satisfy conditions (7.2.1), (7.2.3), (7.2.6) and either (7.2.7)-(7.2.9) or (7.2.10)-(7.2.12). These can be seen as slightly stronger moment requirements than the requirements of subsection 7.3.1.

We have observed in Chapter 3 that certain *UBS* schemes like the paired bootstrap and the delete- d jackknives effectively select subsets of the data in the resample, and model assumptions

are often required to hold on these subsets. The analogue in the present case is assumption (7.2.34).

The next proposition shows that under appropriate conditions the probability of a 'bad' set is small. This set is used for defining the bootstrap estimator for this subsection.

Proposition 7.3.1 *Assume ϕ_{ni} are independent satisfying conditions (7.2.27)-(7.2.30) with $L = 8(1 + \alpha)$ and (7.2.34). Assume $\hat{\beta}_n$ is a solution to (7.1.1) satisfying (7.3.1)-(7.3.2) from Theorem 7.3.1. Let $\mathcal{A}$ be the set on which $m^{-1} \sum_{i \in \mathcal{I}_m} \phi_{1ni}(\hat{\beta}_n) > k_1/2 > 0$ for every such choice of subset $\mathcal{I}_m$ of size m from $\{1, 2, \dots, n\}$ and for every m in $[m_0, n]$. Then $\text{Prob}[\mathcal{A}] > 1 - O(n^{-2})$.*

We define our bootstrap estimator $\hat{\beta}_B$ to be the solution to (7.1.2) under the set $\mathcal{A} \cap \mathcal{W}$, and $\hat{\beta}_n$ otherwise. The set $\mathcal{A}$ is defined in Proposition 7.3.1, and $\mathcal{W}$ is defined in (7.2.6). The UBS variance estimate is $V_{\text{UBS}} = \sigma_n^{-2} E_B(\hat{\beta}_B - \hat{\beta})^2$. This is used for estimating $V_n = E(\hat{\beta} - \beta_0)^2$. In the statement of the next theorem we have used ϕ , ϕ_1 , ϕ_2 respectively for ϕ_{ni} , ϕ_{1ni} , ϕ_{2ni} . The sums range from 1 to n . Also let $g_{1n} = n^{-1} \sum E\phi_1$, $g_{2n} = n^{-1} \sum E\phi_2$.

Theorem 7.3.3 *Assume $\{\phi_{ni}, i = 1, \dots, n\}$ are independent satisfying (7.2.27)-(7.2.31) and (7.2.34) with $L = 8(1 + \alpha)$. Then*

$$\begin{aligned} & ng_{1n}^2(V_{\text{UBS}} - V_n) \\ &= n^{-1} \sum (\phi - E\phi)^2 - \frac{2}{n^2 g_{1n}} \sum \phi \sum \phi \phi_1 - \frac{2}{n^2 g_{1n}} \sum \phi^2 \sum (\phi_1 - E\phi_1) \\ &\quad + \frac{2}{n^3 g_{1n}^2} \sum \phi \sum \phi^2 \sum \phi_2 + O_P(n^{-1}) \end{aligned} \quad (7.3.7)$$

REMARK. Note that in the special case of M -estimation in simple linear regression, we have $\phi_{ni} = -x_i \psi(y_i - \beta x_i)$. It can be seen that in this case (7.3.7) matches with the expression for the jackknife variance estimator established in Bose and Kushary (1996). A similar representation for the variance estimator of the weighted jackknife of Liu and Singh (1992) can also be established. This weighted jackknife is consistent under homoscedasticity of errors but not under heteroscedasticity. From the results on least squares estimator obtained by Liu and Singh (1992) it was plausible that the weighted jackknife would be more efficient than the jackknife if errors are homoscedastic. However, Bose and Kushary (1996) showed that this is not the case with all M -estimation problems, and even in common criterion functions like Tukey's biweight function the jackknife is more efficient than the weighted jackknife. This follows directly from the asymptotic representation (7.3.7). Hence the UBS is not only robust against heteroscedasticity of errors, but also more efficient than the weighted jackknife in some homoscedastic cases.

7.3.3 Higher order accuracy of UBS

In this section we extend the results of Chapter 6 to estimators obtained by solving estimating equations. The solution $\hat{\beta}_n$ of (7.1.1) whose representation was obtained in Theorem 7.3.1 has a bias factor that is not generally negligible in the second order results. A natural question is how to estimate this bias factor, and study the asymptotic properties of the bias corrected statistic.

By calculations similar to those of the proof of Theorem 7.3.3, we obtain that there is a solution $\hat{\beta}_n$ to (7.1.1) that satisfies

$$-[n^{-1} \sum \phi_1][n^{1/2}(\hat{\beta}_n - \beta_0)] = n^{-1/2} \sum \phi - 2^{-1} n^{-1/2} [n^{-1} \sum \phi_2][n^{1/2}(\hat{\beta}_n - \beta_0)]^2 + r_n$$

where $Er_n^2 = O(n^{-(1+\alpha)})$. Since the algebra is repetitive, we omit the details. The technical conditions required to obtain this are stated later on. The bias correction to this statistic can be done by estimating the second term on the right hand side of the above expression. This requires an estimate for $E[n^{1/2}(\hat{\beta}_n - \beta_0)]^2$. One such estimate would be V_{UBS} from Theorem 7.3.3. However, a different variance estimator, like the simple plug-in $\hat{g}_n^2 = n^{-1} \sum \phi_{ni}^2(\hat{\beta}_n)$ may also be used. Let $g_n^2 = n^{-1} \sum \phi_{ni}^2$ and $\hat{\gamma}_{kn} = n^{-1} \sum \phi_{kni}(\hat{\beta}_n)$, $k = 1, 2$. After some routine algebra, we obtain the following bias corrected studentised estimator

$$T_n = \hat{\gamma}_{1n} \hat{g}_n^{-1} [n^{1/2}(\hat{\beta}_n - \beta_0)] - 2^{-1} n^{-1/2} \hat{\gamma}_{1n}^{-2} \hat{g}_n^{-1} \hat{\gamma}_{2n} V_{UBS} = -n^{-1/2} g_n^{-1} \sum \phi_{ni} + r_n \quad (7.3.8)$$

where $Er_n^2 = O(n^{-(1+\alpha)})$.

For the bootstrap case, the bias correction is done as follows: Define the quantity $\hat{g}_{nB}^2 = n^{-1} \sum W_i^2 \phi_{ni}^2(\hat{\beta}_n)$ and $g_{nB}^2 = n^{-1} \sum W_i^2 \phi_{ni}^2$. Then the bootstrap bias corrected, studentised statistic is given by

$$\begin{aligned} T_{nB} &= \hat{\gamma}_{1n} \hat{g}_{nB}^{-1} [\sigma_n^{-1} n^{1/2} (\hat{\beta}_B - \hat{\beta}_n)] + 2^{-1} n^{-1/2} \sigma_n \hat{g}_{nB}^{-1} \hat{\gamma}_{2n} [\sigma_n^{-1} n^{1/2} (\hat{\beta}_B - \hat{\beta}_n)]^2 \\ &= -n^{-1/2} g_{nB}^{-1} \sum W_i \phi_{ni} + r_{nB} \end{aligned} \quad (7.3.9)$$

where $P_B[|r_{nB}| > kn^{-1/2}(\log n)^{-1}] \rightarrow 0$ almost surely.

The slight difference in the ‘bias correction terms’ in (7.3.8) and (7.3.9) is because in the former we used an estimator for $[n^{1/2}(\hat{\beta}_n - \beta_0)]^2$, whereas in the latter case $[\sigma_n^{-1} n^{1/2}(\hat{\beta}_B - \hat{\beta}_n)]^2$ was directly used. The leading terms of T_n and T_{nB} are approximated by studentised sample sums of possibly non-identical but independent random variables. Edgeworth expansions for such cases have been given in Chapter 6. The model conditions for this section is given below. They are similar to the conditions of Chapter 6.

We assume (7.2.26) is satisfied for some $\alpha \in (0, 1]$, and the model satisfies conditions (7.2.27)-(7.2.30) with $L = 16$ and (7.2.32)-(7.2.33). For any constant $C > 1$, let $p_i(C) = Prob[\phi_{ni}^2 \in [C^{-1}, C]]$. Additionally, assume:

Assumption A $\{\phi_{ni}\}$ is a sequence of independent, absolutely continuous random variables defined on the real line. Let the density of ϕ_{ni} be $f_{ni}(\cdot)$. Define the function $h_{ni}(x) = \min\{f_{ni}(x), f_{ni}(-x)\}$, and let $H_n(x) = n^{-1} \sum_{i=1}^n h_{ni}(x)$. We assume that for all large n , $\exists$ a real valued function $g(x) \geq 0$ such that $\mathcal{A} = \{x : H_n(x) \geq g(x)\}$ has positive measure, and $0 < \int_{\mathcal{A}} g(x) dx < \infty$.

Assumption B For some $c \in (0, 1)$, there exists a constant $C > 1$ such that $n^{-1} \sum p_i(C) \geq c$ for all large n .

For the bootstrap, we assume that the *UBS* weights are *i.i.d.* absolutely continuous, and satisfy (7.2.1), (7.2.7), (7.2.13)-(7.2.15). Additionally, assume:

Assumption C Assume that the distribution of W_1 has a density $f(x)$ such that for some positive numbers x_0 and b , we have $f(x) > b$ for all $x \in (-x_0, x_0)$.

Note that assumptions **A**, **B** and **C** are same as those of Chapter 6. In this set-up, we have the following result by making use of the Edgeworth expansions for studentised sample sums of possibly non-identical but independent random variables. We omit the details of the proof of this theorem which are based on arguments similar to those of Chapter 6.

Theorem 7.3.4 Let $F_{n1}(x) = \text{Prob}[T_n \leq x]$ and $F_{nB1}(x) = \text{P}_B[T_{nB} \leq x]$. Under the assumptions of this section, we have

$$\sup_x |F_{nB1}(x) - F_{n1}(x)| = o(n^{-1/2}) \text{ almost surely} \quad (7.3.10)$$

REMARK 1 Corollary 7.3.2 is a weaker result compared to Theorem 7.3.4 but it was obtained under weaker model assumptions. Also, there the convergence was in probability and the rate of convergence not known.

REMARK 2 A comparison with Lahiri (1992) is in order. Lahiri considers *M*-estimation in linear regression with non-random regressors and *i.i.d.* errors, whereas our model set-up is much broader. When considering *M*-estimators in linear regression, we allow random regressors and heteroscedastic errors in our framework. A quantity γ_n defined in Lahiri (1992, C.6), which is dependent on the design variables, appears as a term in the various convergence rates of Lahiri. But for us it maybe random and thus cannot appear in the convergence rates. Lahiri obtains second order consistency of residual based bootstrap techniques. The resampling techniques we consider may be seen as generalisations of the paired bootstrap. The moment assumptions of Lahiri are of lower order than what we assume, and no bias correction is needed for estimators considered by him. The assumption of absolute continuity of weights was not required by Lahiri. By assuming absolute continuity of the weights we ignore many important resampling estimators.

Thus Lahiri makes more relaxed assumptions than us and obtains second order accuracy in estimating the distribution of M -estimators in linear regression, for a residual based bootstrap. Our set of problems is much larger, the assumptions more stringent and the resampling schemes different from Lahiri. Theorem 7.3.4 is perhaps less appealing than that of Lahiri, since a bias correction of the statistic was needed and strong assumptions were made. However, it is still useful for higher order accurate confidence intervals and other applications. It is plausible that with more effort Theorem 7.3.4 can be proved under much weaker assumptions like strong-nonlattiteness instead of absolute continuity.

7.3.4 Dimension asymptotics

In this subsection we consider the case of $p \geq 1$, and in fact allow the dimension $p \rightarrow \infty$ as the data size $n \rightarrow \infty$. We obtain the consistency of UBS distribution estimator, generalising the result of subsection 7.3.1. The asymptotic representation of variance and second order accuracy of distribution estimation may also be considered, but with additional complications in technicalities involved.

In this subsection, we obtain our results under a general framework involving both p and n . The mutual relation between these two sequence is inherent in the assumptions made. Recall that $\sup_{||c||=1} \sum_{i=1}^n E(c^T \phi_{ni})^2 = O(a_n^2)$ and $s_n^2 = p^{-2}[(\sum_{i=1}^n E\phi_{1ni})^{-1}c]^T[\sum_{i=1}^n E\phi_{ni}\phi_{ni}^T]$. The model assumptions are given in Section 7.2.

The following result is an analog of Theorem 7.3.1.

Theorem 7.3.5 *Under (B0)-(B3) exists a sequence $\{\hat{\beta}_n\}$ of solutions of (7.1.1) such that*

(a) *if $pa_n^{-1} \rightarrow 0$, then*

$$a_n p^{-1/2} ||(\hat{\beta}_n - \beta_0)|| = O_P(1) \text{ and} \quad (7.3.11)$$

$$a_n^{-1} p^{-1/2} (\sum_{i=1}^n E\phi_{1ni})^T (\hat{\beta}_n - \beta_0) = -a_n^{-1} p^{-1/2} \sum_{i=1}^n \phi_{ni} + r_n \quad (7.3.12)$$

where $E||r_n||^2 = O(a_n^{-2} p^2)$.

(b) *if $p^3 a_n^{-2} \rightarrow 0$, then for any $c \in \mathbb{R}^p$ with $||c|| = 1$,*

$$s_n^{-1} c^T (\hat{\beta}_n - \beta_0) = -s_n^{-1} [(\sum_{i=1}^n E\phi_{1ni})^{-1} c]^T \phi_{ni} + o_P(1) \quad (7.3.13)$$

$$= \sum_{i=1}^n X_{ni} + o_P(1) \quad (7.3.14)$$

(c) *if $pa_n^{-2} \rightarrow 0$ and (B4) holds, then*

$$a_n^{-1} p^{-1/2} (\sum_{i=1}^n \phi_{1ni})^T (\hat{\beta}_n - \beta_0) = -a_n^{-1} p^{-1/2} \sum_{i=1}^n \phi_{ni} + r_n \quad (7.3.15)$$

where $E||r_n||^2 = O(a_n^{-2} p^2)$.

Part (c) of the above theorem is actually proved under slightly weaker assumptions. The conditions listed in (B4) is required only for $\lambda = \beta_0$. In many examples satisfying (B0)-(B5), one has $a_n = n^{1/2}$ and $p^2/n \rightarrow 0$ is assumed, so that one has $\hat{\beta}_n = \beta_0 + O_P(n^{-1/2}p^{1/2})$. In many cases $p^2/n \rightarrow 0$ can be improved, even upto $p/n \rightarrow 0$. Part (c) of the above theorem achieves this under some additional conditions. The assumption (B5) ensures that $\sum_i X_{ni} \xrightarrow{D} N(0, 1)$ by Theorem 3.2 of Hall and Heyde (1980). The next corollary is immediate from this.

Corollary 7.3.3 *Assume conditions of Theorem 7.3.5 (b), and (B5). Then for any $c \in \mathbb{R}^p$ with $\|c\| = 1$, we have*

$$s_n^{-1}c^T(\hat{\beta}_n - \beta_0) \xrightarrow{D} N(0, 1) \quad (7.3.16)$$

Suppose $\{\hat{\beta}_n\}$ is that sequence of solutions to (7.1.1) which satisfies (7.3.11). Let $F_n(x) = \text{Prob}[s_n^{-1}c^T(\hat{\beta}_n - \beta_0) \leq x]$. In order to bootstrap, we use weights similar to ones used in subsection 7.3.1. The next theorem is an analog of Theorem 7.3.2.

Theorem 7.3.6 *Assume the conditions (B0)-(B3) and the bootstrap weights satisfy conditions (7.2.1) - (7.2.3). Then there exists a sequence $\{\hat{\beta}_B\}$ of solutions of (7.1.2) such that*

(a) *if $p/a_n \rightarrow 0$, then for fixed $\epsilon, \delta > 0$, $\exists K > 0$ and integer n_0 such that for all $n \geq n_0$*

$$\text{Prob}[\text{P}_B(\sigma_n^{-1}a_n p^{-1/2}|\hat{\beta}_B - \hat{\beta}_n| \leq K) < 1 - \epsilon] < \delta \quad (7.3.17)$$

(b) *if $p/a_n \rightarrow 0$, then*

$$\sigma_n^{-1}a_n^{-1}p^{-1/2}\left(\sum_{i=1}^n E\phi_{1ni}\right)^T(\hat{\beta}_B - \hat{\beta}_n) = -a_n^{-1}p^{-1/2}\sum_{i=1}^n W_i\phi_{ni}(\hat{\beta}_n) + r_{nB1} \quad (7.3.18)$$

$$(7.3.19)$$

where $\|r_{nB1}\| = o_{PB}(p^{-1})$.

(c) *if $p^3a_n^{-2} \rightarrow 0$, then for any $c \in \mathbb{R}^p$ with $\|c\| = 1$,*

$$s_n^{-1}\sigma_n^{-1}c^T(\hat{\beta}_B - \hat{\beta}_n) = \sum_{i=1}^n W_iX_{ni} + o_{PB}(1) \quad (7.3.20)$$

(d) *if $pa_n^{-2} \rightarrow 0$ and (B4) holds, then*

$$\sigma_n^{-1}a_n^{-1}p^{-1/2}\left(\sum_{i=1}^n \phi_{1ni}\right)^T(\hat{\beta}_B - \hat{\beta}_n) = -a_n^{-1}p^{-1/2}\sum_{i=1}^n W_i\phi_{ni}(\hat{\beta}_n) + r_{nB1} \quad (7.3.21)$$

where $\|r_{nB1}\| = O_{PB}(\sigma_n a_n^{-1}p)$.

Let $F_{Bn}(x) = \text{P}_B[s_n^{-1}\sigma_n^{-1}c^T(\hat{\beta}_B - \hat{\beta}_n) \leq x]$.

Corollary 7.3.4 Let $\{w_{ni}, i = 1, \dots, n\}$ be exchangeable for every n . Assume the conditions of Theorem 7.3.6, part (c). Also assume (B5) and conditions (7.2.4)-(7.2.5) on the bootstrap weights. Then

$$\sup_x |F_{Bn}(x) - F_n(x)| \rightarrow 0 \text{ in probability} \quad (7.3.22)$$

The proof of the corollary is omitted.

REMARK The bootstrap variance representation result of subsection 7.3.2 and the second order accuracy result of subsection 7.3.3 can be obtained in p dimensions also. However, the algebra is tedious. Also, the conditions needed on ϕ_{ni} and its derivatives are not simple.

7.4 Examples of UBS resampling techniques

In this section we discuss some special cases of UBS resampling techniques. In each such case, we identify a random mechanism, and show how the UBS conditions (7.2.1)-(7.2.15) relate to it. Apart from these, the only other conditions assumed on UBS weights in this thesis are (4.2.14) and (4.2.15) of Chapter 4. These two assumptions are satisfied in all the examples we consider. Also, some of our results assume that the weights are exchangeable or independent. In all the examples we consider, weights are exchangeable, and cases of independence will be particularly mentioned.

Example 7.4.1 THE CLASSICAL BOOTSTRAP AND ITS VARIATIONS

(a) Consider the vector of bootstrap weights $\mathbf{w}_n = (w_{n:1}, \dots, w_{n:n})$ to be a random sample from $Multinomial(n, 1/n, \dots, 1/n)$. These weights can be interpreted as simple random sampling with replacement of the functionals to minimise, thus this is essentially the classical bootstrap of Efron (1979). All the conditions except for (7.2.10) and (7.2.13) holds in this case. Thus these weights satisfy the conditions required for asymptotic representation of the bootstrap variance estimator and consistency of the bootstrap distribution estimator.

In case of least squares in linear regression, (7.1.2) can be seen as the paired bootstrap. This also clarifies the correct way of carrying out a 'paired bootstrap' in case of polynomial regression, or non linear regression. Thus if the model is $y_i = \beta_0^T g(\mathbf{x}_i) + e_i$, and β_0 is estimated by least squared errors, then the 'paired bootstrap' to use is to draw simple random sample with replacement from the pairs $(y_i, g(\mathbf{x}_i))$, and not $(y_i, \mathbf{x}_i)$. These two procedures do not result in identical estimating equations in the bootstrap world. Let $\xi_i = (\xi_{i1}, \dots, \xi_{im})$ be distributed as $Multinomial(1, 1/n, \dots, 1/n)$. Then by paired bootstrap on $(y_i, g(\mathbf{x}_i))$, from the i^{th} bootstrap sample we get the 'squared error' term $\sum_{j=1}^n \xi_{ij} (y_j - \beta_0^T g(\mathbf{x}_j))^2$, whereas by using the pairs $(y_i, \mathbf{x}_i)$ we get the squared error term $\sum_{j=1}^n (\xi_{ij}^{1/2} y_j - \beta_0^T g(\xi_{ij}^{1/2} \mathbf{x}_j))^2$. The latter can lead to inconsistency, see Stute, Gonzalez Manteiga and Presedo Quindimil (1998), and Shao and Tu (1995, page 354) for two such examples.

(b) THE m OUT OF n BOOTSTRAP Consider the m out of n bootstrap, where $m \rightarrow \infty$, $m/n \rightarrow 0$. This is carried out by selecting m data points randomly out of the n available. If the selection is without replacement, this scheme can be identified with the delete- $(n-m)$ jackknife. If the selection is with replacement, the weights are a random sample from $Multinomial(m, 1/n, \dots, 1/n)$. Both these bootstraps are special cases of *UBS*, and after appropriate scaling of the weights, (7.2.1)-(7.2.5) is satisfied. This establishes first order accuracy of the distribution estimator. Consistency of the variance estimator can also be obtained.

(c) (DOUBLE BOOTSTRAP OR NESTED BOOTSTRAP) The double bootstrap or nested bootstrap is a common technique that is employed for bootstrap confidence intervals of studentised statistics with unknown variances. Here the bootstrap weights $(w_{n1}, \dots, w_{nn})$ follow $Multinomial(n, M_{n1}/n, \dots, M_{nn}/n)$ conditional on $M_n = (M_{n1}, \dots, M_{nn})$, which in turn follows $Multinomial(n, 1/n, \dots, 1/n)$. Thus this is a two stage sampling scheme, where the model for the second stage sampling depends on the outcome of the first stage.

The behaviour of this bootstrap is similar to the classical bootstrap described in Example 7.4.1, but this is more likely to concentrate on fewer data points than the classical bootstrap.

(d) POLYA-EGGENBERGER BOOTSTRAP This technique is similar to the double bootstrap, except that none of the conditional probabilities for the second stage sampling is zero. In this case the weights $(w_{n1}, \dots, w_{nn})$ are from the $Multinomial(n, D_{n1}/n, \dots, D_{nn}/n)$ distribution conditional on the random vector $D_n = (D_{n1}, \dots, D_{nn})$, which follows $Dirichlet_n(\alpha, \dots, \alpha)$ for some parameter $\alpha > 0$.

The behaviour of this bootstrap is similar to the classical bootstrap and the double bootstrap.

(e) THE MULTIVARIATE HYPERGEOMETRIC BOOTSTRAP In this example, the weights come from the multivariate hypergeometric distribution

$$Prob[w_{n1} = k_1, \dots, w_{nn} = k_n] = \binom{nK}{n}^{-1} \binom{K}{k_1} \dots \binom{K}{k_n}; \quad \sum_{i=1}^n k_i = n, \quad 0 \leq k_i \leq K$$

where K is some integer parameter.

The behaviour of this bootstrap is similar to the classical bootstrap, or the double bootstrap or Polya-Eggenberger bootstrap.

The m out of n bootstrap is studied in details in Bickel and Sakov (1999), and the variations (c)-(e) of the classical bootstrap have been mentioned in Praestgaard and Wellner (1993).

Example 7.4.2 THE DELETE- d JACKKNIVES For some integer d in $\{1, \dots, n\}$, consider the n dimensional vectors $\eta_{n:i_1, i_2, \dots, i_d}$ defined as follows: The j^{th} coordinate of $\eta_{n:i_1, \dots, i_d}$ is 0 if j is one of $i_1, \dots, i_d$, else it is $n/(n-d)$. Thus there are $\binom{n}{d}$ such vectors. Consider w_n to be a sample from the set of η , each of the $\binom{n}{d}$ vectors having the same probability of being selected.

Such weights can be identified with the delete- d jackknife weights. It can be seen that (7.2.1), (7.2.3) always holds. If $n - d \rightarrow \infty$ then (7.2.2) also holds, and (7.2.6) holds for $m_0 = n - d$.

If $d/n \rightarrow c \in (0, 1)$, then (7.2.4)-(7.2.5), (7.2.7)-(7.2.9) and (7.2.15) also hold. If $d/n \rightarrow 0$, then (7.2.10)-(7.2.12) hold. Thus if $n - d \rightarrow \infty$, the basic representation of the jackknife estimator holds. If further, $d/n \rightarrow c \in (0, 1)$, then first order consistency of the delete- d jackknife distribution estimator is established. However, if $d/n \rightarrow 0$, (7.2.5) is violated, and the jackknife histogram is no longer consistent distribution estimator. However, for variance estimation, the jackknives are consistent. A higher order accuracy of the jackknife distribution may not be expected in general, since a crucial skewness condition (7.2.14) fails to hold.

The bootstrap equation (7.1.2) is a new and interesting way to consider jackknife in time series models, say in autoregressions. The standard interpretation associated with the jackknife, that of 'deleting d observations' and considering the rest does not make sense in time series, since that way the crucial temporal dependence among the observations is lost. Earlier, Kunsch (1989) suggested a jackknife based on block size of the block bootstrap. Our suggestion is different, we 'jackknife' the estimating equations.

Example 7.4.3 THE BAYESIAN BOOTSTRAP Consider $\mathbf{w}_n$ to be a random sample from $\text{Dirichlet}(\alpha, \dots, \alpha)$ distribution. All the conditions on weights except for (7.2.10) and (7.2.14) are satisfied by these weights. This shows first order accuracy of distribution estimate and consistency of variance estimate from the 'Bayesian bootstrap'.

If $\alpha = 4$, then (7.2.14) is also satisfied. Dirichlet weights are exchangeable but not independent, however the i^{th} such weight can be expressed as $nX_i/(\sum X_i)$ for some i.i.d. sequence X_i . This representation is often found useful in studying the second order accuracy of Dirichlet bootstrap. We discuss this further in Example 7.4.4.

Example 7.4.4 THE BOOTSTRAP CLONE METHOD Consider the weights w_{ni} to be i.i.d. random variables, satisfying moment and support restrictions imposed by (7.2.1)-(7.2.15). Absolute continuity and restrictions on the density of the weights may also be imposed. Such weights may be used to get first and second order accurate distribution estimates and variance estimate. Very often as a variation of this, generalised bootstrap is done by considering weights of the form $w_i = nX_i/(\sum X_i)$, for i.i.d. absolutely continuous random variables X_i . This is the idea behind the 'bootstrap clone' method of Lo (1987, 1988, 1991). A recent application of this method is given in James (1997). In (7.1.2), in most situations one may replace such w_i by X_i without loss. Thus second order accurate distribution estimates and variance estimates may be obtained for all such weights. This covers the case of Dirichlet weights also, discussed in Example 2.3.2.

Example 7.4.5 THE WEIGHTED LIKELIHOOD BOOTSTRAP In case $\phi_{ni}(\cdot)$ has a log-likelihood interpretation, the weighted likelihood bootstrap of Newton and Raftery (1994) is also a special

case of *UBS*. This can be used for approximate Bayesian inference as suggested by Newton and Raftery (1994). We have only looked at the frequentist properties of the bootstrap in this thesis. A completely different treatment for bootstrapping likelihoods has been pursued by Davison, Hinkley and Worton (1992). See also Mammen (1996).

7.5 Examples of martingale estimating equations

In this section we discuss some examples of familiar problems in statistics. In subsection 7.5.1 the parameter estimators are obtained by solving martingale equations and in subsection 7.5.2 the set-up is slightly generalised to include examples from the semiparametric framework. In each example, we mention some routine assumptions that are made. These turn out to be sufficient to ensure our technical assumptions. We discuss in some details the problems of M -estimation, least squares estimation in non-linear regression and in generalised linear models. A number of other problems use variations of these.

7.5.1 Finite dimensional parameter problems

Resampling for M -estimators was profoundly studied by Lahiri (1992), and before that by Mammen (1989), who also incorporated dimension asymptotics aspect in his study. In non-linear regression, resampling was studied in Huet and Jolivet (1989); Huet, Jolivet and Messeean (1990) and Gruet, Huet and Jolivet (1993). In generalised linear models resampling was studied by Simonoff and Tsai (1988), Moulton and Zeger (1989, 1991), Sauermann (1989) and Lee (1990). Shao (1992a,b,c) studied the properties of jackknife variance estimators in non-linear regression and generalised linear models. However, almost entirely the literature has concentrated on residual based resampling. As has already been pointed out, the *UBS* may be seen as a generalisation of the paired bootstrap and delete- d jackknives. We keep our discussion non-technical with the understanding that in specific examples, appropriate assumptions have to be made and the technical steps have to be verified. In the examples, some natural conditions are suggested to verify our technical assumptions.

Example 7.5.1 M -ESTIMATION IN LINEAR REGRESSION

Let $y_i = x_i\beta_0 + e_i$ be the linear model and β_0 be the parameter to be estimated. In this example, we have $\phi_{ni}(y_i, x_i, \lambda) = -x_i\psi(y_i - \lambda x_i)$ for some function $\psi(\cdot)$. We assume that the covariate observations x_i are non-random, and the errors e_i are independent. The formulation with random covariates is similar, with some additional complexity. We assume that the function $\psi(\cdot)$ is twice continuously differentiable, with the first and second derivatives denoted by $\psi_1(\cdot)$, $\psi_2(\cdot)$ respectively.

In the following assumptions, k_i and K_i are used as notations for different constants.

(M0) Assume that for all i , $0 < k_1 < |x_i| < K_1 < \infty$. This assumption was made in Liu and Singh (1992), for example.

(M1) For all i , $E\psi(e_i) = 0$.

(M2) For all i , $0 < k_2 < E\psi^2(e_i) < K_2 < \infty$, and $E\psi^4(e_i) < \infty$.

(M3) For all i , $\psi_1(e_i) > 0$, $0 < k_3 < E\psi_1(e_i)$ and $0 < E\psi_1^2(e_i) < K_3 < \infty$.

(M4) For some $\delta > 0$, $\exists$ a function $M(\cdot)$ such that for all i , $\sup_{|t| < \delta} |\psi_2(e_i + t)| < M(e_i)$, and $\sup_i EM^2(e_i) < \infty$.

(M5) $\{\psi(e_i)\}$ satisfy the Lindeberg condition.

The above assumptions ensure that assumptions (A0)-(A5) hold. Note that conditions (M0)-(M5) are only sufficient conditions. For some robust regression techniques $\psi_2(\cdot)$ is assumed to be bounded, and this ensures (M4).

Suppose $\gamma_{Nn} = n^{-1} \sum_{i=1}^n x_i^2 E\psi^2(e_i)$ and $\gamma_{Dn} = n^{-1} \sum_{i=1}^n x_i^2 E\psi_1(e_i)$. The asymptotic variance of the estimator is given by the limit of $\gamma_0 = \lim_{n \rightarrow \infty} n\gamma_{Nn}/\gamma_{Dn}^2$. The simplified expressions for this is

$$\gamma_0 = \lim_n \frac{n \sum_{i=1}^n x_i^2 E\psi^2(e_i)}{[\sum_{i=1}^n x_i^2 E\psi_1(e_i)]^2} \text{ generally, and} \quad (7.5.1)$$

$$\gamma_0 = \lim_n \frac{n E\psi^2(e_1)}{\sum_{i=1}^n x_i^2 E^2\psi_1(e_i)} \text{ for homoscedastic errors} \quad (7.5.2)$$

For this example, the first order consistency of *UBS* distribution approximation holds. With higher moment assumptions, the variance representation and second order accuracy of distribution can also hold.

Example 7.5.2 LEAST SQUARES IN NON-LINEAR REGRESSION AND GLM

In nonlinear regression we have the model

$$y_i = f(\mathbf{x}_i, \beta_0) + e_i$$

for some known function $f(\cdot)$ and in generalised least squares, we have

$$\mu(y_i) = \mathbf{x}_i \beta_0 + e_i$$

for some link function $\mu(\cdot)$. Suppose the estimator is obtained by least squares. The problem of least squares in nonlinear regression and generalised linear models are very similar, and technically they can be treated in a unified fashion. See Bose and Sengupta (1992) for some details on this.

For either of the problems, we have $\phi_{ni}(y_i, \mathbf{x}_i, \lambda) = -f_1(\mathbf{x}_i, \lambda)(y_i - f(\mathbf{x}_i, \lambda))$ for some function f whose k^{th} partial derivative with respect to λ is written as f_k .

We again assume non-random covariates $\mathbf{x}_i$ and independent errors e_i . We assume that for all $\mathbf{x}_i$, the functions $f_1(\mathbf{x}_i, \lambda)$, $f_2(\mathbf{x}_i, \lambda)$, $f_3(\mathbf{x}_i, \lambda)$ exist. In the following assumptions, c_i and C_i are used as notations for different constants.

(NG0) Assume that for all i , $0 < c_4 < |x_i| < C_4 < \infty$.

(NG1) For all i , $Ee_i = 0$.

(NG2) For all i , $0 < c_2 < Ee_i^2$ and $Ee_i^4 < \infty$.

(NG3) Assume that for all i , the functions $f_k(\mathbf{x}_i, \lambda)$, $k = 1, 2, 3$ are bounded.

(NG4) For all i , $0 < c_1 < |f_1(\mathbf{x}_i, \lambda)| < C_1 < \infty$.

(NG5) For all i , the functions $g(\mathbf{x}_i, \lambda)$ is continuous at β_0 for $g = f, f_1, f_2, f_3$.

(NG6) Assume that $f_1(\mathbf{x}_i, \beta_0)e_i$ satisfies the Lindeberg condition.

Conditions (NG0)-(NG6), ensure that assumptions (A0)-(A5) hold. It may be checked that conditions (NG0)-(NG5) are satisfied in commonly used generalised linear models, see McCullagh and Nelder (1989). The error conditions (NG1) and (NG2), and (NG6) are usual assumptions. Note that all the usual link functions are injective functions, and in our framework, the function $f(\mathbf{x}_i, \lambda)$ is actually the inverse of a link function. For the link functions described in McCullagh and Nelder (1989, pages 30-32), it is a small matter to check that conditions (NG3), (NG4), (NG6) hold, if the covariate variable x satisfies (NG0).

Suppose $\gamma_{Nn} = n^{-1} \sum_{i=1}^n f_1^2(\mathbf{x}_i, \beta_0) Ee_i^2$ and $\gamma_{Dn} = n^{-1} \sum_{i=1}^n f_1^2(\mathbf{x}_i, \beta_0)$. The asymptotic variance of the estimator is given by the limit of $\gamma_0 = \lim_{n \rightarrow \infty} n\gamma_{Nn}/\gamma_{Dn}^2$. This can be simplified as

$$\gamma_1 = \lim_n \frac{n \sum_{i=1}^n f_1^2(\mathbf{x}_i, \beta_0) Ee_i^2}{[\sum_{i=1}^n f_1^2(\mathbf{x}_i, \beta_0)]^2} \text{ generally, and} \quad (7.5.3)$$

$$\gamma_2 = \lim_n \frac{n Ee_i^2}{\sum_{i=1}^n f_1^2(\mathbf{x}_i, \beta_0)} \text{ for homoscedastic errors} \quad (7.5.4)$$

We have only discussed above about models with independent errors. Non-linear regression models with martingale difference errors have been considered (Lai (1994)). Conditions of a different flavour appear in Wu (1981). The framework of both these papers overlap with ours.

The various extensions of generalised linear models (see Jiang (1998) for example) may also be treated in our set-up.

The different *UBS* consistency results may be seen to hold in all such cases.

Example 7.5.3 MODELS WITH NUISANCE PARAMETERS The following problem has been of considerable interest (Neyman and Scott (1948), Bhanja and Ghosh (1992), Cox (1993)): Suppose X_i , $i = 1, \dots, n$ are independent with mean μ and variance τ_i^2 , where μ is the parameter of interest. We may use the ordinary least squares, and thus have $\phi_{ni} = -(X_i - \mu)$. The different sufficient conditions of our set-up can be satisfied in this example by assuming $\sup_n EX_n^{16}$ is bounded. This ensures first and second order accuracy of *UBS* distribution estimator, and the *UBS* variance representation result. A more careful analysis specialised to this problem only shows that actually a much lower moment is sufficient.

Example 7.5.4 AUTOREGRESSIVE PROCESSES Suppose we have a stationary process $\{X_n\}$ satisfying $X_n = \theta_1 X_{n-1} + \dots + \theta_p X_{n-p} + e_n$, where $\{e_n\}$ are *i.i.d.* and the polynomial $1 - \theta_1 z - \dots - \theta_p z^p$ has its roots outside the unit disc over the complex plane. The parameter of interest is $\beta_0 = (\theta_1, \dots, \theta_p)$, whose estimate may be obtained by various techniques, see Brockwell and Davis (1986) for details.

Assume $p = 1$ and least squares estimation technique for simplicity. Thus we have $X_t = \theta X_{t-1} + e_t$ where $|\theta| < 1$ and $\{e_n\}$ are *i.i.d.*. Here we have $\phi_{ni} = -X_{i-1}e_i$ whose martingale structure is easily derived. If fourth moment of e_1 is bounded, other conditions may also be easily checked. The *UBS* is thus a different way of bootstrapping in autoregression, that does not use estimated innovations.

One can use general *M*-estimation also in time series models. Recently Allen and Datta (1999) has used estimated innovations based approach for bootstrapping in such problems. Some modified forms of the innovations have been used by them, which is comparable to the modifications required on residuals for use in bootstrapping *M*-estimators in linear regression as in Lahiri (1992). Our technique of resampling is perhaps a simpler approach.

Example 7.5.5 CO-INTEGRATED PROCESSES This example is on multivariate time-series data. A univariate time series $\{Y_t\}$ is integrated d times (I_d) if differencing it d times results in a stationary, invertible *ARMA* process. A multivariate time series process $\{X_t = (X_{1t}, \dots, X_{(p+1)t})\}$ is co-integrated of order (d, b) if (i) $\{X_{kt}\}$ is I_d and (ii) there exists a non-zero vector a such that $a^T X_t$ is $I_{(d-b)}$ for $b > 0$. The vector a is called the co-integrating vector and is the parameter of interest. For a discussion on these models see Engle and Granger (1987). Co-integrated time series have applications in many economic problems. An important feature here is the study of equilibrium, that is, when different economic variables behave in a way that preserves some economic balance between them, which translated in mathematical terms, mean that the variables are I_d for some $d > 0$, and $d = b$, thus the co-integrated process is stationary.

Consider a simple example of a co-integrated process: where we have $X_{2t} = \beta_0 X_{1t} + Y_t$, where X_{1t} is $I(1)$ and Y_t is $AR(1)$. Suppose ordinary least squares on (X_{1t}, X_{2t}) is used and the parameter estimator is $\hat{\beta}_n = (\sum X_{1t}^2)^{-1} \sum X_{1t} X_{2t}$. In this example we have $a_n = n$, thus the

usual $n^{1/2}$ -convergence phenomenon is violated. Let $Y_t = \theta e_{t-1} + e_t$ and $X_{1t} = X_{1(t-1)} + u_t$, and assume that $\{u_t\}$ and $\{e_t\}$ are independent, each with mean zero and finite non-zero variance. Our model conditions (A0)-(A4) are satisfied by assuming bounded fourth moment of X_{1t} and Y_t .

The asymptotics for estimating the co-integrating vector and its bootstrap may be derived from our set-up.

Example 7.5.6 ROBUST ESTIMATORS Robust estimators of location, scale and regression are typically M -estimators like the ones considered in Example 7.5.1, and may be similarly treated. Other type of robust estimators like L -estimators and R -estimators are also closely related to M -estimators. See Jurekova and Sen (1996) for details on these issues. Simultaneous robust estimators of location and scale, and scale estimators like S -estimators may be found in Lopuhaa (1991).

A weighted likelihood based approach was adopted by Markatou, Basu and Lindsay (1998) for diagnostics and robust estimation of parameters. Our set-up covers such inference based on likelihoods, where ϕ_{ni} has the interpretation of being derivative of (weighted or unweighted) log-likelihood terms. Note that the martingale structure is particularly important in such situations.

Generalised estimating equations, often based on modifications of likelihood or 'quasi-likelihood', have also been used in special situations, for example, for clustered data (Hall and Severini (1998)) or data missing completely at random (Chen and Little (1999)). Our set-up adapts for such problems also.

In all these examples, the assumptions (A0)-(A1) on the function ϕ_{ni} and its first derivative is standard, and the second derivative is often bounded, hence assumption (A2) is satisfied. The moment assumptions in (A3) and the asymptotic normality assumptions of (A4) are also routinely satisfied.

Example 7.5.7 LOCAL ESTIMATING EQUATIONS In many semiparametric problems a local polynomial spline fitting technique is used. Such problems also involve estimating a finite number of parameters using estimating equations, and our set-up potentially applies. References for these kind of problems are Carroll, Rupert and Welsh (1998) and Zhou, Shen and Wolfe (1998).

Let us take a simple case: Suppose y_i 's are observations, which are modelled using non-random covariate x_i , and the modelling procedure is that of fitting a local linear spline, at a point x in the interior of the range of x_i 's. This may done by minimising

$$\sum h_x(x_i)(y_i - b_0(x) - x_i b_1(x))^2$$

where h_x is some weight function denoting the neighbourhood of x . The nearest neighbour regression, loess and other curve fitting techniques may be considered in this framework, and

a very general form of this has been discussed in Carroll, Rupert and Welsh (1998). This is a weighted least squares problem, and more general forms of this may be viewed as M -estimation in regression, and our set-up applies here, with assumptions as in Example 7.5.1. Usually in cases similar to this example the entire fitted curve is the object of interest, and not the local parameters $b_0(x)$ or $b_1(x)$. But our bootstrap technique may be useful in deriving confidence bands for the curve, whose pointwise properties may be derived from our results.

7.5.2 Semiparametric inference problems

In semiparametric inference problems, along with the finite dimensional parameter an infinite dimensional nuisance parameter is present. One standard approach is to use a kernel based plug-in for the infinite dimensional parameter, and then use least squares technique for the finite dimensional parameter. Since estimation of the infinite dimensional component involves the entire data, the crucial martingale property of the estimating equations for the finite dimensional parameter estimator is lost. However, a slight generalisation of the set-up of this chapter can be used for such cases, as explained below. We restrict our discussion to the problem of asymptotic normality and consistency of UBS . Higher order results and variance representation will require more careful analysis.

Assume that $\phi_{ni} = m_{ni} + r_{ni}$ where $\{(S_{nj} = \sum_{i=1}^j m_{ni}, \mathcal{F}_{nj}), j = 1, \dots, n\}$ is a martingale for some filtration $\mathcal{F}_{nj}$, and r_{ni} is an 'error' term resulting from using an estimator $\hat{m}(\cdot)$ of the non-parametric component $m(\cdot)$ of the model. We may take $\mathcal{F}_{nj}$ to be the σ -field generated by the first j observations. Suppose the estimator of the finite dimensional parameter is obtained by solving $\sum_{i=n-n_1+1}^n \phi_{ni} = 0$ for some integer n_1 . The technical conditions of Section 7.2 are required to hold for m_{ni} and its derivatives only. An additional requirement is that $R_{nn_1} = \sum_{i=n-n_1+1}^n r_{ni} = o_P(n_1 a_{n_1}^{-1})$ and is $\mathcal{F}_{n,n}$ measurable. Moment conditions on $R_{nn_1}^2$ may also be imposed. In the typical case it is required that $n_1 \rightarrow \infty$ and $n_1/n \rightarrow 0$, and one has $a_n = n_1^{1/2}$. The bootstrap estimator is obtained by solving $\sum_{i=n-n_1+1}^n w_{ni} \phi_{ni} = 0$ for the same integer n_1 .

The solution to $\sum_{i=n-n_1+1}^n \phi_{ni} = 0$ and $\sum_{i=n-n_1+1}^n m_{ni} = 0$ would be close if R_{nn_1} is small. By a careful reading of the proofs it can be seen that our asymptotic representation theorems only use the property $O((\sum_{i=n-n_1+1}^n \phi_{ni})^{2k}) = O(\sum_{i=n-n_1+1}^n \phi_{ni}^{2k})$ and this is available in the more general set-up also. The asymptotic normality of the estimator and consistency of the UBS are also easily established using the martingale difference structure of $\{m_{ni}\}$ and asymptotic negligibility of sums of r_{ni} 's.

We consider three examples below in this slightly general framework.

Example 7.5.8 LOW DIMENSIONAL COMPONENT ESTIMATION IN SEMIPARAMETRIC ADDITIVE REGRESSION MODEL Consider the regression problem: $y_i = f(\mathbf{x}_i, \beta_0) + m(Z_i) + e_i$, where $f(\cdot)$ is a known function, $m(\cdot)$ is an unknown function, β_0 is finite dimensional parameter, e_i 's are independent. Sometimes $f(\cdot)$ is linear, resulting in the model being a partial linear model. These models have generated interest owing to their wide applicability. See Fan, Hardle and Mammen (1998) for some results on these. The partial linear model has been considered in Shi (1998) and in much greater generality in Mammen and Van de Geer (1997).

In order to estimate β one may use least squares criterion and minimise $\sum_i (y_i - f(\mathbf{x}_i, \beta_0) - m(Z_i))^2$. The properties of such estimators may be studied in our set-up. However, since m is not known and has to be estimated, this framework is not the one that is adopted in practice and the minimiser of the above is only of theoretical interest.

Let us consider the partial linear model and suppose that the parameter of interest is β_0 . The estimation technique is as follows: for any value of β in the above equation, the estimation of $m(\cdot)$ is by using non-parametric regression of $y_i - \mathbf{x}_i^T \beta$ on Z_i , to yield a (possibly kernel based) non-parametric estimator $\hat{m}(\cdot)$, which is then used as a plug-in and the parameter β_0 estimated by least squares. Using the technique described at the beginning of this subsection, asymptotics and bootstrap for an estimator of β_0 may be obtained.

In some special cases with algebraic manipulations it can be shown that taking $n_1 = n$ also works, as in the proof of Theorem 1.1 of Shi (1998). More complicated setup has been analysed in Mammen and Van de Geer (1997).

It is often difficult to decide if the component $m(Z_i, \tau)$ is required or whether the low dimensional component suffices for modelling. This is an unrelated but important issue, and a test of parametric versus semiparametric modeling in the generalised linear models has been recently developed by Hardle, Mammen and Muller (1998).

Example 7.5.9 ESTIMATION IN PARAMETRICALLY LINKED ADDITIVE MODELS Consider the set-up $y_i = \mu + \sum_{j=1}^J m_j(X_j) + e_i$, where m_j 's are unknown functions. For identifiability, set $E m_j(X_j) = 0$, $j = 1, \dots, J$. This is the set-up of the additive nonparametric regression, studied in Hastie and Tibshirani (1990), Fan, Hardle, Mammen (1998) and many others. Motivated from non-parametric GARCH model, Carroll, Hardle and Mammen (1999) consider $m_j(x) = \beta_0^{j-1} m_1(x)$, and simultaneous estimation of $m_1(\cdot)$ and β_0 . Carroll, Hardle and Mammen (1999) discuss examples from financial data and autoregression, and also cite several references for other applications.

As in Example 7.5.8, the nonparametric component m_1 is estimated from the data. In the present case, suppose $\mu = 0$ and assume that $\hat{m}_1(x)$ is some nonparametric plug in, and the parameter β_0 is estimated by minimising

$$\sum_{i=1}^{n_1} (y_i - \sum_{j=1}^J \beta_0^{j-1} \hat{m}_1(x_j))^2$$

The estimator depends on the choice of the non-parametric plug-in, but in applications the value of the estimator does not seem to vary much for different choices of the non-parametric plug-in. The asymptotics and bootstrap for the estimator of β_0 can be derived from the more general version of our set-up.

In a similar way our set-up may be used for inference about the parametric component in semiparametric proportional odds regression model, as in Yang and Prentice (1999). The approach in this kind of modelling is based on deriving estimating equations for the parametric component using plug-ins for the non-parametric component (see equation (9) in Yang and Prentice (1999)).

Example 7.5.10 EXPONENTIAL FAMILY DENSITY ESTIMATES Suppose the data $y_1, \dots, y_n$ are *i.i.d.* observations from an absolutely continuous distribution with density $g(x) = g_0(x)\exp(b_0 + t(x)b_1)$. Here $g_0(x)$ is a *carrier density*, and the exponential component ensures that the density belongs to exponential family with sufficient statistic $t(x)$. The parameter b_0 is a normalising factor, whereas b_1 is the finite dimensional parameter of the exponential family. The problem of density estimation in this family have been considered in Efron and Tibshirani (1996). This semiparametric model is remarkably versatile in applicability and mathematical tractability. More classical density estimation procedures often require post-repair of the mean, variance etc. (see Jones (1993)), which are in-built in the present model. The carrier $g_0(x)$ is usually uniform density, but other densities can also be used. This is effective in capturing the local properties of the data, whereas the exponential term captures the global features.

The problem of density estimation in these cases reduces to estimation of $\beta_0 = (b_0, b_1)$. The problem is essentially similar to generalised linear models using a Poisson regression approach. The method is as follows: partition the sample space into K cells, and consider the data $s = (s_1, \dots, s_K)$ of the vector of cell counts, *i.e.*, s_k is the number of original observations in the k^{th} cell. The vector s obviously can be modelled as multinomial data, and some orthogonality of parameters based arguments may be used to model s_k as independent *Poisson*($\mu_k(\beta_0)$), where $\mu_k(\beta_0) = \mu_{k0} \exp(b_0 + t_k b_1)$. Here μ_{k0} is the probability of the k^{th} cell under the carrier density, and t_k is the evaluation of the sufficient statistic at some convenient point inside the k^{th} cell. Then defining $\mathbf{x}_k = (1, t_k)$, $\mathbf{X}$ to be the matrix whose k^{th} row is $\mathbf{x}_k$, and $\mu(\beta_0)$ to be the vector whose k^{th} entry is $\mu_k(\beta_0)$; we have the following estimating equation:

$$\mathbf{X}^T(s - \mu(\beta_0)) = 0$$

which is a generalised linear model. The conditions from Example 7.5.2 can be adapted here for asymptotics and bootstrap of the estimator of β_0 . The more general model framework described at the beginning of this subsection is applicable in this case also.

As with Example 7.5.9, a nonparametric plug-in is used for the carrier density. The choice of K and the non-parametric plug-in does not seem to matter much in practice. See comments

by Efron and Tibshirani (1996) on this matter. The problem of locally parametric density estimation, as considered by Hjort and Jones (1996) is also similarly treated. This particular technique is also widely applicable and versatile. Here the finite set of parameters of interest may be estimated by techniques similar to those of M -estimation.

Example 7.5.11 REGRESSION WITH CENSORED OBSERVATIONS This example is from typical regression problems that arise in the context of survival analysis. We consider only one of a large variety of techniques, namely, the proportional hazards model. Details about such models can be found in Fahrmeir and Tutz (1994). The data generally consists of the triplet $(t_i, \delta_i, \mathbf{x}_i)$, where t_i is an observed time point, δ_i is the indicator on whether the i^{th} observation is uncensored, and $\mathbf{x}_i$ is the observation on the covariate. The proportional hazards model is given by the hazard function $\lambda(t) = \lambda_0(t) \exp(\mathbf{x}_i^T \beta_0)$ where λ_0 is a baseline hazard function and β_0 is the parameter of interest. Modelling is often done by using a plug-in for λ_0 , like an exponential or Weibull hazard function. This is similar to Examples 7.5.9 or 7.5.10. The rest of the modelling technique is similar to generalised linear models. Conditions from Example 7.5.2 can be adapted in this set-up.

7.6 Computational aspects and simulation results

In a typical case (7.1.1) is difficult to solve, and would require heavy computational machinery. For the bootstrap also, an extremely computation intensive Monte Carlo method is generally used. Thus it is very routine in resampling to draw samples from the distribution of $(w_1, \dots, w_n)$ say a large B number of times, compute the value of the bootstrap statistic for each of these B cases, and finally use the histogram of the B bootstrap statistic values. If obtaining the estimator for a single data set of size n is a computationally involved operation, bootstrap usually compounds this by requiring the entire process to be repeated B times, for large B . In this section we indulge in some heuristics, which is motivated at bypassing and reducing the additional computational burden imposed by bootstrap. We suggest a different way of computing the bootstrap statistics that uses no iteration, and is unrelated with the computation of the original estimator. We offer no mathematical justification about this different method of computation of bootstrap estimates, however, we later on remark why it may be expected to work. Some simulations are also carried out to verify that the new method of computation leads to expected results.

In order to numerically solve (7.1.1), one very common and useful technique is the *iteratively reweighted least squares* algorithm. As Bloomfield and Steigler (1983) remark, the *IRLS* is an important tool for calculating M -estimators. For non-linear regression and generalised linear models, variations of this is the most commonly used technique for computation, see Seber and Wild (1989) and McCullagh and Nelder (1989). There is a connection between the *IRLS*

algorithm and the Newton-Raphson technique or its variant the scoring method, see Seber and Wild (1989).

The *IRLS* algorithm, or Newton-Raphson method essentially uses the following steps:

- (Step 1) Rewrite $\phi_{ni}(\lambda) = \phi_{ni}(y_{ni}, \mathbf{x}_{ni}, \lambda)$ as a product of two functions, one of which is linear in λ . Thus $\phi_{ni}(\lambda) = g_{ni}(y_{ni}, \mathbf{x}_{ni}, \lambda)[h_{1ni}(y_{ni}, \mathbf{x}_{ni}, \lambda) - \lambda h_{2ni}(y_{ni}, \mathbf{x}_{ni}, \lambda)]$. For example, in M -estimation $\phi_{ni}(\lambda) = x_i \psi(y_i - \lambda^T \mathbf{x}_i)$, so we may take $g_{ni}(\lambda) = \phi_{ni}(\lambda)/(y_i - \lambda^T \mathbf{x}_i)$, $h_1 = y_i$ and $h_2 = \mathbf{x}_i$. For least squares in nonlinear regression, by a Taylor series expansion and standard approximations used in Newton-Raphson method, we have $\phi_{ni}(\lambda) = f_1(\lambda_0, \mathbf{x}_i)(y_i - f(\lambda_0, \mathbf{x}_i) - f_1(\lambda_0, \mathbf{x}_i)(\lambda - \lambda_0))$, so g_{ni} , h_{1ni} and h_{2ni} are obvious.
- (Step 2) Get an initial estimate λ_0 . In case of non-linear regression, this is the point about which the Taylor series expansion is taken. Generally, evaluate $g_{ni}(\lambda)$, $h_{1ni}(\lambda)$ and $h_{2ni}(\lambda)$ at $\lambda = \lambda_0$. Then a transformed set of equations is given by $\sum_{i=1}^n g_{ni}(\lambda_0)[h_{1ni}(\lambda_0) - \lambda h_{2ni}(\lambda_0)]$, which is only a linear equation in λ . Solve this to get λ_1 , plug in λ_1 in place of λ_0 and iterate this step. At the convergence of the iteration a numerical solution (say λ_∞) is obtained, which is treated as the solution to (7.1.1).

Our technique of computing *UBS* resampling estimates is as follows: store the values of $g_{ni}(\lambda)$, $h_{1ni}(\lambda)$ and $h_{2ni}(\lambda)$ at $\lambda = \lambda_\infty$ and treat these as fixed constants. Then a transformed bootstrap equation in place of (7.1.2) is

$$\sum_{i=1}^n w_{ni} g_{ni}(\lambda_\infty)[h_{1ni}(\lambda_\infty) - \lambda h_{2ni}(\lambda_\infty)],$$

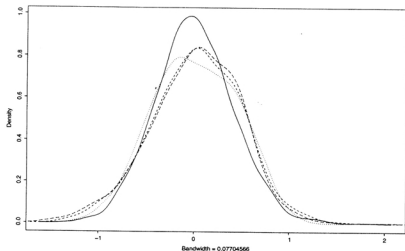
which is also linear in λ . Treat the solution to this as the bootstrap estimate. Thus once a numerical solution for the actual value is obtained, each bootstrap Monte Carlo step requires solving one linear equation only.

For the computation of the estimate by *IRLS* or Newton-Raphson numerical technique, the solution at each iteration is a weighted average, where the weights are also random and depend on the choice of the initial estimate. The choice of the initial estimate is often very important. The bootstrap may be carried either by the new technique suggested above, or by repeating the steps of the iteration. There are two reasons to expect that either method of bootstrap computation should give similar results. First, in case of using iteration for bootstrap, the natural initial estimate to take is λ_∞ , in which case our method is the one-step solution. Second, in the bootstrap world, notice that the functions $g_{ni}(\lambda)$, $h_{1ni}(\lambda)$ and $h_{2ni}(\lambda)$ are no longer random. One step solutions for resampling has been suggested earlier in different context by Shao (1992a,b), Lee (1990) and others.

We now present a simple simulation example to show the performance of the *UBS* techniques. We take the model $y_i = \beta_1 + \exp(-\beta_2 x_i) + e_i$, where $\beta_1 = 10$, $\beta_2 = 1.25$, x_i 's are

i.i.d. from $Uniform(0, 1)$, and e_i 's are *i.i.d.* from $Normal(0, 1)$. A sample of size $n = 30$ was taken, and (β_1, β_2) estimated by using least squared errors. This requires iteration, and in order to get an initial estimate of the parameter we used the following steps: we computed $y_0 = 0.9 \min_{1 \leq i \leq n} y_i$, then used ordinary least squares to regress $\log(y_i - y_0)$ on $-\beta_2 x_i$, to get an estimate β_{20} of β_2 . Then we took the average of $y_i - \exp(-\beta_{20} x_i)$ as β_{10} , the initial estimate of β_1 . The constant 0.9 used in the definition of y_0 was fixed arbitrarily.

Figure 7.1: Different bootstrap density estimators



In Figure 7.1 we plotted the density of the least squares estimator of β_2 , simulated from 1000 trials of the above experiment. By fixing one such dataset, we obtained different *UBS* estimators also for the same quantity. The bold line shows the density of the least squares estimator, the dashed line the bootstrap estimator using $Multinomial(n, 1/n, \dots, 1/n)$ weights, the double-dash line is the one step iteration free estimator described in this section for the same bootstrap weights; and the dotted line is *UBS* estimator with *i.i.d.* $Uniform(1/2, 3/2)$ weights. In each case a kernel smoothing with Gaussian kernel was used, and the common bandwidth used is quoted in the figure. The figure shows that all the three bootstrap methods yield equally good results.

A number of different models were tried for simulation. The broad conclusion is that the different choices of *UBS* weights yield similar results, but the performance of the non-iterative one-step method is variable.

7.7 Proofs

Proof of Theorem 7.3.1 Note that given any $\epsilon > 0$, $\exists K > 0$ and an integer n_0 such that for all $n \geq n_0$

$$\text{Prob}[|a_n^{-1} \sum_{i=1}^n \phi_{ni}| > K] < \epsilon/3 \quad (7.7.1)$$

This is easily established using Chebyshev's inequality, (A0) and (A3). Let

$$\gamma_{1n} = a_n^{-2} \sum_{i=1}^n \phi_{1ni}.$$

Next note that given any constants $\epsilon, C, K > 0$, for all n large enough,

$$\text{Prob}[\sup_{|t|=C} |S_n(t)| > K] < \epsilon/3 \quad \text{where} \quad (7.7.2)$$

$$S_n(t) = a_n^{-1} \sum_{i=1}^n [\phi_{ni}(\beta_0 + a_n^{-1}t) - \phi_{ni}(\beta_0)] - \gamma_{1n}t$$

Instead of (7.7.2), we actually prove the much stronger result

$$E[\sup_{|t| \leq C} |S_n(t)|^2] = O(a_n^{-2}) \quad (7.7.3)$$

By Taylor series expansion we have

$$\sum_{i=1}^n [\phi_{ni}(\beta_0 + a_n^{-1}t) - \phi_{ni}(\beta_0)] = a_n^{-1}t \sum_{i=1}^n \phi_{1ni}(\beta_0) + \frac{1}{2}a_n^{-2}t^2 \sum_{i=1}^n \phi_{2ni}(\beta_1)$$

where β_1 lies in between β_0 and $\beta_0 + a_n^{-1}t$. Then $S_n(t) = \frac{1}{2}a_n^{-3}t^2 \sum_{i=1}^n \phi_{2ni}(\beta_1)$. Thus we have for all large n , $\sup_{|t| \leq C} |S_n(t)| \leq 2^{-1}a_n^{-3}C^2 \sum_{i=1}^n M_{2ni}$ using (A2). Thus

$$\begin{aligned} E[\sup_{|t| \leq C} |S_n(t)|^2] &\leq 4^{-1}a_n^{-6}C^4 E(\sum_{i=1}^n M_{2ni})^2 \\ &= O(a_n^{-2}) \text{ using (A2)} \end{aligned}$$

This completes the proof of (7.7.3). Note that

$$\inf_{|t|=C} \{a_n^{-1}t \sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1}t)\} \geq -C \sup_{|t|=C} |S_n(t)| + C^2\gamma_{1n} - Ca_n^{-1}|\sum_{i=1}^n \phi_{ni}| \quad (7.7.4)$$

From (7.7.1), (7.7.2) and (7.7.4) we have, choosing C large enough,

$$\text{Prob}[\inf_{|t|=C} \{a_n^{-1}t \sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1}t)\} > 0]$$

$$\begin{aligned}
&\geq \text{Prob}[a_n^{-1} |\sum_{i=1}^n \phi_{ni}| + \sup_{|t|=C} |S_n(t)| \leq C\gamma_{1n}] \\
&= 1 - \text{Prob}[a_n^{-1} |\sum_{i=1}^n \phi_{ni}| + \sup_{|t|=C} |S_n(t)| > C\gamma_{1n}] \\
&\geq 1 - \text{Prob}[a_n^{-1} |\sum_{i=1}^n \phi_{ni}| > Ck_2/4] - \text{Prob}[\sup_{|t|=C} |S_n(t)| > Ck_2/4] \\
&\quad - \text{Prob}[a_n^{-2} |\sum (\phi_{1ni} - E\phi_{1ni})| > k_2/2] \\
&\geq 1 - \epsilon \text{ for all } n \text{ sufficiently large}
\end{aligned}$$

Using continuity of $\sum_{i=1}^n \phi_{ni}(\lambda)$ in λ , this means that for fixed $\epsilon > 0$ for all n sufficiently large $\exists C$ large such that

$$\sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1}t) = 0 \text{ has a root } t = T_n \text{ in } |t| \leq C \text{ with probability } > 1 - \epsilon \quad (7.7.5)$$

Putting $\hat{\beta}_n = \beta_0 + a_n^{-1}T_n$, we get a solution to (7.1.1) which satisfies, for fixed $\epsilon > 0$, $\text{Prob}[a_n|\hat{\beta}_n - \beta_0| \leq C] \geq 1 - \epsilon$ for all n large enough. This shows (7.3.1). Now notice that with this C fixed, we have actually obtained in (7.7.3) that with $t = T_n$

$$a_n\gamma_{1n}(\hat{\beta}_n - \beta_0) = -a_n^{-1} \sum_{i=1}^n \phi_{ni} + r_{n1}$$

where $Er_{n1}^2 = O(a_n^{-2})$. Now a little bit more algebra and conditions (A1) and (A3) yields (7.3.2). ■

Proof of Theorem 7.3.2 The techniques used in proving (7.3.4) and (7.3.5) are not much different from what was used to prove (7.3.1) and (7.3.2). First fix $\epsilon, \delta > 0$. Using Chebyshev's inequality, (7.2.1) and (7.2.3) it can be easily shown that $\hat{\gamma}_{1n}^{-1} = O_P(1)$ and hence for some constant k ,

$$\text{P}_B[|a_n^{-1} \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)| > C\hat{\gamma}_{1n}\sigma_n/2] \leq \frac{k\hat{g}_n}{C^2\hat{\gamma}_{1n}^2} (= U_{1C} \text{ say}) \quad (7.7.6)$$

Thus this bootstrap probability is bounded by a random variable depending on C , which we call U_{1C} . By choosing C large enough and using some algebra, it can be shown that $\text{Prob}[U_{1C} > \epsilon/2] < \delta/2$. Fix such a C and define

$$S_{nB}(t) = a_n^{-1} \sum_{i=1}^n w_i [\phi_{ni}(\hat{\beta}_n + a_n^{-1}t) - \phi_{ni}(\hat{\beta}_n)] - \hat{\gamma}_{1n}t$$

Using Taylor series expansion, for some $\hat{\beta}_{n1}$ between $\hat{\beta}_n$ and $\hat{\beta}_n + a_n^{-1}t$, we have

$$\begin{aligned}
& S_{nB}(t) \\
&= a_n^{-2} t \sum_{i=1}^n w_i \phi_{1ni}(\hat{\beta}_n) - \hat{\gamma}_{1n} t + \frac{1}{2} a_n^{-3} t^2 \sum_{i=1}^n w_i \phi_{2ni}(\hat{\beta}_{n1}) \\
&= \sigma_n a_n^{-2} t \sum_{i=1}^n W_i \phi_{1ni}(\hat{\beta}_n) + \frac{1}{2} a_n^{-3} t^2 \sum_{i=1}^n w_i \phi_{2ni}(\hat{\beta}_{n1}) \\
&= S_{nB1}(t) + S_{nB2}(t) \text{ say}
\end{aligned}$$

Consequently,

$$\begin{aligned}
& P_B \left[\sup_{|t|=C\sigma_n} |S_{nB}(t)| > C\hat{\gamma}_{1n}\sigma_n/2 \right] \\
&\leq P_B \left[\sup_{|t|=C\sigma_n} |S_{nB1}(t)| > C\hat{\gamma}_{1n}\sigma_n/4 \right] + P_B \left[\sup_{|t|=C\sigma_n} |S_{nB2}(t)| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \\
&= U_{2C} \text{ say}
\end{aligned} \tag{7.7.7}$$

Now

$$\begin{aligned}
& P_B \left[\sup_{|t|=C\sigma_n} |S_{nB1}(t)| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \\
&= P_B \left[\sigma_n^2 a_n^{-2} C \left| \sum_{i=1}^n W_i \phi_{1ni}(\hat{\beta}_n) \right| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \\
&\leq \frac{k\sigma_n^2 \sum \phi_{1ni}^2(\hat{\beta}_n)}{a_n^4 \hat{\gamma}_{1n}^2} \\
&= O_P(\sigma_n^2 a_n^{-2}) \text{ using (A1), (A2) and (A3)} \\
&= o_P(1)
\end{aligned}$$

$$\begin{aligned}
& P_B \left[\sup_{|t|=C\sigma_n} |S_{nB2}(t)| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \\
&= P_B \left[\sup_{|t|=C\sigma_n} \left| \frac{1}{2} a_n^{-3} t^2 \sum_{i=1}^n w_i \phi_{2ni}(\hat{\beta}_{n1}) \right| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \\
&\leq P_B \left[\frac{1}{2} a_n^{-3} \sup_{|t|=C\sigma_n} t^2 \sum_{i=1}^n w_i |\phi_{2ni}(\hat{\beta}_{n1})| > C\hat{\gamma}_{1n}\sigma_n/4 \right] \text{ since } w_i \geq 0 \text{ for all } i \\
&= P_B \left[\frac{1}{2} a_n^{-3} \sup_{|t|=C\sigma_n} t^2 \sum_{i=1}^n w_i |\phi_{2ni}(\hat{\beta}_{n1})| > C\hat{\gamma}_{1n}\sigma_n/4 \right] (I_{\{|\hat{\beta}-\beta_0| < \delta_0/2\}} + I_{\{|\hat{\beta}-\beta_0| \geq \delta_0/2\}}) \\
&\leq P_B \left[\frac{1}{2} a_n^{-3} \sup_{|t|=C\sigma_n} t^2 \sum_{i=1}^n w_i |\phi_{2ni}(\hat{\beta}_{n1})| > C\hat{\gamma}_{1n}\sigma_n/4 \right] I_{\{|\hat{\beta}-\beta_0| < \delta_0/2\}} + I_{\{|\hat{\beta}-\beta_0| \geq \delta_0/2\}} \\
&\leq P_B \left[\frac{1}{2} a_n^{-3} \sup_{|t|=C\sigma_n} t^2 \sum_{i=1}^n w_i |\phi_{2ni}(\hat{\beta}_{n1})| > C\hat{\gamma}_{1n}\sigma_n/4 \right] I_{\{|\hat{\beta}-\beta_0| < \delta_0/2\}} + O_P(a_n^{-1})
\end{aligned}$$

$$\begin{aligned}
&\leq P_B[\frac{1}{2}a_n^{-3} \sup_{|t|=C\sigma_n} t^2 \sum_{i=1}^n w_i M_{2ni} > C\hat{\gamma}_{1n}\sigma_n/4] + O_P(a_n^{-1}) \text{ using (A2)} \\
&\leq P_B[\frac{1}{2}a_n^{-3} C^2 \sigma_n^2 \sum_{i=1}^n w_i M_{2ni} > C\hat{\gamma}_{1n}\sigma_n/4] + O_P(a_n^{-1}) \\
&\leq 2a_n^{-3} C \sigma_n \hat{\gamma}_{1n} \sum_{i=1}^n M_{2ni} + O_P(a_n^{-1}) \\
&= O_P(\sigma_n a_n^{-3} c_n) + O_P(a_n^{-1}) \text{ using (A2) and (A3)} \\
&= o_P(1)
\end{aligned}$$

So for all n large enough, $Prob[U_{2C} > \epsilon/2] < \delta/2$. Also, as in (7.7.4),

$$\begin{aligned}
&\inf_{|t|=C\sigma_n} \{a_n^{-1}t \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n + a_n^{-1}t)\} \\
&\geq -C\sigma_n \sup_{|t|=C\sigma_n} |S_{nB}(t)| + C^2 \sigma_n^2 \hat{\gamma}_{1n} - C\sigma_n |a_n^{-1} \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)|
\end{aligned} \tag{7.7.8}$$

Thus

$$\begin{aligned}
&P_B[\inf_{|t|=C\sigma_n} \{a_n^{-1}t \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n + a_n^{-1}t)\} > 0] \\
&\geq P_B[|a_n^{-1} \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)| + \sup_{|t|=C\sigma_n} |S_{nB}(t)| \leq C\hat{\gamma}_{1n}\sigma_n] \\
&= 1 - P_B[|a_n^{-1} \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)| + \sup_{|t|=C\sigma_n} |S_{nB}(t)| > C\hat{\gamma}_{1n}\sigma_n] \\
&\geq 1 - P_B[|a_n^{-1} \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)| > C\hat{\gamma}_{1n}\sigma_n/2] - P_B[\inf_{|t|=C\sigma_n} |S_{nB}(t)| > C\hat{\gamma}_{1n}\sigma_n/2] \\
&= 1 - U_{1C} - U_{2C}
\end{aligned}$$

This establishes (7.3.4), since $Prob[U_{iC} > \epsilon/2] < \delta/2$, $i = 1, 2$ from 7.7.6) and (7.7.7). By modifying slightly the above argument it is easy to establish that

$$\sup_{|t| \leq C\sigma_n} |S_{nB}(t)| = \sigma_n r_{nB1} \tag{7.7.9}$$

where $P_B(a_n |r_{nB1}| > \epsilon) = O_P(1)$ for any $\epsilon > 0$, then it follows that

$$\hat{\gamma}_{1n} \sigma_n^{-1} a_n (\hat{\beta}_B - \hat{\beta}_n) = -a_n^{-1} \sum_{i=1}^n W_i \phi_{ni}(\hat{\beta}_n) + O_{PB}(a_n^{-1})$$

With some more algebra, (7.3.5) follows. ■

In order to prove Corollary 7.3.2, we need Lemma 4.6 of Praetgaard and Wellner (1993), which is quoted as Theorem 5.3.1 in Chapter 5.

Proof of Corollary 7.3.2 It can be checked that with $a_{mj} = -n^{1/2}a_n^{-1}s_n-1\phi_{ni}$, the conditions of Theorem 5.3.1 is satisfied with $\tau^2 = 1$, $c^2 = 1$. Using Lemma 3.1 and 4.7 of Praestgaard and Wellner (1993), the conditions of Theorem 5.3.1 are verified.

As in Chapter 5, in place of Theorem 5.3.1, slight variation of the model conditions and assumptions on resampling weights allows us to use Theorem 3.3 of Arenal-Gutierrez and Matran. Under some circumstances this is a more general result. ■

Proof of Theorem 7.3.3 The details of the algebra involved in this proof is similar to those of Theorems 7.3.1 and 7.3.2, and we omit many of the details.

Let us concentrate on the set $\mathcal{A} \cap \mathcal{W}$ only, since the contribution from the complement of this set is negligible. Under the conditions of the theorem, we have that

$$ng_{1n}^2 V_n = n^{-1} \sum E\phi^2 + O(n^{-1})$$

Define

$$\begin{aligned} U_{nB}(t) &= \sigma_n^{-1} n^{-1/2} \sum_{i=1}^n w_i [\phi_{ni}(\hat{\beta}_n + \sigma_n n^{-1/2} t) - \phi_{ni}(\hat{\beta}_n)] \\ &\quad - n^{-1} t \sum_{i=1}^n w_i \phi_{1ni}(\hat{\beta}_n) - 2^{-1} \sigma_n n^{-3/2} t^2 \sum_{i=1}^n w_i \phi_{2ni}(\hat{\beta}_n) \end{aligned}$$

Using (7.2.26), and proceeding along similar lines as with $S_{nB}(t)$ in the proof of Theorem 7.3.2, we can show that

$$E_B \left[\sup_{|t| \leq C\sigma_n} |U_{nB}(t)| \right]^2 = O_P(n^{-2})$$

Now, under $\mathcal{A} \cap \mathcal{W}$, we may plug in $t = \sigma_n^{-1} n^{1/2} (\hat{\beta}_B - \hat{\beta})$ in $U_{nB}(t)$, and after quite some algebra we arrive at

$$\begin{aligned} &-g_{1n} \sigma_n^{-1} n^{1/2} (\hat{\beta}_B - \hat{\beta}) \\ &= n^{-1/2} \sum W_i \phi_{ni} - n^{-2} g_{1n}^{-1} \sum \phi_{ni} \sum W_i \phi_{1ni} \\ &\quad - n^{-2} g_{1n}^{-1} \sum (\phi_{1ni} - E\phi_{1ni}) \sum W_i \phi_{ni} + n^{-5/2} g_{1n}^{-2} \sum \phi_{ni} \sum \phi_{2ni} \sum W_i \phi_{ni} \\ &\quad + \sigma_n n^{-3/2} g_{1n}^{-1} \sum W_i \phi_{ni} \sum W_i \phi_{1ni} + 2^{-1} \sigma_n n^{-3/2} g_{2n} g_{1n}^{-2} (\sum W_i \phi_{ni})^2 + R_{nB} \\ &= C_n + T_{1n} + T_{2n} + T_{3n} + T_{4n} + T_{5n} + R_{nB} \text{ say} \end{aligned}$$

where $E_B R_{nB}^2 = O_P(n^{-(1+\alpha)})$. This derivation follows from obtaining Taylor series expansion about β_0 for quantities involving $\hat{\beta}_n$ and $\hat{\beta}_B$, then using the moment properties and asymptotic expansion of $\hat{\beta}_n - \beta_0$.

Now it can be easily checked that $E_B C_n^2 = O_P(1)$, and $E_B T_{in}^2 = O_P(n^{-1})$, for $i = 1, \dots, 5$. In the cross product, by direct computation $E_B C_n T_{in} = O_P(n^{-1})$ for $i = 4, 5$, and hence

$$ng_{1n}^2 V_{UBS} = E_B C_n^2 + 2E_B C_n (T_{1n} + T_{2n} + T_{3n}) + O_P(n^{-1})$$

Now the rest of the proof follows by calculating the above moments. ■

Proof of Theorem 7.3.5 From (7.2.37), one can easily show $\sum_{i=1}^n E(\|\phi_{ni}\|^2) = O(a_n^2 p)$, by using any orthonormal basis of $\mathbb{R}^p$. Then we show that given any $\epsilon > 0$, $\exists K > 0$ and an integer n_0 such that for all $n \geq n_0$

$$Prob[\|a_n^{-1} p^{-1/2} \sum_{i=1}^n \phi_{ni}\| > K] < \epsilon/2 \quad (7.7.10)$$

This is easily established using Chebyshev's inequality. Let

$$S_n(t) = a_n^{-1} p^{-1/2} \sum_{i=1}^n [\phi_{ni}(\beta_0 + a_n^{-1} p^{1/2} t) - \phi_{ni}(\beta_0)] - G_{1n}^T t$$

Let $M_{1n} = \sum_{i,j=1}^n \sum_{a=1}^p (\phi_{1nia} - E\phi_{1nia})(\phi_{1nja} - E\phi_{1nja})^T$. Then $E\lambda_{\max}(M_{1n}) = O(a_n^2 p^2)$ follows from (B1). Also, (B3) ensures that G_{1n} is non-singular. Because of (7.2.35), using (B1) and (B2) we have that

$$\|S_n(t)\|^2 \leq 2a_n^{-4} \|t\|^2 \lambda_{\max}(M_{1n}) + 2^{-1} a_n^{-6} p \|t\|^4 \sum_{i,j=1}^n \sum_{a=1}^p \lambda_{\max}(M_{2nia}) \lambda_{\max}(M_{2nja})$$

from which we have that

$$[\sup_{\|t\| \leq C} \|S_n(t)\|]^2 \leq 2a_n^{-4} C^2 \lambda_{\max}(M_{1n}) + 2^{-1} a_n^{-6} p C^4 \sum_{i,j=1}^n \sum_{a=1}^p \lambda_{\max}(M_{2nia}) \lambda_{\max}(M_{2nja})$$

Now using (B1) and (B2) we have that $E[\sup_{\|t\| \leq C} \|S_n(t)\|]^2 = O(a_n^{-2} p^2)$. Note that

$$\inf_{|t|=C} \{a_n^{-1} p^{-1/2} t^T \sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1} p^{1/2} t)\} \geq -C \sup_{|t|=C} \|S_n(t)\| + C^2 l_{1n} - C a_n^{-1} p^{-1/2} \|\sum_{i=1}^n \phi_{ni}\|$$

where $l_{1n} = \lambda_{\min}(2^{-1}(G_{1n} + G_{1n}^T))$. By choosing C large enough, from the previous calculations we have that

$$\begin{aligned} & Prob[\inf_{|t|=C} \{a_n^{-1} p^{-1/2} t^T \sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1} p^{1/2} t)\} > 0] \\ & \geq Prob[a_n^{-1} p^{-1/2} \|\sum_{i=1}^n \phi_{ni}\| + \sup_{|t|=C} \|S_n(t)\| < C l_{1n}] \\ & = 1 - Prob[a_n^{-1} p^{-1/2} \|\sum_{i=1}^n \phi_{ni}\| + \sup_{|t|=C} \|S_n(t)\| > C l_{1n}] \\ & \geq 1 - Prob[a_n^{-1} p^{-1/2} \|\sum_{i=1}^n \phi_{ni}\| > C k_2/2] - Prob[\sup_{|t|=C} \|S_n(t)\| > C k_2/2] \\ & \geq 1 - \epsilon \text{ for all } n \text{ sufficiently large} \end{aligned}$$

Using continuity of $\sum_{i=1}^n \phi_{ni}(\lambda)$ in λ , by Theorem 6.4.3 of Ortega and Rheinboldt (1970) this means that for fixed $\epsilon > 0$ for all n sufficiently large $\exists C$ large such that

$$\sum_{i=1}^n \phi_{ni}(\beta_0 + a_n^{-1} p^{1/2} t) = 0 \text{ has a root } T_n \text{ in } |t| \leq C \text{ with probability } > 1 - \epsilon \quad (7.7.11)$$

Putting $\hat{\beta}_n = \beta_0 + a_n^{-1} p^{1/2} T_n$, we get a solution to (7.1.1) which satisfies, for fixed $\epsilon > 0$, $\text{Prob}[a_n p^{-1/2} \|\hat{\beta}_n - \beta_0\| \leq C] \geq 1 - \epsilon$ for all n large enough. This shows (7.3.11). Now notice that with this C fixed, we have actually shown that with $t = T_n$

$$a_n p^{-1/2} G_{1n}(\hat{\beta}_n - \beta_0) = -a_n^{-1} p^{-1/2} \sum_{i=1}^n \phi_{ni} + r_{n1}$$

where $E r_{n1}^2 = O(a_n^{-2} p^2)$. The representations (7.3.14) now follows easily, and (7.3.15) is obtained by slightly modifying the above steps. ■

Corollary 7.3.3 is immediate.

Proof of Theorem 7.3.6 We prove (7.3.17) and (7.3.21) first.

Since $\sum_{i=1}^n \phi_{ni}(\hat{\beta}_n) = 0$, we have for some constant $k > 0$

$$\begin{aligned} & P_B[\sigma_n^{-1} p^{-1/2} a_n^{-1} \|\sum w_i \phi_{ni}(\hat{\beta}_n)\| > K] \\ &= P_B[p^{-1/2} a_n^{-1} \|\sum W_i \phi_{ni}(\hat{\beta}_n)\| > K] \\ &\leq k K^{-2} p^{-1} a_n^{-2} \sum \|\phi_{ni}(\hat{\beta}_n)\|^2 \end{aligned}$$

Using (7.2.35), we have that

$$\phi_{nia}(\hat{\beta}_n + t) = \phi_{nia} + \phi_{1nia}^T t + 2^{-1} t^T H_{2nia}(\beta_2) t$$

for some $\beta_2 = \beta_0 + c(\hat{\beta}_n - \beta_0)$. On the set $\{\|\hat{\beta}_n - \beta_0\| < \delta_0/2\}$, with a little algebra, this leads to

$$\begin{aligned} \sum_{i=1}^n \|\phi_{ni}(\hat{\beta}_n)\|^2 &\leq k \left[\sum_{i=1}^n \|\phi_{ni}\|^2 + \|\hat{\beta}_n - \beta_0\|^2 \sum_{i=1}^n \sum_{a=1}^p \|\phi_{1nia}\|^2 \right. \\ &\quad \left. + \|\hat{\beta}_n - \beta_0\|^4 \sum_{i=1}^n \sum_{a=1}^p \lambda_{max}^2(M_{2nia}) \right] \end{aligned}$$

Thus

$$\begin{aligned} & P_B[\sigma_n^{-1} p^{-1/2} a_n^{-1} \|\sum w_i \phi_{ni}(\hat{\beta}_n)\| > K] \\ &\leq k K^{-2} p^{-1} a_n^{-2} \left[\sum_{i=1}^n \|\phi_{ni}\|^2 + \|\hat{\beta}_n - \beta_0\|^2 \sum_{i=1}^n \sum_{a=1}^p \|\phi_{1nia}\|^2 \right. \\ &\quad \left. + \|\hat{\beta}_n - \beta_0\|^4 \sum_{i=1}^n \sum_{a=1}^p \lambda_{max}^2(M_{2nia}) \right] \\ &= U_K \end{aligned}$$

Note that $U_K = K^{-2}O_P(1)$, so that for fixed $\delta_1, \delta_2 > 0$, by choosing K large enough we have,

$$Prob[P_B[\sigma_n^{-1}p^{-1/2}a_n^{-1}||\sum w_i\phi_{ni}(\hat{\beta}_n)|| > K] > \delta_1] < \delta_2$$

From (7.2.35) we have that

$$\phi_{nia}(\hat{\beta}_n + \sigma_n p^{1/2} a_n^{-1} t) = \phi_{nia}(\hat{\beta}_n) + \sigma_n p^{1/2} a_n^{-1} \phi_{1nia}^T t + 2^{-1} \sigma_n^2 p a_n^{-2} t^T H_{2nia}(\beta_1) t$$

where $\beta_1 = \hat{\beta}_n + cp^{1/2}a_n^{-1}t$ for some $0 < c < 1$. Let

$$S_{nB}(t) = \sigma_n^{-1} p^{-1/2} a_n^{-1} \sum_{i=1}^n w_i [\phi_{ni}(\hat{\beta}_n + \sigma_n p^{1/2} a_n^{-1} t) - \phi_{ni}(\hat{\beta}_n)] - \Gamma_{1n}^T(\hat{\beta}_n) t$$

We have the a^{th} element of the vector $S_{nB}(t)$ given by

$$\begin{aligned} S_{nBa}(t) &= a_n^{-2} \sum w_i \phi_{1nia}^T(\hat{\beta}_n) t + 2^{-1} \sigma_n p^{1/2} a_n^{-3} t^T \sum w_i H_{2nia}(\beta_1) t - \Gamma_{1n}^T(\hat{\beta}_n) t \\ &= a_n^{-2} \sigma_n \sum W_i \phi_{1nia}^T(\hat{\beta}_n) t + 2^{-1} \sigma_n p^{1/2} a_n^{-3} t^T \sum w_i H_{2nia}(\beta_1) t \end{aligned}$$

On the set $\{||t|| \leq C\} \cap \{||\hat{\beta}_n - \beta_0|| < \delta_0/2\}$, we thus have for large n

$$\begin{aligned} &||S_{nB}(t)||^2 \\ &\leq 2\sigma_n^2 a_n^{-4} C^2 \lambda_{max} \left(\sum_{a=1}^p \sum_{i,j=1}^n W_i W_j \phi_{1nia}(\hat{\beta}_n) \phi_{1nja}(\hat{\beta}_n)^T \right) \\ &\quad + \sigma_n^2 p a_n^{-6} C^4 \sum_{a=1}^p \left(\sum_{i=1}^n w_i \lambda_{max}(M_{2nia}) \right)^2 \\ &= T_1 + T_2 \text{ say} \end{aligned}$$

Define $b_{n,p} = \sigma_n^{-1} p^{-1} a_n$. Thus we have

$$\begin{aligned} &P_B[b_{n,p} \sup_{||t|| \leq C} ||S_{nB}(t)|| > 2K] \\ &\leq \sum_{j=1}^2 P_B[b_{n,p}^2 T_j > K] + O_P(a_n^{-1} p^{1/2}) \end{aligned}$$

Now on the set $\{||\hat{\beta}_n - \beta_0|| < \delta_0/2\}$,

$$\begin{aligned} &P_B[2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 \lambda_{max} \left(\sum_{a=1}^p \sum_{i,j=1}^n W_i W_j \phi_{1nia}(\hat{\beta}_n) \phi_{1nja}(\hat{\beta}_n)^T \right) > K] \\ &\leq 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} E_B \lambda_{max} \left(\sum_{a=1}^p \sum_{i,j=1}^n W_i W_j \phi_{1nia}(\hat{\beta}_n) \phi_{1nja}(\hat{\beta}_n)^T \right) \\ &\leq 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} \sum_{a=1}^p \sum_{i,j=1}^n tr(E_B W_i W_j \phi_{1nia}(\hat{\beta}_n) \phi_{1nja}(\hat{\beta}_n)^T) \end{aligned}$$

$$\begin{aligned}
&\leq 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} \sum_{a=1}^p \sum_{i=1}^n \text{tr}(\phi_{1nia}(\hat{\beta}_n) \phi_{1nia}(\hat{\beta}_n)^T) \\
&\quad + 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} c_{11} \text{tr}(\sum_{a=1}^p \sum_{i,j=1, i \neq j}^n \phi_{1nia}(\hat{\beta}_n) \phi_{1nja}(\hat{\beta}_n)^T) \\
&\leq 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} \sum_{a=1}^p \sum_{i=1}^n \|\phi_{1nia}(\hat{\beta}_n)\|^2 \\
&\quad + 2\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} c_{11} \sum_{a=1}^p \sum_{i,j=1, i \neq j}^n \|\phi_{1nia}(\hat{\beta}_n)\| \|\phi_{1nja}(\hat{\beta}_n)\| \\
&\leq k\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} \sum_{a=1}^p \sum_{i=1}^n \|\phi_{1nia}(\hat{\beta}_n)\|^2 \\
&\leq k\sigma_n^2 b_{n,p}^2 a_n^{-4} C^2 K^{-1} \sum_{a=1}^p \sum_{i=1}^n [\|\phi_{1nia}\|^2 + \|\hat{\beta}_n - \beta_0\|^2 \lambda_{\max}^2(M_{2nia})] \\
&= O_P(1)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_B[\sigma_n^2 p b_{n,p}^2 a_n^{-6} C^4 \sum_{a=1}^p (\sum_{i=1}^n w_i \lambda_{\max}(M_{2nia}))^2 > K] \\
&= \sigma_n^2 p b_{n,p}^2 a_n^{-6} C^4 K^{-1} \sum_{a=1}^p \mathbb{E}_B(\sum_{i=1}^n w_i \lambda_{\max}(M_{2nia}))^2 \\
&= \sigma_n^2 p b_{n,p}^2 a_n^{-6} C^4 K^{-1} (n + k\sigma_n^2) \sum_{a=1}^p \sum_{i=1}^n \lambda_{\max}^2(M_{2nia}) \\
&= O_P(1)
\end{aligned}$$

Thus we have that $\mathbb{P}_B[b_{n,p} \sup_{\|t\| \leq C} \|S_{nB}(t)\| > 2K] = O_P(1)$.

Now observe that

$$\begin{aligned}
&\inf_{|t|=C} \{\sigma_n^{-1} a_n^{-1} p^{-1/2} t^T \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n + \sigma_n a_n^{-1} p^{1/2} t)\} \\
&\geq -C \sup_{|t|=C} \|S_n(t)\| + C^2 \hat{l}_{1n} - C \sigma_n^{-1} a_n^{-1} p^{-1/2} \|\sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n)\|
\end{aligned}$$

where $\hat{l}_{1n} = \lambda_{\min}(2^{-1}(\Gamma_{1n}(\hat{\beta}_n) + \Gamma_{1n}^T(\hat{\beta}_n)))$. By choosing C large enough, from the previous calculations we have that on the set $\{\|\hat{\beta}_n - \beta_0\| < \delta_0/2\}$,

$$\begin{aligned}
&\mathbb{P}_B[\inf_{|t|=C} \{\sigma_n^{-1} a_n^{-1} p^{-1/2} t^T \sum_{i=1}^n w_i \phi_{ni}(\hat{\beta}_n + \sigma_n a_n^{-1} p^{1/2} t)\} > 0] \\
&\geq \mathbb{P}_B[\sigma_n^{-1} a_n^{-1} p^{-1/2} \|\sum_{i=1}^n \phi_{ni}(\hat{\beta}_n)\| + \sup_{|t|=C} \|S_n(t)\| < C \hat{l}_{1n}]
\end{aligned}$$

$$\begin{aligned}
&= 1 - \mathbb{P}_B[a_n^{-1}p^{-1/2}|\sum_{i=1}^n \phi_{ni}| + \sup_{|t|=C} ||S_n(t)|| > C\hat{l}_{1n}] \\
&\geq 1 - \mathbb{P}_B[\sigma_n^{-1}a_n^{-1}p^{-1/2}|\sum_{i=1}^n w_i\phi_{ni}| > Ck_2/2] - \mathbb{P}_B[\sup_{|t|=C} ||S_n(t)|| > Ck_2/2]
\end{aligned}$$

Thus for fixed $\delta_1, \delta_2 > 0$, we have that for C large enough for all large n ,

$$\begin{aligned}
&\text{Prob}[\mathbb{P}_B[\inf_{|t|=C} \{\sigma_n^{-1}a_n^{-1}p^{-1/2}t^T \sum_{i=1}^n w_i\phi_{ni}(\hat{\beta}_n + \sigma_n a_n^{-1}p^{1/2}t)\} > 0] < 1 - \delta_1] \\
&\leq \text{Prob}[\mathbb{P}_B[\sigma_n^{-1}a_n^{-1}p^{-1/2}|\sum_{i=1}^n w_i\phi_{ni}| > Ck_2/2] > \delta_1/2] \\
&\quad + \text{Prob}[\mathbb{P}_B[\sup_{|t|=C} ||S_n(t)|| > Ck_2/2] > \delta_1/2] + O(a_n^{-1}p^{1/2}) \\
&\leq \delta_2
\end{aligned}$$

Using continuity of $\sum_{i=1}^n w_i\phi_{ni}(\lambda)$ in λ , by Theorem 6.4.3 of Ortega and Rheinboldt (1970) this means that for fixed $\epsilon, \delta > 0$ for all n sufficiently large $\exists C$ large such that the bootstrap probability that $\sum_{i=1}^n w_i\phi_{ni}(\hat{\beta}_n + \sigma_n a_n^{-1}p^{1/2}t) = 0$ has a root T_n in $|t| \leq C$ is $< 1 - \epsilon$ with a probability $< \delta$. Putting $\hat{\beta}_B = \hat{\beta}_n + \sigma_n a_n^{-1}p^{1/2}T_n$, we get a solution to (7.1.2) which satisfies, for fixed $\epsilon, \delta > 0$, $\text{Prob}[\mathbb{P}_B[\sigma_n^{-1}a_n p^{-1/2}||\hat{\beta}_B - \hat{\beta}_n|| \leq C] < 1 - \epsilon] < \delta$ for all n large enough. This shows (7.3.17). Now notice that with this C fixed, we have actually shown that with $t = T_n$

$$\sigma_n^{-1}a_n p^{-1/2}\Gamma_{1n}(\hat{\beta}_n)(\hat{\beta}_B - \hat{\beta}_n) = -a_n^{-1}p^{-1/2} \sum_{i=1}^n W_i\phi_{ni}(\hat{\beta}_n) + r_{nB1}$$

where $||r_{nB1}|| = O_{PB}(\sigma_n a_n^{-1}p)$. This shows (7.3.21). With slight modification to the definition $S_{nB}(t)$, and some more calculations one can obtain

$$\sigma_n^{-1}a_n p^{-1/2}G_{1n}(\hat{\beta}_B - \hat{\beta}_n) = -a_n^{-1}p^{-1/2} \sum_{i=1}^n W_i\phi_{ni}(\hat{\beta}_n) + r_{nB1}$$

where $||r_{nB1}|| = O_{PB}(p^{-1})$, without using (B4). Now using (7.2.35) again, after some algebra, (7.3.18) is obtained. The representation (7.3.20) follows easily from this. ■

Chapter 8

GENERALISED BOOTSTRAP FOR MINIMUM CONTRAST (M_m) ESTIMATORS

8.1 Introduction

In Chapter 7 we have discussed *UBS* resampling technique in estimating equations. The results for that chapter can also be adapted for estimators obtained by minimising functionals, as long as the functionals are smooth in the parameter. However, there are many interesting examples where estimators are obtained by minimising non-smooth functionals. One way to compensate for the lack of smoothness is to assume that the function to minimise is convex in the parameter. By exploiting the properties of convex functions and some smoothness in expectations, asymptotic properties of these estimators may be established. Recent references in this direction include Haberman (1989), Niemiro (1992), and Bose (1998). In this chapter we study *UBS* resampling in such cases, where the estimator is obtained by minimising convex functionals.

Let $Z_1, \dots, Z_m$ be m independent and identically distributed copies of a $\mathbf{Z}$ valued random variable where $\mathbf{Z}$ is an arbitrary measure space and let $f(a, z)$ be a real measurable function defined for $a \in \mathbb{R}^d$, $z \in \mathbf{Z}^m$. Consider the function

$$Q(a) = Ef(a, Z_1, \dots, Z_m) \quad (8.1.1)$$

Let $a_* \in \mathbb{R}^d$ be such that

$$Q(a_*) = \min_a Q(a) \quad (8.1.2)$$

and suppose a_* is the unknown parameter, assumed to be unique, that is to be estimated from the data. If the probability distribution of Z_1 is unknown but an *i.i.d.* sample $Z_1, \dots, Z_n$ is available, then we can consider a sample analog of (8.1.1), namely,

$$Q_n(a) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} f(a, Z_{i_1}, Z_{i_2}, \dots, Z_{i_m}) \quad (8.1.3)$$

and minimise $Q_n(a)$ instead of $Q(a)$. Denote by a_n such a point depending on the sample that

$$Q(a_n) = \min_a Q_n(a) \quad (8.1.4)$$

The statistic a_n , which can be measurably (but not necessarily uniquely) selected, is a minimum contrast estimator a_n of the parameter a_* .

The maximum likelihood estimator obtained from *i.i.d.* observations when the likelihood is log-concave is an example. Other examples include (i) Oja median (Oja (1983)), (ii) univariate location estimators of Maritz et. al. (1977), (iii) univariate Hodges-Lehmann estimators of location, (iv) a univariate robust scale estimator of Bickel and Lehmann (1979), (v) a regression coefficient estimator of Theil (see Hollander and Wolfe (1973)) (vi) U quantiles (Chowdhury and Serfling (1988)) (vii) L_1 median (viii) geometric quantiles of Chaudhuri(1996) (ix) Hodges-Lehmann versions of (vii) and (viii), (x) LAD regression estimator and (xi) Haldane's spatial median.

The aim in this chapter is to study resampling estimates of a_n . The resampling in the present set-up is slightly different from the method used earlier. It turns out that UBS resampling for this situation must involve weights that are indexed by m -tuple vectors. In Section 8.3 we consider the special case of $m = 1$ and prove an asymptotic representation of the bootstrap estimator similar to Chapter 7. Consistency results for the UBS techniques follow from this representation. The general m case is discussed in Section 8.4 and similar results are obtained. We then discuss how the traditional bootstrap and jackknife relate to UBS in this special situation. The broad generality of the resampling schemes allows us to consider many non-traditional forms of resampling also. We discuss one such form which has the dual advantage of being computationally faster and less space consuming than traditional techniques, as well as being simpler mathematically. In Section 8.5 we compare the new and traditional forms of resampling through some simulation examples. Our last example in Section 8.5 is that of a real life data problem, which had been analysed earlier. We show how the methods of this chapter lead to a different and simpler approach to this dataset.

This chapter is based on Chatterjee and Bose (1999d).

8.2 Preliminaries

Let $g(a, z)$ be a subgradient of $f(a, z)$, that is

$$f(a, z) + (b - a)^T g(a, z) \leq f(b, z) \quad (8.2.1)$$

hold for all $a, b \in \mathbb{R}^d$, $z \in \mathbb{Z}^m$. Let $D(a)$ and $D^2(a)$ be the gradient and second derivative of $Q(a)$. Let $H = D^2(a_*)$. We make the following assumptions.

(MC1) $f(a, z)$ is convex with respect to a for each fixed z .

(MC2) $Q(a)$ is well defined, that is, the expectation in (8.1.1) exists and is finite for all a .

(MC3) a_* satisfying (8.1.2) exists and is unique.

(MC4) $E\|g(a, Z_1, \dots, Z_m)\|^2 < \infty$ for all a in a neighbourhood of a_* .

(MC5) $Q(a)$ is twice differentiable at a_* and H is positive definite.

The second assumption may hold only for some subset of $\mathbb{R}^d$, then all the results discussed here are valid for points in the interior of that subset. The above assumptions correspond respectively to conditions (i) – (iv) and (vi) of Bose (1998). In this set-up, Haberman (1989) and Niemiro (1992) (for $m = 1$), and Bose (1998) (for general m) have proved

Theorem 8.2.1 (Haberman (1989), Niemiro (1992), Bose (1998)) *Under the model assumptions (MC1)-(MC5),*

$$n^{1/2}(a_n - a_*) = -n^{1/2}H^{-1}S_n + o_P(1). \quad (8.2.2)$$

where $S_n = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} g(a, Z_{i_1}, Z_{i_2}, \dots, Z_{i_m})$.

From the above representation, asymptotic normality of the minimum contrast estimator easily follows. However, since typically the asymptotic variance of the U -statistic S_n will not be known or easily computable, resampling estimates for the distribution of a_n or its asymptotic variance is of interest.

We now explain the resampling scheme. Consider the triangular sequence of real valued non-negative random variables $\{w_{n:i_1 i_2 \dots i_m}, i_1, i_2, \dots, i_m \text{ distinct}, i_j \in \{1, 2, \dots, n\}; j = 1, \dots, m, n = 1, 2, \dots\}$. We use these as ‘bootstrap weights’. We discuss conditions on these weights later on. The bootstrap equivalent of a_n is obtained by minimizing

$$Q_{nB}(a) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} w_{n:i_1 i_2 \dots i_m} f(a, Z_{i_1}, Z_{i_2}, \dots, Z_{i_m}) \quad (8.2.3)$$

Let $\{a_{nB}\}$ be chosen in a measurable way satisfying

$$Q_{nB}(a_{nB}) = \min_a Q_{nB}(a). \quad (8.2.4)$$

We establish an asymptotic representation of a_{nB} , similar to (8.2.2) in Theorems 8.3.1 and 8.4.1. This allows us to prove consistency of the UBS resampling schemes for estimating the distribution of a_n or its asymptotic variance. We then discuss how to make a choice of the weights $w_{n:i_1 \dots i_m}$. Although our results are discussed for general bootstrap weights $w_{n:i_1 \dots i_m}$, it is perhaps more convenient and practical to use random variables $w_{n:1}, \dots, w_{n:n}$, and then have $w_{n:i_1 \dots i_m} = \gamma(w_{n:i_1}, \dots, w_{n:i_m})$ for some symmetric function $\gamma(\cdot)$. Standard resampling schemes like the classical bootstrap or the delete- d jackknife can be identified in this restricted set-up. Also, by varying the function $\gamma(\cdot)$, we get new and interesting resampling techniques.

Some special cases of minimum contrast estimators have been treated earlier. Arcones and Gine (1992b) have proved consistency of the classical bootstrap using empirical processes for

some minimum contrast estimators, where they did not use convexity. Rao and Zhao (1992) have also established consistency of some generalised bootstrap estimators, where the bootstrap weights are *i.i.d.* random variables.

The bootstrap and jackknife for U -statistics can also be studied in this framework. Consistency results for the usual bootstrap were obtained by Bickel and Freedman (1981), Athreya, Ghosh, Low and Sen (1984) for normalised and studentised U -statistics respectively. Later, Helmers (1991) proved second order accuracy of the bootstrap approximation for the distribution of a non-degenerate U -statistics. A general consistency result, regardless of the degree of degeneracy of the U -statistics, was obtained by Arcones and Gine (1992b), as long as the degree of degeneracy was known. A number of other authors have also studied bootstrap of U -statistics, see references in the papers mentioned above and also the book by Shao and Tu (1995). Generalised bootstrap of a particular form has been studied by Huskova and Janssen (1993a,b) and Janssen (1994), which may be seen as a special case of UBS , as also the usual bootstrap and jackknife. From our set-up one can obtain the first order consistency results for UBS for non-degenerate U -statistics with expectation zero. If the expectation of the U -statistics is not zero, the results of this chapter can be extended by routine techniques but some additional assumptions are needed on UBS weights.

In this chapter we do not study higher order results. We also do not consider cases where degenerate U -statistics arise.

8.3 Bootstrap representation results for $m = 1$

In this section we consider the $m = 1$ case separately for the sake of transparency. We establish a representation result similar to Theorem 8.2.1 for the bootstrap estimator, and obtain consistency results from it.

Let $\{w_{n;i}, i = 1, \dots, n, n = 1, 2, \dots\}$ be a triangular sequence of non-negative, row-wise exchangeable random variables, independent of $\{X_1, \dots, X_n\}$. We use the notations P_B , E_B , V_B to respectively denote probabilities, expectations and variances of these 'bootstrap weights', conditional on the given data $X_1, \dots, X_n$. We henceforth drop the first suffix in the weights $w_{n;i}$ and denote it by w_i . The notations $\sigma_n^2 = V_B w_i$, and $W_i = \sigma_n^{-1}(w_i - 1)$ will also be used. Since rowwise exchangeability of $w_{n;i}$'s is assumed, we may unambiguously adopt the notation

$$c_{abc\dots} = E_B W_1^a W_2^b W_3^c \dots$$

The notations k, K will be generic for constants. The following are a set of conditions on the weights:

$$E_B w_1 = 1 \tag{8.3.1}$$

$$K > \sigma_n^2 > k > 0 \tag{8.3.2}$$

$$c_{11} = O(n^{-1}) \quad (8.3.3)$$

$$c_{22} \rightarrow 1 \quad (8.3.4)$$

$$\sup_n c_4 < \infty \quad (8.3.5)$$

Theorem 8.3.1 *Suppose the assumptions (MC1)-(MC5) hold. Assume that UBS resampling is carried by minimising (8.2.3) by using weights satisfying (8.3.1)-(8.3.3). Then*

$$\sigma_n^{-1} n^{1/2} (a_{nB} - a_n) = -n^{-1/2} H^{-1} S_{nB} + r_{nB} \quad (8.3.6)$$

where $S_{nB} = \sum W_i g(a_*, Z_i)$ and $P_B[||r_{nB}|| > \epsilon] = o_P(1)$ for any $\epsilon > 0$.

In order to prove this theorem, we use Lemma 4 of Niemiro (1992). We quote this lemma below.

A set $B \subset \mathbb{R}^d$ is called a δ -triangulation of A , if every $a \in A$ is equal to a convex combination $\sum \lambda_i b_i$ of points $b_i \in B$, such that $|b_i - a| < \delta$.

Lemma 8.3.1 (Niemiro (1992)) *Let $A \subset A_0$ be convex subsets of $\mathbb{R}^d$ such that $|a - b| > 2\delta$ whenever $a \in A$ and $b \notin A_0$. Assume B is a δ -triangulation of A_0 , function h satisfies the Lipschitz condition $|h(a) - h(b)| \leq L|a - b|$ for $a, b \in A_0$, and function h' is convex on A_0 . If*

$$\sup_{b \in B} |h(b) - h'(b)| < \epsilon$$

then

$$\sup_{a \in A} |h(a) - h'(a)| < 5L\delta + 3\epsilon$$

Proof of Theorem 8.3.1 In order to prove this theorem, let $a_* = 0$ and $Q(a_*) = 0$, without loss of generality. Now define $X_{ni} = [f(n^{-1/2}a, Z_i) - f(0, Z_i)] - n^{-1/2}a^T g(0, Z_i)$ and $X_{nBi} = w_i X_{ni}$. Then we have $E_B X_{nBi} = X_{ni}$ and $\sum X_{nBi} = nQ_{nB}(n^{-1/2}a) - nQ_{nB}(0) - n^{-1/2}a^T S_{nW}$ where $S_{nW} = \sum w_i g(0, Z_i)$. In order to make explicit the fact that X_{ni} is a function of a , we sometimes write $X_{ni}(a)$ for X_{ni} .

Fix $\delta_1, \delta_2 > 0$. We first prove two facts.

$$\text{for any } M > 0 \quad P_B\left[\sup_{||a|| \leq M} \sigma_n^{-1} \left|\sum X_{nBi} - a^T H a / 2\right| > \delta_2\right] = o_P(1) \quad (8.3.7)$$

$$\exists M > 0 \text{ such that } Prob[P_B[||n^{-1/2} S_{nW}|| \geq M] > \delta_1] < \delta_2 \quad (8.3.8)$$

PROOF OF (8.3.7). Fix an $M > 0$. For fixed $\delta_* > 0$, get $\delta > 0$ and $\epsilon > 0$ such that $M_1 = M + (2\delta)^{1/2}$ and $\delta_* > 5M_1 \lambda_{\max}(H)\delta + 3\epsilon$, where $\lambda_{\max}(H)$ is the maximum eigenvalue

of H . Consider the set $A_1 = \{a : \|a\| \leq M_1\}$ and let $B = \{b_1, \dots, b_N\}$ be a finite δ -triangulation of A_1 . Note that with $A = \{a : \|a\| \leq M\}$, $L = M_1 \lambda_{\max}(H)$, $h(a) = a^T H a / 2$ and $h'(a) = \sum w_i X_{ni}(a)$, all the conditions of the above Lemma 8.3.1 are satisfied. Now

$$\begin{aligned} & P_B \left[\sup_{\|a\| \leq M} \sigma_n^{-1} \left| \sum X_{nBi} - a^T H a / 2 \right| > \delta_* \right] \\ & \leq P_B \left[\sup_{\|a\| \leq M} \sigma_n^{-1} \left| \sum X_{nBi} - a^T H a / 2 \right| > 5L\delta + 3\epsilon \right] \\ & \leq P_B \left[\sup_{a \in B} \sigma_n^{-1} \left| \sum X_{nBi} - a^T H a / 2 \right| > \epsilon \right] \end{aligned} \quad (8.3.9)$$

$$\begin{aligned} & \leq \sum_{j=1}^N P_B \left[\sigma_n^{-1} \left| \sum w_i X_{ni}(b_j) - b_j^T H b_j / 2 \right| > \epsilon \right] \\ & \leq \sum_{j=1}^N P_B \left[\left| \sum W_i X_{ni}(b_j) \right| > \epsilon / 2 \right] + \sum_{j=1}^N I_{\{\sigma_n^{-1} \left| \sum X_{ni}(b_j) - b_j^T H b_j / 2 \right| > \epsilon / 2\}} \end{aligned} \quad (8.3.10)$$

$$\leq \sum_{j=1}^N k \sum_{i=1}^n X_{ni}^2(b_j) + \sum_{j=1}^N I_{\{\sigma_n^{-1} \left| \sum X_{ni}(b_j) - b_j^T H b_j / 2 \right| > \epsilon / 2\}} \quad (8.3.11)$$

$$= o_P(1) \quad (8.3.12)$$

In the above calculations (8.3.9) follows from Lemma 8.3.1. Observe that in (8.3.11) the sums indexed by j run over finitely many points, so it is enough to show the probability rate (8.3.12) for each fixed b . It has been proved by Niemiro (1992, page 1522) that for fixed b , $\sum_i X_{ni}^2(b) = o_P(1)$ and $\sum_i X_{ni}(b) - b^T H b / 2 = o_P(1)$, hence (8.3.12) follows by using the lower bound in (8.3.2). This proves (8.3.7).

PROOF OF (8.3.8). Note that

$$\begin{aligned} P_B[\|n^{-1/2} S_{nW}\| > M] & \leq \frac{1}{M^2 n} E_B \left\| \sum_{i=1}^n w_i g(0, Z_i) \right\|^2 \\ & \leq \frac{2}{M^2 n} [\sigma_n^2 E_B \left\| \sum_{i=1}^n W_i g(0, Z_i) \right\|^2 + \left\| \sum_{i=1}^n g(0, Z_i) \right\|^2] \\ & \leq \frac{K \sigma_n^2}{M^2 n} \sum_{i=1}^n \|g(0, Z_i)\|^2 + \frac{2}{M^2 n} \left\| \sum_{i=1}^n g(0, Z_i) \right\|^2 \\ & = U_M \text{ say} \end{aligned}$$

The constant K in the last step is obtained by using the upper bound condition in (8.3.2) and (8.3.3). Now fix any two constants $\epsilon, \delta > 0$. By choosing M large enough, we have

$$\begin{aligned} & Prob[P_B[\sigma_n^{-1} \|n^{-1/2} S_{nW}\| < M] > \epsilon] \\ & \leq Prob[\epsilon < U_M] \\ & < \delta \end{aligned} \quad (8.3.13)$$

by using (MC4). This proves (8.3.8).

Fix a $\delta_0 > 0$. Now on the set $\mathcal{A}$ where both

$$\sup_{\|a\| \leq M} \sigma_n^{-1} \left\| \sum_{i=1}^n X_{nBi} - \frac{1}{2} a^T H a \right\| < \delta_0 \quad (8.3.14)$$

$$\|n^{-1/2} H^{-1} S_{nW}\| < M - 1 \quad (8.3.15)$$

hold, the convex function $nQ_{nB}(n^{-1/2}a) - nQ_{nB}(0)$ assumes at $n^{-1/2}H^{-1}S_{nW}$ a value less than its values on the sphere $\|a + n^{-1/2}H^{-1}S_{nW}\| = \kappa\sigma_n\delta_0^{1/2}$, where $\kappa = 2\lambda_{\min}^{1/2}(H)$. The global minimiser of this function is $n^{1/2}a_{nB}$, hence we have

$$n^{1/2}\sigma_n^{-1}a_{nB} = -n^{-1/2}\sigma_n^{-1}H^{-1}S_{nW} + r_{nB3} \quad (8.3.16)$$

Since δ_0 is arbitrary, and because of (8.3.7) and (8.3.8), we have $P_B[\|r_{nB3}\| > \delta] = o_P(1)$ for any $\delta > 0$.

Now divide (8.2.2) by σ_n and subtract it from (8.3.16) to get the result. This step again uses (8.3.2). ■

For any $c \in \mathbb{R}^d$ with $\|c\| = 1$, let

$$F_n(x) = \text{Prob}[n^{1/2}c^T(a_n - a_*) \leq x] \quad \text{and}$$

$$F_B(x) = P_B[\sigma_n^{-1}n^{1/2}c^T(a_{nB} - a_n) \leq x],$$

where P_B is the probability conditional on the data. The following corollary establishes the consistency of the resampling technique for estimating F_n . This also shows consistency for estimating the distribution of $n^{1/2}(a_n - a_*)$.

Corollary 8.3.1 (Consistency of UBS distribution estimate)

Assume the conditions of Theorem 8.3.1. Assume also (8.3.4) - (8.3.5). Then

$$\sup_{x \in \mathbb{R}} |F_B(x) - F_n(x)| = o_P(1) \quad \text{as } n \rightarrow \infty$$

Proof of Corollary 8.3.1 Let $h(0, Z_i) = c^T H^{-1} g(0, Z_i)$. Then for $v = Eh^2(0, Z_1)$, $\sup_x |F_n(x) - N(0, v)| \rightarrow 0$ follows from Theorem 8.2.1. Also observe that given the condition on the bootstrap weights, Lemma 4.6 of Praestgaard and Wellner (1993) (see Theorem 5.3.1) may be applied to obtain $\sup_x |F_B(x) - N(0, v)| = o_P(1)$ as $n \rightarrow \infty$. This lemma has also been used in Chapters 5 and 7, where we have shown in details how the different conditions of it are satisfied, hence we do not go into the details here. ■

A related problem is to estimate the asymptotic variance of $n^{1/2}c^T(a_n - a_*)$ for any vector $c \in \mathbb{R}^d$ with $\|c\| = 1$. Let $Eg(0, Z)g^T(0, Z) = V$. Then the asymptotic variance of $n^{1/2}c^T(a_n - a_*)$ is given by $v = c^T H^{-1} V H^{-1} c$. Let the bootstrap variance estimator be $v_B = \sigma_n^{-2} E_B[n^{1/2}c^T(a_{nB} - a_n)]^2$. The proof of Theorem 8.3.1 can be slightly modified in (8.3.7) and (8.3.8) by using bootstrap expectations in place of bootstrap probabilities. With such a modification, the following result is obtained as a corollary. Details are omitted.

Corollary 8.3.2 (Consistency of UBS variance estimate) *Assume the conditions of Theorem 8.3.1. Then*

$$v_B - v = o_P(1) \text{ as } n \rightarrow \infty$$

Note that Corollary 8.3.2 establishes consistency of the bootstrap variance estimator, and is weaker than the asymptotic representation of variance that were considered in earlier chapters.

8.4 Results for the general m case

In this section our main goal is to establish the results similar to Section 8.3 for general m . For convenience, let us fix the notation $S = \{(i_1, \dots, i_m) : 1 \leq i_1 < i_2 < \dots < i_m \leq n\}$, and a typical element of S is $s = (i_1, \dots, i_m)$. We use the notation $|s \cap t| = k$ to denote two typical subsets $s = \{i_1, \dots, i_m\}$ and $t = \{j_1, \dots, j_m\}$ of size m from $\{1, \dots, n\}$, which have exactly k elements in common. We often use the same notation s for both $s = (i_1, \dots, i_m) \in S$ and $s = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$. Also, let $N = \binom{n}{m}$. Let $Z_s = (Z_{i_1}, \dots, Z_{i_m})$ when $s = (i_1, \dots, i_m) \in S$. The notation $\sum_{|s \cap t|=j}$ denotes sum over all s and t for which $|s \cap t| = j$ holds.

The bootstrap weights w_s are non-negative random variables. We assume that for all $s \in S$, the distribution of w_s is the same. Define $\xi_n^2 = E_B(w_s - 1)^2$ and let $W_s = \xi_n^{-1}(w_s - 1)$. Assume that $E_B W_s W_t$ is a function of $|s \cap t|$ only, and let $c_j = E_B W_s W_t$ whenever $|s \cap t| = j$, for $j = 0, 1, \dots, m$.

We assume that the following conditions are satisfied:

$$K > \xi_n^2 > k > 0 \quad (8.4.1)$$

$$c_0 = o(1) \quad (8.4.2)$$

$$|c_j| = O(1), \quad j = 1, \dots, m. \quad (8.4.3)$$

Define $g_1(0, Z_1) = E[g(0, Z_1, \dots, Z_m) | Z_1]$ and $f_i = \binom{n-1}{m-1}^{-1} \sum_{s: i \in s} W_s$ where the sum runs over all $s = (i_1, \dots, i_m) \in S$ such that i is one of $i_1, \dots, i_m$.

Theorem 8.4.1 *Suppose the assumptions (MC1)-(MC5) hold. Assume that resampling is carried by minimising (8.2.3) by using weights satisfying (8.4.1)-(8.4.2). Then*

$$\xi_n^{-1} n^{1/2} (a_{nB} - a_n) = -n^{1/2} H^{-1} S_{nB} + r_{nB} \quad (8.4.4)$$

$$= -m n^{-1/2} \sum f_i H^{-1} g_1(0, Z_i) + R_{nB} \quad (8.4.5)$$

where $S_{nB} = \binom{n}{m}^{-1} \sum_{s \in S} W_s g(a_*, Z_s)$, $P_B[||r_{nB}|| > \epsilon] = o_P(1)$ and $P_B[||R_{nB}|| > \epsilon] = o_P(1)$ for every $\epsilon > 0$.

Proof of Theorem 8.4.1 The first part of this proof is similar to that of Theorem 8.3.1. Define $X_{ns} = f(n^{-1/2} a, Z_s) - f(0, Z_s) - n^{-1/2} a^T g(0, Z_s)$, and let $X_{nBs} = w_s X_{ns}$, $S_{nB} =$

$N^{-1} \sum_s g(0, Z_s)$. We have $nQ_{nB}(n^{-1/2}a) - nQ_{nB}(0) - n^{1/2}a^T S_{nW} - a^T Ha/2 = nN^{-1} \sum X_{nBs} - a^T Ha/2$. Then arguments similar to those of Theorem 8.3.1 yields (8.4.4) once it is established that for any fixed $\epsilon, \delta > 0$

$$\text{for any } M > 0 \quad \mathbb{P}_B[\sup_{||a|| \leq M} |nN^{-1} \sum X_{nBs} - a^T Ha/2| > \delta] = o_P(1) \quad (8.4.6)$$

$$\exists M > 0 \text{ such that } \text{Prob}[\mathbb{P}_B[||n^{1/2}S_{nW}|| > M] > \epsilon] < \delta \quad (8.4.7)$$

The proof of these are similar to that of (8.3.7) and (8.3.8). By carefully following the arguments for those, we only have to show

$$n^2 N^{-2} \mathbb{E}_B(\sum W_s X_{ns})^2 = o_P(1) \quad (8.4.8)$$

$$nN^{-2} \mathbb{E}_B||\sum w_s g(0, Z_s)||^2 = O_P(1) \quad (8.4.9)$$

For (8.4.8), we have

$$\begin{aligned} n^2 N^{-2} \mathbb{E}_B(\sum W_s X_{ns})^2 &= n^2 N^{-2} \sum_{j=0}^m c_j \sum_{|s \cap t|=j} X_{ns} X_{nt} \\ &= n^2 N^{-2} c_0 (\sum_s X_{ns})^2 + n^2 N^{-2} \sum_{j=1}^m c_j \sum_{|s \cap t|=j} X_{ns} X_{nt} \end{aligned}$$

A little algebra shows that $nN^{-1} \sum X_{ns} = o_P(1)$. The first term is $o_P(1)$ from this and (8.4.2). For the other term, first note that the sum over j is finite, and from (8.4.3), we only need to show for every fixed $j = 1, \dots, m$, $n^2 N^{-2} \sum_{|s \cap t|=j} X_{ns} X_{nt} = o_P(1)$. For this, we have

$$\begin{aligned} &n^2 N^{-2} \sum_{|s \cap t|=j} X_{ns} X_{nt} \\ &\leq nN^{-2} \sum_{|s \cap t|=j} [a^T (g(n^{-1/2}a, Z_s) - g(0, Z_s))] [a^T (g(n^{-1/2}a, Z_t) - g(0, Z_t))] \end{aligned}$$

Since the number of terms in $\sum_{|s \cap t|=j}$ is $O(n^{-1}N^2)$, it is enough to show for any $s, t \in S$

$$a^T (g(n^{-1/2}a, Z_s) - g(0, Z_s)) a^T (g(n^{-1/2}a, Z_t) - g(0, Z_t)) = o_P(1)$$

Observe that $a^T (g(n^{-1/2}a, Z_s) - g(0, Z_s))$ are non-negative random variables that are non-increasing in n , and their limit is 0. This establishes the above. Details of this argument is similar to those of (4.11) of Niemiro (1992). The proof of (8.4.9) is along the same lines. This proves (8.4.4).

In order to get the representation (8.4.5), we have to show

$$n^{1/2} N^{-1} \sum W_s g(0, Z_s) = mn^{-1/2} \sum_i f_i g_1(0, Z_i) + R_{nB1},$$

where for any $\delta > 0$, we have $\mathbb{P}_B[||R_{nB1}|| > \delta] = o_P(1)$.

Let $h(0, Z_s) = g(0, Z_s) - \sum_{j=1}^m g_1(0, Z_{i_j})$. This is a kernel of a first order degenerate U -statistics. Then we have $W_s g(0, Z_s) = W_s \sum_{j=1}^m g_1(0, Z_{i_j}) + W_s h(0, Z_s)$. Also, it can be shown that $\sum_s \sum_{j=1}^m W_s g_1(0, Z_{i_j}) = \binom{n-1}{m-1} \sum_{i=1}^n f_i g_1(0, Z_i)$. Also $E(|N^{-1} \sum h(0, Z_s)|)^{-2} = O(n^{-2})$ (see Serfling (1980, page 188) for a proof of this when $d = 1$). Now using this and (8.4.3), after some algebra we obtain $P_B[|n^{1/2} N^{-1} \sum W_s h(0, Z_s)| > \delta] = o_P(1)$, and this yields (8.4.5). ■

As in Section 8.3, the representation (8.4.5) is useful for showing consistency of UBS , both for variance estimation as well as distributional approximation. From (8.4.2) we have that the asymptotic mean of f_1 is 0. Let its asymptotic variance be $v^2(m)$. This will in general be a function of m . Then the appropriate standardised bootstrap statistic is $n^{1/2} \xi_n^{-1} v_m^{-1} (a_{nB} - a_n)$.

For any $c \in \mathbb{R}^d$ with $\|c\| = 1$, let

$$\begin{aligned} v &= E[n^{1/2} c^T (a_n - a_*)]^2 \\ F_n(x) &= Prob[n^{1/2} c^T (a_n - a_*) \leq x] \\ v_B &= E_B[n^{1/2} \xi_n^{-1} v_m^{-1} (a_{nB} - a_n)]^2 \\ F_B(x) &= P_B[n^{1/2} \xi_n^{-1} v_m^{-1} (a_{nB} - a_n) \leq x]. \end{aligned}$$

Then we have

Corollary 8.4.1 Assume the conditions of Theorem 8.4.1.

(a) (Consistency of UBS variance estimate)

$$v_B - v = o_P(1) \text{ as } n \rightarrow \infty$$

(b) (Consistency of UBS distribution estimate) Assume, in addition to the above conditions that $f_{n;i}$'s are exchangeable with $\sup_n E_B f_{n;i}^4 < \infty$ and $E_B f_{n;i}^2 f_{n;j}^2 \rightarrow 1$ for $i \neq j$ as $n \rightarrow \infty$. Then

$$\sup_{x \in \mathbb{R}} |F_B(x) - F(x)| = o_P(1) \text{ as } n \rightarrow \infty$$

The proof of the above results are similar to the proof of their corresponding results in Section 8.3, and we omit the details.

REMARK 1. All the examples of UBS bootstrap techniques discussed in Chapter 2 may be used in the set-up of the present chapter for $m = 1$, except for the delete- d jackknife where $d/n \rightarrow 0$ or 1, for which (8.3.2) is not satisfied.

REMARK 2. In case of n_1 out of n bootstrap with replacement, the upper bound in (8.3.2) is violated. However, using the particular properties of this bootstrap, the representation (8.3.6) may still be derived. Notice that the only place where the upper bound of (8.3.2) is used is in the proof of (8.3.8). We have to show that $n^{-1} E_B \|\sum w_i g(0, Z_i)\|^2$ is bounded in probability. Now note that in this particular example, the marginal distribution of w_i is as follows: it is 0 with a probability $(n - n_1)/n$, and $nn_1^{-1} U_i$ with probability n_1/n , where U_i is a *binomial*($n_1, 1/n$)

random variable. The joint distribution of (w_i, w_j) may be similarly calculated. Now routine computations by expanding the squared term and then computing individual moments shows that $n^{-1}E_B[\sum w_i g(0, Z_i)]^2$ is bounded in probability. Thus (8.3.6) holds.

REMARK 3. For the general m set-up, although the discussion in Section 8.4 is based on weights $w_{i_1, \dots, i_m}$, it is convenient to start with some $w_1, \dots, w_n$, and define w_s as a function of $w_{i_1, \dots, i_m}$. Two such functions can be easily defined, (2a) $w_s = \prod_{j=1}^m w_{n+i_j}$ and (2b) $w_s = m^{-1} \sum w_{n+i_j}$.

For weights described in Chapter 2, with either definition of w_s , the conditions (8.4.1) – (8.4.3) are satisfied. With definition (2a) moments of products of w_i 's of various orders are involved and conditions are difficult to check unless m is small. With weights from Example 2.3.1 and 2.3.3 respectively, and definition (2a), one gets back the classical bootstrap and jackknife schemes in this set-up.

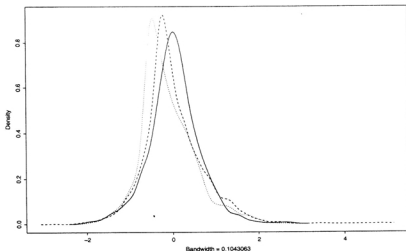
A natural question is how f_i and W_s relate to each other. It may be observed that in both the definitions of w_s ((2a) and (2b)) the difference between W_s and an appropriately scaled f_i (where the scaling factor is a function of m only) is negligible in the representation (8.4.5). Again the algebra is complicated in case of (2a), and may be verified with small values of m . There are no such difficulties with (2b). With the usual bootstrap and jackknife random variables as weights and definition (2b), we have alternative forms of the usual bootstrap and jackknife in case of U -functionals that have not been studied before.

8.5 Computational examples

We now present some computational examples to show how the proposed bootstrap performs in practice. The first four examples are based on simulated data, while the last one is based on a real life problem. First, as discussed above, bootstrap weights for general m may be generated from those for $m = 1$ by using either a multiplicative (2a) or additive (2b) function. The multiplicative form (2a) is more traditional. However, it can be easily shown that the computational time and space complexity of the problem greatly reduces if (2b) is used. Also, computations may be further simplified if *i.i.d.* weights from some *uniform* distribution is used, since such weights can be generated very easily and very fast in modern computers. Our examples in this section shows that in terms of accuracy, there is no essential difference between (2a) and (2b).

Let $(X_{i1}, X_{i2})^T$, $i = 1, \dots, n$ be *i.i.d.* observations from $\mathbb{R}^2$ valued random vector, and the p^{th} Oja median of the distribution of $(X_1, X_2)^T$ is defined as that $\theta = (\theta_1 \theta_2)^T \in \mathbb{R}^2$ that minimises $E\Delta_p$, where $\Delta_p = |(X_1 - \theta_1)(X_{j2} - \theta_2) - (X_{i2} - \theta_2)(X_{j1} - \theta_1)|^p$. Note that Δ_p is convex in θ for $p \geq 1$. The Oja medians of different order p are affine equivariant location parameters for multivariate distributions, with many nice properties associated with them, see Oja (1983) for details. Suppose our parameter of interest is the Oja median ($p = 1$), and we

Figure 8.1: Bivariate normal distribution



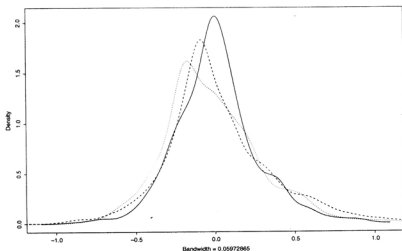
want to test the hypothesis $\theta_1 = \theta_2$. Suppose a random sample $\{(X_{i1}, X_{i2})^T, i = 1, \dots, n\}$ is taken. The sample Oja median is the solution $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^T$ to the following minimisation problem:

$$\min_{\theta} \sum_{i < j} |(X_{i1} - \theta_1)(X_{j2} - \theta_2) - (X_{i2} - \theta_2)(X_{j1} - \theta_1)|$$

The test statistic is taken to be $T = (\hat{\theta}_1 - \hat{\theta}_2)/\sqrt{2}$, which is a linear combination $c^T \hat{\theta}$, with $\|c\| = 1$. The behaviour of the density of T under null hypothesis is of interest, which may be estimated using bootstrap methods. In Examples 1-4, we simulate data from four different bivariate distributions to study this problem.

For each distribution, we generate a data of size n and calculate T_n . The values of n used is given later. The simulation is repeated 1000 times, and from these 1000 values of T , a kernel smoothed density estimate of the true distribution of T is computed. In practice we have only one data set. Then for one of these 1000 data sets, three separate bootstrap computations are carried out. The bootstrap sample size B was taken to be 5000 for each of the methods. These are, (i) the classical bootstrap method ((2a) with $Multinomial(n, 1/n, \dots, 1/n)$ weights); (ii) a new 'random sampling from data' method given by (2b) with $Multinomial(n, 1/n, \dots, 1/n)$ weights; (iii) a 'uniform weights bootstrap' method given by (2b) with *i.i.d.* weights from $Uniform(1/2, 3/2)$ distribution. Kernel smoothed density estimates of T are obtained for each of the three bootstrap methods. The results are displayed in Figures 8.1-8.4. The results for the third bootstrap method are not reported since they are similar to the results from the other

Figure 8.2: Spherically contoured distribution



two bootstrap techniques. Each figure contains the overlayed plots of the three densities, the simulated density in bold line, the classical bootstrap estimate from (i) in broken line and the new bootstrap from (ii) in dotted line. The bandwidth used is quoted in the figures.

Instead of making a comparison of densities we could have chosen some other graphical or quantitative criteria. Such comparison were also made, and since the results are similar they are not reported here. We now describe the four different distributions used. In each case, we have $\theta = 0 \in \mathbb{R}^2$.

EXAMPLE 1 The data is simulated from a bivariate normal distribution. A sample of size 50 is obtained from the bivariate normal distribution

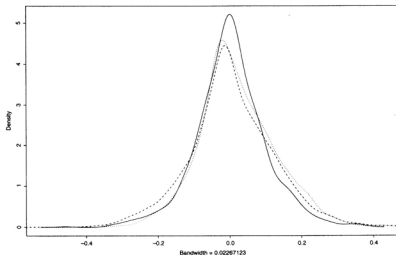
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 9 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

The results are reported in Figure 8.1.

EXAMPLE 2 The data is a sample of size 100 from a spherically contoured distribution. For $(X, Y)^T$ following that distribution, we take $X = rU_1$ and $Y = rU_2$, where $r \sim \text{Gamma}(10)$, and $(U_1, U_2)^T$ are uniformly distributed over the unit sphere in $\mathbb{R}^2$. The results are displayed in Figure 8.2.

EXAMPLE 3 The data is a sample of size 100 from an elliptically contoured distribution.

Figure 8.3: Elliptically contoured distribution



For $(X, Y)^T$ from that distribution, we have

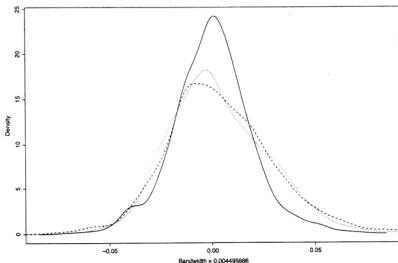
$$\begin{pmatrix} X \\ Y \end{pmatrix} = r A \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \quad \text{where } A = 9 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

$r \sim \text{Gamma}(2, 2)$ and $(U_1, U_2)^T$ is uniformly distributed over the unit sphere in $\mathbb{R}^2$. The results are reported in Figure 8.3.

EXAMPLE 4 The data is a sample of size 100 from the bivariate t -distribution, whose density is given by $g(t) = c[1 + t^T t/10]^{-6}$, $t \in \mathbb{R}^2$, and $c = (10\pi)^{-1}[\Gamma(5)]^{-1}\Gamma(6)$. The results are reported in Figure 8.4.

In all the above four examples, the classical and both the new bootstrap methods perform equally well, and since the new bootstrap techniques requires much less time and computer memory, they are preferable.

EXAMPLE 5 The lipid data (Scott (1992), page 275) is data on 371 males evaluated for chest pain, of which 51 patients show no evidence of heart disease, and 320 patients have arterial occlusion. The concentration of plasma cholesterol and plasma triglycerids are reported. The two variables have strong non-linear interaction, outlying observations are possibly present in the data, sample plots show incompatibility with normality assumption, see Scott (1992) for discussion, comments and further references. All these factors lead to a complicated analysis

Figure 8.4: Bivariate t -distribution

of the data, see Scott *et al* (1978, 1980). If μ_1 and μ_2 respectively represent some location parameter for the bivariate distribution of the two plasma lipids from patients without arterial occlusion and with it, the problem is to test $\mu_1 = \mu_2$. Because of possible non-normality and presence of outlying observations, some multivariate median is preferable to the mean as a location parameter. We first took the Oja median ($p = 1$) as our parameter. We generated 5000 classical bootstrap sample values of μ_1 to generate a distribution of the sample Oja median for patients with no arterial occlusion. The sample value of $\|\mu_2\|$ was found to lie above the 99th percentile of the distribution of $\|\mu_1\|$. The experiment was repeated with the other two bootstrap schemes described earlier in this section to get same conclusions. We then repeated the entire scheme by using the spatial median as the parameter. For this case also, for all the three resampling schemes the value of $\|\mu_2\|$ lie above the 99th percentile of the bootstrap histogram. Whereas the Oja median has the advantage of being affine equivariant, the spatial median has a high breakdown point. Thus, across the two choices of multivariate location and three choices of resampling scheme, the hypothesis $\mu_1 = \mu_2$ was rejected. This is a simple verification of the well established fact that plasma lipids influence coronary diseases. Note that we did not require the elaborate data analysis that was employed earlier for the same problem. This serves to illustrate that the *UBS* technique can reduce complexity of the statistical problem and be extremely useful in practice.

References

- ALLEN, M. AND DATTA, S. (1999) A note on bootstrapping M -estimators in ARMA models, *Jour. Time Series Anal.*, **20**, 365-379.
- ARCONES, M. A. AND GINE, E. (1992a) On the bootstrap of U and V statistics, *Ann. Statist.*, **20**, 655-674.
- ARCONES, M. A. AND GINE, E. (1992b) On the bootstrap of M -estimators and other statistical functionals, in *Exploring the limits of bootstrap*, (LePage, R. and Billard, L. ed.), 13-48, Wiley, New York.
- ARENAL-GUTIERREZ, E. AND MATRAN, C. (1996) A zero-one law approach to the central limit theorem for the weighted bootstrap mean, *Ann. Prob.*, **24**, 532-540.
- ATHREYA, K. B., GHOSH, M., LOW, L. Y. AND SEN, P. K. (1984) Laws of large numbers for bootstrapped U -statistics, *J. Statist. Plann. Inference*, **9**, 185-194.
- BABU, G. J. AND BAI, Z. (1992) Edgeworth expansions for errors-in-variables models, *Jour. Mult. Anal.*, **42**, 226-244.
- BABU, G. J. AND BAI, Z. (1993) Edgeworth expansions of a function of sample means under minimal moment conditions and partial Cramer's condition, *Sankhyā, Series A*, **55**, 244-258.
- BABU, G. J. AND BAI, Z. (1996) Mixtures of global and local Edgeworth expansions and their applications, *Jour. Mult. Anal.*, **59**, 282-306.
- BABU, G. J. AND SINGH, K. (1983) Inference on means using the bootstrap, *Ann. Statist.*, **11**, 999-1003.
- BABU, G. J. AND SINGH, K. (1984) Asymptotic representations related to jackknifing and bootstrapping L -statistics, *Sankhyā, Series A*, **46**, 195-206.
- BAI, Z. AND RAO, C. R. (1991) Edgeworth expansion of a function of sample means, *Ann. Statist.*, **19**, 1295-1315.

- BAI, Z. AND ZHAO, L. (1986) Edgeworth expansions of distribution function of independent random variables, *Scientica Sinica A*, **29**, 1-22.
- BARBE, P. AND BERTAIL, P. (1995) *The Weighted Bootstrap*. Lecture Notes in Statistics, Volume **98**, Springer - Verlag, New York.
- BERA, A. K. AND HIGGINS, M. L. (1993) A survey of ARCH models: properties, estimation and testing, *Jour. of Economic Surveys*, **7**, 305-366.
- BHATTACHARYA, R. N. AND RANGA RAO, R. (1976) *Normal approximation and asymptotic expansions*, John Wiley, New York.
- BHANJA, J. AND GHOSH, J. K. (1992) Efficient estimation with many nuisance parameters, *Sankhyā, Series A*, **54**, 1-39.
- BICKEL, P. J. AND FREEDMAN, D. A. (1981) Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196-1217.
- BICKEL, P. J. AND FREEDMAN, D. A. (1983) Bootstrapping regression models with many parameters. In *Festschrift for Erich L. Lehmann* (P. Bickel, K. Doksum and J. Hodges, Jr., eds.), 28-48, Wadsworth, Belmont, Calif.
- BICKEL, P.J. AND LEHMANN, E.L. (1979) Descriptive statistics for nonparametric models, IV. Spread. In *Contributions to Statistics* Ed. J. Jureckova. pp. 33-40. Academia, Prague.
- BICKEL, P.J. AND SAKOV, A. (1999) On the choice of m in m out of n bootstrap in estimation problems, *preprint*.
- BLOOMFIELD, P. AND STIGLER, W. L. (1983) *Least absolute deviations Theory, applications and algorithms*, Birkhauser, Boston.
- BOOS, D. AND MONAHAN, J. F. (1986) Bootstrap methods using prior information. *Biometrika*. **73**. 77-83.
- BOSE, A. (1998) Bahadur representation of M_m estimates, *Ann. Statist.*, **26**, 771-777.
- BOSE, A. AND KUSHARY, D. (1996) Jackknife and weighted jackknife estimation of the variance of M -estimators in linear regression, *Tech. Report No. 96-12*, Dept. of Statistics, Purdue University.
- BOSE, A. AND SENGUPTA, D. (1992) Asymptotic properties of maximum likelihood and least squares estimates in nonlinear regression and generalised linear models, *Tech. Report No. 20/91*, Stat-Math Unit, Indian Statistical Institute, Calcutta.

- BROCKWELL, P. J., AND DAVIS, R. A. (1986) *Time series: theory and methods*, Springer-Verlag, New York.
- CARROLL, R. J., HARDLE, W. AND MAMMEN, E. (1999) Estimation in an additive model when components are linked parametrically, *preprint available from <http://sfb.wiwi.hu-berlin.de/>*.
- CARROLL, R. J., RUPERT, D. AND WELSH, A. H. (1998) Local estimating equations, *Jour. Amer. Statist. Assoc.*, **93**, 214-227.
- CARLSTEIN, E. (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *Ann. Statist.*, **14**, 1171-1194.
- CHATTERJEE, S. AND BOSE, A. (1997) Second order comparison of different resampling schemes in linear regression, Tech. Report No. 15/97, Stat-Math Unit, Indian Statistical Institute, Calcutta.
- CHATTERJEE, S. (1998) Another look at the jackknife: further examples of generalised bootstrap. *Statist. Prob. Letters*, **40**, 307-319.
- CHATTERJEE, S. AND BOSE, A. (1998a) Second order comparison of different bootstrap and jackknife methods in multiple regression, Tech. Report No. 10/98, Stat-Math Unit, Indian Statistical Institute, Calcutta.
- CHATTERJEE, S. AND BOSE, A. (1998b) Second order comparison of different resampling techniques in regression, *submitted for publication*.
- CHATTERJEE, S. AND BOSE, A. (1998c) Dimension asymptotics for the uncorrelated weights bootstrap in linear regression. Tech. Report No. 12/98, Stat-Math Unit, Indian Statistical Institute, Calcutta.
- CHATTERJEE, S. AND BOSE, A. (1999a) Variance estimation in high dimensional regression models, To appear in *Statistica Sinica*.
- CHATTERJEE, S. AND BOSE, A. (1999b) On the uncorrelated weights bootstrap in non-linear problems, Tech. Report No. 1/99, Stat-Math Unit, Indian Statistical Institute, Calcutta.
- CHATTERJEE, S. AND BOSE, A. (1999c) Second order accuracy of generalised bootstrap in linear regression, Tech. Report No. 4/99, Stat-Math Unit, Indian Statistical Institute, Calcutta.
- CHATTERJEE, S. AND BOSE, A. (1999d) Generalised bootstrap for minimum contrast estimators, Tech. Report No. 5/99, Stat-Math Unit, Indian Statistical Institute, Calcutta.

- CHOUDHURY, J. AND SERFLING, R.J. (1988) Generalized order statistics, Bahadur representations, and sequential nonparametric fixed width confidence intervals, *Jour. Statist. Plann. Inference*, **19**, 269 - 282.
- CHAUDHURI, P. (1996) On a geometric notion of quantiles for multivariate data, *J. Amer. Statist. Assoc.*, **91**, 862 - 872.
- CHEN, H. Y., AND LITTLE, R. (1999) A test of missing completely at random for generalised estimating equations with missing data, *Biometrika*, **86**, 1-14.
- COX, D. R. (1993) Unbiased estimating equations derived from statistics that are functions of a parameter, *Biometrika*, **83**, 433-440.
- CSORGO, S. AND MASON, D. M. (1989) Bootstrap empirical functions, *Ann. Statist.*, **17**, 1447 - 1471.
- DAVISON, A. C., HINKLEY, D. V. (1997) *Bootstrap methods and their applications*, Cambridge University Press, Cambridge.
- DAVISON, A. C., HINKLEY, D. V., AND WORTON, B. J., (1992) Bootstrap likelihoods, *Biometrika*, **79**, 113-130.
- EFRON, B. (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 101-118.
- EFRON, B. (1982) *The Jackknife, the Bootstrap and other resampling plans*, SIAM, Philadelphia.
- EFRON, B. (1992) Six questions raised by the bootstrap, in *Exploring the limits of bootstrap*, (ed. LePage and Billard), Wiley, New York.
- EFRON, B. AND LE PAGE, R. (1992) Introduction to bootstrap, in *Exploring the limits of bootstrap*, (ed. LePage and Billard), Wiley, New York.
- EFRON, B. AND TIBSHIRANI, R. (1993) *An introduction to the bootstrap*, Chapman and Hall, New York.
- EFRON, B. AND TIBSHIRANI, R. (1996) Using specially designed exponential families for density estimation, *Ann. Statist.*, **24**, 2431-2461.
- EFRON, B. AND TIBSHIRANI, R. (1998) The problem of regions, *Ann. Statist.*, **26**, 1687-1718.
- ENGLE, R. F. AND GRANGER, C. W. J. (1987) Co-integration and error-correction: representation, estimation and testing, *Econometrica*, **55**, 251-276.

- FAN, J., HARDLE, W. AND MAMMEN, E. (1998) Direct estimation of low dimensional components in additive models, *Ann. Statist.*, **26**, 943-971.
- FAHRMEIR, L. AND TUTZ, G. (1994) *Multivariate statistical modelling based on generalised linear models*, Springer-Verlag, New York.
- FELLER, W. (1968) *An introduction to probability theory and its applications, Volume-1*, 3rd edition, Wiley Eastern, New Delhi.
- FREEDMAN, D. A. (1981) Bootstrapping regression models, *Ann. Statist.*, **9**, 1218-1228.
- FREEDMAN, D. A. (1984) On bootstrapping two stage least squares estimators in stationary linear models, *Ann. Statist.*, **12**, 827-852.
- FREEDMAN, D. A. AND PETERS, S. C. (1984), Bootstrapping a regression equation: some empirical results, *J. Amer. Statist. Assoc.*, **79**, 97-106.
- GINE, E. (1997) Lectures on some aspects of the bootstrap, in *Lectures on probability theory and statistics, Lecture notes in mathematics, 1665*, Springer-Verlag, New York.
- GINE, E. AND ZINN, J. (1989) Necessary conditions for the bootstrap of the mean, *Ann. Statist.*, **17**, 684-691.
- GINE, E. AND ZINN, J. (1990) Bootstrap general empirical measures, *Ann. Prob.*, **18**, 851-869.
- GOTZE, F. AND KUNSCH, H. R. (1996) Second order correctness of the blockwise bootstrap for stationary observations, *Ann. Statist.*, **24**, 1914-1933.
- GRUET, M. A., HUET, S. AND JOLIVET, E. (1993) Practical use of bootstrap in regression, *Computer Intensive Methods in Regression*, 150-166, Stat. Comp. Physica, Heidelberg.
- HABERMAN, S. J. (1977a) Log-linear and frequency tables with small expected cell counts, *Ann. Statist.*, **5**, 1148-1169.
- HABERMAN, S. J. (1977b) Maximum likelihood estimates in exponential response models, *Ann. Statist.*, **5**, 815-841.
- HABERMAN, S. J. (1989) Concavity and estimation, *Ann. Statist.*, **17**, 1631-1661.
- HALL, P. (1987) Edgeworth expansion for Student's t statistic under minimal moment conditions, *Ann. Probab.*, **15**, 920-931.
- HALL, P. (1992) *The bootstrap and Edgeworth expansion*, Springer-Verlag, New York.

- HALL, P. AND HEYDE, C. C. (1980) *Martingale limit theory and its application*, Academic Press, New York.
- HALL, P. AND MAMMEN, E. (1994) On general resampling algorithms and their performance in distribution estimation, *Ann. Statist.*, **22**, 2011-2030.
- HALL, D. B., AND SEVERINI, T. A. (1998) Extended generalised estimating equations for clustered data, *Jour. Amer. Statist. Assoc.*, **93**, 1365-1375.
- HARDLE, W. AND MARRON, S. (1991) Bootstrap simultaneous error bars for non-parametric regression. *Ann. Statist.* **19**, 778-796.
- HARDLE, W., MAMMEN, E. AND MULLER, M. (1998) Testing parametric versus semiparametric modeling in generalised linear models, *Jour. Amer. Statist. Assoc.*, **93**, 1461-1474.
- HASLETT, J. AND HAYES, K. (1998) Residuals for the linear model with general covariance structure. *J. Roy. Statist. Soc. Series B*, **60**, 201-215.
- HASTIE, T. J. AND TIBSHIRANI, R. (1990) *Generalised additive models*, Chapman and Hall, New York.
- HELMERS, R. (1991) On the Edgeworth expansion and the bootstrap approximation for a studentised U -statistic, *Ann. Statist.*, **19**, 470-484.
- HINKLEY, D. V. (1977) Jackknifing in unbalanced situations. *Technometrics*, **19**, 285-292.
- HJORT, N. L. AND JONES, M. C. (1996) Locally parametric nonparametric density estimation, *Ann. Statist.*, **24**, 1619-1647.
- HOLLANDER, M. AND WOLFE, D.A. (1973) *Nonparametrical Statistical Methods*, John Wiley and Sons, New York.
- HUBER, P. J. (1981) *Robust Statistics*, Wiley, New York.
- HUET, S. AND JOLIVET, E. (1989) Exactitude au second order des intervalles de confiance bootstrap pour les parametres d'un modele de regression non lineaire, *C. R. Acad. Sci. Paris Ser. I. Math.*, **308**, 429-432.
- HUET, S., JOLIVET, E. AND MESSEAN, A. (1990) Some simulation results about confidence intervals and bootstrap methods in non-linear regression. *Statistics*, **21**, 369-432.
- HUSKOVA, M. AND JANSSEN, P. (1993a) Generalised bootstrap for studentised U -statistics: a rank statistics approach, *Statist. Prob. Lett.*, **16**, 225-233.

- HUSKOVA, M. AND JANSSEN, P. (1993b) Consistency of generalised bootstrap for degenerate U -statistics, *Ann. Statist.*, **21**, 1811-1823.
- JAMES, L. F. (1997) A study of a class of weighted bootstraps for censored data, *Ann. Statist.*, **25**, 1595-1621.
- JANSSEN, P. (1994) Weighted bootstrap of U -statistics, *J. Statist. Plann. Inference*, **38**, 31-42.
- JIANG, J. (1998) Consistent estimators in generalised linear mixed models, *Jour. Amer. Statist. Assoc.*, **93**, 720-729.
- JONES, M. C. (1993) Kernel density estimation when the bandwidth is large, *Australian Jour. Statist.*, **35**, 319-326.
- JURECOVA, J. AND SEN, P. K. (1996) *Robust statistical procedures*, Wiley, New York.
- KOUL, H. L. (1992) *Weighted empiricals and linear models*, IMS lecture notes monograph series, Vol. **21**.
- KOUL, H. L. (1996) Asymptotics of some estimators and sequential residual empiricals in nonlinear time series, *Ann. Statist.*, **24**, 380-404.
- KUNSCH, H. R. (1989) The jackknife and the bootstrap for general stationary observations, *Ann. Statist.*, **17**, 1217-1241.
- LAHIRI, S. N. (1992) Bootstrapping M -estimators of a multiple linear regression parameter, *Ann. Statist.*, **20**, 1548-1570.
- LAHIRI, S. N. (1996) On Edgeworth expansion and moving block bootstrap for studentised M -estimators in multiple linear regression models, *Jour. Mult. Anal.*, **56**, 42-59.
- LAHIRI, S. N. (1999) Theoretical comparisons of block bootstrap methods, *Ann. Statist.*, **27**, 386-404.
- LAI, T. L. (1994) Asymptotic properties of nonlinear least squares estimates in stochastic regression models, *Ann. Statist.*, **22**, 1917-1930.
- LEE, K. W. (1990) Bootstrapping logistic regression models with random regressors, *Comm. Statist. A*, **19**, 2527-2539.
- LIU, R. Y. (1988) Bootstrap procedures under some non-i.i.d. models, *Ann. Statist.*, **16**, 1696-1708.

- LIU, R. Y. AND SINGH, K. (1992) Efficiency and robustness in resampling. *Ann. Statist.*, **20**, 370-384.
- LIU, R. Y. AND SINGH, K. (1992a) Moving block jackknife and bootstrap capture weak dependence, in *Exploring the limits of bootstrap*, (ed. LePage and Billard), Wiley, New York.
- LO, A. Y. (1987) A large sample study of the Bayesian bootstrap, *Ann. Statist.*, **15**, 360-375.
- LO, A. Y. (1988) A Bayesian bootstrap for a finite population, *Ann. Statist.*, **16**, 1684-1695.
- LO, A. Y. (1991) Bayesian bootstrap clones and a biometry function. *Sankhyā, Ser. A*, **53**, 320-333.
- LOPUHAA, H. P. (1991) Breakdown point and asymptotic properties of multivariate S -estimators and τ -estimators: A summary, in *Directions in robust statistics and diagnostics, Part I*, (Stahel, W. and Weisberg, S. ed.), 167-182, Springer-Verlag, New York.
- MAMMEN, E. (1989) Asymptotics with increasing dimension for robust regression with applications to the bootstrap, *Ann. Statist.*, **17**, 382-400.
- MAMMEN, E. (1992a) *When Does Bootstrap Work ? Asymptotic results and simulations*, Lecture Notes in Statistics, Volume **77**, Springer - Verlag, New York.
- MAMMEN, E. (1992b) Bootstrap, wild bootstrap and asymptotic normality, *Prob. Theory and related fields*, **93**, 439-455.
- MAMMEN, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models, *Ann. Statist.*, **21**, 255-285.
- MAMMEN, E. (1995) Bootstrap, wild bootstrap and generalised bootstrap, *preprint available from [ftp://statlab.uni-heidelberg.de/pub/reports/by-series/beitrag11.ps](http://statlab.uni-heidelberg.de/pub/reports/by-series/beitrag11.ps)*.
- MAMMEN, E. (1996) Empirical process of residuals for high dimensional linear models, *Ann. Statist.*, **24**, 307-335.
- MAMMEN, E. AND VAN DE GEER, S. (1997) Penalised quasi-likelihood estimation in partial linear models, *Ann. Statist.*, **25**, 1014-1035.
- MARITZ, J. S., WU, M. AND STAUDTE, R. G. (1977) A location estimator based on U -statistic, *Ann. Statist.*, **5**, 779-786.
- MARKATOU, M., BASU, A. AND LINDSAY, B. G. (1998) Weighted likelihood equations with bootstrap root search, *Jour. Amer. Statist. Assoc.*, **93**, 740-750.

- MASON, D. A. AND NEWTON, M. A. (1992) A rank statistics approach to the consistency of a general bootstrap, *Ann. Statist.*, **20**, 1611-1624.
- MCCULLAGH, P. AND NELDER, J. A. (1989) *Generalised Linear Models*, second edition, Chapman and Hall, London.
- MILLER, R. G. (1964) A trustworthy jackknife, *Ann. Math. Statist.*, **35**, 1594-1605.
- MOULTON, L. H. AND ZEGER, S. L. (1989) Analysing repeated measures on generalised linear models via the bootstrap, *Biometrics*, **45**, 381-394.
- MOULTON, L. H. AND ZEGER, S. L. (1991) Bootstrapping generalised linear models, *Compu. Statist. Data Anal.*, **11**, 53-63.
- MULIERE, P. AND WALKER, S. (1998) Extending the family of Bayesian bootstraps and exchangeable urn schemes, *Jour. Royal Statist. Soc. Series B*, **60**, 175-182.
- NAVIDI, W. (1989) Edgeworth expansions for bootstrapping regression models, *Ann. Statist.*, **17**, 1472-1478.
- NIEMIRO, W. (1992) Asymptotics for M -estimators defined by convex minimization, *Ann. Statist.*, **20**, 1514-1533.
- NEWTON, M. A. AND RAFTERY, A. E. (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion), *Jour. Royal Statist. Soc., B*, **56**, 3-48.
- NEYMAN, J. AND SCOTT, E. L. (1948) Consistent estimates based on partially consistent observations, *Econometrica*, **16**, 1-32.
- OJA, H. (1983) Descriptive statistics for multivariate distribution, *Statist. Prob. Letters*, **1**, 327 - 333.
- ORTEGA, J. M. AND RHEINOLDT, W. C. (1970) *Iterative solution of nonlinear equations in several variables*, Academic Press, New York.
- POLITIS AND ROMANO (1992) A general resampling scheme for triangular arrays of α -mixing random variables with application to the problem of spectral density estimation, *Ann. Statist.*, **20**, 1985-2007.
- POLITIS, D. N. AND ROMANO, J. P. (1994) The stationary bootstrap, *Jour. Amer. Statist. Assoc.*, **89**, 1303-1313.
- POLITIS, D. N., ROMANO, J. P. AND WOLF, M. (1999) *Subsampling*, Springer-Verlag, New York.

- PORTNOY, S. (1984) Asymptotic behaviour of M -estimators of p regression parameters when p^2/n is large, I: Consistency, *Ann. Statist.*, **12**, 1298-1309.
- PORTNOY, S. (1985) Asymptotic behaviour of M -estimators of p regression parameters when p^2/n is large, II: Asymptotic normality, *Ann. Statist.*, **13**, 1403-1417.
- PORTNOY, S. (1988) Asymptotic behaviour of the empirical distribution of M -estimator residuals from a regression model with many parameters, *Ann. Statist.*, **14**, 1152-1170.
- PRAESTGAARD, J. AND WELLNER, J. A. (1993) Exchangeably weighted bootstrap of the general empirical process, *Ann. Prob.*, **21**, 2053-2086.
- QUENOUILLE, M. (1949) Approximation tests of correlation in time series. *J. Royal Statist. Soc. B*, **11**, 18-84.
- QUENOUILLE, M. (1956) Notes on bias in estimation, *Biometrika*, **43**, 353-360.
- QUMSIYEH, M. B. (1990) Edgeworth expansion in regression models, *Jour. of Mult. Anal.*, **35**, 86-101.
- RAO, C. R. AND ZHAO, L. C. (1992) Approximations to the distribution of M -estimates in linear models by randomly weighted bootstrap, *Sankhyā, Series A*, **54**, 323-331.
- RAWLINGS, J. O., PANTULA, S. G. AND DICKEY, D. A. (1998) *Applied regression analysis: a research tool*, 2nd edition, Springer-Verlag, New York.
- RUBIN, D. B. (1981) The Bayesian bootstrap, *Ann. Statist.*, **9**, 130-134.
- SAUERMAN, W. (1989) Bootstrapping the maximum likelihood estimator in high-dimensional log-linear models, *Ann. Statist.*, **17**, 1198-1216.
- SCOTT, D. A. (1992) *Multivariate density estimation*, John Wiley, New York.
- SCOTT, D. A., GORRY, G. A., HOFFMAN, R. G., BARBORIAK, J. J., GOTTO, A. M. (1980) A new approach for evaluating risk factors in coronary artery disease: a study of lipid concentrations and severity of disease in 1847 males, *Circulation*, **62**, 477-484.
- SCOTT, D. A., GOTTO, A. M., COLE, J. S., GORRY, G. A. (1978) Plasma lipids as collateral risk factors in coronary artery disease: a study of 371 males with chest pain, *Journal of Chronic Diseases*, **31**, 337-345.
- SEBER, G. A. F. AND WILD, C. J. (1989) *Nonlinear regression*, Wiley, New York.
- SERFLING, R. J. (1980) *Approximation theorems of mathematical statistics*, John Wiley, New York.

- SHAO, J. (1987) Sampling and resampling: an efficient approximation to jackknife variance estimators in linear models. *Chinese J. Appl. Prob. Statist.*, **3**, 368-379.
- SHAO, J. (1988) Consistency of the jackknife estimators of the variance of sample quantiles, *Comm. Statist. A*, **17**, 3017-3028.
- SHAO, J. (1989) The efficiency and consistency of approximations to the jackknife variance estimators, *Jour. Amer. Statist. Assoc.*, **84**, 114-119.
- SHAO, J. (1992a) One-step jackknife for M -estimators computed using Newton's method, *Ann. Inst. Statist. Math.*, **44**, 687-701.
- SHAO, J. (1992b) Consistency of least squares estimator and its jackknife variance estimator in nonlinear models, *Canadian J. Statist.*, **20**, 415-428.
- SHAO, J. (1992c) Jackknifing in generalised linear models, *Ann. Inst. Statist. Math.*, **44**, 673-686.
- SHAO, J. AND CHEN, Y. (1998) Bootstrapping sample quantiles based on complex survey data under hot deck imputation, *Statistica Sinica*, **8**, 1071-1086.
- SHAO, J. AND TU, D. (1995) *The jackknife and bootstrap*, Springer-Verlag, New York.
- SHAO, J. AND WU, C. F. J. (1987) Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Ann. Statist.*, **15**, 1563-1579.
- SHAO, J. AND WU, C. F. J. (1989) A general theory for jackknife variance estimation, *Ann. Statist.*, **17**, 1176-1197.
- SHI, J. (1998) Edgeworth expansion in partial linear models, *Acta Math. Sinica (New Series)*, **14**, 303-314.
- SHORACK, G. (1982) Bootstrapping robust regression, *Comm. Statist. A*, **11**, 961-972.
- SINGH, K. (1981) On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187-1195.
- SIMONOFF, J. S. AND TSAI, C. (1989) Jackknifing and bootstrapping quasi-likelihood estimators, *J. Statist. Compu. Simul*, **30**, 213-232.
- STUTE, W., GONZALEZ MANTEIGA, W., AND PRESEDO QUINDIMIL, M. (1998) Bootstrap approximations in model checks for regression, *Jour. Amer. Statist. Assoc.*, **93**, 141-149.
- WENG, C. S. (1989) On a second order asymptotic property of the Bayesian bootstrap mean. *Ann. Statist.* **17**. 705-710.

- WU, C. F. J. (1981) Asymptotic theory of nonlinear least squares estimation, *Ann. Statist.*, **9**, 501-513.
- WU, C. F. J. (1986) Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14**, 1295.
- WU, C. F. J. (1990) On the asymptotic properties of the jackknife histogram, *Ann. Statist.*, **18**, 1438-1452.
- YANG, S. AND PRENTICE, R. L. (1999) Semiparametric inference in the proportional odds regression model, *Jour. Amer. Statist. Assoc.*, **94**, 125-136.
- YIN, Y. Q., BAI, D., AND KRISHNAIAH, P. R. (1988) On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Prob. Theory Related Fields*, **78**, 509-521.
- ZHENG, Z. AND TU, D. (1988) Random weighting method in regression models. *Sci. Sinica Ser. A*, **31**, 1442 - 1459.
- ZHOU, S., SHEN, X., AND WOLFE, D. A. (1998) Local asymptotics for regression splines and confidence regions, *Ann. Statist.*, **26**, 1760-1782.