Estimation of the Amount of Sparsity in Normal Mixture Models

by

Yibo Wang

A thesis submitted to
the Faculty of Graduate Studies and Postdoctoral Affairs
in partial fulfilment of
the requirements for the degree of
Master of Science

in

Probability and Statistics

Ottawa-Carleton Institute for
Mathematics and Statistics

Carleton University
Ottawa, Ontario, Canada

©2016, Yibo Wang

Abstract

In this thesis, we are interested in estimating the amount of sparsity in sparse normal mixture models. The estimation problem in hand emerges naturally in the context of variable selection in high-dimensional settings. Handling this problem, we have modified the well-known procedure of Cai et al. (2007) on estimating the proportion of nonzero means in normal mixture models in such a way that the new procedure is more efficient in theory, less complicated in construction, and less asymptotic in applications. The analytical findings obtained in the thesis are supported via simulations. The simulation study testifies that the new procedure not only displays a better convergence rate but also requires smaller sample size in order to work as designed. Theorem 7 of Chapter 3 that establishes the upper bound on the risk of the newly proposed estimator is the main result of the thesis. This result shows the superiority of our estimator over the estimator of Cai et al. (2007), and it is new.

The thesis is organized as follows. The Introduction provides a brief overview of the problem of our interest. Chapter 1 gives a detailed motivation of our study. In Chapter 2, the procedure proposed by Cai et al. (2007) is presented and the relevant asymptotic results are formulated. The modified procedure, along with the necessary facts related to our construction, is introduced in Chapter 3. The main result of the thesis, Theorem 7, is also stated in Chapter 3. A small-scale simulation study is presented in Chapter 4. The proof of Theorem 7 is given in Chapter 5. The proofs of lemmas that are used to prove the main result are collected in Chapter 6. Possible directions for future study are outlined in the Conclusion. The R codes are provided in the Appendix.

Acknowledgements

I would like to thank my supervisor, Dr. Natalia Stepanova, for her guidance and support during this research and for introducing me to the interesting world of high-dimensional statistics.

I would also like to thank Drs. Vladimir Pestov, Evangelos Kranakis and Mohamedou Ould Haye for acting on my graduate committee.

I am especially grateful to my family for their great support during the last two years and for all their love throughout my life.

Contents

List of Tables	vi
List of Figures	vii
Abbreviation	viii
Introduction	1
1 Motivation	12
2 The CJL procedure	18
3 Modification of the CJL procedure	24
3.1 Preliminaries	24
3.2 Construction of the modified CJL estimator	29
3.3 Main result	31
4 Simulations	33
4.1 Algorithms	33
4.2 Numerical Results	37
5 Proof of the main result	40
6 Proofs of Lemmas	49
Conclusion	56

List of Tables

1.1	Three regions of variable selection	17
3.1	Percentage points of $H(t) = \mathbf{P}(\sup_{0 < u < 1} B(u) /q(u) \leq t)$	27
4.1	Numerical summary for the CJL estimation procedure	37
4.2	Numerical summary for the new estimation procedure	38

List of Figures

1	<i>Three regions of the β-r plane. The detection boundary, $\rho^*(\beta)$, separates the detection region from the no detection region. In the detection region, one can simply use the likelihood ratio test (LRT) to identify the alternative hypothesis H_1.</i>	4
2	<i>Three regions connected to the problem of variable selection. In the region of Exact Selection, all relevant components can be consistently recovered. In the region of Almost Full Selection, it is possible to have the number of misclassified components negligibly small as compared to the number of all relevant components. In the region of No Selection, all variable selection procedures fail completely, but for $\rho^*(\beta) < r < \beta$ it is yet possible to detect the presence of nonzero means.</i>	9
3.1	<i>Confidence bands for simulated data. The solid green line is the true cdf of a normal mixture model with $\beta = 1/3$ and $r = 3/4$. The solid blue lines above and below the green line are a 99 percent CJL confidence band. The red dashed lines are a 99 percent CsCsHM confidence band.</i>	28

Nomenclature

Abbreviation

a.s.	almost surely
iid	independently and identically distributed
cdf	cumulative distribution function
pdf	probability density function
edf	empirical distribution function
CJL	Cai-Jin-Low
EFKP	Erdős-Feller-Kolmogorov-Petrovski
CsCsHM	Csörgő-Csörgő-Horváth-Mason
SDP	standard deviation proportional

Notation

$\lim_{n \rightarrow \infty} a_n = a$	a is the limit of the sequence a_n
$\limsup_{n \rightarrow \infty} a_n$	the superior limit of the sequence a_n , $\inf_{n \geq 1} \sup_{k \geq n} a_k$
$\liminf_{n \rightarrow \infty} a_n$	the inferior limit of the sequence a_n , $\sup_{n \geq 1} \inf_{k \geq n} a_k$
$a_n \asymp b_n$	$0 < \liminf_{n \rightarrow \infty} a_n/b_n \leq \limsup_{n \rightarrow \infty} a_n/b_n < \infty$
$a_n \sim b_n$	$\lim_{n \rightarrow \infty} a_n/b_n = 1$
$\log a$	natural logarithm (base e) of the number a
$\mathbb{R}$	set of the real number
$\mathbb{R}^+$	set of the positive real numbers
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$N(0, 1)$	standard normal distribution in $\mathbb{R}$
$N(0, I_{n \times n})$	standard normal distribution in $\mathbb{R}^n$
$U(a, b)$	uniform distribution on the interval (a, b)

$\phi(\cdot)$	pdf of the standard normal distribution
$\Phi(\cdot)$	cdf of the standard normal distribution
$\mathbb{I}(A)$	indicator function of the set A
A^c	the complement of a set A
$\{a, b\}$	the set consisting of the elements a and b
$[x]$	integer part of the real number x
$f(x) = O(g(x))$ for $x \rightarrow a$	“big O ” means: $f(x)/g(x) \rightarrow C$ as $x \rightarrow a$, constant $C \neq 0$
$f(x) = o(g(x))$ for $x \rightarrow a$	“small o ” means: $f(x)/g(x) \rightarrow 0$ as $x \rightarrow a$
$\xrightarrow{\text{P}}$	convergence in probability
$\xrightarrow{\text{d}}$	convergence in distribution
$\stackrel{\text{d}}{=}$	equals in distribution
$\stackrel{\text{def}}{=}$	equals by definition
$\{B(u), 0 \leq u \leq 1\}$	Brownian bridge
$X_{(k)}$	the k th order statistic
$\forall x$	for all elements x

Introduction

Recently, there has been a spurt of statistical interest in developing test and estimation procedures for multiple testing situations where most of the component problems correspond to the null, or nothing going on, and there may be a small fraction of component problems where something small might be going on. In certain statistical applications, we may need to identify some “moderately” large coefficients, which cannot be detected with high probability individually. For example, when dealing with the Taiwanese-American Occultation Survey (TAOS), as described in Meinshausen and Rice [23], it may not be possible to tell whether an occultation has actually occurred for a particular star at a particular time, but it is feasible to estimate the fraction of occultations that have occurred for a period of time. The TAOS targeted at detecting small (of approximately 1 km in diameter) objects in the Kuiper belt, a region of the solar system beyond the orbit of Neptune. During the implementation of this project, more than 1000 small objects with diameter larger than 100 km have been detected by using large telescopes. The amount of Kuiper Belt Objects (KBOs) with diameters in the range of 500 m to 30 km, nevertheless, remained a tough task owing that the bodies of KBOs are believed to be so small that they cannot reflect enough light back to Earth to be directly observed. The aim of TAOS was to estimate the number of KBOs through observing occultation. An occultation is an event that happens when one object is hidden by another object that passes between it and the observer. Usually, it may not be possible to say whether an occultation has actually occurred for a particular star at a particular time, but it is feasible to estimate the fraction of occultation that have occurred over a period of time. Four dedicated robotic telescopes were operated simultaneously to monitor 2000 to

3000 stars in the Kuiper Belt every clear night for a couple of years and then the combined results were used to test for an occultation of each star around every 0.20 seconds, resulting in the order of 10^{11} tests per year. Then, the interest was on estimating the number of occultations, or the occultation rate so that one can obtain the distribution of KBOs. More details regarding TAOS, the project that motivated Meinshausen and Rice [23] and Cai et al. [6] for their studies, can be found in Chen et al. [8] and Liang et al. [21].

In recent years, statisticians have developed various tools and methods to handle such problems. Ingster [16] pioneered the exploration of subtle testing problems in Gaussian models (see also his book [17]). In his 1997 publication, Ingster strived to identify a range of cases where the amount of nonzero means is so small that the nonzero parts cannot exhibit noticeable effect on how the data is distributed. Through this work, Ingster presented surprisingly rich theory that was further exploited by Jager and Wellner [19], Haupt et al. [15] and many other researchers. In particular, Donoho and Jin [11] adapted the results of Ingster [16] to a special setting where there is a small fraction of non-null hypotheses. In their work, Donoho and Jin studied the problem of testing the hypothesis

$$H_0 : X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1),$$

which may be viewed as the intersection $\cap_{j=1}^n H_{0,j}$ of independently tested hypothesis $H_{0,j} : X_j \sim N(0, 1)$, versus the alternative (more precisely, versus a sequence of alternatives)

$$H_1 : X_1, \dots, X_n \stackrel{\text{iid}}{\sim} (1 - \varepsilon_n)N(0, 1) + \varepsilon_n N(\mu_n, 1),$$

where $\varepsilon_n = n^{-\beta}$ with $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ with $0 < r < 1$. In what follows, testing H_0 versus H_1 will be referred to as a *signal detection problem*. Such problem has been analyzed in great details (see, for example, Benjamini and Hochberg [2, 3], Meinshausen and Bühlmann [22], Meinshausen and Rice [23], Nettleton and Hwang [24], Storey [28], etc.). The subtlety of this problem is seen from the following well known property of a normal distribution: if $\xi_1, \xi_2, \dots$ is a sequence of iid standard normal random variables, then (see,

for example, Lemma 1 in Cui [10])

$$\mathbf{P}\left(\max_{1 \leq i \leq n} |\xi_i| \geq \sqrt{2 \log n}\right) \rightarrow 0, \quad n \rightarrow \infty. \quad (1)$$

Therefore, in general, the smaller the value of r is, the harder is to separate the observations with nonzero means from those with zero means. Thus, the problem of testing H_0 versus H_1 is a rather delicate problem. The model specified by the hypothesis H_1 is known in the literature as a *two-point normal mixture model*.

As known from the work of Ingster [16] (see also Donoho and Jin [11]), the “phase space” of possible values of (β, r) in the above testing problem, which is sometimes called the *parameter space*, is divided into two regions: the first region, known as the *detection region*, consists of those values of (β, r) for which the null and alternative hypotheses separate asymptotically; the second region, called the *no detection region*, consists of those values of (β, r) for which the null and alternative hypotheses merge asymptotically. There exist slight variations in understanding the phrase that “two hypotheses separate (merge) asymptotically”. The following definition is often used.

Definition 1. Consider testing the null hypothesis H versus the alternative hypothesis A at level α . We say that the hypotheses H and A *separate asymptotically* if there exists a consistent test procedure for testing H versus A of asymptotic level α ; otherwise, if no consistent test procedure of asymptotic level α exists, we say that the hypotheses H and A *merge asymptotically*.

In connection with testing H_0 versus H_1 , the values of β ranging from 0 to $1/2$ (the case of moderate sparsity) and the values of r that are larger than 1 are not of much interest as they lead to an easy problem (see Remark 1 below). An interesting problem is to separate H_1 from H_0 when $\beta \in (1/2, 1)$ (the case of high sparsity) and $r \in (0, 1)$.

To describe the detection region precisely, consider the following function

$$\rho^*(\beta) = \begin{cases} 0, & 0 < \beta \leq 1/2, \\ \beta - 1/2, & 1/2 < \beta \leq 3/4, \\ (1 - \sqrt{1 - \beta})^2, & 3/4 < \beta < 1. \end{cases} \quad (2)$$

This function, known in the literature as a *detection boundary*, was found by Ingster [16] who showed that if $r > \rho^*(\beta)$ then H_0 and H_1 separate asymptotically, whereas if $r < \rho^*(\beta)$ then H_0 and H_1 merge asymptotically. Thus, the detection region in the problem of testing H_0 versus H_1 is as follows:

$$\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1, \rho^*(\beta) < r < 1\}.$$

The “phase space” of the signal detection problem in hand is presented in Figure 1.

Figure 1: Three regions of the β – r plane. The detection boundary, $\rho^*(\beta)$, separates the detection region from the no detection region. In the detection region, one can simply use the likelihood ratio test (LRT) to identify the alternative hypothesis H_1 .

Remark 1. With regard to testing H_0 versus H_1 , let us note the following. Consider the calibration of the parameters ε_n and μ_n in the hypothesis H_1 as $\varepsilon_n = n^{-\beta}$, $0 < \beta < 1$, and $\mu_n = \sqrt{2r \log n}$, $0 < r < 1$. If $0 < \beta \leq 1/2$, then the hypotheses H_0 and H_1 trivially separate asymptotically for all $0 < r < 1$, see formula (2). This explains the reason why in the estimation procedure of Cai et al. [6] presented in Chapter 2 only the value of β between $1/2$ and 1 are of interest. To make the detection problem for the case of $0 < \beta \leq 1/2$ non-trivial, statisticians consider different calibration: $\varepsilon_n = n^{-\beta}$, $0 < \beta \leq 1/2$, and $\mu_n = n^{-r}$, $0 < r < 1/2$. With this calibration, the detection boundary has the form $\rho^*(\beta) = 1/2 - \beta$ for $0 < \beta \leq 1/2$, and the detection region becomes $\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta \leq 1/2, 0 < r < \rho^*(\beta)\}$. See, for example, Section 2 of Cai et al. [7].

In this thesis, we consider a two-point normal mixture model that was first studied by Ingster [16] and then by Donoho and Jin [11] and many other authors. The model has the following form (see the hypothesis H_1 above):

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} (1 - \varepsilon_n)N(0, 1) + \varepsilon_n N(\mu_n, 1), \quad (3)$$

where $\varepsilon_n = n^{-\beta}$ with $0 < \beta < 1$, and $\mu_n = \sqrt{2r \log n}$ with $0 < r < 4$. The parameters β and r are unknown. Notice that nonzero means are *sparse* and *moderately large*. In addition to testing mixtures, another problem of interest is that of variable selection. When dealing with the problem of variable selection in this setting, one aims at identifying those of the observations among $X_1, \dots, X_n$ that have nonzero means.

Using the methodology of Donoho and Jin [11], Cai et al. [6] proposed a method for constructing a lower confidence bound for the fraction of nonzero components in model (3). However, their estimation procedure is quite complex and, in order to work as designed, requires a large sample size of $n = 10^9$ and greater. In this thesis, using the result of Stepanova and Pavlenko [27] on the asymptotically correct confidence band for a continuous cumulative distribution function (cdf), we modify the idea of Cai et al. [6] and show that the new procedure, while being analytically and computationally simpler, dominates the existing

procedure. In our work, the parameter space of possible values of (β, r) is connected to a related variable selection (not the signal detection) problem with sparse components. For this problem, some values of (β, r) outside of the region $\{(\beta, r) \in \mathbb{R}^2 : 1/2 < \beta < 1, 0 < r < 1\}$ are also of great interest. Namely, the parameter space considered in this thesis is given by (see Chapter 1 for details)

$$\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1, 0 < r < 4\}.$$

Similar settings for the problem of variable selection were studied by Genovese et al. [14] and later, in a somewhat different setup, by Cui [10]. It turned out that, for the purpose of variable selection, the values of β between 0 and 1 (not only between 1/2 and 1) and the values of r between 0 and 4 (not only between 0 and 1) are all of interest. Cui [10] studied the problem of reconstruction of the components of a vector $X = (X_1, \dots, X_n)$ in a Gaussian sequence model

$$X = m + \xi, \tag{4}$$

where $\xi = (\xi_1, \dots, \xi_n) \sim N(0, I_{n \times n})$ and the mean vector $m = (m_1, \dots, m_n)$ is sparse and is of the form

$$m = a\eta, \quad a = a_n \in \mathbb{R}^+, \quad \eta = (\eta_1, \dots, \eta_n) \in H_{n,\beta},$$

with the set $H_{n,\beta}$ defined for some constants $0 < c \leq 1 \leq C < \infty$ as

$$H_{n,\beta} = \{\eta = (\eta_1, \dots, \eta_n) : \eta_j \in \{0, 1\}, \quad cn^{1-\beta} \leq \sum_{j=1}^n \eta_j \leq Cn^{1-\beta}\}. \tag{5}$$

If we allow the constants c and C in the definition of $H_{n,\beta}$ to depend on n in such a way that $c = 1 + o(1)$ and $C = 1 + o(1)$ as $n \rightarrow \infty$, then the fraction of nonzero means in model (4) becomes

$$n^{-1} \sum_{j=1}^n \eta_j \sim n^{-\beta}.$$

This makes the problem of estimating the fraction of nonzero means in model (4) asymptotically equivalent to that of estimating the fraction of nonzero means in the model specified by the hypothesis H_1 . In this model, the quantity

$$K_n = K_n(\beta) \stackrel{\text{def}}{=} [n^{1-\beta}]$$

is the number of nonzero means and $(n - K_n)$ is the number of zero means. Notice that if n is very large, that is, if $n \rightarrow \infty$, then $K_n = o(n)$. This indicates that a nonzero component of our model is *sparse*. The degree of sparsity is controlled by the *sparsity parameter* β . Again, the fact that nonzero means are only “moderately” large makes the problems of recovering and estimating the fraction of nonzero means non-trivial.

In connection with the problem of variable selection in model (4), a new “phase diagram” that partitions the parameter space $\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1, 0 < r < 4\}$ into three regions in which: (i) exact variable selection is possible (asymptotically), (ii) almost full variables selection is possible (asymptotically), and (iii) successful variable selection is impossible, has been recently obtained (see Cui [10] and Genovese et al. [14]). *Exact variable selection* refers to the situation where the set of all relevant components can be consistently recovered (asymptotically). *Almost full variable selection* stands for the possibility of having the number of misclassified components negligibly small as compared to the number of all relevant components. The latter strategy requires milder restrictions on a statistical model and can be used in the situations where exact recovery is impossible. If neither exact nor almost full variable selection can be achieved, we speak of “no variable selection” when the optimal risk is as large as the number of relevant components and any variable selection procedure fails completely.

In a particular case, when the parameter a in model (4) has the form $a = \sqrt{2r \log n}$ with $r > 0$, Theorems 2–5 in [10] (compare with Theorems 9 and 10 in [14]) give the following

partition of the parameter space $\{(\beta, r) : 0 < \beta < 1, 0 < r < 4\}$. For $\beta \in (0, 1)$, put

$$\rho_1^*(\beta) = \beta \quad \text{and} \quad \rho_2^*(\beta) = \left(1 + \sqrt{1 - \beta}\right)^2. \quad (6)$$

By analogy with the detection boundary $\rho^*(\beta)$ given in (2), we call these two functions the *selection boundaries*. If $r < \rho_1^*(\beta)$, then successful variable selection is impossible; if $\rho_1^*(\beta) < r < \rho_2^*(\beta)$, then almost full variable selection is possible (asymptotically); finally, if $r > \rho_2^*(\beta)$, then exact variable selection is possible (asymptotically). That is, variable selection (exact and almost full) can be achieved when the mean value $a = \sqrt{2r \log n}$ is not very small. One can restrict our attention to the values of r between 0 and 4 because for $r \geq 4$ exact variable selection is always possible, regardless of the value of $\beta \in (0, 1)$. Figure 2 displays three different selection regions on the β - r plane.

For the selection procedures that yield asymptotically exact and almost full component selection in model (4), we refer to Theorems 2 and 4 in Cui [10], stated as Theorems 1 and 3 in Chapter 1 of this thesis. Both these procedures have a simple thresholding nature. At the same time, being computationally simple, they are unimprovable (in the asymptotically minimax sense). Note that the exact selection procedure in Theorem 2 of Cui [10] (see Theorem 1 in Chapter 1 of this thesis) is independent of β ; at the same time, the almost full selection procedure in Theorem 4 of Cui [10] (see Theorem 3 in Chapter 1 of the present work) depends on the sparsity parameter β , which is, generally speaking, unknown. In addition, the above cited conditions in terms of the selection boundaries $\rho_1^*(\beta)$ and $\rho_2^*(\beta)$ under which exact variable selection in Theorem 1 of Chapter 1 and almost full variable selection in Theorem 3 of Chapter 1 are achievable also involve the parameter β . All this motivates us for finding an efficient estimator for the parameter β in the context of variable selection. Clearly, once we get a good estimator for $\varepsilon_n = n^{-\beta}$, we immediately get a good estimator for β . More details on optimal variable selection in a Gaussian sequence model is given in Chapter 1.

The purpose of this work is to propose a good estimator for the fraction of nonzero

Figure 2: Three regions connected to the problem of variable selection. In the region of *Exact Selection*, all relevant components can be consistently recovered. In the region of *Almost Full Selection*, it is possible to have the number of misclassified components negligibly small as compared to the number of all relevant components. In the region of *No Selection*, all variable selection procedures fail completely, but for $\rho^*(\beta) < r < \beta$ it is yet possible to detect the presence of nonzero means.

means in the framework of the above variable selection problem. Two estimators of this kind, that were introduced in connection with similar multiple testing problems, are well known: one estimator is due to Meinshausen and Rice [23] and the other one is due to Cai et al. [6]. In case of a normal mixture model, the second estimator is known to perform better as compared to the first one. At the same time, the estimator introduced by Meinshausen and Rice [23] works quite well for a larger family of mixture models, not only for the normal

mixture model as in (3). Both estimators are “strongly asymptotic”, that is, in order to work reasonably well they require very large sample sizes.

At the outset of this work, the goal was to propose, in the context of variable selection, an alternative to the estimator of Cai et al. [6] with the following three properties:

- (i) a new estimator has a simpler form and is easier to use in applications;
- (ii) a new estimator has a better rate of convergence;
- (iii) a new estimator performs reasonably well even for moderate sample sizes.

In the process of working on the thesis, we found that, with the approach taken in this work, goal (iii) is hardly achievable. Nevertheless, the new estimator introduced in Section 3.2 satisfies the desirable properties (i) and (ii).

In order to provide a good estimator for $\varepsilon_n = n^{-\beta}$ when $0 < \beta < r < 4$, we modify the estimation procedure of Cai et al. [6] by employing the result of Stepanova and Pavlenko [27] that gives a family of optimally adaptive test statistics for some class of multiple testing problems, including the problem of testing H_0 versus H_1 . The proposed estimator is computationally simpler than the Cai-Jin-Low (CJL) estimator introduced in Cai et al. [6] and, in the region $\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < r < 4\}$, has better statistical properties than the former. In the simulation study of Chapter 4, the superiority of our estimator over the CJL estimator for normal mixture models is quite visible, even though, in theory, the difference between their rates of convergence is not that big (compare the upper bound on the risk in Theorem 5 of Chapter 2 with that in Theorem 7 of Section 3.3).

The thesis is structured as follows. Chapter 1 provides detailed motivation for our study. In Chapter 2, we present the CJL estimation procedure from the work of Cai et al. [6] and provide results on its quality. Also, an important condition for successful implementation of the CJL method is discussed. In Chapter 3, the core part of this thesis, the ideas and tools behind our approach to estimating the amount of sparsity in normal mixture models are presented, the construction of the modified CJL estimator is performed, and the upper bound on its risk is established. We find that, in the region where variable selection is

possible, the new estimator dominates the CJL estimator. This finding is supported by the numerical results presented in Chapter 4. In particular, in the selection region, the numerical performance of our procedure is noticeably better than that of the CJL procedure as the (estimated) risk of the former is smaller than the (estimated) risk of the latter. In addition to that, as shown numerically, the new procedure is slightly “less asymptotic” than the CJL procedure. The proof of the upper bound on the risk of our estimator, which constitutes the main result of this work, is given in Chapter 5. Chapter 6 contains the proofs of Lemmas 1–3 that are used in the proof of Theorem 7 in Chapter 5. The results of the study and some ideas of their extensions are summarized in the Conclusion.

Chapter 1

Motivation

Over the last two decades, new methods for making efficient statistical inference in sparse high-dimensional models, where most classical methods of statistics fail to work and therefore are not applicable, have been developed. Inference problems that somehow involve a very large number of unknown parameters are called *high-dimensional problems*. Examples of the types of scenarios where high-dimensional problems arise include, among others, goodness-of-fit tests, multiple testing, classification, and variable selection. In making efficient high-dimensional inference, *sparsity* plays a key role. When one knows that many or even most of the components of the large-dimensional parameter are small, or even zero, one speaks of a sparse inference problem. The estimation problem that is handled in this thesis may be viewed as a sparse high-dimensional problem. Initially, it emerged in the context of multiple testing, as described in the Introduction, and is treated in the thesis in a slightly different context of variable selection.

As an example of a variable selection problem in high dimensions, we consider the situation described in Jensen, Chen, and Stoeckert [20]. In their experiment, the researchers have a collection of approximately 40 transcription factor proteins (TFs) that respond to regulate cellular processes in yeast by promoting or repressing transcription of specific genes. However, which of the genes in the yeast data are regulated by each of the transcription factors remains unknown. Therefore, the goal of the analysis in Jensen, Chen, and Stoeckert

[20] is to discover the corresponding network of gene–TF relationships, which is known as a gene regulatory network. Each gene, however, is regulated by only a small subset of the TFs. Under this circumstance, the problem of interest becomes the problem of variable selection in a sparse setting. The available data consists of gene expression measures for approximately 6000 genes in yeast across several hundred experiments, as well as expression measures for each of the approximately 40 transcription factors.

In order to connect the estimation problem under study to that of variable selection, we first describe the results of Cui [10] on variable selection in a Gaussian sequence model.

Let $X = (X_1, X_2, \dots, X_n)$ be a random vector distributed according to model (4). As seen from Theorems 1–4 stated below, the nuisance parameter a that appears in this model may be thought as the parameter that controls the “strength of a signal”. In this setup, a typical problem of variable selection consists in selecting the relevant components of vector m , that is, in estimating the vector $\eta = (\eta_1, \dots, \eta_n) \in H_{n,\beta}$. For estimating $\eta = (\eta_1, \dots, \eta_n)$, a selector is commonly used. Also, the loss of estimating η by means of a selector $\hat{\eta}$ is often characterized by the Hamming distance.

Definition 2. An estimator $\hat{\eta}(X) = (\hat{\eta}_1(X), \dots, \hat{\eta}_n(X))$ of a vector $\eta = (\eta_1, \dots, \eta_n) \in H_{n,\beta}$ that takes values on $\{0, 1\}^n = \underbrace{\{0, 1\} \times \dots \times \{0, 1\}}_{n \text{ times}}$ is called a *selector*.

Definition 3. The *Hamming distance* between a vector $\eta \in H_{n,\beta}$ and a selector $\hat{\eta}$ measures the distance between η and $\hat{\eta}$ by counting the number of positions at which $\hat{\eta}$ and η differ and is defined as follows:

$$|\hat{\eta} - \eta| \stackrel{\text{def}}{=} \sum_{j=1}^n |\hat{\eta}_j - \eta_j|.$$

With the Hamming distance as a loss function, a common measure of quality is the

maximum risk defined as

$$R(\hat{\eta}) \stackrel{\text{def}}{=} \sup_{\eta \in H_{n,\beta}} \mathbf{E}|\hat{\eta} - \eta|,$$

where the set $H_{n,\beta}$ is given in (5).

Definition 4. A selector $\hat{\eta} = \hat{\eta}(X)$ is called an *exact selector* (asymptotically) if

$$\limsup_{n \rightarrow \infty} R(\hat{\eta}) = 0.$$

Definition 5. A selector $\hat{\eta} = \hat{\eta}(X)$ is called an *almost full selector* (asymptotically) if

$$\limsup_{n \rightarrow \infty} n^{\beta-1} R(\hat{\eta}) = 0.$$

Let us comment on these two definitions. Roughly, a selector $\hat{\eta}$ is *exact* if as $n \rightarrow \infty$

$$R(\hat{\eta}) = o(1).$$

Such a selector can correctly identify all the relevant components of a vector X in model (4). Likewise, a selector $\hat{\eta}$ is *almost full* if as $n \rightarrow \infty$

$$R(\hat{\eta}) = o(n^{1-\beta}).$$

An almost full selector can correctly identify most of all relevant components of X , but not all of them, and its maximum risk is small as compared to the number of relevant components

Now we cite the results from Cui [10] on exact and almost full variable selection in model (4). These results provide optimal (in the asymptotically minimax sense) selectors for both regimes and yield the partition of the set of possible values of (β, a) into three regions where (i) exact variable selection is possible (asymptotically), (ii) almost full variable selection is possible (asymptotically), and (iii) successful variable selection is impossible.

Before stating the results, we present the selectors proposed by Cui [10]. For some $\Delta = \Delta_n > 0$ satisfying the condition

$$\Delta \rightarrow 0 \quad \text{and} \quad \Delta \log n \rightarrow \infty,$$

let η^* be the selector defined by the formula

$$\eta^* = (\eta_1^*, \dots, \eta_n^*), \quad \eta_j^* = \eta_j^*(X) = \mathbb{I} \left(X_j > \sqrt{(2 + \Delta) \log n} \right), \quad j = 1, \dots, n;$$

and let η^{**} be the selector given by

$$\eta^{**} = (\eta_1^{**}, \dots, \eta_n^{**}), \quad \eta_j^{**} = \eta_j^{**}(X) = \mathbb{I} \left(X_j > \sqrt{(2\beta + \Delta) \log n} \right), \quad j = 1, \dots, n.$$

It will now be seen that η^* and η^{**} are asymptotically exact and almost full selectors, respectively.

Theorem 1 (Theorem 2 on page 13 of Cui [10]). *Assume that $\beta \in (0, 1)$ and that the parameter a in model (4) satisfies the condition*

$$\liminf_{n \rightarrow \infty} \frac{a}{\sqrt{\log n}} > \sqrt{2} \left(1 + \sqrt{1 - \beta} \right). \quad (1.1)$$

Then

$$\liminf_{n \rightarrow \infty} \sup_{\eta \in H_{n, \beta}} \mathbf{E}_\eta |\eta^* - \eta| = 0,$$

that is, the selector η^* provides exact variable selection.

Theorem 2 (Theorem 3 on page 14 of Cui [10]). *Assume that $\beta \in (0, 1)$ and that the parameter a in model (4) satisfies the condition*

$$\limsup_{n \rightarrow \infty} \frac{a}{\sqrt{\log n}} < \sqrt{2} \left(1 + \sqrt{1 - \beta} \right).$$

Then

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{n,\beta}} \mathbf{E}_{\eta} |\tilde{\eta} - \eta| > 0,$$

where the infimum is over all possible selectors $\tilde{\eta}$, that is, exact variable selection is impossible.

Theorem 3 (Theorem 4 on page 15 of Cui [10]). *Assume that $\beta \in (0, 1)$ and that the parameter a in model (4) satisfies the condition*

$$\liminf_{n \rightarrow \infty} \frac{a}{\sqrt{\log n}} > \sqrt{2\beta}. \quad (1.2)$$

Then

$$\limsup_{n \rightarrow \infty} \sup_{\eta \in H_{n,\beta}} n^{\beta-1} \mathbf{E}_{\eta} |\eta^{**} - \eta| = 0,$$

that is, the selector η^{**} provides almost full variable selection.

Theorem 4 (Theorem 5 on page 16 of Cui [10]). *Assume that $\beta \in (0, 1)$ and that the parameter a in model (4) satisfies the condition*

$$\limsup_{n \rightarrow \infty} \frac{a}{\sqrt{\log n}} < \sqrt{2\beta}.$$

Then

$$\liminf_{n \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{n,\beta}} n^{\beta-1} \mathbf{E}_{\eta} |\tilde{\eta} - \eta| > 0,$$

that is, almost full variable selection is impossible.

Theorems 1–4 give the *sharp selection boundaries*, i.e., the necessary and sufficient conditions for the possibility of successful variable selection in the exact and almost full regimes. These conditions describe the variable selection problem as a whole rather than a particu-

lar method. Therefore they can serve as benchmarks for practitioners who need to design experiments.

Writing the parameter a in model (4) in the form $a = \sqrt{2r \log n}$ with $r > 0$, we arrive at the three different regions, as depicted in Figure 2, and the following summary in our Table 1.1 that is analogous to Table 1.1 in Cui [10].

Table 1.1: Three regions of variable selection

Selection type	Condition on r as a function of β	Selection threshold	Maximum risk
Exact	$\rho_2^*(\beta) < r < \infty$	$\sqrt{(2 + \delta) \log n}$	$o(1)$
Almost Full	$\rho_1^*(\beta) < r < \rho_2^*(\beta)$	$\sqrt{(2\beta + \delta) \log n}$	$o(n^{1-\beta})$
No selection	$0 < r < \rho_1^*(\beta)$	None	$O(n^{1-\beta})$

Here the functions ρ_1^* and ρ_2^* are given by (6).

Now, we are ready to spell out the motivation for the present study. Observe that the exact selection procedure based on η^* is independent of β ; at the same time, the almost full selection procedure based on η^{**} depends on the sparsity parameter β , which is, generally speaking, unknown. In addition, the conditions under which exact variable selection in Theorem 1 and almost full variable selection in Theorem 3 are achievable (see formulas (1.1) and (1.2)) also involve the parameter β . All this motivates us for finding an efficient estimator for β in the context of variable selection. Clearly, once we get a good estimator ε_n^* for $\varepsilon_n = n^{-\beta}$, we immediately obtain a good estimator β_n^* for β in the form

$$\beta_n^* = \frac{\log(1/\varepsilon_n^*)}{\log n},$$

see Remark 5 in Chapter 3. Our estimator of ε_n is a modification of the CJL estimator. In Chapter 2, we describe the CJL procedure in detail, and in Chapter 3 we introduce our estimation procedure.

Chapter 2

The CJL procedure

Return to the two-point mixture model (3) and denote by Φ and ϕ the cdf and probability density function (pdf) of a standard normal distribution $N(0, 1)$, respectively. In terms of a common cdf $F(t)$ of the observations $X_1, \dots, X_n$, the model can be rewritten as follows:

$$F(t) = (1 - \varepsilon_n)\Phi(t) + \varepsilon_n\Phi(t - \mu_n), \quad t \in \mathbb{R}, \quad (2.1)$$

where the parameters μ_n and ε_n are as before. Based on the representation of $F(t)$ as in (2.1), the CJL procedure first estimates the mean μ_n , and then uses the estimated mean to estimate ε_n , the fraction of nonzero mean normal distributions. To obtain the estimators for μ_n and ε_n , the CJL procedure suggests to follow the five steps as given below.

1. Pick two distinct points $0 \leq t < t'$ and consider the function

$$D(\mu; t, t') \stackrel{\text{def}}{=} \frac{\Phi(t) - \Phi(t - \mu)}{\Phi(t') - \Phi(t' - \mu)}. \quad (2.2)$$

The function $D(\mu; t, t')$ is a positive and continuous function of μ . Moreover, it can be shown (see Lemma 8.1 of Cai et al. [5]) that $D(\mu; t, t')$ is strictly decreasing in $\mu > 0$

for any $t < t'$. Hence the parameters ε_n and μ_n are uniquely determined by

$$\varepsilon_n = \frac{\Phi(t) - F(t)}{\Phi(t) - \Phi(t - \mu_n)} \quad \text{and} \quad D(\mu_n; t, t') = \frac{\Phi(t) - F(t)}{\Phi(t') - F(t')},$$

respectively. Also, it is easy to see that

$$\inf_{\mu > 0} D(\mu_n; t, t') = \frac{\Phi(t)}{\Phi(t')} < \frac{\Phi(t) - F(t)}{\Phi(t') - F(t')} < \sup_{\mu > 0} D(\mu_n; t, t') = \frac{\phi(t)}{\phi(t')}. \quad (2.3)$$

2. Let

$$\mathbb{F}_n(t) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq t), \quad t \in \mathbb{R}, \quad (2.4)$$

be the empirical distribution function (edf) based on $X_1, \dots, X_n$. Though, in theory, the estimators of ε_n and μ_n can be obtained via replacing $F(t)$ and $F(t')$ by $\mathbb{F}_n(t)$ and $\mathbb{F}_n(t')$, respectively, in practice, this idea is not good because the resulting estimators depend strongly on the choice of t and t' and for some choices of t and t' may perform poorly. Therefore the CJL procedure uses slightly biased estimators of $F(t)$ and $F(t')$ to obtain good estimators of ε_n and μ_n . Namely, the CJL procedure prescribes to choose a value a_n (in order to get an efficient estimator for ε_n , a_n must be chosen to satisfy $a_n \rightarrow \infty$ as $n \rightarrow \infty$), and solve the equation

$$\frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}} = a_n, \quad (2.5)$$

Solving this equation yields two solutions, $\mathbb{F}_{a_n}^+(t)$ and $\mathbb{F}_{a_n}^-(t)$, given by

$$\mathbb{F}_{a_n}^\pm(t) = \frac{2\mathbb{F}_n(t) + a_n^2/n \pm \sqrt{a_n^2/n + (4\mathbb{F}_n(t) - 4F_n^2(t))(a_n/\sqrt{n})}}{2(1 + a_n^2/n)}. \quad (2.6)$$

Note that, by construction, $\mathbb{F}_{a_n}^-(t) \leq F(t) \leq \mathbb{F}_{a_n}^+(t)$ if and only if $\frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}} \leq a_n$.

Now, by taking a_n to be the $(1 - \alpha)$ th quantile of $\sup_{t \in S_n} \frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}}$, one arrives at

the asymptotic $100(1 - \alpha)\%$ confidence band for $F(t)$ of the form $[\mathbb{F}_{a_n}^-(t), \mathbb{F}_{a_n}^+(t)]$ for $t \in S_n$. In Cai et al. [6], the set S_n is taken to be the interval $[0, \sqrt{2 \log n}]$ (see Section 7 in Cai et al. [6] and Section 4.1 of this thesis for details).

3. Observe that in Cai et al. [6] model (3) is considered with $0 < r < 1$ only, giving the restriction $0 < \mu_n < \sqrt{2 \log n}$. Therefore the CJL procedure suggests further to consider the interval $[0, \sqrt{2 \log n}]$, pick equidistant grid points $t_j = (j - 1)/\sqrt{2 \log n}$, $1 \leq j \leq 2 \log n + 1$, and apply steps 1 and 2 to each pair t_j and t_{j+1} of adjacent points. Let $\hat{\mu}_{a_n}^j$ be the solution of the equation in step 1, that is, of the equation

$$D(\mu; t_j, t_{j+1}) = \frac{\Phi(t_j) - \mathbb{F}_{a_n}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{a_n}^-(t_{j+1})},$$

where the functions $\mathbb{F}_{a_n}^+(t)$ and $\mathbb{F}_{a_n}^-(t)$ are given by (2.6), provided that such a solution exists. If there is no solution, one sets $\hat{\varepsilon}_{a_n}^j$, which is defined below in step 4, equal to zero. Note that for those t 's for which the condition $\frac{\sqrt{n}|\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1-F(t))}} \leq a_n$ holds, one has that with high probability $\mathbb{F}_{a_n}^-(t) \leq F(t) \leq \mathbb{F}_{a_n}^+(t)$ and hence

$$\frac{\Phi(t_j) - \mathbb{F}_{a_n}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{a_n}^-(t_{j+1})} \leq \frac{\Phi(t) - F(t)}{\Phi(t') - F(t')},$$

which yields $\hat{\mu}_{a_n}^j > \mu_n$. That is, for all j the estimator $\hat{\mu}_{a_n}^j$ (if exists) overestimates the mean value μ_n with high probability.

4. Having produced the (biased) estimator $\hat{\mu}_{a_n}^j$ of μ_n , the CJL procedure then says to use step 1 to obtain an estimator for ε_n as follows:

$$\hat{\varepsilon}_{a_n}^j = \frac{\Phi(t_j) - \mathbb{F}_{a_n}^+(t_j)}{\Phi(t_j) - \Phi(t_j - \hat{\mu}_{a_n}^j)}.$$

By construction, one sees that $\hat{\varepsilon}_{a_n}^j \leq \varepsilon_n$ with high probability.

5. The resulting CJL estimator $\hat{\varepsilon}_{a_n}^*$ of ε_n is defined by taking the maximum of $\hat{\varepsilon}_{a_n}^j$:

$$\hat{\varepsilon}_{a_n}^* = \max_{1 \leq j \leq 2 \log n} \hat{\varepsilon}_{a_n}^j. \quad (2.7)$$

Remark 2. We should note that, as recommended on page 2427 of Cai et al. [6], in step 3 of the above procedure, for each pair (t_j, t_{j+1}) the condition, cf. (2.3),

$$\frac{\Phi(t_j)}{\Phi(t_{j+1})} < \frac{\Phi(t_j) - \mathbb{F}_{a_n}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{a_n}^-(t_{j+1})} < \frac{\phi(t_j)}{\phi(t_{j+1})}. \quad (2.8)$$

must be checked. If it does not hold, $\hat{\mu}_{a_n}^j$ is not a good estimator of μ_n and then, in step 4, one should take $\hat{\varepsilon}_{a_n}^j = 0$. We completely agree with this recommendation as we have seen the importance of the validity of (2.8) in our empirical study whose results are presented in Section 4.2. We also wish to note that (2.8) consistently holds only for very large n such as $n = 10^7$ and greater. The necessity to have condition (2.8) valid is the main reason why the CJL estimator $\hat{\varepsilon}_{a_n}^*$ is “strongly asymptotic”, that is, works well only in high-dimensional settings.

Going through steps 1 to 5 of the above algorithm gives an estimator $\hat{\varepsilon}_{a_n}^*$ that, with high probability, underestimates the fraction ε_n of nonzero means in model (3). Let us now cite two results from Cai et al. [6] showing that the estimator $\hat{\varepsilon}_{a_n}^*$ as in (2.7) is nearly optimal. Nearly optimality refers to the fact that, inside the detection region, for all values of $1/2 < \beta < 1$ and $0 < r < 1$ the upper bound on the quadratic risk of the ratio $\hat{\varepsilon}_{a_n}^*/\varepsilon_n$ differs from the lower bound on the corresponding minimax risk by a multiple of a logarithmic factor. That is, in assessing the quality of their estimator, Cai et al. [6] follow the *minimax approach* that prescribes the statistician to look at the worst that can happen and try to minimize the maximal expected loss.

Theorem 5 (Theorem 4.1 in Cai et al. (2007)). *Suppose that $F(t)$ is defined as in (2.1) with $\mu_n = \sqrt{2r \log n}$, where $0 < r < 1$, $1/2 < \beta < 1$, and $r > \rho^*(\beta)$ so that (β, r) falls into the interior part of the detection region. Set $a_n = 4\sqrt{2\pi}(\log n)^{3/2}$. Then for all*

sufficiently large n the estimator $\hat{\varepsilon}_{a_n}^*$ in (2.7) satisfies

$$\mathbf{E} \left(\frac{\hat{\varepsilon}_{a_n}^*}{\varepsilon_n} - 1 \right)^2 \leq \begin{cases} C(\beta, r) (\log n)^{5.5} n^{-1-2r+2\beta}, & \beta \geq 3r, \\ C(\beta, r) (\log n)^{5.5} n^{-1+(\beta+r)^2/(4r)}, & r < \beta < 3r, \\ C(\beta, r) (\log n)^4 n^{-1+\beta}, & \beta \leq r, \end{cases}$$

where $C(\beta, r)$ is a generic constant depending on β and r .

The next theorem says that, for a relatively small region of the values of (ε_n, μ_n) , the minimax risk is bounded from below by a quantity that differs from the upper bound of Theorem 5 by a logarithmic factor. Thus, from the asymptotically minimax point of view, the estimator $\hat{\varepsilon}_{a_n}^*$ is a good estimator.

Theorem 6 (Theorem 4.2 in Cai et al. (2007)). *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} (1 - \varepsilon_n) N(0, 1) + \varepsilon_n N(\mu_n, 1)$. For $0 < r < 1$, $1/2 < \beta < 1$, $a_1, a_2 > 0$ and $b_2 > b_1 > 0$, set $\Omega_n = \{(\varepsilon_n, \mu_n) : b_1 n^{-\beta} \leq \varepsilon_n \leq b_2 n^{-\beta}, \sqrt{2r \log n} - a_1/\log n \leq \mu_n \leq \sqrt{2r \log n} + a_2/\log n\}$. Then for all sufficiently large n and some constant $c > 0$*

$$\inf_{\tilde{\varepsilon}_n} \sup_{(\varepsilon_n, \mu_n) \in \Omega_n} \mathbf{E} \left(\frac{\tilde{\varepsilon}_n}{\varepsilon_n} - 1 \right)^2 \geq \begin{cases} c (\log n) n^{-1-2r+2\beta}, & \beta \geq 3r, \\ c (\log n)^{5/2} n^{-1+(\beta+r)^2/(4r)}, & r < \beta < 3r, \\ cn^{-1+\beta}, & \beta \leq r, \end{cases}$$

where the infimum is over all possible estimators $\tilde{\varepsilon}_n$ of ε_n .

Remark 3. Under the assumptions of Theorem 5, when $1/2 < \beta \leq r < 1$, the rate of convergence of the normalized estimator $\hat{\varepsilon}_{a_n}^*$ is $r_n = n^{-(1-\beta)/2} \log^2 n$, which means that for a finite constant $C = C(\beta, r) > 0$ and all sufficiently large n

$$\sup_{(\varepsilon_n, \mu_n) \in \mathcal{P}_n} \mathbf{E} \left(\frac{\hat{\varepsilon}_{a_n}^*}{\varepsilon_n} - 1 \right)^2 \leq Cr_n^2,$$

where $\mathcal{P}_n = \{(\varepsilon_n, \mu_n) : \varepsilon_n = n^{-\beta} \text{ with } 1/2 < \beta < 1, r_n = \sqrt{2r \log n} \text{ with } 0 < r < 1\}$. Next, Theorem 6 shows that for all sufficiently large n , the *minimax risk* associated with the set

$\Omega_n \subset \mathcal{P}_n$ satisfies

$$\inf_{\tilde{\varepsilon}_n} \sup_{(\varepsilon_n, \mu_n) \in \Omega_n} \mathbf{E} \left(\frac{\tilde{\varepsilon}_n}{\varepsilon_n} - 1 \right)^2 \geq cn^{-(1-\beta)}$$

with some constant $c > 0$. This implies that under the assumptions of Theorem 5 the estimator $\hat{\varepsilon}_{a_n}^*$ nearly (up to a logarithmic factor) attains the optimal rate of convergence $n^{-(1-\beta)/2}$ associated with the set Ω_n .

Remark 4. Inspection of the proofs of Theorems 5 and 6 that are given by Cai et al. [6] shows that the results of these theorems established for the region $1/2 < \beta \leq r < 1$ continue to hold for a larger region $0 < \beta \leq r < 1$.

In this thesis, keeping in mind the problem of variable selection rather than that of signal detection, we are interested in the region $0 < \beta < r < 4$. In Chapter 3 we establish an analog of Theorem 5 when $\beta < r$ and show that, in the region where variable selection is possible, our estimator dominates the CJL estimator $\hat{\varepsilon}_{a_n}^*$.

Chapter 3

Modification of the CJL procedure

3.1 Preliminaries

As seen from equation (2.5), statistical properties of the CJL estimator $\hat{\varepsilon}_{a_n}^*$ depend on the properties of the empirical process $\{\sqrt{n}(\mathbb{F}_n(t) - F(t)), -\infty < t < \infty\}$ standardized by the weight function $\delta(F(t))$, where $\delta(u)$ is the so-called *standard deviation proportional (SDP)* weight function

$$\delta(u) = \sqrt{u(1-u)}, \quad 0 < u < 1.$$

The standardized process $\{\sqrt{n}(\mathbb{F}_n(t) - F(t)) / \sqrt{F(t)(1-F(t))}, -\infty < t < \infty\}$ inflates when $F(t)$ approaches zero and one, creating additional technical problems when constructing the CJL estimator $\hat{\varepsilon}_{a_n}^*$ and examining its statistical properties. For this reason, the quantity a_n in (2.6) has to be taken tending to infinity (see Theorem 5 in Chapter 2). As shown in Stepanova and Pavlenko [27], for certain goodness-of-fit testing problems, a good alternative to the sup-functional

$$\sup_{0 < F(t) < 1} \frac{\sqrt{n}(\mathbb{F}_n(t) - F(t))}{\sqrt{F(t)(1-F(t))}}, \quad (3.1)$$

that is used by Cai et al. [6] for the construction of their estimator $\hat{\varepsilon}_{a_n}^*$, is provided by the functional (see Csörgő et al. [9])

$$\sup_{0 < F(t) < 1} \frac{\sqrt{n} (\mathbb{F}_n(t) - F(t))}{q(F(t))}, \quad (3.2)$$

with weight function

$$q(u) = \sqrt{u(1-u) \log \log (1/(u(1-u)))}, \quad 0 < u < 1, \quad (3.3)$$

that belongs to the family of *Erdős-Feller-Kolmogorov-Petrovski (EFKP) upper-class functions* of a Brownian bridge (for the precise definition of EFKP upper-class functions, examples, and their use in the theory of empirical processes, we refer to Csörgő et al. [9]).

Under H_0 , the sup-functional in (3.1) tends to infinity in probability and even almost surely. The almost sure rate at which the statistic in (3.1) blows up is considered in Chapter 16 of Shorack and Wellner [26]. Note also that, under H_0 , when normalized as in Eicker [13] and Jaeschke [18], the statistic in (3.1) has an extreme value distribution in the limit. However, the convergence of the null distribution of (3.1) to the extreme value distribution is, in general, slow. For more details, we refer to Section 2 of Stepanova and Pavlenko [27]. At the same time, the sup-functional in (3.2) is finite with probability one (see Csörgő et al. [9] for details). Its null limit distribution is given by the following result.

Fact 1 (Consequence of Theorem 4.2.3 of Csörgő et al. [9]). Let $\{B(u), 0 \leq u \leq 1\}$ be a Brownian bridge and let the function $q(u)$ be given by (3.3). Then as $n \rightarrow \infty$

$$\sup_{0 < F(t) < 1} \frac{\sqrt{n} |\mathbb{F}_n(t) - F(t)|}{q(F(t))} \xrightarrow{d} \sup_{0 < u < 1} |B(u)|/q(u).$$

The convergence result in Fact 1 continues to hold when $q(F(t))$ is replaced by $q(\mathbb{F}_n(t))$ (see Remark 3.1 in Stepanova and Pavlenko [27]). The distribution of the limiting random variable $\sup_{0 < u < 1} |B(u)|/q(u)$ was tabulated in Eastwood and Eastwood [12] and later, using somewhat different method, in Orasch and Pouliot [25].

The key difference between the procedure proposed in this thasis and the CJL procedure is that we use a simultaneous envelope, or confidence band, based on the sup-functional as in (3.2) rather than on the sup-functional as in (3.1). This confidence band has been recently studied in Stepanova and Pavlenko [27] and found to perform very well.

As in Stepanova and Pavlenko [27], using the notation

$$X_{(1)} = \min(X_1, \dots, X_n), \quad X_{(n)} = \max(X_1, \dots, X_n),$$

where $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F(t)$, setting

$$H(t) = \mathbf{P} \left(\sup_{0 < u < 1} |B(u)|/q(u) \leq t \right), \quad (3.4)$$

and denoting by c_α the $(1 - \alpha)$ th quantile of the limit cdf $H(t)$, we obtain

$$\begin{aligned} 1 - \alpha &= \lim_{n \rightarrow \infty} \mathbf{P} \left(\sup_{0 < F(t) < 1} \frac{\sqrt{n} |\mathbb{F}_n(t) - F(t)|}{q(\mathbb{F}_n(t))} \leq c_\alpha \right) \\ &= \lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbb{F}_n(t) - \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)) \leq F(t) \leq \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)), \forall t \in [X_{(1)}, X_{(n)}] \right). \end{aligned}$$

Since $F(t)$ takes on its values in $[0, 1]$, it now follows that

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P} \left(\max \left\{ 0, \mathbb{F}_n(t) - \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)) \right\} \leq F(t) \leq \right. \\ \left. \leq \min \left\{ 1, \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)) \right\}, \forall t \in [X_{(1)}, X_{(n)}] \right) = 1 - \alpha. \quad (3.5) \end{aligned}$$

This gives an asymptotically correct $100(1 - \alpha)\%$ confidence band $[\mathbb{F}_{n,\alpha}^-(t), \mathbb{F}_{n,\alpha}^+(t)]$, called in [27] the *Csörgő–Csörgő–Horváth–Mason (CsCsHM) confidence band*, for $F(t)$ on the interval $t \in [X_{(1)}, X_{(n)}]$, where

$$\mathbb{F}_{n,\alpha}^-(t) = \max \left\{ 0, \mathbb{F}_n(t) - \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)) \right\}, \quad \mathbb{F}_{n,\alpha}^+(t) = \min \left\{ 1, \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)) \right\}, \quad (3.6)$$

and, given $\alpha \in (0, 1)$, the value of c_α is obtained from Table III in Orasch and Pouliot [25]. For instance, $c_{0.05} = 4.57$. Table 3.1 below copies a part of Table III from Orasch and Pouliot [25].

Table 3.1: Percentage points of $H(t) = \mathbf{P}(\sup_{0 < u < 1} |B(u)|/q(u) \leq t)$

c_α	$H(c_\alpha)$	α
2.80	0.50	0.50
4.12	0.90	0.1
4.19	0.91	0.09
4.25	0.92	0.08
4.35	0.93	0.07
4.45	0.94	0.06
4.57	0.95	0.05
4.72	0.96	0.04
4.89	0.97	0.03
5.16	0.98	0.02
5.53	0.99	0.01

Thus, in our estimating ε_n , the role of the functions $\mathbb{F}_{a_n}^-(t)$ and $\mathbb{F}_{a_n}^+(t)$ as in (2.6) will be taken by the functions $\mathbb{F}_{n,\alpha}^-(t)$ and $\mathbb{F}_{n,\alpha}^+(t)$ as defined in (3.6). In Stepanova and Pavlenko [27], the asymptotically correct $100(1 - \alpha)\%$ confidence band $[\mathbb{F}_{n,\alpha}^-(t), \mathbb{F}_{n,\alpha}^+(t)]$ was numerically compared to two common confidence bands (at the same level of confidence): one is based on the convergence result for the two-sided Kolmogorov–Smirnov test statistics and the other is on the 1979 results of Eicker [13] and Jaeschke [18]. Numerical simulations showed that, even for moderate sample sizes, when compared to the Kolmogorov–Smirnov confidence band, the CsCsHM confidence band is of the same length “in the middle” and is shorter on the tails. Also, the CsCsHM confidence band outperforms the Eicker–Jaeschke confidence band “in the middle” and does a similar job on the tails.

It is also important for us that the CsCsHM confidence band given by (3.6) works better as compared the confidence band in (2.6) proposed in Cai et al. [6]. As seen from Figure 3.1, the CsCsHM confidence band outperforms the confidence band from Cai et al. [6].

Figure 3.1: *Confidence bands for simulated data. The solid green line is the true cdf of a normal mixture model with $\beta = 1/3$ and $r = 3/4$. The solid blue lines above and below the green line are a 99 percent CIL confidence band. The red dashed lines are a 99 percent CsCsHM confidence band.*

Now, let us cite one well-known probability inequality (see, for example, Appendix A.2 in Bickel and Doksum [4]).

Fact 2 (Bonferroni's Inequality). For n events $A_1, A_2, \dots, A_n$ we have

$$\mathbf{P} \left(\bigcap_{i=1}^n A_i \right) \geq 1 - \sum_{i=1}^n \mathbf{P}(A_i^c).$$

We shall also need the following result that provides an asymptotic confidence band for the underlying cdf $F(t)$ on the interval $[0, \mu_n]$. To prove this result, relation (3.5) and Fact 2 will be used.

Lemma 1. Let $X_1, X_2, \dots$ be a sequence of iid random variables from a two-point mixture model having a common cdf $F(t)$ as in (2.1) with $\varepsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$, where $0 < \beta < 1$ and $0 < r < 4$. Then for any $\alpha \in (0, 1)$

$$\lim_{n \rightarrow \infty} \mathbf{P} \left(\mathbb{F}_{n,\alpha}^-(t) \leq F(t) \leq \mathbb{F}_{n,\alpha}^+(t), \forall t \in [0, \mu_n] \right) \geq 1 - 2\alpha,$$

where $\mathbb{F}_{n,\alpha}^+(t)$ and $\mathbb{F}_{n,\alpha}^-(t)$ are as in (3.6).

The proof of Lemma 1 is given in Chapter 6.

3.2 Construction of the modified CJL estimator

With Lemma 1 available, we are now ready to present our modification of the CJL procedure. Namely, we propose the following algorithm for constructing our estimator.

- (a) Pick small $\alpha \in (0, 1)$ and consider $[\mathbb{F}_{n,\alpha}^-(t), \mathbb{F}_{n,\alpha}^+(t)]$, an asymptotically correct $100(1 - \alpha)\%$ confidence band for $F(t)$ defined by formula (3.6). Take two points $0 < t < t'$ and define $\hat{\mu}_n = \hat{\mu}_{n,\alpha}$ as a solution (if exists) of the equation

$$D(\mu; t, t') = \frac{\Phi(t) - \mathbb{F}_{n,\alpha}^+(t)}{\Phi(t') - \mathbb{F}_{n,\alpha}^-(t')}. \quad (3.7)$$

Note that with high probability (of at least $1 - 2\alpha$) one has $\hat{\mu}_n \geq \mu_n$ for all large enough

n . Then, we set

$$\hat{\varepsilon}_n = \hat{\varepsilon}_{n,\alpha} = \frac{\Phi(t) - \mathbb{F}_{n,\alpha}^+(t)}{\Phi(t) - \Phi(t - \hat{\mu}_n)} \quad (3.8)$$

if the solution $\hat{\mu}_n$ exists and $\hat{\varepsilon}_n = 0$ otherwise. As easily seen, with high probability (of at least $1 - 2\alpha$) one has $\hat{\varepsilon}_n \leq \varepsilon_n$ for all large enough n .

(b) Consider the interval $[0, \sqrt{8 \log n}]$, pick equidistant grid points

$$t_j = (j - 1)/\sqrt{8 \log n}, \quad 1 \leq j \leq 8 \log n + 1,$$

and apply the procedure in step (a) to each pair of adjacent points $(t, t') = (t_j, t_{j+1})$.

Let the index J_0 be such that

$$t_{J_0+1} < X_{(n)} \leq t_{J_0+2}.$$

Then, for all $j = 1, \dots, J_0$, define the estimator $\hat{\mu}_n^j = \hat{\mu}_{n,\alpha}^j = \hat{\mu}_{n,\alpha}(t_j, t_{j+1})$ of μ_n as in step (a), that is, as a solution (if exists) of the equation

$$D(\mu; t_j, t_{j+1}) = \frac{\Phi(t_j) - \mathbb{F}_{n,\alpha}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{n,\alpha}^-(t_{j+1})}.$$

For each obtained estimator $\hat{\mu}_n^j$, we need to check whether or not the following inequalities hold true (see Remark 2 in Chapter 2):

$$\frac{\Phi(t_j)}{\Phi(t_{j+1})} \leq \frac{\Phi(t_j) - \mathbb{F}_{n,\alpha}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{n,\alpha}^-(t_{j+1})} \leq \frac{\phi(t_j)}{\phi(t_{j+1})}. \quad (3.9)$$

If (3.9) does not hold, then equation does not give a good estimate for μ_n and we take $\hat{\varepsilon}_n^j = 0$. If the inequalities in (3.9) hold, we put

$$\hat{\varepsilon}_n^j = \hat{\varepsilon}_{n,\alpha}(t_j, t_{j+1}) = \frac{\Phi(t_j) - \mathbb{F}_{n,\alpha}^+(t_j)}{\Phi(t_j) - \Phi(t_j - \hat{\mu}_n^j)}. \quad (3.10)$$

By construction, with high probability $\hat{\varepsilon}_n^j \leq \varepsilon_n$ provided n is sufficiently large.

- (c) Finally, having obtained the estimators $\hat{\varepsilon}_n^j$, $j = 1, \dots, J_0$, we define an estimator $\hat{\varepsilon}_n$ of ε_n by the formula

$$\hat{\varepsilon}_n = \hat{\varepsilon}_{n,\alpha} = \max_{1 \leq j \leq J_0} \hat{\varepsilon}_n^j. \quad (3.11)$$

Clearly, for all sufficiently large n , one has $\hat{\varepsilon}_n \leq \varepsilon_n$ with high probability. In the sequel, we will write $\hat{\varepsilon}_n$ instead of $\hat{\varepsilon}_{n,\alpha}$, suppressing the dependence of $\hat{\varepsilon}_n$ on the value of α .

3.3 Main result

In connection with the proposed estimator $\hat{\varepsilon}_n$, we have the following upper bound on the risk of $\hat{\varepsilon}_n/\varepsilon_n$. Let us stress that the estimator $\hat{\varepsilon}_n$ as in (3.11) is dependent on α , where $(1 - \alpha)$ is a prescribed confidence level of the CsCsHM confidence band in step (a) of our construction in Section 3.2. To make our estimator analytically more efficient, we need to take $\alpha = \alpha_n \rightarrow 0$ moderately (polynomially) fast, for example, $\alpha = O(n^{-2})$.

Theorem 7. *Let $X_1, X_2, \dots$ be a sequence of iid random variables from a continuous distribution with cdf $F(t)$ as in (2.1), where $\varepsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$ with $0 < \beta < r < 4$ so that (β, r) falls into the interior of the region where variable selection is possible. Then, for any $\alpha \in (0, 1)$ and all sufficiently large n , the estimator $\hat{\varepsilon}_n$ as in (3.11) satisfies*

$$\mathbf{E} \left(\frac{\hat{\varepsilon}_n}{\varepsilon_n} - 1 \right)^2 \leq 2\alpha n^{2\beta} + \frac{C(\beta, r, \alpha) \log n \times \log \log n}{n^{1-\beta}}, \quad (3.12)$$

where $C(\beta, r, \alpha)$ is a positive constant depending on β , r , and α . Next, choosing $\alpha = \alpha_n = o(n^{-1-\beta})$ decaying to zero at a polynomial rate gives that for all sufficiently large n

$$\mathbf{E} \left(\frac{\hat{\varepsilon}_n}{\varepsilon_n} - 1 \right)^2 \leq \frac{C(\beta, r) \log n \times \log \log n}{n^{1-\beta}}, \quad (3.13)$$

where $C(\beta, r)$ is a positive constant depending on β and r .

Comparing the upper bound (3.13), that is true for all β and r such that $0 < \beta < r < 4$, with the statement of Theorem 5 for the region where $0 < \beta < r < 1$ and noting that $\log \log n = o(\log^p n)$ for any $p > 0$, we conclude that, in the region where variable selection is possible, the new estimator $\hat{\varepsilon}_n$ dominates the CJL estimator $\hat{\varepsilon}_{a_n}^*$. Also, in view of Theorem 6, the estimator $\hat{\varepsilon}_n$ is nearly *rate optimal*. The proof of Theorem 7 will be given in Chapter 5.

Remark 5. By Theorem 7, the estimator

$$\hat{\beta}_n = \frac{\log(1/\hat{\varepsilon}_n)}{\log n},$$

which is constructed by taking a sufficiently small value of $\alpha = \alpha_n$, for example, $\alpha = O(n^{-2})$, in step (a) of the algorithm of Section 3.2, is a consistent estimator of the sparsity parameter β . Indeed, by (3.13) and Markov's inequality, for any $\epsilon > 0$

$$\mathbf{P}(|\hat{\varepsilon}_n/\varepsilon_n - 1| \geq \epsilon) \leq \frac{\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2}{\epsilon^2} \rightarrow 0, \quad n \rightarrow \infty,$$

From this, for any $\epsilon > 0$

$$\begin{aligned} \mathbf{P}(|\hat{\beta}_n - \beta| \geq \epsilon) &= \mathbf{P}\left(\left|\frac{\log(1/\hat{\varepsilon}_n)}{\log n} - \beta\right| \geq \epsilon\right) \\ &= \mathbf{P}(|\log(1/\hat{\varepsilon}_n) - \beta \log n| \geq \epsilon \log n) = \mathbf{P}(|\log(\hat{\varepsilon}_n/\varepsilon_n)| \geq \epsilon \log n) \\ &= 1 - \mathbf{P}(-\epsilon \log n < \log(\hat{\varepsilon}_n/\varepsilon_n) < \epsilon \log n) \\ &= 1 - \mathbf{P}(n^{-\epsilon} - 1 < \hat{\varepsilon}_n/\varepsilon_n - 1 < n^\epsilon - 1) \\ &= 1 - (1 + o(1)) = o(1), \quad n \rightarrow \infty. \end{aligned}$$

where, to get the last equality, we used the facts that $n^{-\epsilon} - 1 \rightarrow -1$ and $n^\epsilon - 1 \rightarrow \infty$ as $n \rightarrow \infty$. Thus, we conclude that

$$\hat{\beta}_n \xrightarrow{\mathbf{P}} \beta, \quad n \rightarrow \infty.$$

Chapter 4

Simulations

To assess the performance of the estimation procedures presented in Chapter 2 and introduced in Chapter 3, we carry out a small-scale simulation study and compare numerically the normalized risk of $\hat{\varepsilon}_n$ as in (3.11) with that of the CJL estimator $\hat{\varepsilon}_{a_n}^*$ as in (2.7). The mean squared error is used to quantify the quality of estimation and highlight the advantages of employing the procedure proposed in this thesis. In our experiment, we simulate n random samples with cdf given by formula (2.1). We pick sample size $n = 10^8$ so that our model is a sparse high-dimensional model. One choice of (β, r) such that $0 < \beta < r < 1$ and various values of $\alpha \in (0, 1)$ are explored in our experiment. The simulation study is run using R statistical packages. The R code of all the functions is provided in the Appendix.

4.1 Algorithms

We now present two algorithms for calculating and comparing the observed values and estimated risks of our estimator $\hat{\varepsilon}_n$ and the CJL estimator $\hat{\varepsilon}_{a_n}^*$. The comparison is based on $M = 10$ independent cycles of simulations.

Before presenting the first algorithm, we shall remark on the choice of the quantity a_n , which in the CJL procedure may vary depending on a purpose, see Section 3.1 of Cai et al. [6]. For instance, to prove their Theorem 4.1 (see Theorem 5 in Chapter 1), Cai et al.

[6] take $a_n = 4\sqrt{2\pi} \log^{3/2} n$ regardless of the value of α , whereas for the simulation study they use a_n that is found as follows. For a chosen $\alpha = \alpha_n \in (0, 1)$, let a_n be such that $\mathbf{P}(Y_n \leq a_n) = 1 - \alpha$, where

$$Y_n \stackrel{d}{=} \max_{0 \leq t \leq \sqrt{2 \log n}} \frac{\sqrt{n} |\mathbb{F}_n(t) - F(t)|}{\sqrt{F(t)(1 - F(t))}},$$

which can be equivalently written as

$$Y_n \stackrel{d}{=} \max_{F(0) \leq t \leq F(\sqrt{2 \log n})} \frac{\sqrt{n} |\mathbb{U}_n(t) - t|}{\sqrt{t(1 - t)}},$$

where $\mathbb{U}_n(t)$ is the edf based on n independent uniform $U(0, 1)$ observations. Since the cdf $F(t)$ in (2.1) depends on the parameters β and r that are generally unknown, Cai et al. [6] suggest to use the statistic

$$W_n \stackrel{d}{=} \max_{\Phi(0) \leq t \leq \Phi(\sqrt{2 \log n})} \frac{\sqrt{n} |\mathbb{U}_n(t) - t|}{\sqrt{t(1 - t)}},$$

for which, as they found numerically,

$$\mathbf{P}(W_n \leq a_n) \approx 1 - \alpha, \tag{4.1}$$

to find a_n . For several choices of α , the (approximate) value of a_n , obtained as the $(1 - \alpha)$ th quantile of W_n , are given in Table 2 of Cai et al. [6]. We use this table to obtain the values of normalized a_n in Table 4.1 below. Such a variation in the choice of a_n in the CJL procedure is due to inappropriate choice of the confidence band as in (2.6). This choice complicates the whole construction, makes the proofs of main results in Cai et al. [6] unnecessary long, and results in technically not very simple simulations. The modified procedure proposed in this thesis, with the CsCsHM confidence band (3.6) instead of the confidence band (2.6), is free of the above listed disadvantages. In the context of estimating $\varepsilon_n = n^{-\beta}$, the superiority of using the CsCsHM confidence band over using the confidence band as in (2.6) has been

explained in Section 3.1. This is also seen from the results in Tables 4.1 and 4.2 below.

Algorithm 1

1. Choose a large value of n and values of $0 < \beta < 1$ and $0 < r < 1$. Calculate the values of $n\varepsilon_n = n^{1-\beta}$ and $\mu_n = \sqrt{2r \log n}$. Generate a sample from the normal mixture model (3) with $K = \lceil n^{1-\beta} \rceil$ normal $N(0, 1)$ observations and $(n - K)$ normal $N(\mu_n, 1)$ observations.
2. Choose a small number $\alpha \in (0, 1)$ and apply formula (4.1) to determine a_n . Next, for each grid point

$$t_j = (j - 1)/\sqrt{2 \log n}, \quad 1 \leq j \leq 2 \log n + 1,$$

use formula (2.6) to find the values of $\mathbb{F}_{a_n}^+(t_j)$ and $\mathbb{F}_{a_n}^-(t_j)$.

3. Follow steps 3–5 of the algorithm in Chapter 2. Namely, for all $j = 1, \dots, 2 \log n + 1$, find the value of $\hat{\mu}_{a_n}^j$, then the value of $\hat{\varepsilon}_{a_n}^j$, and, finally, the value of $\hat{\varepsilon}_{a_n}^* = \max_{1 \leq j \leq 2 \log n} \hat{\varepsilon}_{a_n}^j$. When calculating $\hat{\mu}_{a_n}^j$ as a solution of the equation

$$\frac{\Phi(t_j) - F(t_j)}{\Phi(t_{j+1}) - F(t_{j+1})} = \frac{\Phi(t_j) - \mathbb{F}_{a_n}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{a_n}^-(t_{j+1})},$$

with $F(t)$ given by formula (2.1), use the Newton–Raphson method (see, for example, [1]). (This method starts with a function f defined over the set of real numbers x , the function’s derivative f' , and an initial guess x_0 for a root of the function f . If the derivative of f exists and, at the point x_0 , is not equal to 0, a better approximation x_1 is obtained through $x_1 = x_0 - f(x_0)/f'(x_0)$. This process is repeated as $x_{n+1} = x_n - f(x_n)/f'(x_n)$ for $n = 1, 2, \dots$, until a certain accurate value is obtained.) For those indices j for which the inequalities in (2.8) are not satisfied, set $\hat{\varepsilon}_{a_n}^j = 0$. If there are no indices j for which (2.8) holds, increase the value of the sample size n and start the algorithm from the beginning.

4. Choose a large positive number M and repeat steps 1–3 M times. Let $\hat{\varepsilon}_{a_n, m}^*$ be

the value of $\hat{\varepsilon}_{a_n}^*$ obtained on the m th replication, $m = 1, \dots, M$. Find the average value of $\hat{\varepsilon}_{a_n}^*$ over M replications, that is, $M^{-1} \sum_{m=1}^M \hat{\varepsilon}_{a_n,m}^*$, and use it as an estimate of ε_n . Find the (estimated) value of the risk $\mathbf{E}(\hat{\varepsilon}_{a_n}^*/\varepsilon_n - 1)^2$ by calculating $M^{-1} \sum_{m=1}^M (\hat{\varepsilon}_{a_n,m}^*/\varepsilon_n - 1)^2$.

Now, we give an analogous algorithm for the modified estimator $\hat{\varepsilon}_n$ proposed in this thesis.

Algorithm 2

- I. Choose a large value of n and values of $0 < \beta < 1$ and $0 < r < 1$. Calculate the values of $n\varepsilon_n = n^{1-\beta}$ and $\mu_n = \sqrt{2r \log n}$. Generate a sample from the normal mixture model (3) with $K = \lceil n^{1-\beta} \rceil$ normal $N(0, 1)$ observations and $(n - K)$ normal $N(\mu_n, 1)$ observations.
- II. Choose a small value $\alpha \in (0, 1)$ and use Table 3.1 in Chapter 2 to find the $(1 - \alpha)$ th quantile c_α of the function $H(t)$ defined by formula (3.4). Then, for all grid points

$$t_j = (j - 1)/\sqrt{8 \log n}, \quad 1 \leq j \leq 8 \log n + 1,$$

calculate $\mathbb{F}_{n,\alpha}^-(t_j)$ and $\mathbb{F}_{n,\alpha}^+(t_j)$ by applying formulas in (3.6).

- III. Follow steps (a)–(c) of the algorithm in Section 3.2. Namely, for all $j = 1, \dots, J_0$, where J_0 is such that $t_{J_0+1} < X_{(n)} \leq t_{J_0+2}$, find the value of $\hat{\mu}_n^j$, then the value of $\hat{\varepsilon}_n^j$, and, finally, the value of $\hat{\varepsilon}_n = \max_{1 \leq j \leq J_0} \hat{\varepsilon}_n^j$. When calculating $\hat{\mu}_n^j$ as a solution of the equation

$$\frac{\Phi(t_j) - F(t_j)}{\Phi(t_{j+1}) - F(t_{j+1})} = \frac{\Phi(t_j) - \mathbb{F}_{n,\alpha}^+(t_j)}{\Phi(t_{j+1}) - \mathbb{F}_{n,\alpha}^-(t_{j+1})},$$

with $F(t)$ given by formula (2.1), use the Newton–Raphson method. For those indices j for which the inequalities in (3.9) are not satisfied, set $\hat{\varepsilon}_n^j = 0$. If there are no indices j for which (3.9) holds, increase the value of the sample size n and start the algorithm from the beginning.

- IV. Choose a large positive number M and repeat steps 1–3 M times. Let $\hat{\varepsilon}_{n,m}$ be the value of $\hat{\varepsilon}_n$ obtained on the m th replication, $m = 1, \dots, M$. Find the average value of $\hat{\varepsilon}_n$ over M replications, that is, $M^{-1} \sum_{m=1}^M \hat{\varepsilon}_{n,m}$, and use it as an estimate of ε_n . Find the (estimated) value of the risk $\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2$ by calculating $M^{-1} \sum_{m=1}^M (\hat{\varepsilon}_{n,m}/\varepsilon_n - 1)^2$.

The results obtained by running Algorithms 1 and 2 are summarized in the next section.

4.2 Numerical Results

In this section, we present the summary of our empirical study. We know that the new estimator $\hat{\varepsilon}_n$ is slightly better than the CJL estimator $\hat{\varepsilon}_{a_n}^*$ (see Theorem 5 in Chapter 2 and Theorem 7 in Section 3.3). To compare the performance of $\hat{\varepsilon}_n$ and $\hat{\varepsilon}_{a_n}^*$ numerically, we run Algorithms 1 and 2 from Section 4.1 and summarize the obtained results in Tables 4.1 and 4.2. In our experiment, we generate a random sample of size $n = 10^8$ from the mixture model in hand with $\beta = 1/3$ and $r = 3/4$ to ensure that the model is high-dimensional and sparse. To calculate the observed values of $\hat{\varepsilon}_{a_n}^*$ and $\hat{\varepsilon}_n$ and their quadratic risks, we run $M = 10$ independent cycles of simulations. The estimators $\hat{\beta}_{a_n}^*$ and $\hat{\beta}_n$, whose observed values are also presented in Tables 4.1 and 4.2, are related to $\hat{\varepsilon}_{a_n}^*$ and $\hat{\varepsilon}_n$ through the formulas (recall that $\varepsilon_n = n^{-\beta}$)

$$\hat{\beta}_{a_n}^* = \frac{\log(1/\hat{\varepsilon}_{a_n}^*)}{\log n} \quad \text{and} \quad \hat{\beta}_n = \frac{\log(1/\hat{\varepsilon}_n)}{\log n}.$$

Note that, with high probability, both estimators $\hat{\beta}_{a_n}^*$ and $\hat{\beta}_n$ overestimate the true β .

Table 4.1: Numerical summary for the CJL estimation procedure

α	0.005	0.01	0.025	0.05	0.1	0.5
$a_n/\sqrt{2 \log \log n}$	2.077	1.911	1.660	1.509	1.338	0.918
ε_n	0.00215	0.00215	0.00215	0.00215	0.00215	0.00215
$\hat{\varepsilon}_{a_n}^*$	0.00157	0.00159	0.00168	0.00169	0.00171	0.00179
$\mathbf{E}(\hat{\varepsilon}_{a_n}^*/\varepsilon_n - 1)^2$	0.07312	0.06830	0.04742	0.04533	0.04188	0.02902
$\hat{\beta}_{a_n}^*$	0.35045	0.34978	0.34667	0.34652	0.34633	0.34347

$n = 10^8$, $M = 10$, $\beta = 1/3$, and $r = 3/4$.

Table 4.2: Numerical summary for the new estimation procedure

α	0.01	0.02	0.03	0.04	0.05	0.1
c_α	5.53	5.16	4.89	4.72	4.57	4.12
ε_n	0.00215	0.00215	0.00215	0.00215	0.00215	0.00215
$\hat{\varepsilon}_n$	0.00199	0.00201	0.00202	0.00205	0.00206	0.00210
$\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2$	0.006130	0.004541	0.002831	0.002513	0.001823	0.0007504
$\hat{\beta}_n$	0.33775	0.33710	0.33683	0.33603	0.33577	0.33472

$n = 10^8$, $M = 10$, $\beta = 1/3$, and $r = 3/4$.

Let us comment on the results reported in Tables 4.1 and 4.2. First, as seen from Tables 4.1 and 4.2, for the same values of level α , the risk of the CJL estimator is noticeably larger than that of our estimator. This finding is in agreement with the statements of Theorems 5 and 7 in Chapters 2 and 3. In addition, we notice that, for the same values of level α , the estimates of β obtained by using the modified procedure are closer to the true value of β .

Next, Tables 4.1 and 4.2 show that as α gets larger, the values of both estimators $\hat{\varepsilon}_{a_n}^*$ and $\hat{\varepsilon}_n$ increase and their risks $\mathbf{E}(\hat{\varepsilon}_{a_n}^*/\varepsilon_n - 1)^2$ and $\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2$ decrease, which means a better quality with a larger α . This phenomenon can be explained by the fact that increasing the value of α will boost the chance of underestimation while, at the same time, avoid significantly increasing the chance of overestimation, thus making the value of estimator closer to the real value of ε_n . In terms of the parameter β , the observed values of $\hat{\beta}_{a_n}^*$ and $\hat{\beta}_n$ decrease as α increases. It is seen from Tables 4.1 and 4.2 that $\hat{\beta}_n$ is a more accurate estimator of β than $\hat{\beta}_{a_n}^*$.

When examining Tables 4.1 and 4.2 further, we find that while the values of the estimators $\hat{\varepsilon}_{a_n}^*$ and $\hat{\varepsilon}_n$ increase slowly as α gets larger, $\hat{\varepsilon}_n$ certainly displays larger decrease in the risk. For example, as α increases from 0.01 to 0.05, the risk of the CJL procedure decreases approximately by 1.5 while the risk of the modified procedure decreases by 2.5. This phenomenon is supported by the fact that the weight function $q(u)$ as in (3.3) that we use to construct a confidence band needed for our estimation procedure is better than the SDP weight function $\delta(u) = \sqrt{u(1-u)}$ used by Cai et al. [6] for their confidence band.

To ensure nearly optimal rates of convergence of the CJL estimator $\hat{\varepsilon}_{a_n}^*$, the quantity

a_n used in the CJL procedure needs to be chosen large enough, i.e., $a_n = 4\sqrt{2\pi}\log^{3/2} n$ (see Theorems 5 and 6 in Chapter 2). However, running Algorithm 1 with such a large a_n requires a very large sample size, or very high-dimensional setting, otherwise condition (2.8) does not hold. In our setting, with $n = 10^8$, the experiment on calculating the estimates of $\hat{\varepsilon}_{a_n}^*$ fails to yield any result because we find that condition (2.8) is not valid for $n = 10^8$. (For this reason, in Table 4.1, in order to find the values of $a_n/\sqrt{\log \log n}$ for our choice of $n = 10^8$, we use the (recalculated) values from Table 2 from Cai et al. [6].) At the same time, Algorithm 2 produces the values of $\hat{\varepsilon}_n$ even for $n = 10^8$. In this sense, our estimator $\hat{\varepsilon}_n$ is slightly “less asymptotic” and works well in a slightly less high-dimensional setup.

In summary, the results of our numerical study are in agreement with the theoretical findings, according to which the estimation procedure proposed in this thesis outperforms the one in Cai et al. [6].

Chapter 5

Proof of the main result

The proof of our main result goes along the lines of that of Theorem 5.3 in Cai et al. [6], with the necessary modifications done.

Proof of Theorem 7. We first prove the upper bound (3.12), from which the upper bound (3.13) will follow. Recall that, by construction, with high probability $\hat{\mu}_n^j \geq \mu_n$ and $\hat{\varepsilon}_n^j \leq \varepsilon_n$ for all $j = 1, \dots, J_0$. For further analysis, we define the event

$$\mathbb{A}_{n,\alpha} \stackrel{\text{def}}{=} \{\mathbb{F}_{n,\alpha}^-(t) \leq F(t) \leq \mathbb{F}_{n,\alpha}^+(t), \forall t \in [0, \mu_n]\}, \quad (5.1)$$

where $\mathbb{F}_{n,\alpha}^+(t)$ and $\mathbb{F}_{n,\alpha}^-(t)$ are given in (3.6). By Lemma 1, for any $\alpha \in (0, 1)$

$$\mathbf{P}(\mathbb{A}_{n,\alpha}) \geq 1 - 2\alpha,$$

provided n is large enough. Consider the quadratic risk $\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2$. We have

$$\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2 = \mathbf{E}[(\hat{\varepsilon}_n/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha})] + \mathbf{E}[(\hat{\varepsilon}_n/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha}^c)], \quad (5.2)$$

where the event $\mathbb{A}_{n,\alpha}$ is given in (5.1). Let the index j^* be chosen to satisfy

$$t_{j^*} \leq \mu_n < t_{j^*+1}.$$

Then, over the event $\mathbb{A}_{n,\alpha}$,

$$\hat{\varepsilon}_n^{j*} \leq \hat{\varepsilon}_n \leq \varepsilon_n,$$

and hence

$$(1 - \hat{\varepsilon}_n/\varepsilon_n)^2 \leq (1 - \hat{\varepsilon}_n^{j*}/\varepsilon_n)^2.$$

Note that any reasonable estimator of $\varepsilon_n = n^{-\beta} = o(1)$ cannot be larger than one; unreasonable estimators are removed by checking the inequalities in (3.9). Therefore, using (5.2), we may continue

$$\begin{aligned} \mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2 &\leq \mathbf{E}(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \\ &= \mathbf{E}\left[(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha})\right] + \mathbf{E}\left[(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha}^c)\right] \\ &\leq \mathbf{E}\left[(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha})\right] + \mathbf{P}(\mathbb{A}_{n,\alpha}^c) \varepsilon_n^{-2}, \end{aligned}$$

where, by Lemma 1, $\mathbf{P}(\mathbb{A}_{n,\alpha}^c) \leq 2\alpha$. Hence, recalling that $\varepsilon_n = n^{-\beta}$, we can write

$$\mathbf{E}(\hat{\varepsilon}_n/\varepsilon_n - 1)^2 \leq \mathbf{E}(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \leq \mathbf{E}\left[(\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1)^2 \mathbb{I}(\mathbb{A}_{n,\alpha})\right] + (2\alpha)n^{-2\beta} \stackrel{\text{def}}{=} T_1 + T_2. \quad (5.3)$$

At this point, we observe that the following modification of Lemma 1 with $\alpha = \alpha_n \rightarrow 0$ moderately (polynomially) fast holds true: let $X_1, X_2, \dots$ be a sequence of iid random variables from a two-point mixture model having a common cdf $F(t)$ as in (2.1) with $\varepsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$, where $0 < \beta < 1$ and $0 < r < 4$; then for any $\alpha = \alpha_n \rightarrow 0$ such that $\alpha_n = O(n^{-\nu})$ for some $\nu > 0$ and all sufficiently large n

$$\mathbf{P}(\mathbb{F}_{n,\alpha_n}^-(t) \leq F(t) \leq \mathbb{F}_{n,\alpha_n}^+(t), \forall t \in [0, \mu_n]) \geq 1 - 2\alpha_n, \quad (5.4)$$

where $\mathbb{F}_{n,\alpha}^+(t)$ and $\mathbb{F}_{n,\alpha}^-(t)$ are as defined in (3.6). The validity of (5.4) is easy to verify by going through the proof of Lemma 1 in Chapter 6.

Back to (5.3), if we now choose $\alpha = \alpha_n$ that decreases to zero polynomially and such that $\alpha = \alpha_n = o(n^{-1-\beta})$, we get

$$T_2 = o(n^{2\beta-1-\beta}) = o(n^{\beta-1}) = o(1). \quad (5.5)$$

Hence it remains to obtain a good upper bound on the term T_1 . When analyzing T_1 , we shall always limit our attention to the event $\mathbb{A}_{n,\alpha}$.

By definition of $\hat{\varepsilon}_n^{j*}$ (see formula (3.10)), the following identity holds true

$$\hat{\varepsilon}_n^{j*}/\varepsilon_n - 1 = \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)}{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})} \left(\frac{F(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - F(t_{j*})} - \frac{\Phi(t_{j*} - \mu_n) - \Phi(t_{j*} - \hat{\mu}_n^{j*})}{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)} \right). \quad (5.6)$$

Indeed,

$$\begin{aligned} \hat{\varepsilon}_n^{j*}/\varepsilon_n - 1 &= \frac{\Phi(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})} \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)}{\Phi(t_{j*}) - F(t_{j*})} - 1 \\ &= \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)}{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})} \frac{\Phi(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - F(t_{j*})} - 1 \\ &= \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)}{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})} \left(\frac{\Phi(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - F(t_{j*})} - \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})}{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)} \right). \end{aligned}$$

Now noting that

$$\begin{aligned} &\frac{F(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - F(t_{j*})} - \frac{\Phi(t_{j*} - \mu_n) - \Phi(t_{j*} - \hat{\mu}_n^{j*})}{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)} \\ &= \frac{\Phi(t_{j*}) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*}) - F(t_{j*})} - \frac{\Phi(t_{j*}) - \Phi(t_{j*} - \hat{\mu}_n^{j*})}{\Phi(t_{j*}) - \Phi(t_{j*} - \mu_n)}, \end{aligned}$$

we arrive at (5.6).

Next, by the monotonicity of $\Phi(t)$,

$$\Phi(t_{j*} - \mu_n) \geq \Phi(t_{j*} - \hat{\mu}_n^{j*}),$$

and hence

$$\frac{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \hat{\mu}_n^{j^*})} \leq 1. \quad (5.7)$$

Observe that for all $a, b \in \mathbb{R}$

$$|a \pm b| \leq |a| + |b| \leq 2 \max(|a|, |b|),$$

and hence

$$|a \pm b|^2 \leq 4 \max(|a|^2, |b|^2) \leq 4(|a|^2 + |b|^2). \quad (5.8)$$

Combining (5.6), (5.7), and (5.8), we obtain

$$\begin{aligned} (\hat{\varepsilon}_n^{j^*} / \varepsilon_n - 1)^2 &\leq 4 \left(\frac{F(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 + 4 \left(\frac{\Phi(t_{j^*} - \mu_n) - \Phi(t_{j^*} - \hat{\mu}_n^{j^*})}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} \right)^2 \\ &\stackrel{\text{def}}{=} I_1 + I_2. \end{aligned} \quad (5.9)$$

In order to estimate the term I_1 from above, we notice the following consequence of Lemma 1, which is true for both fixed $\alpha \in (0, 1)$ and $\alpha = \alpha_n \rightarrow 0$ moderately (polynomially) fast: for all sufficiently large n (over the event $\mathbb{A}_{n,\alpha}$)

$$|\mathbb{F}_{n,\alpha}^\pm(t) - F(t)| \leq \frac{2c_\alpha q(F(t))}{\sqrt{n}}, \quad (5.10)$$

where $\mathbb{F}_{n,\alpha}^+(t)$ and $\mathbb{F}_{n,\alpha}^-(t)$ are defined as in (3.6) and $q(t)$ is given in (3.3). Indeed, it follows from (3.6) that

$$\mathbb{F}_{n,\alpha}^-(t) \geq \mathbb{F}_n(t) - \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)),$$

and

$$\mathbb{F}_{n,\alpha}^+(t) \leq \mathbb{F}_n(t) + \frac{c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)).$$

As a result,

$$|\mathbb{F}_{n,\alpha}^\pm(t) - F(t)| \leq |\mathbb{F}_{n,\alpha}^+(t) - \mathbb{F}_{n,\alpha}^-(t)| \leq \frac{2c_\alpha}{\sqrt{n}} q(\mathbb{F}_n(t)),$$

and (5.10) is verified. It now follows from (5.10) that

$$I_1 = 4 \left(\frac{F(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 \leq \frac{16c_\alpha^2 q^2(F(t_{j^*}))}{n(\Phi(t_{j^*}) - F(t_{j^*}))^2}, \quad (5.11)$$

where, using formula (3.3),

$$\frac{q^2(F(t_{j^*}))}{(\Phi(t_{j^*}) - F(t_{j^*}))^2} = \frac{F(t_{j^*})(1 - F(t_{j^*})) \log \log \left(\frac{1}{F(t_{j^*})(1 - F(t_{j^*}))} \right)}{(\Phi(t_{j^*}) - F(t_{j^*}))^2}. \quad (5.12)$$

Note that the function $F(t)(1 - F(t))$ is strictly decreasing when $F(t) \geq 1/2$. Therefore, taking into account the fact that $1/2 \leq F(t_{j^*}) < F(\mu_n)$ for all sufficiently large n , we can write $F(t_{j^*})(1 - F(t_{j^*})) \geq F(\mu_n)(1 - F(\mu_n))$, and hence

$$\log \log \left(\frac{1}{F(t_{j^*})(1 - F(t_{j^*}))} \right) \leq \log \log \left(\frac{1}{F(\mu_n)(1 - F(\mu_n))} \right), \quad (5.13)$$

provided n is sufficiently large. Now we need the following result whose proof is given in Chapter 6, cf. Lemma 8.1 in Cai et al. [6].

Lemma 2. *Let $F(t)$ be the cdf as in (2.1) with $\varepsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r \log n}$, where $0 < \beta < 1$ and $0 < \beta < r < 4$. Then we have as $n \rightarrow \infty$*

$$\frac{F(\mu_n)(1 - F(\mu_n))}{(\Phi(\mu_n) - F(\mu_n))^2} = 2n^\beta (1 + o(1)).$$

Moreover, for any t with the property $|t - \mu_n| \leq c/\sqrt{\log n}$ for some constant $c > 0$, there

exists a constant $C = C(\beta, r, c) > 0$ such that

$$\frac{F(t)(1 - F(t))}{(\Phi(t) - F(t))^2} \leq C \frac{F(\mu_n)(1 - F(\mu_n))}{(\Phi(\mu_n) - F(\mu_n))^2}.$$

Recall that, in the modified estimation procedure, the grid points are chosen so that $t_{j+1} - t_j = (8 \log n)^{-1/2}$. Therefore, by the fact that $t_{j^*} \leq \mu_n < t_{j^*+1}$ and Lemma 2, there exists a constant $C = C(\beta, r) > 0$ such that for all sufficiently large n

$$\frac{F(t_{j^*})(1 - F(t_{j^*}))}{(\Phi(t_{j^*}) - F(t_{j^*}))^2} \leq C \frac{F(\mu_n)(1 - F(\mu_n))}{(\Phi(\mu_n) - F(\mu_n))^2}.$$

From this and (5.13), we conclude that the expression in (5.12) satisfies

$$\frac{F(t_{j^*})(1 - F(t_{j^*})) \log \log \left(\frac{1}{F(t_{j^*})(1 - F(t_{j^*}))} \right)}{(\Phi(t_{j^*}) - F(t_{j^*}))^2} \leq C \frac{F(\mu_n)(1 - F(\mu_n)) \log \log \left(\frac{1}{F(\mu_n)(1 - F(\mu_n))} \right)}{(\Phi(\mu_n) - F(\mu_n))^2}.$$

Therefore, taking into account (5.11) and (5.12), we obtain with some constant $C_1 = C_1(\beta, r, \alpha) > 0$ that

$$I_1 = 4 \left(\frac{F(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 \leq \frac{C_1}{n} \frac{F(\mu_n)(1 - F(\mu_n))}{(\Phi(\mu_n) - F(\mu_n))^2} \log \log \left(\frac{1}{F(\mu_n)(1 - F(\mu_n))} \right). \quad (5.14)$$

We shall now analyze the right-hand side of (5.14). First, consider the difference $\Phi(\mu_n) - F(\mu_n)$. By (2.1), we have

$$\begin{aligned} \Phi(\mu_n) - F(\mu_n) &= \Phi(\mu_n) - (1 - \varepsilon_n)\Phi(\mu_n) - \varepsilon_n\Phi(0) \\ &= \Phi(\mu_n) - \Phi(\mu_n) + \varepsilon_n\Phi(\mu_n) - \varepsilon_n\Phi(0) \\ &= \varepsilon_n\Phi(\mu_n) - \varepsilon_n/2 \\ &= \varepsilon_n/2 - \varepsilon_n(1 - \Phi(\mu_n)). \end{aligned}$$

Note that as $x \rightarrow \infty$

$$1 - \Phi(x) \sim \frac{\phi(x)}{x}. \quad (5.15)$$

In the view of (5.15), as $n \rightarrow \infty$,

$$1 - \Phi(\mu_n) \sim \frac{\phi(\mu_n)}{\mu_n} = \frac{\exp(-r \log n)}{\sqrt{2\pi} \sqrt{2r \log n}} = \frac{1}{2n^r \sqrt{\pi r \log n}}, \quad (5.16)$$

and hence

$$\begin{aligned} \Phi(\mu_n) - F(\mu_n) &= \frac{\varepsilon_n}{2} - \varepsilon_n (1 - \Phi(\mu_n)) \\ &\sim \frac{n^{-\beta}}{2} \left(1 - \frac{1}{n^r \sqrt{\pi r \log n}} \right) \sim \frac{n^{-\beta}}{2}. \end{aligned} \quad (5.17)$$

Next, we obtain an asymptotic expression for $\log \log (1 / (F(\mu_n) (1 - F(\mu_n))))$. Using representation (5.16), we can write

$$1 - F(\mu_n) = (1 - \Phi(\mu_n))(1 - \varepsilon_n) + \varepsilon_n/2 \sim n^{-\beta}/2. \quad (5.18)$$

From this,

$$\log \log (1 / (F(\mu_n) (1 - F(\mu_n)))) \sim \log \log ((1 - F(\mu_n))^{-1}) \sim \log \log (2n^\beta) \sim \log \log n. \quad (5.19)$$

Now, combining (5.14) with formulas (5.17)–(5.19), we obtain that (over the event $\mathbb{A}_{n,\alpha}$) the term I_1 on the right-hand side of (5.9) satisfies

$$I_1 = 4 \left(\frac{F(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right)^2 = O \left(\frac{\log \log n}{n^{1-\beta}} \right). \quad (5.20)$$

Notice that here and below the notation $a_n = O(b_n)$ means that $a_n/b_n \rightarrow C(\beta, r, \alpha)$ as $n \rightarrow \infty$ for some constant $C(\beta, r, \alpha) \neq 0$ that, in general, depends on the parameters β, r

and α (in case of fixed $\alpha \in (0, 1)$).

Let us now derive a required upper bound on the (main) term I_2 on the right-hand side of (5.9). Recall that our attention is restricted to the event $\mathbb{A}_{n,\alpha}$. On this event, as $n \rightarrow \infty$

$$\frac{\Phi(t_{j^*} - \hat{\mu}_n^{j^*}) - \Phi(t_{j^*} - \mu_n)}{\mu_n - \hat{\mu}_n^{j^*}} \sim \phi(t_{j^*} - \mu_n),$$

and hence

$$I_2 = 4 \left(\frac{\Phi(t_{j^*} - \mu_n) - \Phi(t_{j^*} - \hat{\mu}_n^{j^*})}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} \right)^2 \sim 4 \left(\frac{\phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} \right)^2 (\mu_n - \hat{\mu}_n^{j^*})^2. \quad (5.21)$$

Before studying the right-hand side of (5.21), observe that as $n \rightarrow \infty$

$$t_{j^*} \rightarrow \infty \quad \text{and} \quad t_{j^*} - \mu_n \rightarrow -\infty.$$

Therefore $\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n) \rightarrow 1$ and, since the standard normal pdf ϕ is a bounded function, we obtain as $n \rightarrow \infty$

$$\frac{\phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} = O(1).$$

Thus (see representation (5.21)) the behaviour of I_2 is controlled by the term $(\mu_n - \hat{\mu}_n^{j^*})^2$.

In connection with this term, we have following result whose proof is given in Chapter 6.

Lemma 3. *Under the assumptions of Theorem 7, there exists a constant $C = C(\beta, r, \alpha) > 0$ such that for all sufficiently large n*

$$\mathbf{E} \left((\hat{\mu}_n^{j^*} - \mu_n) \mathbb{I}(\mathbb{A}_{n,\alpha}) \right)^2 \leq \frac{C \log n \times \log \log n}{n^{1-\beta}}.$$

This inequality continues to hold for $\alpha = \alpha_n \rightarrow 0$ moderately (polynomially) fast with a constant $C = C(\beta, r) > 0$.

In the view of Lemma 3, taking into account (5.17) and (5.19), we get

$$\mathbf{E}(I_2\mathbb{I}(\mathbb{A}_{n,\alpha})) = O\left(\frac{\log n \times \log \log n}{n^{1-\beta}}\right).$$

Therefore, using relations (5.3), (5.9), and (5.20), we obtain

$$\begin{aligned} T_1 &= \mathbf{E}\left(\left(\hat{\varepsilon}_n^{j^*}/\varepsilon_n - 1\right)^2 \mathbb{I}(\mathbb{A}_{n,\alpha})\right) = \mathbf{E}(I_1\mathbb{I}(\mathbb{A}_{n,\alpha})) + \mathbf{E}(I_2\mathbb{I}(\mathbb{A}_{n,\alpha})) \\ &= O\left(\frac{\log \log n}{n^{1-\beta}}\right) + O\left(\frac{\log n \times \log \log n}{n^{1-\beta}}\right) \\ &= O\left(\frac{\log n \times \log \log n}{n^{1-\beta}}\right). \end{aligned}$$

From this, using (5.3)–(5.5) and recalling that the notation $a_n = O(b_n)$ means that $a_n/b_n \rightarrow C(\beta, r, \alpha)$ as $n \rightarrow \infty$ for some constant $C(\beta, r, \alpha) \neq 0$, we arrive at (3.12). Then, in view of relation (5.4), our proof also goes through with $\alpha = \alpha_n \rightarrow 0$ moderately (polynomially) fast. Hence, if we choose $\alpha = \alpha_n = o(n^{-1-\beta})$ that decays to zero as a power of n , the upper bound (3.12) implies (3.13). The proof is completed. $\square$

Chapter 6

Proofs of Lemmas

This section contains the proofs of Lemmas 1–3 that were used to prove our main result, Theorem 7. Lemma 1 is stated in Section 3.1 and Lemmas 2 and 3 are stated in the proof of Theorem 7 in Chapter 5.

Proof of Lemma 1. In view of (3.5) and Fact 2 from Section 3.1, it is sufficient to show that for any $\alpha \in (0, 1)$, there exists a number n^* such that for all $n \geq n^*$,

$$\mathbf{P}(X_{(n)} \geq \mu_n) \geq 1 - \alpha/2, \quad (6.1)$$

and

$$\mathbf{P}(X_{(1)} \leq 0) \geq 1 - \alpha/2. \quad (6.2)$$

First, we show that inequality (6.1) holds true. Dealing instead of model (3) with the equivalent model (4) in which $a = \sqrt{2r \log n}$ and $n^{-1} \sum_{j=1}^n \eta_j \sim n^{-\beta}$, observe that

$$X_{(n)} = \max_{1 \leq j \leq n} X_j = \max(\max_{j: \eta_j=1} X_j, \max_{j: \eta_j=0} X_j).$$

Now denoting by $Z_1, Z_2, \dots$ a sequence of iid $N(0, 1)$ random variables and recalling that

$K_n = \lceil n^{1-\beta} \rceil \rightarrow \infty$ as $n \rightarrow \infty$, we have

$$\begin{aligned} \mathbf{P}(X_{(n)} \geq \mu_n) &\geq \mathbf{P}\left(\max_{j:\eta_j=1} X_j \geq \mu_n\right) = \mathbf{P}\left(\max_{j:\eta_j=1} (X_j - \mu_n) + \mu_n \geq \mu_n\right) \\ &= \mathbf{P}\left(\max_{1 \leq j \leq K_n} Z_j \geq 0\right) = 1 - \mathbf{P}\left(\max_{1 \leq j \leq K_n} Z_j \leq 0\right) = 1 - (\Phi(0))^{K_n} \\ &= 1 - (1/2)^{K_n} \rightarrow 1. \end{aligned}$$

Hence, for a given $\alpha \in (0, 1)$ one can always find a large number N such that $\mathbf{P}(X_{(n)} \geq \mu_n) \geq 1 - \alpha/2$ for all $n \geq N$.

Now, using similar arguments, we shall prove inequality (6.2). To this end, observe that

$$X_{(1)} = \min_{1 \leq j \leq n} X_j = -\max_{1 \leq j \leq n} (-X_j).$$

Therefore

$$\begin{aligned} \mathbf{P}(X_{(1)} \leq 0) &= \mathbf{P}\left(-\max_{1 \leq j \leq n} (-X_j) \leq 0\right) = \mathbf{P}\left(\max_{1 \leq j \leq n} (-X_j) \geq 0\right) \\ &= \mathbf{P}\left(\max_{j:\eta_j=1} (-X_j), \max_{j:\eta_j=0} (-X_j) \geq 0\right) \\ &\geq \mathbf{P}\left(\max_{j:\eta_j=0} (-X_j) \geq 0\right) = \mathbf{P}\left(\max_{1 \leq j \leq n-K_n} (-Z_j) \geq 0\right). \end{aligned}$$

Since $N(0, 1) \stackrel{d}{=} -N(0, 1)$, it then follows that as $n \rightarrow \infty$

$$\begin{aligned} \mathbf{P}(X_{(1)} \leq 0) &\geq 1 - \mathbf{P}\left(\max_{1 \leq j \leq n-K_n} Z_j \leq 0\right) \\ &= 1 - (1/2)^{n-K_n} = 1 - (1/2)^{n(1+o(1))} \rightarrow 1. \end{aligned}$$

Hence, for a given $\alpha \in (0, 1)$ one can always find a large number M such that $\mathbf{P}(X_{(1)} \leq 0) \geq 1 - \alpha/2$ for all $n \geq M$. Thus, both inequalities (6.1) and (6.2) hold true with $n^* =$

$\max(N, M)$. From this, using Fact 2, for all $n \geq n^*$

$$\mathbf{P}(X_{(1)} \leq 0 \text{ and } X_{(n)} \geq \mu_n) \geq 1 - \alpha. \quad (6.3)$$

Finally, the combination of (3.5), (6.3), and Fact 2 yields the statement of the lemma. $\square$

Proof of Lemma 2. Relations (5.17) and (5.18) imply

$$\frac{F(\mu_n)(1 - F(\mu_n))}{(\Phi(\mu_n) - F(\mu_n))^2} \sim 2n^\beta, \quad n \rightarrow \infty,$$

and the first claim of the lemma is proved. It remains to prove the second claim. Let t be such that for some constant $c > 0$

$$|t - \mu_n| \leq c/\sqrt{\log n}.$$

Then, using relation (5.15), it is easy to show that

$$1 - \Phi(t) \sim 1 - \Phi(\mu_n), \quad n \rightarrow \infty.$$

Applying this asymptotic relation, we can write as $n \rightarrow \infty$

$$\begin{aligned} F(t) &= (1 - \varepsilon_n)\Phi(t) + \varepsilon_n\Phi(t - \mu_n) \\ &\sim (1 - \varepsilon_n)\Phi(\mu_n) + \varepsilon_n\Phi(\mu_n) + R_n \\ &= F(\mu_n) + R_n, \end{aligned} \quad (6.4)$$

where for all t such that $|t - \mu_n| \leq c/\sqrt{\log n}$

$$R_n \stackrel{\text{def}}{=} \varepsilon_n (\Phi(t - \mu_n) - \Phi(\mu_n)) \asymp n^{-\beta}, \quad n \rightarrow \infty. \quad (6.5)$$

In view of (6.4), we have

$$\frac{F(t)(1-F(t))}{(\Phi(t)-F(t))^2} = \frac{(F(\mu_n)+R_n)(1-F(\mu_n)-R_n)}{(\Phi(\mu_n)-F(\mu_n)-R_n)^2}, \quad (6.6)$$

where by (5.17) and (5.18)

$$1-F(\mu_n) \sim \Phi(\mu_n)-F(\mu_n) \sim n^{-\beta}/2. \quad (6.7)$$

Representation (6.6) together with (6.5) and (6.7) now gives us that for some constant $C > 0$

$$\frac{F(t)(1-F(t))}{(\Phi(t)-F(t))^2} \leq C \frac{(F(\mu_n)(1-F(\mu_n)))}{(\Phi(\mu_n)-F(\mu_n))^2}$$

The lemma is proved. $\square$

Proof of Lemma 3. The proof is similar to that of Lemma 8.3 in Cai et al. [6], with necessary modification done. Without further notice, we always confine our attention to the event $\mathbb{A}_{n,\alpha}$ given by formula (5.1). By definition, μ_n and $\hat{\mu}_n^{j*}$ are solutions of the respective equations

$$D(\mu_n) = \frac{\Phi(t_{j*}) - F(t_{j*})}{\Phi(t_{j*+1}) - F(t_{j*+1})} \quad \text{and} \quad D(\hat{\mu}_n^{j*}) = \frac{\Phi(t_j) - \mathbb{F}_{n,\alpha}^+(t_{j*})}{\Phi(t_{j*+1}) - \mathbb{F}_{n,\alpha}^-(t_{j*+1})}, \quad (6.8)$$

with the functions $D(\mu_n) = D(\mu_n; t_{j*}, t_{j*+1})$ and $D(\hat{\mu}_n^{j*}) = D(\hat{\mu}_n^{j*}; t_{j*}, t_{j*+1})$ defined in accordance with (2.2). Set

$$\delta_n \stackrel{\text{def}}{=} 1 - \frac{D(\mu_n)}{D(\hat{\mu}_n^{j*})},$$

and observe that $\delta_n \leq 0$. If we show that (a) δ_n is algebraically small as $n \rightarrow \infty$, and (b) $D'(\mu_n)/D(\mu_n) = O(t_{j*} - t_{j*+1})$, we obtain that an algebraically small perturbation on the right-hand side of the equations in (6.8) only results in an algebraically small perturbation of the corresponding solution.

First, by (6.8), δ_n can be written in the form

$$\delta_n = 1 - \frac{(\Phi(t_{j^*}) - F(t_{j^*})) (\Phi(t_{j^*+1}) - \mathbb{F}_{n,\alpha}^-(t_{j^*+1}))}{(\Phi(t_{j^*+1}) - F(t_{j^*+1})) (\Phi(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*}))}. \quad (6.9)$$

Now note that for any $a > 0$, $b > 0$,

$$\left| \frac{1+a}{1+b} - 1 \right| \leq a+b. \quad (6.10)$$

Therefore, if we choose

$$a = \frac{\Phi(t_{j^*+1}) - \mathbb{F}_{n,\alpha}^-(t_{j^*+1})}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} - 1 > 0 \quad \text{and} \quad b = \frac{\Phi(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} - 1 > 0,$$

we get from (6.9) and (6.10) that

$$\begin{aligned} |\delta_n| &\leq \left| \frac{\Phi(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} - 1 \right| + \left| \frac{\Phi(t_{j^*+1}) - \mathbb{F}_{n,\alpha}^-(t_{j^*+1})}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} - 1 \right| \\ &= \left| \frac{F(t_{j^*}) - \mathbb{F}_{n,\alpha}^+(t_{j^*})}{\Phi(t_{j^*}) - F(t_{j^*})} \right| + \left| \frac{F(t_{j^*+1}) - \mathbb{F}_{n,\alpha}^-(t_{j^*+1})}{\Phi(t_{j^*+1}) - F(t_{j^*+1})} \right|. \end{aligned}$$

From this, the application of (5.10) gives that with some constant $C > 0$

$$|\delta_n| \leq C \left(\frac{q(F(t_{j^*}))}{\sqrt{n}(\Phi(t_{j^*}) - F(t_{j^*}))} + \frac{q(F(t_{j^*+1}))}{\sqrt{n}(\Phi(t_{j^*+1}) - F(t_{j^*+1}))} \right).$$

Now, by the choice of t_{j^*} and t_{j^*+1} (recall that $t_{j^*} \leq \mu_n < t_{j^*+1}$) and using the arguments leading to the statement of Lemma 2, we obtain for some constant $C = C(\beta, r) > 0$ as $n \rightarrow \infty$

$$|\delta_n| \leq C \frac{q(\mu_n)}{\sqrt{n}(\Phi(\mu_n) - F(\mu_n))} = O \left(\sqrt{\frac{\log \log n}{n^{1-\beta}}} \right) = o(1). \quad (6.11)$$

This shows, in particular, that δ_n is algebraically small, and hence, by the continuity of $D(\mu_n)$, on the event $\mathbb{A}_{n,\alpha}$, the quantities $\hat{\mu}_n^{j^*}$ and μ_n are close to each other.

We now move on to analyzing the term $D'(\mu_n)/D(\mu_n)$. Recall that $D(\mu_n)$ is defined as

(see formula (2.2))

$$D(\mu_n) = \frac{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)}.$$

By simple algebra we have

$$D'(\mu_n) = \frac{\phi(t_{j^*} - \mu_n) (\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)) - \phi(t_{j^*+1} - \mu_n) (\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n))}{(\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n))^2},$$

and hence

$$\begin{aligned} \frac{D'(\mu_n)}{D(\mu_n)} &= \frac{\phi(t_{j^*} - \mu_n) (\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)) - \phi(t_{j^*+1} - \mu_n) (\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n))}{(\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)) (\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n))} \\ &= \frac{\phi(t_{j^*} - \mu_n)}{\Phi(t_{j^*}) - \Phi(t_{j^*} - \mu_n)} - \frac{\phi(t_{j^*+1} - \mu_n)}{\Phi(t_{j^*+1}) - \Phi(t_{j^*+1} - \mu_n)}. \end{aligned}$$

From this, using formula (5.15) and Taylor's expansion, it is easy to show that the ratio $D'(\mu_n)/D(\mu_n)$ satisfies

$$0 > \frac{D'(\mu_n)}{D(\mu_n)} = O(t_{j^*+1} - t_{j^*}) = O(\log^{-1/2} n). \quad (6.12)$$

It now follows from (6.12) that (on the event $\mathbb{A}_{n,\alpha}$)

$$\frac{D(\hat{\mu}_n^{j^*}) - D(\mu_n)}{\hat{\mu}_n^{j^*} - \mu_n} \frac{1}{D(\mu_n)} = O(\log^{-1/2} n),$$

implying

$$\hat{\mu}_n^{j^*} - \mu_n \asymp \frac{D(\hat{\mu}_n^{j^*}) - D(\mu_n)}{D(\mu_n)} \log^{1/2} n,$$

where by (6.11) and the fact that $\delta_n \leq 0$

$$\left| \frac{D(\hat{\mu}_n^{j^*}) - D(\mu_n)}{D(\mu_n)} \right| = \left| \frac{D(\hat{\mu}_n^{j^*})}{D(\mu_n)} - 1 \right| = \left| \frac{1}{1 - \delta_n} - 1 \right| = \left| \frac{\delta_n}{1 - \delta_n} \right| \leq |\delta_n| = O\left(\sqrt{\frac{\log \log n}{n^{1-\beta}}}\right).$$

Therefore as $n \rightarrow \infty$

$$\hat{\mu}_n^{j^*} - \mu_n = O\left(\sqrt{\frac{\log n \times \log \log n}{n^{1-\beta}}}\right).$$

Finally, noting that $\mathbf{P}(\mathbb{A}_{n,\alpha}) \leq 1$, we get for some constant $C = C(\beta, r, \alpha) > 0$

$$\mathbf{E}\left((\hat{\mu}_n^{j^*} - \mu_n)\mathbb{I}(\mathbb{A}_{n,\alpha})\right)^2 \leq C \frac{\log n \times \log \log n}{n^{1-\beta}},$$

and for fixed $\alpha \in (0, 1)$ the lemma is proved. Now, assume that $\alpha = \alpha_n$ decays to zero at a polynomial rate. Then, due to (5.4), the same arguments, with $\mathbb{A}_{n,\alpha_n}$ in place of $\mathbb{A}_{n,\alpha}$, give also the statement of Lemma 3 for $\alpha = \alpha_n \rightarrow 0$. $\square$

Conclusion

In this thesis, we study the problem of estimating the fraction of nonzero means in sparse normal mixture models. We propose an improved version of the well-known CJL procedure from Cai et al. [6] with a faster rate of convergence. This becomes possible by utilizing the result of Stepanova and Pavlenko [27] on the asymptotically correct confidence band for a continuous cdf. This result, in turn, is based on a fundamental work of Csörgő et al. [9]. When compared to the CJL estimator, our estimator is slightly less asymptotic, that is, works well for smaller sample sizes. It is also more efficient, that is, has a better convergence rate than the CJL estimator, and simpler in construction. The superiority of the modified procedure is shown analytically and is also demonstrated numerically. Thus, in this thesis, we have achieved the initial goal of our study as stated in the Introduction.

For future work, one could try to construct an optimal procedure whose maximum risk achieves the lower bound $O(n^{-1+\beta})$ as given in Theorem 6 of Chapter 2. For this, an entirely different approach, not related to the function $D(\mu; t, t')$ in (2.2) should be applied. Another possible direction is to find a procedure that would work comparably well for moderate sample sizes like $n = 10^3$. It would be also of interest to extend the methodology to non-normal sparse mixtures, like the ones considered in Donoho and Jin [11] and in Meinshausen and Rice [23]. Meinshausen and Rice [23] considered a more general mixture model that can be written as $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} (1 - \varepsilon_n)N(0, 1) + \varepsilon_n G$, where G is an arbitrary distribution. We conjecture that, as long as the tail probabilities of a non-normal mixture behave like those of the normal mixture (3), the upper bound on the maximum risk of $\hat{\varepsilon}_n$ will be the same as in Theorem 7.

Bibliography

- [1] Ben A. (1966). A Newton-Raphson method for the solution of systems of equations. *J. Math. Anal. Appl.*, **15**, 243-252.
- [2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.
- [3] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple hypothesis testing with independent statistics. *J. Educ. Behav. Stat.*, **25**, 60–83.
- [4] Bickel, P. J. and Doksum, R. A. (2001). Mathematical statistics. Basic ideas and selected topics. Vol. I, 2nd ed., Prentice-Hall, New Jersey.
- [5] Cai, T. T., Jin, J., and Low, M. (2006). Estimation and confidence sets for sparse normal models. <http://arxiv.org/abs/math/0612623v1>.
- [6] Cai, T. T., Jin, J., and Low, M. (2007). Estimation and confidence sets for sparse normal models. *Ann. Statist.*, **35**, 2421–2449.
- [7] Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc. B*, **73**, 629–662.
- [8] Chen, W., Zhang, Z., King, S., Alcock, C., Byun, Y., Cook, K., Dave, R., Giammarco, J., Lee, T., Lehner, M., Liang, C., Lissauer, J., Marshall, S., De Pater, I., Porrata, R., Rice, J., Wang, A., Wang, S. and Wen, C. (2003). Fast CCD photometry in the Taiwanese-American Occultation Survey. *Baltic Astronomy*, **12**, 568-573.

- [9] Csörgő, M., Csörgő, S., Horváth, L., and Mason, D. (1986). Weighted empirical and quantile processes. *Ann. Probab.*, **14**, 31–85.
- [10] Cui, X. (2014). Optimal component selection in high dimensions. MSc Thesis. Carleton University.
- [11] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.
- [12] Eastwood, B. J. and Eastwood, V. R. (1998). Tabulating weighted functionals of Brownian bridges via Monte Carlo simulation. In: *Asymptotic methods in probability and statistics — A volume in honour of Miklós Csörgő* (ed. B. Szyszkowicz), pp. 707–719. Elsevier Science B.V., Amsterdam.
- [13] Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *Ann. Statist.*, **7**, 116–138.
- [14] Genovese, C. R., Jin J., Wasserman L., and Yao Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13**, 2107–2143.
- [15] Haupt, J., Castro, R., and Nowak, R. (2008). Adaptive discovery of sparse signals in noise. In: *Proc. 42nd Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA, October 2008* (ed. M. B. Matthews), pp. 1727–1731. New York, NY: IEEE.
- [16] Ingster, Yu. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Meth. Statist.*, **6**, 47–69.
- [17] Ingster, Yu. I. and Suslina, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lect. Notes Statist., Vol. **169**, Springer-Verlag, New York.
- [18] Jaeschke, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.*, **7**, 108–115.

- [19] Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.*, **35**, 2018–2053.
- [20] Jensen, S. T., Chen, G., and Stoeckert, C. J., Jr. (2007). Bayesian variable selection and data integration for biological regulatory networks. *Ann. Appl. Statist.*, **1**, 612–633.
- [21] Liang, G.-L., Rice, J., De Pater, I., Alcock, C., Axelrod, T., Wang, A., and Marshall, S. (2004). Statistical methods for detecting stellar occultation by Kuiper Belt objects: the Taiwanese-American occultation survey. *Statist. Sci.*, **19**, 265–273.
- [22] Meinshausen, N. and Bühlmann, P. (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, **92**, 893–907.
- [23] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34**, 373–393.
- [24] Nettleton, D. and Hwang, J. (2003). Estimating the number of false null hypotheses when conducting many tests. Technical report, Dept. Statistics, Iowa State Univ.
- [25] Orasch, M. and Pouliot, W. (2004). Tabulating weighted sup-norm functionals used in change-point problem. *J. Stat. Comput. Simul.*, **74**, 249–276.
- [26] Shorack, G. R. and Wellner, J. A. (1986). Empirical processes with applications to statistics. Wiley, New York.
- [27] Stepanova N. and Pavlenko T. (2016). Goodness-of-fit tests based on sup-functionals of weighted empirical processes. <http://arxiv.org/abs/1406.0526>.
- [28] Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.

Appendix: R codes

```
1 # Yibo Wang
2 # 2016 Carleton University
3 #
4 # Estimation of the Amount of Sparsity in Normal Mixture Models
5 # Programming code
6 # The goal of such a program is to determine epsilon_n by means of simulation
7
8 #####
9 #####
10 # The CJL procedure
11 #####
12 #####
13 # Global constants
14 # n, the sample size
15 # mu, the theoretical mean of the second normal term in the normal mixture
    model
16 # varep, the coefficient of the second normal term in the normal mixture model
17 # b, the sparsity parameter which controls the amount of nonzero means
18 # a, the quantile of the weight function
19 n = 10^(8)
20 b = 1/3
21 r = 3/4
22 mu = sqrt(2*r*log(n))
23 varep = n^(-b)
24 Alpha = a*sqrt(2*log(log(n)))
```

```

25 #####
26 # Normal mixture model sampling
27 # The method for sampling from normal mixture model
28 mixsample = function(n, varep, mu)
29 {
30   # These are *cumulative* probabilities
31   # Generate random variables from a bernoulli distribution to select the
      relevant distribution
32   dists = rbinom(n,1,1-varep)
33   # This is where the actual data are generated, it's a if->then
34   # statement, followed by the normal distributions you are interested in
35   data = vector(length = n)
36   for(i in 1:n)
37   {
38     if(dists[i] == 1 )
39     {
40       data[i] = rnorm(1, mean=0, sd=1)
41     } else {
42       data[i] = rnorm(1, mean=mu, sd=1)
43     }
44   }
45   return(data)
46 }
47
48 # Normal mixture model: theoretical cumulative distribution function cdf
49 twopointcdf = function(epsn, mu, t)
50 {
51   Ft = (1-epsn)*pnorm(t) + epsn*pnorm(t-mu)
52   return(Ft)
53 }
54
55 # Normal mixture model empirical cdf
56 twopointecdf = function(samp, t)

```

```

57 {
58   n = length(samp)
59   rat = (1/n)*sum(samp <= t)
60   return(rat)
61 }
62
63 # D of mu, t, tprime function with theoretical cdf
64 dcdcf = function(mu, t, tprime)
65 {
66   ratio = (pnorm(t) - pnorm(t-mu))/(pnorm(tprime)-pnorm(tprime-mu))
67   return(ratio)
68 }
69
70 # F functions for both max and min
71 Fminus = function(Alpha, samp, t)
72 {
73   a = (Alpha)^(2)
74   n = length(samp)
75   rh = (2*twopointecdf(samp, t)+(a/n)-sqrt((a/n)+4*twopointecdf(samp, t)-4*((
       twopointecdf(samp, t))^(2))))*sqrt(a/n))/(2*(1+(a/n)))
76   return(rh)
77 }
78 Fplus = function(Alpha, samp, t)
79 {
80   a = (Alpha)^(2)
81   n = length(samp)
82   rh = (2*twopointecdf(samp, t)+(a/n)+sqrt((a/n)+4*twopointecdf(samp, t)-4*((
       twopointecdf(samp, t))^(2))))*sqrt(a/n))/(2*(1+(a/n)))
83   return(rh)
84 }
85
86 # The slightly biased function with Fplus and Fminus
87 dplusminus = function(t, tprime, Alpha, samp)

```

```

88 {
89   result = (pnorm(t) - Fplus(Alpha,samp,t))/(pnorm(tprime) - Fminus(Alpha,samp
      ,tprime))
90   return(result)
91 }
92
93 # Obtain the grid points t
94 tfind = function(samp)
95 {
96   # sample size
97   n = length(samp)
98   # g = 0.2
99   # Initialize the t vector, containing all the tj's
100  t = vector()
101  for (j in 1:(2*log(n) +1))
102  {
103    t[j] = (j-1)/sqrt(2*log(n))
104  }
105  return(t)
106 }
107
108 # Functions for calculating the ratio of cdf
109 # of standard normal distribution for each pair of j an j+1
110 RATIO = function(samp)
111 {
112   # Initialize a void vector for containing the ratio
113   t = vector()
114   # Employ function tfind to obtain the grid
115   t = tfind(samp)
116   # Initialize a void vector for storing R[j]
117   R = vector()
118   # Loop from 1 to length(t)-1
119   for (j in 1 : length(t)-1)

```



```

120  {
121      R[j] = pnorm(t[j])/pnorm(t[j+1])
122  }
123  return(R)
124 }

125 # Function for computing the value of dnorm(t[j])/dnorm(t[j+1])
126 ratio = function(samp)
127 {
128     # Initialize a void vector for the storage of grid points
129     t = vector()
130     # Employ function tfind to obtain the grid
131     t = tfind(samp)
132     # Initialize a void vector for storing R[j]
133     R = vector()
134     # Loop from 1 to length(t)-1
135     for (j in 1 : length(t)-1)
136     {
137         r[j] = dnorm(t[j])/dnorm(t[j+1])
138     }
139     return(r)
140 }

141

142 # Obtain the values of dplusminus for each pair of (t[j],t[j+1])
143 Dvalue = function(Alpha,samp)
144 {
145     # Initialize void vectors for storing the value of D and t
146     D = vector()
147     t = vector()
148     # Employ jfind and tfind function to obtain jo and t vector respectively
149     t = tfind(samp)
150     # Loop from 1 to length(t)-1
151     for (j in 1 : length(t)-1)
152     {

```

```

153   D[j] = dplusminus(t[j], t[j+1],Alpha,samp)
154 }
155 return (D)
156 }
157 # Find the overestimated mu
158 Mufind = function(Alpha,samp)
159 {
160   # Initialize the t vector,containing the gird points of t
161   t = vector()
162   t = tfind(samp)
163   # Initialize the D vector
164   # D = vector()
165   # Obtain the values of dplusminus for each j
166   # Initialize the Omu vector for containing all the overestimated mu for each
       j
167   Omu = vector()
168   # Calculate D value
169   D = Dvalue(Alpha,samp)
170   # Find the overestimated mu
171   # Loop from 1 to length(Con)
172   for (j in 1:length(t)-1)
173   {
174     Omu[j] = tryCatch({Newton(function(mu) (D[j]-(1/(pnorm(t[j+1])-pnorm(t[j
+1]-mu)))*(pnorm(t[j])-pnorm(t[j]-mu))),x0 = 5.2)},
175               warning = function(w)
176               {
177                 return(w = 0)
178               },
179               error =function(e)
180               {
181                 return(e = 0)
182               })
183 }

```

```

184     return(Omu)
185 }
186 # The unbiased D value
187 Dtrue = function(epsn ,mun,samp)
188 {
189     D = vector()
190     t = tfind(samp)
191     for (j in 1: length(t)-1)
192     {
193         D[j] = (pnorm(t[j]) - twopointcdf(epsn ,mun,t[j]))/(pnorm(t[j+1]) -
            twopointcdf(epsn ,mun,t[j+1]))
194     }
195     return(D)
196 }
197 # Find the underestimated epsn
198 Epsnfind = function (Alpha, n, varep, mu)
199 {
200     samp = mixsample(n, varep, mu)
201     # Initialize a void vector to store t
202     t = vector()
203     # Obtain the grid points t
204     t = tfind(samp)
205     # Employ function RATIO and ratio respectively to obtain R and r
206     R = RATIO(samp)
207     r = ratio(samp)
208     #  $d = Dtrue(epsn, mu, samp)$ 
209     # Initialize vector Omu, containing all the overestimated mu for each j
210     Omu = vector()
211     # Obtain the value of overestimated mu
212     Omu = Mufind(Alpha,samp)
213     # Obtain D value
214     D = Dvalue(Alpha,samp)
215     # Initialize a vector to store underestimated epsn

```

```

216 Epsn = vector()
217 # Select the index j which enables D[j] satisfy the condition
218 Con = which(D >= R & D <= r)
219 # Obtain the estimators for the amount of nonzero means (varep)
220 for(j in 1 : length(Con))
221 {
222     Epsn[j] = (pnorm(t[Con[j]])-Fplus(Alpha,samp,t[Con[j]]))/(pnorm(t[Con[
        j]])-pnorm(t[Con[j]]-Omu[Con[j]]))
223
224 }
225 return(max(Epsn))
226 }
227 #####
228 #####
229 # The modified procedure
230 #####
231 #####
232 # Global constants
233 # n, the sample size
234 # mu, the theoretical mean of the second normal term in the normal mixture
       model
235 # varep, the coefficient of the second normal term in the normal mixture model
236 # b, the sparsity parameter which controls the amount of nonzero means
237 # C, the quantile of the limit of cdf H, which will change with alpha
238 n = 10^(8)
239 b = 1/3
240 r = 3/4
241 mu = sqrt(2*r*log(n))
242 varep = n^(-b)
243 C
244 #####
245 # Normal mixture model sampling
246 # The method for sampling from normal mixture model

```

```

247 mixsample = function(n,varep,mu)
248 {
249   # These are *cumulative* probabilities
250   # Generate random variables from a bernoulli distribution to select the
       relevant distribution
251   dists = rbinom(n,1,1-varep)
252   # This is where the actual data are generated, it's a if->then
253   # statement, followed by the normal distributions you are interested in
254   data = vector(length = n)
255   for(i in 1:n)
256   {
257     if(dists[i] == 1 )
258     {
259       data[i] = rnorm(1, mean=0, sd=1)
260     } else {
261       data[i] = rnorm(1, mean=mu, sd=1)
262     }
263   }
264   return(data)
265 }
266
267 # Two point-mixture model theoretical cumulative distribution function (cdf)
268 twopointcdf = function(epsn, mun, t)
269 {
270   Ft = (1-epsn)*pnorm(t) + epsn*pnorm(t-mun)
271   return(Ft)
272 }
273
274 # Normal mixture model empirical cdf
275 twopointecdf = function(samp,t)
276 {
277   rat = (1/length(samp))*sum(samp <= t)
278   return(rat)

```

```

279 }
280
281 # Functions for calculating the ratio of cdf of standard normal distribution
      for each pair of j and j+1
282 RATIO = function(samp)
283 {
284   # Initialize a void vector for containing the value of ratio
285   t = vector()
286   # Employ function tfind to obtain the grid
287   t = tfind(samp)
288   # Obtain the index jo
289   jo = jfind(samp)
290   # Initialize a void vector for storing R[j]
291   R = vector()
292   # Loop from 1 to jo
293   for (j in 1 : jo)
294   {
295     R[j] = pnorm(t[j]) / pnorm(t[j+1])
296   }
297   return(R)
298 }
299 # Function for computing the ratio of dnorm(t[j])/dnorm(t[j+1])
300 ratio = function(samp)
301 {
302   # Initialize a void vector for the storage of grid points
303   t = vector()
304   # Employ function tfind to obtain the grid
305   t = tfind(samp)
306   # Obtain index jo
307   jo = jfind(samp)
308   # Initialize a void vector for storing R[j]
309   R = vector()
310   # Loop from 1 to jo

```

```

311   for (j in 1 : jo)
312   {
313     r[j] = dnorm(t[j])/dnorm(t[j+1])
314   }
315   return(r)
316 }
317 # EFKP upper-class function q used to construct weighted brownian bridge
318 qweight = function(u)
319 {
320   return(sqrt(u*(1-u)*log(log(1/(u*(1-u))))))
321 }
322
323 # MF functions for both max and min
324
325 MFminus = function(C,samp,t)
326 {
327   n = length(samp)
328   rh = twopointecdf(samp,t) - (C/sqrt(n))*qweight(twopointecdf(samp,t))
329   return(max(0,rh))
330 }
331
332 MFplus = function(C,samp,t)
333 {
334   n = length(samp)
335   rh = twopointecdf(samp,t) + (C/sqrt(n))*qweight(twopointecdf(samp,t))
336   return(min(1,rh))
337 }
338
339 # Compute the biased D value with MFplus and MFminus
340 dplusminus = function(t,tprime,C,samp)
341 {
342   result = (pnorm(t) - MFplus(C,samp,t))/(pnorm(tprime)-MFminus(C,samp,tprime)
343   )

```

```

343   return(result)
344 }
345 # Obtain the grid points t
346 tfind = function(samp)
347 {
348   n = length(samp) #sample size
349   # Initialize the t vector, containing all the tj's
350   t = vector()
351   for (j in 1 : (8*log(n)+1))
352   {
353     t[j] = (j-1)/(sqrt(8*log(n)))
354   }
355   return(t)
356 }
357
358 # Find jo which satisfies t[jo+1] <= max(samp) <= t[jo+2]
359 jfind = function(samp)
360 {
361   t = vector()
362   t = tfind(samp)
363   for (j in 1 : length(t))
364   {
365     if (max(samp) >= t[j+1] & max(samp) <= t[j+2])
366     {
367       break
368     }
369     jo = j
370   }
371   return(jo)
372 }
373
374 # Compute the values of function dplusminus
375 Dvalue = function (C,samp)

```



```

376 {
377   # Initialize a void vector for storing the value of D
378   # Initialize a void vector for containing the values of t
379   D = vector()
380   t = vector()
381   # Employ jfind and tfind functon to obtain jo and t vector respectively
382   t = tfind(samp)
383   jo = jfind(samp)
384   for (j in 1 : jo)
385   {
386     D[j] = dplusminus(t[j],t[j+1],C,samp)
387   }
388   return (D)
389 }
390
391
392 # Find the overestimated mu
393 Mufind = function(C,samp)
394 {
395   # Initialize the Omu vector,containing the gird points of t
396   t = vector()
397   # Employ jfind and tfind functon to obtain jo and t vector respectively
398   t = tfind(samp)
399   # Initialize a void vector for storing dplusminus
400   D = vector()
401   # Use function jfind to get jo
402   jo = jfind(samp)
403   # Obtain the values of dplusminus for each j
404   D = Dvalue(C,samp)
405   # Initialize vector Omu for containing all the overestimated mu
     corresponding to each j
406   Omu = vector()
407   # Loop from 1 to jo

```

```

408 # Find the value of overestimated mu
409 for (j in 1: jo)
410 {
411   Omu[j] = tryCatch({Newton(function(mu) (D[j] - (1/(pnorm(t[j+1]) - pnorm(t[j]
+1] - mu))) * (pnorm(t[j]) - pnorm(t[j] - mu))), x0 = 4.6)},
412                       warning = function(w)
413                       {
414                           return(w = 0)
415                       },
416                       error = function(e)
417                       {
418                           return(e = 0)
419                       })
420 }
421 return(Omu)
422 }
423 # Unbiased D value with respect to mixture model
424 Dtrue = function(epsn, mun, samp)
425 {
426   D = vector()
427   t = tfind(samp)
428   jo = jfind(samp)
429   for (j in 1: jo)
430   {
431     D[j] = (pnorm(t[j]) - twopointcdf(epsn, mun, t[j])) / (pnorm(t[j+1]) -
twopointcdf(epsn, mun, t[j+1]))
432   }
433   return(D)
434 }
435 # Find the underestimated varep
436 Epsnfind = function (C, n, varep, mu)
437 {
438   samp = mixsample(n, varep, mu)

```

```

439  # Initialize a void vector to store t
440  t = vector()
441  # Obtain the grid points t
442  t = tfind(samp)
443  R = RATIO(samp)
444  r = ratio(samp)
445  # Initialize vector Omu, containing all the overestimated mu for each j
446  Omu = vector()
447  # Initialize vector D, containing all the values of dplusminus for each j
448  D = vector()
449  # Employ function Dvalue to obtain the values of dplusminus for each j
450  D = Dvalue(C,samp)
451  # Obtain the value of mu
452  Omu = Mufind(C,samp)
453  # Initialize a vector to store underestimated epsn
454  Epsn = vector()
455  # d = Dtrue(epsn, mu, samp)
456  # Select the index j which enables D[j] satisfy the condition as below
457  Con = which(D >= R & D <= r)
458  for(j in 1 : length(Con))
459  {
460    Epsn[j] = (pnorm(t[Con[j]]) - MFplus(C,samp,t[Con[j]])) / (pnorm(t[Con[j]]) -
      pnorm(t[Con[j]] - Omu[Con[j]]))
461
462  }
463  return(max(Epsn))
464 }
465 #####
466 #####
467 # Newton method
468 #####
469 #####
470 # f, the function that needs to be solved

```

```

471 # tol, the tolerant error
472 # x0, the initial value
473 # N, the maximum iteration number
474 Newton = function(f, tol = 1E-12, x0, N = 20000)
475 {
476   # h, the step size used for searching
477   h = 0.00001
478   i = 1; x1 = x0
479   # Initialize a void vector
480   p = vector()
481   while (i <= N) {
482     df.dx = (f(x0+h)-f(x0))/h
483     x1 = (x0 - (f(x0)/df.dx))
484     p[i] = x1
485     i = i + 1
486     # when certain condition is reached, jump out of the loop
487     if (abs(x1-x0) < tol) break
488     x0 = x1
489   }
490   return(p[i-1])
491 }
492
493 #####
494 #####
495 # The plot of confidence band
496 #####
497 #####
498
499 par(mfrow = c(3,1))
500 # Defined functions based on sample1 with sample size 100
501 x = seq(from = -4, to = 4, by = 0.00001)
502 f1 = function(x) MFminus(C, sample1, x)
503 g1 = function(x) MFplus(C, sample1, x)

```

```

504 u1 = function(x)Fminus(Alpha,sample1,x)
505 v1 = function(x)Fplus(Alpha,sample1,x)
506 p = function(x)twopointcdf(varep,mu,x)
507 # Plot functions respectively
508 plot(x,lapply(x,u),xlab =" n=100",ylab="confidence band value", ylim=c(0,1),
      type="l", col= "blue",lwd =1)
509 lines(x,lapply(x,v), type = "l",col = "blue",lwd =1)
510 lines(x,lapply(x,f), lty = 2,col = "red")
511 lines(x,lapply(x,g), lty = 2,col = "red")
512 lines(x,lapply(x,p),lty=1,col="green")
513 #####
514 # Define functions based on sample2 with the sample size 500
515 f2 = function(x)MFminus(C,sample2,x)
516 g2 = function(x)MFplus(C,sample2,x)
517 u2 = function(x)Fminus(Alpha,sample2,x)
518 v2 = function(x)Fplus(Alpha,sample2,x)
519 p = function(x)twopointcdf(varep,mu,x)
520 # Plot the functions respectively
521 plot(x,lapply(x,u),xlab =" n=500",ylab="confidence band value", ylim=c(0,1),
      type="l", col= "blue",lwd =1)
522 lines(x,lapply(x,v), type = "l",col = "blue",lwd =1)
523 lines(x,lapply(x,f), lty = 2,col = "red")
524 lines(x,lapply(x,g), lty = 2,col = "red")
525 lines(x,lapply(x,p),lty=1,col="green")
526 #####
527 # Define functions based on sample 3 with the sample size 1000
528 f3 = function(x)MFminus(C,sample3,x)
529 g3 = function(x)MFplus(C,sample3,x)
530 u3 = function(x)Fminus(Alpha,sample3,x)
531 v3 = function(x)Fplus(Alpha,sample3,x)
532 p = function(x)twopointcdf(varep,mu,x)
533 # Plot the functions respectively
534 plot(x,lapply(x,u),xlab =" n=1000",ylab="confidence band value", ylim=c(0,1),

```

```
    type="l", col= "blue",lwd =1)
535 lines(x,lapply(x,v), type = "l",col = "blue",lwd =1)
536 lines(x,lapply(x,f), lty = 2,col = "red")
537 lines(x,lapply(x,g), lty = 2,col = "red")
538 lines(x,lapply(x,p),lty=1,col="green")
```