Optimal Component Selection in High Dimensions

by

Xin Cui

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

Mathematics

Ottawa-Carleton Institute for
Mathematics and Statistics

Carleton University
Ottawa, Ontario

© 2014 Xin Cui

Abstract

It is now the modern trend and reality in various fields of life and science that the data sets to be analyzed are high dimensional, and the number of observations is much smaller than their dimension. As the classical statistical methods are not designed to deal with big or high-dimensional data, the problem of developing new methods of high-dimensional statistical analysis is very important. In many statistical applications such as analysis of microarray data, signal recovery, and functional magnetic resonance imaging, the focus is often on identifying and estimating a relatively few significant components from a high-dimensional vector. In this thesis, we study the problem of *component* or *variable selection* in a normal mixture model based on a single high-dimensional observation. The goal is to examine the possibilities and limitations of optimal component selection in a two-point normal mixture model and, whenever possible, to construct optimal (non-improvable) selection procedures. In addition to that, the problem of estimating an unknown parameter that determines the sparsity pattern of the data is addressed. The main theoretical findings of the thesis obtained in Chapter 1, and partly in Chapter 2, are supported by the simulation study; the results of the simulation study are presented in Chapter 3. The main results of the thesis, Theorems 2–5, are new.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Natalia Stepanova for her support during my graduate studies, patience, enthusiasm, and knowledge. Her guidance helped me very much to conduct research and write this thesis.

Also, I would like to thank my parents Bensheng Cui and Ping Wang for supporting me throughout my life.

Notation

$\{a_n\}$	A sequence of elements $a_1, a_2, \dots$
$\lim_{n \rightarrow \infty} a_n = a$	a is the limit of the sequence a_n
$a_n \sim b_n$	Means that $\lim_{n \rightarrow \infty} a_n/b_n = 1$
$\limsup_{n \rightarrow \infty} a_n$	The superior limit, $\limsup_{n \rightarrow \infty} a_n = \inf_{n \geq 1} \sup_{k \geq n} a_k$
$\liminf_{n \rightarrow \infty} a_n$	The inferior limit, $\liminf_{n \rightarrow \infty} a_n = \sup_{n \geq 1} \inf_{k \geq n} a_k$
$\log a$	Natural logarithm (base e) of the number a
$\mathbb{R}$	Set of the real number
$\mathbb{R}^d$	d -dimensional Eudidean vector space
$\mathbb{R}_+$	Set of the non-negative real numbers
$N(a, \sigma^2)$	Normal distribution with mean a and variance σ^2
$N(0, I_d)$	Standard normal distribution in $\mathbb{R}^d$
$\varphi(\cdot)$	Density function of the distribution $N(0, 1)$
$\Phi(\cdot)$	Distribution function of the distribution $N(0, 1)$
$L(\theta, \phi)$	The loss function in a decision problem with parameter θ and decision rule ϕ
$\xrightarrow{P}$	Convergence in probability
$\xrightarrow{d}$	Convergence in distribution
$f(x) = O(g(x))$ for $x \rightarrow a$	Symbol “big O ” means: $f(x)/g(x) \rightarrow C$ ($C = \text{const}, C \neq 0$) as $x \rightarrow a$
$f(x) = o(g(x))$ for $x \rightarrow a$	Symbol “small o ” means: $f(x)/g(x) \rightarrow 0$, as $x \rightarrow a$
$\mathbb{1}(A)$	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
$[x]$	The integer part of the real number x

Contents

Abstract	ii
Acknowledgements	iii
Notation	vi
Contents	viii
List of Tables	ix
List of Figures	x
Introduction	1
1 Exact and almost full component selection	3
1.1 Problem statement	3
1.2 Preliminaries	6
1.3 Main results	10
1.3.1 Exact selection: upper and lower bounds on the risk	12
1.3.2 Almost full selection: upper and lower bounds on the risk	14
1.4 Remarks and comments on Theorems 2–5	16
1.5 Proofs of Theorems 2–5	17

2	Estimation of the sparsity index β	31
2.1	Estimation of β by using the Meinshausen-Rice approach	32
2.2	Estimation of β in the exact selection regime	37
2.3	Numerical comparison and conclusions	42
3	Simulation studies	44
3.1	Exact and almost full selection	44
3.2	Estimation of the sparsity index β	45
3.3	Numerical results	47
	Conclusion	50
	Appendix A Proofs of Lemmas	52
	Appendix B Simulation R Codes	55
	Bibliography	58

List of Tables

1.1	Summary of selection models	17
3.1	Selection and estimation in the exact selection regime	47
3.2	Selection and estimation in the almost full selection regime	48

List of Figures

1.1	Three selection regions	18
3.1	Six histograms of the simulated normal mixture data	49

Introduction

In the traditional data analysis, we consider the observations that are measured on a limited number of well chosen components. Today the trend is towards more and more components in a single observation. The growth of dimensionality of the observation vector now is the most concerned issue with the traditional statistical approaches. More efforts are made in studying the new methods of high-dimensional data analysis, and, in recent years, high-dimensional data is becoming more and more common place in diverse fields of science, economics, engineering, and finance. The wide applications of high-dimensional analysis lead us to pursuit more effective approaches for data collection and processing of all kinds. The sparsity in high-dimensional data characterizes various contemporary problems in statistics. In this thesis, which is partially inspired by the conference talk of Donoho [4], we study the sparsity behaviour in order to approach component selection models by taking advantage of the “blessing of the high dimension”. We consider the situation where a single d -dimensional normal vector of a large dimension d is observed and the data have a sparse component with non-zero mean. The sparse structure is determined by the *sparsity index* $\beta \in (0, 1)$. To detect the *informative*, i.e., large, components with different degree of accuracy, we use two approaches, called *exact selection* and *almost full selection*. Exact selection is more accurate but less feasible due to the restrictions on the model required. Almost full selection provides less accuracy but also with less restrictions on the model. In this thesis, we construct

the selection procedures that provide exact and almost full selection of the informative components of a sparse high-dimensional normal vector. We show that these procedures are optimal in the asymptotically minimax sense as dimension d grows to infinity (see Theorems 2–5).

Furthermore, we estimate the sparsity parameter $\beta \in (0, 1)$ that controls the proportion of informative components. Using the results of Meinshausen and Rice [8], we can estimate the proportion of informative components, which, in turn, gives an estimator of the sparsity parameter β . Furthermore, in the exact selection scenario, we propose a new estimator of the sparsity parameter β . The simulation study conducted in the thesis shows that our estimator is better as compared to the Meinshausen–Rice estimator, provided the exact selection conditions are fulfilled. Simulations are run to support our analytical findings and to make comparisons between theory and practice. In our study, we only consider the case of a single high-dimensional normal observation. In a more complex case, when a whole sample of size n of d -dimensional observations is available, in order to construct optimal selection procedures, certain relationship between the sample size n and the dimension d must be satisfied. See, for example, Section 4 of [5], where a similar problem is studied in somewhat different settings.

The thesis consists of three main parts. In Chapter 1, we study the problem of exact and almost full optimal selection as dimension d grows indefinitely. In Chapter 2, we discuss two different methods of estimation of a sparsity parameter β . Chapter 3 contains the results of simulation studies. Appendix A collects the proofs of auxiliary results. The R codes used to run simulations are given in Appendix B.

Chapter 1

Exact and almost full component selection

1.1 Problem statement

Consider a situation in which a small fraction of informative components is hidden among a large number of the noisy components. In such a situation, it is hard to make accurate and reliable selection. The goal is to study how to detect informative components of a sparse high-dimensional vector in an optimal way. In this thesis, we investigate the possibilities and limitations in the problem of such selection in a Gaussian framework.

Suppose that a d -dimensional normal vector $X = (X_1, \dots, X_d)$ is observed. We assume that the vector X has the form

$$X = \mu + \varepsilon, \tag{1.1}$$

where $\varepsilon \sim N(0, I_d)$ is the noisy vector with I_d being a $d \times d$ identity matrix, and μ is a

sparse mean vector given by

$$\mu = a\eta, \quad a = a(d, \beta) \in \mathbb{R}_+, \quad \eta = (\eta_1, \dots, \eta_d) \in H_{d,\beta}$$

with the set $H_{d,\beta}$ of d -dimensional vectors defined for some constant $C_1 > 1$ as

$$H_{d,\beta} = \left\{ \eta = (\eta_1, \dots, \eta_d) : \eta_j \in \{0, 1\}, \quad \sum_{j=1}^d \eta_j \leq C_1 d^{1-\beta} \right\}. \quad (1.2)$$

The parameter $\beta \in (0, 1)$ is usually called the *sparsity index*. The parameter a is a “nuisance parameter” in this problem. We assume that β is unknown and that $d \rightarrow \infty$.

We shall also need the set

$$H_{d,\beta}^\pm = \left\{ \eta = (\eta_1, \dots, \eta_d) : \eta_j \in \{0, 1\}, \quad C_0 d^{1-\beta} \leq \sum_{j=1}^d \eta_j \leq C_1 d^{1-\beta} \right\}, \quad (1.3)$$

with constants $0 < C_0 < 1 < C_1 < \infty$ and $\beta \in (0, 1)$, where C_1 is the same as in (1.2).

It is clear that $H_{d,\beta}^\pm \subset H_{d,\beta}$. The sparsity assumption (1.3), which is quite common when dealing with spares high-dimensional models, was introduced in Ingster et al. [6] in connection with a high-dimensional classification problem.

According to model (1.1), each component X_j , $j = 1, \dots, d$, is composed by the noisy variable ε_j , and the index variable η_j . If $\eta_j = 1$, then X_j is a component with distribution $N(a, 1)$, otherwise it has distribution $N(0, 1)$. Generally speaking, this model is high-dimensional when d is very large, and, when the majority of the components of η are zero, the model is sparse. The quantity $d^{1-\beta}$ is roughly the number of the non-zero components of η , and $d^{-\beta}$ is roughly the proportion of non-zero components of η .

In this thesis, we study the problem of optimal component selection of a high-dimensional sparse normal vector X in model (1.1). Based on the observed value of

X , we wish to determine which components of X are $N(a, 1)$ distributed. Our selection procedures will be based on an estimator $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d)$ of $\eta = (\eta_1, \dots, \eta_d)$ whose components take on just two values zero and one. Such an estimator will be called a *selector*.

We measure the accuracy of a selector $\hat{\eta} = \hat{\eta}(X)$ by looking at its maximum risk. Let

$$|\hat{\eta} - \eta| = \sum_{j=1}^d |\hat{\eta}_j - \eta_j|$$

be the *Hamming distance*, which counts the number of positions at which $\hat{\eta}$ differs from η . Then the *maximum risk* of a selector $\hat{\eta}$ of vector η in model (1.1) is defined as follows:

$$R(\hat{\eta}) = \sup_{\eta \in H_{d,\beta}} E|\hat{\eta} - \eta| = \sup_{\eta \in H_{d,\beta}} E \sum_{j=1}^d |\hat{\eta}_j - \eta_j|.$$

The *minimax risk* is then defined by

$$\inf_{\tilde{\eta}} R(\tilde{\eta}) = \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}} E|\tilde{\eta} - \eta| = \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}} E \sum_{j=1}^d |\tilde{\eta}_j - \eta_j|,$$

where the infimum is taken over all selectors $\tilde{\eta}$ based on X . We wish to find a selector $\hat{\eta}$ whose maximum risk as $d \rightarrow \infty$ is as small as possible. A selector $\hat{\eta}$ that satisfies

$$R(\hat{\eta}) = o(1), \quad d \rightarrow \infty, \tag{1.4}$$

will be called the *exact selector*. A selector $\tilde{\eta}$ that satisfies

$$R(\tilde{\eta}) = o(d^{1-\beta}), \quad d \rightarrow \infty, \tag{1.5}$$

will be called the *almost full selector*.

Asymptotically, as $d \rightarrow \infty$, the selector $\hat{\eta}$ that satisfies (1.4) selects correctly all the informative components of vector X , whereas the selector $\tilde{\eta}$ that satisfies (1.5) selects correctly almost all informative components of X . The conditions that make the exact and almost full selection possible are expressed in terms of the parameter a . In Section 1.3, we show that if the parameter $a = a(d, \beta)$ is such that

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2}(1 + \sqrt{1 - \beta}),$$

then there exists a selector $\hat{\eta}$ for which relation (1.4) holds true. Relation (1.4) says that, asymptotically, the exact selection can be achieved by means of $\hat{\eta}$. Next, we show that if the parameter a satisfies a less restrictive condition

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2\beta},$$

then there exists a selector $\tilde{\eta}$ for which relation (1.5) holds true, that is, asymptotically almost full selection is possible by means of $\tilde{\eta}$. These two statements will be proved in this thesis. We also show (see Theorem 5 in Section 1.3) that if the parameter a is such that

$$\limsup_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} < \sqrt{2\beta},$$

no selection procedure exists that provides either exact or almost full selection.

1.2 Preliminaries

This section collects some common notations and results of probability theory and mathematical statistics that will be used in this thesis. These results can be found in many

textbooks on mathematical statistics (see, e.g., [3], [7], and [9]).

1. Minimax and Bayesian Estimation. Let $Y_1, \dots, Y_n$ be independent identically distributed (iid) random variables with common distribution function (df) $F_\theta(x)$, $x \in \mathbb{R}$, indexed by an unknown real-valued parameter $\theta \in \Theta \subset \mathbb{R}$. Suppose we wish to estimate θ by an estimator $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$. We measure the quality of an estimator by using a loss function. A *loss function* $L(\theta, \hat{\theta}_n)$ is a nonnegative function that generally increases as the “distance” between $\hat{\theta}_n$ and θ increases. (In Chapter 2 the role of loss function will be taken by the Hamming distance.) If the statistician chooses the estimator $\hat{\theta}_n$, then his/her expected loss will be

$$R(\theta, \hat{\theta}_n) = E_\theta(L(\theta, \hat{\theta}_n)), \quad \theta \in \Theta.$$

The function $R(\theta, \hat{\theta}_n)$ is called the *risk function* or *risk*. It represents the average loss that will be incurred if the estimator $\hat{\theta}_n$ is used. Generally speaking, the smaller the risk is, the better the estimator is.

Among various criteria of a “good” estimator, the minimax and Bayesian criteria are commonly used. An estimator $\hat{\theta}_n$ is called *minimax* if for every $n \geq 1$

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_n) = \inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}_n),$$

where the infimum is taken over all estimators $\tilde{\theta}_n$ of θ based on $Y_1, \dots, Y_n$. In words, the minimax estimator minimizes the maximum risk. The minimax principle gives much weight to the maximum risk. When applying this principle, we wish to guard ourselves from the worst that can happen. An estimator $\hat{\theta}_n$ that satisfies a less restrictive condition

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_n) \sim \inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}_n), \quad n \rightarrow \infty,$$

is called *asymptotically minimax*.

Another common concept of estimation theory is that of a Bayes estimator. In a Bayesian framework, the parameter θ is assumed to be (a realization of) a random variable with probability density function (pdf) or probability mass function (pmf) π , called the *prior pdf* or *pmf*. If $\pi(\theta)$ is the prior pdf of $\theta \in \Theta$, then the Bayes risk of $\hat{\theta}_n$ is given by

$$R_B(\hat{\theta}_n) = \int_{\Theta} R(\theta, \hat{\theta}_n) \pi(\theta) d\theta.$$

An estimator $\hat{\theta}_n$ is called *Bayesian* with respect to prior π if it minimizes the Bayes risk, that is, if

$$R_B(\hat{\theta}_n) = \inf_{\tilde{\theta}_n} R_B(\tilde{\theta}_n).$$

2. Bayes Tests. Let $Y = (Y_1, \dots, Y_n)$ be a random sample with pdf $f(y; \theta)$, $\theta \in \Theta \subseteq \mathbb{R}$, $y \in \mathbb{R}^n$. Consider the problem of testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are two given values from the parameter space Θ . Assume that θ is a realization of a random variable θ with prior pmf π given by

$$\pi(\theta = \theta_0) = p_0, \quad \pi(\theta = \theta_1) = p_1. \quad (1.6)$$

where $p_1 = 1 - p_0$ and $0 < p_0 < 1$. Let ϕ be a decision function with critical region B taking the values 0 and 1, that is, $\phi(Y) = \mathbb{I}(Y \in B)$. The function ϕ is called a *non-randomized decision function* or *non-randomized test function*. Assume that the loss

function $L(\theta, \phi)$ has the form

$$L(\theta, \phi) = \begin{cases} 0, & \text{if } \theta \in \theta_0 \text{ and } \phi = 0 \quad \text{or} \quad \theta \in \theta_1 \text{ and } \phi = 1, \\ L_1, & \text{if } \theta \in \theta_0 \text{ and } \phi = 1, \\ L_2, & \text{if } \theta \in \theta_1 \text{ and } \phi = 0. \end{cases}$$

The corresponding risk function is equal to

$$R(\theta, \phi) = L(\theta, 1)P_\theta(Y \in B) + L(\theta, 0)P_\theta(Y \in B^c), \quad (1.7)$$

or

$$R(\theta, \phi) = \begin{cases} L_1 P_\theta(Y \in B), & \text{if } \theta = \theta_0 \\ L_2 P_\theta(Y \in B^c), & \text{if } \theta = \theta_1. \end{cases}$$

In connection with testing two simple hypotheses as above, there exists a Bayes test function, which is actually a likelihood ratio test. Namely, the following result holds true.

Theorem 1 (Theorem 8 on p. 310 of [9]). *Let $Y = (Y_1, \dots, Y_n)$ be a random sample with pdf $f(y; \theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$, $y \in \mathbb{R}^n$, and let π be a probability distribution on Θ as in (1.6). Then for testing the hypothesis $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ there exists a Bayes decision function, or test, ϕ_π corresponding to the prior distribution π , that is, a decision function which minimizes the Bayes risk*

$$R(\phi) = p_0 R(\theta_0, \phi) + p_1 R(\theta_1, \phi), \quad (1.8)$$

and is given by

$$\phi_\pi(y) = \begin{cases} 1, & \text{if } f(y; \theta_1) > \frac{p_0 L_1}{p_1 L_2} f(y; \theta_0), \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the Bayes test with respect to prior π as in (1.6) has the critical region

$$B = \left\{ y \in \mathbb{R}^n : f(y; \theta_1) > \frac{p_0 L_1}{p_1 L_2} f(y; \theta_0) \right\}.$$

Theorem 1 was found to be a very useful tool in various minimax estimation problems. We will use this result to prove Theorems 3 and 5.

3. Markov's Inequality. Let Y be a random variable taking on non-negative values, and let $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a monotonically increasing function. Then for any $y > 0$ (see, e.g., Section 4.7 of [1])

$$P(Y \geq y) \leq \frac{E(h(Y))}{h(y)}.$$

Note that Chebyshev's inequality is a particular case of Markov's inequality. The former is obtained by taking $Y = |\xi - E(\xi)|$, where ξ is a random variable with finite variance, and $h(y) = y^2$.

1.3 Main results

Consider model (1.1). For both selection strategies of interest, i.e., exact selection and almost full selection, before exploring in detail the possibilities of constructing optimal, i.e., asymptotically minimax, selection procedures, we will discuss the upper bound on the maximum risk of a selector $\hat{\eta} = \hat{\eta}(X)$ and the lower bound on the minimax risk under

slightly different conditions on vector η . Due to the nature of sparsity, total number of non-zero components $\sum_{j=1}^d \eta_j$ is small, as compared to d . We consider two similar sets of conditions on η : (a) $\eta \in H_{d,\beta}$ and (b) $\eta \in H_{d,\beta}^\pm$ (see (1.2) and (1.3) for the definition). Since $H_{d,\beta}^\pm \subseteq H_{d,\beta}$, it follows that

$$\sup_{\eta \in H_{d,\beta}^\pm} E_\eta |\eta - \hat{\eta}| \leq \sup_{\eta \in H_{d,\beta}} E_\eta |\eta - \hat{\eta}|.$$

So we get that

$$\limsup_{d \rightarrow \infty} \sup_{\eta \in H_{d,\beta}^\pm} E_\eta |\eta - \hat{\eta}| = 0,$$

as long as

$$\limsup_{d \rightarrow \infty} \sup_{\eta \in H_{d,\beta}} E_\eta |\eta - \hat{\eta}| = 0.$$

Also, we will have

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}} E_\eta |\eta - \tilde{\eta}| > 0,$$

as long as

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^\pm} E_\eta |\eta - \tilde{\eta}| > 0.$$

In words, the upper bound on the maximum risk over $H_{d,\beta}$ implies the upper bound on the maximum risk over $H_{d,\beta}^\pm$. Furthermore, the lower bound on the minimax risk with the supremum over $H_{d,\beta}^\pm$ implies the lower bound on the minimax risk with the supremum over $H_{d,\beta}$. Thus, in order to obtain the upper and lower bounds for both kinds of suprema, over $H_{d,\beta}$ and over $H_{d,\beta}^\pm$, we only need to consider the upper bound for $H_{d,\beta}$ and the lower bound for $H_{d,\beta}^\pm$.

1.3.1 Exact selection: upper and lower bounds on the risk

In this subsection, we show that under the following condition on the parameter $a = a(d, \beta)$ in model (1.1):

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2}(1 + \sqrt{1 - \beta}),$$

exact selection is possible. To this end, we define a selector $\hat{\eta} = \hat{\eta}(X)$ by the formula

$$\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d), \quad \hat{\eta}_j = \hat{\eta}_j(X) = \mathbb{1}(X_j > \sqrt{(2 + \delta) \log d}), \quad j = 1, \dots, d, \quad (1.9)$$

with some small number $\delta \in (0, 1)$. Specifically, we assume that $\delta = \delta(d) > 0$ satisfies

$$\delta \rightarrow 0 \quad \text{and} \quad \delta \log d \rightarrow \infty, \quad \text{as } d \rightarrow \infty.$$

This technical assumption on the behaviour of δ is needed to prove Theorems 2 and 4 below. Under this assumption, for any constant $c > 0$ one has $d^{-c\delta} = o(1)$ as $d \rightarrow \infty$.

First of all, one might wonder where this exact selector comes from. The idea behind selector (1.9) is simple. We need to consider the most influence that the ε_j 's with distribution $N(0, 1)$ can have to obstruct the determinations of the informative components of X . That is why we need to find the bound for $\max_{1 \leq j \leq d} |\varepsilon_j|$ when $d \rightarrow \infty$. Then, if we set the threshold level larger than $\max_{1 \leq j \leq d} |\varepsilon_j|$, in expectation, we can avoid all the interference from the noisy variables. Then, we define a selector with the threshold value larger than the obtained upper bound. We expect that the thus constructed selector will provide the required selection. The construction of the selector as in (1.9) is based on the following well-known property of a standard normal distribution.

Lemma 1. *Let $\xi_1, \xi_2 \dots$ be a sequence of iid normal $N(0, 1)$ random variables. Then*

$$P\left(\max_{1 \leq j \leq d} |\xi_j| \geq \sqrt{2 \log d}\right) = o(1), \quad d \rightarrow \infty.$$

Lemma 1 will be proved in Appendix A. Having Lemma 1 available, we can decide whether or not a given component of vector X is informative or non-informative by comparing it to $\sqrt{(2 + \delta) \log d}$ with some small $\delta \in (0, 1)$.

With Lemma 1 in mind, we shall now discuss the choice of interesting values of the nuisance parameter a . For $j = 1, \dots, d$, if the observed value of X_j is larger than $\sqrt{(2 + \delta) \log d}$, we set $\hat{\eta}_j = 1$, and X_j will be considered as $N(a, 1)$ distributed. Thus, the parameter a should be at least as large as $\sqrt{(2 + \delta) \log d}$ for exact selection. We will see, however, that this is not enough. Namely, the following result holds true.

Theorem 2. *If the parameter a in model (1.1) satisfies the condition*

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2}(1 + \sqrt{1 - \beta}), \quad (1.10)$$

then

$$\limsup_{d \rightarrow \infty} \sup_{\eta \in H_{d,\beta}} E_\eta |\eta - \hat{\eta}| = 0. \quad (1.11)$$

In words, the selector $\hat{\eta}$ as in (1.9) provides exact component selection for vector X in model (1.1).

Remark 1. Notice that the upper bound (1.11) implies $\limsup_{d \rightarrow \infty} \sup_{\eta \in H_{d,\beta}^\pm} E|\eta - \hat{\eta}| = 0$. In words, the selector $\hat{\eta}$ is also exact with respect to the maximum risk when the maximum is taken over $H_{d,\beta}^\pm$.

Theorem 3. *If the parameter a in model (1.1) is such that*

$$\limsup_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} < \sqrt{2}(1 + \sqrt{1 - \beta}),$$

then

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^{\pm}} E_{\eta} |\eta - \tilde{\eta}| > 0,$$

where the infimum is taken over all selectors $\tilde{\eta}$ based on X .

This theorem says that, under the above condition on the parameter a , exact variable selection cannot be achieved, which will lead us to consider almost full selection.

Remark 2. The lower bound on the minimax risk obtained in Theorem 3 implies that $\limsup_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}} E |\eta - \tilde{\eta}| > 0$. Thus, under the conditions of Theorem 3, no exact selection is also possible when the supremum is taken over all $\eta \in H_{d,\beta}^{\pm}$.

1.3.2 Almost full selection: upper and lower bounds on the risk

As mentioned above, the value of the parameter a should be large enough to make the exact selection possible. If the threshold is set to be less than $\sqrt{2 \log d}$, it could be smaller than the value of $\max_{1 \leq j \leq d} |\varepsilon_j|$. In this case, the mean value may not have a significant effect on the upper extremes and we may not be able to detect the informative components from the Gaussian random noise directly. In other words, if condition (1.10) on the parameter a does not hold, which means that

$$\limsup_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} < \sqrt{2}(1 + \sqrt{1 - \beta}),$$

we no longer can achieve exact selection. When it happens, we can try to use the almost full selection approach, under which one asymptotically selects correctly most of the informative components of vector X .

Recall that a selector $\tilde{\eta} = \tilde{\eta}(X)$ yields almost full selection in the problem at hand if one has

$$R(\tilde{\eta}) = \sup_{\eta \in H_{d,\beta}} E_{\eta} |\eta - \tilde{\eta}| = o(d^{1-\beta}), \quad d \rightarrow \infty.$$

This means that the maximum risk of $\tilde{\eta}$ is small as compared to the total number of the non-zero components $\sum_{j=1}^d \eta_j = O(d^{1-\beta})$. Now we will show that the selector given by

$$\tilde{\eta} = (\tilde{\eta}_1, \dots, \tilde{\eta}_d), \quad \tilde{\eta}_j = \tilde{\eta}_j(X) = \mathbb{1} \left(X_j > \sqrt{(2\beta + \delta) \log d} \right), \quad j = 1, \dots, d, \quad (1.12)$$

where δ is the same as in the definition (1.9), is of this kind.

Theorem 4. *If the parameter a in model (1.1) is such that*

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2\beta}, \quad (1.13)$$

then

$$\limsup_{d \rightarrow \infty} \sup_{\eta \in H_{d,\beta}} d^{\beta-1} E_{\eta} |\eta - \tilde{\eta}| = 0.$$

In words, the selector $\tilde{\eta}$ provides almost full selection.

Next theorem gives, under certain condition on the parameter a , the lower bound on the minimax risk normalized by $d^{\beta-1}$.

Theorem 5. *If the parameter a in model (1.1) is such that*

$$\limsup_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} < \sqrt{2\beta}, \quad (1.14)$$

then

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^{\pm}} d^{\beta-1} E_{\eta} |\eta - \tilde{\eta}| > 0,$$

that is, neither exact nor almost full selection is possible.

It follows from Theorem 5 that, under condition (1.14), none of the selection strategy can provide satisfied results with a reasonably small risk. The noisy components and the “signal” components are mixed and can not be separated by detecting the informative components of X . Before giving the proofs of Theorems 2–5, we pause to comment on the results obtained.

1.4 Remarks and comments on Theorems 2–5

According to Theorem 2, the selector $\hat{\eta}$ as in (1.9) satisfies

$$R(\hat{\eta}) = o(1), \quad d \rightarrow \infty,$$

that is, for all large enough d , $\hat{\eta}$ selects correctly all informative components of a normal vector X in model (1.1). Also, by Theorem 4, the selector $\tilde{\eta}$ as in (1.12) satisfies

$$R(\tilde{\eta}) = o(d^{1-\beta}), \quad d \rightarrow \infty,$$

that is, in the limit, $\tilde{\eta}$ selects correctly most of the informative components of vector X in model (1.1).

To give a graphical representation of Theorems 2–5, we define the functions

$$\phi_1(\beta) = \sqrt{2}(1 + \sqrt{1 - \beta}), \quad \phi_2(\beta) = \sqrt{2\beta}, \quad 0 < \beta < 1,$$

and assume that the parameter a is of the form

$$a = \gamma\sqrt{\log d}, \quad 0 < \gamma < \infty.$$

Table 1.1 provides the selection summary depending on the values of γ .

Table 1.1: Summary of selection models

Selection	Condition on γ	Selection threshold	Maximum risk
Exact	$\phi_1(\beta) < \gamma < \infty$	$\sqrt{(2 + \delta) \log d}$	$o(1)$
Almost	$\phi_2(\beta) < \gamma < \phi_1(\beta)$	$\sqrt{(2\beta + \delta) \log d}$	$o(d^{1-\beta})$
No selection	$0 < \gamma < \phi_2(\beta)$	None	$O(d^{1-\beta})$

When $\gamma > \phi_1(\beta) = \sqrt{2}(1 + \sqrt{1 - \beta})$ exact selection is available, whereas when $\gamma > \phi_2(\beta) = \sqrt{2\beta}$ almost full selection is available. If $\gamma < \phi_2(\beta) = \sqrt{2\beta}$ neither exact nor almost full selection is available. Figure 1.1 illustrates this phenomenon graphically.

1.5 Proofs of Theorems 2–5

In this section, we shall prove Theorems 2–5 that constitute the main results of the thesis.

Below we will use the well-known relation (see, e.g., p. 414 in [1])

$$\Phi(-t) \sim \frac{e^{-t^2/2}}{\sqrt{2\pi}t}, \quad t \rightarrow \infty, \quad (1.15)$$

Figure 1.1: Three selection regions

as well as the inequality

$$\Phi(-t) \leq e^{-t^2/2}, \quad t \rightarrow \infty, \quad (1.16)$$

which follows from (1.15).

Proof of Theorem 2. Let the selector $\hat{\eta}$ be defined by (1.9). Consider the maximum risk of $\hat{\eta}$, which is given by

$$R(\hat{\eta}) = \sup_{\eta \in H_{d,\beta}} E_{\eta} |\hat{\eta} - \eta|.$$

Recalling that the number of indices j for which $\eta_j = 1$ is less than or equal to $C_1 d^{1-\beta}$, we have

$$\begin{aligned}
R(\hat{\eta}) &= \sup_{\eta \in H_{d,\beta}} E_\eta |\hat{\eta} - \eta| = \sup_{\eta \in H_{d,\beta}} E_\eta \sum_{j=1}^d |\hat{\eta}_j - \eta_j| = \sup_{\eta \in H_{d,\beta}} \sum_{j=1}^d E_{\eta_j} |\hat{\eta}_j - \eta_j| \\
&= \sup_{\eta \in H_{d,\beta}} \left(\sum_{j:\eta_j=0} E_{\eta_j} |\hat{\eta}_j - \eta_j| + \sum_{j:\eta_j=1} E_{\eta_j} |\hat{\eta}_j - \eta_j| \right) \\
&= \sup_{\eta \in H_{d,\beta}} \left(\sum_{j:\eta_j=0} P_{\eta_j} \left(X_j > \sqrt{(2+\delta) \log d} \right) + \sum_{j:\eta_j=1} P_{\eta_j} \left(X_j \leq \sqrt{(2+\delta) \log d} \right) \right) \\
&\leq d\Phi \left(-\sqrt{(2+\delta) \log d} \right) + C_1 d^{1-\beta} \Phi \left(\sqrt{(2+\delta) \log d} - a \right) \\
&=: I_1 + I_2.
\end{aligned} \tag{1.17}$$

By using (1.15), we obtain that the first term I_1 satisfies as $d \rightarrow \infty$

$$\begin{aligned}
I_1 &= d\Phi \left(-\sqrt{(2+\delta) \log d} \right) = O \left(d \exp \left(\left(-\frac{2+\delta}{2} \right) \log d \right) \frac{1}{\sqrt{\log d}} \right) \\
&= O \left(d^{-\delta/2} \log^{-1/2} d \right) = o(1).
\end{aligned}$$

Consider the second term I_2 . Note that for all sufficiently large d we can rewrite $\sqrt{(2+\delta) \log d}$ in the form $\sqrt{2 \log d} (1 + \delta_1)$ with some $\delta_1 = o(1)$. Next, by assumption

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2} (1 + \sqrt{1-\beta}).$$

This implies that for some positive $\delta_2 = o(1)$ and all sufficiently large d

$$a > \sqrt{2} (1 + \sqrt{1-\beta}) \sqrt{\log d} (1 + \delta_2),$$

where we may take δ_2 satisfying $\delta_2 > \delta_1$ and such that $d^{-c(\delta_2 - \delta_1)} = o(1)$ for any $c > 0$.

Now, using (1.16), the term I_2 is estimated from above as follows:

$$\begin{aligned}
I_2 &= C_1 d^{1-\beta} \Phi \left(- \left(a - \sqrt{(2+\delta) \log d} \right) \right) = C_1 d^{1-\beta} \Phi \left(- \left(a - \sqrt{2 \log d} (1 + \delta_1) \right) \right) \\
&\leq C_1 d^{1-\beta} \exp \left(- \left(a - \sqrt{2 \log d} (1 + \delta_1) \right)^2 \right) \\
&\leq C_1 d^{1-\beta} \exp \left(- \left(\left(1 + \sqrt{1+\beta} \right) (1 + \delta_2) - (1 + \delta_1) \right)^2 \log d \right) \\
&= C_1 d^{1-\beta} d^{-\left((1+\sqrt{1+\beta})(1+\delta_2) - (1+\delta_1) \right)^2} \leq C_1 d^{1-\beta} d^{-(\sqrt{1+\beta} + (\delta_2 - \delta_1))^2} = o(1).
\end{aligned}$$

Thus, under condition (1.10), $I_2 = o(1)$ and we conclude that

$$R(\hat{\eta}) = I_1 + I_2 = o(1) + o(1) = o(1).$$

This means that asymptotically the selector $\hat{\eta}$ selects correctly all the informative components of vector X when $\eta \in H_{d,\beta}$, and hence when $\eta \in H_{d,\beta}^\pm$. $\square$

Proof of Theorem 3. To prove the lower bound on the minimax risk, we bound the minimax risk from below by the Bayes risk, with a properly chosen prior distribution on η . Observe that, if $\pi(\eta)$ is the prior pdf of $\eta = (\eta_1, \dots, \eta_d)$ on a d -dimensional set $\mathfrak{X}^d$, then the maximum risk of a selector $\hat{\eta}$ is always greater than or equal to the Bayes risk:

$$\max_{\eta \in \mathfrak{X}^d} R(\eta, \hat{\eta}) \geq \int_{\mathfrak{X}^d} R(\eta, \hat{\eta}) \pi(\eta) d\eta. \quad (1.18)$$

With model (1.1) under consideration, the most natural choice of a prior distribution of $\eta = (\eta_1, \dots, \eta_d)$ is as follows:

$$\eta_1, \dots, \eta_d \stackrel{iid}{\sim} \text{Bernoulli}(p), \quad p = \frac{d^{1-\beta}}{d} = d^{-\beta}.$$

Roughly, we may interpret p as the probability of occurrences of non-zero components

of vector η . Let δ_x be the probability measure that has a total mass of 1 at point x , and let π_j denote the prior distribution of η_j , $j = 1, \dots, d$. Then, we can write

$$\pi_j = p\delta_1 + (1 - p)\delta_0 \quad j = 1, \dots, d.$$

It is clear that $E(\eta_j) = p$, $\text{Var}(\eta_j) = p(1 - p)$, and that the sum $\sum_{j=1}^d \eta_j$ has a binomial $\text{Bin}(d, p)$ distribution.

The above introduced prior distribution of vector $\eta = (\eta_1, \dots, \eta_d)$ is concentrated on the set $\{0, 1\}^d = \underbrace{\{0, 1\} \times \dots \times \{0, 1\}}_{d \text{ times}}$ and has the product form

$$\pi(\eta) = \prod_{j=1}^d \pi_j(\eta_j), \quad (1.19)$$

where $\pi_j = p\delta_1 + (1 - p)\delta_0$, $j = 1, \dots, d$. That is, if we now define the event $\mathfrak{X}^d$ as

$$\mathfrak{X}^d = \{\eta \in \{0, 1\}^d\},$$

then $\pi(\mathfrak{X}^d) = 1$. Next, consider the event

$$B_{d,\beta} := \{\eta \in H_{d,\beta}^\pm\} \subseteq \mathfrak{X}^d.$$

To estimate the minimax risk by the Bayes risk from below as in (1.18), we shall need the following result.

Lemma 2. *Let vector $\eta = (\eta_1, \dots, \eta_d)$ have the prior distribution π as in (1.19).*

Then for any $0 < c_0 < 1 < c_1 < \infty$, as $d \rightarrow \infty$

$$P\left(\sum_{j=1}^d \eta_j \in [c_0 d^{1-\beta}, c_1 d^{1-\beta}]\right) \geq 1 - o(d^{-1}),$$

The proof of Lemma 2 is given in Appendix A.

Now, we define the probability measure $\hat{\pi}$ on $B_{d,\beta}$ by the relation

$$\hat{\pi}(A) = \pi(A|B_{d,\beta}) = \frac{\pi(A \cap B_{d,\beta})}{\pi(B_{d,\beta})},$$

where, in view of Lemma 2,

$$\pi(B_{d,\beta}) = \pi(\eta \in H_{d,\beta}^\pm) = \pi\left(\sum_{j=1}^d \eta_j \in [C_0 d^{1-\beta}, C_1 d^{1-\beta}]\right) \geq 1 - o(d^{-1}).$$

Therefore, using inequality (1.18), as $d \rightarrow \infty$, we can estimate the minimax risk from below as follows:

$$\begin{aligned} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^\pm} E_\eta |\eta - \tilde{\eta}| &\geq \inf_{\tilde{\eta}} \int_{B_{d,\beta}} E_\eta |\eta - \tilde{\eta}| d\hat{\pi}(\eta) \\ &= \inf_{\tilde{\eta}} \frac{1}{\pi(B_{d,\beta})} \int_{B_{d,\beta}} E_\eta |\eta - \tilde{\eta}| d\pi(\eta) \\ &= \inf_{\tilde{\eta}} (1 + o(d^{-1})) \left(\int_{\mathcal{X}^d} E_\eta |\eta - \tilde{\eta}| d\pi(\eta) - \int_{B_{d,\beta}^c} E_\eta |\eta - \tilde{\eta}| d\pi(\eta) \right) \\ &\geq (1 + o(d^{-1})) \left(\inf_{\tilde{\eta}} \int_{\mathcal{X}^d} E_\eta |\eta - \tilde{\eta}| d\pi(\eta) - \sup_{\tilde{\eta}} \int_{B_{d,\beta}^c} E_\eta |\eta - \tilde{\eta}| d\pi(\eta) \right). \end{aligned}$$

We now show that $\inf_{\tilde{\eta}} \int_{\mathcal{X}^d} E_\eta |\eta - \tilde{\eta}| d\pi(\eta)$ is the main term, and $\sup_{\tilde{\eta}} \int_{B_{d,\beta}^c} E_\eta |\eta - \tilde{\eta}| d\pi(\eta)$

is the remainder term. We have,

$$\begin{aligned} \sup_{\tilde{\eta}} \int_{B_{d,\beta}^c} E_{\eta} |\eta - \tilde{\eta}| d\pi(\eta) &\leq \sup_{\tilde{\eta}} \max_{\sum_j \eta_j \notin [C_0 d^{1-\beta}, C_1 d^{1-\beta}]} \left(\sum_{j=1}^d E_{\eta_j} |\eta_j - \tilde{\eta}_j| \right) \cdot \pi(B_{d,\beta}^c) \\ &\leq \sup_{\tilde{\eta}} \max_{\sum_j \eta_j \notin [C_0 d^{1-\beta}, C_1 d^{1-\beta}]} \sum_{j=1}^d (E_{\eta_j} |\eta_j| + E_{\eta_j} |\tilde{\eta}_j|) \cdot o(d^{-1}), \end{aligned}$$

where $E_{\eta_j} |\eta_j| + E_{\eta_j} |\tilde{\eta}_j| \leq 2$. Therefore

$$\sup_{\tilde{\eta}} \int_{B_{d,\beta}^c} E_{\eta} |\eta - \tilde{\eta}| d\pi(\eta) \leq O(d) o(d^{-1}) = o(1).$$

Now we consider the main term. When $d \rightarrow \infty$,

$$\begin{aligned} \inf_{\tilde{\eta}} \int_{\mathfrak{X}^d} E_{\eta} |\eta - \tilde{\eta}| d\pi(\eta) &= \inf_{\tilde{\eta}} \int_{\mathfrak{X}^d} E_{\eta} \left(\sum_{j=1}^d |\eta_j - \tilde{\eta}_j| \right) d\pi(\eta) \\ &= \inf_{\tilde{\eta}} \int_{\mathfrak{X}^d} \left(\sum_{j=1}^d E_{\eta_j} |\eta_j - \tilde{\eta}_j| \right) d\pi(\eta) = \inf_{\tilde{\eta}} E_{\pi} \sum_{j=1}^d E_{\eta_j} |\eta_j - \tilde{\eta}_j| = \inf_{\tilde{\eta}} \sum_{j=1}^d E_{\pi} E_{\eta_j} |\eta_j - \tilde{\eta}_j| \\ &\geq \sum_{j=1}^d \inf_{\tilde{\eta}} E_{\pi} E_{\eta_j} |\eta_j - \tilde{\eta}_j| = d \cdot \inf_{\tilde{\eta}_1} E_{\pi_1} E_{\eta_1} |\eta_1 - \tilde{\eta}_1|, \end{aligned} \tag{1.20}$$

where

$$\eta_1 = \begin{cases} 1, & \text{with prob. } p, \\ 0, & \text{with prob. } 1 - p, \end{cases}$$

that is, the prior distribution of η_1 is $\pi_1 \equiv \text{Bernoulli}(p)$, and $\tilde{\eta}_1$ is an arbitrary selector of η_1 based on X_1 , the first component of vector X in model (1.1). Since the variable $\tilde{\eta}_1$ takes on only two values 0 and 1, we may think of $\tilde{\eta}_1$ as a (non-randomized) test function

and hence we may view the quantity

$$\inf_{\tilde{\eta}_1} E_{\pi_1} E_{\eta_1} |\eta_1 - \tilde{\eta}_1| = \inf_{\tilde{\eta}_1} [pE_a(1 - \tilde{\eta}_1) + (1 - p)E_0(\tilde{\eta}_1)]$$

as the *minimal Bayes risk* in the problem testing two simple hypotheses. Namely, the term

$$pE_a(1 - \tilde{\eta}_1) + (1 - p)E_0(\tilde{\eta}_1)$$

is the Bayes risk in the problem of testing

$$H_0 : v = 0 \quad \text{vs.} \quad H_1 : v = a$$

in the model

$$X_1 = v + \varepsilon_1, \quad \varepsilon_1 \sim N(0, 1) \quad \text{where } v = a\eta_1, \quad a \in \mathbb{R}_+.$$

Given a test $\tilde{\eta}_1$, the hypothesis H_0 is accepted when $\tilde{\eta}_1 = 0$ and rejected when $\tilde{\eta}_1 = 1$.

The loss function corresponding to the above Bayes risk is given by

$$L(v, \tilde{\eta}_1) = \begin{cases} 0, & \text{if } v = 0 \text{ and } \tilde{\eta}_1 = 0, \text{ or } v = a \text{ and } \tilde{\eta}_1 = 1, \\ 1, & \text{if } v = 0 \text{ and } \tilde{\eta}_1 = 1, \\ 1, & \text{if } v = a \text{ and } \tilde{\eta}_1 = 0. \end{cases}$$

Indeed, choosing the prior distribution π on v such that $\pi(v = a) = p$, $\pi(v = 0) = 1 - p$, we get the corresponding risk function to be equal to (see formula (1.7))

$$R(v, \tilde{\eta}_1) = L(v, 1)P_v(\tilde{\eta}_1 = 1) + L(v, 0)P_v(\tilde{\eta}_1 = 0),$$

or

$$R(v, \tilde{\eta}_1) = \begin{cases} P_0(\tilde{\eta}_1 = 1), & \text{if } v = 0, \\ P_a(\tilde{\eta}_1 = 0), & \text{if } v = a. \end{cases}$$

The average (Bayes) risk of $\tilde{\eta}_1$ is then equal to (see formula (1.8))

$$\begin{aligned} R(\tilde{\eta}_1) &= R(a, \tilde{\eta}_1)\pi(v = a) + R(0, \tilde{\eta}_1)\pi(v = 0) = pP_a(\tilde{\eta}_1 = 0) + (1 - p)P_0(\tilde{\eta}_1 = 1) \\ &= pE_a(1 - \tilde{\eta}_1) + (1 - p)E_0(\tilde{\eta}_1). \end{aligned}$$

Notice that we can rewrite the above testing problem in the form

$$H_0 : X_1 \sim f_0(x) \quad \text{vs.} \quad H_1 : X_1 \sim f_a(x),$$

where $f_0(x)$ and $f_a(x)$ are the pdf's of an $N(0, 1)$ and an $N(a, 1)$, respectively. The Bayes test that minimizes the average risk $R(\tilde{\eta}_1)$ has the rejection region given by (see Theorem 1 in Section 1.2)

$$B = \{x \in \mathbb{R} : f_a(x) > C f_0(x)\}, \quad C = (1 - p)/p.$$

Thus the minimal value of the Bayes risk is attained for the test based on a weighted likelihood ratio. That is, the Bayes test is given by

$$\eta^{\text{opt}}(X) = \mathbb{1} \left(\frac{p f_a(X)}{(1 - p) f_0(X)} > 1 \right).$$

Taking this into account, we may write

$$\begin{aligned}
\inf_{\tilde{\eta}_1} R(\tilde{\eta}_1) &= \inf_{\tilde{\eta}_1} (pP_a(\tilde{\eta}_1 = 0) + (1-p)P_0(\tilde{\eta}_1 = 1)) \\
&= pP_a(\eta^{\text{opt}}(X) = 0) + (1-p)P_0(\eta^{\text{opt}}(X) = 1) \\
&= pP_a\left(\frac{pf_a(X)}{(1-p)f_0(X)} \leq 1\right) + (1-p)P_0\left(\frac{pf_a(X)}{(1-p)f_0(X)} > 1\right),
\end{aligned}$$

where

$$f_0(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad f_a(x) = \frac{1}{\sqrt{2\pi}}e^{-(x-a)^2/2}.$$

Consider the inequality

$$\frac{p}{1-p} \frac{f_a(x)}{f_0(x)} > 1, \quad \text{or equivalently,} \quad \frac{p}{1-p} \exp\left(-\frac{a^2}{2} + x\right) > 1,$$

Solving it for x , we obtain

$$x > \frac{\log(\frac{1-p}{p})}{a} + \frac{a}{2}.$$

Denote for simplicity

$$T = \frac{\log(\frac{1-p}{p})}{a} + \frac{a}{2}, \tag{1.21}$$

where

$$a = a(d, \beta) \quad \text{and} \quad p = d^{-\beta} = o(1), \quad d \rightarrow \infty,$$

and note that

$$\log \left(\frac{1-p}{p} \right) = \log(d^\beta - 1) \sim \log d^\beta = \beta \log d.$$

Therefore, recalling (1.20), we get

$$\begin{aligned} \inf_{\tilde{\eta}} \int_{\mathfrak{X}^d} E_{\eta} |\eta - \tilde{\eta}| d\pi(\eta) &= d \inf_{\tilde{\eta}_1} E_{\pi_1} E_{\eta_1} |\eta_1 - \tilde{\eta}_1| \\ &= d (pP_a(X \leq T) + (1-p)P_0(X > T)) = d (p\Phi(T-a) + (1-p)\Phi(-T)) \\ &= d^{1-\beta}\Phi(T-a) + d(1-d^{-\beta})\Phi(-T) =: J_1 + J_2, \end{aligned} \tag{1.22}$$

where

$$J_1 = d^{1-\beta}\Phi(T-a), \quad J_2 = d(1-d^{-\beta})\Phi(-T).$$

with T defined as in (1.21). The terms J_1 and J_2 are both non-negative, so if for all sufficiently large d either one of them is bounded away from zero, we can conclude that for all sufficiently large d the left side of (1.22) is strictly positive.

For simplicity of derivations, we take a in the form $a = \gamma\sqrt{\log d}$, $\gamma > 0$. Then condition (1.13) on a gives us the following restriction on γ :

$$\gamma < \sqrt{2}(1 + \sqrt{1-\beta}). \tag{1.23}$$

Now, let us consider J_1 . We have

$$\begin{aligned} J_1 &= d^{1-\beta}\Phi(T-a) = d^{1-\beta}\Phi \left(\frac{\log(\frac{1-p}{p})}{a} - \frac{a}{2} \right) \\ &= d^{1-\beta}\Phi \left(\left(\frac{\beta}{\gamma} - \frac{\gamma}{2} \right) \sqrt{\log d} (1 + o(1)) \right) \end{aligned}$$

If $(\beta/\gamma - \gamma/2) \geq 0$, then $J_1 \rightarrow \infty$ as $d \rightarrow \infty$, and we are done. If $(\beta/\gamma - \gamma/2) < 0$, then by (1.15) we obtain

$$\begin{aligned} J_1 &\sim \frac{d^{1-\beta}}{\sqrt{2}} \exp \left\{ -\frac{1}{2} \left(\frac{\gamma}{2} - \frac{\beta}{\gamma} \right)^2 \log d \right\} / \left(\frac{\gamma}{2} - \frac{\beta}{\gamma} \right) \sqrt{\log d} \\ &= O \left(\frac{d^{1-\beta-\frac{1}{2}(\frac{\gamma}{2}-\frac{\beta}{\gamma})^2}}{\log^{1/2} d} \right). \end{aligned} \quad (1.24)$$

Recall that for any $p > 0$ and $q > 0$,

$$\log^p d = o(d^q), \quad d \rightarrow \infty.$$

In our case, $p = \frac{1}{2} > 0$, and $q = 1 - \beta - \frac{1}{2}(\frac{\gamma}{2} + \frac{\beta}{\gamma})^2 > 0$. The latter inequality is true in view of (1.23). From this, we get by (1.24) that $J_1 \rightarrow \infty$ as $d \rightarrow \infty$, and therefore J_1 is strictly positive for all sufficiently large d . Thus, putting everything together, we can write

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^{\pm}} E_{\eta} |\eta - \tilde{\eta}| > 0,$$

i.e., in the limit, the minimax risk is strictly positive and exact selection is impossible.

This completes the proof of Theorem 3. $\square$

Proof of Theorem 4. The proof is similar to that of Theorem 2. We essentially need to show that as $d \rightarrow \infty$

$$R(\tilde{\eta}) = \sup_{\eta \in H_{d,\beta}} E_{\eta} |\tilde{\eta} - \eta| = o(d^{1-\beta}).$$

As in Theorem 2, we can write (compare to (1.17))

$$\begin{aligned} R(\tilde{\eta}) &\leq d\Phi\left(-\sqrt{(2\beta + \delta)\log d}\right) + C_1 d^{1-\beta}\Phi\left(\sqrt{(2\beta + \delta)\log d} - a\right) \\ &=: L_1 + L_2. \end{aligned}$$

Estimating the first term from above by using (1.16), we have

$$\begin{aligned} L_1 &= d \cdot \Phi(-\sqrt{(2\beta + \delta)\log d}) \leq d \cdot e^{-((2\beta + \delta)\log d)/2} = d^{1-(2\beta + \delta)/2} \\ &= d^{1-\beta-\delta/2} = o(d^{1-\beta}). \end{aligned}$$

Consider the second term L_2 and note that condition (1.13) implies that for all sufficiently large d

$$a > \sqrt{(2\beta + \delta)\log d}.$$

Hence

$$\Phi\left(\sqrt{(2\beta + \delta)\log d} - a\right) \rightarrow 0.$$

and we get $L_2 = o(d^{1-\beta})$. Therefore

$$R(\tilde{\eta}) = L_1 + L_2 = o(d^{1-\beta}),$$

and the proof is complete. $\square$

Proof of Theorem 5. Acting as in the proof of Theorem 3, we can write (compare to relations (1.20) and (1.22))

$$\inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^{\pm}} E_{\eta} |\eta - \tilde{\eta}| = d^{1-\beta} \Phi(T - a) + d(1 - d^{-\beta}) \Phi(-T) =: M_1 + M_2, \quad (1.25)$$

where, assuming that the parameter a has the form $a = \gamma\sqrt{\log d}$, we can represent the

first term as

$$\begin{aligned} M_1 &= d^{1-\beta} \Phi(T - a) = d^{1-\beta} \Phi \left(\left(\frac{\beta}{\gamma} - \frac{\gamma}{2} \right) \sqrt{\log d} (1 + o(1)) \right) \\ &= d^{1-\beta} \Phi \left(\left(\frac{2\beta - \gamma^2}{2\gamma} \right) \sqrt{\log d} (1 + o(1)) \right). \end{aligned}$$

Note that by assumption $\gamma < \sqrt{2\beta}$. Therefore $(2\beta - \gamma^2)/2\gamma > 0$ and hence

$$\Phi \left(\left(\frac{2\beta - \gamma^2}{2\gamma} \right) \sqrt{\log d} (1 + o(1)) \right) \rightarrow 1, \quad d \rightarrow \infty.$$

This implies that $M_1 = O(d^{1-\beta})$. Since the other term M_2 is non-negative, it follows from (1.25) that

$$\liminf_{d \rightarrow \infty} \inf_{\tilde{\eta}} \sup_{\eta \in H_{d,\beta}^{\pm}} d^{\beta-1} E_{\eta} |\eta - \tilde{\eta}| > 0.$$

This completes the proof of Theorem 5. $\square$

Chapter 2

Estimation of the sparsity index β

Back to model (1.1) with $\eta \in H_{d,\beta}^\pm$, recall that β is the (unknown) sparsity index and that $d^{-\beta}$ is the proportion of the informative components of vector X . The conditions, under which the selection procedures in Theorems 2 and 4 work as designed, require the knowledge of β . Therefore, if one attempts to use the selectors $\hat{\eta}$ and $\tilde{\eta}$ as in (1.10) and (1.13) in applications, one needs either to adapt the selection procedures to unknown β or to estimate β consistently. Adaptation is typically done without estimating the unknown parameters involved, but the construction of optimal adaptive procedures is a challenging problem. In this thesis, we choose another strategy and estimate the parameter β consistently. In this section, we consider two estimators of β . One of them is essentially due to Meinshausen and Rice [8] who estimated the number of false null hypotheses among a very large number of independently tested hypotheses. The other estimator is a new estimator proposed in this thesis.

For the technical reason, when estimating β , we need to narrow down the set $\eta \in H_{d,\beta}^\pm$

by assuming that the vector $\eta = (\eta_1, \dots, \eta_d)$ in model (1.1) is such that

$$\sum_{j=1}^d \eta_j \sim d^{1-\beta}, \quad d \rightarrow \infty. \quad (2.1)$$

This essentially corresponds to assuming that η belongs to the set (compare to (1.3))

$$\left\{ \eta = (\eta_1, \dots, \eta_d) : \eta_j \in \{0, 1\}, \quad (1 - \tau)d^{1-\beta} \leq \sum_{j=1}^d \eta_j \leq (1 + \tau)d^{1-\beta} \right\}$$

with a positive $\tau = o(1)$ as $d \rightarrow \infty$. Note that the Meinshausen–Rice estimator discussed below was constructed under assumption (2.1).

2.1 Estimation of β by using the Meinshausen-Rice approach

Meinshausen and Rice [8] proposed a consistent estimator to estimate the proportion of false null hypotheses among a large number of independently tested hypotheses.

Let $F_0(t)$ denote the uniform $U(0, 1)$ df. In [8] Meinshausen and Rice considered the problem of testing the global null hypothesis

$$\bigcap_{i=1}^d \mathcal{H}_{0i} : U_1, \dots, U_d \stackrel{iid}{\sim} F_0(t),$$

where $\mathcal{H}_{0i} : U_i \sim F_0(t)$, $i = 1, \dots, d$, are independent null hypothesis, against the upper-tailed alternative

$$\mathcal{H}_{1,d} : U_1, \dots, U_d \stackrel{iid}{\sim} (1 - \varepsilon_d)F_0(t) + \varepsilon_d G(t) > F_0(t),$$

where $G \neq F_0$ is a df from some family of continuous dfs that includes F_0 , and $\varepsilon_d \sim d^{-\beta}$ with $\beta \in (0, 1)$. The problem of interest in [8] was as follows: to construct an estimator $\hat{\varepsilon}_d$ of ε_d with the property

$$P(\hat{\varepsilon}_d \leq \varepsilon_d) \geq 1 - \alpha \quad (2.2)$$

for a specified $\alpha \in (0, 1)$; then the global null hypothesis that there are no false null hypothesis can be tested at level α by rejecting when $\hat{\varepsilon}_d > 0$. The Meinshausen–Rice estimator of $\varepsilon_d \sim d^{-\beta}$ that satisfies (2.2) is given by (see Section 2 of [8] for details)

$$\hat{\varepsilon}_d = \hat{\varepsilon}_d(U_1, \dots, U_d) = \sup_{t \in (0, 1)} \frac{U_d(t) - t - \lambda_{d, \alpha} \delta(t)}{1 - t}, \quad (2.3)$$

where

$$U_d(t) = d^{-1} \sum_{i=1}^d \mathbb{1}(U_i \leq t), \quad \delta(t) = \sqrt{t(1-t)}, \quad t \in (0, 1),$$

and for a specified number $\alpha \in (0, 1)$

$$\begin{aligned} \lambda_{d, \alpha} &= A_d^{-1}(E^{-1}(1 - \alpha) + B_d), \quad E(x) = \exp(-\exp(-x)), \quad x \in \mathbb{R}, \\ A_d &= \sqrt{2d \log \log d}, \quad B_d = 2 \log \log d + \frac{1}{2} \log \log \log d - \frac{1}{2} \log(4\pi). \end{aligned}$$

Note that in [8] some other choices of the function $\delta(t)$ are also possible. The estimator $\hat{\varepsilon}_d$ measures, after suitable normalization, the closeness of the empirical df $U_d(t)$ to the null df $F_0(t)$.

To connect our problem of estimating β in model (1.1) to that in [8], we assume that there are many independent null hypothesis $H_{0i} : X_i \sim \Phi(x)$, $i = 1, \dots, d$, and we are

interested in rejecting the joint null hypothesis

$$\bigcap_{i=1}^d H_{0i} : X_1, \dots, X_d \stackrel{iid}{\sim} \Phi(x)$$

in favour of the lower-tailed alternative

$$H_{1,d} : X_1, \dots, X_d \stackrel{iid}{\sim} (1 - \varepsilon_d)\Phi(x) + \varepsilon_d\Phi(x - a_d) < \Phi(x),$$

where $\varepsilon_d \sim d^{-\beta}$ is the proportion of null hypothesis that are false, and

$$a_d = \gamma\sqrt{\log d}, \quad 0 < \gamma < \infty.$$

Note that under $H_{1,d}$ the data have a sparse component with non-zero mean at a_d .

In our case, we need to modify the above estimator of ε_d because the observations X_i in our model (1.1) are normal:

$$X_1, \dots, X_d \stackrel{iid}{\sim} (1 - \varepsilon_d)\Phi(x) + \varepsilon_d\Phi(x - a_d), \quad (2.4)$$

where $a_d = \gamma\sqrt{\log d}$ with $0 < \gamma < \infty$ and $\varepsilon_d \sim d^{-\beta}$ with $\beta \in (0, 1)$, and the alternative is lower-tailed. This is done by noting that, under the global null hypothesis $\bigcap_{i=1}^d \mathcal{H}_{0i}$, the transformed observations $1 - \Phi(X_i)$, $i = 1, \dots, d$, are iid uniform $U(0, 1)$. Then, in terms of the X_i s, choosing $\delta(t) = \sqrt{t(1-t)}$, the Meinshausen–Rice estimator of $\varepsilon_d \sim d^{-\beta}$ can be written as (see, for example, Section 6 of [2])

$$\hat{\varepsilon}_d = \hat{\varepsilon}_d(X_1, \dots, X_d) = \sup_{-\infty < x < \infty} \frac{\Phi(x) - F_d(x) - \lambda_{d,\alpha}\sqrt{\Phi(x)(1 - \Phi(x))}}{\Phi(x)} \quad (2.5)$$

where $F_d(x) = d^{-1} \sum_{i=1}^d \mathbb{I}(X_i \leq x)$ and $\lambda_{d,\alpha}$ is as in (2.3). In applications, instead of

the supremum over the whole real line the supremum over $(X_{(1)}, X_{(d)})$ is recommended, where $X_{(1)} = \min_{1 \leq i \leq d} X_i$ and $X_{(d)} = \max_{1 \leq i \leq d} X_i$ (see Section 4 of [8] for details).

The next result says that, under certain conditions on the underlying df, $\hat{\varepsilon}_d = \hat{\varepsilon}_d(U_1, \dots, U_d)$ is a consistent estimator of $\varepsilon_d \sim d^{-\beta}$ with $0 < \beta < 1/2$.

Theorem 6 (Corollary of Theorem 2 in [8]). *Consider the above mixture model $U_1, \dots, U_d \stackrel{iid}{\sim} (1 - \varepsilon_d)F_0(t) + \varepsilon_d G(t)$, where $a_d = \gamma\sqrt{\log d}$ with $0 < \gamma < \infty$ and $\varepsilon_d \sim d^{-\beta}$ with $\beta \in (0, 1/2)$. Let $G(t)$, $0 \leq t \leq 1$, be a continuous df and let $\inf_{t \in (0,1)} G'(t) = 0$. Let $\hat{\varepsilon}_d$ be the estimator of ε_d as in (2.3) with $\alpha = \alpha_d$ vanishing slowly enough, i.e., satisfying the condition $\limsup_{d \rightarrow \infty} (d/\log d)^{1/2} \lambda_{d, \alpha_d} < \infty$. Then*

$$\frac{\hat{\varepsilon}_d}{\varepsilon_d} \xrightarrow{P} 1, \quad d \rightarrow \infty.$$

In order to reduce the normal model (2.4) of interest to that in Theorem 6, we transform the random variables X_i by setting $U_i = 1 - \Phi(X_i)$, $i = 1, \dots, d$. The transformed model becomes

$$U_1, \dots, U_d \stackrel{iid}{\sim} (1 - \varepsilon_d)F_0(t) + \varepsilon_d (1 - \Phi(\Phi^{-1}(1 - t) - a_d)) > F_0(t).$$

Thus, for the transformed normal model, the df $G(t)$ in Theorem 6 is equal to

$$G(t) = 1 - \Phi(\Phi^{-1}(1 - t) - a_d), \quad a_d = \gamma\sqrt{\log d}, \quad 0 < \gamma < \infty, \quad 0 < t < 1,$$

and the conditions of Theorem 6 are fulfilled because

$$G'(t) = \frac{\varphi(\Phi^{-1}(1 - t) - a_d)}{\varphi(\Phi^{-1}(1 - t))},$$

and hence

$$\inf_{t \in (0,1)} G'(t) = \inf_{x \in \mathbb{R}} \frac{\varphi(x - a_d)}{\varphi(x)} = e^{-a_d^2/2} \inf_{x \in \mathbb{R}} e^{a_d x} = 0.$$

The case of $1/2 \leq \beta < 1$ is more delicate as in this case $\varepsilon_d \sim d^{-\beta}$ is very small when d is very large. Nevertheless, under certain conditions of the underlying df, consistent estimation of ε_d is possible, as follows from the following theorem.

Theorem 7 (Corollary of Theorem 3 in [8].) *Consider the normal mixture model $X_1, \dots, X_d \stackrel{iid}{\sim} (1 - \varepsilon_d)\Phi(x) + \varepsilon_d\Phi(x - a_d)$, where $a_d = \gamma\sqrt{\log d}$ with $0 < \gamma < \infty$ and $\varepsilon_d = d^{-\beta}$ with $\beta \in [1/2, 1)$. Let $\hat{\varepsilon}_d$ be the estimator of ε_d as in (2.5) with $\alpha = \alpha_d$ vanishing slowly enough, i.e., satisfying the condition $\limsup_{d \rightarrow \infty} (d/\log d)^{1/2} \lambda_{d, \alpha_d} < \infty$. Then*

$$\frac{\hat{\varepsilon}_d}{\varepsilon_d} \xrightarrow{P} 1, \quad d \rightarrow \infty.$$

Remark 3. Theorems 2 and 3 in [8] assume that $a_d = \gamma\sqrt{\log d}$ with $0 < \gamma < \sqrt{2}$. However, the inspection of their proofs shows that the results continue to hold for all $0 < \gamma < \infty$. This is one of the reason why we consider here the Meinshausen–Rice estimator and do not consider the estimator of ε_d proposed by Cai, Jin, and Low in [2]: the latter estimator of ε_d was shown to perform very well in the normal mixture model (2.4), where $a_d = \gamma\sqrt{\log d}$ with $0 < \gamma < \sqrt{2}$ (see Theorems 4.1 and 4.2 in [2]). However, the extension of these theorems to all $0 < \gamma < \infty$ is not easy, if at all possible.

Theorems 6 and 7 give a receipt for estimating β consistently. Indeed, we can estimate $\varepsilon_d \sim d^{-\beta}$ by means of the consistent estimator $\hat{\varepsilon}_d$ given by (2.5). Hence in order to get

an estimator of β , we need to solve the equation $\hat{\varepsilon}_d = d^{-\beta}$ for β :

$$\hat{\beta}_d^{\text{MR}} = \frac{\log(1/\hat{\varepsilon}_d)}{\log d}. \quad (2.6)$$

It is easy to see that $\hat{\beta}_d^{\text{MR}}$ is a consistent estimator of β . Indeed, for any $\epsilon > 0$, as $d \rightarrow \infty$,

$$\begin{aligned} P\left(|\hat{\beta}_d^{\text{MR}} - \beta| \geq \epsilon\right) &= P\left(\left|\frac{\log(1/\hat{\varepsilon}_d)}{\log d} - \beta\right| \geq \epsilon\right) \\ &= P(|\log(1/\hat{\varepsilon}_d) - \beta \log d| \geq \epsilon \log d) = P(|\log(\hat{\varepsilon}_d/d^{-\beta})| \geq \epsilon \log d) \\ &= 1 - P(-\epsilon \log d < \log(\hat{\varepsilon}_d/d^{-\beta}) < \epsilon \log d) \\ &= 1 - P(d^{-\epsilon} < \hat{\varepsilon}_d/d^{-\beta} < d^{\epsilon}) = o(1), \end{aligned} \quad (2.7)$$

where the last equality follows from Theorems 6 and 7, and the fact that $d^{-\epsilon} \rightarrow 0$ and $d^{\epsilon} \rightarrow \infty$ as $d \rightarrow \infty$.

The Meinshausen-Rice estimator $\hat{\beta}_d^{\text{MR}}$ can be used in both exact and almost full selection regimes. In the former regime, we can propose a better estimator of β . This estimator is introduced in Section 2.2 together with a study of its basic properties.

2.2 Estimation of β in the exact selection regime

Assume that the exact selection is possible, that is, that condition (1.10) on the parameter a in model (1.1) is fulfilled. By assumption, the total number of non-zero components of η is $d^{1-\beta}(1 + o(1))$. In order to estimate β consistently, consider a natural estimator of $\sum_j \eta_j \sim d^{1-\beta}$ given by

$$\hat{K}_d = \sum_{j=1}^d \mathbb{1}\left(X_j > \sqrt{(2 + \delta) \log d}\right)$$

where $\delta \in (0, 1)$ is the same as in (1.9). To obtain an estimator for β , we solve the equation $\hat{K}_d = d^{1-\beta}$ for β . We have

$$\log \hat{K}_d = (1 - \beta) \log d \quad \Rightarrow \quad \frac{\log \hat{K}_d}{\log d} = 1 - \beta \quad \Rightarrow \quad \beta = \frac{\log(d/\hat{K}_d)}{\log d}.$$

Thus, a natural estimator of β is

$$\hat{\beta}_d = \frac{\log(d/\hat{K}_d)}{\log d}. \quad (2.8)$$

The following result holds true.

Lemma 3. *Under the conditions of Theorem 2, assuming that (2.1) holds true, one has*

$$\hat{\beta}_d \xrightarrow{P} \beta, \quad d \rightarrow \infty,$$

that is, $\hat{\beta}_d$ is a consistent estimator of β .

Proof of Lemma 3. If we show that, as $d \rightarrow \infty$

$$\frac{E(\hat{K}_d)}{d^{1-\beta}} = 1 + o(1) \quad \text{and} \quad \text{Var}(\hat{K}_d) = o(d^{2-2\beta}), \quad (2.9)$$

then, by Markov's inequality (see Section 1.2), we obtain that for any $\epsilon > 0$, as $d \rightarrow \infty$,

$$\begin{aligned} P\left(\left|\frac{\hat{K}_d}{d^{1-\beta}} - 1\right| \geq \epsilon\right) &\leq \frac{E\left(\hat{K}_d/d^{1-\beta} - 1\right)^2}{\epsilon^2} \\ &= \frac{\text{Var}\left(\hat{K}_d/d^{1-\beta}\right) + \left(E\left(\hat{K}_d/d^{1-\beta}\right) - 1\right)^2}{\epsilon^2} = \frac{o(1)}{\epsilon^2} = o(1), \end{aligned}$$

or equivalently,

$$\frac{\hat{K}_d}{d^{1-\beta}} \xrightarrow{P} 1. \quad (2.10)$$

Then, using the arguments similar to those in (2.7), we obtain that

$$\hat{\beta}_d \xrightarrow{P} \beta, \quad d \rightarrow \infty.$$

Let us prove the relations in (2.9). First, recalling (2.1),

$$\begin{aligned} E(\hat{K}_d) &= \sum_{j=1}^d E \mathbb{1} \left(X_j > \sqrt{(2+\delta) \log d} \right) = \sum_{j=1}^d P \left(X_j > \sqrt{(2+\delta) \log d} \right) \\ &= \left(\sum_{j:\eta_j=0} + \sum_{j:\eta_j=1} \right) P \left(X_j > \sqrt{(2+\delta) \log d} \right) \\ &\sim (d - d^{1-\beta}) P_0 \left(X_j > \sqrt{(2+\delta) \log d} \right) + d^{1-\beta} P_a \left(X_j > \sqrt{(2+\delta) \log d} \right) \\ &=: D_1 + D_2. \end{aligned}$$

The first term satisfies

$$D_1 \leq d \cdot \exp \left(-\frac{2+\delta}{2} \log d \right) = O(d^{-\frac{\delta}{2}}) = o(1), \quad d \rightarrow \infty.$$

For the second term, denoting by Z a normal $N(0,1)$ random variable, we have

$$\begin{aligned} D_2 &= d^{1-\beta} \cdot P_a \left(X_j - a > \sqrt{(2+\delta) \log d} - a \right) = d^{1-\beta} P \left(Z > \sqrt{(2+\delta) \log d} - a \right) \\ &= d^{1-\beta} \left(1 - P \left(Z < \sqrt{(2+\delta) \log d} - a \right) \right) \\ &= d^{1-\beta} - d^{1-\beta} \cdot P \left(Z \leq - \left(a - \sqrt{(2+\delta) \log d} \right) \right), \end{aligned}$$

where by assumption

$$\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2}(1 + \sqrt{1 - \beta}).$$

From this, noting that $\left(a - \sqrt{(2 + \delta) \log d}\right) \rightarrow \infty$ as $d \rightarrow \infty$, we get

$$D_2 = d^{1-\beta}(1 + o(1)).$$

Summarizing,

$$E(\hat{K}_d) \sim D_1 + D_2 = o(1) + d^{1-\beta}(1 + o(1)) \sim d^{1-\beta}, \quad d \rightarrow \infty,$$

and the first relation in (2.9) is proved.

In order to prove the second relation in (2.9), recall that the optimal exact selector $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d)$ consists of the components $\hat{\eta}_j$ given by (see formula (1.9))

$$\hat{\eta}_j = \mathbb{1} \left(X_j > \sqrt{(2 + \delta) \log d} \right), \quad j = 1, \dots, d.$$

It is clear that $\hat{\eta}_j \sim \text{Bernoulli}(p_j)$ with

$$p_j = p_0 \quad \text{if} \quad \eta_j = 0, \quad \text{and} \quad p_j = p_a \quad \text{if} \quad \eta_j = 1,$$

where

$$\begin{aligned} p_0 &:= P_0 \left(X_j > \sqrt{(2 + \delta) \log d} \right) = P \left(N(0, 1) > \sqrt{(2 + \delta) \log d} \right), \\ p_a &:= P_a \left(X_j > \sqrt{(2 + \delta) \log d} \right) = P \left(N(a, 1) > \sqrt{(2 + \delta) \log d} \right). \end{aligned}$$

Note that the estimator $\hat{K}_d$ of $d^{1-\beta}$ can be written as

$$\hat{K}_d = \sum_{j:\eta_j=1} \hat{\eta}_j + \sum_{j:\eta_j=0} \hat{\eta}_j,$$

where $\sum_{j:\eta_j=1} \hat{\eta}_j$ and $\sum_{j:\eta_j=0} \hat{\eta}_j$ have a binomial $Bin(\sum_j \eta_j, p_a)$ distribution and a binomial $Bin(d - \sum_j \eta_j, p_0)$ distribution, respectively. Therefore, using the independence of the $\hat{\eta}_j$ s and condition (2.1), we obtain

$$\begin{aligned} \text{Var}(\hat{K}_d) &= \text{Var}\left(\sum_{j:\eta_j=0} \hat{\eta}_j\right) + \text{Var}\left(\sum_{j:\eta_j=1} \hat{\eta}_j\right) \\ &\sim (d - d^{1-\beta})p_0(1 - p_0) + d^{1-\beta}p_a(1 - p_a) \\ &=: J_1 + J_2, \end{aligned}$$

where

$$\begin{aligned} J_1 &= (d - d^{1-\beta})p_0(1 - p_0) \leq dp_0(1 - p_0) \\ &= d\Phi\left(\sqrt{(2 + \delta)\log d}\right)\Phi\left(-\sqrt{(2 + \delta)\log d}\right) \\ &\leq d\Phi\left(-\sqrt{(2 + \delta)\log d}\right), \end{aligned}$$

and the last inequality is due to the fact

$$\Phi\left(\sqrt{(2 + \delta)\log d}\right) \leq 1.$$

Since, by inequality (1.16),

$$\Phi\left(-\sqrt{(2 + \delta)\log d}\right) \leq \exp\left(-\frac{(2 + \delta)\log d}{2}\right) = O\left(d^{-(1+\frac{\delta}{2})}\right),$$

if follows that

$$J_1 = O(d^{-\frac{\delta}{2}}) = o(1), \quad d \rightarrow \infty.$$

Next, consider the term D_2 . Similarly, we have

$$\begin{aligned} D_2 &= d^{1-\beta} p_a(1 - p_a) = d\Phi\left(\sqrt{(2+\delta)\log d} - a\right)\Phi\left(a - \sqrt{(2+\delta)\log d}\right) \\ &\leq d\Phi\left(-(a - \sqrt{(2+\delta)\log d})\right) \leq d\exp\left(-\left(a - \sqrt{(2+\delta)\log d}\right)^2/2\right) = o(1), \end{aligned}$$

because by assumption $a > \sqrt{2}(1 + \sqrt{1-\beta})\sqrt{\log d}$, provided d is large enough.

Putting everything together, we obtain

$$\text{Var}(\hat{K}_d) \sim D_1 + D_2 = o(1),$$

and even a stronger version of the second relation in (2.9) is proved. The proof of Lemma 3 is complete. $\square$

In conclusion, as an explanation of our approach, note that if we can guarantee the accuracy of the exact selection, the proportion of non-zero components that we estimate will be close to such a proportion in the data.

2.3 Numerical comparison and conclusions

In Chapter 3 we shall demonstrate numerically the good performance of our optimal selectors $\hat{\eta}$ and $\tilde{\eta}$. Also, assuming that the exact selection is possible, we compare by simulation the Meinshausen and Rice approach with our approach to estimating β . As expected, with a larger mean a of the informative components of X and a higher dimension d , we can get a more accurate estimate of β with both methods. More importantly,

as follows from our simulation study, in the exact selection scenario, i.e., if the parameter a satisfies condition (1.10), the estimator $\hat{\beta}_d$ proposed in Section 2.2 is more accurate than the Meinshausen-Rice estimator $\hat{\beta}_d^{\text{MR}}$. For instance, taking $d = 1000$, $a = 6.69$, and $\beta = 0.4$ and conducting simulations, we obtained that our estimate of β is equal to 0.4003, whereas the Meinshausen-Rice estimate is equal to 0.4031.

In the almost full selection regime, we cannot in general use $\hat{\beta}_d$ to estimate β , as we cannot guarantee that condition (1.13) is satisfied, in which case we cannot guarantee the consistency of $\hat{\beta}_d$ as an estimator of β . More numerical results will be presented in Chapter 3.

Chapter 3

Simulation studies

3.1 Exact and almost full selection

Recall that in Chapter 2 for the purpose of estimating β we simplified the initial setup, according to which $\eta \in H_{d,\beta}^\pm$ (or $\eta \in H_{d,\beta}$), and assumed that the vector η is such that $\sum_{j=1}^d \eta_j \sim d^{1-\beta}$ (see formula (2.1)). In order to construct the exact and almost full selectors and compute their risks, we shall use the following algorithm.

1. Choose a large value of d (such as 1000, 10000, ...), and a value of $\beta \in (0, 1)$ to make sure that $d^{-\beta}$ is small (< 0.1), and set the value of a to satisfy condition (1.10) and condition (1.13) for the exact and almost full regimes, respectively.
2. Generate a d -dimensional vector $(x_1, \dots, x_d)$ from a normal mixture distribution with $k = [d^{1-\beta}]$ normal $N(0, 1)$ components and $(d-k)$ normal $N(a, 1)$ components, and record the vector $\eta = (\eta_1, \dots, \eta_d)$, where $\eta_j = 0$ if x_j is normally $N(0, 1)$ distributed, and $\eta_j = 1$ if x_j is normally $N(a, 1)$ distributed.
3. Use the data obtained in step 2 to compute the selectors $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d)$ and $\tilde{\eta} = (\tilde{\eta}_1, \dots, \tilde{\eta}_d)$, where $\hat{\eta}_j = \mathbb{1}(x_j > \sqrt{(2 + \delta_1) \log d})$, $\tilde{\eta}_j = \mathbb{1}(x_j > \sqrt{(2\beta + \delta_2) \log d})$,

and δ_1 and δ_2 are small positive numbers.

4. Make comparisons between the original vectors and the selector vectors by calculating the Hamming distances $|\hat{\eta} - \eta| = \sum_{j=1}^d |\hat{\eta}_j - \eta_j|$ and $|\tilde{\eta} - \eta| = \sum_{j=1}^d |\tilde{\eta}_j - \eta_j|$.
5. Choose a large positive number M and repeat steps 2 to 4 M times. Find the average (over M replications) Hamming distances for both exact and almost full selectors; this gives the estimates of $E_\eta|\eta - \hat{\eta}|$ and $E_\eta|\eta - \tilde{\eta}|$.
6. Use the quantities obtained in step 5 to estimate the maximum risks

$$R(\hat{\eta}) = \sup_{\eta: \sum_j \eta_j = k} E_\eta|\eta - \hat{\eta}| \quad \text{and} \quad R(\tilde{\eta}) = \sup_{\eta: \sum_j \eta_j = k} E_\eta|\eta - \tilde{\eta}|.$$

As a comment on step 6, we observe that, by the symmetry of the Hamming distance, each of the $\binom{d}{k}$ combinations of $(d - k)$ zero mean normal random variables and k non-zero mean normal random variables generated in step 2 will produce the same values of $E_\eta|\eta - \hat{\eta}|$ and $E_\eta|\eta - \tilde{\eta}|$. Hence, the values obtained in step 5 should be taken as the estimates of $R(\hat{\eta})$ and $R(\tilde{\eta})$.

The results obtained by using this algorithm are presented in Tables 3.1 and 3.2 below.

3.2 Estimation of the sparsity index β

In the exact selection scenario, estimation of the parameter β is simple. Once we obtain the selector $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_d)$, the total number of its components $\hat{K}_d = \sum_{j=1}^d \hat{\eta}_j$ gives an estimator of $d^{1-\beta}$. The corresponding estimator of β is (see (2.8))

$$\hat{\beta}_d = \frac{\log(d/\hat{K}_d)}{\log d} = 1 - \frac{\log \hat{K}_d}{\log d},$$

as discussed in the previous chapter.

In order to compute the Meinshausen–Rice estimator $\hat{\beta}_d^{\text{MR}}$ given by (2.6), we shall use the algorithm below. Recall that the Meinshausen–Rice estimator works for both exact selection and almost full selection regimes.

- I. Choose a large value of d (such as 1000, 10000, ...), and a value of $\beta \in (0, 1)$ to make sure that $d^{-\beta}$ is small (< 0.1), and set the value of a to satisfy condition (1.10) or condition (1.13). Generate a d -dimensional vector $x = (x_1, \dots, x_d)$ from a normal mixture distribution with $k = \lceil d^{1-\beta} \rceil$ normal $N(0, 1)$ components and $(d - k)$ normal $N(a, 1)$ components.
- II. Find the minimum component $x_{(1)}$ and the maximum component $x_{(d)}$ of the vector x . Make a uniform grid over the interval $[x_{(1)}, x_{(d)}]$ with the distance between the grid points u_i and u_{i+1} equal to a specified small number (such as 0.001).
- III. Using formula (2.5), calculate

$$\hat{\varepsilon}_d^{(i)} = \frac{\Phi(u_i) - F_d(u_i) - \lambda_{d,\alpha} \sqrt{\Phi(u_i)(1 - \Phi(u_i))}}{\Phi(u_i)},$$

and find $\hat{\varepsilon}_d := \max_i \hat{\varepsilon}_d^{(i)}$. Then calculate the quantity

$$\hat{\beta}_d = \frac{\log(1/\hat{\varepsilon}_d)}{\log d}.$$

- IV. Choose a large positive number M and repeat steps I to III M times. Find the average value of $\hat{\beta}_d$ over M replications, and use it as an estimate of β .

The results obtained by using this algorithm are presented in Tables 3.1 and 3.2 below.

3.3 Numerical results

To illustrate how good the proposed selectors and estimators work, we apply the algorithms from Sections 3.1 and 3.2 with several choices of d (e.g., 1000, 10000, 100000) and several choices of β (e.g., 0.4, 0.5, 0.6). Table 3.1 contains the values of the maximum risk $R(\hat{\eta})$ of the exact selector $\hat{\eta}$, and the estimates of β obtained by using our estimator $\hat{\beta}_d$ and the Meinshausen–Rice estimator $\hat{\beta}_d^{\text{MR}}$. Table 3.2 contains the values of the maximum risk $R(\tilde{\eta})$ and the normalized maximum risk $d^{\beta-1}R(\tilde{\eta})$ of the almost full selector $\tilde{\eta}$ and the estimates of β obtained by using the Meinshausen–Rice estimator $\hat{\beta}_d^{\text{MR}}$. By ‘Ratio’ we mean the quantity $d^{-\beta}$, which is the ratio of the informative components to all d components of the observed normal vector. By ‘Condition’ we mean the condition on the parameter a from Theorem 2 (in Table 3.1) and from Theorem 4 (in Table 3.2).

Table 3.1: Selection and estimation in the exact selection regime

(d, β)	Ratio	Condition	Chosen a	$R(\hat{\eta})$	$\hat{\beta}_d$	$\hat{\beta}_d^{\text{MR}}$
(1000,0.4)	0.0631	$a > 6.5960$	6.70	0.2342	0.4003	0.4031
(1000,0.5)	0.0316	$a > 6.3452$	6.50	0.2854	0.4982	0.4973
(1000,0.6)	0.0158	$a > 6.0677$	6.30	0.2363	0.5989	0.5988
(10000,0.4)	0.0251	$a > 7.6165$	7.72	0.2530	0.4001	0.4002
(10000,0.5)	0.0100	$a > 7.3268$	7.62	0.1941	0.5000	0.5001
(10000,0.6)	0.0030	$a > 7.0064$	7.34	0.2633	0.5994	0.5996
(100000,0.4)	0.0100	$a > 8.5154$	8.65	0.2618	0.4000	0.4000
(100000,0.5)	0.0032	$a > 8.1916$	8.45	0.2314	0.5000	0.5003
(100000,0.6)	0.0010	$a > 7.8334$	8.23	0.1622	0.6000	0.6001

As seen from Table 3.2, with the chosen values of d and β , the value of $d^{\beta-1}R(\tilde{\eta})$ ranges between 0.099 to 0.261, depending on the value of a . Compared to the exact selection regime, in the almost full selection regime, the values of a were chosen farther from the lower bound on a obtained from the condition $\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2\beta}$. In the exact selection regime, we can choose the value of a just slightly above the lower bound

Table 3.2: Selection and estimation in the almost full selection regime

(d, β)	Ratio	Condition	Chosen a	$R(\tilde{\eta})$	$d^{\beta-1}R(\tilde{\eta})$	$\hat{\beta}_d^{\text{MR}}$
(1000,0.4)	0.0631	$a > 2.3508$	4.47	16.6960	0.2646	0.4139
(1000,0.5)	0.0316	$a > 2.6283$	4.78	6.0321	0.1909	0.5099
(1000,0.6)	0.0159	$a > 2.8791$	4.98	2.1632	0.1361	0.6121
(10000,0.4)	0.0251	$a > 2.7145$	5.22	46.8310	0.1866	0.4074
(10000,0.5)	0.0100	$a > 3.0349$	5.33	16.2601	0.1626	0.5076
(10000,0.6)	0.0040	$a > 3.3245$	5.62	4.0121	0.1337	0.6063
(100000,0.4)	0.0100	$a > 3.0349$	5.61	216.3438	0.2163	0.4030
(100000,0.5)	0.0032	$a > 3.3931$	5.85	48.4632	0.1514	0.5052
(100000,0.6)	0.0010	$a > 3.7169$	6.11	9.8814	0.0988	0.6057

on a , obtained from the condition $\liminf_{d \rightarrow \infty} \frac{a}{\sqrt{\log d}} > \sqrt{2}(1 + \sqrt{1 - \beta})$, to have the maximum risk $R(\hat{\eta})$ below a certain level, e.g., below 0.3. In theory, we assume that the dimension d grows indefinitely. However, when conducting simulations, we have obvious limitations on choosing the values of d (the largest d we choose is 100000). As a result, in order to retain the risk at a low level, we need to choose the parameter δ in formula (1.12) large enough.

In the exact selection regime, when estimating the sparsity index β by the method proposed in Section 2.2, we obtain the estimates that are very close to the estimated parameter β . Note in passing that as the value of a gets farther from the lower bound $\sqrt{2}(1 + \sqrt{1 - \beta})\sqrt{\log d}$ the accuracy improves. In the almost full selection regime, we use the Meinshausen–Rice estimator $\hat{\beta}_d^{\text{MR}}$ to estimate β . Again, the accuracy improves as a gets farther from the lower bound $\sqrt{2\beta \log d}$.

Notice that the exact and almost full selection procedures based on $\hat{\eta}$ and $\tilde{\eta}$ are thresholding in nature, i.e., each procedure selects the components that are larger than the specified threshold level to be the informative components. The following histograms give us a visual representation of how the selectors $\hat{\eta}$ and $\tilde{\eta}$ work for several choices of the values of d and β . In both exact and almost full selection regimes, a vertical line represents

the selection threshold and shows how the selectors $\hat{\eta}$ and $\tilde{\eta}$ divide the normal mixture data into two groups, the noisy $N(0, 1)$ variables the informative $N(a, 1)$ variables.

Figure 3.1: Six histograms of the simulated normal mixture data in the exact and almost selection regimes. The vertical bars represent the threshold levels of the exact and almost full optimal selectors.

Conclusion

In this thesis, we discussed in detail two variable selection strategies in case of observing a single high-dimensional sparse normal vector. The difficulty of the problem is due to high dimensionality. In addition, as the goal was to construct *optimal* selection procedures, we had to obtain the lower bounds on the minimax risk (see Theorems 3 and 5), which is always a more difficult task as compared to obtaining the upper bounds on the maximum risk (see Theorems 2 and 4).

In Chapter 1 we determined the selection boundaries that divide the ‘phase space’, i.e., the β - r plane, into three regions: (1) the region where exact selection is possible, (2) the region where almost full selection is possible, and (3) the region where no selection is possible (see Figure 1.1). For the first two regions, we constructed the selection procedures and showed that they are optimal (in the asymptotically minimax sense).

In order to use the proposed selection procedures in practice, one needs to know the value of the sparsity index β . This is especially crucial for the almost full selection procedure based on the selector $\tilde{\eta}$ given by (1.12) as it has β in its definition. Therefore another problem addressed in the thesis is that of estimating the parameter β . In this connection, in Chapter 2 we considered two consistent estimators of β , one of which is due to Meinshausen and Rice [8], and the other is a new one suggested in this thesis. For both these estimators, we examined their basic asymptotic properties.

In Chapter 3 we presented the results of our simulation studies that confirm good

theoretical properties of the selectors and estimators under consideration. In particular, we found numerically that, in the region where exact selection is possible, our estimator of β is better as compared to that of Meinshausen and Rice.

We wish to stress that the results obtained in Chapters 1 and 2 are theoretical in nature. Therefore their implementation in practice is not straightforward. Numerically, due to the limitation of the device, we can hardly simulate the data with dimensions more than millions, so that the simulation results have some deficiency from the theory, especially in the almost full selection case with a relatively lower dimension (such as $d = 1000$). In that case, we need to take the number δ in the definition of the selector $\tilde{\eta}$ given by (1.12) large enough to reduce its maximum risk to a certain level; however, as the dimension d increases, the simulation results show that the risk reduces. In the exact selection case, the maximum risk is small even when the dimension d is not very high. This is explained by the fact that when β is not close to one, the ‘exact selection region’ does not merge the ‘no selection region’, where nothing can be done.

The work done in this thesis can be extended in several ways. For instance, instead of a single high-dimensional normal vector, we may consider a sample of n high-dimensional not necessarily normal vectors that have (nearly) the same tail behaviour as normal vectors, and study a similar selection problem when n relates to d in one or another way. Another possible direction of future studies is to adapt the almost full selection procedure based on the selector $\tilde{\eta}$ as in (1.12) to an unknown parameter β .

Appendix A

Proof of Lemma 1. Using the symmetry of the ξ_j 's and relation (1.15), we get

$$\begin{aligned}
P\left(\max_{1 \leq j \leq d} |\xi_j| \geq \sqrt{2 \log d}\right) &\leq P\left(\bigcup_{j=1}^d \{|\xi_j| \geq \sqrt{2 \log d}\}\right) \leq \sum_{j=1}^d P\left(|\xi_j| \geq \sqrt{2 \log d}\right) \\
&= dP\left(|\xi_j| \geq \sqrt{2 \log d}\right) = 2dP\left(\xi_j \geq \sqrt{2 \log d}\right) \\
&= 2d\Phi\left(-\sqrt{2 \log d}\right) \sim \frac{2}{\sqrt{2\pi}\sqrt{2 \log d}} d e^{-\log d} \\
&= O\left(\frac{1}{\sqrt{\log d}}\right) = o(1).
\end{aligned}$$

That is, we have shown that

$$P\left(\max_{1 \leq j \leq d} |\xi_j| \geq \sqrt{2 \log d}\right) \rightarrow 0, \quad d \rightarrow \infty,$$

and Lemma 1 is proved. $\square$

Proof of Lemma 2. We need to show that for any $0 < c_0 < 1 < c_1 < \infty$

$$P\left(\sum_{j=1}^d \eta_j \in [c_0 d^{1-\beta}, c_1 d^{1-\beta}]\right) \geq 1 - o(d^{-1}),$$

where $\eta = (\eta_1, \dots, \eta_d)$ has the prior distribution π given by (1.19). To this end, we shall use the Bernstein inequality, which says that if $Y_1, \dots, Y_n$ are iid random variable such that

$$E(Y_i) = 0, \quad E(Y_i^2) = \sigma^2, \quad i = 1, \dots, n,$$

$$|Y_i| \leq M \text{ for some } M > 0,$$

then for any $y > 0$

$$P\left(\left|\sum_{i=1}^n Y_i\right| \geq y\right) \leq 2 \exp\left(-\frac{y^2}{2(\sigma^2 n + \frac{My}{3})}\right). \quad (3.1)$$

Here, we apply this inequality to the random variables $Y_j = \eta_j - p$, $j = 1, \dots, d$, with $p = d^{-\beta}$. We can do that because

$$\mathbb{E}(Y_j) = 0, \quad \mathbb{E}(Y_j^2) = \text{Var}(\eta_j) = p(1 - p).$$

$$|Y_j| \leq |\eta_j| + |p| \leq 1 + p \leq 2.$$

Let $0 < c_0 < 1 < c_1$, in our case,

$$\begin{aligned} P\left(\sum_{j=1}^d \eta_j \in [c_0 d^{1-\beta}, c_1 d^{1-\beta}]\right) &= P\left(c_0 d^{1-\beta} - d^{1-\beta} \leq \sum_{j=1}^d \eta_j - d^{1-\beta} \leq c_1 d^{1-\beta} - d^{1-\beta}\right) \\ &= P\left(c_0 d^{1-\beta} - d^{1-\beta} \leq \sum_{j=1}^d (\eta_j - p) \leq c_1 d^{1-\beta} - d^{1-\beta}\right) \\ &= P\left((c_0 - 1)d^{1-\beta} \leq \sum_{j=1}^d (\eta_j - p) \leq (c_1 - 1)d^{1-\beta}\right). \end{aligned}$$

Denoting $\tau = \min(c_1 - 1, 1 - c_0)$, we get

$$\begin{aligned} P\left(\sum_{j=1}^d \eta_j \in [c_0 d^{1-\beta}, c_1 d^{1-\beta}]\right) &\geq P\left(-\tau d^{1-\beta} \leq \sum_{j=1}^d (\eta_j - p) \leq \tau d^{1-\beta}\right) \\ &= 1 - P\left(\left|\sum_{j=1}^d (\eta_j - p)\right| \geq \tau d^{1-\beta}\right). \end{aligned}$$

Applying the Bernstein inequality to the right-hand side, we obtain

$$\begin{aligned}
1 - P\left(\left|\sum_{j=1}^d(\eta_j - p)\right| \geq \tau d^{1-\beta}\right) &\geq 1 - 2 \exp\left(\frac{-(\tau d^{1-\beta})^2}{2p(1-p)d + \frac{4\tau d^{1-\beta}}{3}}\right) \\
&= 1 - \exp\left(\frac{-\tau^2 d^{2-2\beta}}{2d^{1-\beta} - 2d^{1-2\beta} + \frac{4\tau d^{1-\beta}}{3}}\right) \\
&= 1 - \exp\left(\frac{-\tau^2 d}{2d^\beta - 2 + \frac{4\tau d^\beta}{3}}\right) \\
&= 1 - \exp(-Cd^{1-\beta}),
\end{aligned}$$

with some positive constant C . From this

$$P\left(\sum_{j=1}^d \eta_j \in [c_0 d^{1-\beta}, c_1 d^{1-\beta}]\right) \geq 1 - \exp(-Cd^{1-\beta}) = 1 - o(d^{-1}).$$

The proof of Lemma 2 is complete. $\square$

Appendix B Simulation R Codes

.

1. Exact selection.

This code is used for computing the maximum risk of the exact selector $\hat{\eta}$.

```
exact=function(d,beta,delta,a,m)
{ n=d^(1-beta)
r=d^(-beta)
y=sample(x=1:d, size=round(n,digits=0), replace=FALSE, prob=NULL)
eta=rep(0,d)
eta[y]=1
R={}
for (i in 1:m)
{ e=rnorm(d,mean=0,sd=1)
x=e+eta*a
q=sqrt((2+delta)*log(d))
etahat=(x>q)*1
R=c(abs(etahat-eta),R) }
risk=sum(R)/m
return(risk) }
```

2. Almost full selection.

This code is used for computing the maximum risk of the almost full selector $\tilde{\eta}$.

```
almost=function(d,beta,delta,a,m)
{ n=d^(1-beta)
r=d^(-beta)
```

```

q=sqrt((2*beta+delta)*log(d))
y=sample(x=1:d, size=round(n,digits=0), replace=FALSE,prob=NULL)
eta=rep(0,d)
eta[y]=1
e=rnorm(d,mean=0,sd=1)
x=e+eta*a
R={}
for (i in 1:m)
{ e=rnorm(d,mean=0,sd=1)
x=e+eta*a
q=sqrt((2+delta)*log(d))
etahat=(x>q)*1
R=c(abs(etahat-eta),R)
}
risk=sum(R)/m
return(risk) }

```

3. Beta estimation.

This code is used for computing the estimate $\hat{\beta}_d^{\text{MR}}$ of the sparsity index β .

```

Ratio=function(d,beta,a)
{ n=d^(1-beta)
y=sample(x=1:d, size=round(n,digits=0), replace=FALSE,prob=NULL)
eta=rep(0,d)
eta[y]=1
e=rnorm(d,mean=0,sd=1)

```



```

x=e+eta*a
ratio={};
for(i in seq(min(x), max(x), 0.001))
{ y=pnorm(i)
deltad=y*(1-y)
Fn=ecdf(x)
z=Fn(i)
ad=sqrt(2*d*log(log(d)))
bd=2*log(log(d))+0.5*log(log(log(d)))-0.5*log(4*pi)
betad=ad^(-1)*(-log(log(1/(1-0.005))))+bd)
ratio=c((y-Fn(i)-sqrt(y*(1-y))*(betad))/y,ratio) }
rhat=max(ratio)
return(ratio) }
ratio1=Ratio(1000,0.4,8.45)
r=max(ratio1)
betahat=-log(r)/log(1000)
betahat

```

Bibliography

- [1] Borovkov, A. A. (2003). *Probability Theory* (4th ed.). Editorial URSS, Moscow.
- [2] Cai, T., Jin, J., and Low, M. (2006). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, **35**, 2421–2449. Dept. Statistics, The Wharton School, Univ. Pennsylvania.
- [3] Casella, G. and Berger, R. (2002). *Statistical Inference* (2nd ed.). Springer, New York.
- [4] Donoho, D. (2000). The curses and blessings of dimensionality. Stanford University. <http://statweb.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html>
- [5] Genovese, C. R., Jin, J., and Wasserman, L. (2009). Revisiting marginal regression. Carnegie Mellon University. <http://arxiv.org/pdf/0911.4080.pdf>
- [6] Ingster Yu. I., Pouet, C., and Tsybakov A. B. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A*, **367**, 4427–4448.
- [7] Jun, S. (2005). *Mathematical Statistics*. Springer, New York.
- [8] 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.

- [9] Roussas, G. (1973). *A First Course in Mathematical Statistics* (2nd ed.). Addison-Wesley.