Sparse Signal Recovery with Subbotin Noise

by

Joshua Miller

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Mathematics with Concentration in Statistics

Carleton University
Ottawa, Ontario

Abstract

In this thesis, we expand upon the work of Cui [4] and earlier authors by studying the problem of variable selection in a high-dimensional setting. Specifically, we consider a single vector $\boldsymbol{X}$ from a sequence model with noise components originating from a generalized normal distribution and determine the regions of exact and almost full selection with respect to a Hamming loss function. We make the routine assumption that the model studied is *sparse*. That is, the informative number of components is tiny in the model relative to the dimension d. An adaptive procedure is also proposed for estimating the signal components of $\boldsymbol{X}$ when the level of sparsity is unknown. A synthetic simulation and empirical study is then presented to showcase the aforementioned results. Lastly, we conclude the thesis by proposing areas of future work.

Acknowledgments

I would like to extend a huge thank you to my thesis supervisor, Professor Natalia Stepanova, for her unwavering support and encouragement these past few years. She has been a great help to me in both my undergraduate and graduate studies. I am truly grateful to have been her student. Without her, I would not be the statistician I am today.

I also feel very lucky to have such a loving and supportive family. My mother, father, and brother have always motivated me to go further in my academic journey. I am deeply thankful for each of them.

If I have seen further it is by standing on the shoulders of giants. - Isaac Newton (1675)

Contents

List of Tables

List of Figures

List of Notation

$a_n = o(b_n)$:	$\lim_{n\to\infty} a_n/b_n = 0$
$a_n = O(b_n)$:	$\lim_{n\to\infty} a_n/b_n = C$, where $C \neq 0$
$a_n \ll b_n$:	$a_n = o(b_n)$
$a_n \gg b_n$:	$b_n = o(a_n)$
$a_n \sim b_n$:	$\lim_{n\to\infty} a_n/b_n = 1$
$\limsup a_n$:	$\limsup a_n = \lim_{n\to\infty} \left(\sup_{m\geq n} a_m \right)$
$\liminf a_n$:	$\liminf a_n = \lim_{n\to\infty} \left(\inf_{m\geq n} a_m \right)$
$\mathcal{N}(\mu, \sigma^2)$:	A normal distribution with mean $\mu \in \mathbb{R}$ and variance σ^2, $\sigma \in \mathbb{R}^+$
$\mathcal{N}_k(\mu, \Sigma)$:	A k-dimensional multivariate normal distribution with mean $\mu \in \mathbb{R}^k$ and covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$
$\mathrm{GN}_\gamma(\mu)$:	A generalized normal distribution with mean $\mu \in \mathbb{R}$ and shape parameter $\gamma \in \mathbb{R}_+$
$\mathrm{GN}_{\lambda,\gamma}(\mu)$:	A generalized normal distribution with mean $\mu \in \mathbb{R}$, shape parameter $\gamma \in \mathbb{R}_+$, and scale parameter $\lambda \in \mathbb{R}_+$
Bernoulli(p) :	A Bernoulli distribution with mean p
Binomial(n, p) :	A binomial distribution with parameters n and p
$X_n \xrightarrow{P} \mu$:	Convergence in probability of a sequence of random variables $\{X_n\}_{n\geq 1}$ to the constant μ
$\mathbb{R}$:	The set of real numbers
$\mathbb{R}_+$:	The set of positive real numbers
$\mathbb{E}(X)$:	The expected value of a random variable X
$\mathbb{E}(X\|Y)$:	The conditional expected value of X given Y
$\|\boldsymbol{V}\|$:	The ℓ_1-norm of a vector $\boldsymbol{V} \in \mathbb{R}^d$

Chapter 1

Introduction

Prior to the advent of scientific computing technology, the application of statistical methods to real world data was oftentimes limited by the lack of available hardware. Thankfully, the rapid evolution of computing power within living memory has made it feasible to study statistical problems where the number of computations involved is exceedingly large. A relatively new field that involves a large amount of computations is that of high-dimensional statistics. In high-dimensional statistics, the aim is to conduct inference given that the dimensionality of the data vastly surpasses the number of observations studied. That is, one considers a collection of random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, where $\boldsymbol{X}_i \in \mathbb{R}^d, n \ll d$, and the goal is to perform inference. The scope of high-dimensional statistics is vast, and it has found application in the fields of engineering, the natural sciences, artificial intelligence, and the social sciences. As a motivating example, similar to that of the cancer patient example in the Introduction of Hastie et al. [8], imagine a population of individuals for which a small number of them suffer from a rare genetic disease. For each individual, some of their genetic information is stored. There is believed to be between 20,000 and 25,000 protein-coding genes contained in human chromosomes [10]. Thus, a dataset storing genetic

information would naturally be of a very high dimension. In this setup, it can be imagined that the vast majority of human genes are not related to the presence of the disease, except for some tiny proportion. The aim could be to "select" those genes that are significantly related to the presence of this disease.

In this thesis, we study the following model. Consider a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ of the form

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

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d), \varepsilon_1, \ldots, \varepsilon_d \overset{\text{i.i.d}}{\sim} F_\epsilon, F_\epsilon$ is the distribution function of some zero mean random variable ϵ, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$, $\eta_j \in \{0, 1\}, j = 1, \ldots, d$, and $\mu = \mu_d \in \mathbb{R}_+$ is some constant which is referred to as the *signal*. In model (1.1), the components of $\boldsymbol{\eta}$ can be interpreted as activating the deterministic signal μ. For $\eta_i = 1$, the component X_i is equal to the signal μ perturbed by random noise ε_i. For $\eta_i = 0$, the random variable X_i is merely random noise ε_i. The recovery of $\boldsymbol{\eta}$ in model (1.1) will be referred to as the *variable selection problem*. An estimator of $\boldsymbol{\eta}$, denoted by $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\boldsymbol{X}) \in \{0, 1\}^d$, will be referred to as a *selector*. We refer to the estimator $\hat{\boldsymbol{\eta}}$ as a selector because it attempts to "select" the variables from $\boldsymbol{X}$ for which there is a signal. The problem shall be studied in the asymptotic case, namely, as d tends to infinity. To have a meaningful selection problem, we will assume that $\mu = \mu_d \to \infty$ as $d \to \infty$.

Remark 1: Model (1.1) corresponds to a single observation of a random vector $\boldsymbol{X}$ of dimension d. Models of this type, with normally distributed noise variables, as well as their modifications written as mixtures of signal and noise distributions, have been previously assumed in a number of statistical studies dealing with estimation, hypothesis testing, and variable selection problems in a high-dimensional setting (see, e.g., Butucea et al. [3] and references therein).

To examine the variable selection problem in a high dimension (when d is very large), it is typically assumed that the vector $\boldsymbol{\eta}$ has a sparse structure. Formalizing this notion, we introduce the parameter $\beta \in (0,1)$ which shall be referred to as the *sparsity index*, along with the sets

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

with $1 < C_1 < \infty$ and

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

where $0 < C_0 < 1 < C_1 < \infty$. The set $H_{d,\beta}^{\pm}$ was first presented in Ingster et al. [9] to study the classification of sparse high-dimensional random vectors. Roughly speaking, if it is assumed that $\boldsymbol{\eta}$ belongs to either $H_{d,\beta}$ or $H_{d,\beta}^{\pm}$, then the number of components of $\boldsymbol{X}$ with signal μ_d will grow at a rate $d^{1-\beta}$ as d increases. The parameter β is named the sparsity index because it fixes the maximum number of components of $\boldsymbol{X}$ that contain the signal μ_d. If β is close to 1, it is referred to as the *highly sparse case*. Conversely, if β is close to 0, one refers to this as the *dense case*. If $\boldsymbol{\eta}$ belongs to either $H_{d,\beta}$ or $H_{d,\beta}^{\pm}$, then it is obvious that $\sum_{i=1}^{d} \eta_i$ is $o(d)$ as d approaches infinity, which implies that the proportion of "signal" components of $\boldsymbol{X}$ will shrink to 0 relative to d. This is why elements of $H_{d,\beta}$ and $H_{d,\beta}^{\pm}$ are thought of as sparse. Unless otherwise stated, it will be assumed in this thesis that the parameter $\boldsymbol{\eta}$ in model (1.1) belongs to $H_{d,\beta}$.

To quantify the performance of a selector, we must also equip ourselves with a notion of its risk, or expected loss. To this end, the *Hamming risk* of a selector $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ is defined as

$$\mathbb{E}_{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}| = \mathbb{E}_{\boldsymbol{\eta}} \left(\sum_{j=1}^{d} |\hat{\eta}_j - \eta_j| \right). \tag{1.2}$$

The quantity in (1.2) corresponds to the expected number of components for which the selector $\hat{\boldsymbol{\eta}}$ is not in agreement with $\boldsymbol{\eta}$. To guard oneself against the worst possible case, it is useful to analyse the maximum Hamming risk of the selector $\hat{\boldsymbol{\eta}}$, namely

$$R_d(\hat{\boldsymbol{\eta}}) = \sup_{\boldsymbol{\eta} \in H} \mathbb{E}_{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}|, \tag{1.3}$$

where H is some parameter space for $\boldsymbol{\eta}$, say, $H = H_{d,\beta}$. A "good" selector should be able to correctly identify the signal components of $\boldsymbol{X}$ as d grows to infinity. With this in mind, we define an *exact selector* $\hat{\boldsymbol{\eta}}$ to be such that

$$\limsup_{d \to \infty} R_d(\hat{\boldsymbol{\eta}}) = 0. \tag{1.4}$$

If there exists a selector $\hat{\boldsymbol{\eta}}$ satisfying (1.4), we say that *exact selection* is possible. Intuitively, if a selector is exact, then as d tends to infinity, it is nearly guaranteed that one can correctly determine all the components of the vector $\boldsymbol{\eta}$ from model (1.1). Unfortunately, as will be demonstrated later on, it is not always possible to construct an exact selector if conditions are not ideal. When exact selection is impossible, an almost full selector is sought after instead. We define an *almost full selector* $\hat{\boldsymbol{\eta}}$ to be a selector which satisfies

$$\limsup_{d \to \infty} d^{\beta - 1} R_d(\hat{\boldsymbol{\eta}}) = 0. \tag{1.5}$$

If there exists a selector $\hat{\boldsymbol{\eta}}$ satisfying (1.5), we say that *almost full selection* is possible. Intuitively, an almost full selector cannot consistently classify all compo-

nents of $\boldsymbol{\eta}$ as d grows, but the number of incorrect selections is tiny compared to the number of signal components, which is of order $d^{1-\beta}$. This is since the definition of an almost full selector is equivalent to the condition $R_d(\hat{\boldsymbol{\eta}}) = o(d^{1-\beta})$ as $d \to \infty$.

Remark 2: A related measure of risk that is often considered in the variable selection problem is the *probability of wrong recovery*. The probability of wrong recovery $R_{\mathrm{WR},d}(\hat{\boldsymbol{\eta}})$ concerns the chance that a selector $\hat{\boldsymbol{\eta}}$ will not agree with $\boldsymbol{\eta}$. In our setup, this can be expressed as

$$R_{\mathrm{WR},d}(\hat{\boldsymbol{\eta}}) = \mathbf{P}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} \neq \boldsymbol{\eta}).$$

In the context of variable selection, the probability of wrong recovery has been used as a condition for the consistency of estimators of regression coefficients in sparse high-dimensional linear regression models. For example, Genovese et al. [6] studied the lasso estimator on a sparse high-dimensional linear regression model in a Bayesian framework, proving that, under certain conditions, the probability of wrong recovery tends to zero. Although the probability of wrong recovery has been used extensively by other authors, we argue that the Hamming risk is a more general measure of risk. Note that

$$\mathbf{P}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} \neq \boldsymbol{\eta}) = \mathbf{P}_{\boldsymbol{\eta}}(|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}| \geq 1) \leq \mathbb{E}_{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}|,$$

where the inequality on the right-hand side is due to Markov's inequality. Thus, in view of the above relation, the probability of wrong recovery is bounded from above by the Hamming risk. This implies that if the Hamming risk tends to zero, the probability of wrong recovery is guaranteed to tend to zero. For this reason, we only consider the Hamming risk as a measure of selector performance in this thesis.

The possibility of exact and almost full variable selection in model (1.1) will be

dependent on the strength of the signal μ, the distribution of the noise ϵ, and the level of sparsity β. If the signal μ is too weak, then the difference between the noise components and signal components of $\boldsymbol{X}$ will be impossible to distinguish. For fixed μ, the noise components would hide the signal and it would become undetectable for larger and larger d. The choice of ϵ is important because some distributions assign more or less weight to their tails. If ϵ has high kurtosis, then the signal components will be more challenging to detect and identify. If there is high sparsity ($\beta \approx 1$), then realizing that any signal exists at all will be highly nontrivial.

The most popular and well-studied distribution for ϵ is the normal distribution. In Cui [4], the variable selection problem was studied in great detail for $\epsilon \sim \mathcal{N}(0,1)$. To study this problem, Cui [4] leveraged the property that if $\varepsilon_1, \varepsilon_2, \ldots$ is a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables, then

$$\lim_{d\to\infty} \mathbf{P}\left(\max_{1\le j\le d} |\varepsilon_j| > \sqrt{2\log d}\right) = 0. \tag{1.6}$$

Property (1.6) implies that the maximum of the noise components of $\boldsymbol{X}$ will exceed $\sqrt{2\log d}$ with probability tending to 0. Thus, if we consider some threshold set slightly above $\sqrt{2\log d}$, then when a component of $\boldsymbol{X}$ is observed to be greater than such a threshold, the signal can be determined to be present with a near certain level of confidence. Furthermore, it is clear that if the signal μ_d is much smaller than $\sqrt{2\log d}$, the components of $\boldsymbol{X}$ containing μ_d will not exceed $\sqrt{2\log d}$ at a reliable rate. In this case, the noise is said to "obstruct" the signal. If the signal strength μ_d is markedly greater than $\sqrt{2\log d}$, nearly all components with the signal present will exceed $\sqrt{2\log d}$. This suggests that with a parameterization $\mu = \mu_d = \sqrt{2r\log d}$ with $r > 0$ in model (1.1), it should be hard but achievable to determine when exact and almost full selection are possible for $\epsilon \sim \mathcal{N}(0,1)$. In fact, the main findings of Cui [4]

were as follows.

Theorem 1 (**Theorems 2 to 5 in [4]**). *Consider the random vector* $\boldsymbol{X}$ *in model* (1.1)*, where* $\epsilon \sim \mathcal{N}(0,1), \boldsymbol{\eta} \in H_{d,\beta}, 0 < \beta < 1$*, and the sequence* $\mu_d = \mu(d,\beta) \in \mathbb{R}_+$ *assigns the strength of the signal to components of* $\boldsymbol{X}$*.*

(I). *If* $\liminf_{d\to\infty} \frac{\mu_d}{\sqrt{\log d}} > \sqrt{2}(1+\sqrt{1-\beta})$*, exact variable selection is possible.*

(II). *If* $\limsup_{d\to\infty} \frac{\mu_d}{\sqrt{\log d}} < \sqrt{2}(1+\sqrt{1-\beta})$*, exact variable selection is impossible.*

(III). *If* $\liminf_{d\to\infty} \frac{\mu_d}{\sqrt{\log d}} > \sqrt{2\beta}$*, almost full variable selection is possible.*

(IV). *If* $\limsup_{d\to\infty} \frac{\mu_d}{\sqrt{\log d}} < \sqrt{2\beta}$*, neither exact nor almost full variable selection are possible.*

If the sequence $\mu_d = \sqrt{2r\log d}$ is chosen in Theorem 1, then conditions I-IV on μ_d imply that the parameter space $\left\{(\beta,r) \in \mathbb{R}^2 : (\beta,r) \in (0,1)\times(0,4)\right\}$ can be divided into three regions. If (β,r) are such that exact selection is possible, then we say that (β,r) falls within the region of *exact selection*. If (β,r) are such that only almost full selection is possible, it is said that (β,r) falls within the region of *almost full selection*. If (β,r) do not fall within the region of exact or almost full selection, it is said that *no selection is possible*. The division of the parameter space into three subregions when $\mu_d = \sqrt{2r\log d}$ is shown in Figure 1.1.

In this thesis, we shall extend the work of Cui [4] and some previous authors by supposing that the components of the noise vector $\boldsymbol{\varepsilon}$ in model (1.1) originate from a generalized normal distribution and study the variable selection problem in depth. This distribution was first mentioned by the Russian mathematician Mikhail Subbotin [16]. A random variable Z is said to have a *generalized normal distribution*, or *Subbotin distribution*, if it has a density function of the form

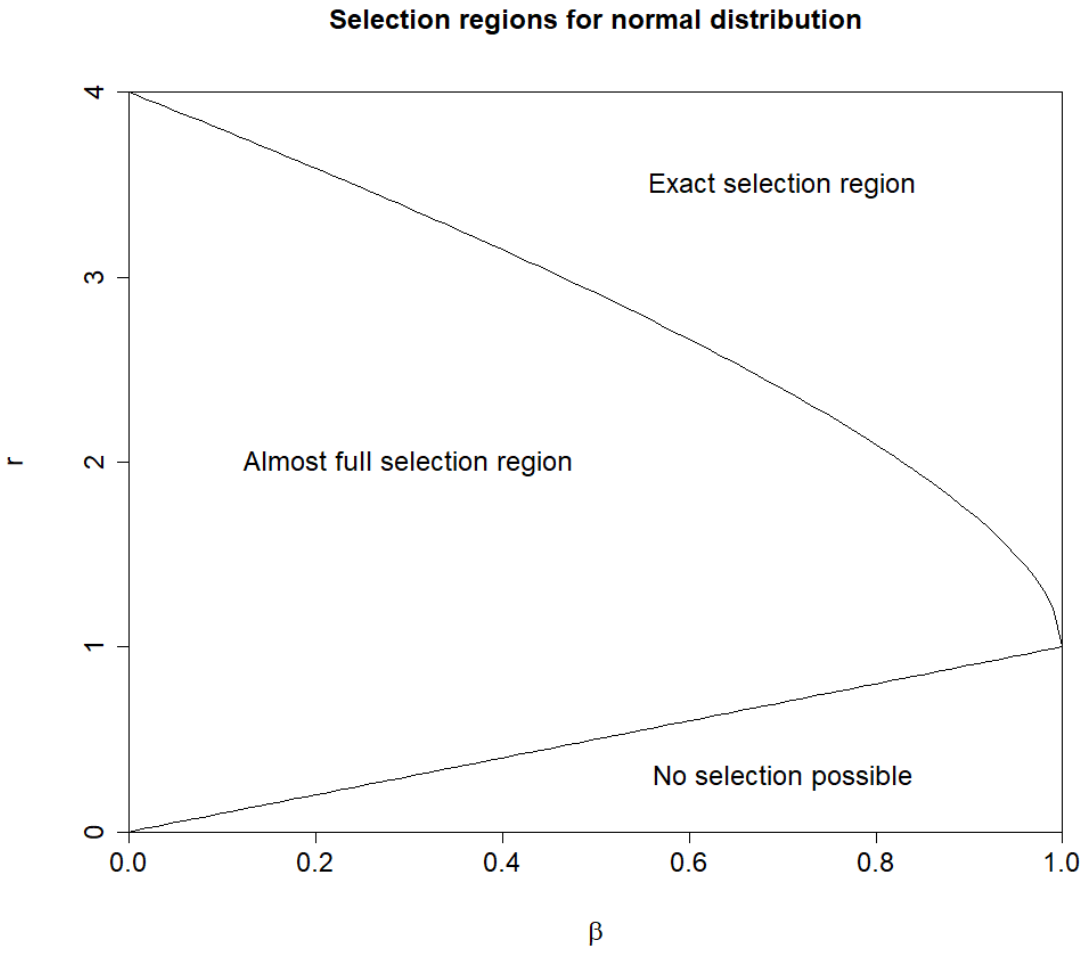

Figure 1.1: *Partition of the parameter space into the regions of variable selection for the normal distribution.*

$$f(z;\mu,\gamma,\lambda) = \frac{\gamma}{2\Gamma\left(\frac{1}{\gamma}\right)\lambda} \exp\left(-\left(\frac{|z-\mu|}{\lambda}\right)^{\gamma}\right), \tag{1.7}$$

where $z \in \mathbb{R}, \mu \in \mathbb{R}, \gamma > 0$, and $\lambda > 0$. To make this density function more accessible in this thesis, we reduce the number of free parameters by 1 by setting λ equal to $\gamma^{1/\gamma}$. This simplifies the density function of Z to the form

$$f(z;\mu,\gamma) = \frac{1}{2\Gamma\left(\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}-1}} \exp\left(-\frac{|z-\mu|^{\gamma}}{\gamma}\right). \tag{1.8}$$

For a visual representation of the shape of f in (1.8), the reader may refer to Figure 1.2. The rich class of densities in (1.8) contains several greater known distributions as special cases. For $\gamma = 1, 2$, we can recognize the p.d.f's as belonging to that of a Laplace distribution and normal distribution, respectively. As mentioned in Section 1 of Dytso et al. [5], when γ tends to infinity, the density function of the Subbotin distribution in (1.7) approaches the uniform distribution on the interval $(\mu - \lambda, \mu + \lambda)$. The generalized normal distribution is particularly useful when one would like to vary the density assigned to the tails. As a function of γ, the kurtosis of the Subbotin distribution can be shown to decrease as γ is increased (see, for example, the derivation of the central moments of the generalized normal distribution in Nadarajah [12]). Given the flexible control on the shape of the tail offered by this distribution, it is well suited to describing data contaminated by outliers. Box and Tiao in Chapter 3 of [2] also detail applications of the Subbotin distribution in a Bayesian framework.

In a manner analogous to Cui [4], we begin our study of the variable selection problem by deriving an asymptotic upper bound on the maximum of $\varepsilon_1, \ldots, \varepsilon_d \overset{\text{i.i.d.}}{\sim} \text{GN}_\gamma(0)$. Observe that for $q > 0$,

$$\begin{aligned}
\mathbf{P}\left(\max_{1\le i\le d} |\varepsilon_i| \ge (\gamma q \log d)^{1/\gamma}\right) &\le \mathbf{P}\left(\cup_{i=1}^{d}\{|\varepsilon_i| \ge (\gamma q \log d)^{1/\gamma}\}\right) \\
&\le \sum_{i=1}^{d} \mathbf{P}\left(|\varepsilon_i| \ge (\gamma q \log d)^{1/\gamma}\right) \\
&\le 2\sum_{i=1}^{d} \mathbf{P}(\varepsilon_i \ge (\gamma q \log d)^{1/\gamma})
\end{aligned}$$

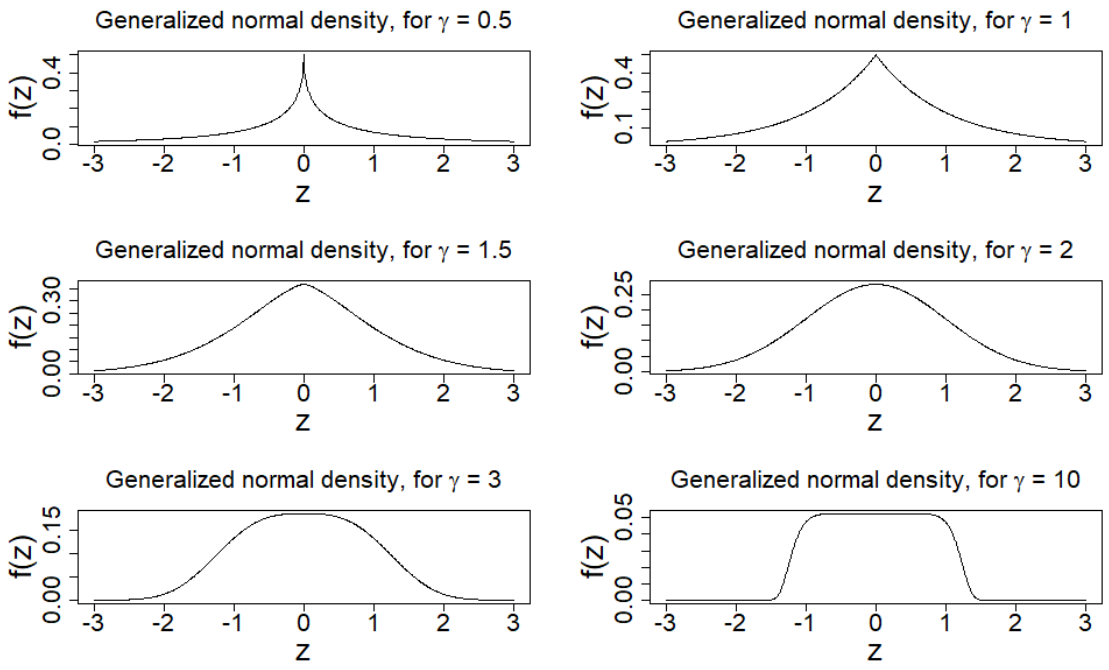

Figure 1.2: *Generalized normal density plot for various choices of* γ.

$$= 2d\mathbf{P}(\varepsilon_1 \geq (\gamma q \log d)^{1/\gamma}).$$

By Property 1 from Appendix A, we have for a known constant $B^*_{1,\gamma} > 0$,

$$2d\mathbf{P}(\varepsilon_1 \geq (\gamma q \log d)^{1/\gamma}) \sim 2B^*_{1,\gamma}(\log d)^{\frac{1}{\gamma}-1} d^{1-q}, \quad d \to \infty.$$

Thus, for any $q > 1$, as $d \to \infty$,

$$\mathbf{P}\left(\max_{1\leq i\leq d} |\varepsilon_i| \geq (\gamma q \log d)^{1/\gamma}\right) \to 0. \tag{1.9}$$

The above relation implies that the maximum of the noise components of $\boldsymbol{X}$ in model (1.1) will exceed $(\gamma \log d)^{1/\gamma}$ with probability tending to 0, assuming $\epsilon \sim \mathrm{GN}_\gamma(0)$. Thus, if we consider some threshold set slightly above $(\gamma \log d)^{1/\gamma}$, the components of $\boldsymbol{X}$ observed to be greater than such a threshold can be predicted to contain the signal with a near certain level of confidence. Also, if the signal μ_d is much smaller

than $(\gamma \log d)^{1/\gamma}$, components of $\boldsymbol{X}$ containing the signal will not exceed the considered threshold at a reliable rate. If the signal is significantly larger than $(\gamma \log d)^{1/\gamma}$, then nearly all signal components of $\boldsymbol{X}$ will exceed a threshold set near $(\gamma \log d)^{1/\gamma}$. In conclusion, property (1.9) will have two consequences for us. Firstly, if we choose a parameterization $\mu_d = (\gamma r \log d)^{1/\gamma}$ with $r > 0$ for the signal, it should be achievable, albeit challenging, to find the regions of exact and almost full selection. Secondly, it would be reasonable to construct a selector $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_d)$ of a sparse vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ such that $\hat{\eta}_i = 1$ when X_i exceeds some threshold near $(\gamma \log d)^{1/\gamma}$, and $\hat{\eta}_i = 0$ otherwise.

Summarizing, the main goals of this thesis will be as follows. Assuming model (1.1) with $\epsilon \sim \mathrm{GN}_\gamma(0), \boldsymbol{\eta} \in H_{d,\beta}$, and $\mu = \mu_d = (\gamma r \log d)^{1/\gamma}$:

(a) For all $0 < \beta < 1$, find some function $\phi_1(\beta, \gamma) > 0$ such that for $r > \phi_1(\beta, \gamma)$, exact variable selection is possible, and for $0 < r < \phi_1(\beta, \gamma)$, exact variable selection is impossible.

(b) For all $0 < \beta < 1$, find some function $0 < \phi_2(\beta, \gamma) < \phi_1(\beta, \gamma)$ such that for $r > \phi_2(\beta, \gamma)$, almost full variable selection is possible, and for $0 < r < \phi_2(\beta, \gamma)$, almost full variable selection is impossible.

(c) Provide concrete examples of exact and almost full selectors of $\boldsymbol{\eta}$. This would include a selector which is adapted to unknown β.

Remark 3: In general, we shall see that for large values of β and small values for r, the variable selection is more challenging. This corresponds to a sparse and weak signal. For small values of β and large values of r, variable selection is more straightforward. This would correspond to a dense and strong signal.

In the following chapter, our focus will be on deriving results in line with (a) through (c). We shall first find the functions $\phi_1(\beta, \gamma)$ and $\phi_2(\beta, \gamma)$. The determination

of $\phi_1(\beta, \gamma)$ and $\phi_2(\beta, \gamma)$ in (a) and (b) will naturally lead us to discover some examples of selectors which achieve exact and almost full selection. In Chapter 3, we then proceed with a simulation study to show that the selectors in (c) perform well in practice. In Chapter 4, we conclude the thesis by suggesting areas of future work.

Chapter 2

Main Results

2.1 Exact variable selection

We begin by deriving the regions for which exact selection is possible.

Theorem 2. *For* $\mu_d = (\gamma r \log d)^{1/\gamma}$ *with* $r > (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma}, \boldsymbol{\eta} \in H_{d,\beta}, 0 < \beta < 1$, *and* $\epsilon \sim GN_{\gamma}(0)$ *in model* (1.1)*, exact selection is possible. That is, there exists a selector* $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_d)$ *of* $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d) \in H_{d,\beta}$ *satisfying*

$$\limsup_{d\to\infty} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}} |\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}| = 0. \tag{2.1}$$

Proof: Consider a selector $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \ldots, \hat{\eta}_d)$ of $\boldsymbol{\eta} \in H_{d,\beta}$, with components

$$\hat{\eta}_j = \mathbb{1}(X_j > ((\gamma + \delta) \log d)^{1/\gamma}), \quad j = 1, \ldots, d, \tag{2.2}$$

where $\delta = \delta(d) > 0$ is such that $\delta \to 0$ and $\frac{\delta \log d}{\log\log d} \to \infty$ as $d \to \infty$. An upper bound on the supremum of the Hamming risk is as follows:

$$\sup_{\boldsymbol{\eta}\in H_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}} |\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}| = \sup_{\boldsymbol{\eta}\in H_{d,\beta}} \left(\sum_{j=1}^{d} \mathbb{E}_{\eta_j} |\hat{\eta}_j - \eta_j| \right)$$

$$
\begin{aligned}
&\leq \sup_{\boldsymbol{\eta}\in H_{d,\beta}} \left(\sum_{j:\eta_j=0} \mathbf{P}\Big(X_j > ((\gamma+\delta)\log d)^{1/\gamma}\Big) \right) \\
&+ \sup_{\boldsymbol{\eta}\in H_{d,\beta}} \left(\sum_{j:\eta_j=1} \mathbf{P}\Big(X_j \leq ((\gamma+\delta)\log d)^{1/\gamma}\Big) \right) \\
&\leq \Bigg[d\mathbf{P}\Bigg(\mathrm{GN}_\gamma(0) > \left((1+\delta/\gamma)\,\gamma\log d\right)^{1/\gamma} \Bigg) \\
&+ C_1 d^{1-\beta}\mathbf{P}\Bigg(\mathrm{GN}_\gamma((\gamma r\log d)^{\frac{1}{\gamma}}) \leq \left((1+\delta/\gamma)\,\gamma\log d\right)^{1/\gamma} \Bigg) \Bigg].
\end{aligned}
$$

Applying Properties 1 and 3 from Appendix A, as $d\to\infty$, we may continue

$$
\begin{aligned}
\sup_{\boldsymbol{\eta}\in H_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}| &\leq B^*_{1,\gamma}(\log d)^{\frac{1}{\gamma}-1} d^{-\frac{\delta}{\gamma}}(1+o(1)) \\
&\quad + C_1 B^*_{3,\gamma}(\log d)^{\frac{1}{\gamma}-1} d^{1-\beta-\left(r^{\frac{1}{\gamma}}-\left(1+\frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^\gamma}(1+o(1)) \stackrel{\text{def}}{=} I_d^{(1)} + I_d^{(2)},
\end{aligned}
$$

where $B^*_{1,\gamma}$ and $B^*_{3,\gamma}$ are known positive constants, $I_d^{(1)}$ is the first term, and $I_d^{(2)}$ is the second term in the above expression. To proceed, note that having $I_d^{(1)}$ tend to zero is guaranteed by having $(1-\gamma)-\frac{\delta\log d}{\log\log d}$ tend to minus infinity. Thus, recalling our choice of δ in (2.2), we have

$$
I_d^{(1)} = o(1) \quad \text{as } d\to\infty.
$$

Let us now examine $I_d^{(2)}$. To have $I_d^{(2)} = o(1)$ as $d\to\infty$, we need the quantity $1-\beta-(r^{\frac{1}{\gamma}}-(1+\delta)^{\frac{1}{\gamma}})^\gamma$ to be negative for large enough d. Hence, noting that δ is such that $\delta\to 0$ as $d\to\infty$, we may conclude that for $r > (1+(1-\beta)^{\frac{1}{\gamma}})^\gamma$,

$$
I_d^{(2)} = o(1) \quad \text{as } d\to\infty.
$$

Thus, with the sequence $\delta=\delta(d)$ in (2.2) chosen such that $\delta\to 0$ and $\frac{\delta\log d}{\log\log d}\to\infty$ as $d\to\infty$, we have for $r > (1+(1-\beta)^{\frac{1}{\gamma}})^\gamma$ and $0<\beta<1$,

$$\sup_{\boldsymbol{\eta}\in H_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}}|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}| = o(1) \quad \text{as } d \to \infty.$$

The proof of Theorem 2 is complete. □

We now turn our attention to proving when exact variable selection is not possible. In what follows, we shall study the maximum Hamming risk over the parameter space $H^{\pm}_{d,\beta}$ instead of $H_{d,\beta}$. Since $H^{\pm}_{d,\beta} \subset H_{d,\beta}$, the impossibility of exact or almost full selection in $H^{\pm}_{d,\beta}$ will immediately imply the impossibility of exact selection in $H_{d,\beta}$.

Theorem 3. *For* $\mu_d = (\gamma r \log d)^{1/\gamma}$ *with* $\beta < r < (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma}, \boldsymbol{\eta} \in H^{\pm}_{d,\beta}, 0 < \beta < 1$, *and* $\epsilon \sim GN_{\gamma}(0)$ *in model* (1.1)*, exact selection is not possible. That is, for* $\beta < r < (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma}$, *we have*

$$\liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H^{\pm}_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0, \tag{2.3}$$

where the infimum is taken over all selectors $\bar{\boldsymbol{\eta}}$ *of* $\boldsymbol{\eta}$ *based on* $\boldsymbol{X}$ *in model* (1.1).

Proof: Let us view $\boldsymbol{\eta}$ as having some prior distribution and denote the parameter space of $\boldsymbol{\eta}$ as

$$\mathcal{X}^d = \{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d) : \eta_j \in \{0, 1\}, j = 1, \ldots, d\}.$$

Now, consider an arbitrarily chosen probability measure $\mathbf{P}_{\pi}$ which returns the probability that $\boldsymbol{\eta}$ belongs to any subset of $\mathcal{X}^d$. Let $\pi(\cdot)$ be a prior p.m.f. of $\boldsymbol{\eta}$ and observe that

$$\sup_{\boldsymbol{\eta}\in\mathcal{X}^d} \mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| = \sup_{\boldsymbol{\eta}\in\mathcal{X}^d} \mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \int_{\mathcal{X}^d} \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \geq \int_{\mathcal{X}^d} \mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}.$$

Therefore the Bayes risk with respect to any prior distribution for $\boldsymbol{\eta}$ yields a lower bound on the minimax risk. It is natural to consider a prior p.m.f for $\eta_j, j = 1, \ldots, d$,

of the form

$$\pi_j(\eta_j) = \begin{cases} d^{-\beta}, & \eta_j = 1, \\ 1 - d^{-\beta}, & \eta_j = 0, \end{cases}$$

and put $\pi(\boldsymbol{\eta}) = \Pi_{j=1}^{d} \pi_j(\eta_j)$. That is, $\eta_1, \ldots, \eta_d \overset{i.i.d.}{\sim} \text{Bernoulli}(d^{-\beta})$.

Consider the event

$$A = \{\boldsymbol{\eta} : \boldsymbol{\eta} \in H_{d,\beta}^{\pm}\}.$$

Firstly, we show that for the prior p.m.f. of $\boldsymbol{\eta}$ which we have selected, we have

$$\mathbf{P}_\pi(\boldsymbol{\eta} \in A) \geq 1 + o(d^{-1}) \quad \text{as } d \to \infty. \tag{2.4}$$

To this end, note

$$\begin{aligned} \mathbf{P}_\pi\left(\boldsymbol{\eta} \in A\right) &= \mathbf{P}_\pi\left(C_0 d^{1-\beta} \leq \sum_{j=1}^{d} \eta_j \leq C_1 d^{1-\beta}\right) \\ &= \mathbf{P}_\pi\left((C_0 - 1)d^{1-\beta} \leq \sum_{j=1}^{d} (\eta_j - d^{-\beta}) \leq (C_1 - 1)d^{1-\beta}\right) \\ &= \mathbf{P}_\pi\left((C_0 - 1)d^{1-\beta} \leq \sum_{j=1}^{d} (\eta_j - p) \leq (C_1 - 1)d^{1-\beta}\right), \end{aligned}$$

where we set $p = d^{-\beta}$. Letting $\zeta = \min(C_1 - 1, 1 - C_0)$, it can be seen that

$$\begin{aligned} &\mathbf{P}_\pi\left((C_0 - 1)d^{1-\beta} \leq \sum_{j=1}^{d} (\eta_j - p) \leq (C_1 - 1)d^{1-\beta}\right) \\ &\quad \geq \mathbf{P}_\pi\left(-\zeta d^{1-\beta} \leq \sum_{j=1}^{d} (\eta_j - p) \leq \zeta d^{1-\beta}\right) = 1 - \mathbf{P}_\pi\left(\left|\sum_{j=1}^{d} (\eta_j - p)\right| \geq \zeta d^{1-\beta}\right). \end{aligned}$$

Next, since $\mathbb{E}_\pi(\eta_j - p) = 0$ and $|\eta_j - p| \leq 2$ for $j = 1, 2, \ldots, d$, we can proceed by applying Bernstein's first inequality from Appendix A. As $d \to \infty$, we have

$$1 - \mathbf{P}_\pi \left(\left| \sum_{j=1}^{d} (\eta_j - p) \right| \geq \zeta d^{1-\beta} \right) \geq 1 - 2\exp\left(\frac{-(\zeta d^{1-\beta})^2}{2\left(p(1-p)d + \frac{2\zeta d^{1-\beta}}{3}\right)} \right)$$
$$= 1 - 2\exp\left(\frac{-\zeta^2 d^{1-\beta}}{2 - 2d^{-\beta} + \frac{4\zeta}{3}} \right)$$
$$\geq 1 - 2\exp\left(-Qd^{1-\beta}\right),$$

where Q is some positive constant. Thus, we have shown that as $d \to \infty$,

$$\mathbf{P}_\pi\left(\boldsymbol{\eta} \in A\right) \geq 1 - \exp(-Qd^{1-\beta}) = 1 + o(d^{-1}).$$

Now, since we are interested in the minimax risk when $\boldsymbol{\eta} \in H_{d,\beta}^{\pm} \subseteq \mathcal{X}^d$, let us restrict our attention to subsets of the event A. For B in $\mathcal{X}^d$, the set function $\mathbf{P}_{\pi|A}(B) = \frac{\mathbf{P}_\pi(\boldsymbol{\eta} \in B \cap A)}{\mathbf{P}_\pi(\boldsymbol{\eta} \in A)}$ is also a probability measure, so it can be used to determine a lower bound for $\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H_{d,\beta}^{\pm}} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}|$. Denoting the p.m.f. of $\boldsymbol{\eta}$ for the measure $\mathbf{P}_{\pi|A}$ as $\pi(\boldsymbol{\eta}|A)$ and using (2.4), a lower bound on the minimax Hamming risk is given by

$$\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H_{d,\beta}^{\pm}} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \geq \inf_{\bar{\boldsymbol{\eta}}} \int_A \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}|A) d\boldsymbol{\eta}$$
$$= \inf_{\bar{\boldsymbol{\eta}}} \frac{1}{\mathbf{P}_\pi(\boldsymbol{\eta} \in A)} \int_A \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} = (1 + o(d^{-1})) \inf_{\bar{\boldsymbol{\eta}}} \int_A \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}$$
$$\geq (1 + o(d^{-1})) \Big(\inf_{\bar{\boldsymbol{\eta}}} \int_{\boldsymbol{\eta} \in \mathcal{X}^d} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} - \sup_{\bar{\boldsymbol{\eta}}} \int_{A^c} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \Big)$$
$$\stackrel{\text{def}}{=} (1 + o(d^{-1}))(L_d^{(1)} + L_d^{(2)}), \quad d \to \infty. \tag{2.5}$$

To proceed, let us consider the behavior of the terms $L_d^{(1)}$ and $L_d^{(2)}$ separately. Firstly, considering the term $L_2^{(d)}$ and using (2.4),

$$L_2^{(d)} = \sup_{\bar{\boldsymbol{\eta}}} \int_{A^c} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} = \sup_{\bar{\boldsymbol{\eta}}} \int_{A^c} \sum_{j=1}^{d} \mathbb{E}_{\eta_j} |\bar{\eta}_j - \eta_j| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta}$$

$$\leq 2d \int_{A^c} \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} = 2d\mathbf{P}_\pi(\boldsymbol{\eta} \in A^c) = O(d)o(d^{-1}) = o(1). \tag{2.6}$$

Thus, the term $L_2^{(d)}$ vanishes to zero as d increases. Now, viewing the quantity $L_d^{(1)}$, we obtain

$$\begin{aligned} L_d^{(1)} &= \inf_{\bar{\boldsymbol{\eta}}} \int_{\mathcal{X}^d} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} = \inf_{\bar{\boldsymbol{\eta}}} \int_{\mathcal{X}^d} \left(\sum_{j=1}^d \mathbb{E}_{\eta_j} |\bar{\eta}_j - \eta_j| \right) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \\ &= \inf_{\bar{\boldsymbol{\eta}}} \mathbb{E}_\pi \left(\sum_{j=1}^d \mathbb{E}_{\eta_j} |\bar{\eta}_j - \eta_j| \right) = \inf_{\bar{\boldsymbol{\eta}}} \sum_{j=1}^d \mathbb{E}_\pi \left(\mathbb{E}_{\eta_j} |\bar{\eta}_j - \eta_j| \right) \\ &\geq \sum_{j=1}^d \inf_{\bar{\boldsymbol{\eta}}} \mathbb{E}_\pi \mathbb{E}_{\eta_j} |\bar{\eta}_j - \eta_j| = d \inf_{\bar{\eta}_1} \mathbb{E}_\pi \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1|. \end{aligned}$$

Next, note that

$$\begin{aligned} \mathbb{E}_\pi \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1| &= \mathbb{E}_{\mu_d} |1 - \bar{\eta}_1| \pi_1(1) + \mathbb{E}_0 |\bar{\eta}_1| \pi_1(0) \\ &= \mathbf{P}_{\mu_d}(\bar{\eta}_1 = 0) \pi_1(1) + \mathbf{P}_0(\bar{\eta}_1 = 1) \pi_1(0), \end{aligned}$$

where $\mathbb{E}_{\mu_d}$ is expectation with respect to a generalized normal distribution with mean μ_d and $\mathbb{E}_0$ is expectation with respect to a generalized normal distribution with mean 0.

Now, observe that the above quantity is simply the Bayes risk in the problem of testing

$$H_0 : X_1 \sim \mathrm{GN}_\gamma(0) \quad \text{vs} \quad H_1^{(d)} : X_1 \sim \mathrm{GN}_\gamma(\mu_d).$$

Designating f_{μ_d} as the density of a generalized normal distribution with mean $\mu_d > 0$, it is known that a decision function with rejection region

$$\left\{ x \in \mathbb{R} : f_{\mu_d}(x) > \frac{\pi_1(0)}{\pi_1(1)} f_0(x) \right\} \tag{2.7}$$

minimizes the Bayes risk in the above simple hypothesis test (see, for example, Section 13.9 of Roussas [14]). Therefore the minimizer of the Bayes risk will be the function $\eta^* = \mathbb{1}\Big(\frac{pf_{\mu_d}(X_1)}{(1-p)f_0(X_1)} > 1\Big)$, where we set $p = d^{-\beta}$. This can be written as

$$\eta^* = \mathbb{1}\left(\frac{p}{1-p}\exp\left(\frac{|X_1|^\gamma - |X_1-\mu_d|^\gamma}{\gamma}\right) > 1\right),$$

which is equivalent to

$$\eta^* = \mathbb{1}\left(|X_1|^\gamma - |X_1-\mu_d|^\gamma > \gamma\log\left(\frac{1-p}{p}\right)\right).$$

Observe that with $p = d^{-\beta}$ we get $\log\left(\frac{1-p}{p}\right) = \log(d^\beta - 1) = \beta\log d + \log(1-d^{-\beta})$. To proceed further, it will convenient to define the sequence of random variables $Y_d = X_1/\left(\gamma r\log d\right)^{1/\gamma}$. Noting that $X_1 = Y_d\left(\gamma r\log d\right)^{1/\gamma}$, we have

$$\begin{aligned}
&\left\{|X_1|^\gamma - |X_1-\mu_d|^\gamma > \gamma\log\left(\frac{1-p}{p}\right)\right\} \\
&= \left\{|X_1|^\gamma - |X_1-\mu_d|^\gamma > \gamma\beta\log d + \gamma\log(1-d^{-\beta})\right\} \\
&= \left\{|Y_d|^\gamma - |Y_d-1|^\gamma > \beta/r + \Lambda_d\right\},
\end{aligned}$$

where

$$\Lambda_d = \frac{\log(1-d^{-\beta})}{r\log d} < 0$$

satisfies

$$\Lambda_d = O\left(1/(d^\beta\log d)\right) = o(1), \quad d\to\infty.$$

To analyse the values Y_d for which $|Y_d|^\gamma - |Y_d-1|^\gamma > \beta/r + \Lambda_d$, let us define the function

$$h_d(y) \stackrel{\text{def}}{=} |y|^\gamma - |y-1|^\gamma - \beta/r - \Lambda_d = \begin{cases} (-y)^\gamma - (1-y)^\gamma - \beta/r - \Lambda_d, & y \le 0, \\ y^\gamma - (1-y)^\gamma - \beta/r - \Lambda_d, & 0 < y < 1, \\ y^\gamma - (y-1)^\gamma - \beta/r - \Lambda_d, & y \ge 1. \end{cases}$$

Below, we shall distinguish between the cases $\gamma \ge 1$ and $0 < \gamma < 1$.

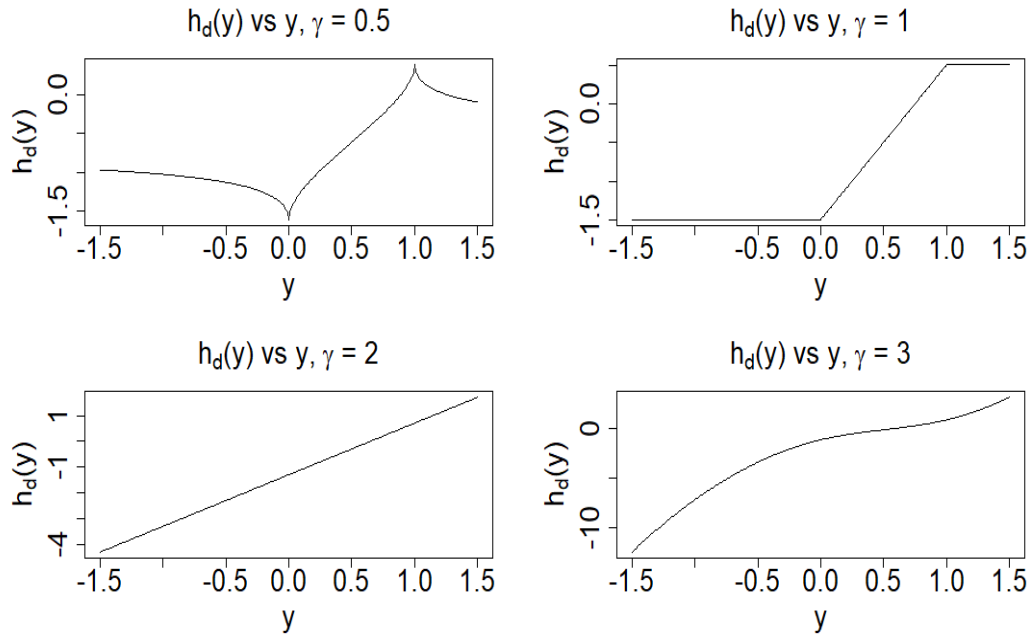

Figure 2.1: *A plot of $h_d(y)$ against y for various choices of γ, with $\beta = 0.5, r = \frac{(1+(1-\beta)^{\frac{1}{\gamma}})^\gamma + \beta}{2}$, and $d = 1000$.*

Case 1: Suppose $\gamma \ge 1$. It can be shown that $h_d(y)$ is non-decreasing on $(-\infty, 0) \cup (1, \infty)$, strictly increasing on $(0, 1)$, and a continuous function of y. Let us show that h_d has a single root. Observe that $h_d(1/2) = -\beta/r - \Lambda_d < 0$ and $h_d(1) = \frac{r-\beta}{r} - \Lambda_d > 0$ for all large enough d. Hence, by the continuity of h_d, for all large enough d there exists $a = a(r, \beta, \gamma) \in (1/2, 1)$ such that $h_d(a) = 0$. In general, it is not possible to solve for a analytically. For our purposes, it is enough to know that a root always

exists for large enough d and lies between $1/2$ and 1. Therefore, for d large,

$$\begin{aligned}\eta^* &= \mathbb{1}\left(|X_1|^\gamma - |X_1 - \mu_d|^\gamma > \gamma \log\left(\frac{1-p}{p}\right)\right) \\ &= \mathbb{1}(Y_d > a) = \mathbb{1}\left(X_1 > a(\gamma r \log d)^{1/\gamma}\right).\end{aligned}$$

Returning now to the Bayes risk, we have as $d \to \infty$,

$$\begin{aligned}d \inf_{\bar{\eta}_1} \mathbb{E}_{\pi_1}\mathbb{E}_{\eta_1}|\bar{\eta}_1 - \eta_1| &= d\Big(\mathbf{P}_{\mu_d}(\eta^* = 0)\pi_1(1) + \mathbf{P}_0(\eta^* = 1)\pi_1(0)\Big) \\ = d\Bigg(p\mathbf{P}_{\mu_d}\left(X_1 \leq a(\gamma r \log d)^{1/\gamma}\right) &+ (1-p)\mathbf{P}_0\left(X_1 > a(\gamma r \log d)^{1/\gamma}\right)\Bigg).\end{aligned}$$

Using Properties 1 and 3 from Appendix A and noting that $p = d^{-\beta}$, as $d \to \infty$,

$$\begin{aligned}d \inf_{\bar{\eta}_1} \mathbb{E}_{\pi_1}\mathbb{E}_{\eta_1}|\bar{\eta}_1 - \eta_1| &= d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma}a)^\gamma}(1 + o(1)) \\ &\quad + d(1 - d^{-\beta}) B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-ra^\gamma}(1 + o(1)) \\ &= d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma}a)^\gamma}(1 + o(1)) \\ &\quad + B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{1-ra^\gamma}(1 + o(1)).\end{aligned}$$

For the above risk to vanish to zero as $d \to \infty$, we must have simultaneously

(a) $1 - \beta - (r^{1/\gamma} - r^{1/\gamma}a)^\gamma < 0$ and

(b) $1 - ra^\gamma < 0$.

Rearranging (a) yields $r > ((1-\beta)^{\frac{1}{\gamma}} + r^{1/\gamma}a)^\gamma$ and (b) implies $r^{1/\gamma}a > 1$. For both (a) and (b) to hold true, it must be that $r > ((1-\beta)^{\frac{1}{\gamma}} + r^{1/\gamma}a)^\gamma > ((1-\beta)^{\frac{1}{\gamma}} + 1)^\gamma$. Thus, if $r < (1 + (1-\beta)^{\frac{1}{\gamma}})^\gamma, \gamma \geq 1$, then

$$d \inf_{\bar{\eta}_1} \mathbb{E}_{\pi}\mathbb{E}_{\eta_1}|\bar{\eta}_1 - \eta_1| \geq C(\log d)^{\frac{1}{\gamma}-1} d^{c'},$$

where $c' > 0, C > 0$. Clearly, this quantity diverges to infinity as d grows, and hence

$L_d^{(1)} \geq C(\log d)^{\frac{1}{\gamma}-1} d^{c'}$, implying, by (2.5) and (2.6),

$$\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H_{d,\beta}^{\pm}} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0$$

for all large enough d.

Case 2: Now suppose that $0 < \gamma < 1$. In this case, it can be shown that $h_d(y)$ is strictly decreasing on $(-\infty, 0) \cup (1, \infty)$ and strictly increasing on $(0, 1)$ (see Figure 2.1). Also,

$$\lim_{y \to \infty} \{y^\gamma - (y-1)^\gamma - \beta/r - \Lambda_d\} = -\beta/r - \Lambda_d$$

and

$$\lim_{y \to -\infty} \{(-y)^\gamma - (1-y)^\gamma - \beta/r - \Lambda_d\} = -\beta/r - \Lambda_d.$$

Furthermore, we know that for large enough d, $h_d(1/2) = -\beta/r - \Lambda_d < 0$, $h_d(1) = \frac{r-\beta}{r} - \Lambda_d > 0$, and h_d is continuous. Combining all of this together, for all large enough d, there exists $a' = a'(r, \beta, \gamma) \in (1/2, 1)$ such that $h_d(a') = 0$ and there exists $b' = b'(r, \beta, \gamma) \in (1, \infty)$ such that $h_d(b') = 0$.

Thus, for d sufficiently large,

$$\begin{aligned} d \inf_{\bar{\eta}_1} \mathbb{E}_{\pi_1} \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1| &= d\Big(\mathbf{P}_{\mu_d}(\eta^* = 0)\pi_1(1) + \mathbf{P}_0(\eta^* = 1)\pi_1(0)\Big) \\ &= d\Bigg\{ p\Bigg(\mathbf{P}_{\mu_d}\left(X_1 \leq a'(\gamma r \log d)^{1/\gamma}\right) \\ &\quad + \mathbf{P}_{\mu_d}\left(X_1 \geq b'(\gamma r \log d)^{1/\gamma}\right) \Bigg) \Bigg\} \\ &\quad + d(1-p)\mathbf{P}_0\left(a'(\gamma r \log d)^{1/\gamma} < X_1 < b'(\gamma r \log d)^{1/\gamma}\right). \end{aligned}$$

As $d \to \infty$, applying Properties 1 through 3 from Appendix A,

$$
\begin{aligned}
d \inf_{\bar{\eta}_1} \mathbb{E}_{\pi_1} \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1| &= d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma} a')^\gamma} (1 + o(1)) \\
&+ d^{1-\beta} B^*_{2,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} b' - r^{1/\gamma})^\gamma} (1 + o(1)) \\
&+ B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} (d^{1-r(a')^\gamma} - d^{1-r(b')^\gamma}) (1 + o(1)) \\
&= d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma} a')^\gamma} (1 + o(1)) \\
&+ d^{1-\beta} B^*_{2,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{\frac{1}{\gamma}} b' - r^{\frac{1}{\gamma}})^\gamma} (1 + o(1)) \\
&+ B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{1-r(a')^\gamma} (1 + o(1)) \\
&\geq d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma} a')^\gamma} (1 + o(1)) \\
&+ B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{1-r(a')^\gamma} (1 + o(1)).
\end{aligned}
$$

Invoking identical arguments as in the case $\gamma \geq 1$, it is clear that the quantity

$$
d^{1-\beta} B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-(r^{1/\gamma} - r^{1/\gamma} a')^\gamma} (1 + o(1)) + B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{1-r(a')^\gamma} (1 + o(1))
$$

does not tend to zero for d large, and hence for $0 < \gamma < 1$ it is also true that

$$
\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H^{\pm}_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0
$$

for all large enough d.

Combining cases 1 and 2 completes the proof. □

2.2 Almost full variable selection

Let us now study the problem of almost full variable selection in model (1.1). We begin by deriving the region of almost full selection.

Theorem 4. *For $\mu_d = (\gamma r \log d)^{1/\gamma}$ with $r > \beta, \boldsymbol{\eta} \in H_{d,\beta}, 0 < \beta < 1$, and $\epsilon \sim GN_\gamma(0)$ in model (1.1), almost full selection is possible. That is, there exists a selector $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)$ such that*

$$\limsup_{d\to\infty} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}| = 0. \tag{2.8}$$

Proof: Consider a selector $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)$ of $\boldsymbol{\eta} \in H_{d,\beta}$, with

$$\tilde{\eta}_j = \mathbb{1}(X_j > ((\gamma\beta + \delta)\log d)^{\frac{1}{\gamma}}), \quad j = 1, \ldots, d, \tag{2.9}$$

where $\delta = \delta(d) > 0$ satisfies $\delta \to 0$ and $\frac{\delta \log d}{\log\log d} \to \infty$ as $d \to \infty$.

Observe that

$$\begin{aligned}
\sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}| &= \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\left(\sum_{j=1}^{d}\mathbb{E}_{\eta_j}|\tilde{\eta}_j - \eta_j|\right) \\
&= \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\left(\sum_{j:\eta_j=0}\mathbf{P}\Big(X_j > ((\gamma+\delta)\log d)^{1/\gamma}\Big)\right) \\
&+ \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\left(\sum_{j:\eta_j=1}\mathbf{P}\Big(X_j \le ((\gamma\beta+\delta)\log d)^{\frac{1}{\gamma}}\Big)\right) \\
&\le d^{\beta}\mathbf{P}\Big(\mathrm{GN}_\gamma(0) > ((\beta+\delta/\gamma)\,\gamma\log d)^{\frac{1}{\gamma}}\Big) \\
&+ C_1\mathbf{P}\Big(\mathrm{GN}_\gamma\left((\gamma r\log d)^{\frac{1}{\gamma}}\right) \le ((\beta+\delta/\gamma)\,\gamma\log d)^{\frac{1}{\gamma}}\Big).
\end{aligned}$$

Applying Properties 1 and 3 from Appendix A and taking d tending to infinity,

$$\begin{aligned}
\sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}| &\le d^{\beta}B^*_{1,\gamma}\,(\log d)^{\frac{1}{\gamma}-1}\,d^{-\left(\beta+\frac{\delta}{\gamma}\right)}(1+o(1)) \\
&+ B^*_{3,\gamma}(\log d)^{\frac{1}{\gamma}-1}d^{-(r^{\frac{1}{\gamma}}-\left(\beta+\frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}})^{\gamma}}(1+o(1)) \stackrel{\text{def}}{=} J_d^{(1)} + J_d^{(2)},
\end{aligned}$$

where $J_d^{(1)}$ is the first term and $J_d^{(2)}$ is the second term in the above expression. Since

we have chosen $\delta = \delta(d) > 0$ such that $\delta \to 0$ and $\frac{\delta \log d}{\log \log d} \to \infty$ as $d \to \infty$, it is clear that $J_d^{(1)} = o(1)$ as $d \to \infty$. For $J_d^{(2)}$ to tend to zero, we need the quantity $-\left(r^{\frac{1}{\gamma}} - \left(\beta + \frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^{\gamma}$ to be negative for d large enough. This is guaranteed to occur with δ tending to zero and $r > \beta$. Hence, for the chosen sequence δ, we have that for $r > \beta, 0 < \beta < 1$,

$$\sup_{\boldsymbol{\eta} \in H_{d,\beta}} d^{\beta-1} \mathbb{E}_{\boldsymbol{\eta}} |\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}| = o(1) \quad \text{as } d \to \infty.$$

Thus, an almost full selector exists and is given by (2.9). □

Remark 4: Note that the threshold of the selector defined in (2.9) is set at a lower level than that of the exact selector defined in (2.2).

We now consider the region for which neither exact nor almost full selection are possible.

Theorem 5. *For $\mu_d = (\gamma r \log d)^{1/\gamma}$ with $0 < r < \beta, \boldsymbol{\eta} \in H_{d,\beta}^{\pm}, 0 < \beta < 1$, and $\epsilon \sim GN_\gamma(0)$ in model (1.1), almost full selection is not possible. That is,*

$$\liminf_{d \to \infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H_{d,\beta}^{\pm}} d^{\beta-1} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0, \tag{2.10}$$

where the infimum is taken over all selectors $\bar{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ based on $\boldsymbol{X}$ in model (1.1).

Proof: Again, our strategy will be to show that $\liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H_{d,\beta}^{\pm}} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0$. In a spirit similar to Theorem 3, we divide the proof into cases.

Case 1: Let us assume $\gamma > 1$. Recall the function $h_d(y) = |y|^\gamma - |y-1|^\gamma - \beta/r - \Lambda_d$. For $0 < r < \beta$, observe that $h_d(1) = \frac{r-\beta}{r} - \Lambda_d < 0$ for d large. Additionally, h_d is strictly increasing. To show h_d has a single root, let us show that $h_d(y)$ approaches ∞ as $y \to \infty$. Note that for $y > 1$,

$$y^\gamma - (y-1)^\gamma = \gamma \int_{y-1}^{y} z^{\gamma-1} dz > \gamma (y-1)^{\gamma-1},$$

which implies $h_d(y) \to \infty$ as $y \to \infty$. Thus, since h_d is continuous, for all large enough d, there exists $\alpha = \alpha(r, \beta, \gamma) \in (1, \infty)$ such that $h_d(\alpha) = 0$. Reproducing the same steps as in the proof of Theorem 3, we obtain as $d \to \infty$,

$$\begin{aligned}
\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H^{\pm}_{d,\beta}} d^{\beta-1} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| &\geq d^\beta \inf_{\bar{\eta}_1} \mathbb{E}_\pi \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1| \\
&= d^\beta \Big(p \mathbf{P}_{\mu_d}(X_1 \leq \alpha(\gamma r \log d)^{1/\gamma}) \\
&\quad + (1-p) \mathbf{P}_0(X_1 > \alpha(\gamma r \log d)^{1/\gamma}) \Big) \\
&\geq \mathbf{P}_{\mu_d}(X_1 \leq \alpha(\gamma r \log d)^{1/\gamma}).
\end{aligned}$$

By symmetry of the generalized normal distribution, $\mathbf{P}_{\mu_d}(X_1 \leq \alpha(\gamma r \log d)^{1/\gamma}) > 1/2$. Thus, for $\gamma > 1$, $0 < r < \beta$, and $0 < \beta < 1$,

$$\liminf_{d \to \infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta} \in H^{\pm}_{d,\beta}} d^{\beta-1} \mathbb{E}_{\boldsymbol{\eta}} |\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 1/2.$$

Case 2: Let us assume $0 < \gamma < 1$. For $0 < r < \beta$, it is clear that $h_d(1) = \frac{r-\beta}{r} - \Lambda_d < 0$ for large enough d. Also, since $\lim_{y \to \infty} h_d(y) = -\beta/r - \Lambda_d, \lim_{y \to -\infty} h_d(y) = -\beta/r - \Lambda_d, h_d(y)$ is decreasing on $(-\infty, 0) \cup (1, \infty), h_d(y)$ is continuous, and $h_d(y)$ is increasing on $(0, 1)$, we must have $\sup_y h_d(y) = h_d(1) < 0$ for all large enough d. This implies that for d large enough, the event $\{|Y_d|^\gamma - |Y_d - 1|^\gamma \leq \beta/r + \Lambda_d\}$ is the certain event Ω. Hence, as $d \to \infty$,

$$\begin{aligned}
d^\beta \inf_{\bar{\eta}_1} \mathbb{E}_\pi \mathbb{E}_{\eta_1} |\bar{\eta}_1 - \eta_1| = d^\beta \Bigg(& \mathbf{P}_{\mu_d} \Big(|Y_d|^\gamma - |Y_d - 1|^\gamma \leq \beta/r + \Lambda_d \Big) \pi_1(1) \\
& + \mathbf{P}_0 \Big(|Y_d|^\gamma - |Y_d - 1|^\gamma > \beta/r + \Lambda_d \Big) \pi_1(0) \Bigg)
\end{aligned}$$

$$= \mathbf{P}_{\mu_d}(\Omega) + d^{\beta}(1 - d^{-\beta})\mathbf{P}_0(\emptyset) = 1,$$

which implies

$$\liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H^{\pm}_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0.$$

Case 3: Let us now assume $\gamma = 1$. In this case, we have

$$\begin{aligned}\inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H^{\pm}_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| &\geq d^{\beta}\Bigg\{p\mathbf{P}_{\mu_d}\left(|X_1| - |X_1 - \mu_d| \leq \log\left(\frac{1-p}{p}\right)\right) \\ &\quad + (1-p)\mathbf{P}_0\left(|X_1| - |X_1 - \mu_d| > \log\left(\frac{1-p}{p}\right)\right)\Bigg\} \\ &\geq \mathbf{P}_{\mu_d}\left(|X_1| - |X_1 - \mu_d| \leq \log\left(\frac{1-p}{p}\right)\right),\end{aligned}$$

where $\log\left(\frac{1-p}{p}\right) \sim \beta \log d$ as $d \to \infty$. Observe that when $X_1 \geq \mu_d$, we have $|X_1| - |X_1 - \mu_d| = r \log d$. By monotonicity of $\mathbf{P}_{\mu_d}$ and noting that $r \log d < \beta \log d$, as $d \to \infty$,

$$\mathbf{P}_{\mu_d}\left(|X_1| - |X_1 - \mu_d| \leq \log\left(\frac{1-p}{p}\right)\right) \geq \mathbf{P}_{\mu_d}(X_1 \geq \mu_d) = 1/2.$$

Thus, for $\gamma = 1, 0 < r < \beta$, and $0 < \beta < 1$,

$$\liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H^{\pm}_{d,\beta}} d^{\beta-1}\mathbb{E}_{\eta}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 1/2.$$

Combining cases 1 to 3 completes the proof of Theorem 5. □

2.3 Comments and extensions to previous results

Summarizing Theorems 2 through 5, if we let $\phi_1(\beta, \gamma) = (1+(1-\beta)^{\frac{1}{\gamma}})^{\gamma}$ and $\phi_2(\beta, \gamma) = \beta$, then for all $0 < \beta < 1$:

1. Exact variable selection is possible for $r > \phi_1(\beta, \gamma) = (1 + (1 - \beta)^{\frac{1}{\gamma}})^{\gamma}$ and impossible for $0 < r < \phi_1(\beta, \gamma) = (1 + (1 - \beta)^{\frac{1}{\gamma}})^{\gamma}$.

2. Almost full variable selection is possible for $r > \phi_2(\beta, \gamma) = \beta$ and impossible for $0 < r < \phi_2(\beta, \gamma) = \beta$.

Remark 5: Note that the impossibility of almost full selection for $0 < r < \beta$ also implies the impossibility of exact selection. This is since

$$\liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} \mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > \liminf_{d\to\infty} \inf_{\bar{\boldsymbol{\eta}}} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\bar{\boldsymbol{\eta}} - \boldsymbol{\eta}| > 0.$$

Furthermore, Theorem 5 implies the impossibility of almost full selection for fixed signal strengths. This is since for fixed $\mu \in \mathbb{R}_+$, there exists some positive integer D such that $\mu < (\gamma\beta \log d)^{\frac{1}{\gamma}}$ for $d > D$. This justifies the requirement that the signal strength μ_d should tend to infinity for the variable selection problem to be meaningful.

Analogous to Theorem 1, the results of Theorems 2 through 5 can be made more general. Specifically, there are many possible sequences μ_d for which exact and almost full selection are possible.

Corollary 1. *Consider the random vector $\boldsymbol{X}$ in model (1.1), where $\epsilon \sim GN_{\gamma}(0), \boldsymbol{\eta} \in H_{d,\beta}, 0 < \beta < 1$, and the sequence $\mu_d = \mu(d, \beta) \in \mathbb{R}_+$ assigns the strength of the signal components of $\boldsymbol{X}$.*

(I). *If $\liminf_{d\to\infty} \frac{\mu_d}{(\gamma \log d)^{1/\gamma}} > 1 + (1 - \beta)^{\frac{1}{\gamma}}$, exact variable selection is possible.*

(II). *If $\limsup_{d\to\infty} \frac{\mu_d}{(\gamma \log d)^{1/\gamma}} < 1 + (1 - \beta)^{\frac{1}{\gamma}}$, exact variable selection is impossible.*

(III). *If $\liminf_{d\to\infty} \frac{\mu_d}{(\gamma \log d)^{1/\gamma}} > \beta^{1/\gamma}$, almost full variable selection is possible.*

(IV). *If $\limsup_{d\to\infty} \frac{\mu_d}{(\gamma \log d)^{1/\gamma}} < \beta^{1/\gamma}$, exact nor almost full variable selection are possible.*

Proof: In view of Theorems 2 through 5, the proof is obvious. □

In addition to the results of Sections 2.1 and 2.2, the selection regions are completely determined for when the density function is represented by (1.7). In fact, if the density function of ϵ in model (1.1) is according to (1.7), then for $\mu_d = \lambda(r \log d)^{1/\gamma}$, the region of exact and almost full selection are identical to those implied by Theorems 2 through 5.

Corollary 2. *For a random variable ϵ with density function f according to* (1.7)*, $\mu_d = \lambda(r \log d)^{1/\gamma}$, and $\boldsymbol{\eta} \in H_{d,\beta}, 0 < \beta < 1$, in model* (1.1)*, the following statements hold true.*

1. *Exact selection is possible for $r > \phi_1(\beta, \gamma) = (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma}$ and impossible for $0 < r < \phi_1(\beta, \gamma) = (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma}$.*

2. *Almost full selection is possible for $r > \phi_2(\beta, \gamma) = \beta$ and impossible for $0 < r < \phi_2(\beta, \gamma) = \beta$.*

Proof: Properties 1 through 3 from Appendix A can be modified slightly. Namely, as $d \to \infty$,

1*. For any $0 < q < \infty, \gamma > 0, \lambda > 0$,
$$\mathbf{P}(\mathrm{GN}_{\lambda,\gamma}(0) > \lambda(q \log d)^{\frac{1}{\gamma}}) = O\left((\log d)^{\frac{1}{\gamma}-1} d^{-q}\right).$$

2*. For any $0 < r < q < \infty, \gamma > 0, \lambda > 0$,
$$\mathbf{P}(\mathrm{GN}_{\lambda,\gamma}(\lambda(r \log d)^{\frac{1}{\gamma}}) > \lambda(q \log d)^{\frac{1}{\gamma}}) = O\left((\log d)^{\frac{1}{\gamma}-1} d^{-\left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^{\gamma}}\right).$$

3*. For any $0 < q < r < \infty, \gamma > 0$,
$$\mathbf{P}(\mathrm{GN}_{\lambda,\gamma}(\lambda(r \log d)^{\frac{1}{\gamma}}) \leq \lambda(q \log d)^{\frac{1}{\gamma}}) = O\left((\log d)^{\frac{1}{\gamma}-1} d^{-\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^{\gamma}}\right).$$

Given the above properties, the corollary is proven by repeating the proofs of Theorems 2 through 5. □

Remark 6: We visually represent the results of Theorems 2 through 5 by plotting a two-dimensional phase diagram of (β, r). Figure 2.2 and Figure 2.3 on pages 31 and 32 display the selection regions for a few choices of γ. As can be seen, the almost full selection boundary $r = \phi_2(\beta, \gamma)$ does not depend on the shape parameter γ; however, the exact selection boundary $r = \phi_1(\beta, \gamma)$ does.

Remark 7: As seen from Theorems 2 through 5 and Corollary 2, for $r > 2^\gamma$ and $0 < \beta < 1$, exact variable selection is trivially possible. Thus, the interesting range for the values of r is $(0, 2^\gamma)$.

2.4 Adaptation to unknown sparsity

In many situations, the sparsity index β is unknown. Observe that the selector $\tilde{\boldsymbol{\eta}}$ in Theorem 4 is reliant on knowledge of β. We now propose a selector which is adapted to unknown β. To construct this selector, we shall act similar to Lepski's method of adaptive estimation (see Lepski [11]). Before introducing the method and proposed selector, further notation must be introduced. Let $M = M_d \ll d$ be a sequence whereby $M \to \infty$ and $\frac{\log d}{M} \to 0$ as $d \to \infty$. Furthermore, consider a positive real number B which can be arbitrarily close to 1 from the left and a sequence of equidistant grid points on $(0, B] \subset (0, 1)$, denoted by $0 < \beta_1 < \beta_2 < \ldots < \beta_M \leq B$, with $\beta_m = m\Delta, m = 1, \ldots, M$, and $\Delta = \Delta_d = \frac{B}{M}$. Observe that by the choice of M, it is obvious that the sequence d^Δ is bounded from above by a finite constant. Thus, let

$$K_{max} \stackrel{\text{def}}{=} \sup_d d^\Delta < \infty.$$

Additionally, define the selector $\tilde{\boldsymbol{\eta}}(\beta_m) = (\tilde{\eta}_1(\beta_m), \ldots, \tilde{\eta}_d(\beta_m))$, where

$$\tilde{\eta}_j(\beta_m) = \mathbb{1}(X_j > ((\gamma\beta_m + \delta)\log d)^{1/\gamma}), \quad j = 1, \ldots, d, \quad m = 1, \ldots, M,$$

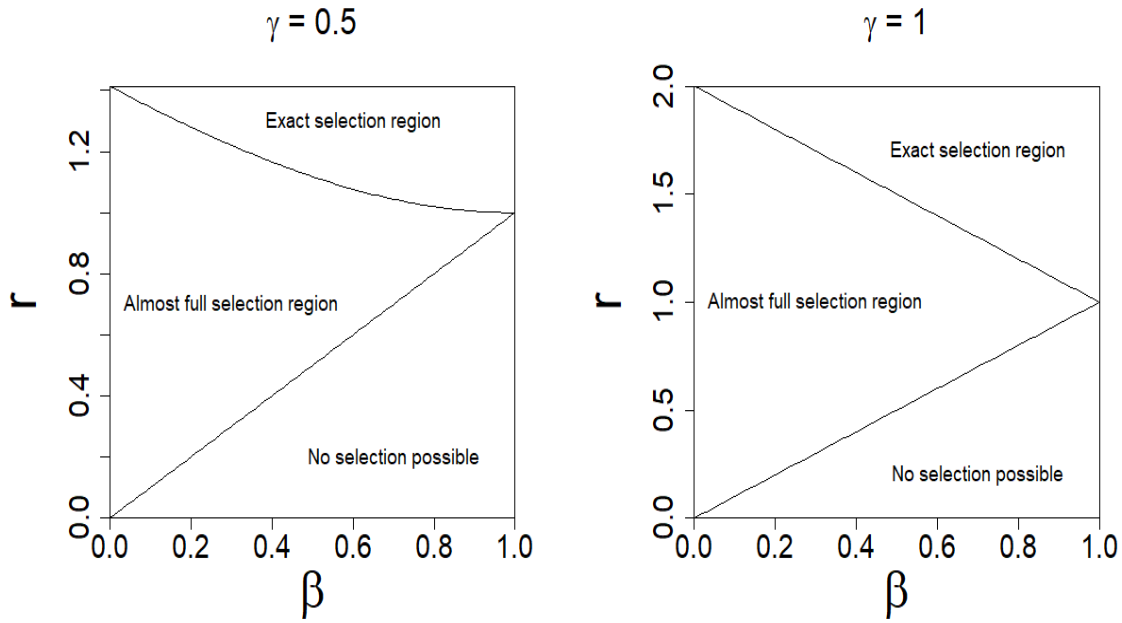

Figure 2.2: *Variable selection regions for* $\gamma = 0.5, 1$.

and the random index

$$\tilde{m} = \max\left\{m \in \{1, \ldots, M\} : |\tilde{\eta}(\beta_m) - \tilde{\eta}(\beta_j)| \leq V_j \text{ for all } j \leq m\right\},$$

where $\tilde{m} = 0$ if the set above is empty, and

a) $\delta = \delta(d) > 0$ is such that $\delta \to 0$ and $\frac{\delta \log d}{\log \log d} \to \infty$ as $d \to \infty$,

b) $V_j = V_{j,d} = d^{1-\beta_j}/\tau_d$, and

c) $\tau_d \to \infty$ with $\tau_d = o\left(d^{\delta/\gamma}(\log d)^{1-\frac{1}{\gamma}}\right)$ as $d \to \infty$.

We are now ready to propose a selector adapted to unknown β. For the proposed selector, we shall assume that $\beta \notin (B, 1)$. This is the price that must be paid for adaptive recovery of the sparsity pattern. Consider the selector $\tilde{\boldsymbol{\eta}}^{\text{ad}} = (\tilde{\eta}_1^{\text{ad}}, \ldots, \tilde{\eta}_d^{\text{ad}})$, where

$$\tilde{\eta}_j^{\text{ad}} = \tilde{\eta}_j(\beta_{\tilde{m}}) = \mathbb{1}(X_j > ((\gamma\beta_{\tilde{m}} + \delta)\log d)^{1/\gamma}), \quad j = 1, \ldots, d.$$

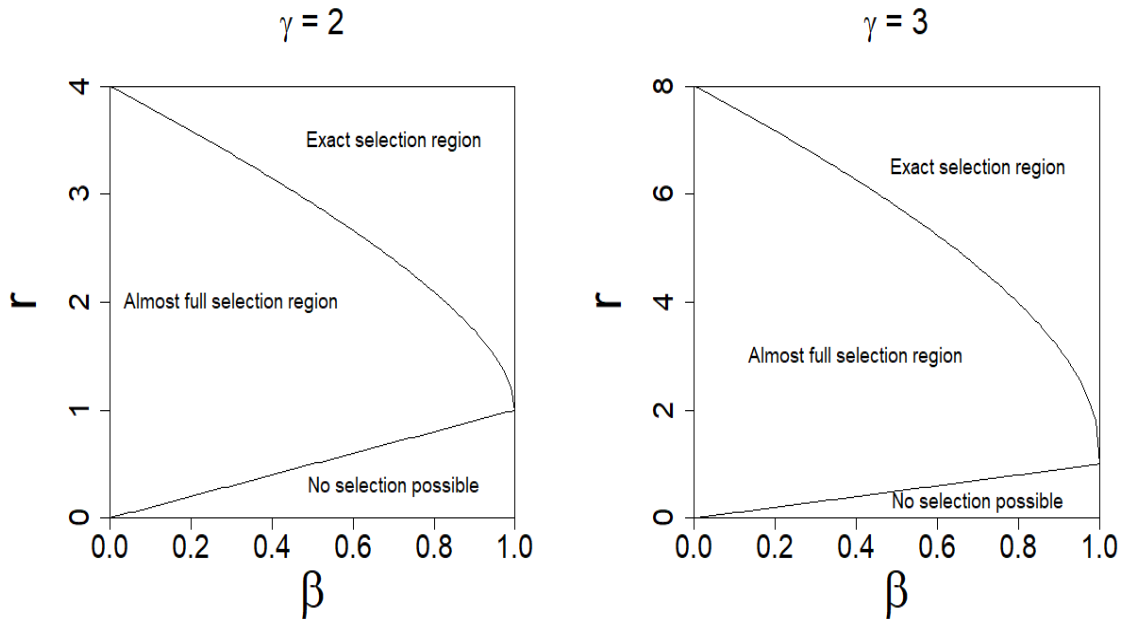

Figure 2.3: *Variable selection regions for* $\gamma = 2, 3$.

We now claim that the selector $\tilde{\boldsymbol{\eta}}^{\mathrm{ad}}$ of $\boldsymbol{\eta}$ in model (1.1) achieves almost full selection.

Theorem 6. *For* $\mu_d = (\gamma r \log d)^{1/\gamma}$ *with* $r > \beta, \epsilon \sim GN_\gamma(0), \boldsymbol{\eta} \in H_{d,\beta}$, *and* $\beta \in (0, B] \subset (0, 1)$ *in model* (1.1), *the selector* $\tilde{\boldsymbol{\eta}}^{\mathrm{ad}}$ *achieves almost full selection. That is,*

$$\limsup_{d\to\infty} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}| = o(1). \tag{2.11}$$

Proof: By the law of total expectation,

$$\begin{aligned}\sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}| &\le \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}\left(|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}|\Big|\tilde{m} \ge m_0\right)\mathbf{P}(\tilde{m} \ge m_0) \\ &+ \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}\left(|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}|\Big|\tilde{m} < m_0\right)\mathbf{P}(\tilde{m} < m_0) \stackrel{\mathrm{def}}{=} S_d^{(1)} + S_d^{(2)},\end{aligned} \tag{2.12}$$

where $m_0 \in \{1, \ldots, M-1\}$ is such that $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$. Let us first consider $S_d^{(1)}$. By the triangle inequality, using the definition of $\tilde{m}$,

$$|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}| = |\tilde{\boldsymbol{\eta}}^{\mathrm{ad}}(\beta_{\tilde{m}}) - \boldsymbol{\eta}| \leq |\tilde{\boldsymbol{\eta}}^{\mathrm{ad}}(\beta_{\tilde{m}}) - \tilde{\boldsymbol{\eta}}(\beta_{m_0})| + |\tilde{\boldsymbol{\eta}}(\beta_{m_0}) - \boldsymbol{\eta}|$$
$$\leq V_{m_0} + |\tilde{\boldsymbol{\eta}}(\beta_{m_0}) - \boldsymbol{\eta}|.$$

Recalling that $V_{m_0} = d^{1-\beta_{m_0}}/\tau_d$, we have

$$d^{\beta-1}V_{m_0} = \frac{d^{\beta-\beta_{m_0}}}{\tau_d} \leq \frac{d^{\Delta}}{\tau_d} \leq \frac{K_{\max}}{\tau_d} = O(\tau_d^{-1}) = o(1) \text{ as } d \to \infty.$$

Additionally, for any non-negative random variable $Y, \mathbb{E}(Y|B)P(B) \leq \mathbb{E}(Y)$. Thus, as $d \to \infty$

$$\begin{aligned}
S_d^{(1)} &\leq \sup_{\boldsymbol{\eta} \in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}}(\beta_{m_0}) - \boldsymbol{\eta}| + O(\tau_d^{-1}) \\
&= \sup_{\boldsymbol{\eta} \in H_{d,\beta}} d^{\beta-1} \sum_{j=1}^{d} \mathbb{E}_{\eta_j}|\tilde{\eta}_j(\beta_{m_0}) - \eta_j| + O(\tau_d^{-1}) \\
&\leq \sup_{\boldsymbol{\eta} \in H_{d,\beta}} d^{\beta-1} \Bigg\{ \sum_{j:\eta_j=0} \mathbf{P}_0\Big(X_j > \Big((\gamma\beta_{m_0} + \delta)\log d\Big)^{1/\gamma}\Big) \\
&+ \sum_{j:\eta_j=1} \mathbf{P}_{\mu_d}\Big(X_j \leq \Big((\gamma\beta_{m_0} + \delta)\log d)^{1/\gamma}\Big)\Big) \Bigg\} + O(\tau_d^{-1}) \\
&\leq d^{\beta}\mathbf{P}_0\Big(X_1 > \Big((\gamma\beta_{m_0} + \delta)\log d\Big)^{1/\gamma}\Big) \\
&+ C_1\mathbf{P}_{\mu_d}\Big(X_1 \leq \Big((\gamma\beta_{m_0} + \delta)\log d\Big)^{1/\gamma}\Big) + O(\tau_d^{-1}) \\
&\stackrel{\mathrm{def}}{=} K_d^{(1)} + K_d^{(2)} + O(\tau_d^{-1}).
\end{aligned}$$

By Property 1 in Appendix A and using condition a) on δ, we obtain

$$\begin{aligned}
K_d^{(1)} &\leq B_{1,\gamma}^* d^{\beta-\beta_{m_0}-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1} \leq B_{1,\gamma}^* d^{\Delta-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1} \\
&\leq B_{1,\gamma}^* K_{\max} d^{-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1} = o(1).
\end{aligned}$$

Turning our attention to $K_d^{(2)}$ and using Property 3 from Appendix A, for all large

enough d,

$$K_d^{(2)} \leq C_1 B_{3,\gamma}^* d^{-\left(r^{1/\gamma}-(\beta_{m_0}+\delta/\gamma)^{1/\gamma}\right)^\gamma} (\log d)^{\frac{1}{\gamma}-1}.$$

Since for all large enough d we have $r > \beta_{m_0} + \delta/\gamma$, it follows that as $d \to \infty$,

$$K_d^{(2)} = o(1).$$

Combining this all together, it may be concluded that as $d \to \infty$,

$$S_d^{(1)} = o(1).$$

Let us now consider the term $S_d^{(2)}$ on the right hand side of (2.12). We have

$$S_d^{(2)} = \sup_{\boldsymbol{\eta} \in H_{d,\beta}} d^{\beta-1} \mathbb{E}_{\boldsymbol{\eta}} \left(|\tilde{\boldsymbol{\eta}}(\beta_{\tilde{m}}) - \boldsymbol{\eta}| \Big| \tilde{m} < m_0 \right) \mathbf{P}(\tilde{m} < m_0) \leq d^\beta \mathbf{P}(\tilde{m} < m_0). \quad (2.13)$$

Note that, by definition of the random index $\tilde{m}$,

$$\begin{aligned}
&\mathbf{P}(\tilde{m} < m_0) = \sum_{k=1}^{m_0-1} \mathbf{P}(\tilde{m} = k) \leq \sum_{k=1}^{m_0-1} \mathbf{P}\left(\exists j \in \{1, \ldots, k\} : |\tilde{\boldsymbol{\eta}}(\beta_{k+1}) - \tilde{\boldsymbol{\eta}}(\beta_j)| > V_j \right) \\
&\leq \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(|\tilde{\boldsymbol{\eta}}(\beta_{k+1}) - \tilde{\boldsymbol{\eta}}(\beta_j)| > V_j \right) = \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{d} |\tilde{\eta}_i(\beta_{k+1}) - \tilde{\eta}_i(\beta_j)| > V_j \right) \\
&= \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{d} \mathbb{1}\left\{ \left((\gamma\beta_j + \delta)\log d\right)^{1/\gamma} < X_i \leq \left((\gamma\beta_{k+1} + \delta)\log d\right)^{1/\gamma} \right\} > V_j \right).
\end{aligned}$$

Defining the event
$A_i = A_{i,j,k+1,d} = \left\{ \left((\gamma\beta_j + \delta)\log d\right)^{1/\gamma} < X_i \leq \left((\gamma\beta_{k+1} + \delta)\log d\right)^{1/\gamma} \right\}$ gives us

$$\mathbf{P}(\tilde{m} < m_0) \leq \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{d} \mathbb{1}(A_i) > V_j \right) = \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{d} W_i > V_j - \sum_{i=1}^{d} \mathbf{P}(A_i) \right), \quad (2.14)$$

where $W_i \stackrel{\text{def}}{=} W_{i,j,k+1,d} = \mathbb{1}(A_i) - \mathbf{P}(A_i)$. Noting that $\mathbb{E}(W_i) = 0$ and $|W_i| \leq 2$, we can proceed by applying Bernstein's second inequality with $H = 2/3$ for the case $t > D_d^2/H$ (see Appendix A). Before doing so, we show that as $d \to \infty$,

$$\sum_{i=1}^{d} \mathbf{P}(A_i) = \sum_{i=1}^{d} \mathbf{P}(A_{i,j,k+1,d}) = o(V_j).$$

To this end, as $d \to \infty$,

$$\begin{aligned}
\sum_{i=1}^{d} \mathbf{P}(A_i) &= \sum_{i:\eta_i=0} \mathbf{P}(A_i) + \sum_{i:\eta_i=1} \mathbf{P}(A_i) \\
&\leq d\mathbf{P}\left(\mathrm{GN}_\gamma(0) > \left((\gamma\beta_j + \delta)\log d\right)^{1/\gamma}\right) \\
&+ C_1 d^{1-\beta}\mathbf{P}\left(\mathrm{GN}_\gamma\left((\gamma r \log d)^{1/\gamma}\right) \leq \left((\gamma\beta_{k+1} + \delta)\log d\right)^{1/\gamma}\right) \\
&= d\mathbf{P}\left(\mathrm{GN}_\gamma(0) > \left(\gamma\left(\beta_j + \delta/\gamma\right)\log d\right)^{1/\gamma}\right) \\
&\quad + C_1 d^{1-\beta}\mathbf{P}\left(\mathrm{GN}_\gamma\left((\gamma r \log d)^{1/\gamma}\right) \leq \left(\gamma\left(\beta_{k+1} + \delta/\gamma\right)\log d\right)^{1/\gamma}\right) \\
&= dB^*_{1,\gamma}\left(\log d\right)^{\frac{1}{\gamma}-1} d^{-\left(\beta_j + \frac{\delta}{\gamma}\right)}(1 + o(1)) \\
&\quad + C_1 d^{1-\beta} B^*_{3,\gamma}\left(\log d\right)^{\frac{1}{\gamma}-1} d^{-\left(r^{\frac{1}{\gamma}} - \left(\beta_{k+1} + \frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^\gamma}(1 + o(1)) \\
&= O\left(d^{1-\beta_j - \frac{\delta}{\gamma}}\left(\log d\right)^{\frac{1}{\gamma}-1}\right) + O\left(d^{1-\beta-\left(r^{\frac{1}{\gamma}} - \left(\beta_{k+1} + \frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^\gamma}\right), \qquad (2.15)
\end{aligned}$$

where the second last equality is due to Properties 1 and 3 from Appendix A. To continue further, observe that by the choice of m_0, for all $j = 1, \ldots, k$, $k = 1, \ldots, m_0 - 1$,

$$\beta_j \leq \beta_{m_0-1} < \beta < r,$$

and also

$$\beta_{k+1} \leq \beta_{m_0} < \beta < r.$$

Therefore, for d sufficiently large,

$$d^{1-\left(\beta_j+\frac{\delta}{\gamma}\right)}(\log d)^{\frac{1}{\gamma}-1} \gg d^{1-\beta-\left(r^{\frac{1}{\gamma}}-\left(\beta_{k+1}+\frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^{\gamma}},$$

and it follows from (2.15) that

$$\begin{aligned}\sum_{i=1}^{d}\mathbf{P}(A_i) &= O\left(d^{1-\beta_j-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1}\right)+O\left(d^{1-\beta-\left(r^{\frac{1}{\gamma}}-\left(\beta_{k+1}+\frac{\delta}{\gamma}\right)^{\frac{1}{\gamma}}\right)^{\gamma}}\right)\\ &= O\left(d^{1-\beta_j-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1}\right).\end{aligned}$$

Finally, recalling that $\tau_d \ll d^{\delta/\gamma}(\log d)^{1-\frac{1}{\gamma}}$, as $d\to\infty$,

$$\sum_{i=1}^{d}\mathbf{P}(A_i) = O\Big(d^{1-\beta_j-\frac{\delta}{\gamma}}(\log d)^{\frac{1}{\gamma}-1}\Big) = o(V_j) = o\left(\frac{d^{1-\beta_j}}{\tau_d}\right). \tag{2.16}$$

Returning now to $\mathbf{P}(\tilde{m}<m_0)$ in (2.14) and using (2.16), as $d\to\infty$,

$$\begin{aligned}\mathbf{P}(\tilde{m}<m_0) &\le \sum_{k=1}^{m_0-1}\sum_{j=1}^{k}\mathbf{P}\left(\sum_{i=1}^{d}W_i > V_j-\sum_{i=1}^{d}\mathbf{P}(A_i)\right)\\ &= \sum_{k=1}^{m_0-1}\sum_{j=1}^{k}\mathbf{P}\left(\sum_{i=1}^{d}W_i > V_j(1+o(1))\right).\end{aligned}$$

As $d\to\infty$, applying Bernstein's second inequality from Appendix A with $H=2/3$ and

$$t = V_j \gg \sum_{i=1}^{d}\mathbf{P}(A_i) > \frac{\sum_{i=1}^{d}\mathbb{E}(W_i^2)}{H}$$

yields

$$\begin{aligned}\sum_{k=1}^{m_0-1}\sum_{j=1}^{k}\mathbf{P}\left(\sum_{i=1}^{d}W_i > V_j(1+o(1))\right) &\le \sum_{k=1}^{m_0-1}\sum_{j=1}^{k}\exp\left(-\frac{V_j(1+o(1))}{4(2/3)}\right)\\ &\le M^2\exp\left(-\frac{3d^{1-\beta_{m_0-1}}}{8\tau_d}(1+o(1))\right).\end{aligned}$$

Thus, for all large enough d,

$$S_d^{(2)} \leq d^{\beta}\mathbf{P}(\tilde{m} < m_0) \leq d^{\beta} M^2 \exp\left(-\frac{3d^{1-\beta_{m_0-1}}}{8\tau_d}\right).$$

Since, by assumption, $1-\beta_{m_0-1}$ is bounded away from zero, $M \ll d$, and $\tau_d \to \infty$ at a rate slower than $d^{1-\beta_{m_0-1}}$, it follows from above that

$$S_d^{(2)} = o(1), \quad \text{as } d \to \infty.$$

Combining everything together, we may conclude that

$$\limsup_{d\to\infty} \sup_{\boldsymbol{\eta}\in H_{d,\beta}} d^{\beta-1}\mathbb{E}_{\boldsymbol{\eta}}|\tilde{\boldsymbol{\eta}}^{\mathrm{ad}} - \boldsymbol{\eta}| = o(1),$$

for all $0 < \beta \leq B < 1, r > \beta$. The proof of Theorem 6 is complete. □

Chapter 3

Simulations

To showcase the exact and almost full selection procedures introduced in this thesis, we conduct three simulations. In the first simulation, we demonstrate the performance of the exact selector from Theorem 2. In the second simulation, we illustrate the behaviour of the almost full selector from Theorem 4. In the final simulation, we exhibit the performance of the adaptive almost full selector from Theorem 6. Rather than presenting the number of incorrect selections $\sum_{i=j}^{d} |\hat{\eta}_j - \eta_j|$, we standardize the risk by showing the proportion $\sum_{i=j}^{d} |\hat{\eta}_j - \eta_j|/d$ instead. The estimated risk values are averaged over 500 simulation cycles.

Simulation 1:

1. Choose some $\gamma \in (0, \infty)$, a large positive integer d, and some $\beta \in (0, 1)$. For $i = 1, 2, \ldots, 500$:

   (a) Let $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$ be such that

$$X_j^{(i)} \sim \eta_j^{(i)} \mathrm{GN}_\gamma(\mu) + (1 - \eta_j^{(i)}) \mathrm{GN}_\gamma(0), \quad j = 1, \ldots, d,$$

   where $\eta_j^{(i)} \sim \text{Bernoulli}(d^{-\beta}), \mu = \mu_d = (\gamma r \log d)^{1/\gamma}$, and

$r = (1 + (1-\beta)^{\frac{1}{\gamma}})^{\gamma} + 1$.

(b) Let $\hat{\boldsymbol{\eta}}^{(i)} = (\hat{\eta}_1^{(i)}, \ldots, \hat{\eta}_d^{(i)})$, where $\hat{\eta}_j^{(i)} = \mathbb{1}(X_j^{(i)} > ((\gamma+\delta)\log d)^{1/\gamma})$, $j = 1, 2, \ldots, d$, and $\delta = (\log d)^{-0.9}$.

(c) Compute $R^{(i)} = \sum_{j=1}^{d} |\hat{\eta}_j^{(i)} - \eta_j^{(i)}|/d$.

2. Compute $\bar{R} = \sum_{i=1}^{500} R^{(i)}/500$.

3. Repeat steps 1 and 2 for many choices of γ, d, and β specified by Theorem 2.

Simulation 2:

1. Choose some $\gamma \in (0, \infty)$, a large positive integer d, and some $\beta \in (0, 1)$. For $i = 1, 2, \ldots, 500$:

(a) Let $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$ be such that

$$X_j^{(i)} \sim \eta_j^{(i)} \mathrm{GN}_\gamma(\mu) + (1 - \eta_j^{(i)}) \mathrm{GN}_\gamma(0), \quad j = 1, \ldots, d,$$

where $\eta_j^{(i)} \sim \text{Bernoulli}(d^{-\beta})$, $\mu = \mu_d = (\gamma r \log d)^{1/\gamma}$, and $r = \frac{(1+(1-\beta)^{\frac{1}{\gamma}})^{\gamma}+\beta}{2}$.

(b) Let $\tilde{\boldsymbol{\eta}}^{(i)} = (\tilde{\eta}_1^{(i)}, \ldots, \tilde{\eta}_d^{(i)})$, where $\tilde{\eta}_j^{(i)} = \mathbb{1}(X_j^{(i)} > ((\gamma\beta+\delta)\log d)^{1/\gamma})$, $j = 1, 2, \ldots, d$, and $\delta = (\log d)^{-0.9}$.

(c) Compute $R^{(i)} = \sum_{j=1}^{d} |\tilde{\eta}_j^{(i)} - \eta_j^{(i)}|/d$.

2. Compute $\bar{R} = \sum_{i=1}^{500} R^{(i)}/500$.

3. Repeat steps 1 and 2 for many choices of γ, d, and β specified by Theorem 4.

Simulation 3:

1. Choose some $\gamma \in (0, \infty)$, a large positive integer d, and some $\beta \in (0, 1)$. For $i = 1, 2, \ldots, 500$:

(a) Let $\boldsymbol{X}^{(i)} = (X_1^{(i)}, \ldots, X_d^{(i)})$ be such that

$$X_j^{(i)} \sim \eta_j^{(i)} \mathrm{GN}_\gamma(\mu) + (1 - \eta_j^{(i)}) \mathrm{GN}_\gamma(0), \quad j = 1, \ldots, d,$$

where $\eta_j^{(i)} \sim \text{Bernoulli}(d^{-\beta})$, $\mu = \mu_d = (\gamma r \log d)^{1/\gamma}$, and $r = \frac{(1+(1-\beta)^{\frac{1}{\gamma}})^\gamma + \beta}{2}$.

(b) Let $\tilde{\boldsymbol{\eta}}^{\mathrm{ad}(i)} = (\tilde{\eta}_1^{\mathrm{ad}(i)}, \ldots, \tilde{\eta}_d^{\mathrm{ad}(i)})$, where $\tilde{\eta}_j^{\mathrm{ad}(i)} = \mathbb{1}(X_j^{(i)} > ((\gamma \beta_{\tilde{m}}^{(i)} + \delta) \log d)^{1/\gamma})$, $j = 1, 2, \ldots, d$, $\beta_{\tilde{m}}^{(i)}$ is the adaptive estimator of β defined in Theorem 6, $M = 100$, and $\delta = (\log d)^{-0.9}$.

(c) Compute $R^{(i)} = \sum_{j=1}^d |\tilde{\eta}_j^{\mathrm{ad}(i)} - \eta_j^{(i)}|/d$.

2. Compute $\bar{R} = \sum_{i=1}^{500} R^{(i)}/500$ and $\bar{\beta}_{\tilde{m}} = \sum_{i=1}^{500} \beta_{\tilde{m}}^{(i)}/500$.

3. Repeat steps 1 and 2 for many choices of γ, d, and β specified by Theorem 6.

Remark 8: At a minimum, the proportion of incorrect selections for a reasonable selector should at least be less than $d^{-\beta}$. Assuming $\eta_i \sim \text{Bernoulli}(d^{-\beta})$, it is always possible to choose a selector with no discriminatory power for which the proportion of incorrect selections is approximately $d^{-\beta}$. To see this, consider a selector $\hat{\boldsymbol{\eta}}_B = (\hat{\eta}_{B,1}, \ldots, \hat{\eta}_{B,d})$, where $\hat{\eta}_{B,j} = 0$ for $j = 1, 2, \ldots, d$. Clearly,

$$\sum_{i=1}^d |\hat{\eta}_{B,j} - \eta_j| = \sum_{i=j}^d \eta_j \sim \text{Binomial}(d, d^{-\beta}),$$

and by the weak law of large numbers,

$$\frac{\sum_{j=1}^d \eta_j}{d} - d^{-\beta} \xrightarrow{P} 0.$$

Thus, if $\eta_j \sim \text{Bernoulli}(d^{-\beta})$ for $j = 1, 2, \ldots, d$, then the proportion of incorrect selections for a selector $\hat{\boldsymbol{\eta}}_B = (\hat{\eta}_{B,1}, \ldots, \hat{\eta}_{B,d})$ is on average $d^{-\beta}$.

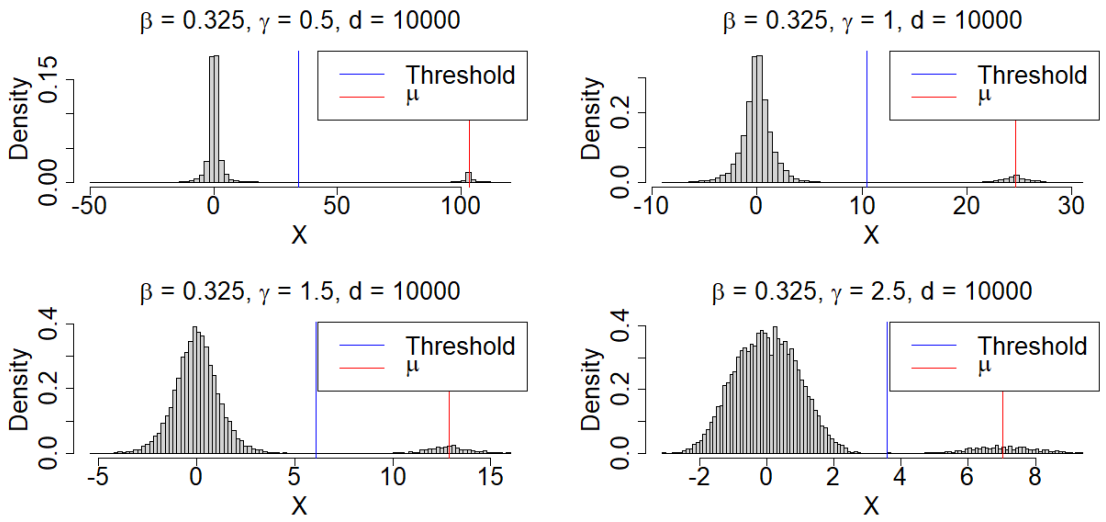

Figure 3.1: *Histograms of generalized normal mixture data from Simulation 1 for various choices of* γ*, with* $d = 10000, \beta = 0.325$*. The selector thresholds and signals are also overlaid. Note that the choice of* $\beta = 0.325$ *makes the selection problem more difficult as compared to the choice of* β *near 0.*

3.1 Comments on simulation results

The results of each simulation are presented in a tabular form on pages 44 to 46. We have also plotted some examples of generalized normal mixture data in Figures 3.1 and 3.2. Viewing Tables 3.1, 3.2, and 3.3, it can be seen that the selectors perform very well in each simulation. For nearly every choice of d, γ, β, and r, the quantity $\bar{R}$ subceeds $d^{-\beta}$. The only case for which it is evident that the selectors do not perform well is when $d = 1000, \beta = 0.667$, and $\gamma = 0.5$ in Simulations 2 and 3. In this case, we see that the quantity $d^{-\beta}$ is 0.01, while the average risk $\bar{R}$ is approximately 0.01297 in Simulation 2, and approximately 0.01247 in Simulation 3. For the case $d = 1000, \beta = 0.667$, and $\gamma = 0.5$, the poor performance of the selectors is likely due to the high level of sparsity, lower dimensionality, and larger kurtosis of the noise distribution (recall that the kurtosis of the generalized normal distribution is larger for smaller γ). Comparing the adaptive selector in Table 3.3 to

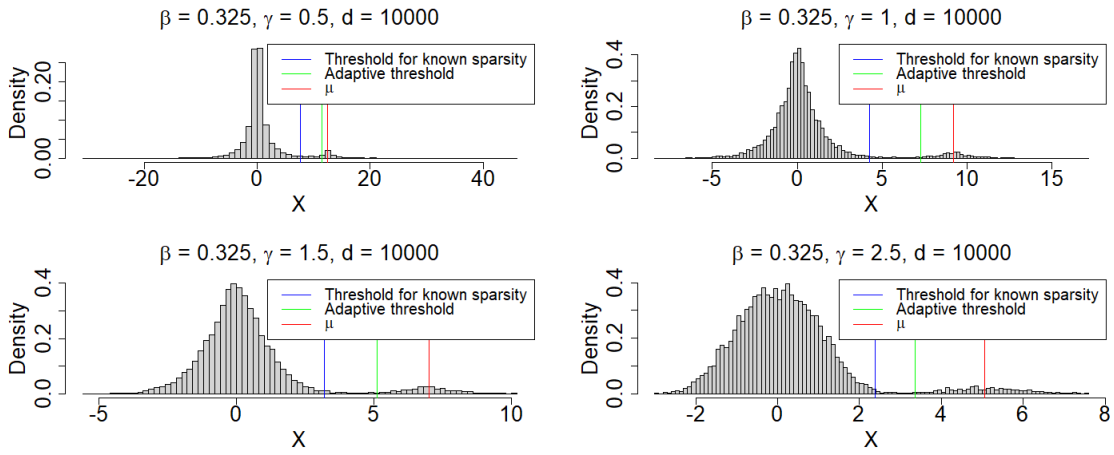

Figure 3.2: *Histograms of generalized normal mixture data from Simulation 3 for various choices of* γ*, with* $d = 10000, \beta = 0.325$*. The non-adaptive and adaptive thresholds, along with signals, are also overlaid.*

the non-adaptive almost full selector in Table 3.2, their does appear to be a minor drop in performance for some choices of d, β, and γ. For example, in the case that $d = 1000, \beta = 0.349$, and $\gamma = 0.5$, it can be seen that $\bar{R}$ is roughly equal to 0.03695 for the non-adaptive almost full selector and $\bar{R}$ is roughly equal to 0.05568 for the adaptive almost full selector. Nonetheless, since the average risk of the non-adaptive selector is nearly always lower than $d^{-\beta}$, this drop in performance is manageable. Viewing the quantity $\bar{\beta}_{\tilde{m}}$ in Table 3.3, we can see that the adaptive estimator of β nearly always exceeds the sparsity index β. This is what we should expect, given the manner in which the adaptive estimator of β was constructed.[1] We can also see that when d is increased with γ and $d^{-\beta}$ held fixed, the value $\bar{R}$ decreases significantly in each simulation. In the exact selection regimen, the value $\bar{R}$ appears to decrease by roughly a factor of 10 when d is increased from 1000 to 10000. In the almost full selection regimen, the value $\bar{R}$ decreases at a slower rate than in the exact selection regimen when d is increased, but it is still clear that the risk is decreasing. For

[1]To see why $\beta_{\tilde{m}}$ should nearly always be greater than β, the reader may re-examine the proof of Theorem 6. Namely, it can be shown that the quantity $\mathbf{P}(\tilde{m} < m_0) = o(1)$ as $d \to \infty$, where m_0 is chosen to have $\beta_{m_0} < \beta \leq \beta_{m_0+1}$.

example, in Simulation 2, with $d = 1000, d^{-\beta} = 0.09$, $\gamma = 0.5$, we have $\bar{R} \approx 0.03695$, and with $d = 10000, d^{-\beta} = 0.09, \gamma = 0.5$, we obtain $\bar{R} \approx 0.02194$. Lastly, observe that even after rescaling the average risk $\bar{R}$ to the average Hamming distance $d\bar{R}$, the strong performance of the selectors in each simulation is quite evident. For instance, with $d = 10000, \beta = 0.325$, and $\gamma = 0.5$, we have $d\bar{R} = 0.51$ in Simulation 1, $d\bar{R} = 143.5$ in Simulation 2, and $d\bar{R} = 144.8$ in Simulation 3. These values are clearly much lower than the expected number of signal components, which is given by $d^{1-\beta} = (10000)^{1-0.325} = 501.2$ for $d = 10000$ and $\beta = 0.325$.

Table 3.1: *Performance of the exact selector defined in Simulation 1.*

d	β	$d^{-\beta}$	γ	r	μ_d	$\phi_1(\beta,\gamma)$	$((\gamma+\delta)\log d)^{1/\gamma}$	$\bar{R}$
1000	0.349	0.09	0.5	2.19	57.39	1.19	21.78	0.000408
10000	0.261	0.09	0.5	2.24	106.71	1.24	34.27	0.000045
1000	0.434	0.05	0.5	2.15	55.10	1.15	21.78	0.000498
10000	0.325	0.05	0.5	2.21	103.24	1.21	34.27	0.000051
1000	0.667	0.01	0.5	2.05	50.33	1.05	21.78	0.000456
10000	0.500	0.01	0.5	2.12	95.14	1.12	34.27	0.000048
1000	0.349	0.09	1	2.65	18.32	1.65	8.12	0.000146
10000	0.261	0.09	1	2.74	25.22	1.74	10.46	0.000012
1000	0.434	0.05	1	2.57	17.73	1.57	8.12	0.000164
10000	0.325	0.05	1	2.67	24.64	1.67	10.46	0.000011
1000	0.667	0.01	1	2.33	16.12	1.33	8.12	0.000144
10000	0.500	0.01	1	2.50	23.03	1.50	10.46	0.000015
1000	0.349	0.09	1.5	3.32	10.57	2.32	5.12	0.000090
10000	0.261	0.09	1.5	3.45	13.14	2.45	6.10	0.000007
1000	0.434	0.05	1.5	3.19	10.29	2.19	5.12	0.000092
10000	0.325	0.05	1.5	3.35	12.90	2.35	6.10	0.000007
1000	0.667	0.01	1.5	2.80	9.45	1.80	5.12	0.000104
10000	0.500	0.01	1.5	3.08	12.19	2.08	6.10	0.000007
1000	0.349	0.09	2	4.27	7.68	3.27	3.88	0.000048
10000	0.261	0.09	2	4.46	9.06	3.46	4.44	0.000005
1000	0.434	0.05	2	4.07	7.50	3.07	3.88	0.000062
10000	0.325	0.05	2	4.32	8.92	3.32	4.44	0.000004
1000	0.667	0.01	2	3.49	6.94	2.49	3.88	0.000092
10000	0.500	0.01	2	3.91	8.49	2.91	4.44	0.000005
1000	0.349	0.09	2.5	5.61	6.23	4.61	3.21	0.000032
10000	0.261	0.09	2.5	5.88	7.12	4.88	3.58	0.000004
1000	0.434	0.05	2.5	5.33	6.10	4.33	3.21	0.000046
10000	0.325	0.05	2.5	5.68	7.03	4.68	3.58	0.000003
1000	0.667	0.01	2.5	4.47	5.69	3.47	3.21	0.000042
10000	0.500	0.01	2.5	5.10	6.73	4.10	3.58	0.000003

Table 3.2: *Performance of the almost full selector defined in Simulation 2.*

d	β	$d^{-\beta}$	γ	r	μ_d	$((\gamma\beta+\delta)\log d)^{1/\gamma}$	$\bar{R}$
1000	0.349	0.09	0.5	0.77	7.09	5.84	0.03695
10000	0.261	0.09	0.5	0.75	12.00	6.02	0.02194
1000	0.434	0.05	0.5	0.79	7.47	7.35	0.03468
10000	0.325	0.05	0.5	0.77	12.44	7.54	0.01435
1000	0.667	0.01	0.5	0.86	8.83	12.36	0.01297
10000	0.500	0.01	0.5	0.81	13.88	12.61	0.00497
1000	0.349	0.09	1	1.00	6.91	3.62	0.01365
10000	0.261	0.09	1	1.00	9.21	3.66	0.01193
1000	0.434	0.05	1	1.00	6.91	4.21	0.00878
10000	0.325	0.05	1	1.00	9.21	4.24	0.00698
1000	0.667	0.01	1	1.00	6.91	5.82	0.00328
10000	0.500	0.01	1	1.00	9.21	5.85	0.00158
1000	0.349	0.09	1.5	1.33	5.76	2.86	0.00917
10000	0.261	0.09	1.5	1.36	7.05	2.87	0.00826
1000	0.434	0.05	1.5	1.31	5.69	3.19	0.00559
10000	0.325	0.05	1.5	1.34	7.00	3.21	0.00457
1000	0.667	0.01	1.5	1.23	5.47	4.04	0.00173
10000	0.500	0.01	1.5	1.29	6.82	4.05	0.00097
1000	0.349	0.09	2	1.81	5.00	2.46	0.00689
10000	0.261	0.09	2	1.86	5.85	2.46	0.00626
1000	0.434	0.05	2	1.75	4.92	2.68	0.00410
10000	0.325	0.05	2	1.82	5.79	2.69	0.00346
1000	0.667	0.01	2	1.58	4.67	3.23	0.00133
10000	0.500	0.01	2	1.71	5.61	3.23	0.00070
1000	0.349	0.09	2.5	2.48	4.49	2.21	0.00553
10000	0.261	0.09	2.5	2.57	5.12	2.21	0.00507
1000	0.434	0.05	2.5	2.38	4.42	2.38	0.00334
10000	0.325	0.05	2.5	2.50	5.06	2.38	0.00273
1000	0.667	0.01	2.5	2.07	4.18	2.77	0.00113
10000	0.500	0.01	2.5	2.30	4.89	2.77	0.00054

Table 3.3: *Performance of the adaptive almost full selector defined in Simulation 3.*

d	β	$d^{-\beta}$	γ	r	μ_d	$\bar{\beta}_{\tilde{m}}$	$((\gamma\bar{\beta}_{\tilde{m}}+\delta)\log d)^{1/\gamma}$	$\bar{R}$
1000	0.349	0.09	0.5	0.77	7.09	0.417	7.04	0.05568
10000	0.261	0.09	0.5	0.75	12.00	0.438	10.66	0.01965
1000	0.434	0.05	0.5	0.79	7.47	0.460	7.85	0.04301
10000	0.325	0.05	0.5	0.77	12.44	0.463	11.44	0.01448
1000	0.667	0.01	0.5	0.86	8.83	0.764	14.84	0.01247
10000	0.500	0.01	0.5	0.81	13.88	0.531	13.66	0.00640
1000	0.349	0.09	1	1.00	6.91	0.670	5.84	0.01624
10000	0.261	0.09	1	1.00	9.21	0.626	7.02	0.00538
1000	0.434	0.05	1	1.00	6.91	0.709	6.11	0.01181
10000	0.325	0.05	1	1.00	9.21	0.657	7.30	0.00399
1000	0.667	0.01	1	1.00	6.91	0.827	6.93	0.00463
10000	0.500	0.01	1	1.00	9.21	0.743	8.09	0.00176
1000	0.349	0.09	1.5	1.33	5.76	0.787	4.44	0.00966
10000	0.261	0.09	1.5	1.36	7.05	0.717	4.99	0.00330
1000	0.434	0.05	1.5	1.31	5.69	0.836	4.60	0.00685
10000	0.325	0.05	1.5	1.34	7.00	0.747	5.11	0.00243
1000	0.667	0.01	1.5	1.23	5.47	0.945	4.95	0.00260
10000	0.500	0.01	1.5	1.29	6.82	0.840	5.49	0.00106
1000	0.349	0.09	2	1.81	5.00	0.864	3.63	0.00717
10000	0.261	0.09	2	1.86	5.85	0.771	3.93	0.00243
1000	0.434	0.05	2	1.75	4.92	0.913	3.72	0.00508
10000	0.325	0.05	2	1.82	5.79	0.804	4.01	0.00178
1000	0.667	0.01	2	1.58	4.67	0.979	3.84	0.00186
10000	0.500	0.01	2	1.71	5.61	0.902	4.23	0.00077
1000	0.349	0.09	2.5	2.48	4.49	0.920	3.11	0.00558
10000	0.261	0.09	2.5	2.57	5.12	0.814	3.31	0.00194
1000	0.434	0.05	2.5	2.38	4.42	0.954	3.16	0.00383
10000	0.325	0.05	2.5	2.50	5.06	0.849	3.37	0.00146
1000	0.667	0.01	2.5	2.07	4.18	0.990	3.20	0.00146
10000	0.500	0.01	2.5	2.30	4.89	0.941	3.50	0.00060

Chapter 4

Conclusion

In this thesis, we completely determined the regions of exact and almost full variable selection for a sparse sequence model with generalized normal noise. We then constructed an adaptive selector that estimates the components of a sparse vector $\boldsymbol{\eta}$ in model (1.1) when the level of sparsity is unknown. In the simulation section, we went on to demonstrate that the selectors from Theorems 2, 4, and 6 perform very well. To our knowledge, this thesis was the first academic study into the use of generalized normal noise in the context of high-dimensional variable selection. Since the generalized normal distribution and its use in high-dimensional statistics is relatively unstudied, we now propose some areas of future work.

4.1 Areas of future work

It would be useful to expand the results of this thesis to a more general model. In [6], a sparse linear regression model of the form

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\nu} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{Z}$ is an $n \times d$ random matrix, $d \gg n$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^\top \in \mathbb{R}^d$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I)$ was studied. If the components of $\boldsymbol{Z}$ are i.i.d. $\mathcal{N}(0, 1/n)$, where $d = n^{1/\theta}, 0 < \theta < \beta < 1$, and for $i = 1, 2, \ldots, d$,

$$\nu_i = \begin{cases} 0, & \text{with probability } 1 - d^{-\beta}, \\ \sqrt{2r \log d}, & \text{with probability } d^{-\beta}, \text{ where } r > 0, \end{cases} \tag{4.2}$$

then it is achievable to prove when exact and almost full selection are possible in model (4.1). In fact, if the regression coefficients are estimated by means of the lasso estimator or a simple thresholding estimator as in Section 4 of [6], exact selection and almost full selection can each be achieved under certain conditions. In the setup of model (4.1), the Hamming risk is measured by comparing the expected number of positions for which $\operatorname{sgn}(\hat{\boldsymbol{\nu}})$ differs from $\operatorname{sgn}(\boldsymbol{\nu})$, where $\hat{\boldsymbol{\nu}}$ is some estimator of $\boldsymbol{\nu}$ and

$$\operatorname{sgn}(x) = I(x > 0) - I(x < 0), \quad x \in \mathbb{R}.$$

As an extension of this thesis, it would be interesting to see if similar results could be proven but for the components of $\boldsymbol{\varepsilon}$ originating from a generalized normal distribution. In this case, the randomised coefficients $\nu_1, \ldots, \nu_d$ could be such that $\nu_i = 0$ with probability $1 - d^{-\beta}$ and equal to $(\gamma r \log d)^{1/\gamma}, r > 0$ with probability $d^{-\beta}$. The problem would then be to study when exact and almost full variable selection are possible. This could in part be done by showing that a simple thresholding estimator achieves exact or almost full variable selection for certain choices of r.

As another extension to this thesis, it would be useful to construct a selector adapted simultaneously to unknown γ and β. In the case that i.i.d. random variables are distributed as $\mathrm{GN}_{\lambda,\gamma}(\mu)$, estimation of γ can be achieved by means of either the method of maximum likelihood or method of moments (see Varanasi and Aazhang [17]). In practice, it would be quite difficult to show that any selector adapted to both unknown γ and β achieves almost full or exact variable selection. There are

three reasons for this. Firstly, any good estimator $\hat{\gamma}$ of γ has a complicated form and typically must be estimated using an iterative method (see Varanasi and Aazhang [17]). Secondly, the components of $\boldsymbol{X}$ in model (1.1) come from two separate distributions, which complicates the estimation of γ. Lastly, the sparsity index β must also be estimated simultaneously. Instead of assuming both β and γ are unknown, it may be more doable, although quite difficult, to consider the case that β is known and γ is unknown. In fact, in the exact selection regimen, it is not crucial to have knowledge of the sparsity index β. This is because the exact selector $\hat{\boldsymbol{\eta}}$ from Theorem 2 only depends on γ and not β.

The study of minimax selectors in the asymptotic and non-asymptotic case was covered extensively by Butucea et al. [3] for sparse high-dimensional Gaussian sequence models. Thus, it would be interesting to search for selectors of $\boldsymbol{\eta}$ in model (1.1) which satisfy the "minimax" property when the noise components originate from a generalized normal distribution. For the non-asymptotic case, one could search for some lower bound on the minimax risk when d is fixed. For example, a bound of the form

$$\inf_{\tilde{\boldsymbol{\eta}}} R_d(\tilde{\boldsymbol{\eta}}) \geq \psi(d, \beta, \gamma, r), \tag{4.3}$$

where $R_d(\tilde{\boldsymbol{\eta}})$ is defined in (1.2). If the lower bound in (4.3) is known to be sharp, that is, there exists at least one selector for which equality holds in (4.3), then one may proceed by finding a selector with maximal risk equal to $\psi(d, \beta, \gamma, r)$. In practice, it may be easier to find a selector which attains the minimax risk in some asymptotic sense. For instance, one could seek out a selector $\tilde{\boldsymbol{\eta}}_{+,d}$ of $\boldsymbol{\eta}$ for which

$$\lim_{d\to\infty} \frac{R_d(\tilde{\boldsymbol{\eta}}_{+,d})}{\psi(d, \beta, \gamma, r)} = 1, \tag{4.4}$$

where the quantity $\psi(d, \beta, \gamma, r)$ is the same in both (4.3) and (4.4).

Appendix A: Supplementary Properties and Inequalities

Let $\Gamma(s,x)$ be the upper incomplete gamma function, defined as

$$\Gamma(s,x) = \int_x^\infty y^{s-1}e^{-y}dy,$$

$s > 0, x > 0$. Let $\kappa_\gamma = \frac{1}{2\Gamma\left(\frac{1}{\gamma}\right)\gamma^{\frac{1}{\gamma}-1}}$ be the normalizing constant in (1.8). In what follows, we shall use the property that $\Gamma(s,x) \sim x^{s-1}e^{-x}$ as $x \to \infty$ (see equation 6.5.32 of Abramowitz and Stegun [1]).

Property 1

For any $0 < q < \infty, \gamma > 0$, as $d \to \infty$,

$$\begin{aligned}
\mathbf{P}(\mathrm{GN}_\gamma(0) > (\gamma q \log d)^{\frac{1}{\gamma}}) &= \kappa_\gamma \int_{(\gamma q \log d)^{\frac{1}{\gamma}}}^\infty \exp\left(-\frac{x^\gamma}{\gamma}\right) dx \\
&= \kappa_\gamma \int_{q\log d}^\infty \exp(-y)\gamma^{\frac{1}{\gamma}-1}y^{\frac{1}{\gamma}-1}dy \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \int_{q\log d}^\infty \exp(-y)y^{\frac{1}{\gamma}-1}dy \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1}\Gamma\left(\frac{1}{\gamma}, q\log d\right) \\
&\sim \kappa_\gamma \gamma^{\frac{1}{\gamma}-1}e^{-q\log d}(q\log d)^{\frac{1}{\gamma}-1}
\end{aligned}$$

$$= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} d^{-q} (q \log d)^{\frac{1}{\gamma}-1}$$
$$= B^*_{1,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-q},$$

where $B^*_{1,\gamma} = \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} q^{\frac{1}{\gamma}-1}$.

Property 2

For any $0 < r < q < \infty, \gamma > 0$, as $d \to \infty$,

$$\begin{aligned}
\mathbf{P}(\mathrm{GN}_\gamma((\gamma r \log d)^{\frac{1}{\gamma}}) > (\gamma q \log d)^{\frac{1}{\gamma}})) &= \kappa_\gamma \int_{(\gamma q \log d)^{\frac{1}{\gamma}}}^{\infty} \exp\left(-\frac{(x - (\gamma r \log d)^{\frac{1}{\gamma}})^\gamma}{\gamma}\right) dx \\
&= \kappa_\gamma \int_{\log d \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma}^{\infty} \exp(-y) \gamma^{\frac{1}{\gamma}-1} y^{\frac{1}{\gamma}-1} dy \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \int_{\log d \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma}^{\infty} \exp(-y) y^{\frac{1}{\gamma}-1} dy \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \Gamma\left(\frac{1}{\gamma}, \log d \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma\right) \\
&\sim \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} e^{-\log d \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma} \left(\log d \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma\right)^{\frac{1}{\gamma}-1} \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} d^{-\left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma} (\log d)^{\frac{1}{\gamma}-1} \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^{1-\gamma} \\
&= B^*_{2,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-\left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^\gamma},
\end{aligned}$$

where $B^*_{2,\gamma} = \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \left(q^{\frac{1}{\gamma}} - r^{\frac{1}{\gamma}}\right)^{1-\gamma}$.

Property 3

For any $0 < q < r < \infty$ and any $\gamma > 0$, using the symmetry argument,

$$\begin{aligned}
&\mathbf{P}(\mathrm{GN}_\gamma((\gamma r \log d)^{\frac{1}{\gamma}}) \leq (\gamma q \log d)^{\frac{1}{\gamma}})) \\
&\qquad = \mathbf{P}(\mathrm{GN}_\gamma((\gamma r \log d)^{\frac{1}{\gamma}}) > 2(\gamma r \log d)^{\frac{1}{\gamma}} - (\gamma q \log d)^{\frac{1}{\gamma}})
\end{aligned}$$

$$
\begin{aligned}
&= \kappa_\gamma \int_{2(\gamma r \log d)^{\frac{1}{\gamma}} - (\gamma q \log d)^{\frac{1}{\gamma}}}^{\infty} \exp\left(-\frac{(x - (\gamma r \log d)^{\frac{1}{\gamma}})^\gamma}{\gamma}\right) dx \\
&= \kappa_\gamma \int_{\log d\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma}^{\infty} \exp(-y) \gamma^{\frac{1}{\gamma}-1} y^{\frac{1}{\gamma}-1} dy \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \Gamma\left(\frac{1}{\gamma}, \log d\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma\right) \\
&\sim \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} e^{-\log d\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma} \left(\log d\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma\right)^{\frac{1}{\gamma}-1}, \quad d \to \infty, \\
&= \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} d^{-\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma} (\log d)^{\frac{1}{\gamma}-1} \left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^{1-\gamma} \\
&= B^*_{3,\gamma} (\log d)^{\frac{1}{\gamma}-1} d^{-\left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^\gamma},
\end{aligned}
$$

where $B^*_{3,\gamma} = \kappa_\gamma \gamma^{\frac{1}{\gamma}-1} \left(r^{\frac{1}{\gamma}} - q^{\frac{1}{\gamma}}\right)^{1-\gamma}$.

Bernstein's first inequality (pp. 855 of Shorack and Wellner [15])

Let $Y_1, \ldots, Y_d$ be i.i.d. random variables with mean 0 and variance σ^2. Furthermore, suppose that $|Y_j| \leq L$ for some $L > 0, j = 1, \ldots, d$. Then for all $y > 0$,

$$
\mathbf{P}\left(\left|\sum_{j=1}^{d} Y_j\right| \geq y\right) \leq 2 \exp\left(\frac{-y^2}{2\left(\sigma^2 d + \frac{Ly}{3}\right)}\right).
$$

Bernstein's second inequality (Theorem 2.8 of Petrov [13])

Let $Y_1, \ldots, Y_d$ be independent random variables such that

1. $\mathbb{E}(Y_j) = 0$, for $j = 1, 2, \ldots, d$;
2. For some $H > 0$ and all $k \geq 2, |\mathbb{E}(Y_j^k)| \leq \frac{\mathbb{E}(Y_j^2)}{2} H^{k-2} k! < \infty$.

If the above conditions hold, then using the notation $S_d = \sum_{j=1}^{d} Y_j$ and $D_d^2 = \sum_{j=1}^{d} \mathbb{E}(Y_j^2)$,

$$\max\left\{\mathbf{P}(S_d \geq t), \mathbf{P}(S_d \leq -t)\right\} \leq \begin{cases} \exp\left(-\frac{t^2}{4D_d^2}\right), & 0 \leq t \leq D_d^2/H, \\ \exp\left(-\frac{t}{4H}\right), & t \geq D_d^2/H. \end{cases}$$

Remark 9: If $\mathbb{E}(Y_j) = 0$ and $|Y_j|$ is almost surely bounded by some positive constant L for $j = 1, \ldots, d$, then above condition 2 holds with $H = L/3$.

Appendix B: R Code

```
library(gnorm)

#Return n observations from a generalized normal
#distribution, with shape parameter gamma and mean mu.
#Note that rgnorm is borrowed from the R package 'gnorm' [7].
rgnormT<-function(n,mu,gamma){
  return(rgnorm(n,mu=mu,alpha=gamma^(1/gamma),beta=gamma))
}

#Generate from a generalized normal mixture model.
mixture<-function(d,beta,r,gamma){
  eta<-rep(0,d)
  u<-runif(d)
  eta<-(u<=(d^(-beta)))
  mu<-(gamma*r*log(d,exp(1)))^(1/gamma)
  generated<-data.frame(eta*mu+rgnormT(d,0,gamma),eta*1)
  colnames(generated)<-c(``data'',``eta'')
  return(generated)
}
```

```
#The signal strength r must exceed the value returned by this
#function for exact variable selection to be possible.
FullSelectionR<-function(beta,gamma){
  region<-(1+((1-beta)^(1/gamma)))^gamma
  return(region)
}

#Sequence delta_d.
delta<-function(d,gamma){
  deltaVal<-1/(log(d)^0.9)
  return(deltaVal)
}

#Threshold for exact selector defined in Theorem 2.
SelectorThreshold<-function(d,gamma){
  return(((gamma+delta(d,gamma))*log(d))^(1/gamma))
}

#Return the dth element of the sequence tau_d.
taud<-function(gamma,d){
  delta<-delta(d,gamma)
  return(d^(delta/(gamma+1)))
}
#Return Vj, where Vj = d^{1-beta_j}/tau_d.
Vj<-function(d,gamma,beta){
```

```
  return((d^(1-beta))/(taud(gamma,d)))
}

######EXACT SELECTION SIMULATION#####
ExactselectionSim<-function(d,beta,gamma,replicates){
  performance<-rep(0,replicates)
  threshold<-SelectorThreshold(d,gamma)
  r<-FullSelectionR(beta,gamma)+1
  for (i in 1:replicates){
    data<-mixture(d,beta,r,gamma)
    performance[i]<-sum(abs((data[,1]>=threshold)*1
                            - data[,2]))/d
  }
  perf<-mean(performance)
  threshold<-((((gamma+delta(d,gamma))*log(d))^(1/gamma)))
  mu<-(gamma*r*log(d))^(1/gamma)
  rGreaterThan<-FullSelectionR(beta,gamma)
  return(c(r,mu,rGreaterThan,threshold,perf))
}

######ALMOST FULL SELECTION SIMULATION######
#Returns selector \tilde{\eta}.
Selector<-function(beta,gamma,X){
  d<-length(X)
  eta<-(X >=((((gamma*beta
                +delta(d,gamma))*log(d))^(1/gamma))))*1
```

```
  return(data.frame(eta))
}

AlmostselectionSim<-function(d,beta,gamma,replicates){
  r<-(FullSelectionR(beta,gamma)+beta)/2
  performance<-rep(0,replicates)
  for (i in 1:replicates){
    dat<-mixture(d,beta,r,gamma)
    etaE<-Selector(beta,gamma,dat[,1])
    performance[i]<-sum(abs(etaE-dat[,2]))/d
  }
  perf<-mean(performance)
  threshold<-((((gamma*beta
                  + delta(d,gamma))*log(d))^(1/gamma)))
  mu<-(gamma*r*log(d))^(1/gamma)
  return(c(r,mu,threshold,perf))
}

######ADAPTIVE SELECTION SIMULATION######
#Find the index tildeM. See Section 2.4 for
#the definition of tildeM.
adaptiveSelectionFindM<-function(B,M,gamma,Xis){
  d<-length(Xis)
  betas<-rep(0,M)
  etaI<-matrix(0,d,M)
  V<-rep(0,M)
```

```
  for (i in 1:M){
    betas[i]<-i*(B/M)
    etaI[,i]<-as.matrix(Selector(betas[i],gamma,Xis))
    V[i]<-Vj(d,gamma,betas[i])
  }
  TildeM<-0
  for (k in 1:M){
    m<-M-k+1
    for(j in 1:m){
      if( sum(abs(etaI[,m]-etaI[,m-j+1])) > V[m-j+1] ){
        break
      }
      if (j==m){
        TildeM<-m
      }
    }
    if (TildeM>0){
      break
    }
  }
  return(betas[TildeM])
}

AdaptiveSelectionSim<-function(d,beta,gamma,B,M,replicates){
  r<-(FullSelectionR(beta,gamma)+beta)/2
  performance<-rep(0,replicates)
```

```
  betaTildeMs<-rep(0,replicates)
  for (i in 1:replicates){
    dat<-mixture(d,beta,r,gamma)
    betaTildeMs[i]<-adaptiveSelectionFindM(B,M,gamma,dat[,1])
    etaE<-Selector(betaTildeMs[i],gamma,dat[,1])
    performance[i]<-sum(abs(etaE-dat[,2]))/d
  }
  meanBeta<-mean(betaTildeMs)
  threshold<-((((gamma*meanBeta
                  +delta(d,gamma))*log(d))^(1/gamma)))
  mu<-(gamma*r*log(d))^(1/gamma)
  perf<-mean(performance)
  return(c(r,mu,meanBeta,threshold,perf))
}

######SIMULATION TABLES######
#For fixed d and prior p on eta, return the value beta.
getbetas<-function(dValues,proportion){
  betas<-rep(0,length(dValues))
  for(i in 1:length(dValues)){
    betas[i]<-log(proportion)/log(dValues[i])
  }
  return(data.frame(dValues,betas,proportion))
}

dbetaP<-function(dVals,proportion){
```

```
  for(i in 1:length(proportions)){
    if (i==1){
      results<-getbetas(dVals,proportion[i])
    }
    if (i>1){
      results<-rbind(results,getbetas(dVals,proportion[i]))
    }
  }
  return(results)
}
#Generate tables which store the results of each simulation.
createTableforSim<-function(ds,props,gam,SelectorType){
  getTable<-dbetaP(ds,props)
  getTable<-cbind(getTable,rep(0,length(getTable[,1])))
  colnames(getTable)<-c(``d'',``beta'',``d^{-beta}'',``gamma'')
  getTable<-do.call(``rbind'',
  replicate(length(gam), getTable, simplify = FALSE))
  pdLength<-length(ds)*length(props)
  for(i in 1:length(gam)){
    getTable[(i-1)*pdLength+1:pdLength,4]<-gam[i]
  }
  if (SelectorType==2){
    SimResults<-matrix(0,nrow(getTable),5)
    colnames(SimResults)<-c(``r'',``mu'',``rGreaterThan'',
                            ``threshold'',``performance'')
    for (i in 1:length(getTable[,1])){
```

```
    SimResults[i,]<-ExactselectionSim(getTable[i,1],
    getTable[i,2],getTable[i,4],500)
  }
  finalTable<-cbind(getTable,SimResults)
}
if (SelectorType==1){
  SimResults<-matrix(0,nrow(getTable),4)
  colnames(SimResults)<-c(``r'',``mu'',``threshold'',
                          ``performance'')
  for (i in 1:length(getTable[,1])){
    SimResults[i,]<-AlmostselectionSim(getTable[i,1],
    getTable[i,2],getTable[i,4],500)
  }
  finalTable<-cbind(getTable,SimResults)
}
if (SelectorType==0){
  SimResults<-matrix(0,nrow(getTable),5)
  colnames(SimResults)<-c(``r'',``mu'',``meanBeta'',
  ``threshold'',``performance'')
  for (i in 1:length(getTable[,1])){
    SimResults[i,]<-AdaptiveSelectionSim(getTable[i,1],
    getTable[i,2],getTable[i,4],0.999,100,500)
  }
  finalTable<-cbind(getTable,SimResults)
}
return(finalTable)
```

```
}

######FUNCTIONS FOR CONSTRUCTING HISTOGRAMS AND VISUALS######
#Plot the density function of the generalized normal
#distribution, for fixed gamma.
plotGNDensity<-function(gamma){
  z<-seq(-3,3,0.01)
  density<-((2*gamma(1/gamma))^(-1))*
  exp(-(abs(z)^gamma)/gamma)
  plot(density~z,type=``l",ylab=``f(z)",
       xlab=``z",main=bquote(``Generalized normal density,
       for "*gamma*`` = "*.(gamma)),cex.lab=2.5,
       cex.main=2,cex.axis=2)
}

#Plot generalized normal mixture data in the case that
#almost full variable selection is possible.
plotMixtureAlmostFull<-function(d,beta,adaptiveBeta,gamma){
  r<-(FullSelectionR(beta,gamma)+beta)/2
  dat<-mixture(d,beta,r,gamma)[,1]
  mu<-(gamma*r*log(d))^(1/gamma)
  thresholdAlmostfullSelector<-((gamma*beta
  +delta(d,gamma))*log(d))^(1/gamma)
  adaptiveThreshold<-((gamma*adaptiveBeta
    +delta(d,gamma))*log(d))^(1/gamma)
  hist(dat,breaks=100,main=bquote(beta*`` = "
```

```
  *.(round(beta,3))*``, "*gamma*`` = "*.(gamma)*``, "
  *d*`` = "*.(d)),xlab=``X",freq=FALSE,
  cex.main=2.0,cex.axis=2.0,cex.lab=2.0)
  abline(v=mu,col="red")
  abline(v=thresholdAlmostfullSelector,col=``blue")
  abline(v=adaptiveThreshold,col=``green")
  legend(``topright", c(``Threshold for known sparsity",
 ``Adaptive threshold", expression(paste(mu))),
        col=c(``blue", ``green",``red"), lwd=1,
        cex=1.5)
return(c(mu,thresholdAlmostfullSelector,adaptiveThreshold))
}

#Plot generalized normal mixture data in the case that exact
#variable selection is possible.
plotMixtureExact<-function(d,beta,gamma){
  r<-FullSelectionR(beta,gamma)+1
  dat<-mixture(d,beta,r,gamma)[,1]
  mu<-(gamma*r*log(d))^(1/gamma)
  thresholdExactSelector<-((gamma+delta(d,gamma))
  *log(d))^(1/gamma)
  hist(dat,breaks=100,main=bquote(beta*`` = "
  *.(round(beta,3))*``, "*gamma
  *`` = "*.(gamma)*``, "*d*`` = "*.(d)),xlab=``X",
  freq=FALSE,cex.lab=2,cex.main=2,cex.axis=2)
  abline(v=mu,col="red")
```

```
  abline(v=thresholdExactSelector,col=``blue")
  legend(``topright", c(``Threshold" ,
  expression(paste(mu))), col=c(``blue", ``red"), lwd=1,
        cex=2)
  return(c(mu,thresholdExactSelector))
}

#Plot variable selection regions for various choices of
#gamma.
plotSelectionRegions<-function(gamma,NoX,NoY,
almostTextX,almostTextY,ExactX,ExactY,Normal){
  beta<-seq(0,1,0.01)
  Exact<-(1+(1-beta)^(1/gamma))^gamma
  if(Normal==0){
    plot(beta,Exact,type=``l",xlim=c(0,1),ylim=c(0,2^gamma),
        ylab=``r",axes=FALSE,xlab=expression(beta),
        cex.lab=3.0, cex.main=2.0,
        main=bquote(gamma*`` = "*.(gamma)))
  }
  if (Normal==1){
    plot(beta,Exact,type=``l",xlim=c(0,1),ylim=c(0,2^gamma),
        ylab=``r",axes=FALSE,xlab=expression(beta),
        cex.lab=3.0, cex.main=2.0,
        main=``Selection regions for normal distribution")
  }
  lines(beta,beta)
```

```
  lines(rep(0,length(beta)),beta*(2^gamma))
  lines(beta,rep(2^gamma,length(beta)))
  lines(beta,rep(0,length(beta)))
  lines(rep(1,length(beta)),beta*(2^gamma))
  axis(1,pos=c(0,0),cex.axis=1.8)
  axis(2,pos=c(0,0),cex.axis=1.8)
  text(NoX,NoY,``No selection possible",cex=1.2)
  text(almostTextX,almostTextY,
       ``Almost full selection region",cex=1.2)
  text(ExactX,ExactY,``Exact selection region",cex=1.2)
}
#Plot the function h_d(y) defined in the proof of Theorem 3.
hdyVsY<-function(gamma,beta,d){
  y<-seq(-1.5,1.5,0.01)
  r<-(FullSelectionR(beta,gamma)+beta)/2
  c<-beta/r
  hy<-(abs(y)^gamma)-(abs(y-1)^gamma)-c
  -(log(1-d^(-beta))/(r*log(d)))
  plot(hy~y,type=``l",ylab=expression(``h"[d]*``(y)"),
  main=bquote(h[d]*``(y) vs y,"~gamma*`` = "
  *.(gamma)),cex.main=2,cex.lab=2,cex.axis=2)
}

######SIMULATION RESULTS AND PLOTS######
#Figure 1.1.
par(mfrow=c(1,1))
```

```
plotSelectionRegions(2,0.7,0.3,0.3,2,0.7,3.5,1)

#Figure 1.2.
par(mfrow=c(3,2),mar=c(5.1, 6, 4.1, 2.1))
plotGNDensity(0.5)
plotGNDensity(1)
plotGNDensity(1.5)
plotGNDensity(2)
plotGNDensity(3)
plotGNDensity(10)



#Figure 2.1.
par(mfrow=c(2,2))
hdyVsY(0.5,0.5,1000)
hdyVsY(1,0.5,1000)
hdyVsY(2,0.5,1000)
hdyVsY(3,0.5,1000)

#Figure 2.2.
par(mfrow=c(1,2))
plotSelectionRegions(0.5,0.7,0.2,0.3,0.7,0.6,1.3,0)
plotSelectionRegions(1,0.7,0.2,0.3,1,0.7,1.7,0)

#Figure 2.3.
plotSelectionRegions(2,0.7,0.3,0.3,2,0.7,3.5,0)
```

```
plotSelectionRegions(3,0.7,0.3,0.4,3,0.7,6.5,0)

beta<-seq(0,1,0.01)
dChoices<-c(1000,10000)
proportions<-c(0.09,0.05,0.01)
gam<-c(0.5,1,1.5,2,2.5)

#Tables for Simulations 1 through 3.
tabforThesisAlmostFull<-createTableforSim(dChoices,
proportions,gam,1)
tabforThesisExact<-createTableforSim(dChoices,
proportions,gam,2)
tabThesisAlmostFullLepski<-createTableforSim(dChoices,
proportions,gam,0)

#Figure 3.1.
par(mfrow=c(2,2))
plotMixtureExact(10000,tabforThesisExact[4,2],0.5)
plotMixtureExact(10000,tabforThesisExact[10,2],1)
plotMixtureExact(10000,tabforThesisExact[16,2],1.5)
plotMixtureExact(10000,tabforThesisExact[28,2],2.5)

#Figure 3.2.
par(mfrow=c(2,2))
plotMixtureAlmostFull(10000,tabforThesisAlmostFull[4,2],
                      tabThesisAlmostFullLepski[4,7],0.5)
```

```
plotMixtureAlmostFull(10000,tabforThesisAlmostFull[10,2],
                      tabThesisAlmostFullLepski[10,7],1)
plotMixtureAlmostFull(10000,tabforThesisAlmostFull[16,2],
                      tabThesisAlmostFullLepski[16,7],1.5)
plotMixtureAlmostFull(10000,tabforThesisAlmostFull[28,2],
                      tabThesisAlmostFullLepski[28,7],2.5)
######END OF CODE######
```

Bibliography

[1] Abramowitz, M. and Stegun, I.A. (1964). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Dover Publications. New York.

[2] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley. Reading, Massachusetts.

[3] Butucea, C., Ndaoud, M., Stepanova, N.A. and Tsybakov, A.B. (2018). Variable selection with Hamming loss. *The Annals of Statistics*, Vol. 46, No. 5, 1837–1875.

[4] Cui, X. (2014). Optimal Component Selection in High Dimensions. MSc Thesis. Carleton University.

[5] Dytso, A., Bustin, R., Poor, H.V. and Shamai, S. (2018). Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, Vol. 5, No. 1, 1–40.

[6] Genovese, C.R., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, Vol. 13, 2107–2143.

[7] Griffin, M. (2018). gnorm: Generalized Normal/Exponential Power Distribution. *R package version 1.0.0.*

[8] Hastie, T., Tibshirani, R. and Wainwright, M. (2016). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. Boca Raton, Florida.

[9] Ingster, Y.I., Pouet, C. and Tsybakov, A.B. (2009). Classification of sparse high-dimensional vectors. *Philisophical Transactions of the Royal Society A*, Vol. 367, No. 1906, 4427–4448.

[10] International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, Vol. 431, 931–945.

[11] Lepski, O.V. (1991). One problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, Vol. 35, No. 3, 454–466.

[12] Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*. Vol. 32, No. 7, 685–694.

[13] Petrov, V.V. (1995). *Limit Theorems of Probability Theory*. Clarendon Press. Oxford.

[14] Roussas, G.G. (1997). *A Course in Mathematical Statistics*. Academic Press. San Diego, California.

[15] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons. New York.

[16] Subbotin, M.T. (1923). On the law of frequency of error. *Matematicheskii Sbornik*. Vol. 31, No. 2, 296–301.

[17] Varanasi, M.K. and Aazhang, B. (1989). Parametric generalized Gaussian density estimation. *The Journal of the Acoustical Society of America*. Vol. 86, 1404–1415.