Classification and feature selection in sparse high-dimensional models

by

Lee Thompson

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

Master of Science

in

Mathematics

Ottawa-Carleton Institute for
Mathematics and Statistics

Carleton University
Ottawa, Ontario, Canada

© 2017, Lee Thompson

Abstract

This work deals with binary classification in a sparse high-dimensional model. Essentially, we have two predetermined subclasses of the normally distributed population and we want to assign a new observation to one of the two subclasses. In this thesis, we look at a classification problem in which the sample size is less than the dimension of the data. The goal is to find good methods for classification that can minimize our chances of misclassification under these circumstances. A typical high-dimensional problem, like the one studied in this thesis, has a large number of unknown parameters and not enough data to make reliable inferences about these parameters. The classical statistical methods were not designed to cope with high-dimensional problems. Therefore, when studying the classification problem of our interest, we have proposed original statistical ideas and tools based on the notion of ‘sparsity’.

Much of the research done on classification in high-dimensional settings has been under the assumption that the covariance matrix of the observed vectors has a diagonal structure. In this thesis, we have a more general condition for this covariance matrix. We only require that the matrix has block-diagonal structure which makes this study more relevant to practical situations.

In this thesis, we propose new classification procedures and investigate one existing classification method from the literature in terms of classification accuracy. Even in a simpler setup of a diagonal covariance matrix, there has been no rigorous justification for the existing method under consideration which makes this research crucial for furthering our understanding of how to minimize the chances of misclassification under the conditions mentioned above.

The motivation behind much of the research done in the realm of high-dimensional statistics is in the pursuit of knowledge. While this is the case here as well, it may also have relevance to the biomedical field as the concept of classification (especially, binary classification) is indistinguishable from the notion of diagnostics in the medical field.

Acknowledgements

I would first like to thank my supervisor, Dr. Natalia Stepanova, for all her help over the past two years. Her guidance, knowledge and patience has been invaluable over that time, but it is her devotion to her students that I have appreciated the most.

I would also like to thank my parents for all their love and encouragement through this endeavor. It is truly amazing to have parents that understand the pursuit of happiness is more important than financial security.

Lastly, I would like to thank Erin. None of this would have been possible without your unwavering support.

Notation

$\exists x$	there exists an element x
$\forall x$	for all elements x
$\mathbb{I}(A)$ or $\mathbb{I}\{A\}$	indicator function of event A
$:=$	equal by definition
$\xrightarrow{d}$	convergence in distribution
$\stackrel{d}{=}$	equality in distribution
$Y \sim \mathbf{P}$	Y is distributed according to the distribution $\mathbf{P}$
$\ll$	much less than
$\gg$	much greater than
$\mathbb{R}$	set of real numbers
$\Phi(\cdot)$	CDF of the standard normal distribution
$\varphi(\cdot)$	PDF of the standard normal distribution
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	p -variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\chi_\nu^2(\gamma)$	chi-square distribution with ν df and noncentrality parameter γ
$F_{\nu_1, \nu_2}(\gamma)$	F distribution with ν_1 and ν_2 df and noncentrality parameter γ
$f = o(g)$	“small o ” means: $f(x)/g(x) \rightarrow 0$ as $x \rightarrow a$
$f = O(g)$	“small O ” means: $f(x)/g(x) \rightarrow C$, constant $C \neq 0$, as $x \rightarrow a$
$\liminf_{n \rightarrow \infty} a_n$	limit inferior of the sequence a_n , $\sup_{n \geq 1} \inf_{k \geq n} a_k$
$\limsup_{n \rightarrow \infty} a_n$	limit superior of the sequence a_n , $\inf_{n \geq 1} \sup_{k \geq n} a_k$
$a_n \sim b_n$	means $\lim_{n \rightarrow \infty} a_n/b_n = 1$
$a_n \asymp b_n$	means $0 < \liminf_{n \rightarrow \infty} (a_n/b_n) \leq \limsup_{n \rightarrow \infty} (a_n/b_n) < \infty$
$\mathbf{E}_{\mathbf{P}}(\cdot)$	expectation with respect to the distribution $\mathbf{P}$
$[x]$	the integer part of the number x
$\ln a$	the natural (base e) logarithm of the number a
$o_p(1)$	a sequence of random variables or vectors that converges to zero in probability
$O_p(1)$	a sequence of random variables or vectors that is bounded in probability
$X_n = o_p(R_n)$	means $X_n = Y_n R_n$ and $Y_n = o_p(1)$
$X_n = O_p(R_n)$	means $X_n = Y_n R_n$ and $Y_n = O_p(1)$

Contents

Introduction	1
1 General problem statement: standard setup	3
2 General problem statement: high-dimensional setup	5
3 Risk measures	9
4 Some useful statistics and their properties	10
5 Definition of the model	13
6 Three regions of β -sparsity	16
6.1 The case of moderate β -sparsity	17
6.2 The case of high β -sparsity	21
6.3 The case of very high β -sparsity	23
7 Classifiers with partially estimated parameters	23
8 Feature selection prior to classification	25
9 Obtaining threshold t^* in the case of $r > \beta$	37
10 Classifiers with selected informative feature blocks	46
11 Obtaining threshold t^* in the case of $r < \beta$	47
12 Classifiers with partially estimated parameters and selected informative feature blocks	55
13 Completely data-driven classifiers	56

14 Numerical analysis	58
14.1 Simulations in the region $\mathbf{D}_1$	59
14.2 Simulations in the region $\mathbf{D}_2$	60
14.3 Simulations in the region $\mathbf{D}_3$	60
15 Informal discussion of the classification problem in the region $\mathbf{D}_3$	62
16 Proofs	68
Conclusion	89
References	91
Appendix	93

Introduction

Over the last two decades, analysis of high-dimensional data has become a common practice in many fields of medicine, engineering and science. Often the task is to predict or classify a p -dimensional vector using the information obtained from a training sample of vectors of the same dimension. The main difficulty is that the dimension p can be very large, in which case it is natural to consider the asymptotic setup when $p \rightarrow \infty$. The situation of very large p is not covered by the classical statistical methods.

This thesis is rooted in a high-dimensional setting in which we attempt to make effective statistical inferences. Generally, a high-dimensional problem has a large number of unknown parameters but not enough data to make reliable inferences about the parameters. In this thesis, we will investigate a classification problem in which the sample size is less than the dimension of the data p . In making effective statistical inference in high-dimensional settings, the assumption of sparsity of the data plays an important role. Sometimes it allows us to reduce the dimension of the problem so that the standard statistical methods become applicable.

This research was inspired by the challenging problem of classification in sparse normal models based on the publications of Ingster et al. [12] and Donoho and Jin [8]. Unlike Ingster et al. [12] and Donoho and Jin [8], who studied a similar problem for normal vectors with independent and identically distributed components, we do not assume that the covariance matrix of the observed vectors is diagonal. Instead, we put a weaker restriction on the covariance matrix to be block-diagonal. Then, having a block-diagonal covariance matrix, we impose the sparsity condition on the data that naturally involves both the means of the observed vectors and their common covariance matrix. The results in Ingster et al. [12] obtained for the diagonal covariance matrix can hardly be extended to the more general model of our interest. Therefore, in this thesis we propose different classification strategies, one of which has some connections with the method in Donoho and Jin [8].

Due to the sparsity in the observed vectors, there are both ‘useful’ and ‘useless’ sub-vectors in terms of classification. The main predicament in classifying sparse normal vectors is selecting sub-vectors that contribute to successful classification while minimizing the probability of misclassification. Prior to classification one has to carefully select as many useful sub-vectors as possible while including as few useless sub-vectors as possible. To do this we apply some recent results from the literature on hypothesis testing and variable selection in sparse models, along with some modifications, to accommodate the more general case of a block-diagonal covariance matrix.

In the process of working on the thesis, we have found that, technically, the high-dimensional problem of our interest turns out to be more involved than expected at the outset. For example, only a justification of a seemingly simple shift from an oracle classifier with known parameters to a similar classifier with partially estimated parameters requires relatively long derivations (see the proof of Proposition 1 in Section 16). At the same time, even in a simpler standard setup, the classical publications and monographs on Multivariate Statistical Analysis, such as the book of Anderson [1], do not contain rigorous justifications of why the well-known data-driven classifiers work well.

The current work should not be treated as a complete study of the problem in hand. The results of the thesis that were found to be consistent with our numerical conclusions of Section 14 but left without rigorous justification, are stated in this thesis as conjectures. In any case, we believe that the results obtained and conjectured in the present work form a good basis for further analytical study of this interesting and challenging problem of classification in high dimensions.

The main results of this thesis, Proposition 1 of Section 7, Propositions 2–3 of Section 9, Proposition 4 of Section 10, and Theorem 5 of Section 12, are new. They provide useful facts and details towards better understanding the problem of binary classification in high dimensions.

1 General problem statement: standard setup

A general classification problem is simple in concept. It is the idea of assigning new observations to categories (subsets of total population) using a training sample where category assignment is known. On a basic level, classification is a type of pattern recognition where we take an educated guess as to what category a new observation belongs using previously vetted information. Successful classification is achieved when an observation is believed to belong to a certain specific category and it turns out to be true. In this work, we are going to consider a binary classification problem with two categories in a high-dimensional setting and how to achieve successful classification.

Let $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ be a training sample of n random vectors which are known to be from a joint distribution of $(\mathbf{X}, Y) \in \mathbb{R}^p \times \{0, 1\}$. We assume that $\mathbf{P}(Y = 0) = \mathbf{P}(Y = 1) = 1/2$, i.e., the classes are equally likely. The sample values are defined as follows:

$$\mathbf{X}_i|Y_i = 0 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \equiv \Pi_1 \quad \text{and} \quad \mathbf{X}_i|Y_i = 1 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \equiv \Pi_2,$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix.

Suppose $\mathbf{X}_0 \in \mathbb{R}^p$ is a new observation from either Π_1 or Π_2 , and we wish to determine from which distribution $\mathbf{X}_0$ originated. The problem is determining the class label Y for $\mathbf{X}_0$. When classifying $\mathbf{X}_0$ as from Π_1 or Π_2 there are four outcomes that arise:

	True population:		
		Π_1	Π_2
Classified as:	Π_1	Correct	Incorrect
	Π_2	Incorrect	Correct

Classification problems are often gauged by their misclassification probabilities, but the cost of misclassification, $c(\cdot|\cdot) \geq 0$, should also be taken into consideration (Section 11.2 of Johnson & Wichern [14]). The cost of misclassification is a predetermined

constant. It can be viewed as a penalty for classifying an observation incorrectly. The cost $c(1|2)$ occurs when an observation classified as Π_1 is actually from Π_2 and the cost $c(2|1)$ occurs when an observation classified as Π_2 is actually from Π_1 . There is no cost when classification is correct, i.e., $c(1|1) = c(2|2) = 0$.

	True population:		
		Π_1	Π_2
	Classified as:	Π_1	Π_2
	Π_1	$c(1 1)$	$c(1 2)$
	Π_2	$c(2 1)$	$c(2 2)$

Consider a more general situation with unequal prior probabilities p_1 and p_2 . Since we are dealing with binary classification we know that $p_1 = 1 - p_2$. If $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are available, then the optimal classification rule that minimizes the chance of misclassification is known to be (see Section 11.3 of Johnson & Wichern [14] and Chapter 6 of Anderson [1])

$$\psi_0(\mathbf{X}_0) = \mathbb{I} \left\{ (\mathbf{X}_0 - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \leq \ln \left[\left(\frac{c(1|2)}{c(2|1)} \right) \left(\frac{p_2}{p_1} \right) \right] \right\}, \quad (1)$$

where $\mathbb{I}\{A\} = 1$ if event A occurs and zero otherwise.

In this work, we assume that $c(1|2) = c(2|1)$, and it has already been stated that the classes are equally likely, which implies $p_1 = p_2 = 1/2$. The optimal classification rule then becomes (see (1))

$$\psi_0(\mathbf{X}_0) = \mathbb{I} \left\{ (\mathbf{X}_0 - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \leq 0 \right\}. \quad (2)$$

This classifier will allocate $\mathbf{X}_0$ to Π_1 when $\psi_0(\mathbf{X}_0) = 1$. It will also allocate $\mathbf{X}_0$ to Π_2 when $\psi_0(\mathbf{X}_0) = 0$. Essentially, we can view $\psi_0(\mathbf{X}_0)$ as an estimator of Y based on $\mathbf{X}_0$. This classifier is constructed as the Bayes rule with equal class prior probabilities which is optimal, yielding the minimum of the total misclassification probability (see Section 11.4 of Johnson & Wichern [14]):

$$\begin{aligned}
\mathcal{E} &:= \min_{0 < \pi < 1} (\pi \mathbf{E}_{\Pi_2}(\psi_0) + (1 - \pi) \mathbf{E}_{\Pi_1}(1 - \psi_0)) \\
&= \frac{1}{2} \mathbf{E}_{\Pi_2}(\psi_0) + \frac{1}{2} \mathbf{E}_{\Pi_1}(1 - \psi_0) = \Phi \left(-\frac{1}{2} \Delta^{1/2} \right), \tag{3}
\end{aligned}$$

where

$$\Delta = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

We see that if Δ is large then $\mathcal{E}$ is small. This means that the risk of misclassification decreases as the distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, represented by Δ , increases. We refer to $\mathcal{E}$ as the *optimum error rate* since it represents the best possible performance that can be achieved in an ideal setting and thus used as a benchmark when evaluating classification accuracy. If $n - 2 \geq p$, one can replace $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ in (2) by their sample counterparts and obtain the corresponding data-driven classifier (see, for example, Section 6 of Anderson [1]).

2 General problem statement: high-dimensional setup

In this thesis, the covariance matrix $\boldsymbol{\Sigma}$ is assumed to have a *block-diagonal structure* which has been sometimes seen in biomedical studies (see, for example, Balding [2] and Donoho and Jin [8]). Namely, we assume that

$$\boldsymbol{\Sigma} = \text{BlkDiag}(\boldsymbol{\Sigma}_{[1]}, \dots, \boldsymbol{\Sigma}_{[b]}),$$

where each block $\boldsymbol{\Sigma}_{[k]}$, $k = 1, \dots, b$, is a $p_0 \times p_0$ matrix with fixed p_0 being a natural

number. This implies that

$$p = bp_0. \quad (4)$$

Next, as we deal with the binary classification problem, we assume that n is even. Let $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, where $\mathbf{X}_k^{(1)} \sim \Pi_1$ and $\mathbf{X}_k^{(2)} \sim \Pi_2$ for $k = 1, \dots, n/2$, be two sub-samples constituting the sample $\mathbf{X}_1, \dots, \mathbf{X}_n$, which are our *labelled training sample* referred to in Section 1. Thinking of all vectors as column-vectors, we shall write for $j = 1, \dots, n/2$

$$\begin{aligned} \mathbf{X}_j^{(1)} &= \left((\mathbf{X}_{j,[1]}^{(1)})^\top, \dots, (\mathbf{X}_{j,[b]}^{(1)})^\top \right)^\top, & \mathbf{X}_j^{(2)} &= \left((\mathbf{X}_{j,[1]}^{(2)})^\top, \dots, (\mathbf{X}_{j,[b]}^{(2)})^\top \right)^\top, \\ \boldsymbol{\mu}_1 &= (\boldsymbol{\mu}_{1,[1]}^\top, \dots, \boldsymbol{\mu}_{1,[b]}^\top)^\top, & \boldsymbol{\mu}_2 &= (\boldsymbol{\mu}_{2,[1]}^\top, \dots, \boldsymbol{\mu}_{2,[b]}^\top)^\top, & \boldsymbol{\mu} &= (\boldsymbol{\mu}_{[1]}^\top, \dots, \boldsymbol{\mu}_{[b]}^\top)^\top. \end{aligned}$$

where we put

$$\boldsymbol{\mu} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1.$$

Below, we will refer to the vector $\mathbf{X}$ as a *feature vector*, with b being the *number of sub-features*. Each feature will consist of p_0 elements. We use b as a driving parameter, and let other parameters be tied to b through fixed parameters. The sub-vectors $\mathbf{X}_{j,[k]}^{(1)}$ and $\mathbf{X}_{j,[k]}^{(2)}$ are in $\mathbb{R}^{p_0}$ for $j = 1, \dots, n/2$ and $k = 1, \dots, b$.

Next, we assume that the sample size $n = n_b$ relates the dimension $p = p_b$ through the relation (recall that $b = p/p_0$)

$$n \sim b^\theta, \quad b \rightarrow \infty, \quad (5)$$

for some (known) parameter $\theta \in (0, 1)$.

If the driving parameter b tends to infinity we have (see (4) and (5))

$$n = o(p)$$

and the classification problem in hand becomes a *high-dimensional problem*. To cope with high dimensionality, some kind of sparsity of the data is often reasonably assumed. Therefore, in addition to this being a high-dimensional problem, we assume that only a small fraction of b features are *informative*. The k th feature block is considered to be informative if

$$\Delta_{[k]} = \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \geq a,$$

for some positive value $a = a_b$. We denote the fraction of informative feature blocks as $\varepsilon = \varepsilon_b$ and assume that

$$\varepsilon = b^{-\beta}$$

for some unknown parameter $0 < \beta < 1$, called the *sparsity parameter*. What we notice is that there is only a small fraction $b^{-\beta}$ of blocks that are informative, which means that there is a high portion, $1 - b^{-\beta}$, of blocks that are non-informative. Seeing as how there is only a small fraction of informative blocks and the labelled training sample is also small, the classical methods introduced in Section 1 generally cannot be used. The informative blocks are considered to be *sparse* and we do not know which blocks provide information leading to successful classification.

Suppose for a moment that we knew which subvectors $\mathbf{X}_{[k]}$ of a vector $\mathbf{X}$ (and hence which subvectors $\mathbf{X}_{j,[k]}^{(1)}$ and $\mathbf{X}_{j,[k]}^{(2)}$ of vectors $\mathbf{X}_j^{(1)}$ and $\mathbf{X}_j^{(2)}$, $j = 1, \dots, n/2$) are informative and define the vector $\omega = (\omega_k)_{k=1, \dots, b}$ as

$$\omega_k = \mathbb{I}(\mathbf{X}_{[k]} \text{ is informative}), \quad k = 1, \dots, b. \quad (6)$$

Then we can alter equation (2) from Section 1 so that we only consider the subvectors of the feature vectors $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$ that are deemed informative. We

define the resulting classifier, which depends on $\mathbf{X}_0$ and $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}, \mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, as

$$\hat{\psi}_0(\mathbf{X}_0) = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \hat{\boldsymbol{\Sigma}}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}, \quad (7)$$

where (recall that $n \sim b^\theta$ for some $0 < \theta < 1$)

$$\hat{\boldsymbol{\mu}}_{1,[k]} = \hat{\boldsymbol{\mu}}_{1,[k],b} = \frac{2}{n} \sum_{j=1}^{n/2} \mathbf{X}_{j,[k]}^{(1)} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{2,[k]} = \hat{\boldsymbol{\mu}}_{2,[k],b} = \frac{2}{n} \sum_{j=1}^{n/2} \mathbf{X}_{j,[k]}^{(2)}, \quad (8)$$

are the sub-sample means, and the estimator $\hat{\boldsymbol{\Sigma}}_{[k]}$ of $\boldsymbol{\Sigma}_{[k]}$ is defined in Section 4 by relation (12). We also put

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1,$$

and write

$$\hat{\boldsymbol{\mu}}_1 = (\hat{\boldsymbol{\mu}}_{1,[1]}^\top, \dots, \hat{\boldsymbol{\mu}}_{1,[b]}^\top)^\top, \quad \hat{\boldsymbol{\mu}}_2 = (\hat{\boldsymbol{\mu}}_{2,[1]}^\top, \dots, \hat{\boldsymbol{\mu}}_{2,[b]}^\top)^\top, \quad \hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_{[1]}^\top, \dots, \hat{\boldsymbol{\mu}}_{[b]}^\top)^\top.$$

If vector $\omega = (\omega_k)_{k=1,\dots,b}$ given by (6) is unknown, it has to be estimated from the sample in a nice way. The quality of the resulting classifier will strongly depend on the quality of estimation of ω .

This classifier modifies on the classification rule from (2), that would fail to work in our settings because $n \ll p$, by not incorporating any non-informative feature blocks. The non-informative blocks increase the risk of misclassification since each non-informative block used in said equation returns a random guess. What we need to do now is to find a procedure for which we can incorporate as many informative feature blocks as possible, while also minimizing our risk of misclassification by excluding as many non-informative feature blocks as possible.

3 Risk measures

Let us denote by $\mathbf{E}_{\Pi_1}$ and $\mathbf{E}_{\Pi_2}$ the expectations with respect to the joint probability distribution of $\mathbf{X}_j^{(1)}, \mathbf{X}_j^{(2)}, j = 1, \dots, n/2$, and $\mathbf{X}_0$ when $\mathbf{X}_0 \sim \Pi_1$ and $\mathbf{X}_0 \sim \Pi_2$, respectively.

In general, the *Bayes risk* of a classifier $\psi = \psi(\mathbf{X}_0, \mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}, \mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)})$ is given by

$$\mathcal{R}_B(\psi) = \pi \mathbf{E}_{\Pi_2}(\psi) + (1 - \pi) \mathbf{E}_{\Pi_1}(1 - \psi), \quad (9)$$

where $0 < \pi < 1$ is a prior probability of the class Π_2 . We can also define the *maximum risk* of ψ by the formula

$$\mathcal{R}_M(\psi) = \max(\mathbf{E}_{\Pi_2}(\psi), \mathbf{E}_{\Pi_1}(1 - \psi)). \quad (10)$$

Here, the expectation $\mathbf{E}_{\Pi_2}(\psi)$ is the error probability of allocating $\mathbf{X}_0$ to Π_1 when $\mathbf{X}_0 \in \Pi_2$. Likewise, $\mathbf{E}_{\Pi_1}(1 - \psi)$ is the error probability of assigning $\mathbf{X}_0$ to Π_2 when $\mathbf{X}_0 \in \Pi_1$. Let $\mathcal{R}(\psi)$ be either the Bayes risk $\mathcal{R}_B(\psi)$ or the maximum risk $\mathcal{R}_M(\psi)$.

Following Ingster et al. [12], we say that *successful classification is possible* if

$$\lim_{b \rightarrow \infty} \inf_{\psi} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_b} \mathcal{R}(\psi) = 0, \quad (11)$$

where the supremum is over all triples $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ from a certain class $\mathcal{S}_b$ consisting of vectors $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ from $\mathbb{R}^p$, and $p \times p$ covariance matrices $\boldsymbol{\Sigma}$, obeying certain restrictions, and the infimum is over all classifiers ψ . The class $\mathcal{S}_b$ of our interest will be specified in Section 5.

Conversely, we say that *successful classification is impossible* if

$$\liminf_{b \rightarrow \infty} \inf_{\psi} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_b} \mathcal{R}(\psi) = \mathcal{R}_{\max},$$

where $\mathcal{R}_{\max} = \min(\pi, 1 - \pi)$ for $\mathcal{R} = \mathcal{R}_B$ with $0 < \pi < 1$ and $\mathcal{R}_{\max} = 1/2$ for $\mathcal{R} = \mathcal{R}_M$. When *successful classification is impossible* we can interpret classification as saying that no decision rule is better than any random guess. If this is the case then any attempt at classification is futile. In this work, we consider in parallel the Bayes risk $\mathcal{R}_B$ with $\pi = 1/2$ and the maximum risk $\mathcal{R}_M$.

4 Some useful statistics and their properties

In order to rigorously introduce the sparse model of our interest, we will need certain statistics obtained from the data $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$. Namely, let us construct b statistics $\hat{T}_1, \dots, \hat{T}_b$ and b statistics $T_1, \dots, T_b$ as follows. For $k = 1, \dots, b$, let $\hat{\boldsymbol{\mu}}_{1,[k]}$ and $\hat{\boldsymbol{\mu}}_{2,[k]}$ be the natural estimators of $\boldsymbol{\mu}_{1,[k]}$ and $\boldsymbol{\mu}_{2,[k]}$ given by (8), that is,

$$\hat{\boldsymbol{\mu}}_{1,[k]} = \frac{2}{n} \sum_{j=1}^{n/2} \mathbf{X}_{j,[k]}^{(1)}, \quad \hat{\boldsymbol{\mu}}_{2,[k]} = \frac{2}{n} \sum_{j=1}^{n/2} \mathbf{X}_{j,[k]}^{(2)},$$

and let

$$\hat{\boldsymbol{\mu}}_{[k]} = \hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}.$$

Next, for $k = 1, \dots, b$, define a pooled estimator $\hat{\boldsymbol{\Sigma}}_{[k]}$ of $\boldsymbol{\Sigma}_{[k]}$ by the formula

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{[k]} = \hat{\boldsymbol{\Sigma}}_{[k],b} &= \frac{1}{n-2} \left(\sum_{j=1}^{n/2} \left(\mathbf{X}_{j,[k]}^{(1)} - \hat{\boldsymbol{\mu}}_{1,[k]} \right) \left(\mathbf{X}_{j,[k]}^{(1)} - \hat{\boldsymbol{\mu}}_{1,[k]} \right)^\top \right. \\ &\quad \left. + \sum_{j=1}^{n/2} \left(\mathbf{X}_{j,[k]}^{(2)} - \hat{\boldsymbol{\mu}}_{2,[k]} \right) \left(\mathbf{X}_{j,[k]}^{(2)} - \hat{\boldsymbol{\mu}}_{2,[k]} \right)^\top \right). \end{aligned} \quad (12)$$

Notice that without the assumption of a block-diagonal structure of $\mathbf{\Sigma}$, we would immediately run into the problem of getting a singular pooled covariance matrix $\hat{\mathbf{\Sigma}}$ for which the usual inverse $\hat{\mathbf{\Sigma}}^{-1}$ does not exist; this fact would dramatically complicate the subsequent analysis. Indeed, to have nonsingular $\hat{\mathbf{\Sigma}}$, we must have $n - 2 \geq p$ (see p. 591 of Johnson and Wichern [14]), which is not true in the high-dimensional settings in hand. This problem is overcome by our assumption on $\mathbf{\Sigma}$, giving, for each k , the nonsingular sample covariance matrix $\hat{\mathbf{\Sigma}}_{[k]}$ as defined in (12).

Now, for $k = 1, \dots, b$, define the statistics

$$\hat{T}_k = \frac{n(n - p_0 - 1)}{4(n - 2)p_0} \hat{\Delta}_k, \quad (13)$$

where

$$\hat{\Delta}_k := \hat{\boldsymbol{\mu}}_{[k]}^\top \hat{\mathbf{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]}$$

is a plug-in estimator of $\Delta_k = \boldsymbol{\mu}_{[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}$. The statistics $\hat{T}_1, \dots, \hat{T}_b$ have the following two properties (see, for example, formulas (8.44)–(8.45) in Section 8 of Izenman [13]):

- (a) $\hat{T}_1, \dots, \hat{T}_b$ are independent;
- (b) for $k = 1, \dots, b$,

$$\hat{T}_k \sim F_{p_0, n - p_0 - 1}(\gamma_k), \quad (14)$$

where

$$\gamma_k := (n/4)\Delta_k = (n/4)\boldsymbol{\mu}_{[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}.$$

In this work, we will refer to γ_k as the *block separation strength* for block k . Here and below, the symbol $F_{\nu_1, \nu_2}(\gamma)$ denotes an F distribution with ν_1 numerator and ν_2 denominator degrees of freedom and noncentrality parameter γ .

Also, assuming that the covariance matrix $\mathbf{\Sigma} = \text{BlkDiag}(\mathbf{\Sigma}_{[1]}, \dots, \mathbf{\Sigma}_{[b]})$ is known, for

$k = 1, \dots, b$, we may consider the statistics

$$T_k = (n/4) \tilde{\Delta}_k, \quad (15)$$

where

$$\tilde{\Delta}_k := \hat{\boldsymbol{\mu}}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]}.$$

It is known (see, for example, formula (8b.2.12) on p. 541 of Rao [16]) that $T_1, \dots, T_b$ are independent, and for $k = 1, \dots, b$,

$$T_k \sim \chi_{p_0}^2(\gamma_k),$$

where the quantity $\gamma_k = (n/4)\Delta_k$ is as in (14) and the symbol $\chi_\nu^2(\gamma)$ denotes a χ^2 distribution with ν degrees of freedom and noncentrality parameter γ .

When $\nu_2 \rightarrow \infty$, the weak limit of $\nu_1 F_{\nu_1, \nu_2}(\lambda)$ is known to be $\chi_{\nu_1}^2(\lambda)$. Therefore, for large b , the distribution of $p_0 \hat{T}_k$ will be ‘close’ to that of T_k , $k = 1, \dots, b$. This observation will be used to construct a classification rule, in which, prior to doing any classification, a careful selection of useful feature blocks has to be done. Analytically, the effective selection (or non-selection) of the k th feature block is performed by comparing the statistic T_k with a suitable threshold. In practice, such a comparison may not be feasible as the use of T_k requires the knowledge of the matrix $\boldsymbol{\Sigma}$. Fortunately, for some values of the parameter θ , the change of T_k for $p_0 \hat{T}_k$ will not visibly affect the quality of selection and hence the quality of the whole classification procedure (see Theorem 5 in Section 12 and Conjecture 1 in Section 13).

Remark 1. In view of the block-diagonal structure of $\boldsymbol{\Sigma}$, the quantity Δ in (3) can be written as

$$\Delta = \sum_{k=1}^b \Delta_k, \quad \text{where} \quad \Delta_k = \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \quad k = 1, \dots, b.$$

Clearly, the smaller the value of Δ is, the harder to distinguish between the two classes, Π_1 and Π_2 , and thus the larger the misclassification error $\mathcal{E}$ is (see formula (3)). Therefore, under our assumptions on Σ , each $\gamma_k = (n/4)\Delta_k$ can be interpreted as a block separation strength, that is, the larger the value of γ_k is, the more the contribution of the k th sub-vector $\mathbf{X}_{[k]}$ of vector $\mathbf{X} = (\mathbf{X}_{[1]}^\top, \dots, \mathbf{X}_{[b]}^\top)^\top$ to separating the classes.

5 Definition of the model

Now, using the fact that the noncentrality parameters of the statistics $\hat{T}_1, \dots, \hat{T}_b$ (and also of the statistics $T_1, \dots, T_b$) introduced in the previous section are equal to $\gamma_k = (n/4)\Delta_k$, $k = 1, \dots, b$, and Remark 1 in Section 4, we arrive at the following definition of a sparse model. By ‘the model’ we mean the initial model, in which n training samples $(\mathbf{X}_i, Y_i)$, $i = 1, \dots, n$, as specified in Section 1, are observed.

Definition 1. Let s and a be two given numbers such that $s \in \{1, \dots, b\}$ and $a > 0$. The model is said to have an (s, a) -sparse block-diagonal structure if the vector $\boldsymbol{\gamma} = (\gamma_k)_{k=1, \dots, b}$ of noncentrality parameters of the statistics $\hat{T}_1, \dots, \hat{T}_b$ defined by formula (13) (or the statistics $T_1, \dots, T_b$ defined by formula (15)) belongs to the set

$$\begin{aligned} \Upsilon_b(s, a) = \{ \boldsymbol{\gamma} \in \mathbb{R}^b : \text{there exists a set } \mathcal{S} \subset \{1, \dots, b\} \text{ with } s \text{ elements} \\ \text{such that } \gamma_k \geq a \text{ for all } k \in \mathcal{S}, \text{ and } \gamma_k = 0 \text{ for all } k \notin \mathcal{S} \}. \end{aligned} \quad (16)$$

In this model, s statistics among $\hat{T}_1, \dots, \hat{T}_b$ have non-central F distributions with non-centrality parameters $\gamma_k \geq a$, and the remaining $(b - s)$ statistics have a central F distribution. Similarly, only s among the statistics $T_1, \dots, T_b$ have non-central χ^2 distributions with non-centrality parameters $\gamma_k \geq a$, whereas the other $(b - s)$ statistics have a central χ^2 distribution.

A suitable range for the parameter a that makes the classification problem interesting is

$$a = 2r \ln b \quad \text{for some } 0 < r < 4.$$

Indeed, the assumption on the block separation strengths $\gamma_k = (n/4)\Delta_k$, $k = 1, \dots, b$, in the form

$$\gamma_k = 2r \ln b$$

implies that for $k = 1, \dots, b$

$$\Delta_k = O(b^{-\theta} \ln b) = o(1), \quad b \rightarrow \infty.$$

This brings us to a non-trivial problem of classification in high dimensions that is closely related to the classification problem for a diagonal matrix $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_{p \times p}$, as studied in Ingster et al. [12] and Donoho and Jin [8]. The restriction on the range of r being the interval $(0, 4)$ is due to a related feature/variable selection problem and is clarified in Section 8.

Also, we assume that the parameter s satisfies, as $b \rightarrow \infty$,

$$s = b^{1-\beta}(1 + o(1)), \quad \text{for some } 0 < \beta < 1,$$

implying $s = o(b)$ as $b \rightarrow \infty$. This type of parametrization for sparsity parameter s is quite common in the literature on variable selection and classification in sparse models. In this case, a small fraction of the statistics $\hat{T}_1, \dots, \hat{T}_b$, formed from the data in the sparse block-diagonal model, have non-central F distributions with non-centrality parameters $\gamma_k \geq a$, whereas the majority of $\hat{T}_1, \dots, \hat{T}_b$ have a central F distribution. Similarly, only a small fraction of the statistics $T_1, \dots, T_b$ have non-central χ^2 distributions with non-centrality parameters $\gamma_k \geq a$, whereas the majority of $T_1, \dots, T_b$ have a central χ^2 distribution. With these arrangements, our definition of an *informative/non-*

informative feature block is as follows (which builds on Definition 1). Recall that the data $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ are n independent copies of the vector $(\mathbf{X}, Y)$ introduced in Section 1.

Definition 2. We say that the k th sub-vector $\mathbf{X}_{[k]}$, $k = 1, \dots, b$, of a normal random vector $\mathbf{X} = (\mathbf{X}_{[1]}^\top, \dots, \mathbf{X}_{[b]}^\top)^\top$ or, what is the same, the k th feature block is *informative* for classification, if the non-centrality parameter γ_k of the statistic $\hat{T}_k$ in (13) (or the statistic T_k in (15)) satisfies $\gamma_k \geq 2r \ln b$ for some $r > 0$. If $\gamma_k = 0$, the k th feature block is said to be *non-informative*.

Note that the model of interest is indexed by two known parameters p_0 and θ and two unknown parameters β and r . The possibility and impossibility of successful classification, as well as the form of a classifier that provides successful classification, will depend on the relationships among the parameters θ , β and r . Depending on the value of β as a function of θ , we distinguish between the cases of moderate, high, and very high β -sparsity (see Section 6). In addition to that, depending on the relationship between r and β , we shall distinguish between the cases of moderate, high, and very high (β, r) -sparsity (see Section 8).

Now, using the above introduced notation, we define the set $\mathcal{S}_b$ of triples $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ as follows (recall that $p_0 \geq 1$ and $0 < \theta < 1$ are known numbers).

Definition 3. For $0 < \beta < 1$, $0 < r < 4$, and $b \geq 1$, put

$$\mathcal{S}_b = \mathcal{S}_{b,\beta,r} = \{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) : \boldsymbol{\mu}_1 \in \mathbb{R}^p, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \boldsymbol{\Sigma} = \text{BlkDiag}(\boldsymbol{\Sigma}_{[1]}, \dots, \boldsymbol{\Sigma}_{[b]}), \text{ and } \boldsymbol{\gamma} = (\gamma_k) \in \Upsilon_b([b^{1-\beta}], 2r \ln b)\},$$

where $p = p_0 b$, $\gamma_k = (b^\theta/4)(\boldsymbol{\mu}_{2,[k]} - \boldsymbol{\mu}_{1,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1}(\boldsymbol{\mu}_{2,[k]} - \boldsymbol{\mu}_{1,[k]})$, and the set $\Upsilon_b(s, a)$ is given by (16).

The goal now is to identify the region(s) inside the *parameter space* $\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1, 0 < r < 4\}$ where successful classification is possible in the sense of definition

(11), and to construct classifier(s) that would provide successful classification is the identified region(s).

In a simpler case of a diagonal covariance matrix $\Sigma = \sigma^2 \mathbf{I}_{p \times p}$, Ingster et al. [12] and Donoho and Jin [8] obtained the boundary, that we call here the *classification boundary*, below which classification is impossible. The classification boundary $r = \rho_3(\beta)$ obtained in their work is given by formula (29). That is, if $r > \rho_3(\beta)$, then successful classification is possible, and if $r < \rho_3(\beta)$, then successful classification is impossible. In our case of a block-diagonal matrix Σ , the classification boundary cannot be lower than the one found in Ingster et al. [12] and Donoho and Jin [8] and, in fact, it is the same. If we know exactly which blocks are informative, the classification boundary will be seen to be $r = 0$ (see Proposition 1 in Section 7).

For technical reasons, we impose the following

Condition A:

for all $k = 1, \dots, b$ for which the noncentrality parameter $\gamma_k \geq 2r \ln b$ for some $0 < r < 4$, at least one of the values $\boldsymbol{\mu}_{1,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{1,[k]}$ and $\boldsymbol{\mu}_{2,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{2,[k]}$ is of the same order of magnitude as $\boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}$ when $b \rightarrow \infty$.

Also, in what follows, we consider the worst case scenario when all nonzero γ_k s are the same and are equal to $2r \ln b$ for some $0 < r < 4$.

6 Three regions of β -sparsity

Consider our sparse model in the sense of Definition 1 with $a = 2r \ln b$ for some $0 < r < 4$ and $s \sim b^{1-\beta}$ for some $0 < \beta < 1$. An important role of the sparsity parameter β in a similar classification problem has been demonstrated by Ingster et al. (2009). Specifically, their scenario (C) corresponds to our model with $\Sigma = \sigma^2 \mathbf{I}_{p \times p}$, $\boldsymbol{\mu}_1 = \mathbf{0}$, and sparse vector $\boldsymbol{\mu}_2 \neq \mathbf{0}$. In case of scenario (C), Ingster et al. (2009) distinguished between the cases of (i) moderate, (ii) high, and (iii) very high β -sparsity, which, in

our notation, can be written as (i) $0 < \beta < (1 - \theta)/2$, (ii) $(1 - \theta)/2 < \beta < 1 - \theta$, and (iii) $1 - \theta < \beta < 1$. We shall now see that, for the sparse block-diagonal model of our interest, the same regions of β -sparsity are to be distinguished.

6.1 The case of moderate β -sparsity

It turns out that, in the case of *moderate β -sparsity*, when $0 < \beta < (1 - \theta)/2$, successful classification is possible without preliminary selecting useful feature blocks and is provided by the classifier, cf. (2),

$$\Psi = \mathbb{I} \left\{ \sum_{k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} \leq 0 \right\},$$

which depends on the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$. (At this point, we assume that $\boldsymbol{\Sigma}$ is known. In Section 13, we will comment on what happens when, in the definition of this and the other classifiers proposed in this thesis, the covariance matrix $\boldsymbol{\Sigma} = \text{BlkDiag}(\boldsymbol{\Sigma}_{[1]}, \dots, \boldsymbol{\Sigma}_{[b]})$ is replaced by its estimator $\hat{\boldsymbol{\Sigma}} = \text{BlkDiag}(\hat{\boldsymbol{\Sigma}}_{[1]}, \dots, \hat{\boldsymbol{\Sigma}}_{[b]})$.)

To show the validity of this claim, we rewrite Ψ as follows:

$$\begin{aligned} \Psi = \mathbb{I} \left\{ \sum_{k=1}^b \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} + \sum_{k=1}^b \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) \right. \\ \left. - \frac{1}{2} \sum_{k=1}^b \left[(\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right] \right. \\ \left. \leq \frac{1}{2} \sum_{k=1}^b (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}. \end{aligned}$$

Then, denoting for all $k = 1, \dots, b$

$$\begin{aligned}\xi_k &= \xi_{k,b} := \mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \\ \eta_k &= \eta_{k,b} := \mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}), \\ \tau_k &= \tau_{k,b} := (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \Sigma_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]},\end{aligned}$$

we may continue

$$\Psi = \mathbb{I} \left\{ \sum_{k=1}^b (\xi_k + \eta_k - \tau_k/2) \leq \frac{1}{2} \sum_{k=1}^b (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}. \quad (17)$$

Our derivations in this subsection have much in common with the proof of Proposition 1, stated in Section 7 and proved in Section 16, to which the reader will occasionally be referred. Although, in the proof of Proposition 1, we deal with the random variables ξ_k , η_k , and τ_k for $k = 1, \dots, b$ such that $\omega_k = 1$, their distributional properties for $k = 1, \dots, b$ such that $\omega_k = 0$ are also immediately obtained.

First, we can see (for details, we refer to the proof of Proposition 1 in Section 16) that, under Π_2 , for all $k = 1, \dots, b$ such that $\omega_k = 1$

$$\xi_k \sim N \left(\boldsymbol{\mu}_{2,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right),$$

and for all $k = 1, \dots, b$ such that $\omega_k = 0$

$$\xi_k = \boldsymbol{\mu}_{2,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \quad \text{with probability one.}$$

(Here we use the usual convention that a random variable Y has a normal distribution with mean m and variance $\sigma^2 = 0$ when $\mathbf{P}(Y = m) = 1$.) Also, for all $k = 1, \dots, b$

$$\eta_k \stackrel{\text{d}}{=} \mathbf{U}_k^\top \mathbf{V}_k,$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are independent normal random vectors satisfying (under Π_2)

$$\mathbf{U}_k \sim N_{p_0} \left(((n/4)\boldsymbol{\Sigma}_{[k]})^{-1/2} \boldsymbol{\mu}_{2,[k]}, (4/n)\mathbf{I} \right), \quad \mathbf{V}_k \sim N_{p_0}(\mathbf{0}, \mathbf{I}),$$

with $\mathbf{I}$ being the $p_0 \times p_0$ identity matrix. Finally (for details, we again refer to the proof of Proposition 1 in Section 16), under Π_2 , for all $k = 1, \dots, b$

$$\tau_k = (4/n)\mathbf{S}_k^\top \mathbf{T}_k + W_k,$$

where $\mathbf{S}_k$ and $\mathbf{T}_k$ are dependent normal vectors such that

$$\mathbf{S}_k \sim N_{p_0} \left(((4/n)\boldsymbol{\Sigma}_{[k]})^{-1/2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]}), \mathbf{I} \right), \quad \mathbf{T}_k \sim N_{p_0}(0, \mathbf{I}),$$

for all $k = 1, \dots, b$ for which $\omega_k = 1$

$$W_k \sim N \left(0, (4/n)\boldsymbol{\mu}_{[k]} \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right),$$

and for all $k = 1, \dots, b$ for which $\omega_k = 0$

$$W_k = 0 \quad \text{with probability one.}$$

It then follows that as $b \rightarrow \infty$

$$\begin{aligned} \sum_{k=1}^b \mathbf{E}_{\Pi_2}(\xi_k + \eta_k - (\tau_k/2)) &= \sum_{k=1}^b \boldsymbol{\mu}_{2,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - (2/n) \sum_{k=1}^b \mathbf{E}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k) \\ &= \left(\sum_{k=1}^b \boldsymbol{\mu}_{2,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) (1 + o(1)), \end{aligned}$$

because $(2/n) \sum_{k=1}^b \mathbf{E}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k) = O(b^{1-\theta})$ and, in view of Condition A, $\sum_{k=1}^b \boldsymbol{\mu}_{2,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} =$

$O(b^{1-\theta} \ln b)$. Therefore, using (17), we can write as $b \rightarrow \infty$

$$\Psi = \mathbb{I} \left\{ \sum_{k=1}^b \left[\left(\xi_k + \eta_k - \frac{\tau_k}{2} \right) - \mathbf{E}_{\Pi_2} \left(\xi_k + \eta_k - \frac{\tau_k}{2} \right) \right] \leq -\frac{1}{2} \left(\sum_{k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) (1 + o(1)) \right\}. \quad (18)$$

The analysis of the variance term $\mathbf{Var}_{\Pi_2}(\xi_k + \eta_k - \tau_k/2)$, similar to that in the proof of Proposition 1 in Section 16, gives that for all $k = 1, \dots, b$

$$\begin{aligned} \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k - \tau_k/2) &= \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k) + (1/4) \mathbf{Var}_{\Pi_2}(\tau_k) - \mathbf{Cov}_{\Pi_2}(\xi_k + \eta_k, \tau_k) \\ &= O(b^{-\theta} \ln b), \end{aligned}$$

and hence as $b \rightarrow \infty$

$$\sum_{k=1}^b \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k - \tau_k/2) = O(b^{1-\theta} \ln b). \quad (19)$$

Now, applying Chebyshev's inequality and noting that by assumption $\sum_{k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]} \boldsymbol{\mu}_{[k]} = O(b^{1-\beta-\theta} \ln b)$, we obtain from (18) and (19) that as $b \rightarrow \infty$

$$\begin{aligned} &\mathbf{E}_{\Pi_2}(\Psi) \\ &= \mathbf{P}_{\Pi_2} \left(\sum_{k=1}^b [(\xi_k + \eta_k - \tau_k/2) - \mathbf{E}_{\Pi_2}(\xi_k + \eta_k - \tau_k/2)] \leq -\frac{1}{2} \left(\sum_{k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]} \boldsymbol{\mu}_{[k]} \right) (1 + o(1)) \right) \\ &\leq \frac{4 \sum_{k=1}^b \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k - \tau_k/2)}{(b^{1-\beta-\theta} \ln b)^2 (1 + o(1))} = \frac{O(b^{1-\theta} \ln b)}{O(b^{2-2\beta-2\theta} \ln^2 b)} \\ &\asymp \frac{1}{b^{1-2\beta-\theta} \ln b}. \end{aligned}$$

It is now seen that if $1 - 2\beta - \theta > 0$, or equivalently $0 < \beta < (1 - \theta)/2$, then

$$\mathbf{E}_{\Pi_2}(\Psi) \rightarrow 0 \quad \text{as } b \rightarrow \infty.$$

Similarly, one can show that if $0 < \beta < (1 - \theta)/2$, then

$$\mathbf{E}_{\Pi_1}(1 - \Psi) \rightarrow 0, \quad b \rightarrow \infty.$$

The obtained results are uniform in $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b,\beta,r}$ for all pairs $(\beta, r) \in (0, (1 - \theta)/2) \times (0, 4)$, where $\mathcal{S}_{b,\beta,r}$ is as in Definition 3 with

$$\begin{aligned} \Upsilon_b(s, a) = \{ & \boldsymbol{\gamma} \in \mathbb{R}^b : \text{there exists a set } \mathcal{S} \subset \{1, \dots, b\} \text{ with } s \text{ elements} \\ & \text{such that } \gamma_k = a \text{ for all } k \in \mathcal{S}, \text{ and } \gamma_k = 0 \text{ for all } k \notin \mathcal{S} \}, \end{aligned}$$

implying that all nonzero γ_k s are the same and are equal to $2r \ln b$ for some $0 < r < 4$.

According to formulas (9) and (10), the risk $\mathcal{R}(\Psi)$ is either the convex combination or the maximum of $\mathbf{E}_{\Pi_2}(\Psi)$ and $\mathbf{E}_{\Pi_1}(1 - \Psi)$. Therefore, uniformly in $(\beta, r) \in (0, (1 - \theta)/2) \times (0, 4)$,

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b,\beta,r}} \mathcal{R}(\Psi) \rightarrow 0, \quad b \rightarrow \infty,$$

provided Condition A is met.

6.2 The case of high β -sparsity

In the case of *high β -sparsity*, when $(1 - \theta)/2 < \beta < 1 - \theta$, prior to doing classification, the selection of useful feature blocks is required. To explain this phenomenon in simpler terms, we assume that $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are known. Then, the information regarding useful feature blocks is available, and we know which of the quantities $\omega_k = \mathbb{I}(\gamma_k \neq 0)$, $k = 1, \dots, b$, are nonzero. Consider the (oracle) classifier, cf. (2),

$$\Psi^*(\mathbf{X}_0) = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\mathbf{X}_{0,[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \leq 0 \right\},$$

where $\omega_k = \mathbb{I}(\gamma_k \neq 0)$ is one of the k th feature block is useful and zero otherwise. Noting that, under Π_2 , for all $k = 1, \dots, b$ for which $\omega_k = 1$

$$\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \sim N \left(\boldsymbol{\mu}_{2,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right),$$

we obtain that

$$\begin{aligned} & \mathbf{E}_{\Pi_2} (\Psi^*(\mathbf{X}_0)) \\ &= \mathbf{P}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b \left[\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - \mathbf{E}_{\Pi_2} \left(\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \right] \leq -\frac{1}{2} \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \\ &\leq \frac{4 \sum_{k=1: \omega_k=1}^b \mathbf{Var}_{\Pi_2} \left(\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right)}{\left(\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right)^2} = \frac{4 \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}}{\left(\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right)^2} \\ &= \frac{O(b^{1-\beta-\theta} \ln b)}{O(b^{2-2\beta-2\theta} \ln^2 b)} \asymp \frac{1}{b^{1-\beta-\theta} \ln b}. \end{aligned}$$

Hence, in the high β -sparsity case when $(1-\theta)/2 < \beta < 1-\theta$ (and hence $1-\beta-\theta > 0$), we obtain that, uniformly in $(\beta, r) \in ((1-\theta)/2, 1-\theta) \times (0, 4)$,

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b,\beta,r}} \mathbf{E}_{\Pi_2} (\Psi^*(\mathbf{X}_0)) \rightarrow 0, \quad b \rightarrow \infty,$$

and, using similar arguments,

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b,\beta,r}} \mathbf{E}_{\Pi_1} (1 - \Psi^*(\mathbf{X}_0)) \rightarrow 0, \quad b \rightarrow \infty.$$

This gives

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b,\beta,r}} \mathcal{R}(\Psi^*) \rightarrow 0, \quad b \rightarrow \infty,$$

uniformly in $(\beta, r) \in ((1-\theta)/2, 1-\theta) \times (0, 4)$.

6.3 The case of very high β -sparsity

Unlike the *high β -sparsity* case, in the case of *very high β -sparsity* when $1 - \theta < \beta < 1$, successful classification is impossible. This follows, for example, from Theorem 6.4 of Ingster et al. [12], where the model under study is asymptotically equivalent to our model with $\Sigma = \sigma^2 \mathbf{I}_{p \times p}$, $\boldsymbol{\mu}_1 = \mathbf{0}$, and sparse vector $\boldsymbol{\mu}_2 \neq \mathbf{0}$. Scenario (C) treated in that theorem corresponds to our assumptions that $n \sim b^\theta$ for some $0 < \theta < 1$ and $b \rightarrow \infty$.

7 Classifiers with partially estimated parameters

Let us now show that replacing $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ by $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$, respectively, will not change the property of the classifier $\Psi^*(\mathbf{X}_0)$ defined in Section 6.2 to have the classification errors tending to zero as $b \rightarrow \infty$.

Our aim now is to show that the risk of the classification rule

$$\hat{\Psi} = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \Sigma_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}, \quad (20)$$

which is a function of the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, tends to zero as b goes to infinity. Recall that we assume that $\mathcal{S}_b = \mathcal{S}_{b,\beta,r}$ is as in Definition 3 with

$$\Upsilon_b(s, a) = \{\boldsymbol{\gamma} \in \mathbb{R}^b : \text{there exists a set } \mathcal{S} \subset \{1, \dots, b\} \text{ with } s \text{ elements} \\ \text{such that } \gamma_k = a \text{ for all } k \in \mathcal{S}, \text{ and } \gamma_k = 0 \text{ for all } k \notin \mathcal{S}\},$$

implying that all nonzero γ_k s are the same and are equal to $2r \ln b$ for some $0 < r < 4$.

Hence

$$\boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} = (\boldsymbol{\mu}_{2,[k]} - \boldsymbol{\mu}_{1,[k]})^\top \Sigma_{[k]}^{-1} (\boldsymbol{\mu}_{2,[k]} - \boldsymbol{\mu}_{1,[k]}) = 8rb^{-\theta} \ln b.$$

In Section 13, we shall also comment on what happens if we replace each indicator $\omega_k = \mathbb{I}(\gamma_k \neq 0)$ by a suitable estimator $\hat{\omega}_k$ and each $\Sigma_{[k]}$ by its pooled estimator $\hat{\Sigma}_{[k]}$. In the classical books on multivariate statistical analysis, the use of sample estimators in place of respective unknown parameters has not been rigorously justified. For example, in Section 6.5 of Anderson [1] we read that substituting the estimators for the parameters “seems intuitively reasonable ... and should give good results”. Here is our result showing that, for all values of $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$, the classification rule $\hat{\Psi}$ in (20) provides successful classification.

Proposition 1. *For the sparse block-diagonal normal model under consideration, assuming that Condition A holds true, uniformly in $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$,*

$$\lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\hat{\Psi}) = 0.$$

Proposition 1 implies that, uniformly in $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$,

$$\liminf_{b \rightarrow \infty} \inf_{\Psi} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\Psi) \leq \lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\hat{\Psi}) = 0.$$

In other words, for all values of $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$, successful classification is possible and is provided by the classifier $\hat{\Psi}$ as in (20). For the proof of Proposition 1, we refer to Section 16.

Remark 2. As mentioned earlier, if we do not know which out of b feature blocks are informative, then within the high β -sparsity region successful classification is possible only for $\rho_3(\beta) < r < 4$, where $\rho_3(\beta)$ is defined by (29). At the same time, as follows from Proposition 1, if we know which feature blocks are informative, then within the high β -sparsity region successful classification is possible for all $0 < r < 4$, and is provided by the classifier $\hat{\Psi}$ defined in (20).

8 Feature selection prior to classification

Throughout the rest of this thesis, we will explore the different options available for achieving the highest level of classification accuracy.

Heuristically, we need to use as many informative features as possible and minimize the number of non-informative, so that we do not turn this classification process into a random guess. To do this, we can first transform the data from the multivariate normal observations to the chi-square distributed observations. Assume, for a while, we know the covariance matrix Σ and consider the statistics as in (15):

$$T_k = (n/4) \tilde{\Delta}_k, \quad k = 1, \dots, b,$$

where

$$\tilde{\Delta}_k = \hat{\boldsymbol{\mu}}_{[k]}^\top \Sigma_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]}.$$

As mentioned in Section 4, $T_1, \dots, T_b$ are independent random variables from a chi-square distribution with p_0 degrees of freedom and non-centrality parameter γ_k (recall $\gamma_k = 0$ if the k th feature is non-informative and $\gamma_k = 2r \ln b$ if the k th feature is informative). Namely,

$$T_k \sim \chi_{p_0}^2(\gamma_k), \quad k = 1, \dots, b,$$

where $\gamma_k = (n/4)\Delta_k$ with $\Delta_k = \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}$, and we assume that all nonzero γ_k are the same and such that $\gamma_k = 2r \ln b$ for some $r > 0$ (the worst case scenario).

As we have assumed that the fraction of informative blocks yielding $\gamma_k = 2r \ln b$ is $\varepsilon_b = b^{-\beta}$ and the fraction of non-informative blocks yielding $\gamma_k = 0$ is $1 - \varepsilon_b = 1 - b^{-\beta}$, we can see that, after sampling with replacement, the transformed observations $T_1, \dots, T_b$

will follow a chi-square mixture model

$$T_1, \dots, T_b \stackrel{\text{iid}}{\sim} (1 - \varepsilon_b)\chi_{p_0}^2(0) + \varepsilon_b\chi_{p_0}^2(\gamma_b), \quad (21)$$

where we want to identify, if possible, all the non-centrally distributed statistics T_k with $\gamma_k = 2r \ln b$ for some $r > 0$.

It is known that any non-central chi-square random variable tends to be larger than a central chi-square random variable with the same number of degrees of freedom:

$$\mathbf{P}(\chi_\nu^2(\gamma) \geq t) > \mathbf{P}(\chi_\nu^2(0) \geq t), \quad \text{for any } t > 0, \gamma > 0.$$

The idea now is that we will only consider the statistics, among $T_1, \dots, T_b$, whose observed values are greater than a certain threshold t^* . We will go into more detail later on about how to choose this threshold. The reason we only want to look at $T_1, \dots, T_b$ that are greater than a specific threshold t^* is because larger values of Δ_k will lead to larger values of T_k with noncentrality parameter $\gamma_k = (n/4)\Delta_k$. If that is the case, then the squared distance $\Delta_{[k]}$ between $\boldsymbol{\mu}_{2,[k]}$ and $\boldsymbol{\mu}_{1,[k]}$ is more noticeable, making the classification of the k th feature to be more accurate since this k th feature is more likely to be informative.

In this thesis, in order to identify informative feature blocks, we will be using some recent results of hypothesis testing and variable selection in high dimensions that were originally developed for a similar normal mixture model, cf. (21),

$$Z_1, \dots, Z_b \stackrel{\text{iid}}{\sim} (1 - \varepsilon_b)N(0, 1) + \varepsilon_b N(\mu_b, 1), \quad (22)$$

where $\varepsilon_b = b^{-\beta}$ with $0 < \beta < 1$ and $\mu_b = \sqrt{2r \ln b}$ with $0 < r < 4$. The parameters β and r are generally unknown. We shall see the connection between models (21) and (22) below through relations (23)–(25). The next result, together with (23) and (25),

will be used to get an idea of the threshold level t^* that eliminates (as many as possible) non-informative feature blocks from the subsequent classification process.

Lemma 1. *If $\xi_1, \xi_2, \dots$ is a sequence of iid normal $N(0, 1)$ random variables, then*

$$\mathbf{P}\left(\max_{1 \leq i \leq b} |\xi_i| > \sqrt{2 \ln b}\right) \rightarrow 0 \text{ as } b \rightarrow \infty.$$

The proof of Lemma 1 can be seen in Section 16. Lemma 1 is a basis for effective variable selection in the normal mixture model (22). Here are some graphs to illustrate Lemma 1.

Figure 1: Random sample of $b = 100$ from $N(0, 1)$

Figure 2: Random sample of $b = 1000$ from $N(0, 1)$

Figure 3: Random sample of $b = 10000$ from $N(0, 1)$

Figure 4: Random sample of $b = 100000$ from $N(0, 1)$

The red vertical lines mark the points $\pm\sqrt{2 \ln b}$. Lemma 1 shows that since $\mu_b = \sqrt{2r \ln b}$ is only moderately large, the normal $N(\mu_b, 1)$ observations in model (22) can be easily mixed with the extreme normal $N(0, 1)$ observations, especially when $0 < r < 1$.

We have the following results that make a bridge between the chi-square mixture model (21) and the normal mixture model (22) (see Section 5 of Donoho and Jin [7]): as $b \rightarrow \infty$

$$\mathbf{P}(\chi_\nu^2(0) > 2s \ln b) = \mathbf{P}(N(0, 1) > \sqrt{2s \ln b})(1 + o(1)) \quad (23)$$

$$= O(b^{-s}(\ln b)^{-1/2}), \quad 0 < s < \infty,$$

$$\mathbf{P}(\chi_\nu^2(2r \ln b) > 2s \ln b) = \mathbf{P}(N(\sqrt{2r \ln b}, 1) > \sqrt{2s \ln b})(1 + o(1)) \quad (24)$$

$$= O(b^{-(\sqrt{s}-\sqrt{r})^2}(\ln b)^{-1/2}), \quad 0 < r < s < \infty,$$

where $\nu \geq 1$ is a fixed number. Relations (23) and (24) show that, asymptotically, the chi-square tail probabilities exhibit behaviour that is analogous to that of the corresponding normal tail probabilities. In addition, we will need the following result, cf. (24).

Lemma 2. *For all $\nu \geq 1$ as $b \rightarrow \infty$*

$$\begin{aligned} \mathbf{P}(\chi_\nu^2(2r \ln b) \leq 2s \ln b) &= \mathbf{P}\left(N(\sqrt{2r \ln b}, 1) \leq \sqrt{2s \ln b}\right)(1 + o(1)) \\ &= O\left(b^{-(\sqrt{s}-\sqrt{r})^2}(\ln b)^{-1/2}\right), \quad 0 < s < r < \infty. \end{aligned} \quad (25)$$

The proof of Lemma 2 can be found in Section 16. The use of relations (23) and (25) makes it possible to translate the variable selection results of Theorems 1–4 below that have been derived for the normal mixture model (22) to the chi-square mixture model (21). The translated results will then be applied to set up a threshold t^* that is expected to separate the informative and non-informative feature blocks.

In this thesis, we are focusing on the interesting case of *high β -sparsity* discussed in

Section 6. In this case we will be concentrating on the classification when $0 < r < 4$ and $(1 - \theta)/2 < \beta < 1 - \theta$, where $\theta \in (0, 1)$ is a known parameter. Since the parameter r has such a dramatic affect on the noncentrality parameter γ , for larger values of r the problem of classification becomes trivial. When $r \geq 4$ exact classification can be achieved without any preliminary work since the non-centrality parameter $\gamma = 2r \ln b$ is large enough to achieve successful classification using rule (2) without any preliminary work (see, for example, Section 2 of Ingster et al. [12]). For $0 < r < 4$ we need to first try to identify which feature blocks are informative, and then use these informative blocks in our classification rule (7) with a good estimator $\hat{\omega}_k$ of ω_k , $k = 1, \dots, b$.

In connection with identifying informative feature blocks, we first consider a more general model from which the normal mixture model (22) follows. Namely, let

$$Z_j = m_j + \varepsilon_j, \quad j = 1, \dots, b, \quad (26)$$

where $\varepsilon_1, \dots, \varepsilon_b \stackrel{iid}{\sim} N(0, 1)$ and $m_j = \eta_j \sqrt{2r \ln b}$ with $\eta = (\eta_1, \dots, \eta_b)$ such that $\eta_j \in \{0, 1\}$ and $\sum_{j=1}^b \eta_j \sim b^{1-\beta}$. Denote the set of such vectors η by $\mathcal{H}_{b,\beta}$. In the context of a variable selection problem, we wish to estimate $\eta = (\eta_1, \dots, \eta_b) \in \mathcal{H}_{b,\beta}$ with an estimator $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_b)$ whose components take only two values, zero and one. We will refer to such an estimator $\hat{\eta}$ as a *selector*.

In connection with model (26), the following four theorems will give us a better idea of what type of selection is possible through different regions of the β - r plane and can be seen in higher detail in Theorems 2–5 of Cui [6]. Theorems 1–4 below are direct consequences of Theorems 2–5 in Cui [6], where a more general model than model (26) is treated.

Theorem 1. *Let $r > (1 + \sqrt{1 - \beta})^2$ and $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_b)$, where $\hat{\eta}_j = \mathbb{I}(Z_j > \sqrt{(2 + \epsilon) \ln b})$, $j = 1, \dots, b$, with $\epsilon = \epsilon_b > 0$ satisfying the properties $\epsilon \rightarrow 0$ and*

$\epsilon \ln b \rightarrow \infty$. Then, uniformly in $0 < \beta < 1$,

$$\limsup_{b \rightarrow \infty} \sup_{\eta \in \mathcal{H}_{b,\beta}} \mathbf{E}_\eta |\hat{\eta} - \eta| = 0.$$

This means that $\sup_{\eta \in \mathcal{H}_{b,\beta}} \mathbf{E}_\eta |\hat{\eta} - \eta| = o(1)$ as $b \rightarrow \infty$, and thus $\hat{\eta}$ provides the so-called *exact variable selection*.

Theorem 2. *If $r < (1 + \sqrt{1 - \beta})^2$ then no selector can provide exact variable selection, that is, uniformly in $0 < \beta < 1$,*

$$\liminf_{b \rightarrow \infty} \inf_{\eta^*} \sup_{\eta \in \mathcal{H}_{b,\beta}} \mathbf{E}_\eta |\eta^* - \eta| > 0,$$

where the infimum is taken over all selectors η^* of a vector η .

Even though exact selection is impossible for $r < (1 + \sqrt{1 - \beta})^2$, we can still do something as shown in the next theorem.

Theorem 3. *Let $r > \beta$ and $\tilde{\eta} = (\tilde{\eta}_1, \dots, \tilde{\eta}_b)$, where $\tilde{\eta}_j = \mathbb{I}(Z_j > \sqrt{(2\beta + \epsilon) \ln b})$, $j = 1, \dots, b$, with $\epsilon = \epsilon_b > 0$ satisfying the properties $\epsilon \rightarrow 0$ and $\epsilon \ln b \rightarrow \infty$. Then, uniformly in $0 < \beta < 1$,*

$$\limsup_{b \rightarrow \infty} \sup_{\eta \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}_\eta |\tilde{\eta} - \eta| = 0.$$

This means that $\sup_{\eta \in \mathcal{H}_{b,\beta}} \mathbf{E}_\eta |\tilde{\eta} - \eta| = o(b^{1-\beta})$ as $b \rightarrow \infty$, and thus $\tilde{\eta}$ provides the so-called *almost full variable selection*. If r is too small, neither exact nor almost full selection is possible as stated in the next theorem.

Theorem 4. *If $r < \beta$, then no selector can provide almost full selection, that is, uniformly in $0 < \beta < 1$,*

$$\liminf_{b \rightarrow \infty} \inf_{\eta^*} \sup_{\eta \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}_\eta |\eta^* - \eta| > 0,$$

where the infimum is taken over all selectors η^* of a vector η .

Although Theorems 1–4 were developed for a normal mixture model, inspection of their proofs show that using relations (23) and (25), we can easily translate these theorems to a chi-square mixture model. These four theorems contribute to the consideration of the following four regions within the *high β -sparsity* case (see Figures 6–9):

$$\begin{aligned}\mathbf{D}_1 &= \{(\beta, r) \in \mathbb{R}^2 : (1 - \theta)/2 < \beta < 1 - \theta, \rho_1(\beta) < r < 4\}; \\ \mathbf{D}_2 &= \{(\beta, r) \in \mathbb{R}^2 : (1 - \theta)/2 < \beta < 1 - \theta, \rho_2(\beta) < r < \rho_1(\beta)\}; \\ \mathbf{D}_3 &= \{(\beta, r) \in \mathbb{R}^2 : (1 - \theta)/2 < \beta < 1 - \theta, \rho_3(\beta) < r < \rho_2(\beta)\}; \\ \mathbf{D}_4 &= \{(\beta, r) \in \mathbb{R}^2 : (1 - \theta)/2 < \beta < 1 - \theta, 0 < r < \rho_3(\beta)\};\end{aligned}$$

where $0 < \theta < 1$ is known, $\rho_3(\beta)$ is defined in (29), and

$$\rho_1(\beta) = (1 + \sqrt{1 - \beta})^2 \quad \text{and} \quad \rho_2(\beta) = \beta. \quad (27)$$

The regions $\mathbf{D}_1$ and $\mathbf{D}_2$ will be called *selection regions* since inside these regions we can identify which observations in model (21) are from the non-central distribution with a certain degree of accuracy. This makes the classification problem much easier in these two regions. In the region $\mathbf{D}_3$ we may not be able to identify the non-central observations among $T_1, \dots, T_b$ but, by means of a hypothesis testing approach, we may decide whether or not $T_1, \dots, T_b$ obey model (21) and find a threshold that is believed to separate the non-central observations from the central ones. Namely, consider testing the null hypothesis

$$H_0 : X_1, \dots, X_n \stackrel{iid}{\sim} \chi_{p_0}^2(0)$$

versus a sequence of alternatives, cf. (21),

$$H_1 : X_1, \dots, X_n \stackrel{iid}{\sim} (1 - \varepsilon)\chi_{p_0}^2(0) + \varepsilon\chi_{p_0}^2(\gamma) \quad (28)$$

where $\varepsilon = \varepsilon_b = b^{-\beta}$ for some $0 < \beta < 1$ and $\gamma = \gamma_b = \sqrt{2r \ln b}$ for some $0 < r < 1$. Then there exists a test statistic (say, a weighted one-sided Kolmogorov-Smirnov statistic of asymptotic level α as defined in Theorem 4.2 of Stepanova and Pavlenko [19]) that distinguishes between H_0 and H_1 . This statistic works well in the case when $r > \rho(\beta)$, where

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

The function $r = \rho(\beta)$ is called the *detection boundary*. It was, for the first time, obtained in Ingster [12]. When $r < \rho(\beta)$ successful hypothesis testing is not possible, i.e., no consistent test exists, as the noncentrality parameter γ is too small. It turns out (see, Ingster et al. [12] and Donoho and Jin [8]) that the classification boundary $r = \rho_3(\beta)$ is just a *scaled detection boundary* $\rho(\beta)$, that is,

$$\rho_3(\beta) = (1 - \theta) \rho\left(\frac{\beta}{1 - \theta}\right) \quad \text{for } (1 - \theta)/2 < \beta < 1 - \theta, \quad (29)$$

or equivalently

$$\rho_3(\beta) = \begin{cases} (1 - \theta) \left(\beta/(1 - \theta) - 1/2 \right), & (1 - \theta)/2 < \beta < 3(1 - \theta)/4, \\ (1 - \theta) \left(1 - \sqrt{1 - \beta/(1 - \theta)} \right)^2, & 3(1 - \theta)/4 < \beta < (1 - \theta). \end{cases}$$

In the region $\mathbf{D}_3$ classification is still possible, but in the region $\mathbf{D}_4$ successful classification is impossible. In the region $\mathbf{D}_4$ the non-centrality parameter $\gamma = 2r \ln b$ is too small (because r is too small) to distinguish the non-central observations from the central ones (see Theorem 6.4 of Ingster et al. [12]).

Here is a visual representation of the combined “phase space” of the variable selection and signal detection problems at hand.

Figure 5: “Phase spaces” of the related variable selection and signal detection problems.

Figures 6–9 give some visual representations of the four regions of classification for the high β -sparsity case and different values of the parameter θ that relates the sample size n to the driving parameter b through the relation $n \sim b^\theta$.

Figure 6: Various regions of classification with $\theta = 0.01$

Figure 7: Various regions of classification with $\theta = 0.25$

Figure 8: Various regions of classification with $\theta = 0.5$

Figure 9: Various regions of classification with $\theta = 0.75$

Figures 6–9 are some examples of the regions with different values of $\theta \in (0, 1)$. It is easy to see that, as the parameter r decreases, the classification problem becomes more difficult since the distance $\Delta_{[k]} = (8r \ln b)/n$, where the k th feature block is informative, is decreased as well. This makes it more difficult to separate the informative feature blocks from the non-informative blocks. Here are some examples of the chi-square mixture model (21) which is derived from the transformed variables $T_1, \dots, T_b$ as in (15).

Figure 10: Central and non-central chi-square distributions in the region $\mathbf{D}_1$ where exact selection is possible.

Figure 11: Central and non-central chi-square distributions in the region $\mathbf{D}_2$ where almost full selection is possible.

Figure 12: Central and non-central chi-square distributions in the region $\mathbf{D}_3$ where neither exact nor almost full selection is possible.

Figure 13: Central and non-central chi-square distributions in the region $\mathbf{D}_4$ where classification is impossible.

Figures 10–13 give a visual representation of the different classification regions, $\mathbf{D}_1$ through $\mathbf{D}_4$. The red function is the probability density function (PDF) of the central chi-square distribution $\chi_{p_0}^2(0)$, while the blue function is the PDF of the non-central chi-square distribution $\chi_{p_0}^2(\gamma)$, where $\gamma = 2r \ln b$. We can see that the problem of feature selection can become very difficult since the non-central chi-square distribution tends to be closer to the corresponding central chi-square distribution as r decreases.

9 Obtaining threshold t^* in the case of $r > \beta$

In this section, we will discuss the methods of obtaining our threshold t^* for the regions $\mathbf{D}_1$ and $\mathbf{D}_2$.

Before we can discuss classification, we need to address the problem of assessing the closeness of a selector $\omega^* = (\omega_k^*)_{k=1,\dots,b}$ to the vector $\omega = (\omega_k)_{k=1,\dots,b}$, where $\omega_k = \mathbb{I}(\mathbf{X}_{[k]} \text{ is informative})$. We do that by using the *Hamming distance*, which counts the number of positions at which ω^* and ω differ:

$$|\omega^* - \omega| = \sum_{k=1}^b |\omega_k^* - \omega_k|.$$

The problem of feature or variable selection with Hamming loss in high dimensions has been extensively studied in the literature (see, for example, [6], [4], [20], [9]), primarily in normal settings, with the majority of the observations coming from a standard normal distribution and a small fraction of observations from a normal distribution with mean $\mu_b \asymp \sqrt{\ln b}$ and variance one. By sampling from the above set of observations with or without replacement, which asymptotically, as $b \rightarrow \infty$, makes no difference, we arrive at the model, cf. (22):

$$t_1, \dots, t_b \stackrel{\text{iid}}{\sim} (1 - \varepsilon_b)N(0, 1) + \varepsilon_b N(\mu_b, 1), \quad (30)$$

where $\varepsilon_b = b^{-\beta}$ with $0 < \beta < 1$, and $\mu_b = \sqrt{2r \ln b}$ with $0 < r < 4$. The parameters β and r are, generally speaking, unknown.

For this model, exact and almost full variable selection (i.e., identification of nonzero mean observations with a certain degree of precision) is known to be possible only if r is not too small, namely, if $r > \beta$ for all $0 < \beta < 1$ (see, for example, Theorem 3 in Section 8). More precisely, if $r > \beta$ then there exists a selector $\hat{\omega} = (\hat{\omega}_1, \dots, \hat{\omega}_b)$ that correctly identifies most of the nonzero mean observations among $t_1, \dots, t_b$ in (30), in

the sense that, uniformly in $0 < \beta < 1$,

$$\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E}|\hat{\omega} - \omega| = o(b^{1-\beta}), \quad b \rightarrow \infty,$$

where

$$\omega = (\omega_k)_{k=1,\dots,b} := (\mathbb{I}(\mu_k \neq 0))_{k=1,\dots,b},$$

with μ_k being the mean of t_k in model (30), and

$$\mathcal{H}_{b,\beta} = \{\omega = (\omega_1, \dots, \omega_b) \in \mathbb{R}^b : \omega_k \in \{0, 1\}, \sum_{k=1}^b \omega_k \sim b^{1-\beta}\} \quad (31)$$

for the set of sparse vectors ω with most components equal to zero (see Theorem 3 in Section 8).

For larger values of r , not only almost full but also exact variable selection is possible. Namely, if $r > \rho_2(\beta)$ where $\rho_2(\beta)$ is given in (27), then there exists a selector $\tilde{\omega}$ that correctly identifies the nonzero mean observations among $t_1, \dots, t_b$, in the sense that the *maximum Hamming risk* of $\tilde{\omega}$ satisfies

$$\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E}|\tilde{\omega} - \omega| = o(1), \quad b \rightarrow \infty,$$

uniformly in $0 < \beta < 1$ (see Theorem 1 in Section 8).

Similarly, one may consider the problem of identifying the relevant (non-central) variables in a chi-square mixture model, cf. (21),

$$t_1, \dots, t_b \stackrel{\text{iid}}{\sim} (1 - \varepsilon_b)\chi_\nu^2(0) + \varepsilon_b\chi_\nu^2(\gamma_b), \quad (32)$$

where $\nu \geq 1$ is a fixed number, $\varepsilon_b = b^{-\beta}$ with $0 < \beta < 1$, and $\gamma_b = 2r \ln b$ with $0 < r < 4$.

The connection between models (30) and (32) can be seen from relations (23) and (25). The use of relation (23) and (25) makes it possible to translate the results on

asymptotically minimax variable selection in the exact and almost full regimes obtained for the normal mixture model (30) (see, for example, Theorems 2 and 4 of Cui (2014) and their proofs) to the case of the chi-square observations as in (32).

In the region $\mathbf{D}_1$, the parameter r is relatively large and the classification problem is relatively easy. This is largely due to the fact that in $\mathbf{D}_1$ we can select informative feature blocks in a nice (asymptotically exact) way. In this region, we can estimate $\omega_k = \mathbb{I}(\gamma_k \neq 0)$ with

$$\hat{\omega}_k = \mathbb{I}(T_k > t^*), \quad k = 1, \dots, b, \quad (33)$$

where $T_1, \dots, T_b$ are given in (15) and

$$t^* = (2 + \epsilon) \ln b, \quad (34)$$

for some $\epsilon = \epsilon_b$ satisfying

$$\epsilon \rightarrow 0 \quad \text{and} \quad \epsilon \ln b \rightarrow \infty. \quad (35)$$

The idea behind choosing the threshold t^* as in (34) is due to Lemma 2 and Theorem 1 obtained for a normal mixture model (22) which, in the context of variable selection, is asymptotically equivalent to model (26). We then can use relations (23) and (25) to translate this selection method to a chi-square mixture model (32). Denote the respective classifier in (7) by $\hat{\psi}$, that is,

$$\hat{\psi} = \mathbb{I} \left\{ \sum_{k=1: \hat{\omega}_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\},$$

where $\hat{\omega}$ is given by (33). The classifier $\hat{\psi}$ is a function of the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$. Note that, at this point, all $\boldsymbol{\Sigma}_{[k]}$ s are

assumed known. Here is an example of the region $\mathbf{D}_1$, where in order to calculate the threshold t^* in (34) we take $\epsilon = 1/\ln \ln b$.

Figure 14: Region $\mathbf{D}_1$ where exact selection is possible with threshold t^* .

The value t^* is calculated using relation (34). We can see that there is a distinct gap between the central and non-central chi-square distributions, thus making it easy to differentiate the statistics T_k belonging to the non-central distribution with noncentrality parameter $\gamma_k = 2r \ln b$ from the statistics T_k belonging to the central distribution with $\gamma_k = 0$. We have the following analog of Theorem 1 in Section 8 for the chi-square mixture model (32).

Proposition 2. *The selector $\hat{\omega}$ given by (33) is an exact selector in the sense that, uniformly in $(\beta, r) \in \{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1, \rho_2(\beta) < r < 4\}$,*

$$\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E}|\hat{\omega} - \omega| = \sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E} \sum_{k=1}^b |\hat{\omega}_k - \omega_k| = o(1), \quad b \rightarrow \infty, \quad (36)$$

where $\omega = (\omega_k)_{k=1,\dots,b} = (\mathbb{I}(\gamma_k \neq 0))_{k=1,\dots,b}$ and the set $\mathcal{H}_{b,\beta}$ is as in (31).

Remark 3. Relation (36) continues to hold for a more general model, from which the chi-square mixture model (32) follows, with b independent statistics $t_1, \dots, t_b$ such that

$[b^{1-\beta}]$ statistics among $t_1, \dots, t_b$ have a non-central chi-square $\chi_\nu^2(\gamma_b)$ distribution and the other $b - [b^{1-\beta}]$ statistics have a central $\chi_\nu^2(0)$ distribution. The proof of Proposition 2 in Section 16 is presented for this more general model.

This means that the maximum Hamming risk of $\hat{\omega}$ is algebraically small in the region $\mathbf{D}_1$. This is the reason for setting t^* as in (34) in this region. The proof of Proposition 2 can be found in Section 16.

Now, if we transition into the region $\mathbf{D}_2$, using the same arguments for thresholding as in Theorem 3 of Section 8, we would set our threshold t^* as

$$t^* = (2\beta + \epsilon) \ln b, \quad (37)$$

and estimate $\omega_k = \mathbb{I}(\gamma_k \neq 0)$ with

$$\tilde{\omega}_k = \tilde{\omega}_k(\beta) = \mathbb{I}(T_k > t^*), \quad k = 1, \dots, b. \quad (38)$$

We use this threshold due to the result of Theorem 3 in Section 8 for a normal mixture model (22) together with relations (23) and (25) to give a similar selection method for a chi-square mixture model (32). We can then denote the respective classifier in (7) by $\tilde{\psi}$, that is,

$$\tilde{\psi} = \mathbb{I} \left\{ \sum_{k=1: \tilde{\omega}_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}.$$

The classifier $\tilde{\psi}$ depends on the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$.

Figure 15 gives an example of the region $\mathbf{D}_2$, where for calculating the threshold t^* in (37) we take $\epsilon = 1/\ln \ln b$.

Figure 15: Region $\mathbf{D}_2$ where almost full selection is possible with threshold t^* .

Unfortunately, β is typically unknown which creates a problem. What we need now is a method for setting up the threshold without knowledge of β . The following procedure gives an *adaptive* version of $\tilde{\omega}(\beta) = (\tilde{\omega}_k(\beta))_{k=1,\dots,b}$, in which the threshold level is set up by a data-driven method.

Initially, we pick an integer $M = M_b$ such that

$$M \rightarrow \infty \quad \text{and} \quad \ln b/M \rightarrow 0,$$

and consider an equidistant grid of points $(1 - \theta)/2 < \beta_1 < \dots < \beta_M < 1 - \theta$, where

$$\beta_m = m\Delta, \quad \Delta \sim \left(\frac{1 - \theta}{2M} \right), \quad m = 1, \dots, M. \quad (39)$$

In view of the above assumptions on M ,

$$\Delta \rightarrow 0, \quad \Delta \ln b \rightarrow 0, \quad b \rightarrow \infty,$$

yielding for all large enough b

$$b^\Delta \leq C, \quad (40)$$

where C is a positive constant.

Next, for all $k = 1, \dots, b$ and $m = 1, \dots, M$, set, cf. (38),

$$\widehat{\omega}_k(\beta_m) = \mathbb{I}(t_k > (2\beta_m + \epsilon) \ln b), \quad (41)$$

where $\epsilon = \epsilon_b > 0$ satisfies (35), and define an adaptive selector $\widehat{\omega}(\beta_{\widehat{m}})$ as

$$\widehat{\omega}(\beta_{\widehat{m}}) = (\widehat{\omega}_1(\beta_{\widehat{m}}), \dots, \widehat{\omega}_b(\beta_{\widehat{m}})), \quad (42)$$

where $\widehat{m}$ is chosen by Lepski's method (see Section 2 of Lepski [15]) as follows:

$$\widehat{m} = \max \{1 \leq m \leq M : |\widehat{\omega}(\beta_m) - \widehat{\omega}(\beta_j)| \leq v_j \text{ for all } j \leq m\}, \quad (43)$$

and set $\widehat{m} = 1$ if the set in (43) is empty. Here the quantities $v_j = v_{j,b}$ are set to be

$$v_j = b^{1-\beta_j} / \tau_b, \quad 1 \leq j \leq m,$$

with a sequence of numbers $\tau_b \rightarrow \infty$ satisfying

$$\tau_b = o(\min(\ln b, b^{\epsilon/2})), \quad \text{as } b \rightarrow \infty,$$

where ϵ satisfies (35).

The idea behind Lepski's procedure is that each estimator $\widehat{\omega}(\beta_j)$ is associated with a confidence interval of a size that is proportional to its standard deviation. When, in the process of iteration, for some index m an estimator $\widehat{\omega}(\beta_m)$ is found to be far enough from

the previous ones, this means that we have detected a bias change, indicating that we are likely close to the true value of β . Algorithmically, Lepski's procedure for choosing $\hat{m}$ works as follows. We start by setting $\hat{m} = 1$ and attempt to increase the value of $\hat{m}$ from 1 to 2. If $|\hat{\omega}(\beta_2) - \hat{\omega}(\beta_1)| \leq v_1$, we set $\hat{m} = 2$; otherwise, we keep $\hat{m}$ equal to 1. In case $\hat{m}$ is increased to 2, we continue the process attempting to increase it further. If $|\hat{\omega}(\beta_3) - \hat{\omega}(\beta_2)| \leq v_2$ and $|\hat{\omega}(\beta_2) - \hat{\omega}(\beta_1)| \leq v_1$, we set $\hat{m} = 3$; otherwise, we keep $\hat{m}$ equal to 2; and so on. Notice that by construction

$$v_1 \geq v_2 \geq \dots \geq v_M.$$

The selector of this type has been previously used in somewhat different settings in Butucea et al. [4]. It can be shown that $\hat{\omega}(\beta_{\hat{m}})$ given by (45) is an *almost exact selector*. Namely, we have the following result for model (32).

Proposition 3. *The selector $\hat{\omega}(\beta_{\hat{m}})$ given by (42) is an almost full selector in the sense that, uniformly in $(\beta, r) \in \{(\beta, r) \in \mathbb{R}^2 : (1 - \theta)/2 < \beta < 1 - \theta, \rho_1(\beta) < r < \rho_2(\beta)\}$,*

$$\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E} |\hat{\omega}(\beta_{\hat{m}}) - \omega| = \sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E} \sum_{k=1}^b |\hat{\omega}_k(\beta_{\hat{m}}) - \omega_k| = o(b^{1-\beta}), \quad b \rightarrow \infty, \quad (44)$$

where $\omega = (\omega_k)_{k=1,\dots,b} = (\mathbb{I}(\gamma_k \neq 0))_{k=1,\dots,b}$ and the set $\mathcal{H}_{b,\beta}$ is as in (31).

Remark 4. Relation (44) continues to hold for a more general model, from which the chi-square mixture model (32) follows, with b independent statistics $t_1, \dots, t_b$ such that $[b^{1-\beta}]$ statistics among $t_1, \dots, t_b$ have a non-central chi-square $\chi_\nu^2(\gamma_b)$ distribution and the other $b - [b^{1-\beta}]$ statistics have a central $\chi_\nu^2(0)$ distribution. The proof of Proposition 3 in Section 16 is given for this more general model.

By Proposition 3, the maximum Hamming risk of $\hat{\omega}(\beta_{\hat{m}})$ is small relative to the number of non-centrally distributed statistics among $T_1, \dots, T_b$. The proof of Proposition 3 can be seen in Section 16. In this proof we can see that if $m_0 \in \{1, \dots, M - 1\}$ is such

that the true $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$, then $\hat{m} \geq m_0$ with high probability.

Now, using Lepski's method, we set our threshold t^* at level, cf. (37),

$$t^* = (2\beta_{\hat{m}} + \epsilon) \ln b, \quad (45)$$

and estimate ω_k with

$$\check{\omega}_k = \mathbb{I}(T_k > t^*), \quad k = 1, \dots, b,$$

where $T_1, \dots, T_b$ are as in (15). We then denote the respective classifier in (7) by $\check{\psi}$, that is,

$$\check{\psi} = \mathbb{I} \left\{ \sum_{k=1: \check{\omega}_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\},$$

where $\check{\psi}$ is a function of the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$.

In Figure 16, we can see an example of our threshold t^* calculated by (45) as compared to the oracle threshold (37).

Figure 16: Region $\mathbf{D}_2$ where almost full selection is possible with threshold t^* as in (45).

In Figure 16 the solid line is the threshold t^* calculated by (37) with true β . The dotted line is the threshold calculated by (45). What we can see is that the estimated threshold (45) over-estimates our original t^* from (37) but it does incorporate almost none of the non-informative features (central chi-square observations T_k with $\gamma_k = 0$) while incorporating most of the informative features (non-central chi-square observations T_k with $\gamma_k \neq 0$).

10 Classifiers with selected informative feature blocks

In this section, using Propositions 2 and 3, we show that in the regions $\mathbf{D}_1$ and $\mathbf{D}_2$ the classifiers in which the informative blocks have been carefully selected provide successful classification.

To this end, consider the statistics $T_1, \dots, T_b$ as defined in (15). The majority of the statistics T_k have a central chi-square $\chi_{p_0}^2(0)$ distribution, whereas a small fraction $\varepsilon_b = b^{-\beta}$ of the T_k s have a non-central $\chi_{p_0}^2(\gamma_k)$ distribution with the same noncentrality parameter $\gamma_k = 2r \ln b$ for some $\beta < r < 4$, $k = 1, \dots, b$. In particular, $T_1, \dots, T_b$ obey the chi-square mixture model (21).

Consider two auxiliary classification rules, one for the region $\mathbf{D}_1$ and the other for the region $\mathbf{D}_2$, with *true* $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ and *estimated* $\omega_k = \mathbb{I}(\gamma_k \neq 0)$, $k = 1, \dots, b$. Specifically, let $\tilde{\omega}_1 = (\tilde{\omega}_{1,k})_{k=1, \dots, b}$ and $\tilde{\omega}_2 = (\tilde{\omega}_{2,k})_{k=1, \dots, b}$ be estimators of a vector $\omega = (\omega_k)_{k=1, \dots, b} \in \mathcal{H}_{b, \beta}$ when $(\beta, r) \in \mathbf{D}_1$ and $(\beta, r) \in \mathbf{D}_2$, respectively, defined in accordance with formulas (33), (34) and (45) as follows:

$$\begin{aligned} \tilde{\omega}_{1,k} &= \mathbb{I}(T_k > (2 + \epsilon) \ln b), & \text{in the region } \mathbf{D}_1, \\ \tilde{\omega}_{2,k} &= \mathbb{I}(T_k > (2\beta_{\tilde{m}} + \epsilon) \ln b), & \text{in the region } \mathbf{D}_2. \end{aligned} \tag{46}$$

Then for $l = 1, 2$ the proposed classifier for the region $\mathbf{D}_l$, which depends on the vector

$\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, is

$$\tilde{\Psi}_l = \mathbb{I} \left\{ \sum_{k=1: \tilde{\omega}_{l,k}=1}^b (\mathbf{X}_{0,[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})/2)^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\boldsymbol{\mu}_{2,[k]} - \boldsymbol{\mu}_{1,[k]}) \leq 0 \right\}.$$

The next result shows that in the region $\mathbf{D}_l$, $l = 1, 2$, the classifier $\tilde{\Psi}_l$, $l = 1, 2$, does its job properly. Again, we consider the worst case scenario and assume that all nonzero noncentrality parameters γ_k are the same and are equal to $2r \ln b$ for some $\beta < r < 4$.

Proposition 4. *For the sparse block-diagonal normal model under consideration, assuming that Condition A holds true, for $l = 1, 2$, uniformly in $(\beta, r) \in \mathbf{D}_l$,*

$$\lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\tilde{\Psi}_l) = 0.$$

The proof of Proposition 4 is presented in Section 16. It is seen from the proof that the major factor in getting the classifiers $\tilde{\Psi}_1$ and $\tilde{\Psi}_2$ that provide successful classification in the regions $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively, is a careful selection of the informative feature blocks as detailed in Propositions 2 and 3 of Section 9.

11 Obtaining threshold t^* in the case of $r < \beta$

It now remains to study the possibility of successful classification in the region $\mathbf{D}_3$. The classification problem in this region is much harder than that in the regions $\mathbf{D}_1$ and $\mathbf{D}_2$. This is because in the region $\mathbf{D}_3$ there can neither be exact selection nor almost full selection, as a result of Theorem 4 and relations (23) and (25). Therefore, based on heuristic arguments, we will only be able to propose a classifier that is believed to work well. Unfortunately, unlike Donoho and Jin [8] who used a similar method and claimed that it provides successful classification, we cannot assert that the classifier proposed in this section, which numerically works slightly better than the procedure of Donoho

and Jin [8], gives successful classification in the region $\mathbf{D}_3$. This is due to the lack of rigorous analytical arguments and also due to the numerical results of Section 14.

Recall that, if Σ is known, the proposed classifier, which is based on the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, has the form

$$\Psi^* = \mathbb{I} \left\{ \sum_{k=1: \omega_k^*=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \Sigma_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}. \quad (47)$$

where $\omega_k^* = \mathbb{I}(T_k > t^*)$ is an estimator of $\omega_k = \mathbb{I}(\gamma_k \neq 0)$. The problem is to find a good threshold t^* that would result in a good estimator ω_k^* of ω_k , $k = 1, \dots, b$.

Following the suggestion of Donoho and Jin [8], in order to set up a threshold for classification purposes, we may consider the problem of testing sparse heterogeneous mixtures. Specifically, based on the fact that the statistics $T_1, \dots, T_b$ introduced in (15) follow the chi-square mixture model, in order to construct a classifier for the region $\mathbf{D}_3$, we consider the problem of testing the hypotheses, cf. (28),

$$H_0 : t_1, \dots, t_b \stackrel{iid}{\sim} \chi_{p_0}^2(0) \quad \text{vs.} \quad H_1 : t_1, \dots, t_b \stackrel{iid}{\sim} (1 - \varepsilon) \chi_{p_0}^2(0) + \varepsilon \chi_{p_0}^2(\gamma), \quad (48)$$

where $\gamma = 2r \ln b$ for some $0 < r < 1$ and $\varepsilon = b^{-\beta}$ for some $0 < \beta < 1$. Recall that in our classification model we have β and r as unknown parameters, θ and p_0 as known parameters, and $b \rightarrow \infty$ as a driving parameter. Now, we denote the cumulative distribution function (CDF) of a chi-square $\chi_\nu^2(\gamma)$ random variable by

$$F_\nu(x; \gamma) = \mathbf{P}(\chi_\nu^2(\gamma) \leq x), \quad x \in \mathbb{R},$$

and transform the statistics t_k , $k = 1, \dots, b$, into the statistics

$$s_k = 1 - F_{p_0}(t_k; 0), \quad k = 1, \dots, b, \quad (49)$$

that take values in $(0, 1)$.

In terms of a common CDF $F(u)$ of the S_k s, the problem of testing H_0 versus H_1 is equivalent to that of testing

$$\mathcal{H}_0 : F(u) = F_0(u), \quad \text{the uniform } U(0, 1) \text{ CDF}$$

versus a sequence of *upper-tailed* alternatives

$$\mathcal{H}_1 : F(u) = F_0(u) + \varepsilon_b \left((1 - u) - F_\nu(F_\nu^{-1}(1 - u; 0); \gamma) \right) > F_0(u).$$

For specific types of alternatives, weighted versions of the one-sided *Kolmogorov-Smirnov* statistic provide a greater power than the original unweighted version, if the weight function is chosen properly. For testing sparse heterogeneous mixtures, such as the one in (48), the weighted Kolmogorov-Smirnov statistics with certain weight functions $q(\cdot)$ on $(0, 1)$ were found to work very well (see, for example, Danoho and Jin [7] and Stepanova and Pavlenko [19]). Donoho and Jin [8] suggested also to use the *weighted Kolmogorov-Smirnov testing* approach for feature selection.

Using their idea, fix a small number $\alpha_0 \in (0, 1/2)$, put

$$S_k = 1 - F_{p_0}(T_k; 0), \quad k = 1, \dots, b,$$

and consider the statistic

$$K_{b,q}^* = \sup_{0 < u < \alpha_0} K_{b,q}(u) = \sup_{0 < u < \alpha_0} \frac{\sqrt{b}(\mathbb{H}_b(u) - u)}{q(\mathbb{H}_b(u))},$$

where

$$K_{b,q}(u) = \frac{\sqrt{b}(\mathbb{H}_b(u) - u)}{q(\mathbb{H}_b(u))}$$

is the so-called *objective function*, with $q(u)$ being a weight function on $(0, 1)$ that can

take multiple forms, and $\mathbb{H}_b(u) = b^{-1} \sum_{i=1}^b \mathbb{I}(S_i \leq u)$ being the empirical distribution function (EDF) based on $S_1, \dots, S_b$.

If we arrange the observations $S_1, \dots, S_b$ in an increasing order to get $S_{(1)} < S_{(2)} < \dots < S_{(b)}$, then for a weight function q that is monotone on $(0, \alpha_0)$ for small α_0 , the statistic $K_{b,q}^*$ can be (asymptotically) equivalently written as

$$K_{b,q}^* = \max_{1 \leq k \leq [\alpha_0 b]} \frac{\sqrt{b}(k/b - S_{(k)})}{q(k/b)}. \quad (50)$$

Now, as suggested by Donoho and Jin [8], we find the observation that maximizes the objective function in (50) which we will use to get our threshold t^* . Denote by k^* the index

$$k^* = \operatorname{argmax}_{1 \leq k \leq [\alpha_0 b]} \frac{\sqrt{b}(k/b - S_{(k)})}{q(k/b)}, \quad k = 1, \dots, b, \quad (51)$$

where $\alpha_0 \in (0, 1/2)$ defines a range of significant levels used for testing. Typically, one takes $\alpha_0 \leq 0.2$. Essentially, we are looking for the largest standardized discrepancy between theory and practice (see page 4452 of [8]). Converting back the distribution on $(0, 1)$ to the chi-square distribution, we define the threshold t^* as

$$t^* = F_{p_0}^{-1}(1 - S_{(k^*)}; 0), \quad k = 1, \dots, b, \quad (52)$$

where $F_\nu^{-1}(x; \gamma)$ is the inverse of the CDF $F_\nu(x; \gamma)$.

In the paper of Donoho and Jin [8], they suggested using the weight function $q(u) = \sqrt{u(1-u)}$, $0 < u < 1$, which is not quite appropriate from the probabilistic point of view (for details, see Csörgő et al. [5] and Stepanova and Pavlenko [19]). When we are dealing with the chi-square mixture model, the weighted Kolmogorov-Smirnov statistics under the alternative H_1 will tend to be larger than under H_0 and greater than zero. This is the case because with high probability each $S_{(k)} < k/b$ when the non-central

observations are transformed to our S_k observations as in (49).

Here are some illustrations of how the objective function $K_{b,q}(S_{(k)}) = \frac{\sqrt{b}(k/b - S_{(k)})}{q(k/b)}$ behaves under the null hypothesis and under the alternative.

Figure 17: Behaviour of the function $K_{b,q}(S_{(k)})$ under $\mathcal{H}_0$ on $(0,1)$.

As we can see in Figure 17, the objective function $K_{b,q}(S_{(k)})$ hovers around zero. This is as we expect since, under $\mathcal{H}_0$, $S_1, \dots, S_b \stackrel{iid}{\sim} U(0,1)$ giving that

$$\mathbf{P}_{\mathcal{H}_0}(S_{(k)} < k/b) \approx 1/2.$$

Figure 18: Behaviour of the function $K_{b,q}(S_{(k)})$ under $\mathcal{H}_1$ on $(0,1)$.

We notice that Figure 18 is strictly positive near zero since we are not dealing with uniformly distributed random variables any more when using the transformation (49) for a chi-square mixture model. That is why we can use the statistic $K_{b,q}^*$ which tends to be larger under H_1 when dealing with the testing problem (48).

Notice that our observations $T_1, \dots, T_b$ from (15) are iid chi-square random variables as given by the alternative H_1 . The issue with setting the threshold using weight $q(u) = \sqrt{u(1-u)}$ is that, no matter which α_0 is chosen, our threshold t^* is usually going to be the maximum of our observations since $K_{b,q}(S_{(k)}) \rightarrow \infty$ as $k/b \rightarrow 0$ with probability one (see Chapter 16 of Shorack and Wellner [17]). This happens because $\sqrt{(k/b)(1-k/b)}$ goes to zero too fast as $k/b \rightarrow 0$. This means that using this method will most likely give only one feature block for classification which poses a problem. In the region $\mathbf{D}_3$, we are dealing with a relatively small non-centrality parameter $\gamma_b = 2r \ln b$, so even if the feature block that was selected is informative our classifier has a high chance of misclassification.

From this point forward, we will denote the weight function above as

$$q_1(u) = \sqrt{u(1-u)}, \quad 0 < u < 1.$$

Here are some other weight functions on $(0,1)$ that we will investigate in terms of their classification accuracy in the region $\mathbf{D}_3$:

$$\begin{aligned} q_2(u) &= (u(1-u))^{1/4}, \\ q_3(u) &= \sqrt{u(1-u)} \ln \ln(1/(u(1-u))), \\ q_4(u) &= \sqrt{u(1-u) \ln \ln(1/(u(1-u)))}. \end{aligned}$$

The function q_1 is a *standard deviation proportional (SDP)* weight function, the functions q_2 and q_3 are the so-called *Chibisov-O'Reilly* weight functions, and the function q_4 is an *Erdős-Feller-Kolmogorov-Petrovski upper-class function of a Brownian bridge*. These

weight functions are well-known in the theory of weighted empirical processes (see Csörgő et al. [5] and references therein for more details). Each of these weight functions, $q_i(u)$ where $i = 2, 3, 4$, decreases to zero near zero and 1, but not as fast as the SDP weight function $q_1(u)$ does. This means that we are less likely to get the maximum of our observations $T_1, \dots, T_b$ as our threshold t^* . Near the origin, these weight functions are ordered in the following manner:

$$q_1(u) < q_4(u) < q_3(u) < q_2(u), \quad \text{as } u \rightarrow 0.$$

This implies that for small $S_{(k)}$

$$K_{b,q_1}(S_{(k)}) > K_{b,q_4}(S_{(k)}) > K_{b,q_3}(S_{(k)}) > K_{b,q_2}(S_{(k)}).$$

The idea of using q_2 , q_3 , or q_4 instead of q_1 is to have test statistics similar to K_{b,q_1}^* that would be finite with probability one under the null hypothesis. With the weight function q_1 this is not the case since, even under the null hypothesis, K_{b,q_1}^* blows up as $b \rightarrow \infty$ (see Chapter 16 of Shorack and Wellner and [17] Stepanova and Pavlenko [19] for more details).

Figures 19 and 20 give an example of the four objective functions $K_{b,q_i}(S_{(k)})$, $i = 1, \dots, 4$, next to each other with the vertical lines representing the values of the statistics K_{b,q_i}^* , $i = 1, \dots, 4$. The thresholds obtained through the transformation seen in relation (52) are shown in Figure 21.

Figure 19: Behaviour of the objective functions $K_{b,q_i}(S_{(k)})$, $i = 1, \dots, 4$, under H_1 on the interval $(0,1)$.

Figure 20: Behaviour of the objective functions $K_{b,q_i}(u)$, $i = 1, \dots, 4$, under H_1 on the interval $(0,0.2)$.

Figure 21: The thresholds $t^* = F_{p_0}^{-1}(1 - S_{(k^*)}; 0)$ for four different weight functions for a chi-square mixture model specified by H_1 in the region $\mathbf{D}_3$.

Thus, our suggestion for the region $\mathbf{D}_3$ is to use the classifier Ψ^* as in (47) with the threshold t^* defined by (51)–(52) with one of the weight functions q_i , $i = 2, 3, 4$. Notice that the smaller the threshold t^* , the more feature blocks are used to estimate the classifier, both informative and non-informative. In the Section 14, we will see which of these thresholds returns the smallest chance of misclassification, which we will refer to as *classification error*.

12 Classifiers with partially estimated parameters and selected informative feature blocks

Back to the classification problem in the regions $\mathbf{D}_1$ and $\mathbf{D}_2$, we now consider almost data-driven classifiers, with Σ being the only known parameter. Namely, based on Proposition 1 of Section 7 and Proposition 4 of Section 10, we propose the following classification rule for the region $\mathbf{D}_l$, $l = 1, 2$:

$$\Psi_l = \mathbb{I} \left\{ \sum_{k=1: \tilde{\omega}_{l,k}=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \Sigma_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}, \quad (53)$$

where $\tilde{\omega}_l = (\tilde{\omega}_{l,k})_{k=1,\dots,b}$ is an estimator of $\omega = (\omega_k)_{k=1,\dots,b}$ with the components defined as in (46), and the only parameter that is assumed known is $\Sigma = \text{BlkDiag}(\Sigma_{[1]}, \dots, \Sigma_{[b]})$. Note that for $l = 1, 2$, Ψ_l is a function of the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$.

The classification procedure Ψ_l , $l = 1, 2$, generalizes (in different ways) the classifiers in Propositions 1 and 4. Indeed, Proposition 1 relates to the classifier $\hat{\Psi}$ with *estimated* mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the *true* vector $\omega = (\omega_k)_{k=1,\dots,b}$, whereas Proposition 4 relates to the classifier $\tilde{\Psi}_l$, $l = 1, 2$, with *estimated* vector $\omega = (\omega_k)_{k=1,\dots,b}$ and the *true* mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Therefore, the combination of the arguments that were used in the proofs Propositions 1 and 4 leads us to the following result.

Theorem 5. *For the sparse block-diagonal normal model under consideration, assuming that Condition A holds true, for $l = 1, 2$, uniformly in $(\beta, r) \in \mathbf{D}_l$,*

$$\lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) \in \mathcal{S}_{b,\beta,r}} \mathcal{R}(\Psi_l) = 0.$$

Theorem 5 implies that for $l = 1, 2$, uniformly in $(\beta, r) \in \mathbf{D}_l$,

$$\liminf_{b \rightarrow \infty} \inf_{\Psi} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\Psi) \leq \lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\Psi_l) = 0.$$

In other words, for all values of $(\beta, r) \in \mathbf{D}_l$, $l = 1, 2$, successful classification is possible and is provided by the classifier Ψ_l as in (53). The proof of Theorem 5 is omitted because it is essentially a combination of the arguments from the proofs of Propositions 1 and 4 that are presented in great detail in Section 16.

13 Completely data-driven classifiers

In the section, based on Theorem 5 and the fact that the weak limit of a random variable $\nu_1 F_{\nu_1, \nu_2}(\gamma)$ for fixed ν_1 and $\nu_2 \rightarrow \infty$ is $\chi_{\nu_1}^2(\gamma)$, we propose completely *adaptive* or *data-driven* classifiers for the regions $\mathbf{D}_1$ and $\mathbf{D}_2$. Namely, for $l = 1, 2$ we put, cf. (53),

$$\Psi_{l, \text{ad}} = \mathbb{I} \left\{ \sum_{k=1: \hat{\omega}_{l,k}=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \hat{\boldsymbol{\Sigma}}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}, \quad (54)$$

where for $k = 1, \dots, b$ the estimators $\hat{\boldsymbol{\mu}}_{1,[k]}^\top$, $\hat{\boldsymbol{\mu}}_{2,[k]}^\top$, and $\hat{\boldsymbol{\Sigma}}_{[k]}$ are given by formulas (8) and (12) and, cf. (46),

$$\begin{aligned} \hat{\omega}_{1,k} &= \mathbb{I}(p_0 \hat{T}_k > (2 + \epsilon) \ln b), & \text{in the region } \mathbf{D}_1, \\ \hat{\omega}_{2,k} &= \mathbb{I}(p_0 \hat{T}_k > (2\beta_{\hat{m}} + \epsilon) \ln b), & \text{in the region } \mathbf{D}_2. \end{aligned} \quad (55)$$

In formula (55), the statistics $\hat{T}_k$, $k = 1, \dots, b$, are defined in (13), the quantity β_m is defined in (39) with $\hat{m}$ as in (43), and $\epsilon = \epsilon_b > 0$ is such that $\epsilon \rightarrow 0$ and $\epsilon \ln b \rightarrow \infty$ as $b \rightarrow \infty$. The classifier in (54) is a function of the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$.

Compared to the classifier Ψ_l , $l = 1, 2$, in (53), in order to include useful and exclude useless feature blocks to/from consideration, the classifier $\Psi_{l,\text{ad}}$, $l = 1, 2$, uses the F -distributed statistics $\hat{T}_1, \dots, \hat{T}_b$ as in (13) rather than the chi-distributed statistics $T_1, \dots, T_b$ as in (15). Unlike the statistics $T_1, \dots, T_b$, whose application assumes the knowledge of the true covariance matrix $\Sigma = \text{BlkDiag}(\Sigma_{[1]}, \dots, \Sigma_{[b]})$, the statistics $\hat{T}_1, \dots, \hat{T}_b$ are entirely data-based. Hence, the classification procedures $\Psi_{1,\text{ad}}$ and $\Psi_{2,\text{ad}}$ are completely data-driven.

The idea behind formula (54) is simple. First, in case of normal observations, $\hat{\mu}_{1,[k]}$, $\hat{\mu}_{2,[k]}$, and $\hat{\Sigma}_{[k]}$ are the most natural estimators of the parameters $\mu_{1,[k]}$, $\mu_{2,[k]}$, and $\Sigma_{[k]}$, respectively. Second, for large b , the distribution of $p_0 \hat{T}_k$, is ‘close’ to that of T_k , $k = 1, \dots, b$. Therefore, for $l = 1, 2$, the difference (in terms of the classification error) between the classifiers with $\hat{\omega}_{l,k}$ as in (55) and similar classifiers with $\tilde{\omega}_{l,k}$, as in (46) is small. The closeness of the CDF of $p_0 \hat{T}_k$ to that of T_k , $k = 1, \dots, b$, is detailed by formula (6.8) in Siotani [18]. This formula gives a representation of the CDF of an $\nu_1 F_{\nu_1, \nu_2}(\gamma)$ random variable with noncentrality parameter $\gamma > 0$ when ν_1 is fixed and $\nu_2 \rightarrow \infty$ in terms of the CDF of a $\chi_{\nu_1}^2(\gamma)$ random variable, with an explicitly written remainder term.

Preliminary calculations with the use of formula (6.8) in Siotani [18] have led us to a preliminary conclusion that, at least for $1/2 < \theta < 1$, the classification error of $\Psi_{l,\text{ad}}$, $l = 1, 2$, tends to zero as $b \rightarrow \infty$. This result is now stated as a conjecture, cf. Theorem 5.

Conjecture 1. *For the sparse block-diagonal normal model under consideration with $1/2 < \theta < 1$, assuming that Condition A holds true, for $l = 1, 2$, uniformly in $(\beta, r) \in \mathbf{D}_l$,*

$$\lim_{b \rightarrow \infty} \sup_{(\mu_1, \mu_2, \Sigma) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\Psi_{l,\text{ad}}) = 0.$$

This conjecture is strongly supported by the numerical results of Section 14.

As for the region $\mathbf{D}_3$, we propose to use the following completely data-driven classifier based on the vector $\mathbf{X}_0$ and the sub-samples $\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n/2}^{(1)}$ and $\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n/2}^{(2)}$, cf. (47):

$$\Psi^{**} = \mathbb{I} \left\{ \sum_{k=1: \omega_k^{**}=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \hat{\boldsymbol{\Sigma}}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}) \leq 0 \right\}.$$

where

$$\omega_k^{**} = \mathbb{I}(p_0 \hat{T}_k > t^*), \quad k = 1, \dots, b,$$

with the statistics $\hat{T}_1, \dots, \hat{T}_b$ defined by (13) and the threshold t^* set up in accordance with formulas (51)–(52), using $q(u) = q_i(u)$, $i = 2, 3, 4$. In the next section, we will see that, numerically, the smallest classification error of Ψ^{**} is obtained when we take either $q(u) = (u(1-u))^{1/4}$ or $q(u) \equiv 1$. The latter choice of $q(u)$ corresponds to thresholding by using the original one-sided Kolmogorov-Smirnov statistic.

14 Numerical analysis

In this section, we will be looking at the numerical results obtained for the feature selection and classification procedures in the scenarios mentioned earlier through $N = 100$ independent cycles of simulations. We will be assessing the results for $\boldsymbol{\Sigma}$ both known and unknown. The following simulations are done with dimensions $b = 1000$ and $p_0 = 3$ implying $p = 3000$ due to limitations regarding the generation of high-dimensional multivariate normal random variables.

In this section, we use the following notation:

$E_{j,b}$ – estimated classification error of the j th simulation

$B_{j,I}$ – number of informative blocks used in the j th simulation

$B_{j,N}$ – number of non-informative blocks used in the j th simulation

t_j^* – the threshold of the j th simulation

where $j = 1, \dots, 100$. Also, we put

$$\bar{E}_b = \frac{1}{N} \sum_{j=1}^N E_{j,b}, \quad \bar{t}^* = \frac{1}{N} \sum_{j=1}^N t_j^*, \quad \bar{B}_I = \frac{1}{N} \sum_{j=1}^N B_{j,I}, \quad \bar{B}_N = \frac{1}{N} \sum_{j=1}^N B_{j,N}.$$

The classification error $\bar{E}_b$ estimates the Bayes risk $\mathcal{R}_B(\psi) = \frac{1}{2}\mathbf{E}_{\Pi_2}(\psi) + \frac{1}{2}\mathbf{E}_{\Pi_1}(1 - \psi)$, where ψ is one of the classifiers proposed in Sections 12 and 13, depending on which of the regions $\mathbf{D}_1$, $\mathbf{D}_2$ or $\mathbf{D}_3$ we are dealing with.

14.1 Simulations in the region $\mathbf{D}_1$

This simulation cycle was completed using the following parameters: $N = 100$, $b = 1000$, $p_0 = 3$, $\theta = 0.5$, $\beta = 0.375$ and $r = 3.6$. This means that the size of the training labelled sample is $n = 32$ and $\varepsilon_b = 0.075$, that is, only 75 of the 1000 feature blocks are informative. The results are presented in Table 1.

Table 1: Numerical summary for the region $\mathbf{D}_1$

Scenario	$\bar{t}^*$	$\bar{B}_I$	$\bar{B}_N$	$\bar{E}_b$
Σ known	27.1657	73.34	0.01	0.000
Σ unknown	9.0552	71.63	0.19	0.000

Interestingly, we see that there it does not matter whether Σ is known or unknown since successful classification is achieved in both cases. It also seems as though exact

selection is not achieved in every iteration but it is rare that a non-informative feature block is included in the classification process.

14.2 Simulations in the region $\mathbf{D}_2$

This simulation cycle was completed using the following parameters: $N = 100$, $b = 1000$, $p_0 = 3$, $\theta = 0.5$, $\beta = 0.375$ and $r = 1.8$. This means that the size of the training labelled sample is $n = 32$ and $\varepsilon_b = 0.075$, that is, only 75 of the 1000 feature blocks are informative. The obtained results are summarized in Table 2.

Table 2: Numerical summary for the region $\mathbf{D}_2$

Scenario	$\bar{t}^*$	$\bar{B}_I$	$\bar{B}_N$	$\bar{E}_b$
Σ known	10.1798	73.41	16.31	0.000
Σ unknown	3.3875	72.66	30.25	0.000

Notice that successful classification can be achieved in the region $\mathbf{D}_2$ by thresholding based on Lepski's method discussed in Section 9. Since almost full selection is possible in this region we consider mostly informative feature blocks which is why our classification error is zero, up to three decimals.

14.3 Simulations in the region $\mathbf{D}_3$

This simulation cycle was completed using the following parameters: $N = 100$, $b = 1000$, $p_0 = 3$, $\theta = 0.5$, $\beta = 0.375$, $r = 0.25$, and $\alpha_0 = 0.2$. This means that the size of the training labelled sample is $n = 32$ and $\varepsilon_b = 0.075$, that is, only 75 of the 1000 feature blocks are informative. The numerical results are collected in Tables 3 and 4.

Table 3: Numerical summary for the region $\mathbf{D}_3$ when Σ is known

Weight	$\bar{t}^*$	$\bar{B}_I$	$\bar{B}_N$	$\bar{E}_b$
q_1	15.8506	10.90	21.42	0.2950
q_2	6.4653	34.20	113.13	0.2046
q_3	9.3644	31.90	120.00	0.2340
q_4	12.4259	21.98	70.40	0.2582

Note that, when Σ is known, looking at all blocks, informative and non-informative, we get the classification error of $\bar{E}_{b,\text{all}} = 0.2145$. Also, the classification error using the original Kolmogorov-Smirnov statistic (with $q \equiv 1$ and $\alpha_0 = 1$) is $\bar{E}_{b,\text{KS}} = 0.2024$.

Table 4: Numerical summary for the region $\mathbf{D}_3$ when Σ is unknown

Weight	$\bar{t}^*$	B_I	B_N	E_b
q_1	6.7090	12.26	32.34	0.3255
q_2	2.5025	32.11	113.58	0.2427
q_3	3.4823	32.55	127.36	0.2564
q_4	5.2440	21.73	75.93	0.2912

Note that, when Σ is unknown, looking at all blocks, informative and non-informative, we get $\bar{E}_{b,\text{all}} = 0.231365$. Also, the classification error using the original Kolmogorov-Smirnov statistic (with $q \equiv 1$ and $\alpha_0 = 1$) is $\bar{E}_{b,\text{KS}} = 0.22943$.

Commenting further on Tables 3 and 4, we notice that the (numerically) best threshold that separates informative blocks from non-informative for subsequent classification is obtained by using the original (unweighted) Kolmogorov-Smirnov statistic. This is quite an unexpected result. Also, we can see that, in the region $\mathbf{D}_3$, the attempt to exclude non-informative blocks from consideration did not, in fact, lead us to the classifier that would be significantly better than the one that uses all the blocks, informative and non-informative. Next, when Σ is unknown our classification error increases but by a small amount which means that these methods can be used in real world scenarios. Also, it turns out that the weight function suggested by Donoho and Jin [8] is the worst

in terms of classification in high-dimensional sparse models. Finally, the best (in terms of the classification error) weight function, out of the four functions considered in the thesis, is the Chibisov–O’Reilly function $q_2(u) = (u(1 - u))^{1/4}$. It gives the lowest error for both known and unknown Σ , which is just slightly larger than the error obtained for the original Kolmogorov-Smirnov statistic. Therefore, our recommendation for the region $\mathbf{D}_3$ is to set up the threshold t^* as in (52) by using the classical one-sided statistic of Kolmogorov and Smirnov. This approach, however, does not give much of the improvement over the classification procedure with all, informative and non-informative, feature blocks included.

15 Informal discussion of the classification problem in the region $\mathbf{D}_3$

In this section, we will attempt to justify the classification results from the region $\mathbf{D}_3$ while explaining what should be looked at more closely. As we can see, the methods discussed in this paper were trumped by the results from the original Kolmogorov-Smirnov test. It is believed that the reason for this lies in the distribution of our classification rule that involves the statistics, cf. (7),

$$U_k = (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}).$$

If $\mathbf{X}_0 \sim \Pi_1$ and the k th block is informative, then U_k is nearly normally distributed with a negative mean μ^- and some known variance σ^2 , that is, $U_k \sim N(\mu^-, \sigma^2)$. If $\mathbf{X}_0 \sim \Pi_2$ and the k th block is informative then U_k is nearly normally distributed with a positive mean μ^+ and the same variance σ^2 , that is, $U_k \sim N(\mu^+, \sigma^2)$. If the k th block is non-informative, then U_k is nearly normally distributed with zero mean regardless of whether $\mathbf{X}_0$ belongs to Π_1 or Π_2 (see Section 6.4 of Anderson [1]). The images on

Figures 22–24 back up what Anderson [1] has indicated.

Figure 22: Estimated density of U_k when the k th block is informative and $\mathbf{X}_0 \sim \Pi_1$.

Figure 23: Estimated density of U_k when the k th block is informative and $\mathbf{X}_0 \sim \Pi_2$.

Figure 24: Estimated density of U_k when the k th block is non-informative.

Suppose we look at the simulated scenario from the region $\mathbf{D}_3$ where $b = 1000$, $p_0 = 3$, $\theta = 0.5$, $\beta = 0.375$ and $r = 0.25$. This means $\varepsilon_b = 0.075$, which means that only 75 of the 1000 feature blocks are informative. If we knew which blocks were informative and we look at the distribution of

$$V_b = \sum_{k=1: \omega_k=1}^b (\mathbf{X}_{0,[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})/2)^\top \hat{\boldsymbol{\Sigma}}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}),$$

where $\omega_k = \mathbb{I}(\mathbf{X}_{[k]} \text{ is informative})$, we can see that it is also nearly normal, roughly because the sum of normal random variables is also a normal random variable itself. Figures 25–27 illustrate the distribution of V_b under the three scenarios (informative from Π_1 , informative from Π_2 , and non-informative).

Figure 25: Estimated density of V_b when looking at all 75 informative blocks when $\mathbf{X}_0 \sim \Pi_1$.

Figure 26: Estimated density of V_b when looking at all 75 informative blocks when $\mathbf{X}_0 \sim \Pi_2$.

Figure 27: Estimated density of V_b when every block is non-informative.

What we notice from Figures 25 and 26 is that, even if we knew which blocks were informative prior to classification, then we still have a chance of misclassification due to the large size of the variance. Since it is not realistic to know which blocks are informative in practical situations, Figures 25 and 26 describe a “best case scenario” for the classification done in $\mathbf{D}_3$. Note that it does not matter which distribution V_b belongs to if the blocks are non-informative.

Now, if we were to look at the distribution of V_b considering all blocks, 75 informative and 925 non-informative, we obtain the following results.

Figure 28: Estimated density of V_b when looking at all blocks when $\mathbf{X}_0 \sim \Pi_1$.

Figure 29: Estimated density of V_b when looking at all blocks when $\mathbf{X}_0 \sim \Pi_2$.

Figures 28–29 make it seem as though the distribution of V_b is a normal mixture model. The following is my attempt at explaining what is going on in very basic terms. Everytime an informative block is used for classification (regardless of whether $\mathbf{X}_0 \sim \Pi_1$ or $\mathbf{X}_0 \sim \Pi_2$), it pushes the mean further away from zero and increases the variance. Everytime a non-informative block is used, it increases the variance without having an impact on the mean.

Recall that in the region $\mathbf{D}_3$ neither exact selection nor almost full selection is possible due to the overlap of the central chi-square distribution and the non-central chi-square distribution, as seen from Figure 30.

Figure 30: Central and non-central chi-square distributions in the region $\mathbf{D}_3$ where neither exact nor almost full selection is possible.

It is believed, however, that there is an optimal threshold t^* that can be achieved where there are enough informative blocks used in our classification rule (7), so that the mean is far enough away from zero without using too many non-informative blocks which only increases the variance of the classification rule. This is why the Kolmogorov-Smirnov threshold held up the best in terms of classification. It seemed to capture enough informative blocks but excluded many non-informative blocks which only increase the variance of V_b and in turn increase the chance of misclassification. This is the area of interest that I believe needs more research.

16 Proofs

In this section, we collect the proofs of all our lemmas and propositions. First, we prove Lemmas 1 and 2 from Section 8.

Proof of Lemma 1. First, we note that

$$1 - \Phi(x) \sim \frac{\varphi(x)}{x}, \quad x \rightarrow \infty. \quad (56)$$

Due to (56) and the symmetry of the ξ_i 's, we get as $b \rightarrow \infty$

$$\begin{aligned} \mathbf{P}\left(\max_{1 \leq i \leq b} |\xi_i| > \sqrt{2 \ln b}\right) &= \mathbf{P}\left(\bigcup_{i=1}^b \{|\xi_i| > \sqrt{2 \ln b}\}\right) \leq \sum_{i=1}^b \mathbf{P}\left(|\xi_i| > \sqrt{2 \ln b}\right) \\ &= b \mathbf{P}\left(|\xi_1| > \sqrt{2 \ln b}\right) = 2b \mathbf{P}\left(\xi_1 > \sqrt{2 \ln b}\right) \\ &= 2b \Phi\left(-\sqrt{2 \ln b}\right) \sim \frac{2}{\sqrt{2\pi} \sqrt{2 \ln b}} b e^{-\ln b} \\ &= O\left(\frac{1}{\sqrt{\ln b}}\right) = o(1). \end{aligned}$$

Therefore

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

□

Proof of Lemma 2. By (56), for $0 < s < r < \infty$ as $b \rightarrow \infty$

$$\begin{aligned} \mathbf{P}\left(N(\sqrt{2r \ln b}, 1) \leq \sqrt{2s \ln b}\right) &= \mathbf{P}\left(N(0, 1) \leq -(\sqrt{r} - \sqrt{s}) \sqrt{2 \ln b}\right) \\ &= \frac{e^{-(\sqrt{r} - \sqrt{s})^2 \ln b}}{\sqrt{2\pi} (\sqrt{r} - \sqrt{s}) \sqrt{2 \ln b}} (1 + o(1)) \\ &= O\left(b^{-(\sqrt{s} - \sqrt{r})^2} \ln^{-1/2} b\right). \end{aligned}$$

In order to prove relation (25) for the chi-square distribution, we shall apply formulas (1) and (2) in [10] that give an expression of the CDF of a non-central chi-square

distribution with odd degrees of freedom in terms of the CDF and PDF of the standard normal distribution. Specifically, if $F_\nu(x; \gamma)$ denotes the CDF of a non-central chi-square $\chi_\nu^2(\gamma)$ distribution, then (see formulas (1) and (2) in [10])

$$F_\nu(x; \gamma) = \Phi(\sqrt{\gamma} + \sqrt{x}) - \Phi(\sqrt{\gamma} - \sqrt{x}), \quad \nu = 1, \gamma > 0, \quad (57)$$

$$F_\nu(x; \gamma) = F_1(x; \gamma) + \sum_{i=1}^k \sum_{j=1}^i \binom{i-1}{j-1} 2^j F_1^{(j)}(x; \gamma), \quad \nu = 2k+1, k = 1, 2, \dots, \gamma > 0, \quad (58)$$

where $F_1^{(j)}(x; \gamma)$ is the j th derivative of $F_1(x; \gamma)$ with respect to γ .

Now, put $a = \sqrt{\gamma} + \sqrt{x}$ and $b = \sqrt{\gamma} - \sqrt{x}$ and observe that formulas (57) and (58) imply that for all odd $\nu \geq 1$ the CDF $F_\nu(x; \gamma)$ is a linear combination of the functions $\Phi(a)$, $\Phi(b)$, $\gamma^{-l} a^m \varphi(a)$, and $\gamma^{-l} b^m \varphi(b)$ with $l = 1/2, 1, 3/2, 2, 5/2, \dots$ and $m = 0, 1, 2, \dots$. Therefore, taking in (57) and (58) $x = 2s \ln b$ and $\gamma = 2r \ln b$ with $0 < s < r < \infty$ and using relation (56) we obtain as $b \rightarrow \infty$

$$\begin{aligned} F_1(2s \ln b; 2r \ln b) &= \Phi\left(\sqrt{2r \ln b} + \sqrt{2s \ln b}\right) - \Phi\left(\sqrt{2r \ln b} - \sqrt{2s \ln b}\right) \\ &= 1 - \Phi\left((\sqrt{r} - \sqrt{s})\sqrt{2 \ln b}\right) - \left(1 - \Phi\left((\sqrt{r} + \sqrt{s})\sqrt{2 \ln b}\right)\right) \\ &\sim \frac{e^{-(\sqrt{r}-\sqrt{s})^2 \ln b}}{\sqrt{2\pi}(\sqrt{r} - \sqrt{s})\sqrt{2 \ln b}} - \frac{e^{-(\sqrt{r}+\sqrt{s})^2 \ln b}}{\sqrt{2\pi}(\sqrt{r} + \sqrt{s})\sqrt{2 \ln b}} \\ &= \frac{b^{-(\sqrt{r}-\sqrt{s})^2}}{2\sqrt{\pi \ln b}(\sqrt{r} - \sqrt{s})}(1 + o(1)), \end{aligned}$$

and also for all odd $\nu \geq 1$ as $b \rightarrow \infty$

$$\begin{aligned} F_\nu(2s \ln b; 2r \ln b) &\asymp F_1(2s \ln b; 2r \ln b) \\ &\asymp b^{-(\sqrt{r}-\sqrt{s})^2} \ln^{-1/2} b. \end{aligned}$$

Finally, noting that for all even $\nu \geq 2$

$$F_\nu(x; \gamma) = \mathbf{P}(\chi_{\nu-1}^2(\gamma) + (N(0, 1))^2 \leq x) \leq \mathbf{P}(\chi_{\nu-1}^2(\gamma) \leq x) = F_{\nu-1}(x; \gamma),$$

we conclude that relation (25) holds for all (odd and even) $\nu \geq 1$. The proof is complete.

□

We now prove Proposition 1 stated in Section 7.

Proof of Proposition 1. We begin by noting that for all $k = 1, \dots, b$

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{1,[k]} &\sim N_{p_0}(\boldsymbol{\mu}_{1,[k]}, (2/n)\boldsymbol{\Sigma}_{[k]}), \quad \hat{\boldsymbol{\mu}}_{2,[k]} \sim N_{p_0}(\boldsymbol{\mu}_{2,[k]}, (2/n)\boldsymbol{\Sigma}_{[k]}), \\ \hat{\boldsymbol{\mu}}_{2,[k]} \pm \hat{\boldsymbol{\mu}}_{1,[k]} &\sim N_{p_0}(\boldsymbol{\mu}_{2,[k]} \pm \boldsymbol{\mu}_{1,[k]}, (4/n)\boldsymbol{\Sigma}_{[k]}).\end{aligned}$$

Recall the the risk $\mathcal{R}(\hat{\Psi})$ of the classifier $\hat{\Psi}$ given by (20) is either the convex combination or the maximum of $\mathbf{E}_{\Pi_2}(\hat{\Psi})$ and $\mathbf{E}_{\Pi_1}(1 - \hat{\Psi})$. Let us first show that, uniformly in $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$,

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_2}(\hat{\Psi}) \rightarrow 0, \quad b \rightarrow \infty.$$

We can rewrite $\hat{\Psi}$ as follows:

$$\begin{aligned}\hat{\Psi} = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} + \sum_{k=1: \omega_k=1}^b \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) \right. \\ \left. - \frac{1}{2} \sum_{k=1: \omega_k=1}^b \left[(\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right] \right. \\ \left. \leq \frac{1}{2} \sum_{k=1: \omega_k=1}^b (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}.\end{aligned}$$

Then, denoting for all $k = 1, \dots, b$ such that $\omega_k = 1$

$$\begin{aligned}\xi_k &= \xi_{k,b} := \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \\ \eta_k &= \eta_{k,b} := \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}), \\ \tau_k &= \tau_{k,b} := (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]},\end{aligned}$$

we may continue

$$\hat{\Psi} = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\xi_k + \eta_k) - \frac{1}{2} \sum_{k=1: \omega_k=1}^b \tau_k \leq \frac{1}{2} \sum_{k=1: \omega_k=1}^b (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}. \quad (59)$$

Let us now show that $\sum_{k=1: \omega_k=1}^b (\xi_k + \eta_k)$ is the main term, and $\frac{1}{2} \sum_{k=1: \omega_k=1}^b \tau_k$ is the remainder term such that as $b \rightarrow \infty$

$$\sum_{k=1: \omega_k=1}^b \tau_k = o_{\mathbf{P}_{\Pi_2}} \left(\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) = o_{\mathbf{P}_{\Pi_2}} (b^{1-\beta-\theta} \ln b), \quad (60)$$

where the second equality is due to the assumption that for all $k = 1, \dots, b$ such that $\omega_k = 1$

$$\boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} = 8rb^{-\theta} \ln b. \quad (61)$$

and the fact that $\sum_{k=1}^b \omega_k \sim b^{1-\beta}$.

First, we examine the distributions of ξ_k and η_k under Π_2 . Note that for all $k = 1, \dots, b$ such that $\omega_k = 1$

$$\xi_k \sim N \left(\boldsymbol{\mu}_{2,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right). \quad (62)$$

Next, recalling that $\hat{\boldsymbol{\mu}}_{[k]} = \hat{\boldsymbol{\mu}}_{2,[k]} - \hat{\boldsymbol{\mu}}_{1,[k]}$, in view of (16), for all $k = 1, \dots, b$ such that $\omega_k = 1$,

$$\eta_k = \left(((n/4)\boldsymbol{\Sigma}_{[k]})^{-1/2} \mathbf{X}_{0,[k]} \right)^\top ((4/n)\boldsymbol{\Sigma}_{[k]})^{-1/2} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) \stackrel{d}{=} \mathbf{U}_k^\top \mathbf{V}_k, \quad (63)$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are independent normal random vectors satisfying (under Π_2)

$$\mathbf{U}_k \sim N_{p_0} \left(\left((n/4) \boldsymbol{\Sigma}_{[k]} \right)^{-1/2} \boldsymbol{\mu}_{2,[k]}, (4/n) \mathbf{I} \right), \quad \mathbf{V}_k \sim N_{p_0}(\mathbf{0}, \mathbf{I}), \quad (64)$$

with $\mathbf{I}$ being the $p_0 \times p_0$ identity matrix.

Therefore, using (63) and (64) and recalling that $n \sim b^\theta$, we can see that, for all $k = 1, \dots, b$ such that $\omega_k = 1$, as $b \rightarrow \infty$

$$\mathbf{E}_{\Pi_2}(\eta_k) = 0, \quad \mathbf{Var}_{\Pi_2}(\eta_k) \asymp b^{-\theta}. \quad (65)$$

Hence, with the use of the Cauchy-Schwarz inequality applied to $\mathbf{Cov}_{\Pi_2}(\xi_k, \eta_k)$, it follows from (61), (62), and (65) that, for all $k = 1, \dots, b$ such that $\omega_k = 1$, as $b \rightarrow \infty$

$$\begin{aligned} \mathbf{E}_{\Pi_2}(\xi_k + \eta_k) &= \boldsymbol{\mu}_{2,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \\ \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k) &= \mathbf{Var}_{\Pi_2}(\xi_k) + \mathbf{Var}_{\Pi_2}(\eta_k) + 2\mathbf{Cov}_{\Pi_2}(\xi_k, \eta_k) \asymp b^{-\theta} \ln b. \end{aligned} \quad (66)$$

From this, the classifier $\hat{\Psi}$ in (59) takes the form

$$\hat{\Psi} = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\xi_k + \eta_k - \mathbf{E}_{\Pi_2}(\xi_k + \eta_k)) - \frac{1}{2} \sum_{k=1: \omega_k=1}^b \tau_k \leq -\frac{1}{2} \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}, \quad (67)$$

and, in order to complete the proof by the application of Chebyshev's inequality, it now remains to verify relation (60).

Let us examine the behaviour of the random variables τ_k . For all $k = 1, \dots, b$ such

that $\omega_k = 1$, we can rewrite τ_k in the form

$$\begin{aligned}
\tau_k &= (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} - (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \\
&+ (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \\
&= (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) + (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]} - \boldsymbol{\mu}_{1,[k]} - \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \\
&=: (4/n) \mathbf{S}_k^\top \mathbf{T}_k + W_k,
\end{aligned} \tag{68}$$

where

$$\begin{aligned}
\mathbf{S}_k &= ((4/n) \boldsymbol{\Sigma}_{[k]})^{-1/2} (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]}) \sim N_{p_0} \left(((4/n) \boldsymbol{\Sigma}_{[k]})^{-1/2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]}), \mathbf{I} \right), \\
\mathbf{T}_k &= ((4/n) \boldsymbol{\Sigma}_{[k]})^{-1/2} (\hat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) \sim N_{p_0}(0, \mathbf{I}), \\
W_k &= (\hat{\boldsymbol{\mu}}_{1,[k]} + \hat{\boldsymbol{\mu}}_{2,[k]} - \boldsymbol{\mu}_{1,[k]} - \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \sim N \left(0, (4/n) \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right).
\end{aligned}$$

Using representation (68), we now obtain the expected value and the variance of each τ_k in the sum in expression (67) under Π_2 .

First, writing $\mathbf{S}_k = (S_{k1}, \dots, S_{kp_0})^\top$ and $\mathbf{T}_k = (T_{k1}, \dots, T_{kp_0})^\top$, we have for all $k = 1, \dots, b$ for which $\omega_k = 1$,

$$\mathbf{E}_{\Pi_2}(\tau_k) = (4/n) \mathbf{E}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k) + \mathbf{E}_{\Pi_2}(W_k) = (4/n) \sum_{i=1}^{p_0} \mathbf{E}_{\Pi_2}(S_{ki} T_{ki}),$$

where, by the Cauchy-Schwarz inequality,

$$|\mathbf{E}_{\Pi_2}(S_{ki} T_{ki})| \leq \sqrt{\mathbf{E}_{\Pi_2}(S_{ki}^2) \mathbf{E}_{\Pi_2}(T_{ki}^2)} = O(1). \tag{69}$$

Hence, for all $k = 1, \dots, b$ for which $\omega_k = 1$, as $b \rightarrow \infty$

$$\mathbf{E}_{\Pi_2}(\tau_k) = O(n^{-1}) = O(b^{-\theta}). \tag{70}$$

Next, for all $k = 1, \dots, b$ for which $\omega_k = 1$, we can write

$$\mathbf{Var}_{\Pi_2}(\tau_k) = (16/n^2) \mathbf{Var}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k) + \mathbf{Var}_{\Pi_2}(W_k) + (8/n) \mathbf{Cov}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k, W_k), \quad (71)$$

and will analyze each term on the right side of (71) separately.

For all $k = 1, \dots, b$ for which $\omega_k = 1$, using (61), we get as $b \rightarrow \infty$

$$\mathbf{Var}_{\Pi_2}(W_k) = (4/n) \boldsymbol{\mu}_{[k]} \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} = O(b^{-2\theta} \ln b). \quad (72)$$

Next, by independence of the components of $\mathbf{S}_k$ and $\mathbf{T}_k$, for all $k = 1, \dots, b$ for which $\omega_k = 1$,

$$\begin{aligned} (16/n^2) \mathbf{Var}_{\Pi_2}(\mathbf{S}_k^\top \mathbf{T}_k) &= (16/n^2) \sum_{i=1}^{p_0} \mathbf{Var}_{\Pi_2}(S_{ki} T_{ki}) \\ &= (16/n^2) \sum_{i=1}^{p_0} \{ \mathbf{E}_{\Pi_2}(S_{ki}^2 T_{ki}^2) - (\mathbf{E}_{\Pi_2}(S_{ki} T_{ki}))^2 \}, \end{aligned} \quad (73)$$

where, by (69), as $b \rightarrow \infty$,

$$(\mathbf{E}_{\Pi_2}(S_{ki} T_{ki}))^2 = O(1),$$

and, by the Cauchy–Schwarz inequality,

$$\mathbf{E}_{\Pi_2}(S_{ki}^2 T_{ki}^2) \leq \sqrt{\mathbf{E}_{\Pi_2}(S_{ki}^4) \mathbf{E}_{\Pi_2}(T_{ki}^4)}.$$

Since each $T_{ki} \sim N(0, 1)$, it follows that $\mathbf{E}_{\Pi_2}(T_{ki}^4) = 3$. Also, $S_{ki} \sim N(m_{ki}, 1)$, where m_{ki} is the i th components of the vector $\mathbf{m}_k := ((4/n) \boldsymbol{\Sigma}_{[k]})^{-1/2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})$ whose squared Euclidian norm satisfies as $b \rightarrow \infty$

$$\mathbf{m}_k^\top \mathbf{m}_k = \sum_{i=1}^{p_0} m_{ki}^2 = \frac{n}{4} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]}) = O(\ln b),$$

where the last equality is due to Condition A stated in Section 5. Hence for all $i = 1, \dots, p_0$ and all $k = 1, \dots, b$ for which $\omega_k = 1$

$$m_{ki}^2 = O(\ln b),$$

and, using well-known formulas for the moments of a normal random variable, we get as $b \rightarrow \infty$

$$\mathbf{E}_{\Pi_2} (S_{ki}^4) = \mathbf{E} ((N(m_{ki}, 1))^4) = m_{ki}^4 + 6m_{ki}^2 + 3 = O(\ln^2 b).$$

From this, as $b \rightarrow \infty$

$$\mathbf{E}_{\Pi_2} (S_{ki}^2 T_{ki}^2) \leq \sqrt{\mathbf{E}_{\Pi_2} (S_{ki}^4) \mathbf{E}_{\Pi_2} (T_{ki}^4)} = \sqrt{O(\ln^2 b) O(1)} = O(\ln b).$$

Summarizing, we obtain from (73) that as $b \rightarrow \infty$

$$(16/n^2) \mathbf{Var}_{\Pi_2} (\mathbf{S}_k^\top \mathbf{T}_k) = O(b^{-2\theta} \ln b),$$

and hence, by (72) and the Cauchy-Schwarz inequality applied to $\mathbf{Cov}_{\Pi_2} (\mathbf{S}_k^\top \mathbf{T}_k, W_k)$, we deduce from (71) that, for all $k = 1, \dots, b$ for which $\omega_k = 1$, as $b \rightarrow \infty$

$$\mathbf{Var}_{\Pi_2} (\tau_k) = O(b^{-2\theta} \ln b), \tag{74}$$

Now, relations (70) and (74), jointly with the assumption $\sum_{k=1}^b \omega_k \sim b^{1-\beta}$, give as $b \rightarrow \infty$

$$\mathbf{E}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b \tau_k \right) = O(b^{1-\beta-\theta}), \quad \mathbf{Var}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b \tau_k \right) = O(b^{1-\beta-2\theta} \ln b).$$

This, together with Chebyshev's inequality, leads to (60) because for any $\epsilon > 0$ as $b \rightarrow \infty$

$$\begin{aligned}
& \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1: \omega_k=1}^b \tau_k \right| \geq \epsilon b^{1-\beta-\theta} \ln b \right) \\
&= \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1: \omega_k=1}^b (\tau_k - \mathbf{E}_{\Pi_2}(\tau_k)) + \sum_{k=1: \omega_k=1}^b \mathbf{E}_{\Pi_2}(\tau_k) \right| \geq \epsilon b^{1-\beta-\theta} \ln b \right) \\
&= \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1: \omega_k=1}^b (\tau_k - \mathbf{E}_{\Pi_2}(\tau_k)) \right| \geq \epsilon b^{1-\beta-\theta} \ln b (1 + o(1)) \right) \\
&\leq \frac{\sum_{k=1: \omega_k=1}^b \mathbf{Var}_{\Pi_2}(\tau_k)}{(b^{1-\beta-\theta} \ln b)^2 (1 + o(1))} \asymp \frac{b^{1-\beta-2\theta} \ln b}{b^{2-2\beta-2\theta} \ln^2 b} = \frac{1}{b^{1-\beta} \ln b} = o(1).
\end{aligned}$$

Therefore, under Π_2 , the classifier in (67) can be written in the form

$$\begin{aligned}
\hat{\Psi} &= \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (\xi_k + \eta_k - \mathbf{E}_{\Pi_2}(\xi_k + \eta_k)) \right. \\
&\quad \left. \leq -\frac{1}{2} \left(\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) (1 + o_{\mathbf{P}_{\Pi_2}}(1)) \right\}.
\end{aligned}$$

Applying again Chebyshev's inequality and using relation (66), uniformly in $(\beta, r) \in ((1-\theta)/2, 1-\theta) \times (0, 4)$, for all sufficiently large b ,

$$\begin{aligned}
& \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_2}(\hat{\Psi}) \\
&\leq \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{P}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b (\xi_k + \eta_k - \mathbf{E}_{\Pi_2}(\xi_k + \eta_k)) \leq -\frac{1}{4} \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \\
&\leq \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \frac{16 \sum_{k=1: \omega_k=1}^b \mathbf{Var}_{\Pi_2}(\xi_k + \eta_k)}{\left(\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right)^2} \\
&\asymp \frac{b^{1-\beta-\theta} \ln b}{(b^{1-\beta-\theta} \ln b)^2} = \frac{1}{b^{1-\beta-\theta} \ln b},
\end{aligned}$$

implying

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_2}(\hat{\Psi}) \rightarrow 0, \quad b \rightarrow \infty, \quad (75)$$

because by assumption $(1 - \theta)/2 < \beta < 1 - \theta$.

The fact that, uniformly in $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$,

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_1}(1 - \hat{\Psi}) \rightarrow 0, \quad b \rightarrow \infty, \quad (76)$$

is proved completely analogously. It now follows from (75) and (76) that, uniformly in $(\beta, r) \in ((1 - \theta)/2, 1 - \theta) \times (0, 4)$,

$$\lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\hat{\Psi}) = 0.$$

The proof is complete. $\square$

Let us now prove Proposition 2 stated in Section 9.

Proof of Proposition 2. Recall that

$$\omega_k = \mathbb{I}(\gamma_k \neq 0), \quad k = 1, \dots, b,$$

where γ_k is the noncentrality parameter of t_k in model (21). By assumption, the number of non-central chi-square statistics among $t_1, \dots, t_b$ is equal to

$$\sum_{k=1}^b \omega_k \sim b^{1-\beta},$$

and hence the number of central chi-square statistics among $t_1, \dots, t_b$ is

$$b - \sum_{k=1}^b \omega_k \sim b - b^{1-\beta}.$$

For the selector $\hat{\omega} = (\hat{\omega}_k)_{k=1}^b$ as given in (33) we can write

$$\begin{aligned}
\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E}|\hat{\omega} - \omega| &= \sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E} \sum_{k=1}^b |\hat{\omega}_k - \omega_k| = \sup_{\omega \in \mathcal{H}_{b,\beta}} \left(\sum_{k=1: \omega_k=0}^b \mathbf{E}|\hat{\omega}_k - \omega_k| + \sum_{k=1: \omega_k=1}^b \mathbf{E}|\hat{\omega}_k - \omega_k| \right) \\
&= \sup_{\omega \in \mathcal{H}_{b,\beta}} \left(\sum_{k=1: \omega_k=0}^b \mathbf{P}(t_k > (2+\epsilon) \ln b) + \sum_{k=1: \omega_k=1}^b \mathbf{P}(t_k \leq (2+\epsilon) \ln b) \right) \\
&\sim (b - b^{1-\beta}) \mathbf{P}(\chi_\nu^2(0) > (2+\epsilon) \ln b) + b^{1-\beta} \mathbf{P}(\chi_\nu^2(\gamma_b) \leq (2+\epsilon) \ln b) \\
&=: K_1 + K_2.
\end{aligned} \tag{77}$$

Applying (23) gives

$$\begin{aligned}
K_1 &= (b - b^{1-\beta}) \mathbf{P}(\chi_\nu^2(0) > (2+\epsilon) \ln b) \leq b \mathbf{P}(\chi_\nu^2(0) > (2+\epsilon) \ln b) \\
&= O(b^{1-(1+\epsilon/2)} \ln^{-1/2} b) = O(b^{-\epsilon/2} \ln^{-1/2} b) = o(1).
\end{aligned} \tag{78}$$

To estimate the second term K_2 , we will use relation (25) and the fact that, in the region $\mathbf{D}_1$,

$$r > \left(1 + \sqrt{1-\beta}\right)^2 > 1,$$

yielding for all sufficiently large b (and hence for all sufficiently small $\epsilon = \epsilon_b$) the relation

$$1 - \beta - (\sqrt{r} - \sqrt{1+\epsilon/2})^2 < 0.$$

Therefore, in view of Lemma 2, as $b \rightarrow \infty$

$$\begin{aligned}
K_2 &= b^{1-\beta} \mathbf{P}(\chi_\nu^2(\gamma_b) \leq (2+\epsilon) \ln b) \\
&= O(b^{1-\beta-(\sqrt{r}-\sqrt{1+\epsilon/2})^2} \ln^{-1/2} b) = o(1).
\end{aligned} \tag{79}$$

Now, combining (77), (78) and (79), we arrive at (36). The proof is complete. $\square$

Remark 3. We now explain the idea behind having ϵ in the definition of the

threshold level $t^* = (2 + \epsilon) \ln b$ of the selector $\hat{\omega}$. Although an analogous selector with threshold $t^* = 2 \ln b$ (without ϵ) will also work (as seen from the proof of Proposition 1), its maximum Hamming risk will decay to zero much slower than that of our selector. Indeed, by looking at (77) we see that, asymptotically, the maximum Hamming risk $\sup_{\omega \in \mathcal{H}_{b,\beta}} \mathbf{E}|\hat{\omega}(\beta_{\hat{m}}) - \omega|$ is composed of the two terms K_1 and K_2 . The term K_2 tends to zero fast enough, as a power of b (see relation (79)). As for the term K_1 , removing ϵ would make it going to zero at a slow logarithmic rate, whereas keeping ϵ makes it tending to zero much faster (see relation (78)).

Proof of Proposition 3. The proof is largely based on the fact that if index $m_0 \in \{1, \dots, M-1\}$ is such that $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$, then with high probability $\hat{m} \geq m_0$. To verify this fact, Bernstein's inequality will be used.

Bernstein's inequality (see, for example, pp. 164–165 of [3]). *If $\mathbb{X}_1, \dots, \mathbb{X}_b$ are independent random variables such that for all $i = 1, \dots, b$ and for some $H > 0$*

$$\mathbf{E}(\mathbb{X}_i) = 0 \quad \text{and} \quad |\mathbf{E}(\mathbb{X}_i^m)| \leq \frac{\mathbf{E}(\mathbb{X}_i^2)}{2} H^{m-2} m! < \infty, \quad m = 2, 3, \dots, \quad (80)$$

then

$$\max \{ \mathbf{P}(\mathbb{S}_b \geq t), \mathbf{P}(\mathbb{S}_b \leq -t) \} \leq \begin{cases} \exp(-t^2/4D_b^2) & \text{if } 0 \leq t \leq D_b^2/H, \\ \exp(-t/4H) & \text{if } t \geq D_b^2/H, \end{cases} \quad (81)$$

where $\mathbb{S}_b = \sum_{i=1}^b \mathbb{X}_i$ and $D_b^2 = \sum_{i=1}^b \mathbf{E}(\mathbb{X}_i^2)$.

Remark 4. Observe that for independent random variables $\mathbb{X}_1, \dots, \mathbb{X}_b$ with the property

$$\mathbf{E}(\mathbb{X}_i) = 0 \quad \text{and} \quad |\mathbb{X}_i| \leq M, \quad i = 1, \dots, b,$$

for some $M > 0$, the Bernstein condition (80) holds with $H = M/3$. Below we will use Bernstein's inequality in the case of $t \geq D_b^2/H$.

Consider the selector

$$\widehat{\omega}(\beta_{\hat{m}}) = (\widehat{\omega}_1(\beta_{\hat{m}}), \dots, \widehat{\omega}_b(\beta_{\hat{m}}))$$

given by formulas (41)–(43) based on the statistics $t_1, \dots, t_b$ as in model (21) with $\nu = p_0$. Let index m_0 be such that the true (but unknown) parameter $\beta \in ((1 - \theta)/2, 1 - \theta)$ satisfies $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$. Using the Law of Total Expectation, we can write

$$\begin{aligned} \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E} |\widehat{\omega}(\beta_{\hat{m}}) - \omega| &\leq \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E} (|\widehat{\omega}(\beta_{\hat{m}}) - \omega| | \hat{m} \geq m_0) \mathbf{P}(\hat{m} \geq m_0) \\ &+ \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E} (|\widehat{\omega}(\beta_{\hat{m}}) - \omega| | \hat{m} < m_0) \mathbf{P}(\hat{m} < m_0) \\ &=: J_1 + J_2. \end{aligned} \tag{82}$$

Consider the term J_1 . When $\hat{m} \geq m_0$, it follows from the triangle inequality and the definition of $\hat{m}$ in (43) that for all $\omega \in \mathcal{H}_{b,\beta}$

$$|\widehat{\omega}(\beta_{\hat{m}}) - \omega| \leq |\widehat{\omega}(\beta_{\hat{m}}) - \widehat{\omega}(\beta_{m_0})| + |\widehat{\omega}(\beta_{m_0}) - \omega| \leq v_{m_0} + |\widehat{\omega}(\beta_{m_0}) - \omega|,$$

where as $b \rightarrow \infty$

$$b^{\beta-1} v_{m_0} = O(\tau_b^{-1}) = o(1).$$

Next, since by the Law of Total Expectation

$$\mathbf{E}(Y) = \mathbf{E}(Y|B)\mathbf{P}(B) + \mathbf{E}(Y|B^c)\mathbf{P}(B^c),$$

it follows that for any non-negative random variable Y , for which the above expectations are well defined,

$$\mathbf{E}(Y|B)\mathbf{P}(B) \leq \mathbf{E}(Y),$$

and hence we can write

$$\begin{aligned}
J_1 &= \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}(|\widehat{\omega}(\beta_{m_0}) - \omega| | \hat{m} \geq m_0) \mathbf{P}(\hat{m} \geq m_0) \leq b^{\beta-1} \mathbf{E}|\widehat{\omega}(\beta_{m_0}) - \omega| \\
&\sim b^{\beta-1} [(b - b^{1-\beta}) \mathbf{P}(\chi_\nu^2(0) > (2\beta_{m_0} + \epsilon) \ln b) + b^{1-\beta} \mathbf{P}(\chi_\nu^2(\gamma_b) \leq (2\beta_{m_0} + \epsilon) \ln b)] \\
&\leq b^\beta \mathbf{P}(\chi_\nu^2(0) > (2\beta_{m_0} + \epsilon) \ln b) + \mathbf{P}(\chi_\nu^2(\gamma_b) \leq (2\beta_{m_0} + \epsilon) \ln b) \\
&=: T_1 + T_2.
\end{aligned} \tag{83}$$

For the first term T_1 , using (23) and (40), we obtain

$$T_1 = O\left(b^{\beta-\beta_{m_0}-\epsilon/2} \ln^{-1/2} b\right) = O\left(b^{\Delta-\epsilon/2} \ln^{-1/2} b\right) = O\left(b^{-\epsilon/2} \ln^{-1/2} b\right) = o(1). \tag{84}$$

Let us now estimate the second term T_2 . For this, we apply relation (25) and the fact that for all sufficiently large b (recall that $\epsilon = \epsilon_b \rightarrow 0$ as $b \rightarrow \infty$)

$$r > \beta > \beta_{m_0} + \epsilon/2,$$

to obtain as $b \rightarrow \infty$

$$T_2 = \mathbf{P}(\chi_{p_0}^2(\gamma_b) \leq (2\beta_{m_0} + \epsilon) \ln b) = O\left(b^{-(\sqrt{r}-\sqrt{\beta_{m_0}+\epsilon/2})^2} \ln^{-1/2} b\right) = o(1), \tag{85}$$

Therefore, using (83)–(85), we obtain, uniformly in $(1 - \theta)/2 < \beta < 1 - \theta$,

$$\begin{aligned}
J_1 &= \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}(|\widehat{\omega}(\beta_{\hat{m}}) - \omega| | \hat{m} \geq m_0) \mathbf{P}(\hat{m} \geq m_0) \\
&\leq b^{\beta-1} v_{m_0} + \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}|\widehat{\omega}(\beta_{m_0}) - \omega| \leq O(\tau_b^{-1}) + T_1 + T_2 = o(1).
\end{aligned} \tag{86}$$

It remains to show that, uniformly in $(1 - \theta)/2 < \beta < 1 - \theta$, it is also true that

$J_2 = o(1)$. Consider the term J_2 . We have

$$J_2 = \sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}(|\widehat{\omega}(\beta_{\hat{m}}) - \omega| | \hat{m} < m_0) \mathbf{P}(\hat{m} < m_0) \leq b^{\beta-1} b \mathbf{P}(\hat{m} < m_0) = b^\beta \mathbf{P}(\hat{m} < m_0).$$

Now, using Bernstein's inequality (81), we show that $\mathbf{P}(\hat{m} < m_0)$ is small yielding $J_2 = o(1)$ as $b \rightarrow \infty$.

Indeed, using the definition of $\hat{m}$,

$$\begin{aligned} & \mathbf{P}(\hat{m} < m_0) \\ &= \sum_{k=1}^{m_0-1} \mathbf{P}(\hat{m} = k) = \sum_{k=1}^{m_0-1} \mathbf{P}(\exists j \in \{1, \dots, k\} : |\widehat{\omega}(\beta_{k+1}) - \widehat{\omega}(\beta_j)| > v_j) \\ &\leq \sum_{k=1}^{m_0-1} \sum_{j=1}^k \mathbf{P}(|\widehat{\omega}(\beta_{k+1}) - \widehat{\omega}(\beta_j)| > v_j) = \sum_{k=1}^{m_0-1} \sum_{j=1}^k \mathbf{P}\left(\sum_{i=1}^b |\widehat{\omega}_i(\beta_{k+1}) - \widehat{\omega}_i(\beta_j)| > v_j\right) \\ &= \sum_{k=1}^{m_0-1} \sum_{j=1}^k \mathbf{P}\left(\sum_{i=1}^b \mathbb{I}\{(2\beta_j + \epsilon) \ln b < t_i \leq (2\beta_{k+1} + \epsilon) \ln b\} > v_j\right). \end{aligned} \quad (87)$$

Now, introducing the events

$$A_i = A_{i,j,k+1} = \{(2\beta_j + \epsilon) \ln b < t_i \leq (2\beta_{k+1} + \epsilon) \ln b\}, \quad i = 1, \dots, b,$$

we obtain from (87) that

$$\mathbf{P}(\hat{m} < m_0) \leq \sum_{k=1}^{m_0-1} \sum_{j=1}^k \mathbf{P}\left(\sum_{i=1}^b \mathbb{I}\{A_i\} > v_j\right) = \sum_{k=1}^{m_0-1} \sum_{j=1}^k \mathbf{P}\left(\sum_{i=1}^b \mathbb{X}_i > v_j - \sum_{i=1}^b \mathbf{P}(A_i)\right), \quad (88)$$

where the random variables $\mathbb{X}_i$ are defined by

$$\mathbb{X}_i = \mathbb{X}_{i,j,k+1} = \mathbb{I}\{A_i\} - \mathbf{P}(A_i), \quad i = 1, \dots, b. \quad (89)$$

In order to apply Bernstein's inequality to the probability $\mathbf{P}\left(\sum_{i=1}^b \mathbb{X}_i > v_j - \sum_{i=1}^b \mathbf{P}(A_i)\right)$

on the right-hand side of (88), we first observe that for all $j = 1, \dots, k$ and $k = 1, \dots, m_0 - 1$

$$\beta_j < \beta_{k+1} \leq \beta_{m_0} < \beta,$$

and hence, recalling that $\epsilon = \epsilon_b$ satisfies (35), for all sufficiently large b

$$\beta_j + \frac{\epsilon}{2} < \beta_{k+1} + \frac{\epsilon}{2} < \beta.$$

This implies that for all large enough b

$$b^{1-(\beta_j+\epsilon/2)} \ln^{-1/2} b \gg b^{1-(\beta_{k+1}+\epsilon/2)} \ln^{-1/2} b \gg b^{1-\beta}.$$

Therefore, using (23), we get

$$\begin{aligned} \sum_{i=1}^b \mathbf{P}(A_i) &= \sum_{i=1}^b [\mathbf{P}(t_i > (2\beta_j + \epsilon) \ln b) - \mathbf{P}(t_i > (2\beta_{k+1} + \epsilon) \ln b)] \\ &\sim (b - b^{1-\beta}) [\mathbf{P}(\chi_\nu^2(0) > (2\beta_j + \epsilon) \ln b) - \mathbf{P}(\chi_\nu^2(0) > (2\beta_{k+1} + \epsilon) \ln b)] \\ &\quad + b^{1-\beta} [\mathbf{P}(\chi_\nu^2(\gamma_b) > (2\beta_j + \epsilon) \ln b) - \mathbf{P}(\chi_\nu^2(\gamma_b) > (2\beta_{k+1} + \epsilon) \ln b)] \\ &= O\left(b^{1-(\beta_j+\epsilon/2)} \ln^{-1/2} b\right) + O\left(b^{1-(\beta_{k+1}+\epsilon/2)} \ln^{-1/2} b\right) + O(b^{1-\beta}) \\ &= O\left(b^{1-(\beta_j+\epsilon/2)} \ln^{-1/2} b\right). \end{aligned} \tag{90}$$

Relation (90) and the definition of v_j , namely:

$$v_j = b^{1-\beta_j} / \tau_b, \quad 1, \dots, b,$$

where

$$\tau_b = o\left(\min(\ln b, b^{\epsilon/2})\right),$$

now show that

$$\sum_{i=1}^b \mathbf{P}(A_i) = o(v_j),$$

and hence as $b \rightarrow \infty$

$$v_j - \sum_{i=1}^b \mathbf{P}(A_i) = v_j(1 + o(1)).$$

Also, since the variance of a random variable taking values in $\{0, 1\}$ is smaller than its expectation, we have by (90) and by the independence of $t_1, \dots, t_b$ that

$$\begin{aligned} \mathbf{Var} \left(\sum_{i=1}^b \mathbb{X}_i \right) &= \mathbf{Var} \left(\sum_{i=1}^b \mathbb{I}\{A_i\} \right) = \sum_{i=1}^b \mathbf{Var} (\mathbb{I}\{A_i\}) \leq \sum_{i=1}^b \mathbf{E} (\mathbb{I}\{A_i\}) \\ &= \sum_{i=1}^b \mathbf{P}(A_i) = O \left(b^{1-(\beta_j+\epsilon/2)} \ln^{-1/2} b \right). \end{aligned}$$

Thus, the random variables $\mathbb{X}_1, \dots, \mathbb{X}_b$ defined in (89) have the following properties:

$$|\mathbb{X}_i| \leq 2, \quad \mathbf{E}(\mathbb{X}_i) = 0, \quad i = 1, \dots, b,$$

and for all $j = 1, \dots, k$ and $k = 1, \dots, m-1$

$$D_b^2 = \sum_{i=1}^b \mathbf{E}(\mathbb{X}_i^2) = \sum_{i=1}^b \mathbf{E}(\mathbb{X}_{i,j,k+1}^2) = \mathbf{Var} \left(\sum_{i=1}^b \mathbb{X}_{i,j,k+1} \right) = O \left(b^{1-(\beta_j+\epsilon/2)} \ln^{-1/2} b \right).$$

Therefore the application of Bernstein's inequality with $t = v_j(1 + o(1))$ and the fact that $\beta_j \leq \beta_{m_0-1}$ for all $j = 1, \dots, k$ and $k = 1, \dots, m_0-1$, give

$$\mathbf{P} \left(\sum_{i=1}^b \mathbb{X}_i > v_j - \sum_{i=1}^b \mathbf{P}(A_i) \right) \leq \exp \left(-\frac{3v_j(1 + o(1))}{8} \right) \leq \exp \left(-\frac{3b^{1-\beta_{m_0-1}}}{8\tau_b} (1 + o(1)) \right).$$

From this and (88) we deduce that for all large enough b

$$\mathbf{P}(\hat{m} < m_0) \leq M^2 \exp \left(-\frac{b^{1-\beta_{m_0-1}}}{4\tau_b} \right),$$

and hence

$$J_2 \leq b^\beta \mathbf{P}(\hat{m} < m_0) \leq b^\beta M^2 \exp\left(-\frac{b^{1-\beta_{m_0-1}}}{4\tau_b}\right) = o(1), \quad (91)$$

where the last equality is due to the fact that

$$0 < \theta < 1 - \beta_{m_0-1} < 1,$$

that is, $1 - \beta_{m_0-1}$ is separated from zero, which follows from the assumptions $(1 - \theta)/2 < \beta < 1 - \theta$ and $\beta_{m_0-1} < \beta_{m_0} < \beta \leq \beta_{m_0+1}$.

Now, combining (82), (86) and (91), we obtain

$$\sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}|\widehat{\omega}(\beta_{\hat{m}}) - \omega| = J_1 + J_2 = o(1), \quad b \rightarrow \infty,$$

uniformly in $(\beta, r) \in \mathbf{D}_2$. This shows that the selector $\widehat{\omega}(\beta_{\hat{m}})$ provides almost full selection in the region $\mathbf{D}_2$. $\square$

Remark 5. As seen from relations (83)–(86), having ϵ in the definition of the threshold level $t^* = (2\beta_{\hat{m}} + \epsilon) \ln b$ of the selector $\widehat{\omega}(\beta_{\hat{m}})$ makes its normalized maximum Hamming risk $\sup_{\omega \in \mathcal{H}_{b,\beta}} b^{\beta-1} \mathbf{E}|\widehat{\omega}(\beta_{\hat{m}}) - \omega|$ tending to zero at a faster rate than that of an analogous selector with threshold $t^* = 2\beta_{\hat{m}} \ln b$ (without ϵ).

Finally, we prove Proposition 4 stated in Section 10.

Proof of Proposition 4. For $l = 1, 2$, the classifier $\tilde{\Psi}_l$ introduced in Section 10 can be written in the form

$$\tilde{\Psi}_l = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b \left(\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - \frac{1}{2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) - \left(\sum_{k=1: \omega_k=1}^b - \sum_{k=1: \tilde{\omega}_{l,k}=1}^b \right) \left(\mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - \frac{1}{2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \leq 0 \right\}.$$

Observe that only those indices k for which either $\omega_k = 1$ and $\tilde{\omega}_{l,k} = 0$ or $\omega_k = 0$ and $\tilde{\omega}_{l,k} = 1$ contribute to the difference of the two sums. With probability one, the second case of $\omega_k = 0$ and $\tilde{\omega}_{l,k} = 1$ gives nothing because, under Π_2 , the random variables

$$v_k = v_{k,b} := \mathbf{X}_{0,[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - \frac{1}{2} (\boldsymbol{\mu}_{1,[k]} + \boldsymbol{\mu}_{2,[k]})^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \quad k = 1, \dots, b,$$

satisfy

$$v_k \sim N \left(\frac{1}{2} \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \quad (92)$$

for all $k = 1, \dots, b$ such that $\omega_k = 1$, and

$$v_k = 0 \quad \text{with probability one,}$$

for all $k = 1, \dots, b$ such that $\omega_k = 0$. Therefore, when $\mathbf{X}_0 \sim \Pi_2$, with probability one,

$$\tilde{\Psi}_l = \mathbb{I} \left\{ \sum_{k=1: \omega_k=1}^b (v_k - \mathbf{E}_{\Pi_2}(v_k)) - \sum_{k=1: \omega_k=1, \tilde{\omega}_{l,k}=0}^b v_k \leq -\frac{1}{2} \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \Sigma_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right\}. \quad (93)$$

Now, for $l = 1, 2$, we denote

$$R_{l,b} := \sum_{k=1: \omega_k=1, \tilde{\omega}_{l,k}=0}^b v_k$$

and show that $R_{l,b}$, $l = 1, 2$, is the remainder term satisfying as $b \rightarrow \infty$

$$R_{l,b} = o_{\mathbf{P}_{\Pi_2}}(b^{1-\beta-\theta} \ln b). \quad (94)$$

For $l = 1, 2$, we have

$$\mathbf{E}_{\Pi_2}(R_{l,b}) = \left(\sum_{k=1: \omega_k=1}^b v_k \mathbb{I}(\omega_k - \tilde{\omega}_{l,k} = 1) \right) = \sum_{k=1: \omega_k=1}^b \mathbf{E}_{\Pi_2}(v_k) \mathbf{P}_{\Pi_2}(\omega_k - \tilde{\omega}_{l,k} = 1),$$

where, by assumption, for all $k = 1, \dots, b$ for which $\omega_k = 1$

$$\mathbf{E}_{\Pi_2}(v_k) = \frac{1}{2} \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} = 8rb^{-\theta} \ln b,$$

for some $0 < \theta < 1$ and some $\beta < r < 4$. Hence, using Markov's inequality, we may continue

$$\begin{aligned} \mathbf{E}_{\Pi_2}(R_{l,b}) &= (8rb^{-\theta} \ln b) \sum_{k=1: \omega_k=1}^b \mathbf{P}_{\Pi_2}(|\omega_k - \tilde{\omega}_{l,k}| = 1) \\ &\leq (8rb^{-\theta} \ln b) \sum_{k=1: \omega_k=1}^b \mathbf{E}_{\Pi_2}|\omega_k - \tilde{\omega}_{l,k}| = O(b^{-\theta} \ln b) \mathbf{E}_{\Pi_2}|\omega - \tilde{\omega}_l|. \end{aligned}$$

From this, applying Propositions 2 and 3, we get as $b \rightarrow \infty$

$$\begin{aligned} \mathbf{E}_{\Pi_2}(R_{1,b}) &= O(b^{-\theta} \ln b) o(1) = o(b^{-\theta} \ln b), & \text{for all } (\beta, r) \in \mathbf{D}_1 \\ \mathbf{E}_{\Pi_2}(R_{2,b}) &= O(b^{-\theta} \ln b) o(b^{1-\beta}) = o(b^{1-\beta-\theta} \ln b), & \text{for all } (\beta, r) \in \mathbf{D}_2. \end{aligned}$$

Similar arguments, together with (92) and the fact that the variance of a random variable that takes only two values zero and 1 cannot be larger than its expectation, lead to

$$\begin{aligned} \mathbf{Var}_{\Pi_2}(R_{1,b}) &= O(b^{-\theta} \ln b) o(1) = o(b^{-\theta} \ln b), & \text{for all } (\beta, r) \in \mathbf{D}_1 \\ \mathbf{Var}_{\Pi_2}(R_{2,b}) &= O(b^{-\theta} \ln b) o(b^{1-\beta}) = o(b^{1-\beta-\theta} \ln b), & \text{for all } (\beta, r) \in \mathbf{D}_2. \end{aligned}$$

Hence, by Chebyshev's inequality, for $l = 1, 2$ and for any $\epsilon > 0$

$$\begin{aligned} \mathbf{P}_{\Pi_2} (|R_{l,b}|/(b^{1-\beta-\theta} \ln b) \geq \epsilon) &= \mathbf{P}_{\Pi_2} (|R_{l,b} - \mathbf{E}_{\Pi_2}(R_{l,b})| \geq \epsilon(b^{1-\beta-\theta} \ln b)(1 + o(1))) \\ &\leq \frac{\mathbf{Var}_{\Pi_2}(R_{l,b})}{(b^{1-\beta-\theta} \ln b)^2(1 + o(1))}, \end{aligned}$$

yielding that for all $(\beta, r) \in \mathbf{D}_1$ as $b \rightarrow \infty$

$$\mathbf{P}_{\Pi_2} (|R_{1,b}|/(b^{1-\beta-\theta} \ln b) \geq \epsilon) = O(b^{-(2-2\beta-\theta)} \ln^{-1} b) = o(1),$$

and, also, for all $(\beta, r) \in \mathbf{D}_2$ as $b \rightarrow \infty$

$$\mathbf{P}_{\Pi_2} (|R_{2,b}|/(b^{1-\beta-\theta} \ln b) \geq \epsilon) = O(b^{-(1-\beta-\theta)} \ln^{-1} b) = o(1).$$

Thus, relation (94) is verified. Using this relation, the fact that

$$\sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} = O(b^{1-\beta-\theta} \ln b), \quad b \rightarrow \infty,$$

and Chebyshev's inequality, we obtain from (93) that, uniformly in $(\beta, r) \in \mathbf{D}_l$, $l = 1, 2$, as $b \rightarrow \infty$

$$\begin{aligned} &\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_2}(\tilde{\Psi}_l) \\ &= \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{P}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b (v_k - \mathbf{E}_{\Pi_2}(v_k)) \leq -\frac{1}{2} \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} (1 + o_{\mathbf{P}_{\Pi_2}}(1)) \right) \\ &\leq \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{P}_{\Pi_2} \left(\sum_{k=1: \omega_k=1}^b (v_k - \mathbf{E}_{\Pi_2}(v_k)) \leq -\frac{1}{4} \sum_{k=1: \omega_k=1}^b \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} \right) \\ &\leq \frac{16 \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \sum_{k=1: \omega_k=1}^b \mathbf{Var}_{\Pi_2}(v_k)}{(8r b^{1-\beta-\theta} \ln b)^2} \asymp \frac{b^{1-\beta-\theta} \ln b}{(b^{1-\beta-\theta} \ln b)^2} \\ &\asymp \frac{1}{b^{1-\beta-\theta} \ln b} = o(1). \end{aligned}$$

Completely analogously, uniformly in $(\beta, r) \in \mathbf{D}_l$, $l = 1, 2$, as $b \rightarrow \infty$

$$\sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathbf{E}_{\Pi_1}(1 - \tilde{\Psi}_l) = o(1).$$

Since the risk $\mathcal{R}(\tilde{\Psi}_l)$ of $\tilde{\Psi}_l$ is either the convex combination or the maximum of $\mathbf{E}_{\Pi_2}(\tilde{\Psi}_l)$ and $\mathbf{E}_{\Pi_1}(1 - \tilde{\Psi}_l)$, it follows that for $l = 1, 2$, uniformly in $(\beta, r) \in \mathbf{D}_l$,

$$\lim_{b \rightarrow \infty} \sup_{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{S}_{b, \beta, r}} \mathcal{R}(\tilde{\Psi}_l) = 0,$$

and the proof is complete. $\square$.

Conclusion

In this thesis, we studied the problem of binary classification in sparse normal models. In many papers in the past, this problem has been treated using a diagonal covariance matrix. In our work, we assumed a more general covariance matrix that must be block-diagonal, which makes this research more relevant to practical situations such as biomedical studies. Although, in reality, the assumption of equally sized blocks of a block-diagonal covariance matrix may be too restrictive, our model can serve as an approximation to various models suggested by real-life problems.

The model under study assumed that only a small part of the data are ‘informative’ and therefore contribute towards successful classification. The key issue was how to extract the informative data from the whole sample because using all the data would result in high classification error. This issue naturally connected the classification problem under study to a relevant feature selection problem.

In this work, we introduced various data-driven classifiers and studied their properties analytically and numerically. We found that, if one can select informative feature

blocks with a certain degree of accuracy, the classification problem at hand becomes easier. For this case, we proposed almost and completely data-driven classifiers and showed that they provide successful classification. The classification problem becomes very difficult when the selection of informative feature blocks is impossible. In this case, we improved on the idea of Donoho and Jin [8] but, unfortunately, were not able to obtain a theoretical justification that the improved idea works very well. At least, numerically, the risks of our modified classifiers, based on the idea of Donoho and Jin [8], are smaller than that of the original classifier in Donoho and Jin [8].

This subject is far from being mastered by anyone in the statistical community, but strides have been made in this thesis. We have obtained new results, such as Theorem 5 of Section 12 (see also Conjecture 1 in Section 13) that will lead to a better understanding of binary classification in high dimensions. While this may be the end of this work, it will not be the end of this research. There are still major steps that need to be taken for a better understanding of the issue of classification when the dimension is larger than the sample size. This subject will eventually be mastered and, when it is, it will truly aid physicians in the discipline of diagnostics.

References

- [1] T. W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.
- [2] D. J. Balding (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.
- [3] S. N. Bernstein, *Probability Theory*, OGIZ, Moscow–Leningrad (1946). In Russian.
- [4] C. Butucea, N. Stepanova, and A. B. Tsybakov (2017). Variable selection with Hamming loss. To appear in the *Ann. Statist.* <http://arxiv.org/abs/1512.01832>.
- [5] M. Csörgő, S. Csörgő, L. Horváth, and D. Mason (1986). Weighted empirical and quantile processes. *Ann. Probab.*, **14**, No. 3, 31–85.
- [6] X. Cui (2014). *Optimal component selection in high dimensions*. MSc Thesis, Carleton University, Ottawa.
- [7] D. Donoho and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, No. 3, 962–994.
- [8] D. Donoho and J. Jin (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. R. Soc. A*, **367**, 4449–4474.
- [9] C. R. Genovese, J. Jin, L. Wasserman, and Z. Yao (2012). A comparison of the Lasso and marginal regression. *J. Mach. Learn. Res.*, **13**, 2107–2143.
- [10] C. P. Han (1975). Some relationships between noncentral chi-squared and normal distribution, *Biometrika*, **62**, No. 1, 213–214.
- [11] Yu. I. Ingster (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Meth. Statist.*, **6**, 47–69.

- [12] Yu. I. Ingster, C. Pouet, and A. B. Tsybakov, (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A*, **367**, 4427–4448.
- [13] A. J. Izenman (2012). *Modern Multivariate Statistical Techniques*. Springer, New York.
- [14] R. A. Johnson and D. W. Wichern (2007). *Applied Multivariate Statistical Analysis Sixth Edition*, Pearson Education Inc, Upper Saddle River, NJ.
- [15] O. V. Lepski (1991). One problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, **35**, No. 3, 454–466.
- [16] C. R. Rao (1973). *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York.
- [17] G. R. Shorack and J. A. Wellner (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
- [18] M. Siotani (1971). An asymptotic expansion of the non-null distribution of Hotelling’s Generalized T_0^2 -statistic. *Ann. Math. Statist.*, **42**, No. 2, 560–571.
- [19] N. Stepanova and T. Pavlenko (2017). Goodness-of-fit tests based on sup-functionals of weighted empirical processes. To appear in the *Theory of Probability and Its Applications*. <http://arxiv.org/abs/1406.0526>.
- [20] A. Y. Zhang and H. H. Zhou (2015). Minimax rates of community detection in stochastic block models. <http://arxiv.org/abs/1507.05313>.

Appendix

R-Code for simulations done in Section 14

```
#####  
###      Thesis Code 13 (Dynamic P0) Chi-Squared      ###  
#####
```

```
library(MASS) # for mvnorm  
library(matrixcalc) #for is.positive.definite  
library(Matrix) #for nearPD  
library(clusterGeneration)  
library(magic)
```

```
Sim<-function(b,beta,r,lambda,sig,p0,theta){  
  library(MASS) # for mvnorm  
  library(matrixcalc) #for is.positive.definite  
  library(Matrix) #for nearPD  
  library(clusterGeneration)  
  library(magic)
```

```
  eps<-(b^(1-beta))  
  eps<-round(eps)  
  Eps<-eps  
  n<-b^theta  
  n<-round(n)  
  if (n%%2!=0){  
    n<-n+1  
  }  
  p<-p0*b  
  ind<-(8*r*(log(b)))/n  
  alpha<-0.2
```

```
  sigma<-matrix(0,p0,p0)  
  for(i in 1:p0){  
    for(j in 1:p0){  
      sigma[i,j]<-exp(-lambda*abs(i-j))  
    }  
  }  
}
```

```
sigma1<-diag(x=1,nrow=p,ncol=p)
```

```
##b=1000##  
x<-list(sigma,sigma,sigma,sigma,sigma,sigma,  
sigma,sigma,sigma,sigma,sigma,sigma,sigma,  
sigma,sigma,sigma,sigma,sigma,sigma,sigma)
```

$$Z < -c(x_1, x_2, \dots, x_{n-1}, x_n)$$

```
sigma_inv<-solve(sigma)
```

```
a0<-startingpoints[1]
```

$$c0=(a0+b0)/2$$
$$c_0 = (a_0 + b_0) / 2$$
$$b_0 < -c_0$$
$$c0=(a0+b0)/2$$

```

} else {

```

$$a_0 < -c_0$$
$$c_0 = (a_0 + b_0) / 2$$

}

$$\}$$

c0

$$\}$$

```
h<-function(x){dchisq(x,p0)-dchisq(x,p0,2*r*log(b)) }
```

```
start<-c(0.01,100)
```

```
t_star_m<-bisection(h, start)
```

```
k<-function(x){df(x,p0,n-p0-1)-df(x,p0,n-p0-1,2*r*log(b)) }
```

```
start<-c(0.01,100)
```

```
t_star_mf<-bisection(k, start)
```

```
f<-function(x){ (x^2)*(sum(sigma_inv))-ind }
```

```
start<-c(0,1000)
```

```
out<-bisection(f, start) #numerically solves for mu
```

```
mul<-NULL
```

```
mu<-NULL
```

```
for(i in 1:(p0*Eps)){
  mul[i]<-0
  mu[i]<-out
}
```

```
for(i in (p0*Eps+1):p){
  mul[i]<-0
  mu[i]<-0
}
```

```
mu2<-mu+mul
```

```
x1<-mvrnorm(n/2, mul, sigma1)
```

```
x2<-mvrnorm(n/2, mu2, sigma1)
```

```
mul_hat<-(2/n)*colSums(x1)
```

```
mean(mul_hat)
```

```
mu2_hat<-(2/n)*colSums(x2)
```

```
mean(mu2_hat)
```

```
mu_hat<-mu2_hat-mul_hat
```

```
mean(mu_hat)
```

```
sigma1_inv<-solve(sigma1)
```

```
delta<-t(mu)%*%sigma1_inv%*%mu
```

```
delta<-as.numeric(delta)
```

```
Siggy<-matrix(0,nrow=p0,ncol=p)
```

```
int<-1
```

```
for(i in 1:p){
  for(j in 1:p){
    if(sigma1[i,j]!=0){
      Siggy[int,i]<-sigma1[i,j]
      int=int+1
    }
    if(int==p0+1){
      int<-1
    }
  }
}
```

```

    }
  }
}

Siggy_hat<-matrix(0,nrow=p0,ncol=p)

for(k in 1:b){

  Sigmahat_k<-matrix(0,nrow=p0,ncol=p0)

  for(i in 1:(n/2)){

    x1new<-x1[i,]
    x1new<-x1new[((k*p0)-(p0-1)):(k*p0)]
    x2new<-x2[i,]
    x2new<-x2new[((k*p0)-(p0-1)):(k*p0)]
    mulnew<-mul_hat[((k*p0)-(p0-1)):(k*p0)]
    mu2new<-mu2_hat[((k*p0)-(p0-1)):(k*p0)]

    Sigmahat_k<-Sigmahat_k+((x1new-mulnew)%*%
    t((x1new-mulnew)+(x2new-mu2new)%*%t((x2new-mu2new)))/(n-2)

  }

  for(i in 1:p0){
    for(j in 1:p0){
      Siggy_hat[p0-(p0-i),(k*p0)-(p0-j)]<-Sigmahat_k[i,j]
    }
  }
}

#####
#####      Transformation to Chi and F      #####
#####

T<-NULL
Gamma<-NULL
Delta_k<-NULL

for(k in 1:b){

  for(i in 1:p0){
    for(j in 1:p0){
      Sigmahat_k[i,j]<-Siggy_hat[(p0-(p0-i)),(k*p0)-(p0-j)]
    }
  }
}

```

```

munew<-mu2_hat[((k*p0)-(p0-1)):(k*p0)]-mu1_hat[((k*p0)-(p0-1)):(k*p0)]
Sigmahat_k_inv<-solve(Sigmahat_k)

T[k]<-(n/4)*as.numeric(t(munew)%*%sigma_inv%*%munew)
Gamma[k]<-(n/4)*(as.numeric(t(munew)%*%sigma_inv%*%munew))
Delta_k[k]<-as.numeric(t(munew)%*%sigma_inv%*%munew)

}

Tf<-NULL
Gammaf<-NULL
Delta_kf<-NULL

for(k in 1:b){

  for(i in 1:p0){
    for(j in 1:p0){
      Sigmahat_k[i,j]<-Siggy_hat[(p0-(p0-i)),(k*p0)-(p0-j)]
    }
  }

  munew<-mu2_hat[((k*p0)-(p0-1)):(k*p0)]-
  mu1_hat[((k*p0)-(p0-1)):(k*p0)]
  Sigmahat_k_inv<-solve(Sigmahat_k)

  Tf[k]<-(n*(n-p0-1))/(4*(n-2)*p0)*as.numeric(t(munew)%*%
  Sigmahat_k_inv%*%munew)
  Gammaf[k]<-(n/4)*(as.numeric(t(munew)%*%
  Sigmahat_k_inv%*%munew))
  Delta_kf[k]<-as.numeric(t(munew)%*%Sigmahat_k_inv%*%munew)

}

#####
#####      Finding t* for D3      #####
#####

ordered_T<-sort(T)
ordered_Tf<-sort(Tf)

s<-NULL
for(i in 1:b){
  s[i]<-1-pchisq(T[i],p0)
}

sf<-NULL

```

```

for(i in 1:b){
  sf[i]<-1-pf(Tf[i],p0,n-p0-1)
}

ordered_s<-sort(s)
ordered_sf<-sort(sf)

g<-NULL
g2<-NULL
HC<-NULL
HC1<-NULL
KS<-NULL
for(i in 1:b){
  g[i]<-sqrt(b)*((1/b)*(i)-ordered_s[i])/
  (sqrt(ordered_s[i]*(1-ordered_s[i])))
  g2[i]<-sqrt(b)*((1/b)*(i)-ordered_s[i])/
  ((ordered_s[i]*(1-ordered_s[i]))^(1/4))
  HC[i]<-sqrt(b)*((1/b)*(i)-ordered_s[i])/
  (sqrt((ordered_s[i]*(1-ordered_s[i]))*
  (log(log(1/((ordered_s[i]*(1-ordered_s[i]))))))^(1/2))))
  HC1[i]<-sqrt(b)*((1/b)*(i)-ordered_s[i])/
  (sqrt((ordered_s[i]*(1-ordered_s[i]))*
  (log(log(1/((ordered_s[i]*(1-ordered_s[i]))))))))
  KS[i]<-sqrt(b)*((1/b)*(i)-ordered_s[i])
}

Epsilon<-1/(sqrt(log(b)))

ordered_s_new<-NULL
g_new<-NULL
g2_new<-NULL
HC_new<-NULL
HC1_new<-NULL
ind<-0
for(i in 1:b){
  if(ordered_s[i]<=alpha){ ## (1-pchisq(t_star_m,p0))
    ind<-ind+1
    ordered_s_new[ind]<-ordered_s[i]
    g_new[ind]<-g[i]
    g2_new[ind]<-g2[i]
    HC_new[ind]<-HC[i]
    HC1_new[ind]<-HC1[i]
  }
}

for(i in 1:length(g_new)){

```



```

  if (g_new[i]==max(g_new)){
    Sk_star_g<-ordered_s_new[i]
  }
  if (g2_new[i]==max(g2_new)){
    Sk_star_g2<-ordered_s_new[i]
  }
  if (HC_new[i]==max(HC_new)){
    Sk_star_HC<-ordered_s_new[i]
  }
  if (HC1_new[i]==max(HC1_new)){
    Sk_star_HC1<-ordered_s_new[i]
  }
}

Sk_star_KS<-ordered_s[which(KS==max(KS))]

gf<-NULL
g2f<-NULL
HCf<-NULL
HC1f<-NULL
KSf<-NULL
for(i in 1:b){
  gf[i]<-sqrt(b)*((1/b)*(i)-ordered_sf[i])/
  (sqrt(ordered_sf[i]*(1-ordered_sf[i])))
  g2f[i]<-sqrt(b)*((1/b)*(i)-ordered_sf[i])/
  ((ordered_sf[i]*(1-ordered_sf[i]))^(1/4))
  HCf[i]<-sqrt(b)*((1/b)*(i)-ordered_sf[i])/
  (sqrt((ordered_sf[i]*(1-ordered_sf[i]))*
  (log(log(1/((ordered_sf[i]*(1-ordered_sf[i]))))))^(1/2)))
  HC1f[i]<-sqrt(b)*((1/b)*(i)-ordered_sf[i])/
  (sqrt((ordered_sf[i]*(1-ordered_sf[i]))*
  (log(log(1/((ordered_sf[i]*(1-ordered_sf[i]))))))))
  KSf[i]<-sqrt(b)*((1/b)*(i)-ordered_sf[i])
}

Epsilon<-1/(sqrt(log(b)))

ordered_sf_new<-NULL
gf_new<-NULL
g2f_new<-NULL
HCf_new<-NULL
HC1f_new<-NULL
indf<-0
for(i in 1:b){
  if(ordered_sf[i]<=alpha){      ## (1-pf(t_star_mf,p0,n-p0-1))
    indf<-indf+1
    ordered_sf_new[indf]<-ordered_sf[i]
  }
}

```

```

    gf_new[indf]<-gf[i]
    g2f_new[indf]<-g2f[i]
    HCf_new[indf]<-HCf[i]
    HC1f_new[indf]<-HC1f[i]
  }
}

for(i in 1:length(gf_new)){
  if(gf_new[i]==max(gf_new)){
    Sk_star_gf<-ordered_sf_new[i]
  }
  if(g2f_new[i]==max(g2f_new)){
    Sk_star_g2f<-ordered_sf_new[i]
  }
  if(HCf_new[i]==max(HCf_new)){
    Sk_star_HC<-ordered_sf_new[i]
  }
  if(HC1f_new[i]==max(HC1f_new)){
    Sk_star_HC1<-ordered_sf_new[i]
  }
}

```

```

Sk_star_KSf<-ordered_sf[which(KSf==max(KSf))]

```

```

oschi<-NULL
gchi<-NULL
g2chi<-NULL
HCchi<-NULL
HC1chi<-NULL
osf<-NULL
gff<-NULL
g2ff<-NULL
HCff<-NULL
HC1ff<-NULL
for(i in 1:b){
  if(ordered_s[i]<=(1-pchisq(t_star_m,p0))){
    oschi[i]<-ordered_s[i]
    gchi[i]<-g[i]
    g2chi[i]<-g2[i]
    HCchi[i]<-HC[i]
    HC1chi[i]<-HC1[i]
  }
  if(ordered_sf[i]<=(1-pf(t_star_mf,p0,n-p0-1))){
    osf[i]<-ordered_sf[i]
    gff[i]<-gf[i]
    g2ff[i]<-g2f[i]
    HCff[i]<-HCf[i]
  }
}

```

```

    HC1ff[i]<-HC1f[i]
  }
}

plot(ordered_s,g,type="l",main="Plot of HC1(s) on (0,1)",
col="red",ylim=c(min(g,g2,HC,HC1),max(g,g2,HC,HC1)))
abline(v=Sk_star_g,col="red")
lines(ordered_s,g2,type="l",main="Plot of HC2(s) on (0,1)",
col="green")
abline(v=Sk_star_g2,col="green")
lines(ordered_s,HC,type="l",main="Plot of HC2(s) on (0,1)",
col="blue")
abline(v=Sk_star_HC,col="blue")
lines(ordered_s,HC1,type="l",main="Plot of HC3(s) on (0,1)",
col="darkgoldenrod1")
abline(v=Sk_star_HC1,col="darkgoldenrod1")
lines(ordered_s,KS,type="l",main="Plot of HC3(s) on (0,1)",
col="deeppink")
abline(v=Sk_star_KS,col="deeppink")

plot(oschi,gchi,type="l",main="Plot of HC1(s) on (0,0.1)",
col="red",ylim=c(min(gchi,g2chi,HCchi,HC1chi),
max(gchi,g2chi,HCchi,HC1chi)),xlim=c(0,alpha))
abline(v=Sk_star_g,col="red")
lines(oschi,g2chi,type="l",main="Plot of HC2(s) on (0,0.1)",
col="green")
abline(v=Sk_star_g2,col="green")
lines(oschi,HCchi,type="l",main="Plot of HC2(s) on (0,0.1)",
col="blue")
abline(v=Sk_star_HC,col="blue")
lines(oschi,HC1chi,type="l",main="Plot of HC3(s) on (0,0.1)",
col="darkgoldenrod1")
abline(v=Sk_star_HC1,col="darkgoldenrod1")

plot(ordered_sf,gf,type="l",main="Plot of HC1(s) on (0,1)",
col="red",ylim=c(min(gf,g2f,HCf,HC1f),max(gf,g2f,HCf,HC1f)))
abline(v=Sk_star_gf,col="red")
lines(ordered_sf,g2f,type="l",main="Plot of HC2(s) on (0,1)",
col="green")
abline(v=Sk_star_g2f,col="green")
lines(ordered_sf,HCf,type="l",main="Plot of HC2(s) on (0,1)",
col="blue")
abline(v=Sk_star_HCf,col="blue")
lines(ordered_sf,HC1f,type="l",main="Plot of HC3(s) on (0,1)",
col="darkgoldenrod1")
abline(v=Sk_star_HC1f,col="darkgoldenrod1")
lines(ordered_sf,KSf,type="l",main="Plot of HC3(s) on (0,1)",

```

```

col="deeppink")
abline(v=Sk_star_KSf,col="deeppink")

plot(osf,gff,type="l",main="Plot of HC1(s) on (0,0.1)",
col="red",ylim=c(min(gff,g2ff,HCff,HC1ff),
max(gff,g2ff,HCff,HC1ff)),xlim=c(0,alpha))
abline(v=Sk_star_gf,col="red")
lines(osf,g2ff,type="l",main="Plot of HC2(s) on (0,0.1)",
col="green")
abline(v=Sk_star_g2f,col="green")
lines(osf,HCff,type="l",main="Plot of HC2(s) on (0,0.1)",
col="blue")
abline(v=Sk_star_HCf,col="blue")
lines(osf,HC1ff,type="l",main="Plot of HC3(s) on (0,0.1)",
col="darkgoldenrod1")
abline(v=Sk_star_HC1f,col="darkgoldenrod1")

func1<-function(x) 1-Sk_star_g-pchisq(x,p0)
func2<-function(x) 1-Sk_star_HC-pchisq(x,p0)
func3<-function(x) 1-Sk_star_HC1-pchisq(x,p0)
func4<-function(x) 1-Sk_star_g2-pchisq(x,p0)
func5<-function(x) 1-Sk_star_KS-pchisq(x,p0)
t_star_g<-bisection(func1,c(0,max(T)))
t_star_HC<-bisection(func2,c(0,max(T)))
t_star_HC1<-bisection(func3,c(0,max(T)))
t_star_g2<-bisection(func4,c(0,max(T)))
t_star_KS<-bisection(func5,c(0,max(T)))

func1<-function(x) 1-Sk_star_gf-pf(x,p0,n-p0-1)
func2<-function(x) 1-Sk_star_HCf-pf(x,p0,n-p0-1)
func3<-function(x) 1-Sk_star_HC1f-pf(x,p0,n-p0-1)
func4<-function(x) 1-Sk_star_g2f-pf(x,p0,n-p0-1)
func5<-function(x) 1-Sk_star_KSf-pf(x,p0,n-p0-1)
t_star_gf<-bisection(func1,c(0,max(Tf)))
t_star_HCf<-bisection(func2,c(0,max(Tf)))
t_star_HC1f<-bisection(func3,c(0,max(Tf)))
t_star_g2f<-bisection(func4,c(0,max(Tf)))
t_star_KSf<-bisection(func5,c(0,max(Tf)))

#####
##### Finding t* for D2 #####
#####

m<-log(b)
m<-round(m)
tau<-(log(b))^(1/2)
gamma<-2*r*log(b)

```

```

BM<-NULL
BMf<-NULL
for (i in 1:m){
  BM[i]<-(1/(m+1))*i*((1-theta)-((1-theta)/2)) +(1-theta)/2
  BMf[i]<-(1/(m+1))*i*((1-theta)-((1-theta)/2)) +(1-theta)/2
}

omega<-matrix(0,m,b)
omegaf<-matrix(0,m,b)
for (i in 1:m){
  for (j in 1:b){
    if (ordered_T[j]>(2*BM[i]+(Eps/b))*log(b)){
      omega[i,j]<-1
    }
    if (ordered_Tf[j]>(2*BMf[i]+(Eps/b))*log(b)/p0){
      omegaf[i,j]<-1
    }
  }
}

omega_sum<-NULL
omegaf_sum<-NULL
for (i in 1:m){
  omega_sum[i]<-sum(omega[i,])
  omegaf_sum[i]<-sum(omegaf[i,])
}

v<-NULL
vf<-NULL
for (i in 1:m){
  v[i]<-b^(1-BM[i])/tau
  vf[i]<-b^(1-BMf[i])/tau
}

betaM<-1-theta
what_m<-m
betaMf<-1-theta
what_mf<-m
for (i in 1:(m-1)){
  if (betaM==(1-theta)){
    if (abs(omega_sum[i+1]-omega_sum[i])>v[i]){
      betaM<-BM[i]
      what_m<-i
    }
  }
}
if (betaMf==(1-theta)){

```

```

    if (abs(omegaf_sum[i+1]-omegaf_sum[i]) > vf[i]) {
      betaMf <- BMf[i]
      what_mf <- i
    }
  }
}

t_star_m <- (2 * ((betaM)) + (1 / (log(log(b))))) * log(b)
t_star_mf <- (2 * ((betaMf)) + (1 / (log(log(b))))) * log(b) / p0

#####
#####      Finding t* for D3      #####
#####

t_star_D1 <- (2 + log(log(b))) * log(b)
t_star_D1f <- (2 + log(log(b))) * log(b) / p0

#####
#####      Histograms of each transformation      #####
#####

x <- seq(from=0, to=max(T), by=0.1)
dist1 <- function(x) {(1-eps/b) * dchisq(x, p0) + (eps/b) *
dchisq(x, p0, 2 * r * log(b))}

hist(T, freq=FALSE, breaks=75, ylim=c(0, 0.3), xlim=c(0, max(T)),
main = "Histogram of T from Chi-Squared")
lines(x, dchisq(x, p0), lwd=2, col="red")
lines(x, dchisq(x, p0, 2 * r * log(b)), lwd=2, col="blue")
lines(x, dist1(x), lwd=2, col="forestgreen")
abline(v=t_star_g, col="yellow", lty=1, lwd=2)
abline(v=t_star_g2, col="hotpink", lty=6, lwd=2)
abline(v=t_star_HC, col="lawngreen", lty=5, lwd=2)
abline(v=t_star_HC1, col="cyan", lty=4, lwd=2)
abline(v=t_star_D1, col="orange", lty=2, lwd=2)
abline(v=t_star_m, col="magenta", lty=3, lwd=2)
abline(v=t_star_KS, col="plum1", lty=3, lwd=2)

x <- seq(from=0, to=max(Tf, t_star_D1f), by=0.1)
dist2 <- function(x) {(1-eps/b) * df(x, p0, n-p0-1) + (eps/b) *
df(x, p0, n-p0-1, 2 * r * log(b))}

hist(Tf, freq=FALSE, breaks=75, ylim=c(0, 0.8),
xlim=c(0, max(Tf, t_star_D1f)), main = "Histogram of T from F")
lines(x, df(x, p0, n-p0-1), lwd=2, type="l", col="red")
lines(x, df(x, p0, n-p0-1, 2 * r * log(b)), lwd=2, col="blue")
lines(x, dist2(x), lwd=2, col="forestgreen")

```

```

abline(v=t_star_gf,col="yellow",lty=1,lwd=2)
abline(v=t_star_g2f,col="hotpink",lty=6,lwd=2)
abline(v=t_star_HCf,col="lawngreen",lty=5,lwd=2)
abline(v=t_star_HC1f,col="cyan",lty=4,lwd=2)
abline(v=t_star_D1f,col="orange",lty=2,lwd=2)
abline(v=t_star_mf,col="magenta",lty=3,lwd=2)
abline(v=t_star_KSf,col="plum1",lty=3,lwd=2)

x<-seq(from=0,to=max(p0*Tf,p0*t_star_D1f),by=0.1)
dist3<-function(x){(1-eps/b)*df(x,p0,n-p0-1)+(eps/b)*
df(x,p0,n-p0-1,2*r*log(b))}

hist(p0*Tf,freq=FALSE,breaks=75,ylim=c(0,0.3),
xlim=c(0,max(p0*Tf,p0*t_star_D1f)),main="Histogram of T from p0F")
lines(x,df(x/p0,p0,n-p0-1)/p0,lwd=2,type="l",col="red")
lines(x,df(x/p0,p0,n-p0-1,2*r*log(b))/p0,lwd=2,col="blue")
lines(x,dist3(x/p0)/p0,lwd=2,col="forestgreen")
abline(v=p0*t_star_gf,col="yellow",lty=1,lwd=2)
abline(v=p0*t_star_g2f,col="hotpink",lty=6,lwd=2)
abline(v=p0*t_star_HCf,col="lawngreen",lty=5,lwd=2)
abline(v=p0*t_star_HC1f,col="cyan",lty=4,lwd=2)
abline(v=p0*t_star_D1f,col="orange",lty=2,lwd=2)
abline(v=p0*t_star_mf,col="magenta",lty=3,lwd=2)
abline(v=p0*t_star_KSf,col="plum1",lty=3,lwd=2)

```

M<-1000

```

x01<-mvrnorm(M,mu1,sigma1)
x02<-mvrnorm(M,mu2,sigma1)

```

```

Psi1_g<-NULL
Psi2_g<-NULL
Psi1_g2<-NULL
Psi2_g2<-NULL
Psi1_HC<-NULL
Psi2_HC<-NULL
Psi1_HC1<-NULL
Psi2_HC1<-NULL

```

```

Psi1_g_new<-NULL
Psi2_g_new<-NULL
Psi1_g2_new<-NULL
Psi2_g2_new<-NULL
Psi1_HC_new<-NULL
Psi2_HC_new<-NULL
Psi1_HC1_new<-NULL

```

Psi2_HC1_new<-NULL

Psi1_all<-NULL

Psi2_all<-NULL

Psi1_allf<-NULL

Psi2_allf<-NULL

Psi1_eps<-NULL

Psi2_eps<-NULL

Psi1_epsf<-NULL

Psi2_epsf<-NULL

Psi1_m<-NULL

Psi2_m<-NULL

Psi1_gf<-NULL

Psi2_gf<-NULL

Psi1_g2f<-NULL

Psi2_g2f<-NULL

Psi1_HCf<-NULL

Psi2_HCf<-NULL

Psi1_HC1f<-NULL

Psi2_HC1f<-NULL

Psi1_gf_new<-NULL

Psi2_gf_new<-NULL

Psi1_g2f_new<-NULL

Psi2_g2f_new<-NULL

Psi1_HCf_new<-NULL

Psi2_HCf_new<-NULL

Psi1_HC1f_new<-NULL

Psi2_HC1f_new<-NULL

Psi1_mf<-NULL

Psi2_mf<-NULL

Psi1_D1<-NULL

Psi2_D1<-NULL

Psi1_D1f<-NULL

Psi2_D1f<-NULL

Psi1_KS<-NULL

Psi2_KS<-NULL


```

Psi1_KSf<-NULL
Psi2_KSf<-NULL

Psi1_Q<-NULL
Psi2_Q<-NULL
Psi1_Of<-NULL
Psi2_Of<-NULL

for(i in 1:M){
  Psi_Sum1_g<-0
  Psi_Sum2_g<-0
  Psi_Sum1_g2<-0
  Psi_Sum2_g2<-0
  Psi_Sum1_HC<-0
  Psi_Sum2_HC<-0
  Psi_Sum1_HC1<-0
  Psi_Sum2_HC1<-0

  Psi_Sum1_g_new<-0
  Psi_Sum2_g_new<-0
  Psi_Sum1_g2_new<-0
  Psi_Sum2_g2_new<-0
  Psi_Sum1_HC_new<-0
  Psi_Sum2_HC_new<-0
  Psi_Sum1_HC1_new<-0
  Psi_Sum2_HC1_new<-0

  Psi_Sum1_all<-0
  Psi_Sum2_all<-0

  Psi_Sum1_allf<-0
  Psi_Sum2_allf<-0

  Psi_Sum1_eps<-0
  Psi_Sum2_eps<-0

  Psi_Sum1_epsf<-0
  Psi_Sum2_epsf<-0

  Psi_Sum1_m<-0
  Psi_Sum2_m<-0

  Psi_Sum1_gf<-0
  Psi_Sum2_gf<-0
  Psi_Sum1_g2f<-0
  Psi_Sum2_g2f<-0
  Psi_Sum1_HCf<-0

```

```

Psi_Sum2_HCf<-0
Psi_Sum1_HC1f<-0
Psi_Sum2_HC1f<-0

Psi_Sum1_gf_new<-0
Psi_Sum2_gf_new<-0
Psi_Sum1_g2f_new<-0
Psi_Sum2_g2f_new<-0
Psi_Sum1_HCf_new<-0
Psi_Sum2_HCf_new<-0
Psi_Sum1_HC1f_new<-0
Psi_Sum2_HC1f_new<-0

Psi_Sum1_mf<-0
Psi_Sum2_mf<-0

Psi_Sum1_D1<-0
Psi_Sum2_D1<-0

Psi_Sum1_D1f<-0
Psi_Sum2_D1f<-0

Psi_Sum1_KS<-0
Psi_Sum2_KS<-0

Psi_Sum1_KSf<-0
Psi_Sum2_KSf<-0

for (k in 1:b){

  x01new<-x01[i,((k*p0)-(p0-1)):(k*p0)]
  x02new<-x02[i,((k*p0)-(p0-1)):(k*p0)]
  mulnew<-mul_hat[(k*p0)-(p0-1):(k*p0)]
  mu2new<-mu2_hat[(k*p0)-(p0-1):(k*p0)]

  for (l in 1:p0){
    for (j in 1:p0){
      Sigmahat_k[l,j]<-Siggy_hat[(p0-(p0-1)),(k*p0)-(p0-j)]
    }
  }

  Sigmahat_k_inv<-solve(Sigmahat_k)

  if (T[k]>=t_star_g){
    Psi_Sum1_g<-Psi_Sum1_g+
      (t*(x01new-(mulnew+mu2new)/2)%%sigma_inv%%(mu2new-mulnew))
    Psi_Sum1_g<-as.numeric(Psi_Sum1_g)
  }
}

```

```

Psi_Sum2_g<-Psi_Sum2_g+
(t(x02new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_g<-as.numeric(Psi_Sum2_g)
}

if(T[k]>=t_star_g2){
Psi_Sum1_g2<-Psi_Sum1_g2+(t(x01new-
(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_g2<-as.numeric(Psi_Sum1_g2)
Psi_Sum2_g2<-Psi_Sum2_g2+(t(x02new-
(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_g2<-as.numeric(Psi_Sum2_g2)
}

if(T[k]>=t_star_HC){
Psi_Sum1_HC<-Psi_Sum1_HC+(t(x01new-
(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_HC<-as.numeric(Psi_Sum1_HC)
Psi_Sum2_HC<-Psi_Sum2_HC+(t(x02new-
(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_HC<-as.numeric(Psi_Sum2_HC)
}

if(T[k]>=t_star_HC1){
Psi_Sum1_HC1<-Psi_Sum1_HC1+
(t(x01new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_HC1<-as.numeric(Psi_Sum1_HC1)
Psi_Sum2_HC1<-Psi_Sum2_HC1+
(t(x02new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_HC1<-as.numeric(Psi_Sum2_HC1)
}

if(T[k]>=t_star_m){
Psi_Sum1_m<-Psi_Sum1_m+
(t(x01new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_m<-as.numeric(Psi_Sum1_m)
Psi_Sum2_m<-Psi_Sum2_m+
(t(x02new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_m<-as.numeric(Psi_Sum2_m)
}

if(T[k]>=t_star_D1){
Psi_Sum1_D1<-Psi_Sum1_D1+
(t(x01new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_D1<-as.numeric(Psi_Sum1_D1)
Psi_Sum2_D1<-Psi_Sum2_D1+
(t(x02new-(mulnew+mu2new)/2)%*%sigma_inv%*(mu2new-mulnew))

```

```

Psi_Sum2_D1<-as.numeric(Psi_Sum2_D1)
}

if(T[k]>=t_star_KS){
  Psi_Sum1_KS<-Psi_Sum1_KS+
  (t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
  Psi_Sum1_KS<-as.numeric(Psi_Sum1_KS)
  Psi_Sum2_KS<-Psi_Sum2_KS+
  (t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
  Psi_Sum2_KS<-as.numeric(Psi_Sum2_KS)
}

Psi_Sum1_all<-Psi_Sum1_all+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_all<-as.numeric(Psi_Sum1_all)
Psi_Sum2_all<-Psi_Sum2_all+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_all<-as.numeric(Psi_Sum2_all)

Psi_Sum1_allf<-Psi_Sum1_allf+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
Psi_Sum1_allf<-as.numeric(Psi_Sum1_allf)
Psi_Sum2_allf<-Psi_Sum2_allf+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
Psi_Sum2_allf<-as.numeric(Psi_Sum2_allf)

if(Tf[k]>=t_star_gf){
  Psi_Sum1_gf<-Psi_Sum1_gf+
  (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
  Psi_Sum1_gf<-as.numeric(Psi_Sum1_gf)
  Psi_Sum2_gf<-Psi_Sum2_gf+
  (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
  Psi_Sum2_gf<-as.numeric(Psi_Sum2_gf)
}

if(Tf[k]>=t_star_g2f){
  Psi_Sum1_g2f<-Psi_Sum1_g2f+
  (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
  Psi_Sum1_g2f<-as.numeric(Psi_Sum1_g2f)
  Psi_Sum2_g2f<-Psi_Sum2_g2f+
  (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
  Psi_Sum2_g2f<-as.numeric(Psi_Sum2_g2f)
}

if(Tf[k]>=t_star_HCf){
  Psi_Sum1_HCf<-Psi_Sum1_HCf+
  (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))

```

```

Psi_Sum1_HCf<-as.numeric(Psi_Sum1_HCf)
Psi_Sum2_HCf<-Psi_Sum2_HCf+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum2_HCf<-as.numeric(Psi_Sum2_HCf)
}

if(Tf[k]>=t_star_HC1f){
Psi_Sum1_HC1f<-Psi_Sum1_HC1f+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum1_HC1f<-as.numeric(Psi_Sum1_HC1f)
Psi_Sum2_HC1f<-Psi_Sum2_HC1f+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum2_HC1f<-as.numeric(Psi_Sum2_HC1f)
}

if(Tf[k]>=t_star_mf){
Psi_Sum1_mf<-Psi_Sum1_mf+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum1_mf<-as.numeric(Psi_Sum1_mf)
Psi_Sum2_mf<-Psi_Sum2_mf+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum2_mf<-as.numeric(Psi_Sum2_mf)
}

if(Tf[k]>=t_star_D1f){
Psi_Sum1_D1f<-Psi_Sum1_D1f+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum1_D1f<-as.numeric(Psi_Sum1_D1f)
Psi_Sum2_D1f<-Psi_Sum2_D1f+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum2_D1f<-as.numeric(Psi_Sum2_D1f)
}

if(Tf[k]>=t_star_KSf){
Psi_Sum1_KSf<-Psi_Sum1_KSf+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum1_KSf<-as.numeric(Psi_Sum1_KSf)
Psi_Sum2_KSf<-Psi_Sum2_KSf+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%(mu2new-mulnew))
Psi_Sum2_KSf<-as.numeric(Psi_Sum2_KSf)
}

if(k<=eps){

if(T[k]>=t_star_g){
Psi_Sum1_g_new<-Psi_Sum1_g_new+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%(mu2new-mulnew))

```

```

Psi_Sum1_g_new<-as.numeric(Psi_Sum1_g_new)
Psi_Sum2_g_new<-Psi_Sum2_g_new+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_g_new<-as.numeric(Psi_Sum2_g_new)
}

if(T[k]>=t_star_g2){
Psi_Sum1_g2_new<-Psi_Sum1_g2_new+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_g2_new<-as.numeric(Psi_Sum1_g2_new)
Psi_Sum2_g2_new<-Psi_Sum2_g2_new+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_g2_new<-as.numeric(Psi_Sum2_g2_new)
}

if(T[k]>=t_star_HC){
Psi_Sum1_HC_new<-Psi_Sum1_HC_new+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_HC_new<-as.numeric(Psi_Sum1_HC_new)
Psi_Sum2_HC_new<-Psi_Sum2_HC_new+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_HC_new<-as.numeric(Psi_Sum2_HC_new)
}

if(T[k]>=t_star_HC1){
Psi_Sum1_HC1_new<-Psi_Sum1_HC1_new+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_HC1_new<-as.numeric(Psi_Sum1_HC1_new)
Psi_Sum2_HC1_new<-Psi_Sum2_HC1_new+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_HC1_new<-as.numeric(Psi_Sum2_HC1_new)
}

Psi_Sum1_eps<-Psi_Sum1_eps+
(t(x01new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum1_eps<-as.numeric(Psi_Sum1_eps)
Psi_Sum2_eps<-Psi_Sum2_eps+
(t(x02new-(mulnew+mu2new)/2)*%sigma_inv%*(mu2new-mulnew))
Psi_Sum2_eps<-as.numeric(Psi_Sum2_eps)

Psi_Sum1_epsf<-Psi_Sum1_epsf+
(t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
Psi_Sum1_epsf<-as.numeric(Psi_Sum1_epsf)
Psi_Sum2_epsf<-Psi_Sum2_epsf+
(t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
Psi_Sum2_epsf<-as.numeric(Psi_Sum2_epsf)

```

```

    if (Tf[k]>=t_star_gf){
      Psi_Sum1_gf_new<-Psi_Sum1_gf_new+
      (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum1_gf_new<-as.numeric(Psi_Sum1_gf_new)
      Psi_Sum2_gf_new<-Psi_Sum2_gf_new+
      (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum2_gf_new<-as.numeric(Psi_Sum2_gf_new)
    }

    if (Tf[k]>=t_star_g2f){
      Psi_Sum1_g2f_new<-Psi_Sum1_g2f_new+
      (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum1_g2f_new<-as.numeric(Psi_Sum1_g2f_new)
      Psi_Sum2_g2f_new<-Psi_Sum2_g2f_new+
      (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum2_g2f_new<-as.numeric(Psi_Sum2_g2f_new)
    }

    if (Tf[k]>=t_star_HCf){
      Psi_Sum1_HCf_new<-Psi_Sum1_HCf_new+
      (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum1_HCf_new<-as.numeric(Psi_Sum1_HCf_new)
      Psi_Sum2_HCf_new<-Psi_Sum2_HCf_new+
      (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum2_HCf_new<-as.numeric(Psi_Sum2_HCf_new)
    }

    if (Tf[k]>=t_star_HC1f){
      Psi_Sum1_HC1f_new<-Psi_Sum1_HC1f_new+
      (t(x01new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum1_HC1f_new<-as.numeric(Psi_Sum1_HC1f_new)
      Psi_Sum2_HC1f_new<-Psi_Sum2_HC1f_new+
      (t(x02new-(mulnew+mu2new)/2)*%Sigmahat_k_inv%*(mu2new-mulnew))
      Psi_Sum2_HC1f_new<-as.numeric(Psi_Sum2_HC1f_new)
    }

  }

}

if (Psi_Sum1_g<=0){
  Psi1_g[i]<-1
}else{
  Psi1_g[i]<-0
}
if (Psi_Sum2_g<=0){
  Psi2_g[i]<-1
}

```

```

} else {
    Psi2_g[ i ] < -0
}

if ( Psi_Sum1_g2 <= 0 ) {
    Psi1_g2[ i ] < -1
} else {
    Psi1_g2[ i ] < -0
}

if ( Psi_Sum2_g2 <= 0 ) {
    Psi2_g2[ i ] < -1
} else {
    Psi2_g2[ i ] < -0
}

if ( Psi_Sum1_HC <= 0 ) {
    Psi1_HC[ i ] < -1
} else {
    Psi1_HC[ i ] < -0
}

if ( Psi_Sum2_HC <= 0 ) {
    Psi2_HC[ i ] < -1
} else {
    Psi2_HC[ i ] < -0
}

if ( Psi_Sum1_HC1 <= 0 ) {
    Psi1_HC1[ i ] < -1
} else {
    Psi1_HC1[ i ] < -0
}

if ( Psi_Sum2_HC1 <= 0 ) {
    Psi2_HC1[ i ] < -1
} else {
    Psi2_HC1[ i ] < -0
}

if ( Psi_Sum1_g_new <= 0 ) {
    Psi1_g_new[ i ] < -1
} else {
    Psi1_g_new[ i ] < -0
}

if ( Psi_Sum2_g_new <= 0 ) {
    Psi2_g_new[ i ] < -1
} else {
    Psi2_g_new[ i ] < -0
}

```



```

if (Psi_Sum1_g2_new<=0){
    Psi1_g2_new[i]<-1
}else{
    Psi1_g2_new[i]<-0
}
if (Psi_Sum2_g2_new<=0){
    Psi2_g2_new[i]<-1
}else{
    Psi2_g2_new[i]<-0
}

if (Psi_Sum1_HC_new<=0){
    Psi1_HC_new[i]<-1
}else{
    Psi1_HC_new[i]<-0
}
if (Psi_Sum2_HC_new<=0){
    Psi2_HC_new[i]<-1
}else{
    Psi2_HC_new[i]<-0
}

if (Psi_Sum1_HC1_new<=0){
    Psi1_HC1_new[i]<-1
}else{
    Psi1_HC1_new[i]<-0
}
if (Psi_Sum2_HC1_new<=0){
    Psi2_HC1_new[i]<-1
}else{
    Psi2_HC1_new[i]<-0
}

if (Psi_Sum1_all<=0){
    Psi1_all[i]<-1
}else{
    Psi1_all[i]<-0
}
if (Psi_Sum2_all<=0){
    Psi2_all[i]<-1
}else{
    Psi2_all[i]<-0
}

if (Psi_Sum1_allf<=0){
    Psi1_allf[i]<-1
}

```

```

} else {
    Psi1_allf [ i ] <- 0
}
if ( Psi_Sum2_allf <= 0 ) {
    Psi2_allf [ i ] <- 1
} else {
    Psi2_allf [ i ] <- 0
}

if ( Psi_Sum1_eps <= 0 ) {
    Psi1_eps [ i ] <- 1
} else {
    Psi1_eps [ i ] <- 0
}
if ( Psi_Sum2_eps <= 0 ) {
    Psi2_eps [ i ] <- 1
} else {
    Psi2_eps [ i ] <- 0
}

if ( Psi_Sum1_epsf <= 0 ) {
    Psi1_epsf [ i ] <- 1
} else {
    Psi1_epsf [ i ] <- 0
}
if ( Psi_Sum2_epsf <= 0 ) {
    Psi2_epsf [ i ] <- 1
} else {
    Psi2_epsf [ i ] <- 0
}

if ( Psi_Sum1_m <= 0 ) {
    Psi1_m [ i ] <- 1
} else {
    Psi1_m [ i ] <- 0
}
if ( Psi_Sum2_m <= 0 ) {
    Psi2_m [ i ] <- 1
} else {
    Psi2_m [ i ] <- 0
}
if ( Psi_Sum1_gf <= 0 ) {
    Psi1_gf [ i ] <- 1
} else {
    Psi1_gf [ i ] <- 0
}
if ( Psi_Sum2_gf <= 0 ) {

```

```

    Psi2_gf[i]<-1
  }else{
    Psi2_gf[i]<-0
  }

  if(Psi_Sum1_g2f<=0){
    Psi1_g2f[i]<-1
  }else{
    Psi1_g2f[i]<-0
  }
  if(Psi_Sum2_g2f<=0){
    Psi2_g2f[i]<-1
  }else{
    Psi2_g2f[i]<-0
  }

  if(Psi_Sum1_HCf<=0){
    Psi1_HCf[i]<-1
  }else{
    Psi1_HCf[i]<-0
  }
  if(Psi_Sum2_HCf<=0){
    Psi2_HCf[i]<-1
  }else{
    Psi2_HCf[i]<-0
  }

  if(Psi_Sum1_HC1f<=0){
    Psi1_HC1f[i]<-1
  }else{
    Psi1_HC1f[i]<-0
  }
  if(Psi_Sum2_HC1f<=0){
    Psi2_HC1f[i]<-1
  }else{
    Psi2_HC1f[i]<-0
  }

  if(Psi_Sum1_gf_new<=0){
    Psi1_gf_new[i]<-1
  }else{
    Psi1_gf_new[i]<-0
  }
  if(Psi_Sum2_gf_new<=0){
    Psi2_gf_new[i]<-1
  }else{
    Psi2_gf_new[i]<-0
  }

```

```

}

if (Psi_Sum1_g2f_new<=0){
    Psi1_g2f_new[ i ]<-1
else{
    Psi1_g2f_new[ i ]<-0
}
if (Psi_Sum2_g2f_new<=0){
    Psi2_g2f_new[ i ]<-1
else{
    Psi2_g2f_new[ i ]<-0
}
if (Psi_Sum1_HCf_new<=0){
    Psi1_HCf_new[ i ]<-1
else{
    Psi1_HCf_new[ i ]<-0
}
if (Psi_Sum2_HCf_new<=0){
    Psi2_HCf_new[ i ]<-1
else{
    Psi2_HCf_new[ i ]<-0
}
if (Psi_Sum1_HC1f_new<=0){
    Psi1_HC1f_new[ i ]<-1
else{
    Psi1_HC1f_new[ i ]<-0
}
if (Psi_Sum2_HC1f_new<=0){
    Psi2_HC1f_new[ i ]<-1
else{
    Psi2_HC1f_new[ i ]<-0
}

if (Psi_Sum1_mf<=0){
    Psi1_mf[ i ]<-1
else{
    Psi1_mf[ i ]<-0
}
if (Psi_Sum2_mf<=0){
    Psi2_mf[ i ]<-1
else{
    Psi2_mf[ i ]<-0
}
if (Psi_Sum1_D1<=0){
    Psi1_D1[ i ]<-1
else{
    Psi1_D1[ i ]<-0

```

```

}
if (Psi_Sum2_D1<=0){
  Psi2_D1[i]<-1
}else{
  Psi2_D1[i]<-0
}
if (Psi_Sum1_D1f<=0){
  Psi1_D1f[i]<-1
}else{
  Psi1_D1f[i]<-0
}
if (Psi_Sum2_D1f<=0){
  Psi2_D1f[i]<-1
}else{
  Psi2_D1f[i]<-0
}
if (Psi_Sum1_KS<=0){
  Psi1_KS[i]<-1
}else{
  Psi1_KS[i]<-0
}
if (Psi_Sum2_KS<=0){
  Psi2_KS[i]<-1
}else{
  Psi2_KS[i]<-0
}
if (Psi_Sum1_KSf<=0){
  Psi1_KSf[i]<-1
}else{
  Psi1_KSf[i]<-0
}
if (Psi_Sum2_KSf<=0){
  Psi2_KSf[i]<-1
}else{
  Psi2_KSf[i]<-0
}
}

ClassError_g<-(1/2)*(M-sum(Psi1_g))/M+(1/2)*(sum(Psi2_g))/M
ClassError_g2<-(1/2)*(M-sum(Psi1_g2))/M+(1/2)*(sum(Psi2_g2))/M
ClassError_HC<-(1/2)*(M-sum(Psi1_HC))/M+(1/2)*(sum(Psi2_HC))/M
ClassError_HC1<-(1/2)*(M-sum(Psi1_HC1))/M+(1/2)*(sum(Psi2_HC1))/M

ClassError_g_new<-(1/2)*(M-sum(Psi1_g_new))/M+(1/2)*
(sum(Psi2_g_new))/M
ClassError_g2_new<-(1/2)*(M-sum(Psi1_g2_new))/M+(1/2)*

```

```

(sum(Psi2_g2_new))/M
ClassError_HC_new<-(1/2)*(M-sum(Psi1_HC_new))/M+(1/2)*
(sum(Psi2_HC_new))/M
ClassError_HC1_new<-(1/2)*(M-sum(Psi1_HC1_new))/M+(1/2)*
(sum(Psi2_HC1_new))/M

ClassError_eps<-(1/2)*(M-sum(Psi1_eps))/M+(1/2)*(sum(Psi2_eps))/M
ClassError_all<-(1/2)*(M-sum(Psi1_all))/M+(1/2)*(sum(Psi2_all))/M

ClassError_epsf<-(1/2)*(M-sum(Psi1_epsf))/M+(1/2)*(sum(Psi2_epsf))/M
ClassError_allf<-(1/2)*(M-sum(Psi1_allf))/M+(1/2)*(sum(Psi2_allf))/M

ClassError_Theo<-pnorm((-1/2)*sqrt(delta))

ClassError_m<-(1/2)*(M-sum(Psi1_m))/M+(1/2)*(sum(Psi2_m))/M

ClassError_gf<-(1/2)*(M-sum(Psi1_gf))/M+(1/2)*(sum(Psi2_gf))/M
ClassError_g2f<-(1/2)*(M-sum(Psi1_g2f))/M+(1/2)*(sum(Psi2_g2f))/M
ClassError_HCf<-(1/2)*(M-sum(Psi1_HCf))/M+(1/2)*(sum(Psi2_HCf))/M
ClassError_HC1f<-(1/2)*(M-sum(Psi1_HC1f))/M+(1/2)*(sum(Psi2_HC1f))/M

ClassError_gf_new<-(1/2)*(M-sum(Psi1_gf_new))/M+(1/2)*
(sum(Psi2_gf_new))/M
ClassError_g2f_new<-(1/2)*(M-sum(Psi1_g2f_new))/M+(1/2)*
(sum(Psi2_g2f_new))/M
ClassError_HCf_new<-(1/2)*(M-sum(Psi1_HCf_new))/M+(1/2)*
(sum(Psi2_HCf_new))/M
ClassError_HC1f_new<-(1/2)*(M-sum(Psi1_HC1f_new))/M+(1/2)*
(sum(Psi2_HC1f_new))/M

ClassError_mf<-(1/2)*(M-sum(Psi1_mf))/M+(1/2)*(sum(Psi2_mf))/M

ClassError_D1<-(1/2)*(M-sum(Psi1_D1))/M+(1/2)*(sum(Psi2_D1))/M
ClassError_D1f<-(1/2)*(M-sum(Psi1_D1f))/M+(1/2)*(sum(Psi2_D1f))/M

ClassError_KS<-(1/2)*(M-sum(Psi1_KS))/M+(1/2)*(sum(Psi2_KS))/M
ClassError_KSf<-(1/2)*(M-sum(Psi1_KSf))/M+(1/2)*(sum(Psi2_KSf))/M

ClassError_O<-(1/2)*(M-sum(Psi1_O))/M+(1/2)*(sum(Psi2_O))/M
ClassError_Of<-(1/2)*(M-sum(Psi1_Of))/M+(1/2)*(sum(Psi2_Of))/M

#####
# Determining Informative Block Vs. Non-Informative Blocks #
#####

vibc_g<-0

```

```

vibc_g2<-0
vibc_HC<-0
vibc_HC1<-0
vibc_m<-0
vibc_D1<-0
vibc_KS<-0
vibf_g<-0
vibf_g2<-0
vibf_HC<-0
vibf_HC1<-0
vibf_m<-0
vibf_D1<-0
vibf_KS<-0
for(i in 1:eps){
  if(T[i]>=t_star_g){
    vibc_g<-vibc_g+1
  }
  if(T[i]>=t_star_g2){
    vibc_g2<-vibc_g2+1
  }
  if(T[i]>=t_star_HC){
    vibc_HC<-vibc_HC+1
  }
  if(T[i]>=t_star_HC1){
    vibc_HC1<-vibc_HC1+1
  }
  if(T[i]>=t_star_m){
    vibc_m<-vibc_m+1
  }
  if(T[i]>=t_star_D1){
    vibc_D1<-vibc_D1+1
  }
  if(T[i]>=t_star_KS){
    vibc_KS<-vibc_KS+1
  }
  if(Tf[i]>=t_star_gf){
    vibf_g<-vibf_g+1
  }
  if(Tf[i]>=t_star_g2f){
    vibf_g2<-vibf_g2+1
  }
  if(Tf[i]>=t_star_HCf){
    vibf_HC<-vibf_HC+1
  }
  if(Tf[i]>=t_star_HC1f){
    vibf_HC1<-vibf_HC1+1
  }
}

```

```

    if (Tf[i] >= t_star_mf) {
        vibf_m <- vibf_m + 1
    }
    if (Tf[i] >= t_star_D1f) {
        vibf_D1 <- vibf_D1 + 1
    }
    if (Tf[i] >= t_star_KSf) {
        vibf_KS <- vibf_KS + 1
    }
}

```

```

vc_g <- 0
vc_g2 <- 0
vc_HC <- 0
vc_HC1 <- 0
vc_m <- 0
vc_D1 <- 0
vc_KS <- 0
vf_g <- 0
vf_g2 <- 0
vf_HC <- 0
vf_HC1 <- 0
vf_m <- 0
vf_D1 <- 0
vf_KS <- 0
for (i in 1:b) {
    if (T[i] >= t_star_g) {
        vc_g <- vc_g + 1
    }
    if (T[i] >= t_star_g2) {
        vc_g2 <- vc_g2 + 1
    }
    if (T[i] >= t_star_HC) {
        vc_HC <- vc_HC + 1
    }
    if (T[i] >= t_star_HC1) {
        vc_HC1 <- vc_HC1 + 1
    }
    if (T[i] >= t_star_m) {
        vc_m <- vc_m + 1
    }
    if (T[i] >= t_star_D1) {
        vc_D1 <- vc_D1 + 1
    }
    if (T[i] >= t_star_KS) {
        vc_KS <- vc_KS + 1
    }
}

```



```

    if (Tf[i] >= t_star_gf) {
      vf_g <- vf_g + 1
    }
    if (Tf[i] >= t_star_g2f) {
      vf_g2 <- vf_g2 + 1
    }
    if (Tf[i] >= t_star_HCf) {
      vf_HC <- vf_HC + 1
    }
    if (Tf[i] >= t_star_HC1f) {
      vf_HC1 <- vf_HC1 + 1
    }
    if (Tf[i] >= t_star_mf) {
      vf_m <- vf_m + 1
    }
    if (Tf[i] >= t_star_D1f) {
      vf_D1 <- vf_D1 + 1
    }
    if (Tf[i] >= t_star_KSf) {
      vf_KS <- vf_KS + 1
    }
  }
}

```

```

vubc_g <- vc_g - vibc_g
vubc_g2 <- vc_g2 - vibc_g2
vubc_HC <- vc_HC - vibc_HC
vubc_HC1 <- vc_HC1 - vibc_HC1
vubc_m <- vc_m - vibc_m
vubc_D1 <- vc_D1 - vibc_D1
vubc_KS <- vc_KS - vibc_KS
vubf_g <- vf_g - vibf_g
vubf_g2 <- vf_g2 - vibf_g2
vubf_HC <- vf_HC - vibf_HC
vubf_HC1 <- vf_HC1 - vibf_HC1
vubf_m <- vf_m - vibf_m
vubf_D1 <- vf_D1 - vibf_D1
vubf_KS <- vf_KS - vibf_KS

```

```

#####
#####      Creating the vector matrix      #####
#####

```

```

Output <- NULL
Output[1] <- p0
Output[2] <- beta
Output[3] <- r
Output[4] <- b

```

```

Output [5]<-M
Output [6]<-eps
Output [7]<-out
Output [8]<-delta

Output [9]<-Sk_star_g
Output [10]<-Sk_star_g2
Output [11]<-Sk_star_HC
Output [12]<-Sk_star_HC1
Output [13]<-Sk_star_gf
Output [14]<-Sk_star_g2f
Output [15]<-Sk_star_HCf
Output [16]<-Sk_star_HC1f

Output [17]<-t_star_g
Output [18]<-t_star_g2
Output [19]<-t_star_HC
Output [20]<-t_star_HC1
Output [21]<-t_star_gf
Output [22]<-t_star_g2f
Output [23]<-t_star_HCf
Output [24]<-t_star_HC1f
Output [25]<-t_star_m
Output [26]<-t_star_mf
Output [27]<-t_star_D1
Output [28]<-t_star_D1f

Output [29]<-ClassError_Theo
Output [30]<-ClassError_eps
Output [31]<-ClassError_all
Output [32]<-ClassError_g
Output [33]<-ClassError_g_new
Output [34]<-ClassError_g2
Output [35]<-ClassError_g2_new
Output [36]<-ClassError_HC
Output [37]<-ClassError_HC_new
Output [38]<-ClassError_HC1
Output [39]<-ClassError_HC1_new
Output [40]<-ClassError_gf
Output [41]<-ClassError_gf_new
Output [42]<-ClassError_g2f
Output [43]<-ClassError_g2f_new
Output [44]<-ClassError_HCf
Output [45]<-ClassError_HCf_new
Output [46]<-ClassError_HC1f
Output [47]<-ClassError_HC1f_new
Output [48]<-ClassError_m

```

```
Output [49]<-ClassError _mf
Output [50]<-ClassError _D1
Output [51]<-ClassError _D1f
```

```
Output [52]<-m
Output [53]<-what _m
Output [54]<-what _mf
```

```
Output [55]<-vc _g
Output [56]<-vibc _g
Output [57]<-vubc _g
Output [58]<-vf _g
Output [59]<-vibf _g
Output [60]<-vubf _g
Output [61]<-vc _g2
Output [62]<-vibc _g2
Output [63]<-vubc _g2
Output [64]<-vf _g2
Output [65]<-vibf _g2
Output [66]<-vubf _g2
Output [67]<-vc _HC
Output [68]<-vibc _HC
Output [69]<-vubc _HC
Output [70]<-vf _HC
Output [71]<-vibf _HC
Output [72]<-vubf _HC
Output [73]<-vc _HC1
Output [74]<-vibc _HC1
Output [75]<-vubc _HC1
Output [76]<-vf _HC1
Output [77]<-vibf _HC1
Output [78]<-vubf _HC1
```

```
Output [79]<-vc _m
Output [80]<-vibc _m
Output [81]<-vubc _m
Output [82]<-vf _m
Output [83]<-vibf _m
Output [84]<-vubf _m
Output [85]<-vc _D1
Output [86]<-vibc _D1
Output [87]<-vubc _D1
Output [88]<-vf _D1
Output [89]<-vibf _D1
Output [90]<-vubf _D1
```

```
Output [91]<-ClassError _O
```

```

Output[92]<-ClassError_Of

Output[93]<-ClassError_KS
Output[94]<-ClassError_KSf

Output[95]<-vc_KS
Output[96]<-vibc_KS
Output[97]<-vubc_KS

Output[98]<-vf_KS
Output[99]<-vibf_KS
Output[100]<-vubf_KS

Output[101]<-ClassError_epsf
Output[102]<-ClassError_allf

return(Output)

}

b<-1000
beta<-0.375
r<-0.06
sig<-0
lambda<-1
p0<-3
theta<-0.5

n<-100
Outcome<-matrix(0,n,102)

for(i in 1:n){
  Outcome[i,]<-Sim(b,beta,r,lambda,sig,p0,theta)
}
}

```