Electronic Journal of Statistics
Vol. 16 (2022) 1952–1996
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1998

Adaptive threshold-based classification of sparse high-dimensional data*

Tatjana Pavlenko

Department of Statistics, Uppsala University,
Box 513, 751 20 Uppsala, Sweden
e-mail: tatjana.pavlenko@statistik.uu.se

Natalia Stepanova and Lee Thompson

School of Mathematics and Statistics, Carleton University,
1125 Colonel By Drive, Ottawa,
Ontario K1S 5B6, Canada
e-mail: nstep@math.carleton.ca; lee.thompson.cu@gmail.com

Abstract: We revisit the problem of designing an efficient binary classifier in a challenging high-dimensional framework. The model under study assumes some local dependence structure among feature variables represented by a block-diagonal covariance matrix with a growing number of blocks of an arbitrary, but fixed size. The blocks correspond to non-overlapping independent groups of strongly correlated features. To assess the relevance of a particular block in predicting the response, we introduce a measure of "signal strength" pertaining to each feature block. This measure is then used to specify a sparse model of our interest. We further propose a threshold-based feature selector which operates as a screen-and-clean scheme integrated into a linear classifier: the data is subject to screening and hard threshold cleaning to filter out the blocks that contain no signals. Asymptotic properties of the proposed classifiers are studied when the sample size n depends on the number of feature blocks b, and the sample size goes to infinity with b at a slower rate than b. The new classifiers, which are fully adaptive to unknown parameters of the model, are shown to perform asymptotically optimally in a large part of the classification region. The numerical study confirms good analytical properties of the new classifiers that compare favorably to the existing threshold-based procedure used in a similar context.

MSC2020 subject classifications: Primary 62H30, 62H12; secondary 62E20.
Keywords and phrases: High-dimensional data, sparse vectors, adaptive threshold-based classification, asymptotically optimal classifier.

Received August 2021.

Contents

1 Introduction 1953

arXiv: 2010.00000

*T. Pavlenko was supported in part by a grant AI4Research from Uppsala University. N. Stepanova was supported by an NSERC grant. L. Thompson was supported in part by an NSERC grant.

2 Model and problem 1955
2.1 Asymptotic regime and sparsity assumption 1957
2.2 Classification regions 1959
3 Classification when $\mathbf{\Sigma}$ is known 1962
3.1 Some useful statistics 1962
3.2 Classification rule in the region $\mathcal{D}_1(\theta)$ 1964
3.3 Classification rule in the region $\mathcal{D}_2(\theta)$ 1968
4 Classification when $\mathbf{\Sigma}$ is unknown 1971
4.1 Classification rule in the region $\mathcal{D}_1^0(\theta)$ 1971
4.2 Classification rule in the region $\mathcal{D}_2^0(\theta)$ 1976
5 Numerical study 1977
6 Concluding remarks 1978
7 Proof of Lemmas 1979
Acknowledgment 1995
References 1995

1. Introduction

Statistical methodology for high-dimensional data is a rapidly growing area where inferential and algorithmic procedures for models with the number of features exceeding the number of observations are of great interest. High-dimensional statistical problems emerge in a variety of applied fields such as genomics and proteomics, cosmology, information technology, finance and banking. Classification is one of the key techniques of high-dimensional statistics where the goal is to predict the categorical class labels of new instances based on past observations.

Despite the abundance of off-the-shelf classifiers with excellent performance in the classical large-sample scenario (examples include support vector machines, AdaBoost, CART, and Artificial Neural Networks classifiers), a straightforward extension of these procedures to high-dimensional settings encounters serious challenges for the following reasons. First, these classifiers fail to explore the sparsity patterns of the high-dimensional data. With a variety of modern experimental techniques that make it possible to automatically measure a high number of features on each subject, the number of individually relevant features, or groups (blocks) of such features, is often a small part of the entire set and is hidden in that set. Incorporating too many noise feature variables with little or no relevance to the classification problem at hand can severely deteriorate the performance accuracy. Second, many classification problems in high dimensions stem from applications where identifying useful features, or groups of features that are *jointly* informative for the class label, is of primary importance. Examples of applications include, among others, the problems of cancer classification with genomics data and disease classification with medical imaging data, where the goal is to design a parsimonious classifier that would not only control the total number of features in the model without noticeable loss of quality but also allow for effective training procedures and good interpretation.

Such type of applications may require the use of feature selection techniques that would effectively operate in high-dimensional settings under various sparsity and weakness assumptions.

These thoughts have motivated us to look at the classification problem in a sparse setup, where only a small fraction of a large number of feature blocks (which are unknown to us) are "useful", and each useful block of feature variables contributes weakly to distinguishing between classes. Aiming at modelling the phenomena of growing dimensions, we use the asymptotic framework that operates over a sequence of classification problems with increasingly many feature blocks and relatively fewer observations.

In general, the classical theory of supervised classification is not designed to work in a sparse framework. Therefore, over the last decades, substantial efforts have been made to develop appropriate alternatives for standard classification procedures, such as linear and quadratic classifiers (see, for example, Ahmad and Pavlenko [1], Aoshima and Yata [2], Chan and Hall [6], Fan et al. [11], Ingster et al. [16]). Several effective classifiers that are suitable for situations where the class-covariance matrices are diagonal have been recently proposed and studied (see, for example, Ingster et al. [16] and Donoho and Jin [10]). Some of these procedures suggest, prior to the classification step, a feature selection step by thresholding. The most recent classification studies pertaining to sparse models have shown that, even under the relatively strong assumption of independence of feature variables, many statistical challenges are yet preserved.

In this paper, we examine a new sparse block-diagonal model reflecting the situation where only a small fraction of feature blocks are useful for classification. Statistical properties of the proposed classifiers depend crucially on the accuracy of a cleaning step that identifies relevant feature blocks. Cleaning is done by means of hard thresholding, with carefully chosen data-driven thresholds, that filter out the blocks containing no signals. The choice of threshold depends on the level of signal separation strength: the weaker the signal, the harder the problem of removing useless feature blocks from the subsequent classification analysis. Where possible, we adopt our newly proposed variable selection techniques to set up an appropriate threshold that would retain all useful feature blocks and perhaps a few useless ones. When the signal strength of useful blocks are too weak to allow feature selection, we propose to use hard thresholding, which is obtained by employing weighted Kolmogorov-Smirnov test statistics with suitably chosen weight functions. These statistics are known to distinguish between the pure noise and the sparse mixtures of noise and signal.

By construction, the proposed classifiers contain a random number of terms, representing classification functions for useful feature blocks, which makes the study of their efficiency properties highly nontrivial. In a large part of the classification region, the proposed classifiers are shown to have the maximum classification error and the Bayes classification error tending to zero as the number of feature blocks increases; for the rest of the classification region, numerical results with simulated data are reported.

In Section 2 we introduce a high-dimensional model of our interest and indicate the fundamental limits of sparse classification. In Section 3 we study the

classification problem at hand when the covariance matrix $\mathbf{\Sigma}$ of the data is known. The more difficult case of unknown $\mathbf{\Sigma}$ is treated in Section 4. Results of the numerical study are summarized in Section 5. Concluding remarks are given in Section 6. Proofs of Lemmas 1–3, which are essential ingredients for the proofs of Theorems 1 and 2, the main results of this work, are deferred to Section 7.

Throughout the paper, the symbol $\chi^2_\nu(\lambda)$ is used for a chi-square random variable with ν degrees of freedom and noncentrality parameter λ. The symbol $F_{\nu_1,\nu_2}(\lambda)$ is used for an F distributed random variable with ν_1 numerator and ν_2 denominator degrees of freedom and noncentrality parameter λ. Φ denotes the cumulative distribution function (cdf) of a normal $N(0,1)$ distribution. The notation $A_b \sim B_b$ means that $\limsup_{b\to\infty} A_b/B_b = 1$. We write $A_b = o(B_b)$ and $A_b = O(B_b)$ for $b \to \infty$ when $\limsup_{b\to\infty} |A_b/B_b| = 0$ and $0 < \liminf_{b\to\infty} |A_b/B_b| \leqslant \limsup_{b\to\infty} |A_b/B_b| < \infty$, respectively. We use the symbol $\log a$ for the natural (base e) logarithm of the number a. For an event A, $\mathbb{I}(A)$ is the indicator of A. We denote by $[\![n]\!]$ the set $\{1,\dots,n\}$ for some $n \in \mathbb{N}$. The Euclidean norm of a vector $\mathbf{x} \in \mathbb{R}^k$, $k \geqslant 1$, is denoted by $\|\mathbf{x}\|$. The stochastic symbols $o_{\mathbf{P}_{\Pi_l}}(1)$ and $O_{\mathbf{P}_{\Pi_l}}(1)$ are short for a sequence of random variables that converge to zero in probability and for a sequence that is bounded in probability, respectively, reflecting the fact that the sequence of random variables involves an observation generated by distribution Π_l, $l \in [\![2]\!]$.

2. Model and problem

Let $\mathbf{X}^{(1)} = (\mathbf{X}_j^{(1)})_{j\in[\![n]\!]}$ and $\mathbf{X}^{(2)} = (\mathbf{X}_j^{(2)})_{j\in[\![n]\!]}$ be random samples drawn from the populations $\Pi_1 \equiv N_p(\mathbf{0},\mathbf{\Sigma})$ and $\Pi_2 \equiv N_p(\boldsymbol{\mu},\mathbf{\Sigma})$, respectively, where $\mathbf{X}_j^{(l)} = (X_{1j}^{(l)},\dots,X_{pj}^{(l)})^\top$ for $j \in [\![n]\!]$ and $l \in [\![2]\!]$. The mean vector $\boldsymbol{\mu} \neq \mathbf{0}$ and the common covariance matrix $\mathbf{\Sigma} = \mathbf{Cov}(\mathbf{X}_j^{(l)})$, $l \in [\![2]\!]$, are generally unknown. Assume further we observe a random vector $\mathbf{X}_0 \in \mathbb{R}^p$, which is independent of $\mathbf{X}^{(l)}$, $l \in [\![2]\!]$, and the distribution of $\mathbf{X}_0$ is known to be either Π_1 (the pure noise) or Π_2 (the signal). The goal is to design a classifier $\psi = \psi\left(\mathbf{X}_0; \mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right)$ that would assign $\mathbf{X}_0$ to either Π_1 or Π_2 and would have small classification error when the dimension p is much larger than the sample size n. The problem of allocating $\mathbf{X}_0$ to either Π_1 or Π_2 is difficult only when Π_1 and Π_2 are "close" to each other. A particular type of closeness for large p is described by the sparsity assumption, which is stated rigorously in Section 2.1 below. Under this assumption, the data is grouped in a large number of blocks, and only a small fraction of the blocks are relevant for classification.

Let $\mathbf{E}_{\Pi_i}$ denote the expectation with respect to the joint distribution of $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}_0$ when $\mathbf{X}_0 \sim \Pi_i$ for $i \in [\![2]\!]$. In the present situation of equally-sized random samples, it is natural to measure the accuracy of ψ by the *Bayes risk* $\pi\mathbf{E}_{\Pi_2}(\psi) + (1-\pi)\mathbf{E}_{\Pi_1}(1-\psi)$ with $\pi = 1/2$, that is, by

$$\mathcal{R}_B(\psi) = (1/2)\mathbf{E}_{\Pi_2}(\psi) + (1/2)\mathbf{E}_{\Pi_1}(1-\psi), \tag{2.1}$$

and also by the *maximum risk*

$$\mathcal{R}_M(\psi) = \max\left(\mathbf{E}_{\Pi_2}(\psi), \mathbf{E}_{\Pi_1}(1-\psi)\right). \tag{2.2}$$

Here, $\mathbf{E}_{\Pi_2}(\psi)$ is the probability of misclassifying $\mathbf{X}_0$ as Π_1 when $\mathbf{X}_0 \in \Pi_2$. Likewise, $\mathbf{E}_{\Pi_1}(1-\psi)$ is the probability of misclassifying $\mathbf{X}_0$ as Π_2 when $\mathbf{X}_0 \in \Pi_1$. In what follows, $\mathcal{R}(\psi)$ will be either the Bayes risk $\mathcal{R}_B(\psi)$ or the maximum risk $\mathcal{R}_M(\psi)$.

Assume that $\mathbf{\Sigma}$ is a block-diagonal matrix of the form $\mathbf{\Sigma} = \mathrm{Diag}\big(\mathbf{\Sigma}_{[1]}, \ldots, \mathbf{\Sigma}_{[b]}\big)$ with each block $\mathbf{\Sigma}_{[k]}$ being symmetric and positive definite. Then, the new observation and each element of the training samples can be split into b feature blocks: for $j \in [\![n]\!]$, $l \in [\![2]\!]$

$$\mathbf{X}_0 = \left(\mathbf{X}_{0,[1]}^\top, \ldots, \mathbf{X}_{0,[b]}^\top\right)^\top, \quad \mathbf{X}_j^{(l)} = \left((\mathbf{X}_{j,[1]}^{(l)})^\top, \ldots, (\mathbf{X}_{j,[b]}^{(l)})^\top\right)^\top.$$

For $k \in [\![b]\!]$ and $b = 2, 3, \ldots$, we define $\hat{\boldsymbol{\mu}}_{[k]} = \hat{\boldsymbol{\mu}}_{[k],b}$ and $\widehat{\mathbf{\Sigma}}_{[k]} = \widehat{\mathbf{\Sigma}}_{[k],b}$ by

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

$$\widehat{\mathbf{\Sigma}}_{[k]} = \frac{1}{2n-1}\left\{\sum_{j=1}^{n}\left(\mathbf{X}_{j,[k]}^{(1)}\right)\left(\mathbf{X}_{j,[k]}^{(1)}\right)^\top + \sum_{j=1}^{n}\left(\mathbf{X}_{j,[k]}^{(2)} - \hat{\boldsymbol{\mu}}_{[k]}\right)\left(\mathbf{X}_{j,[k]}^{(2)} - \hat{\boldsymbol{\mu}}_{[k]}\right)^\top\right\},$$

and take $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_{[1]}^\top, \ldots, \hat{\boldsymbol{\mu}}_{[b]}^\top)^\top$ as an estimator of $\boldsymbol{\mu} = (\boldsymbol{\mu}_{[1]}^\top, \ldots, \boldsymbol{\mu}_{[b]}^\top)^\top$ and $\widehat{\mathbf{\Sigma}} = \mathrm{Diag}(\widehat{\mathbf{\Sigma}}_{[1]}, \ldots, \widehat{\mathbf{\Sigma}}_{[b]})$ as an estimator of $\mathbf{\Sigma} = \mathrm{Diag}(\mathbf{\Sigma}_{[1]}, \ldots, \mathbf{\Sigma}_{[b]})$.

In the case of known $\mathbf{\Sigma}$, we propose to use the classifier $\tilde{\psi}_b = \tilde{\psi}_b(\mathbf{X}_0; \mathbf{X}^{(2)})$ given by

$$\tilde{\psi}_b = \mathbb{I}\left\{\sum_{k=1\,:\,\tilde{\omega}_k=1}^{b} (\mathbf{X}_{0,[k]} - \hat{\boldsymbol{\mu}}_{[k]}/2)^\top \mathbf{\Sigma}_{[k]}^{-1}\hat{\boldsymbol{\mu}}_{[k]} \leqslant 0\right\}, \tag{2.3}$$

where $\tilde{\omega}_k$ is one if the kth feature block of the data is "useful" and zero otherwise. The new observation $\mathbf{X}_0$ is allocated to Π_1 when $\tilde{\psi}_b(\mathbf{X}_0) = 1$ and to Π_2 otherwise. As seen from Theorem 1 in Section 3.2, the risk $\mathcal{R}(\tilde{\psi}_b)$ of $\tilde{\psi}_b$ with suitably chosen $\tilde{\omega}_k$, $k \in [\![b]\!]$, tends to zero when b tends to infinity in a large part of the classification region. Similarly, in the case of unknown $\mathbf{\Sigma}$, we may consider the classifier $\hat{\psi}_b = \hat{\psi}_b(\mathbf{X}_0; \mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ defined as

$$\hat{\psi}_b = \mathbb{I}\left\{\sum_{k=1\,:\,\hat{\omega}_k=1}^{b} (\mathbf{X}_{0,[k]} - \hat{\boldsymbol{\mu}}_{[k]}/2)^\top \widehat{\mathbf{\Sigma}}_{[k]}^{-1}\hat{\boldsymbol{\mu}}_{[k]} \leqslant 0\right\}, \tag{2.4}$$

where $\hat{\omega}_k$ is one if the kth feature block of the data is "useful" and zero otherwise, which allocates $\mathbf{X}_0$ to to Π_1 when $\hat{\psi}_b(\mathbf{X}_0) = 1$ and to Π_2 otherwise. As follows from Theorem 2 in Section 4.1, the risk $\mathcal{R}(\hat{\psi}_b)$ of $\hat{\psi}_b$ with suitably chosen $\hat{\omega}_k$, $k \in [\![b]\!]$, converges to zero as b tends to infinity in a large part of the classification

region. The behavior of $\tilde{\psi}_b$ and $\hat{\psi}_b$ in the remaining part of the classification region, where the selection of useful feature blocks is impossible, is examined in Sections 3.3 and 4.2, and the related numerical results are presented in Section 5. The random functions $\tilde{\omega}_k$ and $\hat{\omega}_k$ are some good estimators of $\omega_k = \mathbb{I}(\Delta_{k,b}^2 \neq 0), k \in [\![b]\!]$, that attempt to remove most of the useless blocks of the data for which $\omega_k = 0$ from further consideration. Due to the technical issues presented by the case of unknown $\mathbf{\Sigma}$, the cases of known and unknown $\mathbf{\Sigma}$ will be treated separately.

2.1. Asymptotic regime and sparsity assumption

We shall design a classifier $\psi = \psi\left(\mathbf{X}_0; \mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right)$ in a high-dimensional framework when (i) the sample size n and the dimension p go to infinity together in such a way that $n = n_p \to \infty$ and $n = o(p)$ as $p \to \infty$, (ii) the covariance matrix $\mathbf{\Sigma}$ is a sparse block-diagonal matrix of the form $\mathbf{\Sigma} = \mathrm{Diag}\left(\mathbf{\Sigma}_{[1]}, \ldots, \mathbf{\Sigma}_{[b]}\right)$, where each block $\mathbf{\Sigma}_{[k]}$ is symmetric and positive definite, and (iii) feature variables that are deemed useful for classification appear in groups (or blocks), according to the structure of $\mathbf{\Sigma}$; the useful feature blocks are *rare* and each block contributes *weakly* to the classification decision.

We first treat the case of equally-sized $p_0 \times p_0$ blocks so that $bp_0 = p$, and then comment on the case of unequally-sized blocks. In modern settings, it is often the case that the dimension p exceeds the number of observations n. In this work, we consider a sequence of classification problems in which p_0 ($p_0 < n$) is a fixed known integer, the number of blocks b is the driving parameter, and n relates to b through

$$n = b^{\theta}(1 + o(1)) \quad \text{as} \quad b \to \infty \tag{2.5}$$

for some known $\theta \in (0, 1)$, implying $n = o(p)$ as $p \to \infty$. (Below, we may think that $n = [b^{\theta}]$.) This assumption yields $\log n \sim \theta \log p$ as $p \to \infty$, which corresponds to scenario (C) in Ingster et al. [16] and refers to as the *regular growth* of dimensionality. This asymptotic approach is also similar to the triangular array setup studied in Greenshtein and Ritov [13]. The number of blocks b of $\mathbf{\Sigma}$ is assumed to be al least 2 because the parametrization used for the model of our interest requires from $\log b$ to be nonzero (see relation (2.8) below).

In an ideal setup, when $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are known, the optimal (using the Bayes risk with equal prior probabilities) classifier $\psi_0 = \psi_0(\mathbf{X}_0)$, which is obtained by employing the likelihood ratio approach, has the risk

$$\mathcal{R}_B(\psi_0) = \Phi\left(-\Delta/2\right),$$

where $\Delta^2 = \boldsymbol{\mu}^{\top}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ is the squared Mahalanobis distance between Π_1 and Π_2. For the block-diagonal matrix $\mathbf{\Sigma} = \mathrm{Diag}\left(\mathbf{\Sigma}_{[1]}, \ldots, \mathbf{\Sigma}_{[b]}\right)$, the squared Mahalanobis distance Δ^2 depends on b and can be expressed as

$$\Delta^2 = \sum_{k=1}^{b} \Delta_{k,b}^2, \quad \Delta_{k,b}^2 = \boldsymbol{\mu}_{[k]}^{\top}\mathbf{\Sigma}_{[k]}^{-1}\boldsymbol{\mu}_{[k]}, \quad k \in [\![b]\!], \tag{2.6}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_{[1]}^\top, \ldots, \boldsymbol{\mu}_{[b]}^\top)^\top$. The quantity $\Delta_{k,b}^2$ is the *signal strength* of the kth block of the data; it measures the contribution of the kth block towards the *total strength of separation* Δ^2 between the populations Π_1 and Π_2. "Large" values of $\Delta_{k,b}^2$ suggest that the kth block is *useful* for classification, and therefore the data $(\mathbf{X}_{j,[k]}^{(l)})_{j\in[\![n]\!]}$, $l \in [\![2]\!]$, should be used for constructing a suitable classification rule; at the same time, "small" values of $\Delta_{k,b}^2$ mean that the kth block of the data is useless for classification and should be removed from further consideration. In this work, we demonstrate how accurate classification can be achieved by means of a classifier that includes an effective screen-and-clean threshold-based feature selector as its integrated part.

To set up a sparse model of our interest, we take two numbers s and a such that $s \in [\![b]\!]$ and $a > 0$, and consider the set of vectors $\mathbf{v} = (v_k)_{k\in[\![b]\!]}$ given by

$$\Gamma_b(s,a) = \{\mathbf{v} \in \mathbb{R}^b : \text{ there exists a set } S \subset [\![b]\!] \text{ with } s \text{ elements such that } v_k \geqslant a \text{ for all } k \in S, \text{ and } v_k = 0 \text{ for all } k \notin S\}. \tag{2.7}$$

The statistical model that consists of observing two independent random samples $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of size n from the respective p-dimensional populations Π_1 and Π_2, where $p = p_0 b$, is said to have an (s,a)-*sparse block-diagonal structure* if a vector $(n\Delta_{k,b}^2)_{k\in[\![b]\!]}$, where $\Delta_{k,b}^2$ is defined in (2.6), belongs to the set $\Gamma_b(s,a)$.

In what follows, both parameters s and a will depend on the driving parameter b. Namely, we assume that the parameter s satisfies as $b \to \infty$

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

implying $s = o(b)$. This type of parametrization for s is quite common in the literature on high-dimensional statistical inference. We speak of β as the *sparsity* parameter. The parameter a cannot be too small (see, for example, Remark 1 in Ingster et al. [16]). A suitable range for a that makes the classification problem at hand interesting is

$$a = a_b = 2r \log b \quad \text{for some} \quad 0 < r < 4, \tag{2.8}$$

that is, the parameter a is only *moderately* large. Indeed, in this case, the squared Mahalanobis distance for the kth block satisfies

$$\Delta_{k,b}^2 = O\left(b^{-\theta} \log b\right) = o(1), \quad b \to \infty. \tag{2.9}$$

This brings us to a nontrivial problem of classification which is closely related to the classification problem for the diagonal matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{p\times p}$, as studied in Ingster et al. [16] and Donoho and Jin [10].

The restriction on the range of r in (2.8) being the interval $(0,4)$ is due to a related feature selection problem; the assumption of $r \geqslant 4$, which corresponds to relation (2.3) in Ingster et al. [16], makes selecting useful blocks obvious and hence the problem of classifying $\mathbf{X}_0$ easy.

We shall now introduce the collection of parameters $\boldsymbol{\mu} = \boldsymbol{\mu}_{p\times 1}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{p\times p}$ of our interest. For $0 < \beta < 1$ and $0 < r < 4$, define the set $\mathbf{M}_{b,\beta,r}$ as follows:

$$\mathbf{M}_{b,\beta,r} = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} = (\boldsymbol{\mu}_{[1]}^\top, \ldots, \boldsymbol{\mu}_{[b]}^\top)^\top \neq \mathbf{0},\ \boldsymbol{\Sigma} = \mathrm{Diag}(\boldsymbol{\Sigma}_{[1]}, \ldots, \boldsymbol{\Sigma}_{[b]}) \text{ is}$$

positive definite and symmetric, and vector $(n\Delta_{k,b})_{k\in[\![b]\!]}$ belongs to the set $\Gamma_b([b^{1-\beta}], 2r\log b)\}$,

where n relates to b through (2.5) and $\Gamma_b(s,a)$ is as in (2.7). Then, our *sparsity assumption* on the model, which is characterized by numbers $\beta \in (0,1)$ and $r \in (0,4)$, says that the pair of parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an element of $\mathbf{M}_{b,\beta,r}$. We say that the kth block of the data is *useful* (for classification) if $n\Delta_{k,b}^2 \geqslant 2r\log b$, and it is *useless* if $n\Delta_{k,b}^2 = 0$. The value $b^{-\beta}$ may be viewed as the 'probability' of occurrence of useful feature blocks among the b blocks available. Thus, very few blocks of features are useful for classification, and the information carried by each of these blocks contributes weakly to the classification decision. This type of sparse model is sometimes referred to in the literature as the *rare* and *weak feature model* (see, for example, Donoho and Jin [10]).

The classifier proposed in (2.3) depends on $\widehat{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}^{-1}$ only through their block-wise products; a similar comment applies to the classifier in (2.4). Therefore, the idea of imposing the sparsity assumption directly on the signal separation strength vector $(n\Delta_{k,b})_{k\in[\![b]\!]}$ is a natural one. This type of sparsity assumption is somewhat weaker and more flexible as compared to some commonly used assumptions that require the sparsity of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$) separately. For instance, Shao et al. [21] proposed some thresholding procedures in which $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are first estimated separately and then plugged into classification rules. In general, in the context of classification, the idea of imposing sparsity assumptions separately on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$) may be inappropriate as there are cases where neither $\boldsymbol{\mu}$ nor $\boldsymbol{\Sigma}^{-1}$ are sparse but $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is (see, for example, Cai and Liu [5]).

Below, we consider the regions inside the *parameter space* $\{(\beta,r)\in\mathbb{R}^2 : 0 < \beta < 1,\ 0 < r < 4\}$ where *successful classification is possible* in the sense that

$$\lim_{b\to\infty}\inf_{\psi}\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathcal{R}(\psi) = 0, \tag{2.10}$$

where $\mathcal{R}(\psi)$ is either $\mathcal{R}_B(\psi)$ or $\mathcal{R}_M(\psi)$ and the infimum is over all measurable functions of $\mathbf{X}_0$ and the training data $\mathbf{X}^{(l)}$, $l\in[\![2]\!]$, with values on $[0,1]$, and construct classifiers that would provide successful classification in part of these regions. Following Ingster et al. [16], we say that a classifier $\psi = \psi_b$ is *asymptotically optimal* if for all β and r such that successful classification is possible, we have

$$\lim_{b\to\infty}\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathcal{R}(\psi) = 0.$$

2.2. Classification regions

Given the sparse model in question, we shall restrict our attention to the most interesting case of *high β-sparsity* with values of β between $(1-\theta)/2$ and $1-\theta$. The reason for this is that a classification problem for "moderately β-sparse" vectors with $\beta \in (0,(1-\theta)/2)$ is easy and not of much interest, whereas successful classification of "very highly β-sparse" vectors with

$\beta \in (1-\theta, 1)$ is impossible (see Remark 1 in Ingster et al. [16]). In the case of moderate β-sparsity, successful classification is possible without preliminary selecting useful feature blocks and is provided, for example, by the classification rule $\psi_b^* = \mathbb{I}\Big\{\sum_{k=1}^b (\mathbf{X}_{0,[k]} - \hat{\boldsymbol{\mu}}_{[k]}/2)^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \hat{\boldsymbol{\mu}}_{[k]} \leqslant 0\Big\}$. Another parameter of the model at hand is r; it may be viewed as the *signal strength* parameter. Depending on the values of r (as a function of β), we will suggest different classification procedures. In general, the larger the value of r, the easier the classification problem.

To be more precise, consider the following function of $\beta \in (0,1)$ (see, for example, Ingster [15] and Donoho and Jin [9]):

$$\rho(\beta) = \begin{cases} 0, & 0 < \beta \leqslant 1/2, \\ \beta - 1/2, & 1/2 < \beta \leqslant 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{2.11}$$

The curve $r = \rho(\beta)$ is often called the *detection boundary*. The *classification boundary* is known to be a rescaled detection boundary of the form (see Ingster et al. [16] and Fan et al. [11])

$$r = \rho^*(\beta), \quad \rho^*(\beta) = (1-\theta)\,\rho\left(\frac{\beta}{1-\theta}\right) \quad \text{for} \quad (1-\theta)/2 < \beta < 1-\theta. \tag{2.12}$$

That is, for all large enough b, successful classification is possible if $r > \rho^*(\beta)$ and it is impossible if $r < \rho^*(\beta)$. The impossibility of classification means that

$$\liminf_{b\to\infty} \inf_{\psi} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in \mathbf{M}_{b,\beta,r}} \mathcal{R}(\psi) = 1/2,$$

that is, no classification rule (or classifier) is better than a simple random guess.

Along with the detection boundary $\rho(\beta)$ defined by (2.11), consider the *selection boundaries* $r = \rho_1(\beta)$ and $r = \rho_2(\beta)$, where (see, for example, Genovese et al. [12] and Ingster and Stepanova [17])

$$\rho_1(\beta) = \beta \quad \text{and} \quad \rho_2(\beta) = \left(1 + \sqrt{1-\beta}\right)^2, \quad 0 < \beta < 1.$$

For all $(1-\theta)/2 < \beta < 1-\theta$ one has $\rho(\beta) < \rho^*(\beta) < \rho_1(\beta) < \rho_2(\beta)$.

Given a parameter $\theta \in (0,1)$ which relates n and b through (2.5), we shall consider the following two regions of the parameter space $\{(\beta, r) \in \mathbb{R}^2 : 0 < \beta < 1,\ 0 < r < 4\}$ where classification is possible:

$$\begin{aligned} \mathcal{D}_1(\theta) &= \{(\beta, r) \in \mathbb{R}^2 : (1-\theta)/2 < \beta < 1-\theta,\ \rho_1(\beta) < r < 4\}, \\ \mathcal{D}_2(\theta) &= \{(\beta, r) \in \mathbb{R}^2 : (1-\theta)/2 < \beta < 1-\theta,\ \rho^*(\beta) < r \leqslant \rho_1(\beta)\}. \end{aligned}$$

Figure 1 displays the two regions: the region $\mathcal{D}_1(\theta) \cup \mathcal{D}_2(\theta)$ where classification is possible and its complement in $((1-\theta)/2, 1-\theta) \times (0,4)$ where classification is impossible, along with the detection boundary $r = \rho(\beta)$ and the selection boundaries $r = \rho_1(\beta)$ and $r = \rho_2(\beta)$.

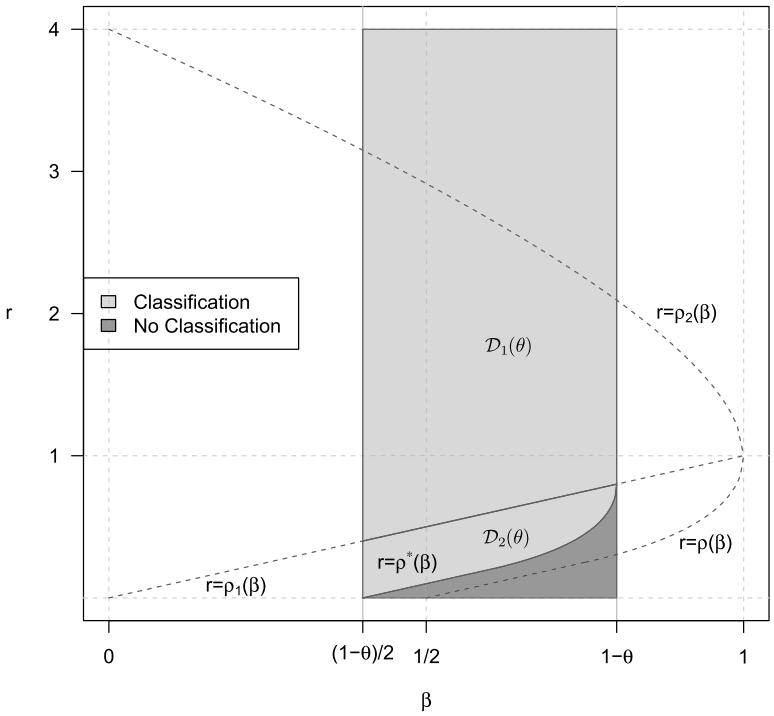

FIG 1. *The curve $r = \rho^*(\beta)$ divides the rectangle $(1-\theta)/2, 1-\theta) \times (0,4)$ into classification and no classification regions. In case of known covariance matrix, the classification region further splits into two subregions $\mathcal{D}_1(\theta)$ and $\mathcal{D}_2(\theta)$.*

In the region $\mathcal{D}_1(\theta)$, we construct an *asymptotically optimal classifier* that is fully data-driven and does not require the knowledge of β to be applied (for details, see Section 3.2). In the region $\mathcal{D}_2(\theta)$, where the classification problem is much harder, based on certain heuristic arguments, we propose a classifier that works well and improves the idea of Donoho and Jin [10] (for details, see Section 3.3). The properties of this classifier are studied numerically in Section 5.

In case of unknown covariance matrix $\mathbf{\Sigma}$, the division of the classification region into two subregions, denoted below by $\mathcal{D}_1^0(\theta)$ and $\mathcal{D}_2^0(\theta)$, is slightly different (see Figure 7 in Section 4.1); this is due to the impact of estimating the true $\mathbf{\Sigma}^{-1}$ by $\widehat{\mathbf{\Sigma}}^{-1}$ on the classification error. Similar to the case of known $\mathbf{\Sigma}$, in the region $\mathcal{D}_1^0(\theta)$ the problem of classification is easier, whereas in the region $\mathcal{D}_2^0(\theta)$ it is more difficult. We first consider in detail the case of known covariance matrix $\mathbf{\Sigma}$ and then extend the obtained results to the case of unknown $\mathbf{\Sigma}$ (see Section 4).

3. Classification when $\mathbf{\Sigma}$ is known

In the present setup, we distinguish between the regions $\mathcal{D}_1(\theta)$ and $\mathcal{D}_2(\theta)$. In the region $\mathcal{D}_1(\theta)$ one can identify useful feature blocks in a precise enough way. In the region $\mathcal{D}_2(\theta)$, where the parameter r (signal strength) is relatively small, the problem of identifying useful blocks is much more difficult.

3.1. Some useful statistics

For $b=2,3,\ldots$, let $\hat{\boldsymbol{\mu}}=(\hat{\boldsymbol{\mu}}_{[1]}^\top,\ldots,\hat{\boldsymbol{\mu}}_{[b]}^\top)^\top$ be the estimator of $\boldsymbol{\mu}=(\boldsymbol{\mu}_{[1]}^\top,\ldots,\boldsymbol{\mu}_{[b]}^\top)^\top$ and let $\widehat{\mathbf{\Sigma}}=\mathrm{Diag}(\widehat{\mathbf{\Sigma}}_{[1]},\ldots,\widehat{\mathbf{\Sigma}}_{[b]})$ be the estimator of $\mathbf{\Sigma}=\mathrm{Diag}(\mathbf{\Sigma}_{[1]},\ldots,\mathbf{\Sigma}_{[b]})$ as above. The random matrix $(2n-1)\widehat{\mathbf{\Sigma}}_{[k]}$ has a (central) Wishart $W_{p_0}(\mathbf{\Sigma}_{[k]},2n-1)$ distribution. The distribution of $\widehat{\mathbf{\Sigma}}_{[k]}^{-1}/(2n-1)$ is called the inverted Wishart distribution, and $\mathbf{E}\left(\widehat{\mathbf{\Sigma}}_{[k]}^{-1}/(2n-1)\right)=(2n-p_0-2)^{-1}\mathbf{\Sigma}_{[k]}^{-1}$. For $k\in[\![b]\!]$ and $b=2,3,\ldots$, we further define

$$\tilde{\Delta}_{k,b}^2=\hat{\boldsymbol{\mu}}_{[k]}^\top\mathbf{\Sigma}_{[k]}^{-1}\hat{\boldsymbol{\mu}}_{[k]},\quad \hat{\Delta}_{k,b}^2=\hat{\boldsymbol{\mu}}_{[k]}^\top\widehat{\mathbf{\Sigma}}_{[k]}^{-1}\hat{\boldsymbol{\mu}}_{[k]}.$$

Then, if $\mathbf{\Sigma}$ is *known*, we may consider a triangular array of statistics

$$\{\tilde{T}_{k,b};\ k\in[\![b]\!],\ b=2,3,\ldots\},\quad \tilde{T}_{k,b}=n\tilde{\Delta}_{k,b}^2. \tag{3.1}$$

The statistics $\tilde{T}_{k,b}$ are independent within each series and

$$\tilde{T}_{k,b}\sim\chi_{p_0}^2(n\Delta_{k,b}^2),\quad k\in[\![b]\!],\ b=2,3,\ldots.$$

The difficulty of identifying useful blocks of the data in the region $\mathcal{D}_2(\theta)$ as compared to the region $\mathcal{D}_1(\theta)$ is seen from Figures 2 and 3. Figure 2 shows a histogram for the chi-square data $\{\tilde{T}_{k,b}:k\in[\![b]\!]\}$, in the region $\mathcal{D}_1(\theta)$, where variable selection is possible. Figure 3 shows a histogram for the data $\{\tilde{T}_{k,b}:k\in[\![b]\!]\}$, in the region $\mathcal{D}_2(\theta)$, where variable selection is impossible (but classification is still possible). On Figures 2 and 3, the central and noncentral chi-square density curves are seen as red and blue lines, respectively.

The failure to classify a new observation $\mathbf{X}_0$ as belonging to either Π_1 or Π_2 outside of the region $\mathcal{D}_1(\theta)\cup\mathcal{D}_2(\theta)$ in the rectangle $((1-\theta)/2,1-\theta)\times(0,4)$ is illustrated by Figure 4, which shows a histogram for the data $\{\tilde{T}_{k,b}:k\in[\![b]\!]\}$, in the region where classification is impossible. In this case, $[b^{1-\beta}]$ noncentral chi-square statistics $\tilde{T}_{k,b}$ get too close to the remaining $b-[b^{1-\beta}]$ central chi-square statistics $\tilde{T}_{k,b}$ to allow successful classification in the sparse regime of our interest.

In a more realistic scenario when $\mathbf{\Sigma}$ is *unknown*, we shall make use of the statistics

$$\{\hat{T}_{k,b};\ k\in[\![b]\!],\ b=2,3,\ldots\},\quad \hat{T}_{k,b}=\frac{(2n-p_0)n}{(2n-1)p_0}\,\hat{\Delta}_{k,b}^2. \tag{3.2}$$

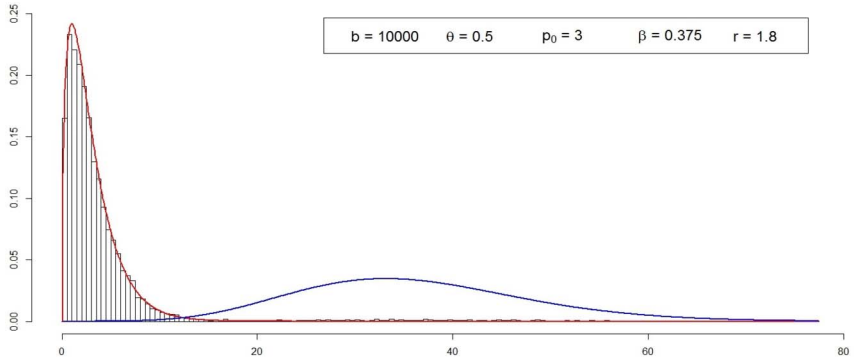

FIG 2. *Histogram for the chi-square data* $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ *when* $(\beta, r) \in \mathcal{D}_1(\theta)$.

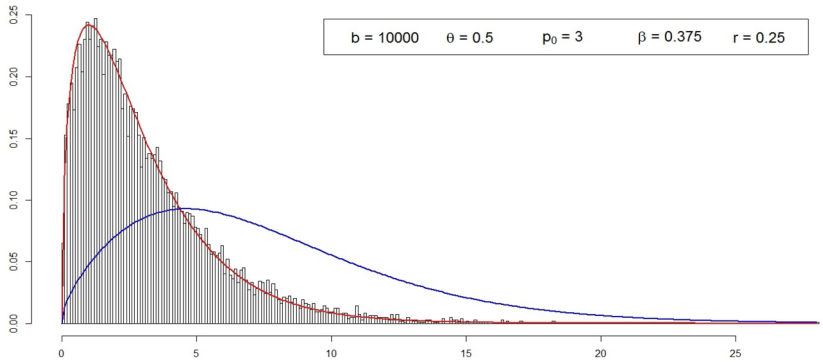

FIG 3. *Histogram for the chi-square data* $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ *when* $(\beta, r) \in \mathcal{D}_2(\theta)$.

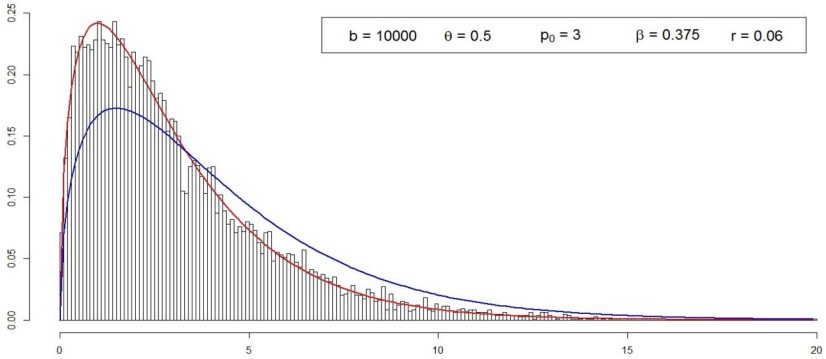

FIG 4. *Histogram for the chi-square data* $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ *when* (β, r) *falls in the complement of* $\mathcal{D}_1(\theta) \cup \mathcal{D}_2(\theta)$ *in the rectangle* $((1-\theta)/2, 1-\theta) \times (0, 4)$.

The statistics $\hat{T}_{k,b}$ are independent within each series and satisfy (see Section 8b of Rao [20])

$$\hat{T}_{k,b} \sim F_{p_0,2n-p_0}(n\Delta_{k,b}^2), \quad k \in [\![b]\!],\ b = 2,3,\ldots.$$

Note in passing that as $b \to \infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=1}^{b} \left(\hat{\Delta}_{k,b}^2 - \tilde{\Delta}_{k,b}^2\right) = o_{\mathbf{P}}(b^{1-\beta-\theta}\log b). \tag{3.3}$$

"In distribution" closeness of $\tilde{T}_{k,b}$ and $p_0\hat{T}_{k,b}$ for large b (for details, see Section 4.1) and the consistency of $\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}$ as an estimator of $\boldsymbol{\Sigma}_{[k]}^{-1}$ allow us to extend the results obtained for the case of known $\boldsymbol{\Sigma}$ to the case of unknown $\boldsymbol{\Sigma}$.

By the sparsity assumption on the model, only $s = [b^{1-\beta}] = o(b)$ statistics among $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ have a noncentral chi-square distribution, and the remaining $(b-s) = b+o(b)$ ones follow a central chi-square distribution. Similarly, only $s = [b^{1-\beta}] = o(b)$ statistics among $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ have a noncentral F distribution, and the remaining $(b-s) = b+o(b)$ ones follow a central F distribution. The noncentrally distributed statistics tend to take larger values as compared to the corresponding centrally distributed statistics. Therefore, "large" values of $\tilde{T}_{k,b}$ and $\hat{T}_{k,b}$ would suggest that the kth block of the data is useful and should be used for classification. These observations lead us to the estimators $\tilde{\omega}_k$ and $\hat{\omega}_k$ of ω_k, $k \in [\![b]\!]$, as given below by (3.4) and (4.1).

3.2. Classification rule in the region $\mathcal{D}_1(\theta)$

Let $\tilde{T}_{k,b}$, $k \in [\![b]\!]$, be the statistics as in (3.1). Consider a classifier $\tilde{\psi}_b$ defined by (2.3) for which

$$\tilde{\omega}_k = \tilde{\omega}_{k,b} = \mathbb{I}\left(\tilde{T}_{k,b} > \tilde{t}\right), \quad k \in [\![b]\!], \tag{3.4}$$

called here a *selector*, is an estimator of $\omega_k = \mathbb{I}(\Delta_{k,b}^2 \neq 0)$, $k \in [\![b]\!]$, with the threshold level $\tilde{t} = \tilde{t}\left(\mathbf{X}^{(2)}\right) > 0$ chosen as follows.

Pick a large number $M = M_b$ such that

$$M \to \infty, \quad b/M \to \infty, \quad (\log b)/M \to 0, \tag{3.5}$$

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

$$\beta_m = m\delta, \quad \delta \sim \frac{1-\theta}{2M}, \quad m \in [\![M]\!]. \tag{3.6}$$

In view of the above assumptions on M,

$$\delta \to 0, \quad \delta \log b \to 0, \quad b \to \infty, \tag{3.7}$$

yielding for all large enough b

$$b^{\delta} \leqslant \text{const.} \tag{3.8}$$

Next, for all $k \in [\![b]\!]$ and $m \in [\![M]\!]$, put

$$\tilde{\omega}_k(\beta_m) = \tilde{\omega}_{k,b}(\beta_m) = \mathbb{I}\left(\tilde{T}_{k,b} > (2\beta_m + \epsilon)\log b\right), \tag{3.9}$$

where $\epsilon = \epsilon_b > 0$ satisfies

$$\epsilon \to 0 \quad \text{and} \quad \epsilon \log b / \log\log b \to \infty \quad \text{as} \quad b \to \infty. \tag{3.10}$$

We define an *adaptive selector* by the formula

$$\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}) = (\tilde{\omega}_1(\beta_{\tilde{m}}), \ldots, \tilde{\omega}_b(\beta_{\tilde{m}})), \tag{3.11}$$

where $\tilde{m} = \tilde{m}_b$ is chosen by Lepski's method (see Section 2 of Lepski [18]) as follows, cf. relation (37) in Butucea and Stepanova [4]:

$$\tilde{m} = \max\{m \in [\![M]\!] : d(\tilde{\boldsymbol{\omega}}(\beta_m), \tilde{\boldsymbol{\omega}}(\beta_j)) \leqslant v_j \ \text{ for all } \ j \leqslant m\}, \tag{3.12}$$

and $\tilde{m} = 1$ if the set in (3.12) is empty. Here $d(\tilde{\boldsymbol{\omega}}, \boldsymbol{\omega}) = \sum_{k=1}^{b} |\tilde{\omega}_k - \omega_k|$ is the *Hamming loss* that counts the number of positions at which $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_k)_{k\in[\![b]\!]}$ and $\boldsymbol{\omega} = (\omega_k)_{k\in[\![b]\!]}$ differ, and the quantities $v_j = v_{j,b}$ are set to be $v_j = b^{1-\beta_j}/\tau_b, j \in [\![m]\!]$, with a sequence of numbers $\tau_b \to \infty$ satisfying $\tau_b = o\left(b^{\epsilon/2}\log^{1-p_0/2} b\right)$ as $b \to \infty$.

Algorithmically, Lepski's procedure for choosing $\tilde{m}$ works as follows. We start by setting $\tilde{m} = 1$ and attempt to increase the value of $\tilde{m}$ from 1 to 2. If $d(\tilde{\boldsymbol{\omega}}(\beta_2), \tilde{\boldsymbol{\omega}}(\beta_1)) \leqslant v_1$, we set $\tilde{m} = 2$; otherwise, we keep $\tilde{m}$ equal to 1. In case $\tilde{m}$ is increased to 2, we continue the process attempting to increase it further. If $d(\tilde{\boldsymbol{\omega}}(\beta_3), \tilde{\boldsymbol{\omega}}(\beta_2)) \leqslant v_2$ and $d(\tilde{\boldsymbol{\omega}}(\beta_3), \tilde{\boldsymbol{\omega}}(\beta_1)) \leqslant v_1$, we set $\tilde{m} = 3$; otherwise, we keep $\tilde{m}$ equal to 2; and so on. By construction $v_1 \geqslant v_2 \geqslant \ldots \geqslant v_M$. It can be seen from the proof of (3.13) below that if $m_0 \in [\![M-1]\!]$ is such that the true $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$, then $\tilde{m} \geqslant m_0$ with high probability.

The next result shows that, asymptotically, $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ identifies correctly most of the noncentrally distributed chi-square statistics among $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$.

Lemma 1. *Consider the* $([b^{1-\beta}], 2r\log b)$*-sparse block-diagonal normal model with known covariance matrix* $\boldsymbol{\Sigma}$*, and let the statistics* $\{\tilde{T}_{k,b} : k \in [\![b]\!], b = 2, 3, \ldots\}$ *be as defined in* (3.1)*. Then, the selector* $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ *in* (3.11) *based on* $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ *with* $\tilde{m}$ *defined by* (3.12) *is an almost full selector in the sense that for all* $(1-\theta)/2 < \beta < 1-\theta$ *and* $\beta < r < 4$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}\, d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}), \boldsymbol{\omega}) = o(b^{1-\beta}), \quad b \to \infty. \tag{3.13}$$

Lemma 1, whose proof is given in Section 7, says that in the region $\mathcal{D}_1(\theta)$ the maximum Hamming risk of $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ is small relative to the number of noncentrally

distributed statistics among $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$. This suggests that in the definition of the classifier $\tilde{\psi}_b$ given by (2.3) and (3.4) the threshold $\tilde{t} = \tilde{t}_b$ should be set at the level

$$\tilde{t} = (2\beta_{\tilde{m}} + \epsilon)\log b. \tag{3.14}$$

The next result shows that in the region $\mathcal{D}_1(\theta)$ the classification rule $\tilde{\psi}_b$ defined by (2.3), (3.4), and (3.14) is asymptotically optimal.

Theorem 1. *Let $\mathbf{X}^{(1)} = (\mathbf{X}_j^{(1)})_{j\in[\![n]\!]}$ and $\mathbf{X}^{(2)} = (\mathbf{X}_j^{(2)})_{j\in[\![n]\!]}$ be training samples of size n in the $([b^{1-\beta}], 2r\log b)$-sparse block-diagonal normal model, with n and b related through (2.5) for a given number $\theta \in (0,1)$, and let $\mathbf{X}_0$ be a new observation to be classified. Assume that the covariance matrix $\mathbf{\Sigma}$ is known. Then, for all $(\beta, r) \in \mathcal{D}_1(\theta)$, the classifier $\tilde{\psi}_b$ defined by (2.3), (3.4), and (3.14) satisfies*

$$\lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathcal{R}(\tilde{\psi}_b) = 0.$$

Proof. For a number $\theta \in (0,1)$, let (β, r) be an arbitrary point in the region $\mathcal{D}_1(\theta)$. It suffices to show that

$$\lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}_{\Pi_2}(\tilde{\psi}_b) = 0, \quad \lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}_{\Pi_1}(1-\tilde{\psi}_b) = 0, \tag{3.15}$$

where $\mathbf{E}_{\Pi_i}$ denotes the expectation with respect to the joint distribution of $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}_0$ when $\mathbf{X}_0 \sim \Pi_i$ for $i \in [\![2]\!]$.

For $b = 2, 3, \ldots$, denote

$$\tilde{V}_k = \tilde{V}_{k,b} = \mathbf{X}_{0,[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \widehat{\boldsymbol{\mu}}_{[k]} - (1/2)\widehat{\boldsymbol{\mu}}_{[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \widehat{\boldsymbol{\mu}}_{[k]}, \quad k \in [\![b]\!], \tag{3.16}$$

and observe that the classifier $\tilde{\psi}_b$ can be expressed as

$$\tilde{\psi}_b = \mathbb{I}\left\{ \sum_{k=1\,:\,\tilde{\omega}_k=1}^{b} \tilde{V}_k \leqslant 0 \right\},$$

where $\tilde{\omega}_k = \tilde{\omega}_k(\beta_{\tilde{m}})$ is the kth component of $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ in (3.11). Denote also

$$V_k = V_{k,b} = \mathbf{X}_{0,[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]} - (1/2)\boldsymbol{\mu}_{[k]}^\top \mathbf{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}, \quad k \in [\![b]\!], \tag{3.17}$$

and note that, under Π_2, $V_k \sim N_{p_0}\left(\Delta_{k,b}^2/2, \Delta_{k,b}^2\right)$ for all $k \in [\![b]\!]$, $b = 2, 3, \ldots$. Therefore, in view of the identity

$$\begin{aligned} \sum_{k=1\,:\,\tilde{\omega}_k=1}^{b} \tilde{V}_k - \sum_{k=1\,:\,\omega_k=1}^{b} V_k \;=\; & \sum_{k=1\,:\,\omega_k=1}^{b} (\tilde{V}_k - V_k) + \sum_{k=1\,:\,\omega_k=0,\tilde{\omega}_k=1}^{b} \tilde{V}_k \\ & - \sum_{k=1\,:\,\omega_k=1,\tilde{\omega}_k=0}^{b} \tilde{V}_k, \end{aligned} \tag{3.18}$$

we can write

$$\left\{\sum_{k=1:\,\tilde{\omega}_k=1}^{b}\tilde{V}_k\leqslant 0\right\}=\left\{\sum_{k=1:\,\omega_k=1}^{b}(V_k-\mathbf{E}_{\Pi_2}(V_k))+\sum_{k=1:\,\omega_k=1}^{b}(\tilde{V}_k-V_k)\right.$$
$$\left.+\sum_{k=1:\,\omega_k=0,\tilde{\omega}_k=1}^{b}\tilde{V}_k-\sum_{k=1:\,\omega_k=1,\tilde{\omega}_k=0}^{b}\tilde{V}_k\leqslant-\frac{1}{2}\sum_{k=1:\,\omega_k=1}^{b}\Delta_{k,b}^2\right\},\quad(3.19)$$

where, by assumption and the fact that $n\sim b^\theta$, for all $(\beta,r)\in\mathcal{D}_1(\theta)$, as $b\to\infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1:\,\omega_k=1}^{b}\Delta_{k,b}^2=O\left(b^{1-\beta-\theta}\log b\right).\quad(3.20)$$

The following result shows that the main contribution to $\sum_{k=1:\,\tilde{\omega}_k=1}^{b}\tilde{V}_k$ is made by the term $\sum_{k=1:\,\omega_k=1}^{b}V_k$.

Lemma 2. *For any $\theta\in(0,1)$ and all $(\beta,r)\in\mathcal{D}_1(\theta)$, as $b\to\infty$*

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1:\,\omega_k=1}^{b}(\tilde{V}_k-V_k)=o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right),\quad(3.21)$$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1:\,\omega_k=1,\tilde{\omega}_k=0}^{b}\tilde{V}_k=o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right),\quad(3.22)$$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1:\,\omega_k=0,\tilde{\omega}_k=1}^{b}\tilde{V}_k=o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right).\quad(3.23)$$

The proof of Lemma 2, where the key role is taken by Lemma 1, is given in Section 7. Next, for $b=2,3,\ldots$, let us introduce the event

$$\mathbb{A}_b=\left\{\left|\sum_{k=1:\,\tilde{\omega}_k=1}^{b}\tilde{V}_k-\sum_{k=1:\,\omega_k=1}^{b}V_k\right|>\frac{1}{4}\sum_{k\,:\,\omega_k=1}^{b}\Delta_{k,b}^2\right\}.$$

Then, by (3.19) and Chebyshev's inequality, for all sufficiently large b

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}\left(\textstyle\sum_{k=1:\,\tilde{\omega}_k=1}^{b}\tilde{V}_k\leqslant 0\right)$$
$$\leqslant\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}\left(\sum_{k=1:\,\omega_k=1}^{b}(V_k-\mathbf{E}_{\Pi_2}(V_k))\leqslant-\frac{1}{4}\sum_{k\,:\,\omega_k=1}^{b}\Delta_{k,b}^2\right)$$
$$+\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}(\mathbb{A}_b)\leqslant\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\frac{\sum_{k=1:\,\omega_k=1}^{b}\Delta_{k,b}^2}{\left((1/4)\sum_{k\,:\,\omega_k=1}^{b}\Delta_{k,b}^2\right)^2}$$
$$+\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}(\mathbb{A}_b).$$

Together with relations (3.18) and (3.20), and Lemma 2 applied to the term $\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2}(\mathbb{A}_b)$, this upper bound yields, as $b \to \infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}_{\Pi_2}(\tilde{\psi}_b) = O\left((b^{1-\beta-\theta}\log b)^{-1}\right) + o(1) = o(1),$$

uniformly in $(\beta, r) \in \mathcal{D}_1(\theta)$ for all $\theta \in (0,1)$. This proves the first relation in (3.15). The second relation in (3.15) is proved completely analogously. The proof is complete. □

Remark 1. Inspection of the proof of Theorem 1 shows that it can be extended to a more realistic case of *unequally-sized* blocks $\boldsymbol{\Sigma}_{[1]}, \ldots, \boldsymbol{\Sigma}_{[b]}$ of respective sizes $p_1 \times p_1, \ldots, p_b \times p_b$, where $p_k \geqslant 3$, $k \in [\![b]\!]$, are uniformly bounded integers such that $\sum_{k=1}^{b} p_k = p$. This is so because, in view of relations (7.2)–(7.4) in the proof of Lemma 1, for large b the tails of the central and noncentral chi-square statistics $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ are not essentially affected by their degrees of freedom.

3.3. Classification rule in the region $\mathcal{D}_2(\theta)$

In the region $\mathcal{D}_2(\theta)$, where the parameter r is small and feature selection is impossible, the classification problem is very hard (see Figure 3). We suggest that, in this region of (β, r)-values, the threshold $\tilde{t}$ of the classifier $\tilde{\psi}_b$ given by (2.3) and (3.4) be chosen by using the weighted Kolmogorov–Smirnov statistic with a suitable weight function q. Namely, we set the threshold $\tilde{t} = \tilde{t}_q$ at the level (see formula (3.29) below)

$$\tilde{t}_q = \tilde{T}_{(b+1-\tilde{k}_q)}$$

where $\tilde{T}_{(k)}$ is the kth order statistic of the data $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ and the index $\tilde{k}_q$ is given by formula (3.28). The idea behind this choice of $\tilde{t}_q$ is tightly connected to the problem of signal detection in sparse chi-square mixtures by means of weighted Kolmogorov-Smirnov tests and is detailed below. It is similar to the suggestion of Donoho and Jin [10]; the main difference is a more general classification model and a different choice of the weight function used.

In view of the intrinsic difficulty of the problem, based on heuristic arguments, we propose a classifier that numerically works well. As seen from Section 5, our classifier $\tilde{\psi}_b$ given by (2.3), (3.4), and (3.29) works numerically better than the procedure of Donoho and Jin [10]. There also exists a Donoho–Jin type classifier proposed by Fan et al. [11] for the situation when the inverse $\boldsymbol{\Sigma}^{-1}$ of a covariance matrix $\boldsymbol{\Sigma}$ admits an "acceptable" estimator which selects useful information by means of truncated higher-criticism thresholding. Its quality, however, is hard to assess because the numerical results in Section 4 of Fan et al. [11] are only given for the region where feature selection is possible and where the classification problem is relatively easy, whereas the obtained analytical results for the whole classification region are "strongly asymptotic", and it is not clear for what p the asymptotics start to give reasonably accurate descriptions of the actual finite sample performance.

Assume that $(\beta, r) \in \mathcal{D}_2(\theta)$ and consider the *worst case scenario* when all nonzero noncentrality parameters $n\Delta_{k,b}^2$, $k \in [\![b]\!]$, are equal to $2r \log b$ with some $\rho^*(\beta) < r < \beta$. Then, asymptotically, the statistics $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ obey a chi-square mixture model

$$\tilde{T}_{1,b}, \dots, \tilde{T}_{b,b} \overset{iid}{\sim} (1 - \varepsilon_b)\chi^2_{p_0}(0) + \varepsilon_b \chi^2_{p_0}(\gamma_b), \tag{3.24}$$

where $\varepsilon_b = b^{-\beta}$ for $0 < \beta < 1$ and $\gamma_b = 2r \log b$ for $\rho^*(\beta) < r < \beta$. Therefore, we may consider an axillary problem of testing the null hypothesis H_0 versus the alternative (more precisely, a sequence of alternatives) $H_{1,b}$ given by

$$\begin{aligned} H_0 &: t_1, \dots, t_b \overset{iid}{\sim} \chi^2_{p_0}(0), \\ H_{1,b} &: t_1, \dots, t_b \overset{iid}{\sim} (1 - \varepsilon_b)\chi^2_{p_0}(0) + \varepsilon_b \chi^2_{p_0}(\gamma_b), \end{aligned}$$

where $\varepsilon_b = b^{-\beta}$ for $0 < \beta < 1$, p_0 is as before, and $\gamma_b = 2r \log b$ for $0 < r < 1$. Next, we transform the statistics t_k to the uniformly distributed on the interval $(0, 1)$ statistics $s_k = 1 - G_{p_0}(t_k; 0)$, $k \in [\![b]\!]$, where $G_\nu(x; \gamma) = \mathbf{P}(\chi^2_\nu(\gamma) \leqslant x)$, $x \in \mathbb{R}$. In terms of a common cdf $F(u)$ of the s_k's, the problem of testing H_0 versus $H_{1,b}$ 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

$$\mathcal{H}_{1,b} : F(u) = F_0(u) + \varepsilon_b \left((1 - u) - G_{p_0}(G^{-1}_{p_0}(1 - u; 0); \gamma_b)\right) > F_0(u).$$

In connection with testing H_0 versus $H_{1,b}$ (or, equivalently, $\mathcal{H}_0$ versus $\mathcal{H}_{1,b}$), consider the function $\rho(\beta)$ defined in (2.11). It is known (see, for example, Section 4 of Stepanova and Pavlenko [24]) that if $r > \rho(\beta)$ then the hypotheses separate asymptotically, whereas if $r < \rho(\beta)$ then these hypotheses merge asymptotically, that is, no consistent test exists. More precisely, let

$$\mathbb{H}_b(u) = b^{-1} \sum_{k=1}^{b} \mathbb{I}(s_k < u), \quad 0 < u < 1,$$

be the empirical distribution function (edf) based on the s_k's and for $\sigma = -1/2, 0, 1/2$ let

$$q_\sigma(u) = \sqrt{u(1-u)}\,(\log\log(1/(u(1-u))))^{1/2+\sigma}, \quad 0 < u < 1. \tag{3.25}$$

The function $q_{-1/2}(u) = \sqrt{u(1-u)}$ is a regularly varying function which is known in the literature as the *standard deviation proportional (SDP) weight function*. The function $q_0(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$ is an *Erdős-Feller-Kolmogorov-Petrovski (EFKP) upper-class function of a Brownian bridge*; the importance of such weight functions in the theory of weighted quantile and empirical processes has been demonstrated by Csörgő et al. [7]. The function

$q_{1/2}(u) = \sqrt{u(1-u)}\log\log(1/(u(1-u)))$ is an example of the *Chibisov–O'Reilly function*. For the use of these three classes of functions in the theory of weighted quantile and empirical processes, we refer to Csörgő et al. [7] and Csörgő and Horváth [8]. It is known (see Donoho and Jin [9] and Stepanova and Pavlenko [24]) that the tests based on the (one-sided) weighted Kolmogorov-Smirnov statistics

$$D_b^+(q_\sigma) = \sup_{0<u<\alpha_0} \frac{\sqrt{b}(\mathbb{H}_b(u)-u)}{q_\sigma(\mathbb{H}_b(u))}, \quad \sigma = -1/2, 0, 1/2, \tag{3.26}$$

where $\alpha_0 \in (0, 1/2)$ is a small number (say, $\alpha_0 = 0.2$) chosen by the statistician, distinguish between $\mathcal{H}_0$ and $\mathcal{H}_{1,b}$ when $r > \rho(\beta)$, with $\mathcal{H}_0$ being rejected for "small" values of $D_b^+(q_\sigma)$. Moreover, the use of weight functions q_0 and $q_{1/2}$ makes the problem of distinguishing between the two hypotheses easier, as compared to using $q_{-1/2}$. This is so because, under $\mathcal{H}_0$, the statistics $D_b^+(q_0)$ and $D_b^+(q_{1/2})$ are finite with probability 1, whereas the statistic $D_b^+(q_{-1/2})$ introduced by Donoho and Jin [9] tends to infinity, in probability and even almost surely, under both $\mathcal{H}_0$ and $\mathcal{H}_{1,b}$, making the problem of separating these two hypotheses relatively hard.

Let $s_{(1)} < s_{(2)} < \ldots < s_{(b)}$ be the order statistics of the sample $s_1, \ldots, s_b$. Then, as each weight function q_σ is monotone on $(0, \alpha_0)$ for small $\alpha_0 \in (0, 0.2)$, the statistic $D_b^+(q_\sigma)$ is asymptotically equivalent to the statistic

$$\mathbf{D}_b^+(q_\sigma) = \max_{1\leqslant k\leqslant[\alpha_0 b]} \frac{\sqrt{b}\,\big(k/b - s_{(k)}\big)}{q_\sigma(k/b)}, \quad \sigma = -1/2, 0, 1/2.$$

In addition to weights $q_\sigma(u)$, $\sigma = -1/2, 0, 1/2$, we also explore one more weight function:

$$q_{1/4}(u) = (u(1-u))^{1/4}, \quad 0 < u < 1, \tag{3.27}$$

which is an example of the Chibisov–O'Reilly function.

As shown in Ingster et al. [16] and Fan et al. [11], in somewhat different yet similar settings, if $\mathcal{H}_0$ and $\mathcal{H}_{1,b}$ are indistinguishable (merge asymptotically), then successful classification cannot be achieved. It can only be achieved in the region of $r > \rho^*(\beta) > \rho(\beta)$ with $\rho^*(\beta)$ as in (2.12). Thus, recalling (3.24), we arrive at the following idea of selecting useful feature blocks by means of $D_b^+(q_\sigma)$-thresholding in the region $\mathcal{D}_2(\theta)$. This idea is similar to that of Donoho and Jin [10] to make feature selection via higher criticism thersholding, that is, by using the statistic, cf. (3.26),

$$\mathrm{HC}_b = \sup_{0<u<\alpha_0} \frac{\sqrt{b}(\mathbb{H}_b(u)-u)}{\sqrt{u(1-u)}},$$

which is the statistic $D_b^+(q_{-1/2})$ in our notation.

First, consider the statistics $S_{k,b} = 1 - G_{p_0}(\tilde{T}_{k,b}; 0)$, $k \in [\![b]\!]$, and note that, under H_0, the transformed statistics $\{S_{k,b} : k \in [\![b]\!], b = 2, 3, \ldots\}$ form a triangular array of iid uniform $U(0,1)$ random variables. Next, denote by $S_{(k)}$ the kth

order statistic of the sample $\{S_{k,b} : k \in [\![b]\!]\}$ and define the index $\tilde{k}_q$ by

$$\tilde{k}_q = \underset{1 \leqslant k \leqslant [\alpha_0 b]}{\arg\max} \frac{\sqrt{b}(k/b - S_{(k)})}{q(k/b)}, \tag{3.28}$$

where q is one of the weight functions q_σ with $\sigma = -1/2, 0, 1/2$, as given in (3.25) or $q_{1/4}$ as in (3.27). Finally, we take $S_{(\tilde{k}_q)}$ as a (random) feature selection threshold, that is, for all $l \in [\![2]\!]$, $j \in [\![n]\!]$, and $k \in [\![b]\!]$, the kth sub-vector $\mathbf{X}^{(l)}_{j,[k]}$ of vector $\mathbf{X}^{(l)}_j$ is deemed useful for classification if $S_{k,b}$ is smaller than $S_{(\tilde{k}_q)}$ or, equivalently, if $\tilde{T}_{k,b}$ is larger than $\tilde{t}_q$, where

$$\tilde{t}_q = G^{-1}_{p_0}(1 - S_{(\tilde{k}_q)}; 0) = \tilde{T}_{(b+1-\tilde{k}_q)}. \tag{3.29}$$

This choice of $\tilde{t}_q$ is motivated by the fact that, in the region $\mathcal{D}_2(\theta)$, $\mathcal{H}_0$ and $\mathcal{H}_{1,b}$ are distinguished by the statistic $D^+_b(q_\sigma)$ (see Section 4 of Stepanova and Pavlenko [24]). Note that the classifier $\tilde{\psi}_b$ defined by (2.3), (3.4), and (3.29) is fully adaptive in the parameters of the model.

The behaviour of the objective function $K_{q,b}(S_{(k)}) := \frac{\sqrt{b}(k/b - S_{(k)})}{q(k/b)}$ on the interval $(0, 0.2)$ for four different weight functions q and the corresponding thresholds $S_{(\tilde{k}_q)}$ are shown on Figure 5. Figure 6 shows a histogram for $\tilde{T}_{k,b}$, $k \in [\![b]\!]$, and the threshold $\tilde{t}_q$ for the four weight functions q of our interest. As seen from Figure 6, the threshold obtained by using the SDP weight function $q_{-1/2}$ (red line) retains too many useless feature blocks for future use in classification, whereas the thresholds that correspond to the Chibisov–O'Reilly weight functions $q_{1/2}$ and $q_{1/4}$ (yellow and green lines) appear to ignore a certain number of useful feature blocks contained in the data. The threshold obtained by using the EFKP upper-class weight function q_0 (blue line) is a compromiser that gives a better classification result (see Table 1 in Section 5).

Note in passing that, in the rare and weak regime in question, various false discovery rate (FDR) controlling multiple testing procedures, including the Benjamini–Hochberg rule, provide very few discoveries and thus lead to high classification error. The desirable properties of FDR controlling procedures in multiple testing have been analytically justified mainly for the situations where rare signals are strong.

4. Classification when $\boldsymbol{\Sigma}$ is unknown

4.1. Classification rule in the region $\mathcal{D}^0_1(\boldsymbol{\theta})$

Let $\hat{T}_{k,b}$, $k \in [\![b]\!]$, be the statistics as in (3.2). In the present settings when $\boldsymbol{\Sigma}$ is unknown, consider a classifier $\hat{\psi}_b$ defined by (2.4) for which

$$\hat{\omega}_k = \hat{\omega}_{k,b} = \mathbb{I}\left(\hat{T}_{k,b} > \hat{t}\right), \quad k \in [\![b]\!], \tag{4.1}$$

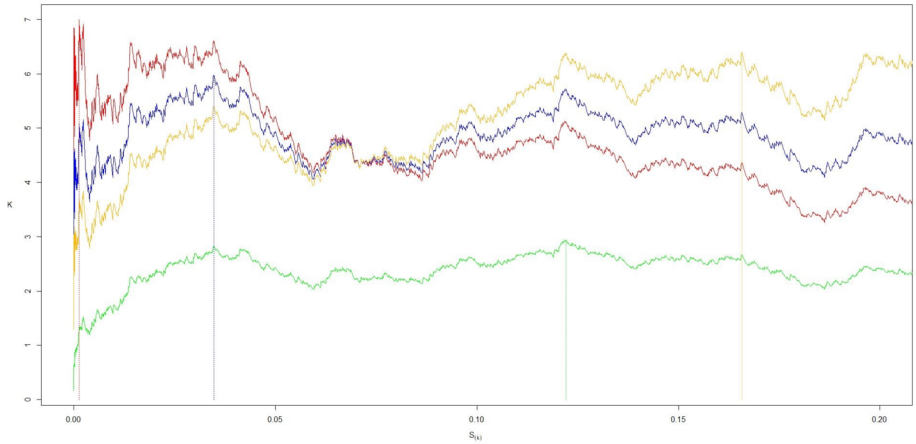

FIG 5. *Objective function $K_{q,b}(S_{(k)})$ on $(0, 0.2)$ with four different weights q for the transformed observations $\{S_{k,b} : k \in [\![b]\!]\}$ in the region $\mathcal{D}_2(\theta)$ with $\theta = 0.5$. The threshold $S_{(\tilde{k}_q)}$ is shown with yellow, green, blue, and red lines when q is $q_{1/2}$, $q_{1/4}$, q_0, and $q_{-1/2}$, respectively.*

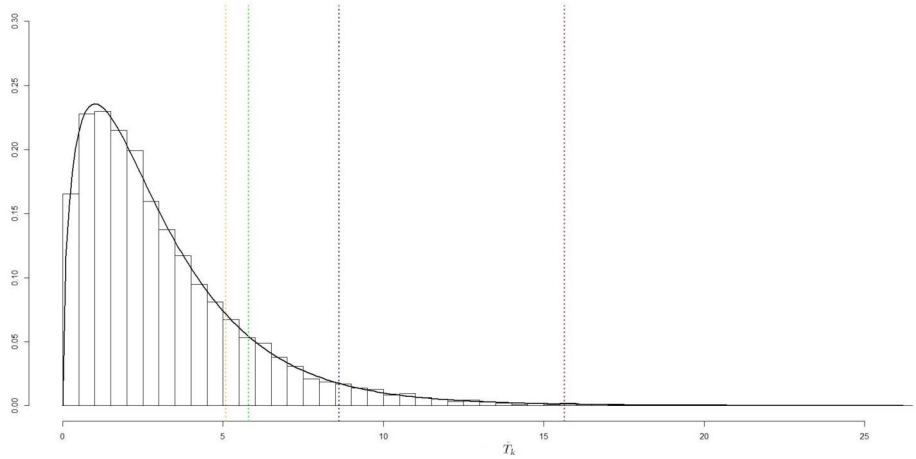

FIG 6. *Threshold $\tilde{t}_q$ as in (3.29) with four different weights q for the chi-square observations $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ in the region $\mathcal{D}_2(\theta)$ with $\theta = 0.5$. The threshold $\tilde{t}_q$ is shown with yellow, green, blue, and red vertical lines when q is $q_{1/2}$, $q_{1/4}$, q_0, and $q_{-1/2}$, respectively.*

with some threshold level $\hat{t} = \hat{t}\left(\mathbf{X}^{(1)}; \mathbf{X}^{(2)}\right) > 0$. We need to set up the threshold $\hat{t}$ in such a way that the maximum (over all $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$ and all $(\beta, r) \in \mathcal{D}_1(\theta)$) risk of $\hat{\psi}_b$ is small when b is large. In the case of unknown $\boldsymbol{\Sigma}$, we will have to narrow down the region $\mathcal{D}_1(\theta)$ of (β, r)-values, which is the price paid for not knowing $\boldsymbol{\Sigma}$.

Note that for all $b = 2, 3, \ldots$ the statistics $\hat{T}_{k,b}$, $k \in [\![b]\!]$, are independent

and there exists a set $S \subset [\![b]\!]$ with $s = [b^{1-\beta}]$ elements such that $\hat{T}_{k,b} \sim F_{p_0,2n-p_0}(n\Delta_{k,b}^2)$ for all $k \in S$, and $\hat{T}_{k,b} \sim F_{p_0,2n-p_0}(0)$ for all $k \notin S$. For $x \in \mathbb{R}$, let

$$\mathbb{F}_{\nu_1,\nu_2}(x;\gamma) = \mathbf{P}(\nu_1 F_{\nu_1,\nu_2}(\gamma) \leqslant x), \quad G_\nu(x;\gamma) = \mathbf{P}(\chi_\nu^2 \leqslant x).$$

Then, it follows from formula (6.8) of Siotani [22] that for any $x \geqslant 0$, any $\nu_1 > 0$, and all large enough ν_2 and γ, with γ tending to infinity not very fast,

$$\begin{aligned}\mathbb{F}_{\nu_1,\nu_2}(x;\gamma) &= G_{\nu_1}(x;\gamma) - \frac{1}{4\nu_2}\{\nu_1(\nu_1-2)G_{\nu_1}(x;\gamma) - 2\nu_1(\nu_1-\gamma)G_{\nu_1+2}(x;\gamma) \\ &\quad + [(\nu_1-\gamma)(\nu_1+2-\gamma) - 2(\nu_1+1)\gamma]\, G_{\nu_1+4}(x;\gamma) \\ &\quad + 2\gamma(\nu_1+2-\gamma)G_{\nu_1+6}(x;\gamma) + \gamma^2 G_{\nu_1+8}(x;\gamma)\} \\ &\quad + O(\nu_2^{-2}\gamma^4). \end{aligned} \tag{4.2}$$

Relation (4.2) shows that for large b, when multiplied by a constant factor p_0, the F-distributed statistics $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ in (3.2) are well approximated by the chi-square statistics $\{\tilde{T}_{k,b} : k \in [\![b]\!]\}$ in (3.1).

For all $0 < \theta < 1$, we now define the two subregions $\mathcal{D}_1^0(\theta)$ and $\mathcal{D}_2^0(\theta)$ of the classification region as follows:

$$\begin{aligned}\mathcal{D}_1^0(\theta) &= \{(\beta,r) \in \mathbb{R}^2 : (1-\theta)/2 < \beta < 1-\theta,\ \beta + \theta/2 < r < 4\} \subset \mathcal{D}_1(\theta), \\ \mathcal{D}_2^0(\theta) &= \{(\beta,r) \in \mathbb{R}^2 : (1-\theta)/2 < \beta < 1-\theta, \rho^*(\beta) < r \leqslant \beta + \theta/2\} \supset \mathcal{D}_2(\theta).\end{aligned}$$

In the region $\mathcal{D}_1^0(\theta)$, we define a selector $\hat{\boldsymbol{\omega}}(\beta_{\hat{m}}) = (\hat{\omega}_k(\beta_{\hat{m}}))_{k\in[\![b]\!]}$ based on $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ similar to the one in (3.11)–(3.12). Namely, we first pick a large number $M = M_b$, the equidistant grid points $(1-\theta)/2 < \beta_1 < \ldots < \beta_M < 1-\theta$, and a small number $\delta = \delta_b$ as in (3.5)–(3.7). Next, for all $k \in [\![b]\!]$ and $m \in [\![M]\!]$, we set, cf. (3.9),

$$\hat{\omega}_k(\beta_m) = \hat{\omega}_{k,b}(\beta_m) = \mathbb{I}\left(\hat{T}_{k,b} > p_0^{-1}(2\beta_m + \epsilon)\log b\right),$$

where $\epsilon = \epsilon_b > 0$ satisfies (3.10), and define an *adaptive selector* by the formula

$$\hat{\boldsymbol{\omega}}(\beta_{\hat{m}}) = (\hat{\omega}_1(\beta_{\hat{m}}), \ldots, \hat{\omega}_b(\beta_{\hat{m}})), \tag{4.3}$$

where $\hat{m} = \hat{m}_b$ is chosen by Lepski's method as follows, cf. (3.12):

$$\hat{m} = \max\{m \in [\![M]\!] : d\,(\hat{\boldsymbol{\omega}}(\beta_m), \hat{\boldsymbol{\omega}}(\beta_j)) \leqslant v_j \ \text{ for all } \ j \leqslant m\}, \tag{4.4}$$

and $\hat{m} = 1$ if the set in (4.4) is empty. Here the quantities $v_j = v_{j,b}$ are set to be $v_j = b^{1-\beta_j}/\tau_b, j \in [\![m]\!]$, with a sequence of numbers $\tau_b \to \infty$ satisfying $\tau_b = o\left(b^{\epsilon/2}\log^{1-p_0/2} b\right)$ as $b \to \infty$.

It is not difficult to show, cf. Lemma 1, that the selector $\hat{\boldsymbol{\omega}}(\beta_{\hat{m}})$ given by (4.3) is an almost full selector in the sense that for all $(1-\theta)/2 < \beta < 1-\theta$ and $\beta < r < 4$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}\, d\,(\hat{\boldsymbol{\omega}}(\beta_{\hat{m}}), \boldsymbol{\omega}) = o(b^{1-\beta}), \quad b \to \infty.$$

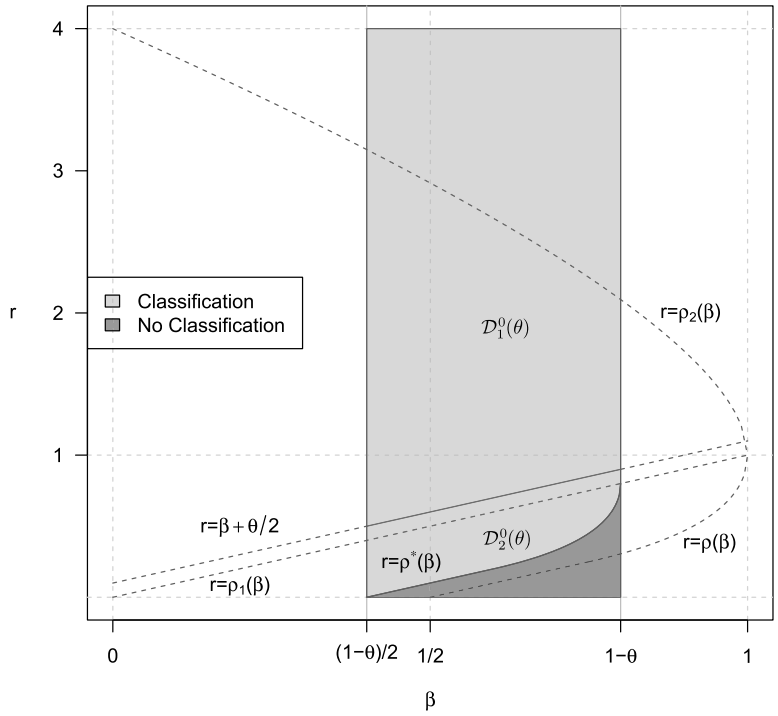

FIG 7. *The curve $r = \rho^*(\beta)$ divides the rectangle $(1-\theta)/2, 1-\theta) \times (0,4)$ into classification and no classification regions. In case of estimated covariance matrix, the classification region splits further into two subregions $\mathcal{D}_1^0(\theta)$ and $\mathcal{D}_2^0(\theta)$.*

For the purpose of classification, however, the threshold that would exclude most of the useless blocks from the classification procedure needs to be higher and, as a result of this, the region where the classifier $\hat{\psi}_b$ does its job properly is narrowed down, as compared to the region $\mathcal{D}_1(\theta)$ where $\tilde{\psi}_b$ works well, to become $\mathcal{D}_1^0(\theta)$. Namely, return to the definition of the classifier $\hat{\psi}_b$ given by (2.4) and (4.1), and define the threshold $\hat{t} = \hat{t}_b$ in (4.1) by

$$\hat{t} = p_0^{-1}(2\beta_{\hat{m}} + \theta + \epsilon)\log b. \tag{4.5}$$

The next theorem is an analogue of Theorem 1 for the case of estimated $\mathbf{\Sigma}$. It shows that in the region $\mathcal{D}_1^0(\theta)$ of (β, r)-values the classification rule $\hat{\psi}_b$ defined by (2.4), (4.1), and (4.5) provides successful classification.

Theorem 2. *Let $\mathbf{X}^{(1)} = (\mathbf{X}_j^{(1)})_{j\in[\![n]\!]}$ and $\mathbf{X}^{(2)} = (\mathbf{X}_j^{(2)})_{j\in[\![n]\!]}$ be training samples of size n in the $([b^{1-\beta}], 2r\log b)$-sparse block-diagonal normal model, with n and b related through (2.5) for a given number $\theta \in (0,1)$, and let $\mathbf{X}_0$ be a new observation to be classified. Assume that the covariance matrix $\mathbf{\Sigma}$ is unknown. Then, for all $(\beta, r) \in \mathcal{D}_1^0(\theta)$, the classifier $\hat{\psi}_b$ defined by (2.4), (4.1), and (4.5)*

satisfies

$$\lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathcal{R}(\hat{\psi}_b) = 0.$$

Proof. The proof of Theorem 2 is similar to that of Theorem 1 yet more technical due the presence of the estimator $\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}$ of $\boldsymbol{\Sigma}_{[k]}^{-1}$, $k \in [\![b]\!]$, in the definition of $\hat{\psi}_b$. For a number $\theta \in (0,1)$, let (β, r) be an arbitrary point in the region $\mathcal{D}_1^0(\theta)$. We need to show that

$$\lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}_{\Pi_2}(\hat{\psi}_b) = 0, \quad \lim_{b\to\infty} \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{E}_{\Pi_1}(1-\hat{\psi}_b) = 0, \tag{4.6}$$

where, as in the proof of Theorem 1, $\mathbf{E}_{\Pi_i}$ denotes the expectation with respect to the joint distribution of $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}_0$ when $\mathbf{X}_0 \sim \Pi_i$ for $i \in [\![2]\!]$. As the proofs of both relations in (4.6) go along the same lines, we shall only prove the first one. Using the notation $\hat{\Delta}_{k,b}^2 = \widehat{\boldsymbol{\mu}}_{[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \widehat{\boldsymbol{\mu}}_{[k]}$, for $k \in [\![b]\!]$, $b = 2, 3, \ldots$, we put

$$\hat{V}_k = \hat{V}_{k,b} = \mathbf{X}_{0,[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} \widehat{\boldsymbol{\mu}}_{[k]} - (1/2)\hat{\Delta}_{k,b}^2. \tag{4.7}$$

Recall also the random variables V_k defined in (3.17) that, under Π_2, satisfy $V_k \sim N_{p_0}(\Delta_{k,b}^2/2, \Delta_{k,b}^2)$ for all $k \in [\![b]\!]$, $b = 2, 3, \ldots$. Then, we can express $\hat{\psi}_b$ as

$$\hat{\psi}_b = \mathbb{I}\left\{ \sum_{k=1\,:\,\hat{\omega}_k=1}^{b} \hat{V}_k \leqslant 0 \right\},$$

where $\hat{\omega}_k = \hat{\omega}_k(\beta_{\hat{m}})$ is the kth component of $\hat{\boldsymbol{\omega}}(\beta_{\hat{m}})$ in (4.3) and, cf. (3.19),

$$\begin{aligned} \left\{ \sum_{k=1\,:\,\hat{\omega}_k=1}^{b} \hat{V}_k \leqslant 0 \right\} = \Bigg\{ & \sum_{k=1\,:\,\omega_k=1}^{b} (V_k - \mathbf{E}_{\Pi_2}(V_k)) + \sum_{k=1\,:\,\omega_k=1}^{b} (\hat{V}_k - V_k) \\ & + \sum_{k=1\,:\,\omega_k=0,\hat{\omega}_k=1}^{b} \hat{V}_k - \sum_{k=1\,:\,\omega_k=1,\hat{\omega}_k=0}^{b} \hat{V}_k \leqslant -\frac{1}{2} \sum_{k=1\,:\,\omega_k=1}^{b} \Delta_{k,b}^2 \Bigg\}, \end{aligned} \tag{4.8}$$

with the term $\sum_{k=1\,:\,\omega_k=1}^{b} \Delta_{k,b}^2$ obeying relation (3.20). As seen from the next result, the main contribution to $\sum_{k=1\,:\,\hat{\omega}_k=1}^{b} \hat{V}_k$ is made by $\sum_{k=1\,:\,\omega_k=1}^{b} V_k$.

Lemma 3. *For all $\theta \in (0,1)$, uniformly in $(\beta, r) \in \mathcal{D}_1^0(\theta)$, as $b \to \infty$*

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=1}^{b} (\hat{V}_k - V_k) = o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right), \tag{4.9}$$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=1,\hat{\omega}_k=0}^{b} \hat{V}_k = o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right), \tag{4.10}$$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=0,\hat{\omega}_k=1}^{b} \hat{V}_k = o_{\mathbf{P}_{\Pi_2}}\left(b^{1-\beta-\theta}\log b\right). \tag{4.11}$$

The proof of Lemma 3 is given in Section 7. Now, with Lemma 3 available, the rest of the proof of Theorem 2 resembles that of Theorem 1 after Lemma 2. □

Remark 2. Inspection of the proof of Theorem 2 shows that it can be extended to the case of blocks $\mathbf{\Sigma}_{[1]}, \dots, \mathbf{\Sigma}_{[b]}$ of different sizes $p_1 \times p_1, \dots, p_b \times p_b$, where $p_k \geqslant 3$, $k \in [\![b]\!]$, are uniformly bounded integers such that $\sum_{k=1}^{b} p_k = p$. Indeed, by (4.2) and relations (7.2)–(7.4), for large b, the tails of the statistics $\hat{T}_{k,b} \sim F_{p_k,2n-p_k}(n\Delta_{k,b}^2)$, $k \in [\![b]\!]$, are not essentially affected by the change of the numerator degree of freedom p_k and the denominator degree of freedom $(2n-p_k)$ by a finite (independent of n) integer number. In this case, the sequence of numbers τ_b which defines the quantities $v_j = b^{1-\beta_j}/\tau_b$, $j \in [\![m]\!]$, in (3.12) and (4.4) is to be chosen to have $\tau_b = o\left(b^{\epsilon/2} \log^{1-\bar{p}/2} b\right)$ as $b \to \infty$, where $\bar{p} = \max(p_1, \dots, p_b)$.

4.2. *Classification rule in the region* $\mathcal{D}_2^0(\boldsymbol{\theta})$

We shall now discuss a suitable choice for the threshold $\hat{t}$ (for notational simplicity, we suppress the dependence of $\hat{t}$ on b) of the classifier $\hat{\psi}_b$ defined by (2.4) and (4.1). By the sparsity assumption on the model, only $s = [b^{1-\beta}]$ statistics among $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ have a noncentral F distribution, whereas the remaining $(b-s) = b + o(b)$ statistics are centrally F distributed. In view of (4.2), for all $k \in [\![b]\!]$ and all large enough b, a central random variable $p_0 F_{p_0,2n-p_0}(0)$ is close in distribution to a $\chi^2_{p_0}(0)$, and a noncentral random variable $p_0 F_{p_0,2n-p_0}(n\Delta_{k,b}^2)$ is close in distribution to a $\chi^2_{p_0}(n\Delta_{k,b}^2)$. Therefore, similar to the case of known $\mathbf{\Sigma}$, we may consider the problem of testing the hypotheses

$$\begin{aligned} \mathbf{H}_0 &: t_1, \dots, t_b \overset{iid}{\sim} F_{p_0,2n-p_0}(0), \\ \mathbf{H}_{1,b} &: t_1, \dots, t_b \overset{iid}{\sim} (1-\varepsilon_b) F_{p_0,2n-p_0}(0) + \varepsilon_b F_{p_0,2n-p_0}(\gamma_b), \end{aligned}$$

where $\varepsilon_b = b^{-\beta}$ for $0 < \beta < 1$ and $\gamma_b = 2r \log b$ for $0 < r < 1$, by means of the weighted Kolmogrov–Smirnov test statistics.

To this end, transform the statistics t_k to the uniformly $U(0,1)$ distributed under $\mathbf{H}_0$ statistics

$$u_k = 1 - F_{p_0,2n-p_0}(t_k; 0), \quad k \in [\![b]\!],$$

where $F_{\nu_1,\nu_2}(x;\gamma) = \mathbf{P}(F_{\nu_1,\nu_2}(\gamma) \leqslant x)$, $x \in \mathbb{R}$. In terms of a common cdf $F(u)$ of the u_k's, the problem of testing $\mathbf{H}_0$ versus $\mathbf{H}_{1,b}$ is equivalent to that of testing

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

versus a sequence of *upper-tailed* alternatives

$$\mathbb{H}_{1,b} : F(u) = F_0(u) + \varepsilon_b \left((1-u) - F_{p_0,2n-p_0}(F^{-1}_{p_0,2n-p_0}(1-u;0);\gamma_b)\right) > F_0(u).$$

Similar to the problem of testing $\mathcal{H}_0$ versus $\mathcal{H}_{1,b}$, the hypotheses $\mathbb{H}_0$ and $\mathbb{H}_{1,b}$ are separated by a weighted Kolmogorov–Smirnov test statistic

$$\mathbf{D}_b^+(q) = \max_{1 \leqslant k \leqslant [\alpha_0 b]} \frac{\sqrt{b}(k/b - u_{(k)})}{q(k/b)},$$

where q is one of the weight functions q_σ, $\sigma = -1/2, 0, 1/2$, of our interest, as defined in (3.25), or function $q_{1/4}$ as in (3.27).

Similar to the choice of threshold $\tilde{t} = \tilde{t}_q$ in Section 3.3, we now choose the threshold $\hat{t} = \hat{t}_q$ to be that order statistic of the sample $\{\hat{T}_{k,b} : k \in [\![b]\!]\}$ for which the objective function under the maximum sign in $\mathbf{D}_b^+(q)$ based on the translated observations $U_{k,b} = 1 - F_{p_0,2n-p_0}(\hat{T}_{k,b}; 0)$, $k \in [\![b]\!]$, in maximized. Namely, using relation (4.2) and the arguments that have led us to the threshold $\tilde{t}_q$ in (3.29), we define the threshold $\hat{t}_q = \hat{t}_q(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ by

$$\hat{t}_q = F^{-1}_{p_0,2n-p_0}(1 - U_{(\hat{k}_q)}; 0).$$

where the index $1 \leqslant \hat{k}_q \leqslant [\alpha_0 b]$ is chosen as, cf. (3.28),

$$\hat{k}_q = \operatorname*{arg\,max}_{1 \leqslant k \leqslant [\alpha_0 b]} \frac{\sqrt{b}(k/b - U_{(k)})}{q(k/b)},$$

and $U_{(k)}$ is the kth order statistic of the sample $\{U_{k,b} : k \in [\![b]\!]\}$. Alternatively, we can write

$$\hat{t}_q = \hat{T}_{(b+1-\hat{k}_q)}. \tag{4.12}$$

If $\hat{T}_{k,b} > \hat{t}_q$, the kth block is rendered useful and hence is retained to contribute to $\hat{\psi}_b$. The classifier $\hat{\psi}_b$ defined by (2.4), (4.1), and (4.12) is fully adaptive in the parameters of the model.

5. Numerical study

Let $\tilde{\psi}_b^{(1)}$ and $\tilde{\psi}_b^{(2)}$ be the classifiers defined by (2.3) and (3.4) with the thresholds as in (3.14) and (3.29), respectively. Also, let $\hat{\psi}_b^{(1)}$ and $\hat{\psi}_b^{(2)}$ be the classifiers given by (2.4) and (4.1) with the thresholds specified by (4.5) and (4.12), respectively.

In the regions $\mathcal{D}_1(\theta)$ and $\mathcal{D}_1^0(\theta)$, with $\theta = 0.5$, $p_0 = 3$, $b = 10^3$ and $\theta = 0.5$, $p_0 = 5$, $b = 10^4$, and various configurations of the parameters β and r, the estimated risks $\mathcal{R}(\tilde{\psi}_b^{(1)})$ and $\mathcal{R}(\hat{\psi}_b^{(1)})$ obtained by averaging over 100 independent cycles of simulations, were found to be zero up to three decimal places. Note that the choice of $\theta = 0.5$ leads to an interesting case when $n = b^{1/2}$ is much smaller than b for large b.

We now present some simulation results related to high-dimensional classification in the regions $\mathcal{D}_2(\theta)$ and $\mathcal{D}_2^0(\theta)$, where feature selection is impossible. Table 1 gives the numerical summary of the performance of the classifiers $\tilde{\psi}_b^{(2)}$

TABLE 1
Estimated risk $\mathcal{R}(\psi) = \frac{1}{2}\mathbf{E}_{\Pi_2}(\psi) + \frac{1}{2}\mathbf{E}_{\Pi_1}(1-\psi)$ *for* $\psi = \tilde{\psi}_b^{(2)}$ *and* $\psi = \hat{\psi}_b^{(2)}$

Weight function	$\mathcal{R}_{\text{est}}(\tilde{\psi}_b^{(2)})$	$\mathcal{R}_{\text{est}}(\hat{\psi}_b^{(2)})$
$q_{-1/2}(u) = \sqrt{u(1-u)}$	0.2786	0.3039
$q_0(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$	0.1579	0.1721
$q_{1/2}(u) = \sqrt{u(1-u)\log\log(1/(u(1-u)))}$	0.2417	0.2418
$q_{1/4}(u) = (u(1-u))^{1/4}$	0.1963	0.1987
All blocks	0.2085	0.2122
Only informative blocks	0.0014	0.0018

(when $\mathbf{\Sigma}$ is known) and $\hat{\psi}_b^{(2)}$ (when $\mathbf{\Sigma}$ is unknown) in the region where variable selection is impossible for four different choices of weight function q in (3.29) and (4.12). To run simulations, we picked $b = 10^4$, $\theta = 0.5$, $p_0 = 3$, $\beta = 0.375$, $r = 0.25$, and averaged the results over 100 simulation cycles. It is seen that, in the region where variable selection is impossible, the classifiers $\tilde{\psi}_b^{(2)}$ and $\hat{\psi}_b^{(2)}$, for which the selection of useful blocks is done by means of weighted Kolmogorov–Smirnov thresholding, works best when the EFKP weight function $q_0(u)$ is used. At the same time, the SDP weight function $q_{-1/2}(u)$ employed by Donoho and Jin [10] in a similar context does not appear to be a good choice.

6. Concluding remarks

This work was inspired by the need for accurate parsimonious classification procedures in sparse high-dimensional settings. Instead of imposing the usual (and often unrealistic) assumption of mutual independence of feature variables, we suggest a different approach by allowing some local dependence, which is modelled by means of a block-diagonal covariance matrix with blocks of possibly different sizes. The assumption on a block-diagonal covariance matrix allows for a variety of within-block covariance structures. The proposed framework has some definite advantages. In particular, it enables to obtain an accurate classifier with incorporated group-wise adaptive feature selector.

The sparse classification model at hand is described by several known parameters, including θ and p_0, and two unknown parameters β and r. For each of the two assumptions regarding the covariance matrix $\mathbf{\Sigma}$ (known or unknown), depending on the location of the point (β, r) inside the classification region, we have proposed two different classifiers: $\tilde{\psi}_b^{(1)}$ and $\tilde{\psi}_b^{(2)}$ when $\mathbf{\Sigma}$ is known, and $\hat{\psi}_b^{(1)}$ and $\hat{\psi}_b^{(2)}$ when $\mathbf{\Sigma}$ is unknown. In certain subregions of the classification region, the classifiers $\tilde{\psi}_b^{(1)}$ and $\hat{\psi}_b^{(1)}$ were shown to be asymptotically optimal in providing successful classification (see Theorems 1 and 2). For small values of r, when the problem of classification is very difficult, the adaptive procedures $\tilde{\psi}_b^{(2)}$ and $\hat{\psi}_b^{(2)}$ were proposed and studied numerically.

Although all our classifiers are adaptive, that is, their definitions do not involve β and r, the application of $\tilde{\psi}_b^{(1)}$ and $\tilde{\psi}_b^{(1)}$ requires that the respective assumptions $r > \beta$ and $r > \beta + \theta/2$ be valid. If one cannot guarantee that r

is large enough to use $\tilde{\psi}_b^{(1)}$ and $\hat{\psi}_b^{(1)}$, we would recommend to use the classifier $\tilde{\psi}_b^{(2)}$ in case of known $\mathbf{\Sigma}$ and the classifier $\hat{\psi}_b^{(2)}$ in case of estimated $\mathbf{\Sigma}$. If we are in a position to assume that $r > 1$, then the classifiers $\tilde{\psi}_b^{(1)}$ and $\hat{\psi}_b^{(1)}$, that work well for both equally-sized and unequally-sized blocks, should be used.

7. Proof of Lemmas

Proof of Lemma 1. The proof is largely based on the fact that if index $m_0 \in [\![M-1]\!]$ is such that $\beta \in (\beta_{m_0}, \beta_{m_0+1}]$, then with high probability $\tilde{m} \geqslant m_0$, where $\tilde{m}$ is given by (3.12). To verify this fact, Bernstein's inequality will be used.

Fact 1. Bernstein's inequality. *If $\mathbb{X}_1, \ldots, \mathbb{X}_b$, $b \in \mathbb{N}$, are independent random variables such that for all $i \in [\![b]\!]$ and for some $H > 0$*

$$\mathbf{E}(\mathbb{X}_i) = 0 \quad \textit{and} \quad |\mathbf{E}(\mathbb{X}_i^m)| \leqslant \frac{\mathbf{E}(\mathbb{X}_i^2)}{2} H^{m-2} m! < \infty, \quad m = 2, 3, \ldots, \tag{7.1}$$

then

$$\max\left\{\mathbf{P}\left(\mathbb{S}_b \geqslant t\right), \mathbf{P}\left(\mathbb{S}_b \leqslant -t\right)\right\} \leqslant \begin{cases} \exp\left(-t^2/4\mathbb{D}_b^2\right) & \textit{if} \quad 0 \leqslant t \leqslant \mathbb{D}_b^2/H, \\ \exp\left(-t/4H\right) & \textit{if} \quad t \geqslant \mathbb{D}_b^2/H, \end{cases}$$

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

Remark 3. (i) This version of Bernstein's inequality can be found on pages 162–166 of [3] and in Section 2.2 of [19]. (ii) For independent random variables $\mathbb{X}_1, \ldots, \mathbb{X}_b$ with the properties $\mathbf{E}(\mathbb{X}_i) = 0$ and $|\mathbb{X}_i| \leqslant L$, $i \in [\![b]\!]$, for some $L > 0$, the Bernstein condition (7.1) holds with $H = L/3$. (iii) Below Bernstein's inequality will be applied for the case of $t \geqslant \mathbb{D}_b^2/H$.

The following asymptotics for the chi-square tail probabilities will also be of great help: for any $\nu \geqslant 1$ as $b \to \infty$

$$\mathbf{P}\left(\chi_\nu^2(0) > 2s\log b\right) = O\left(b^{-s}\log^{\nu/2-1} b\right), \ 0 < s < \infty, \tag{7.2}$$

$$\mathbf{P}\left(\chi_\nu^2(2r\log b) > 2s\log b\right) = O\left(b^{-(\sqrt{s}-\sqrt{r})^2}\log^{-1/2} b\right), \ 0 < r < s < \infty, \tag{7.3}$$

$$\mathbf{P}\left(\chi_\nu^2(2r\log b) \leqslant 2s\log b\right) = O\left(b^{-(\sqrt{s}-\sqrt{r})^2}\log^{-1/2} b\right), \ 0 < s < r < \infty. \tag{7.4}$$

The first two relations follow from formula (5.5) in Donoho and Jin [9]. The third one can be obtained by using relations (1) and (2) in Han [14] which 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.

Consider the selector $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}) = (\tilde{\omega}_1(\beta_{\tilde{m}}), \ldots, \tilde{\omega}_b(\beta_{\tilde{m}}))$ given by (3.9)–(3.12). 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

$$\sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\, d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}), \boldsymbol{\omega})$$

$$\leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\left(d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\boldsymbol{\omega})|\tilde{m}\geqslant m_0\right)\mathbf{P}(\tilde{m}\geqslant m_0)$$
$$+\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\left(d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\boldsymbol{\omega})|\tilde{m}<m_0\right)\mathbf{P}(\tilde{m}<m_0)$$
$$=: I_{1,b}+I_{2,b}, \tag{7.5}$$

where $\boldsymbol{\omega}=(\omega_k)_{k\in[\![b]\!]}=(\mathbb{I}(\Delta_{k,b}^2\neq 0))_{k\in[\![b]\!]}$.

Consider the term $I_{1,b}$. When $\tilde{m}\geqslant m_0$, by the triangle inequality and the definition of $\tilde{m}$ in (3.12), for all $(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}$

$$d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\boldsymbol{\omega})\leqslant d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\tilde{\boldsymbol{\omega}}(\beta_{m_0}))+d(\tilde{\boldsymbol{\omega}}(\beta_{m_0}),\boldsymbol{\omega})\leqslant v_{m_0}+d(\tilde{\boldsymbol{\omega}}(\beta_{m_0}),\boldsymbol{\omega}),$$

where as $b\to\infty$

$$b^{\beta-1}v_{m_0}=O(\tau_b^{-1})=o(1). \tag{7.6}$$

Next, for any non-negative random variable Y, for which the expectations below are well defined, one has $\mathbf{E}(Y|B)\mathbf{P}(B)\leqslant\mathbf{E}(Y)$, and hence we can write

$$\begin{aligned} I_{1,b}&\leqslant b^{\beta-1}v_{m_0}+b^{\beta-1}\mathbf{E}\,d(\tilde{\boldsymbol{\omega}}(\beta_{m_0}),\boldsymbol{\omega})\\ &\leqslant b^{\beta-1}v_{m_0}+b^{\beta}\mathbf{P}\left(\chi^2_{p_0}(0)>(2\beta_{m_0}+\epsilon)\log b\right)\\ &+\mathbf{P}\left(\chi^2_{p_0}(2r\log b)\leqslant(2\beta_{m_0}+\epsilon)\log b\right)=: b^{\beta-1}v_{m_0}+N_{1,b}+N_{2,b}. \end{aligned} \tag{7.7}$$

For the term $N_{1,b}$, using relations (3.8), (3.10), and (7.2), we obtain

$$N_{1,b}=O\left(b^{\beta-\beta_{m_0}-\epsilon/2}\log^{p_0/2-1}b\right)=O\left(b^{\delta-\epsilon/2}\log^{p_0/2-1}b\right)=o(1). \tag{7.8}$$

For the term $N_{2,b}$, by relation (7.4) and the fact that for all sufficiently large b (recall that $\epsilon=\epsilon_b\to 0$ as $b\to\infty$) one has $r>\beta_{m_0}+\epsilon/2$, we obtain as $b\to\infty$

$$N_{2,b}=O\left(b^{-\left(\sqrt{r}-\sqrt{\beta_{m_0}+\epsilon/2}\right)^2}\log^{-1/2}b\right)=o(1). \tag{7.9}$$

Therefore, in view of (7.7)–(7.9), uniformly in $(1-\theta)/2<\beta<1-\theta$ and $\beta<r<4$

$$\begin{aligned} I_{1,b}&=\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\left(d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\boldsymbol{\omega})|\tilde{m}\geqslant m_0\right)\mathbf{P}(\tilde{m}\geqslant m_0)\\ &\leqslant b^{\beta-1}v_{m_0}+\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\,d(\tilde{\boldsymbol{\omega}}(\beta_{m_0}),\boldsymbol{\omega})\\ &\leqslant O(\tau_b^{-1})+N_{1,b}+N_{2,b}=o(1). \end{aligned} \tag{7.10}$$

It remains to show that, uniformly in $(1-\theta)/2<\beta<1-\theta$ and $\beta<r<4$, it is also true for the term $I_{2,b}$ in (7.5) that $I_{2,b}=o(1)$. We have

$$I_{2,b}=\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\left(d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}),\boldsymbol{\omega})|\tilde{m}<m_0\right)\mathbf{P}(\tilde{m}<m_0)\leqslant b^{\beta}\mathbf{P}(\tilde{m}<m_0).$$

Now, using Fact 1, we show that $\mathbf{P}(\tilde{m} < m_0)$ is small yielding $I_{2,b} = o(1)$ as $b \to \infty$. Indeed, recalling (3.12) and writing $\tilde{T}_i$ instead of $\tilde{T}_{i,b}$ for $i \in [\![b]\!]$, we have

$$\begin{aligned}\mathbf{P}(\tilde{m} < m_0) &= \sum_{k=1}^{m_0-1} \mathbf{P}(\tilde{m} = k) = \sum_{k=1}^{m_0-1} \mathbf{P}\left(\exists\, j \in [\![k]\!] : d(\tilde{\boldsymbol{\omega}}(\beta_{k+1}), \tilde{\boldsymbol{\omega}}(\beta_j)) > v_j\right) \\ &\leqslant \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(d(\tilde{\boldsymbol{\omega}}(\beta_{k+1}), \tilde{\boldsymbol{\omega}}(\beta_j)) > v_j\right) \\ &= \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{b} |\tilde{\omega}_i(\beta_{k+1}) - \tilde{\omega}_i(\beta_j)| > v_j\right) \\ &= \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{b} \mathbb{I}\left\{(2\beta_j + \epsilon)\log b < \tilde{T}_i \leqslant (2\beta_{k+1} + \epsilon)\log b\right\} > v_j\right). \end{aligned} \tag{7.11}$$

Now, introducing the events

$$A_i = A_{i,j,k+1,b} = \left\{(2\beta_j + \epsilon)\log b < \tilde{T}_i \leqslant (2\beta_{k+1} + \epsilon)\log b\right\}, \quad i \in [\![b]\!],$$

we obtain from (7.11) that

$$\begin{aligned}\mathbf{P}(\tilde{m} < m_0) &\leqslant \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{b} \mathbb{I}\{A_i\} > v_j\right) \\ &= \sum_{k=1}^{m_0-1} \sum_{j=1}^{k} \mathbf{P}\left(\sum_{i=1}^{b} \mathbb{X}_i > v_j - \sum_{i=1}^{b} \mathbf{P}(A_i)\right), \end{aligned} \tag{7.12}$$

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

$$\mathbb{X}_i = \mathbb{X}_{i,j,k+1,b} \stackrel{\text{def}}{=} \mathbb{I}\{A_i\} - \mathbf{P}(A_i), \quad i \in [\![b]\!]. \tag{7.13}$$

To apply Bernstein's inequality to the term $\mathbf{P}\left(\sum_{i=1}^{b} \mathbb{X}_i > v_j - \sum_{i=1}^{b} \mathbf{P}(A_i)\right)$ on the right-hand side of (7.12), we first show that $\sum_{i=1}^{b} \mathbf{P}(A_i) = o(v_j)$ as $b \to \infty$. Using (7.2) and (7.4) and recalling that $[b^{1-\beta}]$ statistics among $\{\tilde{T}_i : i \in [\![b]\!]\}$ follow a noncentral chi-square distribution and the remaining statistics have a central chi-square distribution, we get for all $j \in [\![k]\!]$ and $k \in [\![m_0 - 1]\!]$ as $b \to \infty$

$$\begin{aligned}\sum_{i=1}^{b} \mathbf{P}(A_i) &= \sum_{i=1:\omega_i=0}^{b} \mathbf{P}(A_i) + \sum_{i=1:\omega_i=1}^{b} \mathbf{P}(A_i) \\ &\leqslant b\mathbf{P}\left(\chi^2_{p_0}(0) > (2\beta_j + \epsilon)\log b\right) + [b^{1-\beta}]\mathbf{P}\left(\chi^2_{p_0}(2r\log b) \leqslant (2\beta_{k+1} + \epsilon)\log b\right) \\ &= O\left(b^{1-(\beta_j+\epsilon/2)} \log^{p_0/2-1} b\right) + O\left(b^{1-\beta-(\sqrt{r}-\sqrt{\beta_{k+1}+\epsilon/2})^2} \log^{-1/2} b\right), \end{aligned}$$

where $\beta_j \leqslant \beta_{m_0-1} < \beta < r$ and $\beta_{k+1} \leqslant \beta_{m_0} < \beta < r$. From this, noting that $b^{1-(\beta_j+\epsilon/2)} \gg b^{1-\beta-(\sqrt{r}-\sqrt{\beta_{k+1}+\epsilon/2})^2}$ for all large enough b gives

$$\sum_{i=1}^{b} \mathbf{P}(A_i) = O\left(b^{1-(\beta_j+\epsilon/2)} \log^{p_0/2-1} b\right), \quad b \to \infty. \tag{7.14}$$

Since by definition $v_j = \tau_b^{-1} b^{1-\beta_j}$ and $\tau_b = o\left(b^{\epsilon/2}\log^{1-p_0/2} b)\right)$, it now follows from (7.14) that for all $j \in [\![k]\!]$ one has $\sum_{i=1}^b \mathbf{P}(A_i) = o(v_j)$ as $b \to \infty$ and hence as $b \to \infty$

$$v_j - \sum_{i=1}^b \mathbf{P}(A_i) = v_j(1 + o(1)), \quad j \in [\![k]\!], \ \ k \in [\![m_0 - 1]\!].$$

Also, since the variance of a random variable taking values in $\{0, 1\}$ is smaller than its expectation, we have by (7.14) and by the independence of $\tilde{T}_1, \dots, \tilde{T}_b$ that as $b \to \infty$

$$\begin{aligned}\mathbf{Var}\left(\sum_{i=1}^b \mathbb{X}_i\right) &= \sum_{i=1}^b \mathbf{Var}\left(\mathbb{I}\{A_i\}\right) \leqslant \sum_{i=1}^b \mathbf{E}\left(\mathbb{I}\{A_i\}\right) = \sum_{i=1}^b \mathbf{P}(A_i) \\ &= O\left(b^{1-(\beta_j+\epsilon/2)}\log^{p_0/2-1} b\right).\end{aligned}$$

Thus, for the random variables $\mathbb{X}_1, \dots, \mathbb{X}_b$ defined in (7.13) we have $|\mathbb{X}_i| \leqslant 2$ and $\mathbf{E}(\mathbb{X}_i) = 0$ for $i \in [\![b]\!]$, and hence for all $j \in [\![k]\!]$ and $k \in [\![m_0 - 1]\!]$

$$\mathbb{D}_b^2 = \sum_{i=1}^b \mathbf{E}(\mathbb{X}_i^2) = \mathbf{Var}\left(\sum_{i=1}^b \mathbb{X}_i\right) = O\left(b^{1-(\beta_j+\epsilon/2)}\log^{p_0/2-1} b\right), \quad b \to \infty.$$

Therefore the application of Bernstein's inequality stated in Fact 1 with $t = v_j(1 + o(1))$, Remark 3, and the fact that $\beta_j \leqslant \beta_{m_0-1}$ for all $j \in [\![k]\!]$ and $k \in [\![m_0 - 1]\!]$ give

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

From this and (7.12) we deduce that for all large enough b

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

and hence

$$I_{2,b} \leqslant b^\beta \mathbf{P}(\tilde{m} < m_0) \leqslant b^\beta M^2 \exp\left(-\frac{b^{1-\beta_{m_0-1}}}{4\tau_b}\right) = o(1), \tag{7.15}$$

where the last equality is due to the fact that $1 - \beta_{m_0-1}$ is separated from zero, which follows from the assumptions $(1 - \theta)/2 < \beta < 1 - \theta$ and $\beta_{m_0-1} < \beta_{m_0} < \beta \leqslant \beta_{m_0+1}$.

Now, combining (7.5), (7.10) and (7.15), we obtain

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} b^{\beta-1}\mathbf{E}\, d(\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}}), \boldsymbol{\omega}) = I_{1,b} + I_{2,b} = o(1), \quad b \to \infty,$$

uniformly in $(\beta, r) \in \mathcal{D}_1(\theta)$ for all $0 < \theta < 1$. This shows that the selector $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ provides almost full selection in the region $\mathcal{D}_1(\theta)$ for all $0 < \theta < 1$. □

Proof of Lemma 2. Throughout the proof, θ is an arbitrary number in the interval $(0,1)$ and (β, r) is an arbitrary point in the region $\mathcal{D}_1(\theta)$. For brevity, we shall write $\Delta_{k,b}^2$ and $\tilde{\Delta}_{k,b}^2$ as Δ_k^2 and $\tilde{\Delta}_k^2$ for $k \in [\![b]\!]$, $b = 2, 3, \ldots$. Let us first check the validity of (3.21). For all $k \in [\![b]\!]$ and $b = 2, 3, \ldots$, consider

$$\tilde{V}_k - V_k = \mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \left(\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]} \right) - \frac{1}{2} \left(\tilde{\Delta}_k^2 - \Delta_k^2 \right), \tag{7.16}$$

where $\Delta_k^2 = \boldsymbol{\mu}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \boldsymbol{\mu}_{[k]}$ and $\tilde{\Delta}_k^2 = \widehat{\boldsymbol{\mu}}_{[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \widehat{\boldsymbol{\mu}}_{[k]}$. It is easy to see that for all $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$

$$\mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k) = -\frac{p_0}{2n}, \quad k \in [\![b]\!], \tag{7.17}$$

and hence (recall that $n \sim b^\theta$ and $\#\{k \in [\![b]\!] : \omega_k = 1\} = [b^{1-\beta}]$)

$$\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k) = -\frac{p_0}{2} b^{1-\beta-\theta}(1 + o(1)), \quad b \to \infty.$$

Therefore, by the triangle and Chebyshev's inequalities, using the block-wise independence of the data, for any $\varepsilon > 0$ and all large enough b

$$\begin{aligned}
&\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1\,:\,\omega_k=1}^{b} (\tilde{V}_k - V_k) \right| \geqslant \varepsilon b^{1-\beta-\theta} \log b \right) \\
&\leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1\,:\,\omega_k=1}^{b} \left\{ (\tilde{V}_k - V_k) - \mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k) \right\} \right| \geqslant \frac{\varepsilon}{2} b^{1-\beta-\theta} \log b \right) \\
&\leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}} \frac{\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k)^2}{((\varepsilon/2) b^{1-\beta-\theta} \log b)^2}.
\end{aligned} \tag{7.18}$$

Consider the numerator on the right side of (7.18). Using relation (7.16), the inequalities $(a-b)^2 \leqslant 2(a^2+b^2)$ and $(\mathbf{u}^\top \mathbf{v})^2 \leqslant \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$, and the independence of $\mathbf{X}_{0,[k]}$ and $\widehat{\boldsymbol{\mu}}_{[k]}$,

$$\begin{aligned}
&\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k)^2 \\
&\leqslant 2 \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2} \left(\mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \left(\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]} \right) \right)^2 + \frac{1}{2} \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E} \left(\tilde{\Delta}_k^2 - \Delta_k^2 \right)^2 \\
&\leqslant 2 \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2} \left(\mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1} \mathbf{X}_{0,[k]} \right) \mathbf{E} \left((\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]})^\top \boldsymbol{\Sigma}_{[k]}^{-1} (\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]}) \right) \\
&\quad + \frac{1}{2} \sum_{k=1\,:\,\omega_k=1}^{b} \left(\frac{1}{n^2} \mathbf{E} \left\{ (\chi_{p_0}^2(n\Delta_k^2))^2 \right\} - \frac{2\Delta_k^2}{n} \mathbf{E} \left\{ \chi_{p_0}^2(n\Delta_k^2) \right\} + (\Delta_k^2)^2 \right)
\end{aligned}$$

From this, using the identity $\mathbf{X}^\top\mathbf{A}\mathbf{X} = \mathrm{Tr}(\mathbf{A}\mathbf{X}\mathbf{X}^\top)$, the fact $\mathbf{E}_{\Pi_2}\left(\mathbf{X}_{0,[k]}\mathbf{X}_{0,[k]}^\top\right) = \mathbf{\Sigma}_{[k]}$, and the equalities $\mathbf{E}(\chi^2_\nu(\lambda)) = \nu+\lambda$ and $\mathbf{Var}(\chi^2_\nu(\lambda)) = 2(\nu+2\lambda)$, we may continue

$$\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k)^2$$

$$\leqslant 2\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\mathrm{Tr}\left(\mathbf{\Sigma}_{[k]}^{-1}\mathbf{X}_{0,[k]}\mathbf{X}_{0,[k]}^\top\right)\mathbf{E}\,\mathrm{Tr}\left(\mathbf{\Sigma}_{[k]}^{-1}(\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]})(\widehat{\boldsymbol{\mu}}_{[k]} - \boldsymbol{\mu}_{[k]})^\top\right)$$

$$+\frac{1}{2}\sum_{k=1\,:\,\omega_k=1}^{b}\left(\frac{2p_0 + 4n\Delta_k^2 + (p_0+n\Delta_k^2)^2}{n^2} - \frac{2\Delta_k^2(p_0+n\Delta_k^2)}{n} + (\Delta_k^2)^2\right)$$

$$= 2\sum_{k=1\,:\,\omega_k=1}^{b}\frac{p_0^2}{n} + \frac{1}{2}\sum_{k=1\,:\,\omega_k=1}^{b}\left(\frac{4\Delta_k^2}{n} + \frac{2p_0}{n^2} + \frac{p_0^2}{n^2}\right),$$

uniformly in $(\boldsymbol{\mu},\mathbf{\Sigma}) \in \mathbf{M}_{b,\beta,r}$. Therefore, as $b\to\infty$

$$\sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}(\tilde{V}_k - V_k)^2 = O\left(b^{1-\beta-\theta}\right) + O\left(b^{1-\beta-2\theta}\log b\right)$$

$$= O\left(b^{1-\beta-\theta}\right). \tag{7.19}$$

The combination of (7.18) and (7.19) now yields that for any $\varepsilon > 0$ and all $(\beta, r) \in D_1(\theta)$, as $b\to\infty$

$$\sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1}^{b}(\tilde{V}_k - V_k)\right| \geqslant \varepsilon b^{1-\beta-\theta}\log b\right) = o(1),$$

and hence (3.21) is proved.

Next, let us verify relation (3.22). For brevity, we shall omit the argument $\beta_{\tilde{m}}$ of the selector $\tilde{\boldsymbol{\omega}}(\beta_{\tilde{m}})$ in (3.11) and write $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_1,\dots,\tilde{\omega}_b)$. First, we have

$$\sup_{(\boldsymbol{\mu},\mathbf{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{E}_{\Pi_2}\left(\sum_{k\,:\,\omega_k=1}^{b}\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k = 0)\right) = o(b^{1-\beta-\theta}\log b), \quad b\to\infty. \tag{7.20}$$

Indeed, using Lemma 1, for all $(\beta, r) \in \mathcal{D}_1(\theta)$, uniformly in $(\boldsymbol{\mu},\mathbf{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, for all large enough b

$$\mathbf{E}_{\Pi_2}\left(\sum_{k\,:\,\omega_k=1}^{b}\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k = 0)\right)$$

$$\leqslant \sum_{k\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left(\mathbf{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\right)^\top\mathbf{E}\left\{\mathbf{\Sigma}_{[k]}^{-1/2}\widehat{\boldsymbol{\mu}}_{[k]}\mathbb{I}\left(n\|\mathbf{\Sigma}_{[k]}^{-1/2}\widehat{\boldsymbol{\mu}}_{[k]}\|^2 \leqslant (2\beta_{\tilde{m}} + \epsilon)\log b\right)\right\}$$

$$\leqslant \frac{C\log b}{n}\sum_{k=1}^{b}\mathbf{P}(|\tilde{\omega}_k - \omega_k| = 1) = \frac{C\log b}{n}\,\mathbf{E}\,d(\tilde{\boldsymbol{\omega}},\boldsymbol{\omega}) = o(b^{1-\beta-\theta}\log b),$$

where $C > 0$ is an absolute constant. From (7.20), by the triangle and Chebyshev's inequalities, using the block-wise independence of the data, for any $\varepsilon > 0$ and all large enough b

$$\begin{aligned}
&\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1}^{b} \tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=0)\right| \geqslant \varepsilon b^{1-\beta-\theta}\log b\right)\\
&\leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1}^{b} \left\{\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=0) - \mathbf{E}_{\Pi_2}(\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=0))\right\}\right|\right.\\
&\left.\geqslant \frac{\varepsilon}{2}\,b^{1-\beta-\theta}\log b\right) \leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \frac{\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left(\tilde{V}_k^2\,\mathbb{I}(\tilde{\omega}_k=0)\right)}{((\varepsilon/2)b^{1-\beta-\theta}\log b)^2}. \qquad (7.21)
\end{aligned}$$

Consider the numerator on the right side of (7.21). Using $(a-b)^2 \leqslant 2(a^2+b^2)$ and $(\mathbf{u}^\top\mathbf{v})^2 \leqslant \|\mathbf{u}\|^2\|\mathbf{v}\|^2$, and noting that $\mathbf{E}_{\Pi_2}\left(\mathbf{X}_{0,[k]}^\top\boldsymbol{\Sigma}_{[k]}^{-1}\mathbf{X}_{0,[k]}\right) = p_0$ and $(2\beta_{\tilde{m}}+\epsilon)\log b \leqslant 3\log b$ when b is large, we obtain for all large enough b

$$\begin{aligned}
&\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left(\tilde{V}_k^2\,\mathbb{I}(\tilde{\omega}_k=0)\right)\\
&= \sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left[\left(\mathbf{X}_{0,[k]}^\top\boldsymbol{\Sigma}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]} - (1/2)\tilde{\Delta}_k^2\right)^2\mathbb{I}(\tilde{\omega}_k=0)\right]\\
&\leqslant 2\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left[\left(\mathbf{X}_{0,[k]}^\top\boldsymbol{\Sigma}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}\right)^2\mathbb{I}(n\tilde{\Delta}_k^2 \leqslant (2\beta_{\tilde{m}}+\epsilon)\log b)\right]\\
&\quad +1/(2n^2)\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}\left[(n\tilde{\Delta}_k^2)^2\,\mathbb{I}(n\tilde{\Delta}_k^2 \leqslant (2\beta_{\tilde{m}}+\epsilon)\log b)\right]\\
&\leqslant (2/n)\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left(\mathbf{X}_{0,[k]}^\top\boldsymbol{\Sigma}_{[k]}^{-1}\mathbf{X}_{0,[k]}\right)\mathbf{E}\left(n\tilde{\Delta}_k^2\,\mathbb{I}(n\tilde{\Delta}_k^2 \leqslant (2\beta_{\tilde{m}}+\epsilon)\log b)\right)\\
&\quad +1/(2n^2)\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}\left((n\tilde{\Delta}_k^2)^2\,\mathbb{I}(n\tilde{\Delta}_k^2 \leqslant (2\beta_{\tilde{m}}+\epsilon)\log b)\right)\\
&\leqslant (2p_0/n)3\log b\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{P}(\tilde{\omega}_k=0) + (3\log b)^2/(2n^2)\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{P}(\tilde{\omega}_k=0)\\
&\leqslant \left(\frac{6p_0\log b}{n} + \frac{9\log^2 b}{2n^2}\right)\sum_{k=1}^{b}\mathbf{P}(|\tilde{\omega}_k-\omega_k|=1) \leqslant \left(\frac{6p_0\log b}{n} + \frac{9\log^2 b}{2n^2}\right)\mathbf{E}\,d(\tilde{\boldsymbol{\omega}},\boldsymbol{\omega}).
\end{aligned}$$

From this, recalling (2.5) and applying Lemma 1, we get for all $(\beta, r) \in \mathcal{D}_1(\theta)$ as $b \to \infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1\,:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left(\tilde{V}_k^2\,\mathbb{I}(\tilde{\omega}_k=0)\right) = o(b^{1-\beta-\theta}\log b). \qquad (7.22)$$

It now follows from (7.21) and (7.22) that all $(\beta, r) \in \mathcal{D}_1(\theta)$ as $b \to \infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1}^{b} \tilde{V}_k\, \mathbb{I}(\tilde{\omega}_k = 0)\right| \geqslant \varepsilon b^{1-\beta-\theta}\log b\right) = o(1),$$

yielding relation (3.22).

It remains to prove (3.23). Aiming again at using Chebyshev's inequality, we first show that $\mathbf{E}_{\Pi_2}\left(\sum_{k=1\,:\,\omega_k=0}^{b} \tilde{V}_k \mathbb{I}(\tilde{\omega}_k = 1)\right) = o\left(b^{1-\beta-\theta}\log b\right)$, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$. To this end, we note that for each k for which $\omega_k = 0$ the statistic $n\tilde{\Delta}_k^2$ has a central chi-square $\chi^2_{p_0}(0)$ distribution with pdf

$$g(x; p_0) = (2^{p_0/2}\Gamma(p_0/2))^{-1} x^{p_0/2-1} e^{-x/2}, \quad \text{for } x > 0.$$

Also, as seen from the proof of Lemma 1, the grid point $\beta_{m_0} = \delta m_0$, which is chosen to have $\beta_{m_0} < \beta \leqslant \beta_{m_0+1}$, satisfies

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

where $M = M_b$ is as in (3.5), that is, the probability $\mathbf{P}(\tilde{m} < m_0)$ decreases to zero at an exponential rate as $b \to \infty$. In particular, for all large enough b

$$\mathbf{P}(\tilde{m} < m_0) \leqslant b^{-2\beta}. \tag{7.23}$$

Therefore, since $\mathbf{E}_{\Pi_2}(\mathbf{X}_{0,[k]}) = \mathbf{0}$ for all those indices k for which $\omega_k = 0$, we have

$$\begin{aligned}
\left|\mathbf{E}_{\Pi_2}\left(\sum_{k=1\,:\,\omega_k=0}^{b} \tilde{V}_k \mathbb{I}(\tilde{\omega}_k = 1)\right)\right| &= \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(n\tilde{\Delta}_k^2 > (2\beta_{\tilde{m}} + \epsilon)\log b)\right) \\
&= \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(n\tilde{\Delta}_k^2 > (2\beta_{\tilde{m}} + \epsilon)\log b, \tilde{m} \geqslant m_0)\right) \\
&\quad + \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(n\tilde{\Delta}_k^2 > (2\beta_{\tilde{m}} + \epsilon)\log b, \tilde{m} < m_0)\right) \\
&\leqslant \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(n\tilde{\Delta}_k^2 > (2\beta_{m_0} + \epsilon)\log b)\right) \\
&\quad + \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(\tilde{m} < m_0)\right).
\end{aligned}$$

Now, applying the Cauchy-Schwarz inequality to $\mathbf{E}\left(n\tilde{\Delta}_k^2\, \mathbb{I}(\tilde{m} < m_0)\right)$, using (7.23) and the asymptotic relation $\int_A^\infty x^\nu e^{-x/2}\, dx \sim 2A^\nu e^{-A/2}$ as $A \to \infty$, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$\left|\mathbf{E}_{\Pi_2}\left(\sum_{k=1\,:\,\omega_k=0}^{b} \tilde{V}_k \mathbb{I}(\tilde{\omega}_k = 1)\right)\right|$$

$$\leqslant \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b}\int_{(2\beta_{m_0}+\epsilon)\log b}^{\infty} x g(x;p_0)\,dx + \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b}\Big(\mathbf{E}\{(n\tilde{\Delta}_k^2)^2\}\mathbf{P}(\tilde{m}<m_0)\Big)^{1/2}$$

$$\leqslant \frac{2}{n}\sum_{k=1\,:\,\omega_k=0}^{b}\frac{(2\beta_{m_0}+\epsilon)^{p_0/2}}{2^{p_0/2}\Gamma(p_0/2)}\,b^{-\beta_{m_0}-\epsilon/2}\log^{p_0/2} b + \frac{1}{2n}\sum_{k=1\,:\,\omega_k=0}^{b}(2p_0+p_0^2)^{1/2}b^{-\beta}$$

$$= O\left(b^{1-\theta-\beta_{m_0}-\epsilon/2}\log^{p_0/2} b\right) + O\left(b^{1-\beta-\theta}\right)$$

$$= O\left(b^{1-\theta-\beta+\delta-\epsilon/2}\log^{p_0/2} b\right) + O(b^{1-\beta-\theta}) = O\left(b^{1-\beta-\theta}\right), \quad b\to\infty,$$

where the last two equalities are due to the fact $0<\beta-\beta_{m_0}<\delta$ and relations (3.8) and (3.10). Thus, as $b\to\infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{E}_{\Pi_2}\left(\sum_{k=1\,:\,\omega_k=0}^{b}\tilde{V}_k\mathbb{I}(\tilde{\omega}_k=1)\right) = o\left(b^{1-\beta-\theta}\log b\right). \tag{7.24}$$

Therefore, by the triangle and Chebyshev's inequalities, using the block-wise independence of the data, for any $\varepsilon>0$ and all large enough b

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=0}^{b}\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=1)\right|\geqslant \varepsilon b^{1-\beta-\theta}\log b\right)$$

$$\leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=0}^{b}\left\{\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=1)-\mathbf{E}_{\Pi_2}(\tilde{V}_k\,\mathbb{I}(\tilde{\omega}_k=1))\right\}\right|\right.$$

$$\left.\geqslant \frac{\varepsilon}{2}\,b^{1-\beta-\theta}\log b\right) \leqslant \sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\frac{\sum_{k=1\,:\,\omega_k=0}^{b}\mathbf{E}_{\Pi_2}\left(\tilde{V}_k^2\,\mathbb{I}(\tilde{\omega}_k=1)\right)}{((\varepsilon/2)b^{1-\beta-\theta}\log b)^2}. \tag{7.25}$$

Consider the numerator on the right side of (7.25). Applying the arguments similar to those that have led us to (7.24), we obtain the relation

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}}\sum_{k=1\,:\,\omega_k=0}^{b}\mathbf{E}_{\Pi_2}\left(\tilde{V}_k^2\,\mathbb{I}(\tilde{\omega}_k=1)\right) = O(b^{1-\beta-\theta}\log b), \quad b\to\infty,$$

which together with (7.25) yields (3.23). Noting that $\theta\in(0,1)$ and $(\beta,r)\in\mathcal{D}_1(\theta)$ were chosen arbitrary completes the proof. □

Proof of Lemma 3. Throughout the proof, θ is an arbitrary number in the interval $(0,1)$ and (β,r) is an arbitrary point in the region $\mathcal{D}_1^0(\theta)$. As in the proof of Lemma 2, for all $k\in[\![b]\!]$ and $b=2,3,\ldots$, we shall write $\Delta_{k,b}^2$, $\tilde{\Delta}_{k,b}^2$, and $\hat{\Delta}_{k,b}^2$ as Δ_k^2, $\tilde{\Delta}_k^2$, and $\hat{\Delta}_k^2$, suppressing the dependence on b. For brevity, we shall also omit the argument $\beta_{\hat{m}}$ of the selector $\hat{\boldsymbol{\omega}}(\beta_{\hat{m}})$ in (4.3) and write $\hat{\boldsymbol{\omega}}=(\hat{\omega}_1,\ldots,\hat{\omega}_b)$. We first verify relation (4.9). Let V_k, $\tilde{V}_k$, and $\hat{V}_k$ be as defined in (3.16), (3.17), and (4.7). We have

$$\sum_{k=1\,:\,\omega_k=1}^{b}\left(\hat{V}_k-V_k\right) = \sum_{k=1\,:\,\omega_k=1}^{b}\left(\hat{V}_k-\tilde{V}_k\right)+\sum_{k=1\,:\,\omega_k=1}^{b}\left(\tilde{V}_k-V_k\right).$$

Therefore, in view of (3.21), it is sufficient to show that as $b \to \infty$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=1}^{b} \left(\hat{V}_k - \tilde{V}_k\right) = o_{\mathbf{P}_{\Pi_2}}(b^{1-\beta-\theta}\log b).$$

Furthermore, due to the identity

$$\hat{V}_k - \tilde{V}_k = \mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]} - \frac{1}{2}\left(\hat{\Delta}_k^2 - \tilde{\Delta}_k^2\right)$$

and relation (3.3), the latter problem is reduced to showing that

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}} \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]} = o_{\mathbf{P}_{\Pi_2}}(b^{1-\beta-\theta}\log b). \quad (7.26)$$

Using Markov's inequality, for any $\varepsilon > 0$

$$\mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right| \geqslant \varepsilon b^{1-\beta-\theta}\log b\right) \leqslant \frac{\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left|\mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right|}{\varepsilon b^{1-\beta-\theta}\log b}. \quad (7.27)$$

Consider the numerator on the right side of (7.27). Using the Cauchy-Schwarz inequality, the identity $\mathbf{X}^{\top}\mathbf{A}\mathbf{X} = \mathrm{Tr}(\mathbf{A}\mathbf{X}\mathbf{X}^{\top})$, and the relations $\mathbf{E}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\right) = \frac{2n-1}{2n-p_0-2}\boldsymbol{\Sigma}_{[k]}^{-1}$, $\mathbf{E}_{\Pi_2}(\hat{\Delta}_k^2) = \frac{(2n-1)(p_0+n\Delta_k^2)}{(2n-p_0-2)n}$, $\mathbf{E}_{\Pi_2}(\tilde{\Delta}_k^2) = \frac{p_0}{n} + \Delta_k^2$, we have, uniformly in $(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$\begin{aligned}
&\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left|\mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right| \\
&\leqslant \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left(\left\|\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)^{1/2}\mathbf{X}_{0,[k]}\right\| \cdot \left\|\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)^{1/2}\widehat{\boldsymbol{\mu}}_{[k]}\right\|\right) \\
&\leqslant \sum_{k=1\,:\,\omega_k=1}^{b} \left(\mathbf{E}_{\Pi_2}\left\{\mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\mathbf{X}_{0,[k]}\right\} \cdot \mathbf{E}\left\{\hat{\Delta}_k^2 - \tilde{\Delta}_k^2\right\}\right)^{1/2} \\
&= \sum_{k=1\,:\,\omega_k=1}^{b} \left(\mathbf{E}_{\Pi_2}\mathrm{Tr}\left\{\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1}\right)\mathbf{X}_{0,[k]}\mathbf{X}_{0,[k]}^{\top}\right\}\right. \\
&\qquad \left. \times \left\{\frac{(2n-1)(p_0+n\Delta_k^2)}{(2n-p_0-2)n} - \frac{p_0}{n} - \Delta_k^2\right\}\right)^{1/2} \\
&= \sum_{k=1\,:\,\omega_k=1}^{b} \left(\mathrm{Tr}\left\{\frac{1+p_0}{2n-p_0-2}\mathbf{I}_{p_0\times p_0}\right\}\left\{\frac{1+p_0}{2n-p_0-2}\left(\frac{p_0}{n} + \Delta_k^2\right)\right\}\right)^{1/2}
\end{aligned}$$

$$= \sum_{k=1\,:\,\omega_k=1}^{b} \frac{p_0(1+p_0)}{(2n-p_0-2)} \left(\frac{1}{n} + \frac{\Delta_k^2}{p_0}\right)^{1/2},$$

and hence, as $b \to \infty$,

$$\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2} \left| \mathbf{X}_{0,[k]}^{\top} \left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1} \right) \widehat{\boldsymbol{\mu}}_{[k]} \right| = O\left(b^{1-\beta-3\theta/2} \log^{1/2} b \right). \tag{7.28}$$

From this and (7.27), for any $\varepsilon > 0$

$$\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}} \mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{X}_{0,[k]}^{\top} \left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}^{-1} \right) \widehat{\boldsymbol{\mu}}_{[k]} \right| \geqslant \varepsilon b^{1-\beta-\theta} \log b \right)$$
$$= \frac{O(b^{1-\beta-3\theta/2} \log^{1/2} b)}{O(b^{1-\beta-\theta} \log b)} = O(b^{-\theta/2} \log^{-1/2} b) = o(1), \quad b \to \infty.$$

This proves relation (7.26) and hence relation (4.9). Next, in order to verify relations (4.10) and (4.11), we shall need the following fact.

Fact 2. Formula (22) from Sitgreaves [23]. *Let* $\mathbf{Y}_1$ *and* $\mathbf{Y}_2$ *be* d*-dimensional* ($d \geqslant 3$) *normal random vectors with expected values* $k_1\boldsymbol{\delta}$ *and* $k_2\boldsymbol{\delta}$*, where* k_1 *and* k_2 *are known scalars, and let* $\mathbf{A}$ *be a* $d \times d$ *symmetric matrix with a Wishart distribution involving* N *degrees of freedom. Assume that* $\mathbf{Y}_1$*,* $\mathbf{Y}_2$*, and* $\mathbf{A}$ *are independently distributed with the same covariance matrix* $\boldsymbol{\Lambda}$*, and define the* 2×2 *matrix* $\mathbf{M}^* = (m_{ij}^*) = \mathbf{Y}^{\top}\mathbf{A}^{-1}\mathbf{Y}$*, where* $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$*. Denote by* $\mathbf{I}$ *the* 2×2 *identity matrix. Then, if* $\boldsymbol{\delta}^{\top}\boldsymbol{\Lambda}^{-1}\boldsymbol{\delta} = 0$*, the pdf of the symmetric matrix* $\mathbf{M}^*$ *(with three non-repeated elements* m_{11}^**,* m_{12}^**, and* m_{22}^**) is equal to*

$$p(m_{11}^*, m_{22}^*, m_{12}^*) = \frac{\Gamma(\frac{1}{2}(N+1))\Gamma(\frac{1}{2}(N+2))}{\Gamma(\frac{1}{2}(N-d+2))\Gamma(\frac{1}{2}(N-d+1))\Gamma(\frac{1}{2}(d-1))\Gamma(\frac{1}{2})\Gamma(\frac{d}{2})}$$
$$\times \frac{|\mathbf{M}^*|^{\frac{1}{2}(d-3)}}{|\mathbf{I}+\mathbf{M}^*|^{\frac{1}{2}(N+2)}}, \quad m_{11}^* \geqslant 0,\ m_{22}^* \geqslant 0,\ |\mathbf{M}^*| \geqslant 0.$$

We shall apply Fact 2 to show that (4.11) holds true. Below, for brevity, we shall write $\hat{T}_k$ instead of $\hat{T}_{k,b}$. Using Markov's inequality, for any $\varepsilon > 0$

$$\mathbf{P}_{\Pi_2} \left(\left| \sum_{k=1\,:\,\omega_k=0,\hat{\omega}_k=1}^{b} \hat{V}_k \right| \geqslant \varepsilon b^{1-\beta-\theta} \log b \right) \leqslant \frac{\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}_{\Pi_2} \left\{ |\hat{V}_k| \mathbb{I}(\hat{\omega}_k = 1) \right\}}{\varepsilon b^{1-\beta-\theta} \log b}. \tag{7.29}$$

Consider the numerator on the right side of (7.29). The arguments between relations (7.23) and (7.24) show that the error of replacing the event $\{\hat{\omega}_k = 1\} = \left\{ \hat{T}_k > p_0^{-1}(2\beta_{\hat{m}} + \theta + \epsilon) \log b \right\}$ by the analogous event with $\beta_{m_0} = \delta m_0$, as defined in the proof of Lemma 1, in place of $\beta_{\hat{m}}$ is negligible. Hence, it is

sufficient to show that, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, as $b \to \infty$

$$\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}_{\Pi_2}\left\{|\hat{V}_k|\mathbb{I}\left(\hat{T}_k > p_0^{-1}(2\beta_{m_0}+\theta+\epsilon)\log b\right)\right\} = o(b^{1-\beta-\theta}\log b).$$

We have

$$\begin{aligned}
&\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}_{\Pi_2}\left\{|\hat{V}_k|\mathbb{I}\left(\hat{T}_k > p_0^{-1}(2\beta_{m_0}+\theta+\epsilon)\log b\right)\right\} \\
&\leqslant \sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}_{\Pi_2}\left\{|\mathbf{X}_{0,[k]}^{\top}\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}|\mathbb{I}(\hat{T}_k > p_0^{-1}(2\beta_{m_0}+\theta+\epsilon)\log b)\right\} \\
&+\frac{1}{2}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left\{\hat{\Delta}_k^2\mathbb{I}(\hat{T}_k > p_0^{-1}(2\beta_{m_0}+\theta+\epsilon)\log b)\right\} =: J_{1,b} + J_{2,b}. \qquad (7.30)
\end{aligned}$$

Consider the term $J_{2,b} = J_{2,b}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta)$. Using the asymptotic relations

$$\Gamma(x) \sim x^{x-1/2}e^{-x}\sqrt{2\pi}, \quad x \to \infty, \qquad (7.31)$$

and

$$\int_{(c\log b)/n}^{\infty} \frac{y^{p_0/2}}{(1+y)^n}\,dy \sim \frac{(c\log b)^{p_0/2}}{n^{p_0/2+1}b^c}, \quad b \to \infty, \qquad (7.32)$$

where n satisfies (2.5) and c is a positive constant, and recalling that if index $k \in [\![b]\!]$ is such that $\omega_k = 0$ then $\hat{T}_k \sim F_{p_0,2n-p_0}(0)$, we can write

$$\begin{aligned}
J_{2,b} &= \frac{1}{2}\frac{(2n-1)p_0}{(2n-p_0)n}\sum_{k=1\,:\,\omega_k=0}^{b} \mathbf{E}\left\{\hat{T}_k\,\mathbb{I}\left(\hat{T}_k > p_0^{-1}(2\beta_{m_0}+\theta+\epsilon)\log b\right)\right\} \\
&= \frac{\left(b-[b^{1-\beta}]\right)(2n-1)p_0\Gamma(n)}{2(2n-p_0)n\Gamma(p_0/2)\Gamma(n-p_0/2)}\left(\frac{p_0}{2n-p_0}\right)^{p_0/2} \\
&\quad\times\int_{\frac{(2\beta_{m_0}+\theta+\epsilon)\log b}{p_0}}^{\infty} x^{p_0/2}\left(1+\frac{p_0}{2n-p_0}x\right)^{-n}dx \\
&= O(bn^{p_0/2})\int_{\frac{(2\beta_{m_0}+\theta+\epsilon)\log b}{2n-p_0}}^{\infty}\frac{y^{p_0/2}}{(1+y)^n}\,dy \\
&= O(bn^{p_0/2})O\left(n^{-(p_0/2+1)}b^{-\beta-\theta/2-\epsilon/2}\log^{p_0/2} b\right) \\
&= O\left(b^{1-\beta-3\theta/2-\epsilon/2}\log^{p_0/2} b\right), \quad b \to \infty.
\end{aligned}$$

From this, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$J_{2,b} = J_{2,b}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta) = o(b^{1-\beta-\theta}\log b), \quad b \to \infty. \qquad (7.33)$$

We shall now consider the term $J_{1,b} = J_{1,b}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta)$ and apply Fact 2, in which we take $\mathbf{Y}_1 = \mathbf{X}_{0,[k]}$, $\mathbf{Y}_2 = \sqrt{n}\widehat{\boldsymbol{\mu}}_{[k]}$, $\mathbf{A} = (2n-1)\widehat{\boldsymbol{\Sigma}}_{[k]}$, $d = p_0$, $N = 2n-1$, $\boldsymbol{\delta} = \mathbf{0}$, $k_1 = 1$, and $k_2 = \sqrt{n}$. Then, the joint pdf of $X_n := \frac{1}{2n-1}\mathbf{X}_{0,[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\mathbf{X}_{0,[k]}$, $Y_n := \frac{n}{2n-1}\widehat{\boldsymbol{\mu}}_{[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}$, and $Z_n := \frac{\sqrt{n}}{2n-1}\mathbf{X}_{0,[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}$ is

$$p_n(x,y,z) = \frac{C(n,p_0)(xy-z^2)^{\frac{1}{2}(p_0-3)}}{((1+x)(1+y)-z^2)^{n+\frac{1}{2}}}, \quad x \geqslant 0,\ y \geqslant 0,\ xy - z^2 \geqslant 0,$$

where, in view of (7.31), the normalizing constant

$$C(n,p_0) = \frac{\Gamma(n)\Gamma(n+\frac{1}{2})}{\Gamma(n-\frac{p_0}{2}+\frac{1}{2})\Gamma(n-\frac{p_0}{2})\Gamma(\frac{p_0}{2}-\frac{1}{2})\Gamma(\frac{1}{2})\Gamma(\frac{p_0}{2})}$$

satisfies $C(n,p_0) = O(n^{p_0})$ as $n \to \infty$. Denoting

$$\eta_b = \frac{(2\beta_{m_0} + \theta + \epsilon)\log b}{2n - p_0}, \tag{7.34}$$

we have for all $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, $b \geqslant 2$,

$$\begin{aligned}
J_{1,b} &= \sum_{k=1\,:\,\omega_k=0}^{b} \frac{2n-1}{\sqrt{n}} \mathbf{E}_{\Pi_2}\left\{ \frac{\sqrt{n}}{2n-1}\left|\mathbf{X}_{0,[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}\right| \right. \\
&\quad \left. \times \mathbb{I}\left(\frac{n}{2n-1}\widehat{\boldsymbol{\mu}}_{[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]} > \frac{(2\beta_{m_0}+\theta+\epsilon)\log b}{2n-p_0} \right) \right\} \\
&= \frac{2n-1}{\sqrt{n}}\left(b - [b^{1-\beta}]\right) C(n,p_0) \iiint\limits_{\substack{x>0, y>\eta_b, xy>z^2, \\ -\infty<z<\infty}} \frac{|z|(xy-z^2)^{(p_0-3)/2}}{((1+x)(1+y)-z^2)^{n+1/2}}\,dx\,dy\,dz \\
&= \frac{2(2n-1)}{\sqrt{n}}\left(b - [b^{1-\beta}]\right) C(n,p_0) \\
&\quad \times \int_{\eta_b}^{\infty} dy \int_0^{\infty} z\,dz \int_{z^2/y}^{\infty} \frac{(xy-z^2)^{(p_0-3)/2}}{((1+x)(1+y)-z^2)^{n+1/2}}\,dx \\
&= \frac{2(2n-1)}{\sqrt{n}}\left(b - [b^{1-\beta}]\right) C(n,p_0) C_1(n,p_0) \\
&\quad \times \int_{\eta_b}^{\infty} \frac{y^{n-1/2}}{(1+y)^{(p_0-1)/2}}\,dy \int_0^{\infty} \frac{z}{(y^2+y+z^2)^{n+1-p_0/2}}\,dz,
\end{aligned}$$

where $C_1(n,p_0) = \frac{\Gamma((p_0-1)/2)}{(n-1/2)(n-3/2)\ldots(n-p_0/2+1)}$ for $p_0 \geqslant 3$. Noting that $C(n,p_0) = O(n^{p_0})$ and $C_1(n,p_0) = O(n^{(1-p_0)/2})$ as $n \to \infty$, and using (2.5) and (7.32), we may continue and obtain that, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$J_{1,b} = \frac{(2n-1)}{\sqrt{n}}\left(b - [b^{1-\beta}]\right) \frac{C(n,p_0)C_1(n,p_0)}{(n-p_0/2)} \int_{\eta_b}^{\infty} \frac{y^{(p_0-1)/2}}{(1+y)^{n-1/2}}\,dy$$

$$= \frac{(2n-1)}{\sqrt{n}} \frac{b\, C(n,p_0) C_1(n,p_0)}{(n-p_0/2)} \frac{\eta_b^{(p_0-1)/2}(1+o(1))}{(n-3/2)(1+\eta_b)^{n-3/2}}$$
$$= O\left(b^{1-\beta_{m_0}-\theta-\epsilon/2}(\log b)^{(p_0-1)/2}\right) = O\left(b^{1-\beta-\theta-\epsilon/2}(\log b)^{(p_0-1)/2}\right)$$
$$= O\left(b^{1-\beta-\theta}\log b\right) O\left(b^{-\epsilon/2}\log^{(p_0-3)/2} b\right), \quad b \to \infty,$$

where, in view of (3.10), $b^{-\epsilon/2}\log^{(p_0-3)/2} b = o(1)$ as $b \to \infty$. Therefore, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$J_{1,b} = J_{1,b}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta) = o(b^{1-\beta-\theta}\log b), \quad b \to \infty. \tag{7.35}$$

It now follows from (7.29), (7.30), (7.33), and (7.35) that relation (4.11) holds true.

Remark 4. As seen from the derivation of (7.35), the choice of threshold $\hat{t} = p_0^{-1}(2\beta_{\hat{m}} + \theta + \epsilon)\log b$ instead of $\hat{t} = p_0^{-1}(2\beta_{\hat{m}} + \epsilon)\log b$, which would be sufficient for selecting useful feature blocks, is done to have relation (4.11) valid, and hence successful classification possible.

Finally, we shall verify the validity of (4.10). Again, using Markov's inequality,

$$\mathbf{P}_{\Pi_2}\left(\left|\sum_{k=1\,:\,\omega_k=1,\hat{\omega}_k=0}^{b} \hat{V}_k\right| \geqslant \varepsilon b^{1-\beta-\theta}\log b\right) \leqslant \frac{\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{|\hat{V}_k|\mathbb{I}(\hat{\omega}_k = 0)\right\}}{\varepsilon b^{1-\beta-\theta}\log b}, \tag{7.36}$$

and hence the problem reduces to showing that, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, the numerator is $o(b^{1-\beta-\theta}\log b)$ as $b \to \infty$. Consider the numerator on the right side of (7.36):

$$\sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{|\hat{V}_k|\mathbb{I}(\hat{\omega}_k = 0)\right\}$$
$$\leqslant \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^{\top}\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}\right| \mathbb{I}\left(p_0\hat{T}_k \leqslant (2\beta_{\hat{m}} + \theta + \epsilon)\log b\right)\right\}$$
$$+ \frac{p_0(2n-1)}{2n(2n-p_0)} \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}\left\{\hat{T}_k \mathbb{I}\left(p_0\hat{T}_k \leqslant (2\beta_{\hat{m}} + \theta + \epsilon)\log b\right)\right\}$$
$$=: L_{1,b}(\beta_{\hat{m}}) + L_{2,b}(\beta_{\hat{m}}). \tag{7.37}$$

Similar to the proof of (3.23) in Lemma 2, for deriving suitable upper bounds on $L_{1,b} = L_{1,b}(\beta_{\hat{m}})$ and $L_{2,b} = L_{2,b}(\beta_{\hat{m}})$, the statistic $\beta_{\hat{m}}$ in the expressions for $L_{1,b}$ and $L_{2,b}$ can be replaced by the (non-random) grid point $\beta_{m_0} = \delta m_0$, which is chosen to have $\beta_{m_0} < \beta \leqslant \beta_{m_0+1}$, for the error of doing that is negligibly small, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$.

Consider the term $L_{2,b}(\beta_{m_0}) = L_{2,b}(\beta_{m_0}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta)$. Recall that $n \sim b^{\theta}$, the number of k's for which $\omega_k = 1$ is $[b^{1-\beta}]$, and $\hat{T}_k = \frac{(2n-p_0)n}{(2n-1)p_0}\hat{\Delta}_k^2 \sim F_{p_0,2n-p_0}(n\Delta_k^2)$.

Then, using relations (4.2) and (7.4), and noting that $(\beta, r) \in \mathcal{D}_1^0(\theta)$, that is, $r > \beta + \theta/2$, we have, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, as $b \to \infty$

$$
\begin{aligned}
L_{2,b}(\beta_{m_0}) &= \frac{p_0(2n-1)}{2n(2n-p_0)} \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}\left\{\hat{T}_k \mathbb{I}\left(p_0\hat{T}_k \leqslant (2\beta_{m_0}+\theta+\epsilon)\log b\right)\right\} \\
&\leqslant \frac{p_0}{2n}\left(\frac{4\log b}{p_0}\right) \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{P}\left(p_0 F_{p_0,2n-p_0}(n\Delta_k^2) \leqslant (2\beta_{m_0}+\theta+\epsilon)\log b\right) \\
&\leqslant \frac{2b^{1-\beta}\log b}{n}\,\mathbf{P}\left(p_0 F_{p_0,2n-p_0}(2r\log b) \leqslant (2\beta_{m_0}+\theta+\epsilon)\log b\right) \\
&= \frac{2b^{1-\beta}\log b}{n}\,\mathbf{P}\left(\chi^2_{p_0}(2r\log b) \leqslant (2\beta_{m_0}+\theta+\epsilon)\log b\right)\left(1+O(b^{-\theta}\log^2 b)\right) \\
&= \frac{2b^{1-\beta}\log b}{n} O\left(b^{-\left(\sqrt{r}-\sqrt{\beta_{m_0}+\theta/2+\epsilon/2}\right)^2}\log^{-1/2} b\right) = o(b^{1-\beta-\theta}\log b).
\end{aligned}
$$

Thus, for all $(\beta, r) \in \mathcal{D}_1^0(\theta)$, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$, as $b \to \infty$,

$$
L_{2,b}(\beta_{m_0}) = L_{2,b}(\beta_{m_0}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta) = o(b^{1-\beta-\theta}\log b),
$$

and hence

$$
L_{2,b}(\beta_{\hat{m}}) = L_{2,b}(\beta_{\hat{m}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta) = o(b^{1-\beta-\theta}\log b). \tag{7.38}
$$

Now, we turn to the analysis of the term $L_{1,b}(\beta_{m_0}) = L_{1,b}(\beta_{m_0}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, r, \theta)$. Denote

$$
B_k = B_{k,b} = \{p_0\hat{T}_k \leqslant (2\beta_{m_0}+\theta+\epsilon)\log b\}, \quad k \in [\![b]\!],\ b = 2, 3, \ldots,
$$

and observe that, in view of relations (4.2) and (7.4), for all $(\beta, r) \in \mathcal{D}_1^0(\theta)$, uniformly in $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbf{M}_{b,\beta,r}$,

$$
\max_{1\leqslant k\leqslant b} \mathbf{P}(B_k) = o(1), \quad b \to \infty. \tag{7.39}
$$

By the triangle and Cauchy-Schwarz inequalities, using the independence of $\mathbf{X}_{0,[k]}$ from $\widehat{\boldsymbol{\mu}}_{[k]}$ and $\mathbb{I}(B_k)$, we get

$$
\begin{aligned}
L_{1,b}(\beta_{m_0}) &= \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^\top \widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}\right| \mathbb{I}\left(B_k\right)\right\} \\
&\leqslant \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^\top \boldsymbol{\Sigma}_{[k]}^{-1}\widehat{\boldsymbol{\mu}}_{[k]}\right| \mathbb{I}\left(B_k\right)\right\} \\
&+ \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^\top \left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1} - \boldsymbol{\Sigma}_{[k]}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right| \mathbb{I}\left(B_k\right)\right\} \\
&\leqslant \sum_{k=1\,:\,\omega_k=1}^{b} \mathbf{E}_{\Pi_2}\left\{\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\|\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\widehat{\boldsymbol{\mu}}_{[k]}\|\mathbb{I}(B_k)\right\}
\end{aligned}
$$

$$+\sum_{k=1:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}-\boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right|\right\}$$

$$\leqslant\sum_{k=1:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\|\left(\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\boldsymbol{\mu}_{[k]}\|\,\mathbf{P}(B_k)\right.$$

$$\left.+\,\mathbf{E}\left\{\|\boldsymbol{\Sigma}_{[k]}^{-1/2}(\widehat{\boldsymbol{\mu}}_{[k]}-\boldsymbol{\mu}_{[k]})\|\mathbb{I}(B_k)\right\}\right)$$

$$+\sum_{k=1:\,\omega_k=1}^{b}\mathbf{E}_{\Pi_2}\left\{\left|\mathbf{X}_{0,[k]}^{\top}\left(\widehat{\boldsymbol{\Sigma}}_{[k]}^{-1}-\boldsymbol{\Sigma}_{[k]}^{-1}\right)\widehat{\boldsymbol{\mu}}_{[k]}\right|\right\}=:K_{1,b}+K_{2,b}. \tag{7.40}$$

Consider the term $K_{1,b}$. Since $\sqrt{n}\boldsymbol{\Sigma}_{[k]}^{-1/2}(\widehat{\boldsymbol{\mu}}_{[k]}-\boldsymbol{\mu}_{[k]})\sim N_{p_0}(\mathbf{0},\mathbf{I}_{p_0\times p_0})$ and $\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\|^2\sim\chi^2_{p_0}(\Delta_k^2)$, it follows that

$$\mathbf{E}_{\Pi_2}\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\|\leqslant\left(\mathbf{E}_{\Pi_2}\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\mathbf{X}_{0,[k]}\|^2\right)^{1/2}=\left(\mathbf{E}_{\Pi_2}(\chi^2_{p_0}(\Delta_k^2))\right)^{1/2}=\sqrt{p_0+\Delta_k^2},$$

and

$$\mathbf{E}\|\sqrt{n}\boldsymbol{\Sigma}_{[k]}^{-1/2}(\widehat{\boldsymbol{\mu}}_{[k]}-\boldsymbol{\mu}_{[k]})\|\leqslant\left(\mathbf{E}\|\sqrt{n}\boldsymbol{\Sigma}_{[k]}^{-1/2}(\widehat{\boldsymbol{\mu}}_{[k]}-\boldsymbol{\mu}_{[k]})\|^2\right)^{1/2}$$
$$=(\mathbf{E}(\chi^2_{p_0}(0))^{1/2}=\sqrt{p_0}.$$

Therefore, noting that $\|\boldsymbol{\Sigma}_{[k]}^{-1/2}\boldsymbol{\mu}_{[k]}\|=\sqrt{\Delta_k^2}$, applying relation (7.39) and the property of absolute continuity of the integral of an integrable function, for all $(\beta,r)\in\mathcal{D}_1^0(\theta)$, uniformly in $(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}$, as $b\to\infty$

$$K_{1,b}\leqslant\sum_{k=1:\,\omega_k=1}^{b}\sqrt{(p_0+\Delta_k^2)}$$

$$\times\left(\sqrt{\Delta_k^2}\,\mathbf{P}(B_k)+\frac{1}{\sqrt{n}}\mathbf{E}\left\{\|\sqrt{n}\boldsymbol{\Sigma}_{[k]}^{-1/2}(\widehat{\boldsymbol{\mu}}_{[k]}-\boldsymbol{\mu}_{[k]})\|\mathbb{I}(B_k)\right\}\right)$$

$$=O(b^{1-\beta})O\left(n^{-1/2}\log^{1/2}\right)\left\{O\left(n^{-1/2}\log^{1/2}b\right)o(1)+O\left(n^{-1/2}\right)o(1)\right\}$$

$$=o\left(b^{1-\beta-\theta}\log b\right). \tag{7.41}$$

The upper bound on $K_{2,b}$, for all $(\beta,r)\in\mathcal{D}_1^0(\theta)$, uniformly in $(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}$, is given by (7.28). It now follows from (7.28), (7.40) and (7.41) that for all $(\beta,r)\in\mathcal{D}_1^0(\theta)$, uniformly in $(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\mathbf{M}_{b,\beta,r}$, $L_{1,b}(\beta_{m_0})=L_{1,b}(\beta_{m_0},\boldsymbol{\mu},\boldsymbol{\Sigma},\beta,r,\theta)=O\left(b^{1-\beta-\theta}\log b\right)o(1)=o(b^{1-\beta-\theta}\log b)$, and hence as $b\to\infty$

$$L_{1,b}(\beta_{\hat{m}})=L_{1,b}(\beta_{\hat{m}},\boldsymbol{\mu},\boldsymbol{\Sigma},\beta,r,\theta)=o(b^{1-\beta-\theta}\log b). \tag{7.42}$$

Finally, the combination of (7.37), (7.38) and (7.42) gives (4.10). The proof of Lemma 3 is complete. □

Acknowledgment

The authors are grateful to the Editor Domenico Marinucci and anonymous referee for many helpful comments on this work.

References

[1] Ahmad, R. M. and Pavlenko, T. (2018). A U-classifier for high-dimensional data for non-normality. *Journal of Multivariate Analysis* **167** 269–283. MR3830646

[2] Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Mathematical Statistics* **66** 983–1010. MR3250825

[3] Bernstein, S. N., *Probability Theory*. OGIZ, Moscow–Leningrad (1946). In Russian.

[4] Butucea, C. and Stepanova, N. (2017). Adaptive variable selection in nonparametric sparse additive models. *Electronic Journal of Statistics* **11** 2321–2357. MR3656494

[5] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577. MR2896857

[6] Chan, Y.-B. and Hall, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika* **96** 469–478. MR2507156

[7] Csörgő, M., Csörgő, S., Horváth, L. and Mason, D. (1986). Weighted empirical and quantile processes. *The Annals of Probability* **14** 31–85. MR0815960

[8] Csörgő, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley, New York. MR1215046

[9] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994. MR2065195

[10] Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A* **367** 4449–4470. MR2546396

[11] Fan, Y., Jin, J. and Yao, Z. (2013). Optimal classification in sparse Gaussian graphic models. *The Annals of Statistics* **41** 2537–2571. MR3161437

[12] Genovese, C. R., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research* **13** 2107–2143. MR2956354

[13] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over-parametrization. *Bernoulli* **10** 971–988. MR2108039

[14] Han, C.-P. (1975). Some relationships between noncentral chi-squared and normal distributions. *Biometrika* **62** 213–214. MR0373095

[15] Ingster, Yu. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Mathematical Methods of Statistics* **6** 47–69. MR1456646

[16] Ingster, Yu. I., Pouet, C. and Tsybakov, A. B. (2009). Classification of sparse high-dimensional vectors. *Philosophical Transactions of the Royal Society A* **367** 4427–4448. MR2546395

[17] Ingster, Yu. I. and Stepanova, N. A. (2014). Adaptive variable selection in nonparametric sparse regression. *Journal of Mathematical Sciences* **199** 184–201. MR3032218

[18] Lepski, O. (1991). One problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications* **35** 454–466. MR1091202

[19] Petrov, V. V. (2004). *Limit Theorems of Probability Theory*. Clarendon Press, Oxford. MR1353441

[20] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley and Sons, New York. MR0346957

[21] Shao, J., Wang, Y., Deng, X. and Wang. S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics* **39** 1241–1265. MR2816353

[22] Siotani, M. (1971). An asymptotic expansion of the non-null distribution of Hotelling's generalized T_0^2-statistic. *The Annals of Mathematical Statistics* **42** 560–571. MR0286202

[23] Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedures. *The Annals of Mathematical Statistics* **23** 263–270. MR0057500

[24] Stepanova, N. and Pavlenko, T. (2018). Goodnes-of-fit tests based on sup-functionals of weighted empirical processes. *Theory of Probability and Its Applications* **63** 358–388. MR3796493