Classification with many classes: Challenges and pluses

Article *in* Journal of Multivariate Analysis · August 2019

DOI: 10.1016/j.jmva.2019.104536

Felix Abramovich
Tel Aviv University
56 PUBLICATIONS 2,404 CITATIONS
SEE PROFILE

Marianna Pensky
University of Central Florida
96 PUBLICATIONS 1,662 CITATIONS
SEE PROFILE

Journal of Multivariate Analysis 174 (2019) 104536

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Classification with many classes: Challenges and pluses

Felix Abramovich [a,*], Marianna Pensky [b]

[a] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel
[b] Department of Mathematics, University of Central Florida, 4393 Andromeda Loop N Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 19 May 2019
Received in revised form 24 July 2019
Accepted 24 July 2019
Available online 2 August 2019

AMS 2010 subject classifications:
primary 62H30
secondary 62H99

Keywords:
Feature selection
High-dimensionality
Misclassification error
Multi-class classification
Sparsity

ABSTRACT

The objective of the paper is to study accuracy of multi-class classification in high-dimensional setting, where the number of classes is also large (“large L, large p, small n” model). While this problem arises in many practical applications and many techniques have been recently developed for its solution, to the best of our knowledge nobody provided a rigorous theoretical analysis of this important setup. The purpose of the present paper is to fill in this gap.

We consider one of the most common settings, classification of high-dimensional normal vectors where, unlike standard assumptions, the number of classes could be large. We derive non-asymptotic conditions on effects of significant features, and the low and the upper bounds for distances between classes required for successful feature selection and classification with a given accuracy. Furthermore, we study an asymptotic setup where the number of classes is diverging with the dimension of feature space and while the number of samples per class is possibly limited. We point out on an interesting and, at first glance, somewhat counter-intuitive phenomenon that a large number of classes may be a “blessing” rather than a “curse” since, in certain settings, the precision of classification can improve as the number of classes grows. This is due to more accurate feature selection since even weaker significant features, which are not sufficiently strong to be manifested in a coarse classification, being shared across the classes, have a stronger impact as the number of classes increases. We supplement our theoretical investigation by a simulation study and a real data example where we again observe the above phenomenon.

1. Introduction

Classification has been studied in many contexts. In the modern era one is usually interested in classifying objects that are described by a large number of features and belong to many different groups. For example the large hand-labeled ImageNet data set http://www.image-net.org/ contains 10,000,000 labeled images depicting more than 10,000 object categories where each image, on the average, is represented by $482 \times 415 \approx 200{,}000$ pixels (see [22] for description and discussion of this data set). The challenge of handling large dimensional data got the name of “large p small n” type of problems which means that dimensionality of parameter space p by far exceeds the sample size n. It is well known that solving problems of this type requires rigorous model selection. In fact, the results of Bickel and Levina [2], Fan and Fan [11], Shao et al. [23] demonstrate that even for the standard case of two classes, classification of high-dimensional normal vectors without feature selection is as bad as just pure random guessing. However, while analysis of

[*] Corresponding author.
E-mail address: felix@tauex.tau.ac.il (F. Abramovich).

https://doi.org/10.1016/j.jmva.2019.104536
0047-259X/© 2019 Elsevier Inc. All rights reserved.

high-dimensional data has become ubiquitous, to the best of our knowledge, there are no theoretical studies that examine the effect of large number of classes on classification accuracy. The objective of the present paper is to fill in this gap.

At first glance, the problem of successful classification when the number of classes is large seems close to impossible. On the other hand, humans have no difficulty in distinguishing between thousands of objects, and the accuracy of state-of-the-art computer vision techniques is approaching human accuracy. In fact, in some settings, the accuracy of classification improves when the number of classes grows. How is this possible? One of the reasons why multi-class classification succeeds is that selection of appropriate features from a large sparse p-dimensional vector becomes easier when the number of classes is growing since even weaker significant features that are not sufficiently strong to be manifested in a coarse classification with a small number of classes may nevertheless have a strong impact as the number of classes grows. Simulation studies in [7] and [21] support such a claim. Arias-Castro, Candès and Plan [1] reported on a similar occurrence for testing in the sparse ANOVA model. Our paper establishes a firm theoretical foundation under the above phenomenon and confirms it via simulation studies and a real data example.

Although there exists an enormous amount of literature on classification, most of the existing theoretical results have been obtained for the binary classification ($L = 2$) (see [4] and references therein for a comprehensive survey). In particular, binary classification of high-dimensional sparse Gaussian vectors was considered in [2,8,9,11,17] and [23] among others.

In the meantime, a significant amount of effort has been spent on designing methods for the multi-class classification in statistical and machine learning literature. We can mention here techniques designed to adjust pairwise classification to multi-class setting [10,14,18], adjustment of the support vector machine technique to the case of several classes [5,19] as well as a variety of approaches to expand the linear regression and the neural networks techniques to accommodate the multi-category setup (see, e.g., [13]). Pan, Wang and Li [20] and Tewari and Bartlett [24] generalized theoretical results for binary classification to the case of multi-class classification and established consistency of the proposed classification procedures. However, all above-mentioned investigations considered only the "small L, large p, small n" setup, where the number of classes was assumed to be *fixed*.

This paper is probably the first attempt to rigorously investigate "large L, large p, small n" classification and the impact of the number of classes on the accuracy of feature selection and classification. In particular, we explore the somewhat counter-intuitive phenomenon, where the large number of classes may become a "blessing" rather than a "curse" for successful classification as more significant features may be revealed. For this purpose, we consider a well-known problem of multi-class classification of high-dimensional normal vectors. We assume that only a subset of truly significant features really contribute to separation between classes (sparsity). For this reason, we carry out feature selection and, following a standard scheme, assign the new observed vector to the closest class w.r.t. the scaled Mahalanobis distance in the space of the selected significant features. Our paper considers a realistic scenario where the number of classes as well as the number of features is large while the number of observations per class is possibly limited ("large L, large p, small n" model). We do not fix the total number of observations since in the real world the experience of each new class comes with its own, usually finite, set of observations.

We start with a non-asymptotic setting and derive the conditions on effects of significant features, and the low and upper bounds for the distances between classes required for successful feature selection and classification with a given accuracy. All the results are obtained with the explicit constants and remain valid for any combination of parameters. Our finite sample study is followed by an asymptotic analysis for a large number of features p, where, unlike previous works, the number of classes L may grow with p while the number of samples per class may grow or stay fixed. Our findings indicate that having larger number of classes aids the feature selection and, hence, can improve classification accuracy. On the other hand, larger number of classes require having larger number of significant features p_1 for their separation which automatically leads to a "large p" setting. Nevertheless, due to increasing point isolation in high-dimensional spaces (see, e.g., [12], Section 1.2.1), those separation conditions become attainable when p is large.

We ought to point out that our paper does not propose a novel methodology for feature selection or classification. Rather than that, it studies one of the most popular Gaussian setting and adapts to the case of a large number of classes a standard general scheme, where feature selection is implemented by a thresholding technique with the properly chosen threshold and classification is carried out on the basis of the minimal Mahalanobis distance (we consider both the known and the unknown covariance matrix scenarios). This is a common widely used general scheme for classification and feature selection in such setting (see, e.g., [11,20] and [23] for similar approaches that differ mostly by selections of thresholds and distances). Nevertheless, the setup is simple enough for derivations of conditions required for successful classification with a specified precision when the number of classes is large. Therefore, in our simulation study we do not compare these simple and well known techniques with the state of the art classification methodologies but instead investigate how these popular procedures perform when p is large and both the number of classes L and the number of significant features p_1 are growing. In particular, simulations support our finding that classification precision can improve when L is increasing. The real data example confirms that the phenomenon above is not due to an artificial construction and is possible in a real life setting.

The rest of the paper is organized as follows. In Section 2 we present the feature selection and multi-class classification procedures and derive the non-asymptotic bounds for their accuracy. An asymptotic analysis is considered in Section 3. Section 4 discusses adaptation of the procedure in the case of the unknown covariance matrix. In Section 5 we illustrate the performance of the proposed approach on simulated and real-data examples. Some concluding remarks are summarized in Section 6. All the proofs are given in Appendix.

2. Feature selection and classification procedure

2.1. Notation and preliminaries

Consider the problem of multi-class classification of p-dimensional normal vectors with L classes:

$$\mathbf{Y}_{li} = \mathbf{m}_l + \boldsymbol{\epsilon}_{li} \tag{1}$$

for $l \in \{1,\dots,L\}$, $i \in \{1,\dots,n_l\}$, where $\mathbf{m}_l \in \mathbb{R}^p$ is the vector of mean effects of p features in the lth class and $\boldsymbol{\epsilon}_{li} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with the common non-singular covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. To clarify the proposed approach we assume meanwhile that $\boldsymbol{\Sigma}$ is known and discuss the situation with the unknown $\boldsymbol{\Sigma}$ in Section 4.

In what follows, we study a realistic scenario where the number of classes as well as the number of features is large while the number of observations per class is possibly limited ("large L, large p, small n" model). We do not fix the total number of observations since in the real world the experience of each new class comes with its own, usually finite, set of observations.

After averaging over repeated observations within each class, model (1) yields

$$\bar{\mathbf{Y}}_l = \mathbf{m}_l + \boldsymbol{\epsilon}_l^*, \tag{2}$$

where $\boldsymbol{\epsilon}_l^* \sim \mathcal{N}(\mathbf{0}_p, n_l^{-1}\boldsymbol{\Sigma})$ and $l \in \{1,\dots,L\}$.

The objective is to assign a new observed feature vector $\mathbf{Y}_0 \in \mathbb{R}^p$ to one of the L classes. Denote

$$N = \sum_{l=1}^L n_l, \quad \rho_l = n_l/(n_l+1), \quad L_1 = L-1, \tag{3}$$

where evidently $1/2 \le \rho_l < 1$.

Since $\mathrm{Var}(\mathbf{Y}_0 - \bar{\mathbf{Y}}_l) = \rho_l^{-1}\boldsymbol{\Sigma}$, we assign $\mathbf{Y}_0$ to the class l with the nearest centroid $\bar{\mathbf{Y}}_l$ w.r.t. to the scaled Mahalanobis distance:

$$\hat{l} = \arg\min_{1\le l\le L} \left\{\rho_l(\mathbf{Y}_0 - \bar{\mathbf{Y}}_l)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_0 - \bar{\mathbf{Y}}_l)\right\}. \tag{4}$$

It is well-known (see, e.g., [2,11] and [23]) that the performance of classification procedures is worsening as the number of features grows (curse of dimensionality). Hence, dimensionality reduction by feature selection prior to classification is crucial for large values of p.

Re-write (2) in terms of the one-way multivariate analysis of variance (MANOVA) model as follows:

$$\bar{\mathbf{Y}}_l = \boldsymbol{\delta} + \boldsymbol{\beta}_l + \boldsymbol{\epsilon}_l^* \tag{5}$$

for $l \in \{1,\dots,L\}$, where $\mathbf{m}_l = \boldsymbol{\delta} + \boldsymbol{\beta}_l$, $\boldsymbol{\delta}$ is the vector of mean main effects of features and β_{lj}, $j \in \{1,\dots,p\}$ is the mean interaction effect of jth feature with lth class, with the standard identifiability conditions $\sum_{l=1}^L \beta_{lj} = 0$ for each j.

The impact of jth feature on classification depends on its variability between the different classes characterized by the interactions β_{lj}, $l \in \{1,\dots,L\}$ in the model (5). The larger are the interactions, the stronger is the impact of the feature. A natural global measure of feature's contribution to classification is then $b_j^2 = \sum_{l=1}^L \beta_{lj}^2$. Note that a feature may still have a strong main effect δ_j but its contribution to classification nevertheless remains weak if it does not vary significantly between classes, that is, if b_j^2 is small. The main goal of feature selection is to identify a sparse subset of significant features for further use in classification.

2.2. Oracle classification

First, we consider an ideal situation where there is an oracle that provides the list of truly significant features with $b_j^2 > 0$. In this case, we would obviously use only those features for classification, thus, reducing the dimensionality of the problem. Define indicator variables $x_j = I\{b_j^2 > 0\}$, and let $p_1 = \sum_{j=1}^p x_j$ and $p_0 = p - p_1$ be, respectively, the numbers of significant and non-significant features. Without loss of generality, we can always order features in such a way that those p_1 significant features are the first ones. The classification procedure (4) then becomes

$$\hat{l} = \operatorname*{argmin}_{1\le l\le L} \left\{\rho_l(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^*)^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^*)\right\}, \tag{6}$$

where $\mathbf{Y}_0^*, \mathbf{Y}_l^* \in \mathbb{R}^{p_1}$ are the truncated versions of $\mathbf{Y}_0$ and $\bar{\mathbf{Y}}_l$ respectively: $Y_{0j}^* = Y_{0j}$ and $Y_{lj}^* = \bar{Y}_{lj}$, $j \in \{1,\dots,p_1\}$, and $\boldsymbol{\Sigma}^* \in \mathbb{R}^{p_1\times p_1}$ is the corresponding upper left sub-matrix of $\boldsymbol{\Sigma}$.

Theorem 1 provides an upper bound for misclassification error of the oracle classification procedure (6):

Theorem 1. *Consider the model (1) and the equivalent model (5). Let $\mathbf{m}_k^* \in \mathbb{R}^{p_1},\ k \in \{1,\dots,L\}$ be the truncated versions of class centers $\mathbf{m}_k$ and assume that for all pairs of classes*

$$(\mathbf{m}_k^* - \mathbf{m}_{k'}^*)^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{m}_k^* - \mathbf{m}_{k'}^*) \geq \frac{8\,\ln(L_1/\alpha)}{\min(\rho_k,\rho_{k'})}\left\{1+\frac{1}{\sqrt{2\min(n_k,n_{k'})}}\left(1+\sqrt{2p_1/\ln(L_1/\alpha)}\right)\right\}, \qquad (7)$$

for some $0<\alpha\leq 1$.

Let a new observation Y_0 from the class l be assigned to the $\hat{l}$th class according to classification rule (6). Then, the misclassification error is

$$\Pr(\hat{l}\neq l)\leq \alpha.$$

Condition (7) verifies that classes should be sufficiently separated from each other (in terms of Mahalanobis distance) to achieve the required classification accuracy. In fact, the requirements in (7) are also essentially necessary. Theorem 2, which is a direct consequence of Fano's lemma for the lower bound of misclassification error (see, e.g., [15], Section 7.1), implies that the first term $O(\ln(L_1/\alpha))$ on the RHS of (7) is unavoidable for successful classification and cannot be significantly improved (in the minimax sense) even in the idealized case, where the class centers $\mathbf{m}_k^*$ are known:

Theorem 2. *Consider the model (1). Let a new observation $\boldsymbol{Y}_0$ be from one of L classes. If*

$$\tilde{\Delta}^2 = \min_{l\neq k}\ (\boldsymbol{m}_l^* - \boldsymbol{m}_k^*)^\top(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{m}_l^* - \boldsymbol{m}_k^*) \leq 2\aleph \ln L_1,$$

for some $\aleph>0$, then

$$\inf_{\psi}\max_{1\leq l\leq L}\Pr_l(\psi(\boldsymbol{Y}_0)\neq l)\geq 1-\aleph-\frac{\ln 2}{\ln L_1},$$

where $\Pr_l$ is the probability evaluated under the assumption that $\boldsymbol{Y}_0$ belongs to the lth class, and the infimum is taken over all classification rules $\psi(\boldsymbol{Y}_0):\boldsymbol{Y}_0\to\{1,\dots,L\}$.

The second term on the RHS of (7) appears due to replacing the unknown p_1-dimensional class centers $\mathbf{m}_k^*$'s by the corresponding within-class sample means $\bar{\mathbf{Y}}_k^*$'s in (6). Indeed, straightforward extension of the results of Theorem 1 of [11] for a general $L\geq 2$ yields that, unless for all pairs (k,k'), $(\mathbf{m}_k^*-\mathbf{m}_{k'}^*)^\top(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{m}_k^*-\mathbf{m}_{k'}^*)\geq C\sqrt{p_1\ln L_1/\min(n_k,n_{k'})}$ for some $C>0$, the curse of dimensionality affects the accumulated error in estimating high-dimensional $\mathbf{m}_k^*$'s and yields classification performance nearly the same as random guessing.

2.3. Feature selection procedure

Consider now classification setup in the MANOVA model (5) with a more realistic scenario, where a set of significant features is unknown and should be identified from the data.

To simplify the calculus and to avoid complications with post-selection inference, we split the data at random into two sets $Y_{lj}^{(1)}$'s and $Y_{lj}^{(2)}$'s in some fixed proportion $\pi\in(0,1)$ (in the simplest case, the sizes of both sets are equal with $\pi=1/2$). Subsequently, use $Y_{lj}^{(1)}$'s for feature selection and $Y_{lj}^{(2)}$'s for classification based on the selected features. More specifically, for lth class, split its n_l observations Y_{lj}'s into two sub-samples of sizes $n_l^{(1)}$ and $n_l^{(2)}$ at the same proportion π, i.e., $n_l^{(1)}=\lfloor \pi n_l\rfloor$, where $\lfloor\cdot\rfloor$ is the integer part, and $n_l^{(2)}=n_l-n_l^{(1)},\ l\in\{1,\dots,L\}$. Denote the total sample sizes of the resulting two sets by $N_1=\sum_{l=1}^L n_l^{(1)}$ and $N_2=\sum_{l=1}^L n_l^{(2)}$, so that $N_1+N_2=N$.

Following our previous arguments, a jth feature is not significant (irrelevant) for classification if it has zero interaction effects with all classes, that is, if $\beta_{lj}=0,\ j\in\{1,\dots,L\}$ or, equivalently, $b_j^2=0$. Then, for each $j=1,\dots,p$ we need to test the null hypothesis $H_{0j}:b_j^2=0$. An obvious test statistic is then

$$\zeta_j=\sigma_j^{-2}\sum_{l=1}^L n_l^{(1)}(\bar{Y}_{lj}^{(1)}-\bar{Y}_{\cdot j}^{(1)})^2, \qquad (8)$$

where $\sigma_j^2=\boldsymbol{\Sigma}_{jj}$ and $\bar{Y}_{\cdot j}^{(1)}=(n_l^{(1)})^{-1}\sum_{l=1}^L Y_{lj}^{(1)}$. Under the null, $\zeta_j\sim\chi^2_{L_1}$, while under the alternative $\zeta_j\sim\chi^2_{L_1;\mu_j}$, where $\chi^2_{L_1;\mu_j}$ is the non-central chi-square distribution with the non-centrality parameter $\mu_j=\sigma_j^{-2}\sum_{l=1}^L n_l^{(1)}\beta_{lj}^2$. Note that unless $\boldsymbol{\Sigma}$ is diagonal, ζ_j's are correlated.

For a given $0<\alpha\leq 1$, define a threshold

$$\lambda=L_1+2\sqrt{L_1\ln(2p/\alpha)}+2\ln(2p/\alpha) \qquad (9)$$

and select the jth feature as significant (reject H_{0j}) if

$$\zeta_j=\sigma_j^{-2}\sum_{l=1}^L n_l^{(1)}(\bar{Y}_{lj}^{(1)}-\bar{Y}_{\cdot j}^{(1)})^2>\lambda. \qquad (10)$$

The following theorem shows that under certain conditions on the minimal required effect for significant features, the proposed feature selection procedure correctly identifies the true (unknown) subset of significant features with probability at least $1-\alpha$:

Theorem 3. *Consider the feature selection procedure* (10) *with the threshold* (9) *for some* $0<\alpha\leq 1$. *Define indicator variables* $\hat{x}_j = I\{\sigma_j^{-2}\ \sum_{l=1}^L n_l^{(1)}(\bar{Y}_{lj}-\bar{Y}_{\cdot j}^{(1)})^2>\lambda\}$ *for* $j\in\{1,\ldots,p\}$. *Let*

$$\mu^* = \min_{1\leq j\leq p_1} \sigma_j^{-2}\ \sum_{l=1}^L n_l^{(1)}\beta_{lj}^2 \tag{11}$$

and assume that for all p_1 *truly significant features one has*

$$\mu^*\geq 4\left(3\ln(2p/\alpha)+\sqrt{L_1\ln(2p/\alpha)}\right). \tag{12}$$

Then,

$$\Pr(\hat{x}=x)\geq 1-\alpha.$$

The condition (12) on the total minimal effect for significant features can be re-formulated in terms of their *average* effect per class:

$$\frac{1}{\sigma_j^2 L}\ \sum_{l=1}^L n_l^{(1)}\beta_{lj}^2\geq 4\left(\frac{3\ln(2p/\alpha)}{L}+\sqrt{\frac{\ln(2p/\alpha)}{L}}\right) \tag{13}$$

for $j\in\{1,\ldots,p_1\}$.

Thus, as the number of classes in model (1) increases, even significant features with weaker effects within each class become manifested and contribute to classification. Effect of a certain feature that remains latent and unnoticed in coarse classification with a small number of classes may be expressed in a finer classification.

2.4. Classification rule and misclassification error

Consider now the classification rule (6) applied on the second set of the data with $\bar{Y}_l^{(2)*}$, where the unknown true x_j are replaced by $\hat{x}_j$ following the proposed feature selection procedure. Let $\hat{p}_1=\sum_{j=1}^p \hat{x}_j$ be the number of features declared significant and $\hat{p}_0=p-\hat{p}_1$. Again, order the features in such a way that those $\hat{p}_1$ features selected as significant are the first ones. Thus, the resulting classification rule can then be presented as follows:

$$\hat{l}=\operatorname*{argmin}_{1\leq l\leq L}\left\{\rho_l\,(\mathbf{Y}_0^*-\bar{\mathbf{Y}}_l^{(2)*})^\top(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^*-\bar{\mathbf{Y}}_l^{(2)*})\right\}, \tag{14}$$

where the truncated vectors $\mathbf{Y}_0^*, \bar{\mathbf{Y}}_l^{(2)*}\in\mathbb{R}^{\hat{p}_1},\ l\in\{1,\ldots,L\}$ are defined now as $Y_{0j}^*=Y_{0j},\ Y_{lj}^{(2)*}=\bar{Y}_{lj}^{(2)},\ j\in\{1,\ldots,\hat{p}_1\}$, and $\boldsymbol{\Sigma}^*\in\mathbb{R}^{\hat{p}_1\times\hat{p}_1}$ is the corresponding upper left sub-matrix of $\boldsymbol{\Sigma}$, and $\rho_l=n_l^{(2)}/(n_l^{(2)}+1)$.

We have

$$\Pr(\hat{l}\neq l)\leq\Pr(\hat{l}\neq l\mid\hat{x}=x)+\Pr(\hat{x}\neq x), \tag{15}$$

where, due to the fact that different data was used for feature selection and classification, by Theorems 1 and 3, each probability on the RHS of (15) is at most α. Thus, the following result holds:

Theorem 4. *Consider the model* (1) *and the corresponding model* (5). *Assume the conditions* (7) *(with* n_l *replaced by* $n_l^{(2)}$*) and* (12) *hold for some* $0<\alpha\leq 1/2$. *Apply feature selection procedure* (10) *and use the selected features for classification via the rule* (14). *Then,*

$$\Pr(\text{correct classification})\geq 1-2\alpha.$$

3. Asymptotic analysis

Conditions (7) and (12) (or (13)) of Theorems 1 and 2, respectively, provide the non-asymptotic lower bounds on the minimal distance between different classes and the minimal effect of significant features required for the perfect feature selection and classification error bounded above by 2α. In order to gain better understanding of these conditions, we consider an asymptotic setup.

Standard asymptotics considered in classification literature assume that the number of features p and the sample sizes n_l increase whereas the number of classes L is fixed (see, e.g., [11,23] for $L=2$ and [20] for a general but fixed L). On the contrary, our study is motivated by the case where the number of classes may also be large ("large L, large p, small n").

Recall that $N = \sum_{l=1}^{L} n_l$ is the total sample size and let the number of features $p \to \infty$. Following [20], assume that all eigenvalues of the $p_1 \times p_1$ covariance matrix of significant features $\mathbf{\Sigma}^*$ are finite and bounded away from zero, i.e., there exist absolute constants τ_1 and τ_2 such that

$$0 < \tau_1 \le \lambda_{\min}(\mathbf{\Sigma}^*) \le \lambda_{\max}(\mathbf{\Sigma}^*) \le \tau_2 < \infty.$$

The samples sizes n_l within classes also grow with p. For simplicity of exposition, we assume that they are of the same asymptotic order and split more or less equally between the two sets ($\pi \sim 1/2$), that is, $n_l^{(1)} \sim n_l^{(2)} \sim n$ for all $l \in \{1, \ldots, L\}$, where $n = N/(2L)$ and $a \sim b$ means $a = b(1 + o(1))$. In such asymptotic setup, $\rho_l \sim 1 - 1/n$, while $\sqrt{1-\rho_l \rho_k} \sim \sqrt{2/n}$. Though the results in the previous section allow one to study various other settings with unequal class sizes, the asymptotic analysis of a vast variety of such possible scenarios is beyond the scope of this paper.

Consider now the condition (7) of Theorems 1 and 4 on the minimal separation Mahalanobis distance between any two class centers as p tends to infinity, while n, the number of significant features p_1 and the number of classes L may increase with p, and α may depend on n, p and L. Thus, (7) yields:

$$\min_{k \neq k'} (\boldsymbol{m}_k^* - \boldsymbol{m}_{k'}^*)^\top (\mathbf{\Sigma}^*)^{-1} (\boldsymbol{m}_k^* - \boldsymbol{m}_{k'}^*) \ge \Delta_*^2 \sim 8 \ln(L_1/\alpha) \left\{ 1 + \frac{1}{\sqrt{2n}} \left(1 + \sqrt{2p_1/\ln(L_1/\alpha)} \right) \right\}. \tag{16}$$

Define

$$\eta_1 = \lim_{p \to \infty} \sqrt{\frac{p_1}{n \ln(L_1/\alpha)}}$$

Depending on η_1, the condition (16) implies two possible asymptotic regimes for Δ_*^2:

$$\Delta_*^2 \sim \begin{cases} 8 \ln\left(\dfrac{L_1}{\alpha}\right)(1+\eta_1), & 0 \le \eta_1 < \infty, \quad \text{sparse regime — small number of significant features,} \\ 8\sqrt{\dfrac{p_1 \ln(L_1/\alpha)}{n}}, & \eta_1 = \infty, \quad \text{dense regime — large number of significant features.} \end{cases}$$

For sparse regime ($\eta_1 < \infty$), the required minimal between-class distance Δ_*^2 grows slowly as $\ln L$ and from Theorem 2 it immediately follows that this is the lowest possible rate for successful classification:

Proposition 1. *Let $L \to \infty$ and $p_1 \to \infty$ as $p \to \infty$. Let a new observation $\boldsymbol{Y}_0$ be from one of L classes. If*

$$\Delta_*^2 \sim 2\,\delta_{p_1} \ln L_1,$$

where $\delta_{p_1} \to 0$ arbitrarily slow as $p \to \infty$, then

$$\lim_{p\to\infty} \inf_{\psi} \max_{1 \le l \le L} \Pr_l(\psi(\boldsymbol{Y}_0) \neq l) = 1,$$

where $\Pr_l$ is the probability evaluated under the assumption that $\boldsymbol{Y}_0$ belongs to the lth class, and the infimum is taken over all classification rules $\psi(\boldsymbol{Y}_0) : \boldsymbol{Y}_0 \to \{1, \ldots, L\}$.

For dense regime, the number of significant features p_1 is large enough for the accumulated error of estimating p_1-dimensional $\mathbf{m}_k^*$'s by $\bar{\mathbf{Y}}_k^{(1)*}$'s to become dominant (see Section 2.2) and the classes should be, therefore, much stronger separated to deal with the curse of dimensionality.

It is natural that for successful classification the between-class distances should grow with L. Note, however, that unless the number of classes increases exponentially with p_1, the growth rate of Δ_*^2 is $o(p_1)$ and the corresponding average per-feature distances $\frac{1}{p_1}(\boldsymbol{m}_k^* - \boldsymbol{m}_{k'}^*)^\top (\mathbf{\Sigma}^*)^{-1} (\boldsymbol{m}_k^* - \boldsymbol{m}_{k'}^*)$ still tend to zero.

Similarly, from the condition (12) in Theorems 3 and 4 on the minimal effect for significant features required for the perfect feature selection, we have asymptotically

$$b_*^2 = \min_{1 \le j \le p_1} \sigma_j^{-2} b_j^2 \sim \frac{4}{n} \left(3\ln(2p/\alpha) + \sqrt{L_1 \ln(2p/\alpha)} \right).$$

Let

$$\eta_2 = \lim_{p\to\infty} \sqrt{\frac{\ln(2p/\alpha)}{L_1}}\,.$$

Then,

$$b_*^2 \sim \begin{cases} 4n^{-1}\sqrt{L_1 \ln(2p/\alpha)}(1+3\eta_2), & 0 \le \eta_2 < \infty, \quad \text{large number of classes,} \\ 12 n^{-1} \ln(2p/\alpha), & \eta_2 = \infty, \quad \text{small number of classes.} \end{cases} \tag{17}$$

and the threshold λ in (9) for feature selection can be presented as

$$\lambda \sim \begin{cases} L_1(1 + 2\eta_2 + 2\eta_2^2), & 0 \le \eta_2 < \infty, \\ 2\ln(2p/\alpha), & \eta_2 = \infty. \end{cases}$$

To gain some insight on the minimal required effect for a significant feature to contribute to classification as the number of classes increases, assume for simplicity that each significant feature has equal effects on each class, that is, β_{lj} in (5) vary only in signs: $\beta_{lj}^2=\beta_j^2$, $l\in\{1,\dots,L\}$. Since $0\le\eta_2<\infty$ implies that L is large, so that $L_1=L-1\sim L$, condition (17) yields as $p\to\infty$:

$$\beta_j^2\sim\begin{cases}4\sigma_j^2\,n^{-1}\,\eta_2(1+3\eta_2), & 0\le\eta_2<\infty, \quad \text{large number of classes,}\\ 12\sigma_j^2\,n^{-1}\,L^{-1}\ln(2p/\alpha), & \eta_2=\infty, \quad \text{small number of classes.}\end{cases} \tag{18}$$

Since η_2 is decreasing with L for a given value of α, the required minimal level for β_j^2 on the RHS of (18) decreases as L grows and, therefore, more significant features become manifested in classification for larger number of classes. Thus, while it might be hard to perform coarse classification with a set of weak features, their impacts grow as one considers finer and finer separation between objects (see also the corresponding remarks at the end of Section 2.3).

Although in this section our goal was to explore the case when $L\to\infty$, calculations above remain valid for a fixed value of L (commonly, $L=2$). In particular, if L is fixed and $n=o(p)$, conditions (16) and (18) are of the form $\Delta_*^2\sim C_1\sqrt{p_1/n}$ and $\beta_j^2\sim C_2n^{-1}\ln(p/\alpha)$, $C_1,C_2>0$ and are similar to those of Theorem 1 and Theorem 3 in [11]. See also the results of [8,9] and [17] for closely related setups.

4. Unknown covariance matrix

So far the covariance matrix $\mathbf{\Sigma}$ was assumed to be known. In practice, however, it should usually be estimated from the data. The standard MLE estimator based on the first sub-sample

$$\widehat{\mathbf{\Sigma}}^{(1)}=\frac{1}{N_1}\sum_{l=1}^{L}\sum_{i=1}^{n_l^{(1)}}\left(\mathbf{Y}_{il}^{(1)}-\bar{\mathbf{Y}}_l^{(1)}\right)\left(\mathbf{Y}_{il}^{(1)}-\bar{\mathbf{Y}}_l^{(1)}\right)^\top \tag{19}$$

and the similar unbiased pooled estimator commonly used in MANOVA behave poorly for high-dimensional data. However, under the sparsity assumption, the proposed classification procedure requires only to estimate the variances σ_j^2 in feature selection procedure (8) and the inverse of the upper left sub-matrix $\mathbf{\Sigma}^*\in\mathbb{R}^{\hat p_1\times\hat p_1}$ of $\mathbf{\Sigma}$ in classification rule (14). Thus, when $p_1\ll p$, a low-dimensional matrix $(\widehat{\mathbf{\Sigma}^*})^{-1}$ may still be a good estimator of the true sub-matrix $(\mathbf{\Sigma}^*)^{-1}$ and (under some additional mild conditions) may be used instead of the latter in (14).

Assume that $p\le\frac{\alpha}{2}\,e^{(N_1-L)/4}$. Replace σ_j^2 in (8) by $\hat\sigma_j^2=\widehat{\mathbf{\Sigma}}_{jj}^{(1)}$ and consider the feature selection procedure (10) with a somewhat larger threshold

$$\lambda_1=\frac{\lambda}{1-\kappa}, \tag{20}$$

where λ is the threshold (9) used for the case of known variances and

$$\kappa=\kappa(p,N_1,L,\alpha)=2\sqrt{\frac{\ln(2p/\alpha)}{N_1-L}}+2\,\frac{\ln(2p/\alpha)}{N_1-L}<1 \tag{21}$$

The following theorem shows that under slightly stronger conditions on the minimal required effect for significant features, the above feature selection procedure with estimated σ_j^2 still controls the probability of correct identification of the true subset of significant features.

Theorem 5. *Let $0<\alpha\le 1/2$ and assume that $p\le\frac{\alpha}{2}\,e^{(N_1-L)/4}$. Define indicator variables*

$$\hat x_j=I\{\hat\sigma_j^{-2}\sum_{l=1}^{L}n_l^{(1)}(\bar Y_{lj}^{(1)}-\bar Y_{\cdot j}^{(1)})^2>\lambda_1\} \tag{22}$$

for $j\in\{1,\dots,p\}$ with λ_1 given in (20). Assume that μ_* in (11) satisfies

$$\mu_*+L_1-2\sqrt{(L_1+2\mu_*)\ln(2p/\alpha)}>\lambda_1(1+\kappa). \tag{23}$$

Then,

$$\Pr(\hat x=x)\ge 1-2\alpha.$$

Consider now the classification procedure (14). In what follows we assume that $\mathbf{\Sigma}^*$ is non-singular. Consider an estimator $\widehat{\mathbf{\Sigma}^*}$ of $\mathbf{\Sigma}^*$ of the form

$$\widehat{\mathbf{\Sigma}^*}=\frac{1}{N_2}\sum_{l=1}^{L}\sum_{i=1}^{n_l^{(2)}}(\mathbf{Y}_{il}^{(2)*}-\bar{\mathbf{Y}}_l^{(2)*})(\mathbf{Y}_{il}^{(2)*}-\bar{\mathbf{Y}}_l^{(2)*})^\top,$$

where $\mathbf{Y}_{il}^{(2)*}$ are the corresponding $\hat p_1$-dimensional truncated versions of $\mathbf{Y}_{il}^{(2)}$.

Assign $\mathbf{Y}_0$ the $\hat{l}$th class by replacing the true (unknown) $(\boldsymbol{\Sigma}^*)^{-1}$ in (14) by $(\widehat{\boldsymbol{\Sigma}^*})^{-1}$:

$$\hat{l} = \underset{1\le l\le L}{\operatorname{argmin}} \left\{ \rho_l (\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*})^\top (\widehat{\boldsymbol{\Sigma}^*})^{-1} (\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*}) \right\}. \tag{24}$$

Then the following version of Theorem 4 holds:

Theorem 6. *Consider the model (1) and the corresponding model (5), where* $p \le \frac{\alpha}{2}\, e^{(N_1-L)/4}$,

$$\max\left\{L, 2\ln\left(\frac{2}{\alpha}\right)\right\} < p_1 < \frac{1}{4C_1}\left(\frac{\lambda_{\min}(\boldsymbol{\Sigma}^*)}{\lambda_{\max}(\boldsymbol{\Sigma}^*)}\right)^4 N_2 \tag{25}$$

for some $0 < \alpha < 1/4$ *and* C_1 *is an absolute constant specified in the proof. Denote*

$$\gamma_{p_1,N_2} = 2\frac{\lambda_{\max}^2(\boldsymbol{\Sigma}^*)}{\lambda_{\min}^2(\boldsymbol{\Sigma}^*)}\sqrt{\frac{C_1 p_1}{N_2}} \tag{26}$$

and note that $\gamma_{p_1,N_2} < 1$ *due to (25). Assume the condition (23) and a somewhat stronger version of the condition (7), namely,*

$$(\mathbf{m}_k^* - \mathbf{m}_{k'}^*)^\top (\boldsymbol{\Sigma}^*)^{-1} (\mathbf{m}_k^* - \mathbf{m}_{k'}^*) \ge \frac{8\ \ln(L_1/\alpha)}{(1-\gamma_{p_1,N_2})\min(\rho_k, \rho_{k'})} \times \left\{1 + \sqrt{\frac{1}{2\min\left(n_k^{(2)}, n_{k'}^{(2)}\right)} + \gamma_{p_1,N_2}^2} \cdot \left(1 + \sqrt{\frac{2p_1}{\ln(L_1/\alpha)}}\right)\right\} \tag{27}$$

Apply feature selection procedure (22) and use the selected features for classification via the rule (24). Then,

$$\Pr(\text{correct classification}) \ge 1 - 4\alpha.$$

Theorem 6 shows that for a sparse setup the proposed classification procedure can still be used when the covariance matrix is unknown and estimated from the data.

5. Examples

In this section we demonstrate the performance of the proposed feature selection and classification procedure on simulated and real-data examples. Its main goal is to illustrate the phenomenon of improving the accuracy as the number of classes grows as discussed in the previous sections.

We found that in practice there is no real need to split the original data and used the entire data set for both feature selection and classification.

5.1. Simulation study

Simulated examples follow the settings presented in [20].

We generated the class means as i.i.d. normal vectors $\boldsymbol{m}_l \sim \mathcal{N}(0, \sigma_m^2 \mathbf{X})$, $l \in \{1, \ldots, L\}$, where $\mathbf{X} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with $x_i = 1$ for p_1 indices and $x_i = 0$ for others. Since the vectors generated in this manner do not necessarily satisfy our assumptions, in order to reduce an impact of a particular choice of vectors $\boldsymbol{m}_l$, we generated M_1 replications of the class means. Furthermore, following the model (2), for each replication of class means $\boldsymbol{m}_l$, $l \in \{1, \ldots, L\}$ we generated M_2 sets of training samples $\bar{Y}_{lji} = m_{lj} + \epsilon_{lji}^*$, $j \in \{1, \ldots, p\}$, $i \in \{1, \ldots, n\}$, where ϵ_{lji}^* are i.i.d. $\mathcal{N}(0, n^{-1}\boldsymbol{\Sigma})$. Finally, for each of $M_1 \cdot M_2$ sets of training samples, we drew a test set of M_3 new vectors from randomly chosen classes as i.i.d. normal vectors $\mathcal{N}(\boldsymbol{m}_l, \boldsymbol{\Sigma})$.

We used the same three choices for covariance matrix $\boldsymbol{\Sigma}$ as in [20]. In Example 1 features were independent, i.e., $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$. In Example 2 we used the autoregressive covariance structure with $\boldsymbol{\Sigma}_{h_1,h_2} = \sigma^2\, 0.5^{|h_1-h_2|}$, while in Example 3 we set $\boldsymbol{\Sigma}_{h_1,h_2} = \sigma^2\,(0.5 + 0.5\, I\{h_1 = h_2\})$, $h_1, h_2 \in \{1, \ldots, p\}$ implying equal variances σ^2 and all covariances equal to $\sigma^2/2$ (compound symmetric structure). We carried out simulations with both the true covariance matrix $\boldsymbol{\Sigma}$ and its MLE $\widehat{\boldsymbol{\Sigma}}$ given by (19). Since the performances of feature selection and classification procedures in both cases were similar, in what follows we present only the results obtained with $\widehat{\boldsymbol{\Sigma}}$.

For each training sample we first carried out the feature selection procedure described above with the threshold λ_1 defined in (20) and $\alpha = 0.05$. Subsequently, we used the selected features for classifying M_3 vectors from the corresponding test set according to the rule (24). In the case when it delivered a non-unique solution, we chose one of the suggested solutions at random.

In all simulations we used $M_1 = M_2 = M_3 = 50$, $p = 500$, $\sigma = 1$ and $n = 20$. Note that classification precision depends on the variance ratio $\tau^2 = \sigma_m^2/(\sigma^2/n)$ that may be viewed as a signal-to-noise ratio. For this reason,

Table 1
Average proportions of false negative features for $p = 500$ and various values of L, p_1 and τ over $M_1 \cdot M_2 = 2500$ training samples.

p_1	τ	Example 1				Example 2			
		$L = 2$	$L = 10$	$L = 20$	$L = 50$	$L = 2$	$L = 10$	$L = 20$	$L = 50$
10	1	1.000	.996	.975	.785	1.000	1.000	.978	.788
	2	.936	.297	.033	.000	.991	.592	.186	.000
	3	.880	.158	.006	.000	.898	.147	.003	.000
50	1	1.000	.995	.976	.785	1.000	.995	.977	.783
	2	.975	.604	.187	.001	.979	.609	.172	.001
	3	.896	.158	.005	.000	.901	.146	.004	.000
100	1	1.000	.996	.975	.784	1.000	.996	.976	.782
	2	.976	.601	.177	.001	.981	.611	.169	.000
	3	.895	.149	.005	.000	.898	.142	.004	.000
200	1	1.000	.995	.976	.783	1.000	.995	.977	.783
	2	.975	.605	.172	.000	.980	.617	.175	.000
	3	.892	.150	.004	.000	.895	.150	.004	.000

we studied performance of feature selection and classification for various combinations of p_1, L and τ. In particular, we used $p_1 = 10, 50, 100, 200$, $L = 2, 10, 20, 50$ and several values of τ depending on p_1.

The results of simulations indicate that for such data generating model (somewhat different from that analyzed in the paper), the threshold λ_1 in (20) (as well as λ in (9) for the known variances) might be too high, especially for small values of τ. The latter led to an over-conservative feature selection procedure. Thus, in all simulations the feature selection procedure did not detect false positive features. The information on the proportions of false negative features (over the total number of significant features) for several combinations of p_1, L and τ over $M_1 \cdot M_2 = 2500$ training samples is summarized in Table 1 for Example 1 and Example 2 (the results for Example 3 were similar and we omit their presentation to save the space). In particular, Table 1 clearly shows that for small values of τ and small L, due to the over-conservative feature selection procedure, almost not a single significant feature has been detected and the resulting classification is then essentially reduced to just a pure random guess. However, for any τ the detection rate improves as L grows. The improvement rate is very fast for $\tau \geq 2$. Thus, for $L = 50$ the vast majority of significant features were detected in spite of high level of noise. As we have mentioned, this improves the classification precision since weaker significant features that remained latent in coarse classification become active and may have a strong impact with increasing L.

For each combination of p_1, L and τ we calculated the corresponding average misclassification errors: see Figs. 1–3 for Examples 1–3, respectively. Figs. 1–3 show similar behavior for all three examples. For any p_1 and L misclassification error tends to zero as τ increases. The decay is faster for larger p_1 – the more significant features, the easier is classification. The figures demonstrate also another interesting phenomenon: for moderate and large p_1, the larger L, the faster is the decay. As we have argued, this is due to the fact that the impact of weaker significant features becomes stronger with increasing L. For small τ (strong noise), misclassification errors are higher for larger number of classes L. This is naturally explained by the failure of feature selection procedure to detect significant features in this case (see comments above), so that the resulting classification is similar to a random guess with a misclassification error $1 - 1/L$ (see Figs. 1–3). However, as τ increases, even the first few detected significant features strongly improve classification precision.

5.2. Real-data example

We applied feature selection techniques discussed above to a data set of communication signals recorded from South American knife fishes of the genus Gymnotus. These nocturnally active freshwater fishes generate pulsed electrostatic fields from electric organ discharges (EODs). The three-dimensional electrostatic EOD fields of Gymnotus can be summarized by two-dimensional head-to-tail waveforms recorded from underwater electrodes placed in front of and behind a fish. EOD waveforms vary among species and are used by genus Gymnotus in order to recognize its own kind for more productive mating and other purposes.

The data set consists of 512-dimensional vectors of the Symmlet-4 discrete wavelet transform coefficients of signals obtained from eight genetically distinct species of Gymnotus (*G. arapaima* (G1), *G. coatesi* (G2), *G. coropinae* (G3), *G. curupira* (G4), *G. jonasi* (G5), *G. mamiraua* (G6), *G. obscurus* (G7), *G. varzea* (G8)) at various stages of their development. In particular, species were divided into six ontogenetic categories: postlarval (J0), small juvenile (J1), large juvenile (J2), immature adult (IA), mature male (M) and mature female (F). The EODs were recorded from 42 of 48 possible combinations of eight species and six categories. There are 677 samples from 42 classes with sizes varying from 3 to 69. The complete description of the data can be found in [6].

As it is evident from [6], there is no expectation that these groups should all be mutually separable: there are considerable overlaps between developmental stages of the same species as well as among juveniles of different species. For this reason, we reduced the number of classes to include only those species/categories that might be potentially separated. In particular, we ran our feature selection and classification procedure with the data sets comprised of 10 to

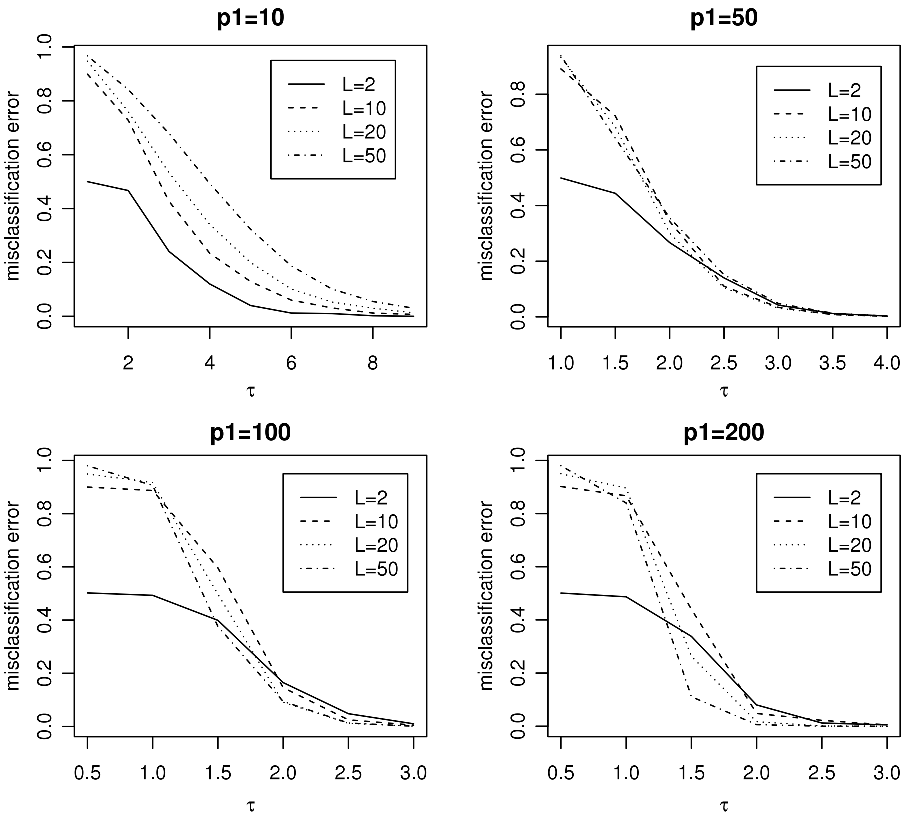

Fig. 1. Average misclassification errors as functions of τ for various combinations of p_1 and L for Example 1.

Table 2
The sample sizes of train (N_{train}) and test (N_{test}) sets, the numbers of selected significant features ($\hat{p}_1$) and misclassification errors with standard errors in brackets averaged over 100 splits for the Gymnotus fish data.

L	N_{train}	N_{test}	$\hat{p}_1$	Misclassification error
10	32	10	67.0	.077 (.006)
11	38	13	68.3	.092 (.006)
12	46	16	65.3	.127 (.007)
13	51	18	67.6	.166 (.007)
14	57	20	83.7	.149 (.006)
15	64	23	87.4	.130 (.006)
16	68	24	86.8	.162 (.007)

16 classes listed in the order they appear: G2-M, G4-M, G5-M, G1-F, G2-F, G5-F, G7-F, G8-F, G2-J1, G4-J1, G2-F, G1-J1, G7-AI, G1-F, G6-M, G7-J1.

We split the respective data sets into training and test parts. For this purpose, in each class we chose at random at most 1/3 of the total number of observations for validation leaving the rest of the data as training samples. Using those training samples, we carried out feature selection and subsequent classification of vectors in the test part of the data set. We repeated the process 100 times for various splits and recorded the average misclassification errors and their standard errors for each of the cases ($L \in \{10, \dots, 16\}$). Table 2 presents results of the study: the average sample sizes of train (N_{train}) and test (N_{test}) sets for each L, the average number of selected significant features ($\hat{p}_1$) and average misclassification error with the corresponding standard errors.

The table shows that when one starts with 10 well separated classes the misclassification error is initially grows when L increases from 10 to 13. However, at $L = 13$ there is a strong jump in the numbers of detected features and the misclassification errors again start to decrease when L grows from 13 to 15 due to better feature selection. For $L > 15$ the misclassification error grows again with L due to poor separation of juvenile Gymnotus EOD waveforms shapes.

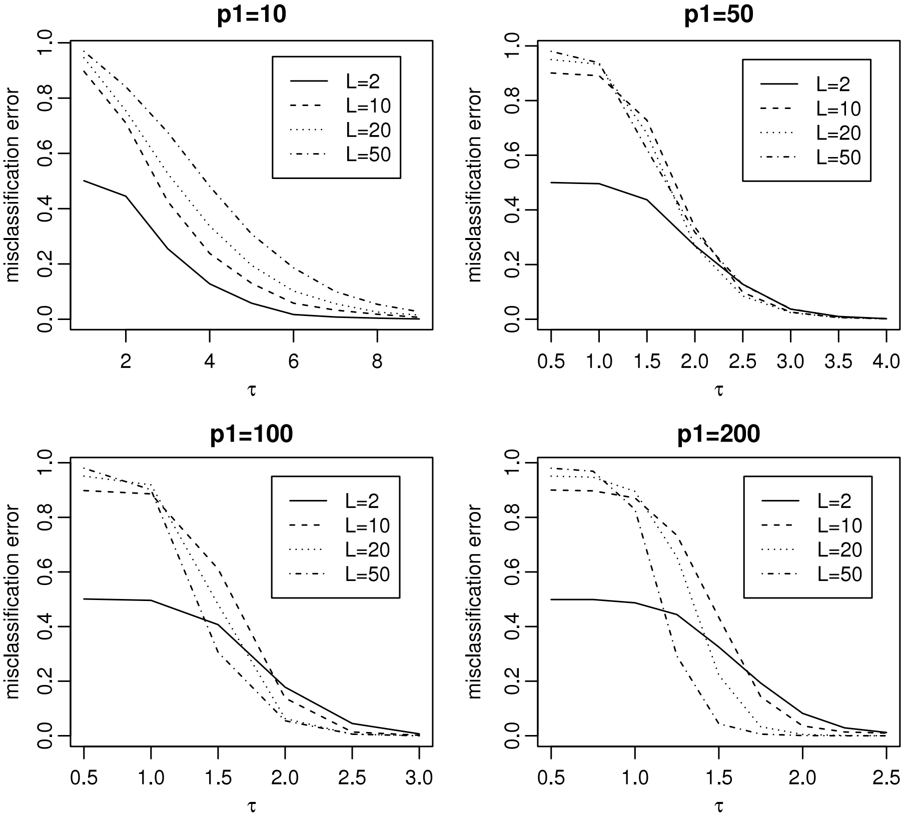

Fig. 2. Average misclassification errors as functions of τ for various combinations of p_1 and L for Example 2.

6. Concluding remarks

The paper considers multi-class classification of high-dimensional normal vectors, where the number of classes may diverge. This is a first attempt to rigorously study "large L, large p, small n" classification problem. Our main goal was not to propose a novel methodology but to explore interesting phenomena arising in such a new setup. In particular, our results indicate that the precision of classification can improve as the number of classes grows. This is, at first glance, a somewhat counter-intuitive conclusion and has not been observed so far due to shortage of literature on multi-class classification. It is explained by the fact that even weaker significant features, that might be undetected for smaller L, being shared across classes, can strongly contribute to successful classification when the number of classes is large. We believe that the results of the paper motivate further investigation of "large L, large p, small n" classification in other, more complicated setups.

The contents of this paper can be extended in a variety of ways. To begin with, an extension to different covariance matrices across the classes is straightforward. One can also allow different supports of sparsity for different clusters and/or relax the Gaussian assumption by considering sub-Gaussian or sub-exponential data in a similar way, though such generalizations will require to re-derive the corresponding conditions for correct classification.

Acknowledgments

Felix Abramovich was supported by the Israel Science Foundation (ISF), grant ISF-589/18. Marianna Pensky was partially supported by National Science Foundation (NSF), USA, grants DMS-1407475 and DMS-1712977. The authors would like to thank Vladimir Koltchinskii and Ruth Heller for valuable remarks, and Will Crampton for providing the data set used for the real data example. Helpful comments of the anonymous referee are gratefully acknowledged.

Appendix

We start from recalling two lemmas of [3] that will be used further in the proofs.

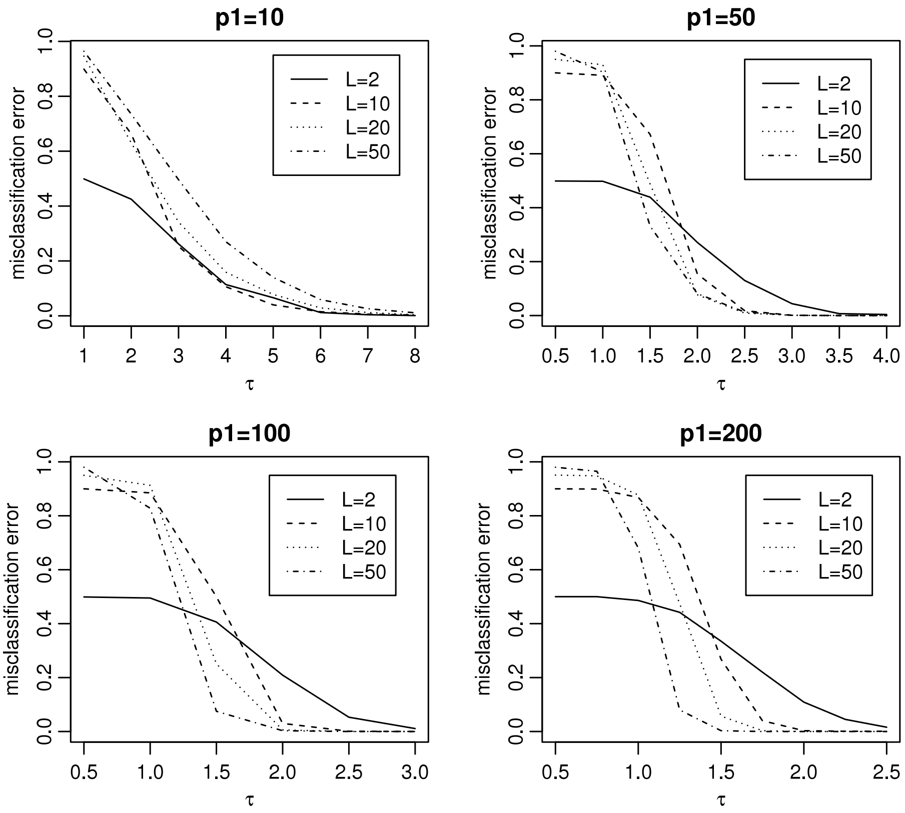

Fig. 3. Average misclassification errors as functions of τ for various combinations of p_1 and L for Example 3.

Lemma 1 (*Lemma 8.1 of [3]*). *Let $\zeta \sim \chi^2_{k,\mu}$, $\mu > 0$. Then, for any $x > 0$*

$$\Pr(\zeta > \mu + k + 2\sqrt{(k+2\mu)x} + 2x) \le e^{-x}$$

and

$$\Pr(\zeta < \mu + k - 2\sqrt{(k+2\mu)x}) \le e^{-x}.$$

Lemma 2 (*Lemma 8.2 of [3]*). *Let X be a random variable such that*

$$\ln\left\{\mathrm{E}\left(e^{sX}\right)\right\} \le \frac{(as)^2}{1-bs} \quad \textit{for} \quad 0 < s < b^{-1},$$

where a and b are positive constants. Then

$$\Pr\left(X \ge 2a\sqrt{x} + bx\right) \le e^{-x} \quad \textit{for all} \quad x > 0.$$

Proof of Theorem 1. Note that

$$\Pr(\hat{l} \ne l) = \sum_{k \ne l} \Pr(\hat{l} = k) \le L_1 \max_{k \ne l} \Pr(\hat{l} = k), \tag{28}$$

For a given $k \ne l$ define a $(2p_1)$-dimensional random vector $\widetilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y}_0^* - \mathbf{Y}_l^* \\ \mathbf{Y}_0^* - \mathbf{Y}_k^* \end{pmatrix}$, where the vectors $\mathbf{Y}_0^*$, $\mathbf{Y}_l^*$ and $\mathbf{Y}_k^*$ are defined just after (6). A straightforward calculus yields

$$\widetilde{\mathbf{Y}} \sim N(\boldsymbol{\theta}, \mathbf{V}) \quad \text{with} \quad \boldsymbol{\theta} = \begin{pmatrix} \mathbf{0}_{p_1} \\ \mathbf{m}_l^* - \mathbf{m}_k^* \end{pmatrix}, \quad \mathbf{V} = \sigma^2 \begin{pmatrix} \rho_l^{-1}\boldsymbol{\Sigma}^* & \boldsymbol{\Sigma}^* \\ \boldsymbol{\Sigma}^* & \rho_k^{-1}\boldsymbol{\Sigma}^* \end{pmatrix}, \tag{29}$$

where ρ_l is defined in (3). Then, it follows from (6) that

$$\Pr(\hat{l} = k) \le \Pr\left(\rho_l(\mathbf{Y}_0^* - \mathbf{Y}_l^*)^\top(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \mathbf{Y}_l^*) > \rho_k(\mathbf{Y}_0^* - \mathbf{Y}_k^*)^\top(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \mathbf{Y}_k^*)\right) = \Pr(\widetilde{\mathbf{Y}}^\top \mathbf{A}\widetilde{\mathbf{Y}} \ge 0),$$

where

$$\mathbf{A} = \begin{pmatrix} \rho_l\,(\boldsymbol{\Sigma}^*)^{-1} & \mathbf{0}_{p_1\times p_1} \\ \mathbf{0}_{p_1\times p_1} & -\rho_k\,(\boldsymbol{\Sigma}^*)^{-1} \end{pmatrix}.$$

Consider a random variable $\xi = \widetilde{\mathbf{Y}}^\top \mathbf{A}\widetilde{\mathbf{Y}}$. Since $\mathbf{V}^{-1}$ is a symmetric positive-definite matrix and $\mathbf{A}$ is symmetric, they can be simultaneously diagonalized, that is, there exists a matrix $\mathbf{W}$, such that $\mathbf{W}^\top\mathbf{V}^{-1}\mathbf{W} = \mathbf{I}$ and $\mathbf{W}^\top\mathbf{A}\mathbf{W} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues $\varphi_j,\ j = 1,\dots,2p_1$ of $\mathbf{R} = \mathbf{VA}$. Then, from the known results on the distribution of quadratic forms of normal variables (e.g., [16]), ξ can be represented as a weighted sum of independent (generally) non-central chi-square variables, namely,

$$\xi = \sum_{j=1}^{2p_1} \varphi_j \chi^2_{1,\eta_j^2}, \tag{30}$$

where $\boldsymbol{\eta}$ is such that $\boldsymbol{\theta} = \mathbf{W}\boldsymbol{\eta}$ with $\boldsymbol{\theta}$ given by (29). By a straightforward matrix calculus, obtain

$$\mathbf{R}^2 = \begin{pmatrix} (1-\rho_k\rho_l)\ \mathbf{I}_{p_1} & \mathbf{0}_{p_1\times p_1} \\ \mathbf{0}_{p_1\times p_1} & (1-\rho_k\rho_l)\ \mathbf{I}_{p_1} \end{pmatrix}$$

and, therefore, all eigenvalues $\varphi_j,\ j\in\{1,\dots,2p_1\}$ of a matrix $\mathbf{R} = \mathbf{VA}$ are of the forms

$$\varphi_j = \pm\varphi_*, \quad \varphi_* = \sqrt{1-\rho_k\rho_l} \tag{31}$$

for $j \in \{1,\dots,2p_1\}$.

Consider now the logarithm of the moment generating function of the centered random variable $\xi - \mathrm{E}(\xi)$, where ξ is defined in (30). We have $\mathrm{E}\xi = \sum_{j=1}^{2p_1}\varphi_j(1+\eta_j^2) = \sum_{j=1}^{2p_1}\varphi_j\eta_j^2$, where recall that $\mathbf{W}\boldsymbol{\eta} = \boldsymbol{\theta}$. Hence, using formula (31), for $s < 1/(2\varphi_*)$, we have

$$\begin{aligned} \ln \mathrm{E}e^{s(\xi - E\xi)} &= \sum_{j=1}^{2p_1} \frac{\eta_j^2\varphi_j s}{1-2\varphi_j s} - \frac{1}{2}\sum_{j=1}^{2p_1}\ln(1-2\varphi_j s) - s\sum_{j=1}^{2p_1}\varphi_j(1+\eta_j^2) \\ &= \sum_{j=1}^{2p_1}\left(\frac{\eta_j^2\varphi_j s}{1-2\varphi_j s} - \eta_j^2\varphi_j s\right) - \frac{1}{2}\sum_{j=1}^{2p_1}\left(\ln(1-2\varphi_j s) + 2\varphi_j s\right) \ \le\ \sum_{j=1}^{2p_1}\frac{2s^2\eta_j^2\varphi_*^2}{1-2\varphi_j s} + \sum_{j=1}^{2p_1}\frac{s^2\varphi_*^2}{1-2\varphi_j s} \\ &\le\ \frac{2s^2}{1-2\varphi_* s}\ \varphi_*^2\|\boldsymbol{\eta}\|^2 + \frac{2s^2\varphi_*^2 p_1}{1-4\varphi_*^2 s^2}\ \le\ \frac{2s^2}{1-2\varphi_* s}\ \varphi_*^2\|\boldsymbol{\eta}\|^2 + \frac{2s^2\varphi_*^2 p_1}{1-2\varphi_* s}\,. \end{aligned}$$

Denote

$$\Delta^2 = (\mathbf{m}_l^* - \mathbf{m}_k^*)^\top(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{m}_l^* - \mathbf{m}_k^*)$$

Using $\mathbf{W}^\top\mathbf{V}^{-1}\mathbf{W} = \mathbf{I}$, $\mathbf{W}^\top\mathbf{A}\mathbf{W} = \boldsymbol{\Lambda}$ and $\mathbf{W}\boldsymbol{\eta} = \boldsymbol{\theta}$, one can verify that $\varphi_*^2\|\boldsymbol{\eta}\|^2 = \boldsymbol{\eta}^\top\boldsymbol{\Lambda}^2\boldsymbol{\eta} = \boldsymbol{\theta}^\top\mathbf{AVA}\boldsymbol{\theta} = \rho_k\,\Delta^2$, where $\boldsymbol{\theta}$ and V are defined in (29). Thus,

$$\ln \mathrm{E}e^{s(\xi - E\xi)} \le \frac{a^2s^2}{1-bs},$$

where $b = 2\varphi_*$ and

$$a = \sqrt{2\rho_k\,\Delta^2 + 2\varphi_*^2 p_1} \le \sqrt{2}\left(\sqrt{\rho_k}\ |\Delta| + \varphi_*\sqrt{p_1}\right).$$

In addition,

$$\mathrm{E}\xi = \boldsymbol{\eta}^\top\boldsymbol{\Lambda}\boldsymbol{\eta} = \boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta} = -\rho_k\ \Delta^2.$$

A straightforward calculus shows that, under the condition (7) of Theorem 1, one has $\rho_k\,\Delta^2 \ge 2a\sqrt{\ln(L_1/\alpha)} + b\ln(L_1/\alpha)$. Then, applying Lemma 2, one obtains

$$\Pr(\xi > 0) \le \Pr\left(\xi \ge -\rho_k\ \Delta^2 + 2a\sqrt{\ln(L_1/\alpha)} + b\ln(L_1/\alpha)\right) \le \frac{\alpha}{L_1}$$

that, together with (28), completes the proof.

Proof of Theorem 3. Let $\hat{p}_{01} = \sum_{j=1}^p I\{\hat{x}_j = 1 \mid x_j = 0\}$ and $\hat{p}_{11} = \sum_{j=1}^p I\{\hat{x}_j = 1 \mid x_j = 1\}$ be the numbers of erroneously and truly identified significant features respectively, where obviously $\hat{p}_{01}$ and $\hat{p}_{11}$ are independent, and $\hat{p}_{01} + \hat{p}_{11} = \hat{p}_1$. Note that

$$\Pr(\hat{x} \ne x) \le \Pr(\hat{p}_{01} > 0) + \Pr(\hat{p}_{11} < p_1).$$

Recall that for $x_j = 0$, the corresponding $\zeta_j \sim \chi^2_{L_1}$. Let $u_j,\ j \in \{1,\ldots,p_0\}$ be any, possibly correlated, $\chi^2_{L_1}$ random variables. Then,

$$\Pr(\hat{p}_{01} > 0) = \Pr\left(\max_{1\le j\le p_0} u_j > \lambda\right) \le p\ \Pr\left(u_j > L_1 + 2\sqrt{L_1 \ln(2p/\alpha)} + 2\ln(2p/\alpha)\right).$$

Apply Lemma 1 for the particular case $\mu = 0$ to obtain

$$\Pr\left(u_j > L_1 + 2\sqrt{L_1 \ln(2p/\alpha)} + 2\ln(2p/\alpha)\right) \le \frac{\alpha}{2p},$$

so that $\Pr(\hat{p}_{01} > 0) \le \alpha/2$. Similarly, let $\mu_* = \min_{1\le j\le p_1} \mu_j = \min_{1\le j\le p_1} \sigma_j^{-2}\ \sum_{l=1}^{L} n_l^{(1)} \beta_{lj}^2$ and consider any, possibly correlated, non-central chi-squared variables $v_j \sim \chi^2_{L_1;\mu_*},\ j \in \{1,\ldots,p_1\}$. We have

$$\Pr(\hat{p}_{11} < p_1) \le \Pr\left(\min_{1\le j\le p_1} v_j \le \lambda\right) \le p\ \Pr\left(v_j < \lambda\right).$$

A straightforward calculus shows that, under the condition (12) on μ_*, one has $\mu_* + L_1 - 2\sqrt{(L_1 + 2\mu_*)\ln(2p/\alpha)} > \lambda$. Thus, Lemma 1 yields $\Pr(v_j < \lambda) \le \alpha/(2p)$ and, therefore, $\Pr(\hat{p}_{11} < p_1) \le \alpha/2$, which completes the proof.

Proof of Theorem 5. We start with the following lemma:

Lemma 3.

$$\Pr\left(\max_{1\le j\le p} \left|\hat{\sigma}_j^2/\sigma^2 - 1\right| \le \kappa\right) \ge 1 - \alpha,$$

where κ was defined in (21).

Let $\mathcal{A}$ be the event $\{\max_{1\le j\le p} \left|\hat{\sigma}_j^2/\sigma^2 - 1\right| \le \kappa\}$ and $I_{\mathcal{A}}$ its indicator. By Lemma 3,

$$\Pr(\hat{x} \ne x) \le \Pr\left((\hat{x} \ne x) I_{\mathcal{A}}\right) + \alpha, \tag{32}$$

where

$$\Pr\left((\hat{x} \ne x) I_{\mathcal{A}}\right) \le \Pr\left((\hat{p}_{01} > 0) I_{\mathcal{A}}\right) + \Pr\left((\hat{p}_{11} < p_1) I_{\mathcal{A}}\right). \tag{33}$$

Let $\hat{\zeta}_j = \hat{\sigma}_j^{-2}\ \sum_{l=1}^{L} n_l^{(1)} (\bar{Y}_{lj}^{(1)} - \bar{Y}_{\cdot j}^{(1)})^2$. Then, on the event $\mathcal{A}$

$$\Pr\left((\hat{\zeta}_j > \lambda_1) I_{\mathcal{A}} \mid x_j = 0\right) = \Pr\left(\left(u_j > \lambda_1\, \hat{\sigma}_j^2/\sigma_j^2\right) I_{\mathcal{A}}\right) \le \Pr(u_j > \lambda),$$

where $u_j \sim \chi^2_{L_1},\ j \in \{1,\ldots,p_0\}$. Hence, following the arguments of Theorem 3, by Lemma 1

$$\Pr\left((\hat{p}_{01} > 0) I_{\mathcal{A}}\right) \le \Pr\left((\max_{1\le j\le p} \hat{\zeta}_j > \lambda_1) I_{\mathcal{A}} \mid x_j = 0\right) \le \Pr(\max_{1\le j\le p_0} u_j > \lambda) \le \frac{\alpha}{2}. \tag{34}$$

Similarly, $\Pr\left((\hat{\zeta}_j < \lambda_1) I_{\mathcal{A}} \mid x_j = 1\right) \le \Pr\left(v_j < \lambda_1(1+\kappa)\right)$, where $v_j \sim \chi^2_{L_1;\mu_*}, j \in \{1,\ldots,p_1\}$. Then, under the condition (12) of the theorem, Lemma 1 yields

$$\Pr\left((\hat{p}_{11} < p_1) I_{\mathcal{A}}\right) \le \Pr\left(\min_{1\le j\le p_1} v_j \le \lambda_1(1+\kappa)\right) \le \frac{\alpha}{2}. \tag{35}$$

Combination of (32)–(35) completes the proof.

Proof of Theorem 6. Assume that $\mathbf{Y}_0$ is from the lth class. From (15) we have $\Pr(\hat{l} \ne l) \le \Pr(\hat{l} \ne l \mid \hat{x} = x) + \Pr(\hat{x} \ne x)$, where $\Pr(\hat{x} \ne x) \le 2\alpha$ by Theorem 5. Consider a set $\Omega = \{\omega : \ \hat{x} = x\}$ with $\Pr(\Omega) \ge 1 - \alpha$. In order to bound above $\Pr(\hat{l} \ne l \mid \hat{x} = x)$ we assume that $\omega \in \Omega$. We will use the following two lemmas:

Lemma 4. *If $\|\widehat{\mathbf{\Sigma}^*} - \mathbf{\Sigma}^*\| \le \lambda_{\min}(\mathbf{\Sigma}^*)/2$, then $\|(\widehat{\mathbf{\Sigma}^*})^{-1} - (\mathbf{\Sigma}^*)^{-1}\| \le 2\, \lambda_{\min}^{-2}(\mathbf{\Sigma}^*)\, \|\widehat{\mathbf{\Sigma}^*} - \mathbf{\Sigma}^*\|$.*

Lemma 5. *Under the condition (25), $\Pr\left(\|\widehat{\mathbf{\Sigma}^*} - \mathbf{\Sigma}^*\| \le \lambda_{\max}(\mathbf{\Sigma}^*)\sqrt{\frac{C_1 p_1}{N_2}}\right) \ge 1 - 2\alpha$.*

From Lemmas 4 and 5 it follows that under (25),

$$\Pr\left(\|(\widehat{\mathbf{\Sigma}^*})^{-1} - (\mathbf{\Sigma}^*)^{-1}\| \le \gamma_{p_1,N_2}\right) \ge 1 - 2\alpha, \tag{36}$$

where γ_{p_1,N_2} is defined in (26). Furthermore, for any $1 \le k \le L$,

$$\frac{(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^*)^\top \left((\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\right)(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^*)}{(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^*)^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^*)} \le \|\boldsymbol{\Sigma}^* \left((\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\right)\| \le \tau_2 \|(\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\| \,. \tag{37}$$

Since the sample mean and the sample covariance matrix are independent in the case of the normal distribution, inequalities (36) and (37) imply that with probability at least $1 - 2\alpha$

$$\begin{aligned}
&\rho_l\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*})^\top (\widehat{\boldsymbol{\Sigma}^*})^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*}) - \rho_k\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})^\top (\widehat{\boldsymbol{\Sigma}^*})^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})\\
&= \rho_l\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*})^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*}) - \rho_k\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})\\
&+ \rho_l\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*})^\top \left((\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\right)(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*}) - \rho_k\,(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})^\top \left((\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\right)(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})\\
&\le \rho_l(1 + \gamma_{p_1,N_2})(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*})^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_l^{(2)*}) - \rho_k(1 - \gamma_{p_1,N_2})(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})^\top (\boldsymbol{\Sigma}^*)^{-1}(\mathbf{Y}_0^* - \bar{\mathbf{Y}}_k^{(2)*})\,.
\end{aligned}$$

Define $\rho_l' = \rho_l(1 + \gamma_{p_1,N_2})$ and $\rho_k' = \rho_k(1 - \gamma_{p_1,N_2})$. In particular, note that $\rho_l'\rho_k' = \rho_l\rho_k(1 - \gamma_{p_1,N_2}^2)$. Repeating the proof of Theorem 1 but with ρ_l' and ρ_k' and under the stronger condition (27), obtain $\Pr(\hat{l} \ne l \mid \hat{x} = x) \le 2\alpha$ that, together with (15) and $\Pr(\hat{x} \ne x) \le 2\alpha$, completes the proof.

Proof of Lemma 3. Note that $\sigma_j^{-2}(N_1 - L)\hat{\sigma}_j^2 \sim \chi^2_{N_1-L}$ and apply Lemma 1 to obtain $\Pr(|\hat{\sigma}_j^2/\sigma^2 - 1| \ge \kappa) \le \alpha/p$ for all $j \in \{1, \dots, p\}$ and, therefore, $\Pr\left(\max_{1\le j\le p} |\hat{\sigma}_j^2/\sigma^2 - 1| \ge \kappa\right) \le \alpha$.

Proof of Lemma 4. Under the condition of the lemma we have

$$\|(\widehat{\boldsymbol{\Sigma}^*})^{-1}\|^{-1} = \min_{\|\mathbf{a}\|=1} \mathbf{a}^\top \widehat{\boldsymbol{\Sigma}^*}\mathbf{a} \ge \min_{\|\mathbf{a}\|=1} \mathbf{a}^\top \boldsymbol{\Sigma}^*\mathbf{a} - \max_{\|\mathbf{a}\|=1} \mathbf{a}^\top (\widehat{\boldsymbol{\Sigma}^*} - \boldsymbol{\Sigma}^*)\mathbf{a} \ge \lambda_{\min}(\boldsymbol{\Sigma}^*)/2$$

and, therefore,

$$\|(\widehat{\boldsymbol{\Sigma}^*})^{-1} - (\boldsymbol{\Sigma}^*)^{-1}\| \le \|(\widehat{\boldsymbol{\Sigma}^*})^{-1}\| \cdot \|\widehat{\boldsymbol{\Sigma}^*} - \boldsymbol{\Sigma}^*\| \cdot \|(\boldsymbol{\Sigma}^*)^{-1}\| \le 2\lambda_{\min}^{-2}(\boldsymbol{\Sigma}^*)\,\|\widehat{\boldsymbol{\Sigma}^*} - \boldsymbol{\Sigma}^*\| \,.$$

Proof of Lemma 5. Define $\mathbf{Z}_{il} = \left(\mathbf{Y}_{il}^*\right)^{(2)} - \mathbf{m}_l^* \sim \mathcal{N}(\mathbf{0}_{p_1}, \boldsymbol{\Sigma}^*)$, $i \in \{1, \dots, n_l^{(2)}\}$, $l \in \{1, \dots, L\}$. The sample covariance matrix is translation invariant and, therefore,

$$\widehat{\boldsymbol{\Sigma}^*} = \frac{1}{N_2}\sum_{l=1}^{L}\sum_{i=1}^{n_l^{(2)}}(\mathbf{Z}_{il} - \bar{\mathbf{Z}}_l)(\mathbf{Z}_{il} - \bar{\mathbf{Z}}_l)^\top = \frac{1}{N_2}\sum_{l=1}^{L}\sum_{i=1}^{n_l^{(2)}}\mathbf{Z}_{il}\mathbf{Z}_{il}^\top - \frac{1}{N_2}\sum_{l=1}^{L} n_l^{(2)}\,\bar{\mathbf{Z}}_l\bar{\mathbf{Z}}_l^\top = \mathbf{S}_1 - \mathbf{S}_2.$$

Thus,

$$\|\widehat{\boldsymbol{\Sigma}^*} - \boldsymbol{\Sigma}^*\| \le \|\mathbf{S}_1 - \boldsymbol{\Sigma}^*\| + \|\mathbf{S}_2\| \,. \tag{38}$$

By Remark 5.51 of [25], under the conditions of the lemma there exists an absolute constant C_0 such that

$$\Pr\left(\|\mathbf{S}_1 - \boldsymbol{\Sigma}^*\| \le \tau_2\sqrt{\frac{C_0 p_1}{N_2}}\right) \ge 1 - \alpha. \tag{39}$$

Consider now S_2. Define a matrix $\bar{\mathbf{Z}} \in \mathbb{R}^{p_1\times L}$ with columns $\bar{\mathbf{Z}}_l$, $l \in \{1, \dots, L\}$ and the diagonal matrix $\mathbf{D} = \operatorname{diag}\left(\sqrt{n_1^{(2)}}, \dots, \sqrt{n_L^{(2)}}\right)$. It is easy to see that $\mathbf{S}_2 = N^{-1}(\bar{\mathbf{Z}}\mathbf{D})(\bar{\mathbf{Z}}\mathbf{D})^\top$ and that matrix $\boldsymbol{\Xi} = (\boldsymbol{\Sigma}^*)^{-1/2}\bar{\mathbf{Z}}\mathbf{D}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. Indeed, columns $\boldsymbol{\Xi}_l = \sqrt{n_l^{(2)}}(\boldsymbol{\Sigma}^*)^{-1/2}\,\bar{\mathbf{Z}}_l$ of matrix $\boldsymbol{\Xi}$ are independent with $\mathrm{Cov}(\boldsymbol{\Xi}_l) = \mathbf{I}_{p_1}$. Hence,

$$\|\mathbf{S}_2\| = N_2^{-1}\,\|\bar{Z}\mathbf{D}\|^2 = N_2^{-1}\,\|\sqrt{\boldsymbol{\Sigma}^*}\,\boldsymbol{\Xi}\|^2 \le N_2^{-1}\,\lambda_{\max}(\boldsymbol{\Sigma}^*)\|\boldsymbol{\Xi}\|^2.$$

Then, by Corollary 5.35 of [25]

$$\Pr\left(\|\mathbf{S}_2\| \le N_2^{-1}\,\lambda_{\max}(\boldsymbol{\Sigma}^*)\left(\sqrt{p_1} + \sqrt{L} + \sqrt{2\ln(2/\alpha)}\right)^2\right) \ge 1 - \alpha$$

that, under (25), yields

$$\Pr\left(\|\mathbf{S}_2\| \le 9\lambda_{\max}(\boldsymbol{\Sigma}^*)N_2^{-1}\,p_1\right) \ge 1 - \alpha. \tag{40}$$

Combination of (38)–(40) completes the proof with $C_1 = \max(\sqrt{C_0}, 9)$.

References

[1] E. Arias-Castro, E.J. Candès, Y. Plan, Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism, Ann. Statist. 39 (2011) 2533–2556.

[2] P. Bickel, E. Levina, Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, Bernoulli 10 (2004) 989–1010.

[3] L. Birgé, An alternative point of view on lepski's method, in: M.C.M. van Zwet, C.A.J. de Gunst, A.W. Klaassen, van der Vaart (Eds.), State of the Art in Probability and Statistics, Festschrift for Willem R, in: Lecture Notes-Monograph Series, vol. 36, Institute of Mathematical Statistics, 2001, pp. 113–133.

[4] S. Boucheron, O. Bousquet, G.G. Lugosi, Theory of classification: a survey of some recent advances, ESAIM: Prob. Statist. 9 (2005) 323–375.

[5] K. Crammer, Y. Singer, On the algorithmic implementation of multiclass kernel-based vector machines, J. Mach. Learn. Res. 2 (2001) 265–292.

[6] W.R.G. Crampton, N.R. Lovejoy, J.C. Waddell, Reproductive character displacement and signal ontogeny in a sympatric assemblage of electric fish, Evolution 65 (2011) 1650–1666.

[7] J. Davis, M. Pensky, W. Crampton, Bayesian feature selection for classification with possibly large number of classes, J. Statist. Plan. Inf. 141 (2011) 3256–3266.

[8] D. Donoho, J. Jin, Feature selection by higher criticism thresholding achieves the optimal phase diagram, Phil. Trans. R. Soc. Ser. A 367 (2009) 4449–4470.

[9] D. Donoho, J. Jin, Impossibility of successful classication when useful features are rare and weak, Proc. Natl. Acad. Sci. 106 (2009) 8859–8864.

[10] S. Escalera, D.M.J. Tax, O. Pujol, P. Radeva, R.P.W. Duin, Multi-class classification in image analysis via error-correcting output codes, in: H. Kwasnicka, L.C. Jain (Eds.), Innovations in Intelligent Image Analysis, Springer-Verlag, Berlin, 2011, pp. 7–29.

[11] J. Fan, Y. Fan, High-dimensional classification using feature annealed independence rules, Ann. Statist. 36 (2008) 2605–2637.

[12] C. Giraud, Introduction to High-Dimensional Statistics, CRC Press, 2015.

[13] M.R. Gupta, S. Bengio, J. Weston, Training highly multiclass classifiers, J. Mach. Learn. Res. 15 (2014) 1461–1492.

[14] S.I. Hill, A. Doucet, A framework for kernel-based multi-category classification, J. Artif. Intell. Res. 30 (2007) 525–564.

[15] I.A. Ibragimov, R.Z. Hasminskii, Statistical Estimation. Asymptotic Theory, Springer-Verlag, New York, 1981.

[16] J.P. Imhof, Computing the distribution of quadratic forms in normal variables, Biometrika 48 (1961) 419–426.

[17] Y.I. Ingster, C. Pouet, A.B. Tsybakov, Classification of sparse high-dimensional vectors, Phil. Trans. R. Soc. Ser. A 367 (2009) 4427–4448.

[18] P. Jain, A. Kapoor, Active learning for large multi-class problems, in: Proc. IEEE Conf. Comput. Vis. and Pattern Recogn. (CVPR), 2009, pp. 762–769.

[19] Y. Lee, Y. Lin, G. Wahba, Multicategory support vector machines theory and application to the classification of microarray data and satellite radiance data, J. Amer. Statist. Assoc. 99 (2004) 67–81.

[20] R. Pan, H. Wang, R. Li, Ultrahigh-dimensional multiclass linear discriminant analysis for pairwise sure independence screening, J. Amer. Statist. Assoc. 111 (2016) 169–179.

[21] N. Parrish, M.R. Gupta, Dimensionality reduction by local discriminative Gaussians, in: Proc. 29th Int. Conf. Mach. Learn, 2012, pp. 559–566.

[22] O. Russakovsky, J. Jia Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C.M. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.

[23] J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis by thresholding for high-dimensional data, Ann. Statist. 39 (2011) 1241–1265.

[24] A. Tewari, P.L. Bartlett, On the consistency of multiclass classification methods, J. Mach. Learn. Res. 8 (2007) 1007–1025.

[25] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in: Y.C. Eldar, G.G. Kutyniok (Eds.), Compressed Sensing. Theory and Applications, Cambridge University Press, 2012, pp. 210–268.