Statistica Sinica Preprint No: SS-2022-0219	
Title	A Locally Adaptive Shrinkage Approach to False Selection Rate Control in High-Dimensional Classification
Manuscript ID	SS-2022-0219
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0219
Complete List of Authors	Bowen Gang, Yuantao Shi and Wenguang Sun
Corresponding Authors	Bowen Gang
E-mails	bgang@fudan.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

A Locally Adaptive Shrinkage Approach to False Selection Rate Control in High-Dimensional Classification

Bowen Gang[1], Yuantao Shi[2] and Wenguang Sun[3]

[1]*Fudan University,* [2]*University of Chicago and* [3]*Zhejiang University*

Abstract:

The uncertainty quantification and error control of classifiers are crucial in many high-consequence decision-making scenarios. We propose a selective classification framework that provides an "indecision" option for any observations that cannot be classified with confidence. The false selection rate (FSR), defined as the expected fraction of erroneous classifications among all definitive classifications, provides a useful error rate notion that trades off a fraction of indecisions for fewer classification errors. We develop a new class of locally adaptive shrinkage and selection (LASS) rules for FSR control in the context of high-dimensional linear discriminant analysis (LDA). LASS is easy-to-analyze, has robust performance across sparse and dense regimes, and controls the FSR under weaker conditions compared to existing methods. The empirical performances of LASS are investigated using both simulated and real data.

Key words and phrases: Classification with confidence, False discovery rate, Linear discriminant analysis, Risk control, Shrinkage estimation.

1. Introduction

Linear discriminant analysis (LDA) has been widely used in classification problems. We focus on the basic setup, which assumes that the observations are p-dimensional vector-valued features that are drawn with equal probability from one of the two multivariate normal distributions:

$$\mathcal{N}(\boldsymbol{\mu}_1, \Sigma) \text{ (class 1) and } \mathcal{N}(\boldsymbol{\mu}_2, \Sigma) \text{ (class 2).} \tag{1.1}$$

Let $\boldsymbol{W} \in \mathcal{R}^p$ be a new observation. Denote $\boldsymbol{\mu} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$ and $\boldsymbol{d} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The procedure that achieves the minimal misclassification risk is Fisher's linear discriminant rule:

$$\delta^F = \mathbb{I}\left\{(\boldsymbol{W} - \boldsymbol{\mu})^\top \Sigma^{-1} \boldsymbol{d} < 0\right\} + 2 \cdot \mathbb{I}\left\{(\boldsymbol{W} - \boldsymbol{\mu})^\top \Sigma^{-1} \boldsymbol{d} \geq 0\right\}, \tag{1.2}$$

which assigns $\boldsymbol{W}$ to class c if $\delta^F = c$, $c = 1, 2$. When $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ are unknown, the common practice is to construct a data-driven LDA rule by obtaining suitable estimates of the unknown quantities in (1.2). In the high-dimensional setting, naive sample estimates become highly unstable, and a plethora of regularized LDA rules have been proposed and shown to achieve substantial improvements in prediction accuracy (Friedman, 1989;

Tibshirani et al., 2003; Witten and Tibshirani, 2009; Cai and Liu, 2011; Shao et al., 2011; Mai et al., 2012; Cai and Zhang, 2019; among others). However, it still remains unknown how to assess the uncertainties and control the decision errors in high-dimensional LDA. This article proposes a selective classification approach to controlling the false selection rate (FSR). We develop a new class of data-driven LDA rules based on locally adaptive shrinkage and selection (LASS), and illustrate how LASS can be deployed in decision-making scenarios to control the FSR at a user-specified level.

1.1 Selective classification and false selection rate

Uncertainty quantification and error control are crucial in many sensitive decision-making scenarios. The decision errors, which can be very expensive to correct, are often unavoidable due to the intrinsic ambiguity of a classification problem. Consider the ideal setting where the multivariate normal parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ are known. Then among all classification rules, the LDA rule (1.2) achieves the minimum classification risk $1-\Phi\left(\frac{1}{2}\sqrt{\boldsymbol{d}^\top\Sigma^{-1}\boldsymbol{d}}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable. However, this minimum risk can still be unacceptably high when the signal to noise ratio $\sqrt{\boldsymbol{d}^\top\Sigma^{-1}\boldsymbol{d}}$ is low. The issue is exacerbated in practice, particularly in high-dimensional settings,

where we must employ "plug-in" rules learned from limited training data.

In contrast with conventional classification algorithms which are forced to make classifications on all new observations, a useful strategy for uncertainty quantification involves providing an *indecision* option (also referred to as abstention or reject option) for any observations which cannot be classified with confidence. The observations with indecisions can then be separately evaluated. This strategy is attractive in practice when the cost of handling indecisions is less than that of fixing a classification error. To see how it aligns with the social and policy objectives, consider a high-consequence classification scenario where one needs to assess the likelihood of a defendant becoming a recidivist. Obviously the social cost of incorrectly classifying a low-risk individual as a recidivist is much higher than that of an indecision – – it is worthy of waiting and collecting additional contextual knowledge of the convicted individual to mitigate the ambiguity. Likewise, in medical screening, a misclassification can result in either missed medical care or unnecessary treatments, both of which can be much more expensive than turning the patient over for a more careful examination/evaluation.

Suppose we observe labeled training data $\mathcal{D}^{train}$. The goal is to predict the classes for m new observations $\mathcal{D}^{test} = \{\boldsymbol{W}_j : 1 \leq j \leq m\}$. This article considers a *selective classification* framework that only makes definitive de-

cisions on a *selected subset* of $\mathcal{D}^{test}$, while the remaining subjects will receive indecisions (i.e. be rejected for further investigation). The reject/indecision option, which is much less expensive to handle, is considered as a wasted opportunity rather than a severe error. We propose to control the false selection rate (FSR), which is the expected fraction of erroneous classifications among all definitive classifications. Selective classification with FSR control provides an effective approach to uncertainty quantification and error control. We demonstrate that with the reject/indecision option, the FSR can be controlled at a user-specified level. When the signal to noise ratio is low, the degree of ambiguity in the classification task can be, in a sense, captured by the fraction of indecisions in $\mathcal{D}^{test}$. Hence, a more powerful data-driven rule, subject to the FSR constraint, translates to a smaller fraction of indecisions, which means that less wasted efforts are needed to perform separate evaluations.

1.2 FSR control via locally adaptive shrinkage and selection

The task of controlling the FSR in high-dimensional LDA is challenging; we start by discussing several limitations of existing works.

First, the methodology and theory of many high-dimensional LDA rules (e.g. Cai and Liu, 2011; Shao et al., 2011; Mai et al., 2012; Cai and Zhang,

2019) critically depend on strong sparsity assumptions, which may not hold in practice. The sparsity assumption is counter-intuitive from the perspective of classification error control. Consider the simple case where all non-zero coordinates in $\boldsymbol{d} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ take the same value. Then a larger l_0 norm of $\boldsymbol{d}$ (i.e. non-sparse setting) virtually implies that the two classes are better separated, and hence, the control of classification risk should become easier. However, many state-of-the-art LDA rules lack theoretical justifications and often do poorly in the supposedly easier non-sparse setting (Section 5). Second, the analysis of the error rate of a classifier often requires a precise quantification of the quality of its outputs, which is in general intractable due to the complexity of contemporary LDA rules. Finally, most learning algorithms are driven by the need of improving prediction performance instead of avoiding high-consequence decision errors. It is unclear how to tailor existing algorithms to trade off a fraction of indecisions for fewer classification errors, and further, how to calibrate suitable data-driven thresholds to control the FSR at a user-specified level.

This article develops a class of FSR rules via a locally adaptive shrinkage and selection (LASS) algorithm. LASS consists of three steps: first estimating a score according to the LDA rule (1.2), second ordering all individuals based on the estimated scores, and finally choosing upper and

lower thresholds to select individuals into the two classes, with the unselected ones assigned to the *indecision* group. We prove theories to establish the asymptotic validity of LASS for FSR control. A key innovation in our method is the construction of intuitive and easy-to-analyze shrinkage factors that are capable of reducing uncertainties with much weaker assumptions on sparsity. LASS provides a principled and theoretically solid LDA rule that has comparable performance with state-of-the-art classification rules (e.g. Cai and Liu, 2011; Shao et al., 2011; Cai and Zhang, 2019) in the sparse setting and substantially better performance under the non-sparse setting. The theoretical adaptiveness of LASS to the unknown sparsity and its robust numerical performance across sparse and dense settings are attractive for practitioners – particularly in real world applications we only "bet on sparsity"; this working assumption (of sparsity) can distort the hardness of the problem, and hence, lead to wrong choices of method.

1.3 Our contributions

Our work makes several contributions. First, selective classification via FSR control provides a useful approach to risk-sensitive decision-making scenarios, where classification errors have high impacts on one's social, economic or health status. Second, we develop a novel shrinkage rule for estimating

the linear discriminant score, which is effective for reducing the uncertainties in high dimensions. It is intuitive, easy-to-analyze and enjoys strong theoretical properties. Third, we derive data-adaptive decision boundaries based on the shrunken LDA rule to select and classify the observations. Theoretical guarantees on FSR control are established under much weaker conditions compared to existing theories on sparse LDA.

1.4 Related works

We discuss several related lines of research to further explain the merits of LASS and place our contributions in context.

The idea of indecision, also referred to as reject option, has been considered in several works in the classification literature (Herbei and Wegkamp, 2006; Franc et al., 2021).The intrinsic ambiguity in classification may also be characterized by employing set-valued classifiers (Lei, 2014; Guan and Tibshirani, 2022). In terms of interpretation, an indecision means that we refrain from making a decision to avoid misclassification, whereas the set-valued output aims to guarantee that the true state would match with one of our output labels with high probability. We extend the indecision notion from the single-decision setup to a multiple-decision one, where the inflation of decision errors becomes a critical issue. The FSR framework

provides a new tool for dealing with the inflated errors when many units must be classified at the same time.

Under the high-dimensional sparse setting, Bickel and Levina (2004) demonstrated that the naive Fisher's rule can perform no better than random guess. A plethora of regularized LDA rules have been proposed to exploit the sparse structure in the data; a few representative works include the shrunken centroid method (Tibshirani et al., 2003), the LPD and AdaLDA rules (Cai and Liu, 2011; Cai and Zhang, 2019), and other penalized or thresholding methods (Shao et al., 2011; Mai et al., 2012). However, as we demonstrate in our numerical studies, these methods do not work well under the non-sparse setting. LASS, which employs an adaptive shrinkage rule with robust performance *across sparse and dense regimes*, is provably valid for error rate control.

Exemplified by the James-Stein estimator (James and Stein, 1992) and Tweedie's formula (e.g. Brown and Greenshtein, 2009; Efron, 2011; Koenker and Mizera, 2014), shrinkage is a powerful and ubiquitous idea in compound estimation. Under the independence assumption (i.e. Σ is a diagonal matrix), the implementation of the LDA rule (1.2) boils down to the compound estimation of $\boldsymbol{\mu}$ and $\boldsymbol{d}$. Efron (2009), Greenshtein and Park (2009) and Dicker and Zhao (2016) proposed empirical Bayes (EB) methods (Tweedie's

formula and g-estimation) to construct "plug-in" LDA rules. EB shrinkage can effectively reduce the uncertainty in high dimensions without the sparsity assumption. However, there are several drawbacks. First, existing EB rules ignore the correlations, which may lead to suboptimal shrinkage factors and hence inferior LDA rules. By contrast, LASS performs shrinkage in a coordinate-wise shrinkage manner, which enjoys strong numerical and theoretical properties under dependence. Second, in contrast with EB "plug-in" rules that are rather complicated to analyze, the uncertainty quantification of LASS is simple, which enables the development of data-driven rules and theory on FSR control.

1.5 Organization and notations

The article is organized as follows. Section 2 presents the problem formulation and derives the oracle rule for FSR control. The data-driven LASS is developed in Section 3 with its theoretical properties established in Section 4. The numerical results are presented in Section 5. Proofs and additional numerical results are relegated to the Supplementary Material.

Summary of notation. Denote $\boldsymbol{\mu} = \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, $\boldsymbol{d} = (d_1, ..., d_p)^\top = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_p)^\top = (1/n_1)\sum_{i=1}^{n_1} \boldsymbol{X}_i$ and $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_p)^\top = (1/n_2)\sum_{i=1}^{n_2} \boldsymbol{Y}_i$. I_p denotes the $p \times p$ identity matrix. For matrix A and $1 \le w \le \infty$, the

matrix l_w norm is defined as $||A||_w = \sup_{|\boldsymbol{x}|_w \leq 1} \|A\boldsymbol{x}\|_w$. When $v \in \mathbb{R}^p$ is a vector, $\|v\|_w$ is the vector l_w norm $\|v\|_w := (\sum_{i=1}^p v_i^w)^{1/w}$. The largest/smallest eigenvalue of A is denoted $\lambda_{max}(A)/\lambda_{min}(A)$.

2. Problem Formulation

This section first introduces a generalized discriminant rule, then defines the false selection rate, and finally outlines the roadmap.

2.1 The generalized discriminant rule

Let $\boldsymbol{W}$ be a new observation and $S := S(\boldsymbol{W})$ be a generic score, with larger (smaller) S indicating a higher chance of being in class 2 (class 1). Suppose we need to classify m new observations $\{\boldsymbol{W}_j : 1 \leq j \leq m\}$, which are drawn with equal probability from (1.1). It is natural to consider the following generalized discriminant rule $\boldsymbol{\delta} = (\delta_j : 1 \leq j \leq m)$, where

$$\delta_j = \mathbb{I}\{S(\boldsymbol{W}_j) < t_l\} + 2 \cdot \mathbb{I}\{S(\boldsymbol{W}_j) \geq t_u\}, \quad 1 \leq j \leq m. \tag{2.1}$$

In the above, t_l and t_u represent the lower and upper thresholds, respectively, with the requirement that $t_l \leq t_u$. A key difference between the two discriminant rules (2.1) and (1.2) is that (1.2) uses $t_l = t_u = 0$ whereas

the generalized rule (2.1) allows $t_l < t_u$. The interval (t_l, t_u) is called an *ambiguity region*. The true class of any observation falling within this region cannot be determined with confidence. The values of t_l and t_u will be determined according to user-specified error rates, which will be discussed in the next subsection. It follows that δ_j defined in (2.1) can take three possible values in the action space $\mathcal{A} = \{0, 1, 2\}$, with $\delta_j = k$ indicating that we classify $\boldsymbol{W}_j$ to class k, $k = 1, 2$, while $\delta_j = 0$ indicating that we make an indecision or rejection option (Herbei and Wegkamp, 2006; Sun and Wei, 2011; Lei, 2014). Denote $\{\theta_j : 1 \leq j \leq m\} \in \{1, 2\}^m$ the unknown true classes. Consider the example of medical screening, where $\theta_j = 1$ ($\theta_j = 2$) indicates that the patient is healthy (sick). Then a patient with $\delta_j = 1$ will not receive the treatment, with $\delta_j = 2$ will receive the treatment, with $\delta_j = 0$ will be followed up for further evaluation.

2.2 False selection rate

In risk-sensitive applications, we view misclassifications as severe errors, the fraction of which should be controlled at a low level. Under the selective inference framework (Benjamini, 2010), the error rate is defined to assess the quality of the selected subset, in which observations receive definitive classifications. By contrast, the indecisions are viewed as wasted opportunities

and used to describe the power notion.

For the binary setting, we may encounter two types of misclassifications: $(\theta = 1, \delta = 2)$ and $(\theta = 2, \delta = 1)$. If the two directions are symmetric, it is natural to consider the false selection rate (FSR):

$$\mathrm{FSR} = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{I}\left(\theta_j \neq \delta_j, \delta_j \neq 0\right)}{\left\{\sum_{j=1}^{m} \mathbb{I}(\delta_j \neq 0)\right\} \vee 1}\right], \tag{2.2}$$

where $x \vee y = \max(x, y)$. The FSR was recently considered in Rava et al. (2021) in a different context (fairness in machine learning), and is analogous to the false discovery rate (Benjamini and Hochberg, 1995) in one-class classification problems (outlier detection) (Bates et al., 2021; Angelopoulos et al., 2021). The FSR reduces to the misclassification rate $\frac{1}{m}\mathbb{E}\{\sum_{j=1}^{m}(\theta_j \neq \delta_j)\}$ if indecisions are not allowed (i.e. $\delta_j \neq 0$ for every $1 \leq j \leq m$).

In the asymmetric situation we may define the class-specific FSR:

$$\mathrm{FSR}^c = \mathbb{E}\left[\frac{\sum_{j=1}^{m} \mathbb{I}(\delta_j = c, \theta_j \neq c)}{\{\sum_{j=1}^{m} \mathbb{I}(\delta_j = c)\} \vee 1}\right], \quad c = 1, 2. \tag{2.3}$$

This provides a useful error notion in applications where one type of error is more sensitive than the other and it is desirable to set different tolerance levels for the two types of errors[1]. To this end, we focus on the setup

[1]The class-specific FSR^c is connected to but fundamentally different from the

allowing class-specific constraints: $\text{FSR}^c \leq \alpha_c$, $c = 1, 2$. As a special case, one may set $\alpha_1 = \alpha_2 = \alpha$. Given there is at least one classification made for each class and if the class-specific constraints $\text{FSR}^c \leq \alpha$ are fulfilled for both $c = 1$ and $c = 2$, then it can be shown that the global constraint $\text{FSR} \leq \alpha$ defined in (2.2) is also fulfilled; a proof this statement is provided in Section S8 of the Supplementary material.

The selective classification framework enables one to control the FSR at a user-specified level, which can be impossible without the indecision option. However, the price we need to pay is the wasted opportunity for performing separate evaluations on the indecisions. The user-specified error bounds α_c reflect our tolerance levels of the associated risks. To simultaneously quantify the degree to which the decisions can be trusted and minimize the wasted efforts, we consider a constrained optimization problem. Let $\text{ECC} = \mathbb{E}\left\{\sum_{j=1}^{m} \mathbb{I}(\theta_j = \delta_j)\right\}$ denote the expected number of correct classifications. The goal is to

$$\text{maximize the ECC subject to } \text{FSR}^c \leq \alpha_c,\ c = 1, 2. \tag{2.4}$$

Neyman-Pearson classification framework (Scott and Nowak, 2005; Rigollet and Tong, 2011) for asymmetric error control. The class-specific FSR^c, as a concept under the selective inference framework, is analogous to the FDR in multiple testing, whereas Neyman-Pearson classification operates under the classical Type I/II error paradigm in single hypothesis testing. Moreover, the two lines of research focus on substantially different issues.

The constrained optimization formulation (2.4) has not been considered under the classification setup, although the idea is related to the multiple testing formulation in Sun and Cai (2007). There are several crucial differences between the two formulations. First, Sun and Cai (2007) proposed to minimize the false nondiscovery rate (FNR) subject to the constraint on FDR. Under this multiple testing setup, we only have one alternative state and the decision takes values in $\{0, 1\}$. In contrast, the selective classification formulation has two alternative states and the decision takes values in $\{0, 1, 2\}$. This requires the development of new optimality theory. Second, in multiple testing each data point corresponds to the value of a one-dimensional summary statistic (e.g. p-value or z-value). In contrast, the observation W in our setup is a high dimensional vector, which makes the theoretical analysis significantly more challenging.

2.3 Oracle rules for FSR control

In this subsection we derive a class of oracle FSR rules. To motivate our methodology, consider an asymptotically equivalent error rate (Supplementary material Section S3), the marginal FSR

$$\mathrm{mFSR}^c = \frac{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c, \theta_j \neq c)\right\}}{\mathbb{E}\left\{\sum_{j=1}^m \mathbb{I}(\delta_j = c)\right\}}. \tag{2.5}$$

We aim to develop a selective classification rule that solves the following constrained optimization problem: maximize the ECC subject to $\text{mFSR}^c \leq \alpha_c$, c=1, 2. Next we prove an intuitive result that the optimal mFSR rule is a thresholding rule based on the optimal LDA function $S_j^\pi \equiv (\boldsymbol{W}_j - \boldsymbol{\mu})^\top \Sigma^{-1} \boldsymbol{d}$ (or its monotone transformations).

Consider a generalized discriminant rule $\boldsymbol{\delta}(t_1, t_2) = (\delta_j : 1 \leq j \leq m)$ of the form (2.1): $\delta_j = \mathbb{I}(1 - T^j < t_1) + 2\mathbb{I}(T^j < t_2)$, $1 \leq j \leq m$, where $T^j := T(\boldsymbol{W}_j) = \mathbb{P}(\theta_j = 1 | \boldsymbol{W}_j) = \dfrac{\exp(S_j^\pi)}{\exp(S_j^\pi) + 1}$, and $t_1, t_2 \in (0, 1)$ are the lower and upper thresholds satisfying $t_1 < 1 - t_2$. As T^j is a monotone transformation of S_j^π, generalized LDA rules based on T^j and S_j^π, with suitably adjusted thresholds, are equivalent. We use T^j instead of S_j^π to facilitate the development of a step-wise algorithm, which is described at the end of this section.

Let $Q^c(t_c)$ be the mFSR^c of $\boldsymbol{\delta}(t_1, t_2)$, $c = 1, 2$. Define the oracle thresholds $t_{OR}^c = \sup\{t : Q^c(t) \leq \alpha_c\}, c = 1, 2$. To avoid assigning an individual to multiple classes, we assume that α_1 and α_2 have been chosen such that both t_{OR}^1 and t_{OR}^2 are less than or equal to 0.5[2]. Define the oracle mFSR

[2]This is an assumption to facilitate theoretical development. If there is overlapping selection in practice, we can simply classify the individual to the class with larger class probability $P(\theta_j = c | \boldsymbol{W}_j), c = 1, 2$.

procedure $\boldsymbol{\delta}_{OR} = (\delta^j_{OR} : 1 \le j \le m)$, where

$$\delta^j_{OR} = \mathbb{I}\left(1 - T^j < t^1_{OR}\right) + 2 \cdot \mathbb{I}\left(T^j < t^2_{OR}\right). \tag{2.6}$$

The next theorem shows that $\boldsymbol{\delta}_{OR}$ is optimal.

Theorem 1. *Let $\mathcal{D}_{\alpha_1,\alpha_2}$ be the collection of all classification rules such that for any $\boldsymbol{\delta} \in \mathcal{D}_{\alpha_1,\alpha_2}$, $mFSR^1_{\boldsymbol{\delta}} \le \alpha_1$ and $mFSR^2_{\boldsymbol{\delta}} \le \alpha_2$. Then $ECC_{\boldsymbol{\delta}} \le ECC_{\boldsymbol{\delta}_{OR}}$ for any $\boldsymbol{\delta} \in \mathcal{D}_{\alpha_1,\alpha_2}$.*

The thresholds t^1_{OR} and t^2_{OR} in the oracle rule (2.6) can be approximately calculated using the following step-wise algorithm. Denote $T^{(i)}$ the ith ordered statistic of $\{T^1, ..., T^m\}$. Let

$$k_1 = \min\left\{1 \le j \le m : \frac{1}{j+1}\sum_{i=0}^{j}\left\{1 - T^{(m-i)}\right\} \le \alpha_1\right\}, k_2 = \max\left\{1 \le j \le m : \frac{1}{j}\sum_{i=1}^{j} T^{(i)} \le \alpha_2\right\}. \tag{2.7}$$

As indicated by the theory in Section 4, t^1_{OR} and t^2_{OR} can be consistently estimated by $\hat{t}^1_{OR} = \min\left(T^{(k_2)}, 0.5\right)$ and $\hat{t}^2_{OR} = \min\left(1 - T^{(m-k_1)}, 0.5\right)$ under mild conditions. Here 0.5 is imposed to avoid overlapping selections. To see why the stepwise algorithm (2.7) makes sense, note that the moving average $\frac{1}{r}\sum_{j=1}^{r} T^{(j)}$ provides an estimate of mFSR^2 when r observations with the smallest T^j are selected to class 2 (cf. Sun and Cai (2007)). Hence, it

follows from (2.7) that $\hat{t}^2_{OR}$ corresponds to the largest threshold such that the estimated FSR^2 is below α_2. The explanation for $\hat{t}^1_{OR}$ is similar.

Denote $\boldsymbol{\delta}^*_{OR} = \big\{\mathbb{I}(1 - T^j < \hat{t}^1_{OR}) + 2 \cdot \mathbb{I}(T^j < \hat{t}^2_{OR}) : 1 \le j \le m\big\}$. The next theorem shows that the stepwise algorithm (2.7) is valid.

Theorem 2. *Consider the oracle setting where T^j are known, $j = 1, \cdots, m$. Then we have $FSR^k(\boldsymbol{\delta}^*_{OR}) \le \alpha_k$, $mFSR^k(\boldsymbol{\delta}^*_{OR}) \le \alpha_k$, for $k = 1, 2$.*

Remark 1. $\boldsymbol{\delta}^*_{OR}$ is asymptotically optimal in the sense that $ECC_{\boldsymbol{\delta}^*_{OR}}/ECC_{\boldsymbol{\delta}_{OR}} \to 1$ as $m \to \infty$. This fact can be proved using similar arguments as those presented in the proof of Theorem 4.

2.4 Issues and roadmap

The FSR control in selective classification, which is closely related to the false discovery rate (FDR, Benjamini and Hochberg, 1995) control in multiple testing, presents unique challenges in high-dimensional inference. In multiple testing the null distribution of the p-values is assumed to be known precisely; hence FDR rules, such as the Benjamini-Hochberg's algorithm, can be derived to determine a proper p-value threshold that upper bounds the FDR. However, in classification the scores (S^π_j or T^j) must be estimated from the training data with noise. For state-of-the-art LDA rules in the high-dimensional setting (Cai and Liu, 2011; Shao et al., 2011; Mai

et al., 2012; Dicker and Zhao, 2016; Cai and Zhang, 2019), the distributions of the estimated scores (and hence p-values) are in general unknown, rendering the uncertainty quantification and analysis of error rate intractable.

We take a different approach and develop a data-driven FSR rule in two steps. In the first step, we provide an efficient and robust score $\hat{S}_j$, which employs a new shrinkage rule that works well across sparse and dense regimes. In the second step, we develop a step-wise algorithm based on $\hat{S}_j$. Owing to the easy-to-analyze shrunken mechanism, we show that the uncertainties in the estimated score and its stochastic contribution to the errors by running the algorithm can be precisely quantified, establishing the theory on FSR control.

3. The Data-Driven LASS Procedure

The key step in estimating the score S_j^π is to develop a good estimate for $\boldsymbol{d} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. In high-dimensional settings, most regularized LDA rules bet on the sparsity of $\boldsymbol{d}$ (e.g. Tibshirani et al., 2003; Shao et al., 2011) to reduce the high variability in the sample estimates. However, the sparsity requirement, which may not hold in practice and often only serves as a working assumption, is counter-intuitive in the sense that the two classes are better separated as $\boldsymbol{d}$ has more nonzero elements. By contrast, Efron

(2009); Greenshtein and Park (2009); Dicker and Zhao (2016) proposed LDA rules based on Tweedie-type shrinkage estimators of $\boldsymbol{d}$, sidestepping the sparsity assumption. Existing non-sparse LDA rules have two limitations. First, Tweedie-type estimates are intractable to analyze, making it difficult to assess the uncertainties in classification. Moreover, Tweedie's formula requires that the elements in $\boldsymbol{d}$ must be independent, which leads to efficiency loss when the dependence structure is highly informative (Cai and Liu, 2011; Shao et al., 2011). We propose an easy-to-analyze shrinkage estimator that overcomes the above limitations. The methodology and illustrative examples are provided in Sections 3.1 and Section S4 of the supplement. The data-driven LASS procedure is presented in Section 3.2.

3.1 Methodology

Let $\bar{X}_k$ and $\bar{Y}_k$ be the kth coordinate of $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$, respectively. We consider a class of shrinkage estimators

$$\hat{\boldsymbol{d}} = (\hat{d}_k : 1 \le k \le p) = \left\{(\bar{X}_k - \bar{Y}_k)q_k : 1 \le k \le p\right\}, \tag{3.1}$$

where $q_k \in (0, 1)$ is a coordinate-wise shrinkage factor. To effectively reduce the uncertainties and quantify the associated misclassification risks, q_k

needs to be designed carefully such that it converges to 1/0 at appropriate rates according to the strength of the signal. The proposed method chooses the following class of q_k:

$$q_k := \frac{g_{1k}\left(|\bar{X}_k - \bar{Y}_k|\right)}{g_0\left(|\bar{X}_k - \bar{Y}_k|\right) + g_{1k}\left(|\bar{X}_k - \bar{Y}_k|\right)}, \tag{3.2}$$

where g_0 and g_{1k} are respectively the density functions of $\mathcal{N}\left(0, \frac{n_1+n_2}{n_1 n_2}\right)$ and $\mathcal{N}\left(\left\{(2+b)\sqrt{\hat{\sigma}_{kk}} + \sqrt{(2+b)^2\hat{\sigma}_{kk}+4}\right\}\sqrt{\frac{(n_1+n_2)}{2n_1n_2}\log p}, \frac{n_1+n_2}{n_1n_2}\right)$, $b > 0$ is a small constant, and $\hat{\sigma}_{kk}$ is the pooled sample variance of $\{X_{ik} : i = 1\ldots, n_1\}$ and $\{Y_{ik} : i = 1\ldots, n_2\}$. The constant $b > 0$ is included in the definition only for theoretical considerations. In practice we can choose $b \approx 0$ or simply set $b = 0$. In all our simulations and data analyses, we report the results with $b = 0.1$.

The behavior of q_k is qualitatively different depending on the strength of d_k. The following proposition shows an intuitively appealing demarcation phenomenon of q_k, implying that the multiplicative shrinkage rule (3.1) produces effects similar to that of hard thresholding rules: strong signals are kept and moderate/weak signals are suppressed.

Proposition 1. *Consider q_k defined in (3.2). Let $a_k = \left\{(2+b)\sqrt{\sigma_{kk}} + \sqrt{(2+b)^2\sigma_{kk}+4}\right\}$ and ϵ an arbitrarily small constant. Define the following three groups:*

$$\mathcal{G}_1 = \left\{1 \le k \le p : |d_k| > (a_k/2 + \epsilon)\sqrt{\tfrac{(n_1+n_2)}{2n_1n_2}\log p}\right\} \textit{ (strong signals);}$$

$$\mathcal{G}_2 = \left\{1 \le k \le p : |d_k| = o(\sqrt{\tfrac{(n_1+n_2)}{2n_1n_2}\log p})\right\} \textit{ (weak signals);}$$

$$\mathcal{G}_3 = \left\{1 \le k \le p : |d_k| < (a_k/2 - \epsilon)\sqrt{\tfrac{(n_1+n_2)}{2n_1n_2}\log p} \textit{ and } k \notin \mathcal{G}_2\right\} \textit{ (moderate signals).}$$

Then there exists $\gamma > 0$ *independent of* p, n_1 *and* n_2 *such that:*

(a). $1 - \mathbb{E}(q_k \mid k \in \mathcal{G}_1) = O(p^{-\gamma})$;

(b). $\mathbb{E}(q_k \mid k \in \mathcal{G}_3) = O(p^{-\gamma})$;

(c). $\mathbb{E}(q_k \mid k \in \mathcal{G}_2) = O(p^{-(1+\gamma)})$.

We mention some merits of the proposed shrinkage rule. First, under the dense regime, the multiplicative factor q_k can produce significantly less noisy estimates than original observations while being capable of retaining more nonzero coordinates than thresholding rules. This leads to shrinkage rules with robust and superior performance at different sparsity levels. Second, unlike LDA rules based on Tweedie's formula (Efron, 2009; Dicker and Zhao, 2016), the coordinate-wise shrinkage scheme in (3.2) does not require the independence between d_k. Finally, the multiplicative rule is easy to analyze and leads to provably valid rules for FSR control.

3.2 The Data-Driven LASS Procedure

We propose to estimate $S_j^\pi = (\boldsymbol{W}_j - \boldsymbol{\mu})^\top \Sigma^{-1}\boldsymbol{d}$ by $\hat{S}_j = \left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}} + \bar{\boldsymbol{Y}}}{2}\right)^\top \hat{\Sigma}^{-1}\hat{\boldsymbol{d}}$. First, $\hat{\Sigma}^{-1}$ is the estimated precision matrix and $\hat{\boldsymbol{d}}$ is the proposed shrinkage estimate (3.1). The estimation of the precision matrix has been intensively studied in the literature; see Liu and Luo (2015), Cai et al. (2016), Loh and Tan (2018), Wang et al. (2013), Sun and Zhang (2013), Yuan (2010) for related works. In our numerical studies we use the ACLIME estimator proposed in Cai et al. (2011). Next, $\hat{\boldsymbol{d}} = ((\bar{X}_1 - \bar{Y}_1)q_1, \ldots, (\bar{X}_p - \bar{Y}_p)q_p)$, where q_k is as defined in (3.2). Let $\hat{T}^j := \dfrac{\exp(\hat{S}_j)}{1+\exp(\hat{S}_j)}$. Denote $\{\hat{T}^{(j)} : 1 \le j \le m\}$ the ordered statistics. Define

$$k_1 = \min\left\{1 \le j \le m : \frac{1}{j+1}\sum_{i=0}^{j}(1 - \hat{T}^{(m-i)}) \le \alpha_1\right\}, \quad k_2 = \max\left\{1 \le j \le m : \frac{1}{j}\sum_{i=1}^{j}\hat{T}^{(i)} \le \alpha_2\right\}. \tag{3.3}$$

The data-driven LASS procedure is given by $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots \hat{\delta}_m)$, where

$$\hat{\delta}_j = \mathbb{I}\left\{1 - \hat{T}^j < \min\left(1 - \hat{T}^{(m-k_1)}, 0.5\right)\right\} + 2 \cdot \mathbb{I}\left\{\hat{T}^j \le \min\left(\hat{T}^{(k_2)}, 0.5\right)\right\}. \tag{3.4}$$

Remark 2. If we choose $\alpha_1 = \alpha_2 = 0.5$, then indecisions are not allowed by (3.4). That is, LASS becomes a classical rule that makes definitive classifications on all individuals. We shall see that LASS is still superior

in both theory and numerical performance compared to existing methods under this classical setup (Corollary 1 in Section 4 and Section 5.1).

4. Theoretical Properties of LASS

This section studies theoretical properties for the data-driven LASS procedure. We focus on the regime of $\frac{n_1+n_2}{n_1 n_2}\log p \to 0$, which requires that the dimension does not grow too fast relative to the sample size. We consider issues on FSR control and optimality in turn. A discussion of connections to existing works is given in Section S7 of the Supplement.

We first state and explain a few conditions needed in our theoretical analysis.

(A1) The covariance matrix $\Sigma = (\sigma_{kl})_{1\le k,l\le p}$ satisfies $0 < \epsilon_0 \le \sigma_{kk} \le 1/\epsilon_0$ for all $1 \le k \le p$, where ϵ_0 is a fixed positive constant.

(A1) is a standard condition in matrix analysis, and is satisfied when the covariance matrix is *well-conditioned*, as assumed in, for example, Bickel and Levina (2008).

(A2) The estimated precision matrix $\hat{\Sigma}^{-1}$ satisfies $\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_2^2 = o(1)$.

Consistent estimation of the precision matrix Σ^{-1} has been studied intensively. Effective estimators and sufficient conditions for consistent estimation have been discussed in a large body of works including Bickel

and Levina (2008), Yuan (2010), Liu and Luo (2015),Cai et al. (2016) and Avella-Medina et al. (2018), among others. A more detailed discussion of (A2) is provided in Supplementary material S9.

(A3) $|\mathcal{G}_1| \geq 1$ and $|\mathcal{G}_3| = O(\frac{n_1 n_2}{n_1+n_2})$, where $\mathcal{G}_1$ and $\mathcal{G}_3$ are defined in Proposition 1 and correspond to collections of strong and moderate coordinates of $\boldsymbol{d}$, respectively.

In (A3), $|\mathcal{G}_1| \geq 1$ provides a sufficient condition under which LASS makes at least one definitive classification with high probability. As opposed to existing works that require the sparsity of both $\mathcal{G}_1$ and $\mathcal{G}_3$, we do not impose an upper bound on $|\mathcal{G}_1|$. Our condition seems to be more sensible because having more strong signals ($d_k \in \mathcal{G}_1$) is helpful for distinguishing the two classes, and what really hurts the performance of LDA rules is an overwhelming number of nonzero elements of moderate strength ($d_k \in \mathcal{G}_3$). The condition $|\mathcal{G}_3| = O(\frac{n_1 n_2}{n_1+n_2})$ corresponds to a weaker notion of sparsity in the sense that the sparsity or approximate sparsity conditions in existing works (e.g. Cai and Liu, 2011) will be violated if $|\mathcal{G}_3| \gg \frac{n_1 n_2}{n_1+n_2}$. We stress that the conventional sparsity notion assumes that there are relatively few signals, while we require that there are relatively few signals *of moderate strength* will fall with the narrow range defined by $\mathcal{G}_3$, which eliminates the need of the counter-intuitive sparsity condition on $\mathcal{G}_1$ as used in existing

works. The superiority of LASS under the dense signal setting is illustrated in our numerical results (Section 5). The next theorem establishes the asymptotic validity of LASS for FSR control.

Theorem 3. *Let $\hat{\boldsymbol{\delta}}$ be the data-driven LASS procedure defined in* (3.4). *Under conditions (A1)-(A3), we have $mFSR^c_{\hat{\boldsymbol{\delta}}} \leq \alpha_c + o(1)$ and $FSR^c_{\hat{\boldsymbol{\delta}}} \leq \alpha_c + o(1)$ for $c = 1, 2$.*

Our conditions on error rate control are substantially different in nature compared to those required by state-of-the-art LDA rules. First, the sparsity of Σ^{-1} is not a necessary condition for our theory on FSR control[3]. Second, our theory needs neither sparsity nor consistent estimation of $\boldsymbol{d}$. In particular if $\sum_{i\in\mathcal{G}_1} d_i^2 \to \infty$, as long as $\sum_{i\in\mathcal{G}_1} \hat{d}_i^2$ also goes to ∞, we can still perfectly separate the two classes under condition (A3). Finally, in contrast with existing works, our theory has no restrictions on the norm of $\boldsymbol{d}$ or $\Sigma^{-1}\boldsymbol{d}$. We stress that estimation and classification are fundamentally different tasks: the assumptions on the norm are natural for estimation problems, but counter-intuitive for classification problems – larger norms make the classification task easier and lead to lower error rates.

Next we investigate the asymptotic optimality of LASS. As the moder-

[3]Note that even when the sparsity of Σ^{-1} is needed for consistent estimation, the sparsity conditions on Σ^{-1} and $\boldsymbol{d}$ usually correspond to fundamentally different notions in scientific studies. Our theory seems to be more sensible as it eliminates the need on the sparsity of $\boldsymbol{d}$.

ate and weak signals have been shrunk to values close to zero, LASS would be asymptotically optimal if weak and moderate signals have negligible effects or strong signals have dominating effects. We formalize this intuition in the next theorem.

Theorem 4. *In addition to conditions (A1)-(A3), if either of the following two condition holds*

(A4) $\sum_{k\notin\mathcal{G}_1} d_k^2 = o(1)$.

(A5) $\sum_{k\in\mathcal{G}_1} d_k^2 \to \infty$.

We have $mFSR^c_{\hat{\boldsymbol{\delta}}} \le \alpha_c + o(1)$, $FSR^c_{\hat{\boldsymbol{\delta}}} \le \alpha_c + o(1)$, *and* $\frac{ECC_{\hat{\boldsymbol{\delta}}}}{ECC_{\boldsymbol{\delta}_{OR}}} = 1 + o(1)$.

If we let $\alpha_1 = \alpha_2 = 0.5$, then the FSR control setup reduces to the classical setup where indecisions are not allowed. Let $\boldsymbol{\delta}$ be a classification rule that only takes on value 1 or 2, define $L(\boldsymbol{\delta}) = \mathbb{P}\{\theta_j \neq \delta_j | (\boldsymbol{X}_i, 1 \le i \le n_1), (\boldsymbol{Y}_i, 1 \le i \le n_2)\}$, $R(\boldsymbol{\delta}) = \mathbb{E}\{L(\boldsymbol{\delta})\}$. A direct consequence of Theorem 4 is given below.

Corollary 1. *(Risk consistency). Suppose we choose* $\alpha_1 = \alpha_2 = 0.5$. *Then under conditions (A1)-(A4), we have* $R(\hat{\boldsymbol{\delta}}) - R(\boldsymbol{\delta}^F) \to 0$, *where* $\boldsymbol{\delta}^F$ *is the oracle Fisher's rule.*

5. Numerical Experiments

This section illustrates the numerical performance of LASS using both simulated and real data. The simulation considers two setups: the conventional

setup that does not allow indecisions (Section 5.1), and the selective classification setup that aims to control the FSR (Section 5.2). Two real data sets are discussed in Section 5.3 and Section S6 of the supplement. In all analyses, LASS is implemented using $b = 0.1$ in (3.2) and the ACLIME method (Cai et al., 2016) is adopted for estimating Σ^{-1}. For the simulated data, we take $n_1 = n_2 = n$.

5.1 Simulation: the conventional setup

We start with the classical setting where no indecisions are allowed. We compare LASS with the following methods: (a) Fisher's rule that uses true $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ^{-1} (denoted "Oracle"). This method serves as the optimal benchmark for all classification rules; (b) The LPD rule proposed by Cai and Liu (2011) (denoted "LPD"), which is implemented using the code provided on the authors' website; (c) The AdaLDA rule proposed by Cai and Zhang (2019) (denoted "AdaLDA"), which is implemented using the code provided on the authors' website; (d) Fisher's rule that uses sample estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ^{-1} (denoted "Naive"). Specifically, S_j^{π} is estimated as $\left(\boldsymbol{W}_j - \frac{\bar{\boldsymbol{X}}+\bar{\boldsymbol{Y}}}{2}\right)^{\top} \hat{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$, where $\hat{\Sigma}^{-1}$ is the Penrose inverse of the sample covariance matrix; (e) The L_1 logistic regression method (denoted "Lasso"). We have followed the approach suggested by Lei (2014), where

the tuning parameter is chosen by cross-validation; (f) The empirical Bayes method proposed in Efron (2009) (denoted "Ebay"). We use the R-package `Ebay` to estimate $\boldsymbol{d}$. Σ^{-1} is estimated by a diagonal matrix where diagonals are the inverses of sample variances. We present numerical results in the next two subsections to show that LASS (a) is comparable to state-of-the-art methods in the sparse case, and (b) substantially outperforms competing methods in the dense case.

5.1.1 Sparse setting

Let $\boldsymbol{\mu}_1 = (0, ..., 0)^\top \in \mathbb{R}^p$, and $\boldsymbol{\mu}_2$ be a vector with the first 10 entries being 0.5, the next 10 $0.1(\log p/n)^{1/2}$, and the rest 0. Consider the following three correlation structures that are widely considered in the literature (Cai and Liu, 2011; Cai and Zhang, 2019; Avella-Medina et al., 2018).

Model 1: Band graph. Let $\Sigma^{-1} = \Omega = (\omega_{ij})_{p\times p}$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.35$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.175$, and $\omega_{ij} = 0$ if $|i-j| > 2$.

Model 2: AR(1) structure. Let $\Sigma^{-1} = \Omega = (\omega_{ij})_{p\times p}$, where $\omega_{ij} = 0.3^{|i-j|}$.

Model 3: Block structure. Let $\Sigma^{-1} = \Omega = (\mathbf{B} + \delta I_p)/(1+\delta)$, where $b_{ij} = b_{ji} = 0.05 \cdot \text{Bernoulli}(0.1)$ for $1 \le i \le p/2, i < j \le p$, $b_{ij} = b_{ji} = 0.05$ for $p/2+1 \le i < j \le p$, $b_{ii} = 1$ for $1 \le i \le p$, and

$$\delta = \max\{-\lambda_{min}(\mathbf{B}), 0\} + 0.1.$$

The size of the training set is $n = 400$, with p varying from 500 to 1000. The mis-classification rate is computed based on $m = 2000$ test points generated from $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma)$ or $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$ with equal probability. We repeat the experiment for 100 times, and report the misclassification rates (in percentage) in Table 1.

p	**Oracle**	**Naive**	**LASS**	**LPD**	**AdaLDA**	**LASSO**	**Ebay**
			Model 1				
500	13.74	29.39	**14.78**	15.78	14.79	15.75	15.93
600	13.64	33.12	14.72	15.52	**14.55**	15.78	15.66
700	13.81	38.51	15.02	15.80	**14.85**	16.10	15.99
800	13.64	47.72	14.87	15.89	**14.81**	16.12	15.62
900	13.70	41.66	**14.44**	17.41	14.75	16.31	15.79
1000	13.70	41.16	**14.59**	18.06	14.73	16.27	15.92
			Model 2				
500	14.82	30.68	15.98	16.61	**15.77**	16.72	16.03
600	14.81	34.45	16.15	16.68	**15.77**	16.84	16.35
700	14.79	39.08	16.21	16.67	**15.86**	16.92	16.36
800	14.71	47.83	16.13	16.77	**15.91**	16.98	16.02
900	14.87	41.79	16.19	18.14	**15.86**	17.08	16.37
1000	14.92	41.25	16.38	18.72	**15.92**	17.11	16.51
			Model 3				
500	21.16	36.20	**22.93**	23.83	23.71	24.21	23.31
600	20.87	39.44	**23.14**	24.23	24.04	24.69	23.48
700	21.00	42.81	**23.52**	24.69	24.49	25.03	23.88
800	20.99	48.59	**23.82**	25.00	24.87	25.28	24.08
900	21.02	44.22	**24.29**	25.78	25.37	26.01	24.48
1000	21.05	43.02	**24.72**	27.04	26.11	26.41	25.08

Table 1: Comparison of average misclassification rate in percentage. The smallest error rate (next to that by the oracle) in each setting has been boldfaced.

We can see that the Naive method can be substantially improved by LPD, AdaLDA and LASSO, all of which make strong assumptions on the sparsity structure of the data generating model. Although no method dominates, LASS and AdaLDA seem to perform the best among all methods

being considered. LASS is comparable to AdaLDA in terms of the overall effectiveness across the three settings. This is impressive since Cai and Zhang (2019) showed AdaLDA is minimax optimal in sparse LDA. We shall see in the next simulation that in the non-sparse setting, LASS substantially outperforms AdaLDA. Similar to LASS, the Ebay method adopts the shrinkage idea and does not make strong assumptions on the sparsity structure. We can see that Ebay performs reasonably well. However, Ebay relies on the independence assumption, has no theoretical guarantee on the convergence of error rate, and is less effective than LASS in all settings.

5.1.2 Dense setting

Consider the three models in the previous section. The choices of $\boldsymbol{\mu}_1$ and Σ are the same, while $\boldsymbol{\mu}_2$ contains more nonzero entries: the first $(p/4)$ entries are 0.4 and the rest are 0. The misclassification rates (in percentage) are summarized in Table 2. As expected, methods that rely heavily on the sparsity assumption of $\boldsymbol{d}$, such as LPD and AdaLDA, do not perform well. We mention a few important patterns in the results: (a) The performance of LPD and AdaLDA deteriorates as p increases. This is undesirable considering that the classification problem seems to have become easier, as is manifested in the improved performance of the oracle rule. In many set-

tings LPD and AdaLDA can be worse than Naive; (b) Ebay does relatively well when p is small but its performance also deteriorates as p increases. The misclassification rates can be much higher than the oracle benchmark; (c) LASS and Lasso substantially outperform competing methods in most settings. The performance of both improves as p increases, exhibiting the same desirable trend as that of the oracle rule. LASS dominates Lasso, and the gap in error rate is substantial in several settings.

5.2 Simulation: FSR control

We now turn to the selective inference setup where the goal is to control the FSR. For Naive, Lasso and Ebay, we first form an estimate for the discriminant, denoted as $\hat{S}_j$, then use $\hat{T}^j = \frac{\exp(\hat{S}_j)}{1+\exp(\hat{S}_j)}$ in (3.3) and (3.4), which serve as the *base algorithm* for FSR control. LPD and AdaLDA are omitted as they only produce the signs of the discriminants and it is unclear how to adjust the algorithms for FSR control.

Next we present results pertaining to FSR control in the sparse settings as considered in Section 5.1.1, but omit the results for the dense settings in Sections 5.1.2. The reason is that when the classification task becomes easy (as indicated by Table 2): the misclassification rate is so low that the FSR framework is no longer needed.

p	**Oracle**	**Naive**	**LASS**	**LPD**	**AdaLDA**	**Lasso**	**Ebay**
			Model 1				
500	0.07	2.95	0.20	5.19	2.09	0.52	**0.12**
600	0.02	4.48	**0.09**	5.00	3.18	0.31	0.42
700	0.01	9.79	**0.04**	11.59	4.80	0.20	0.78
800	0.00	39.79	**0.03**	15.24	6.38	0.16	1.09
900	0.00	12.37	**0.01**	15.45	8.59	0.15	1.30
1000	0.00	8.42	**0.01**	11.29	10.17	0.12	1.54
			Model 2				
500	0.08	3.02	0.23	5.48	2.05	0.54	**0.16**
600	0.03	4.56	**0.10**	4.93	2.81	0.32	0.46
700	0.01	10.08	**0.05**	11.80	4.38	0.23	0.79
800	0.00	39.94	**0.02**	12.69	6.20	0.15	1.15
900	0.00	12.67	**0.01**	15.27	8.14	0.13	1.32
1000	0.00	8.36	**0.01**	11.31	9.42	0.10	1.56
			Model 3				
500	0.26	5.06	**1.32**	6.52	3.37	2.04	1.50
600	0.08	6.35	**0.86**	6.22	3.78	1.37	1.69
700	0.02	11.72	**0.62**	5.34	4.64	0.99	1.92
800	0.01	39.89	**0.44**	5.32	6.36	0.70	2.05
900	0.00	14.08	**0.38**	16.00	8.25	0.61	2.18
1000	0.00	10.05	**0.36**	18.96	10.47	0.51	2.22

Table 2: Comparison of average misclassification rate in percentage. The smallest error rate (next to that by the oracle) in each setting has been boldfaced.

Consider the models in Section 5.1.1. We fix $n = 400$ and vary p from 200 to 800. The target FSR^1 and FSR^2 levels are both set 0.1. The experiment is repeated for 100 times, and the average FSRs (displayed in the first two columns) and power (defined as ECC/m; displayed in the last column) are reported in Figure 1. We mention the following patterns: (a) Both Naive and Ebay fail to control the FSR. The Naive method becomes worse as p increases. This corroborates the analysis in Bickel and Levina (2004), which shows that LDA rules based on sample estimates suffer from high dimensionality; (b) Both Lasso and LASS control the FSR at the

nominal level, showing that our proposed data-driven algorithm (3.3) is effective for FSR control when equipped with reasonably good estimates of the scores; (c) LASS controls the FSR at the nominal level accurately across all settings. Lasso is conservative and has lower power.

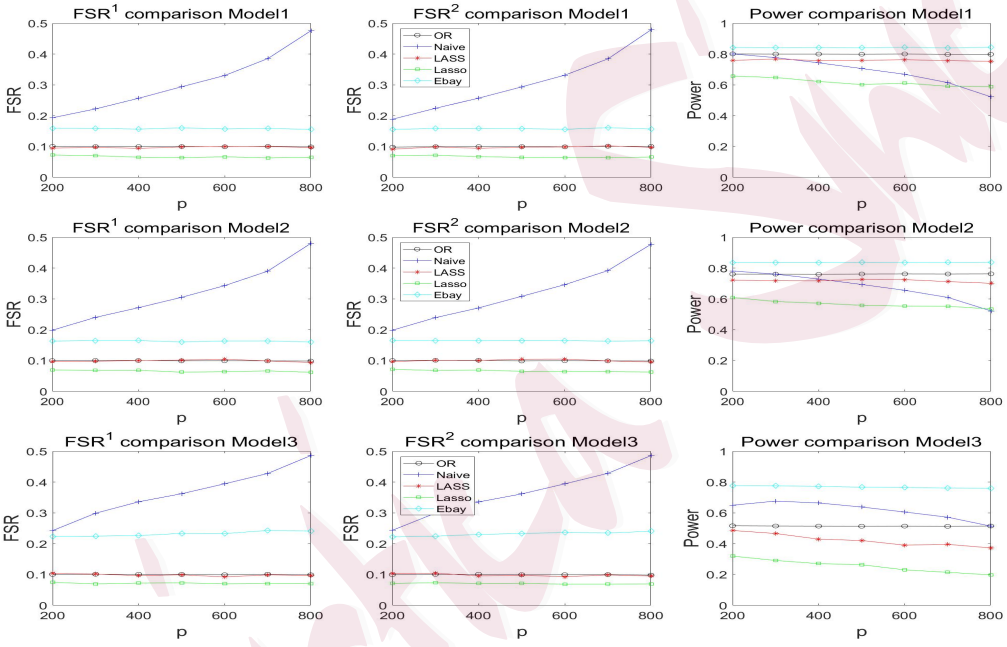

Figure 1: Comparison of FSR and Power. Naive and Ebay fail to control the FSR. LASS controls FSR at the nominal level with the highest Power.

5.3 p53 Mutants Data

Finally we perform classification on p53 mutants data (Danziger et al., 2009), which consist of 16,772 tissue samples, and for each sample a $p = 5,407$-dimensional vector is measured. Among the 16,772 samples, 143 are

Table 3: Misclassification rates of different methods

	LASS	Naive	LPD	AdaLDA	Lasso	Ebay
Misclassification	30.73%	40.24%	32.34%	**30.28**%	31.23%	31.57%

determined as "active" and the rest are determined as "inactive". We randomly select 100 active samples and 100 inactive samples as our training data, then take the rest 43 active samples and 50 random inactive samples as our testing set. To make the classification problem more difficult, an independent $\mathcal{N}(0, 40)$ noise variable is added to each gene in both the training and testing sets.

We follow the previous pre-processing steps: (a) the training data are used to estimate the sample variances, (b) genes with variances greater than 10^2 or smaller than 10^{-2} are dropped, and (c) only top 100 genes with the largest t-statistics are used. The experiment is repeated 50 times with results summarized in Tables 3 and 4.

Table 3 contains the results under the conventional setup. LASS performs as well as the LPD and AdaLDA rules. But all methods have high misclassification rates. Hence we consider FSR control. We set the target FSR levels for both classes to be 0.1. In Table 4 we compare the FSR and power of different methods. We can see that LASS effectively controls the FSR while Naive, Ebay and Lasso fail to do so.

Table 4: FSR and power comparison.

	LASS	Naive	Lasso	Ebay
FSR^1	**10.38**%	41.90%	17.81%	32.01%
FSR^2	**11.51**%	38.75%	16.00%	30.88%
Power	20.60%	59.38%	19.25%	68.41%

6. Discussion and Future Extensions Section

This article develops a selective classification framework for high-dimensional LDA problems. The proposed LASS procedure, which provides an indecision option for any observations that cannot be classified with confidence, controls the FSR at user-specified levels. LASS is easy-to-analyze and has robust performance across sparse and dense regimes.

We conclude the article with several directions for future research. First, it is of interest to relax Condition (A2). Intuitively, if the signal to noise ratio $\sqrt{\boldsymbol{d}^\top \Sigma^{-1} \boldsymbol{d}}$ is high then some errors in estimating $\hat{\Sigma}^{-1}$ and $\hat{\boldsymbol{d}}$ can be tolerated without much degrading the accuracy of LASS-type classifiers. Second, it is desirable to design model-free methods that guarantee FSR control without requiring consistent estimation of class probabilities. Promising ideas include the construction of knockoffs or mirror sequences as done in Barber and Candès (2015); Leung and Sun (2021), or the use of conformal techniques as done in Bates et al. (2021); Guan and Tibshirani (2022). Finally, this work has focused on the situation where both the training and test data come from two classes. It would be of interest to generalize the

framework to handle the multi-class setup, and to develop new inference procedures for detecting novel classes (outliers) in the test data.

Supplementary Materials

The supplement contains the proofs of main theorems, propositions and corollaries, proofs of technical lemmas, an argument establishing the asymptotic equivalence of FSR and mFSR, additional numerical results and illustrations, an example showing the advantage LASS has over LPD, a proof showing class-specific FSR control implies global FSR control and a discussion on condition (A2).

Funding

Bowen Gang's research was supported by STCSM grant 22YF1403000.

References

Angelopoulos, A. N., S. Bates, E. J. Candès, M. I. Jordan, and L. Lei (2021). Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv:2110.01052*, Preprint.

Avella-Medina, M., H. S. Battey, J. Fan, and Q. Li (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika 105*(2), 271–284.

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085.

Bates, S., E. Candès, L. Lei, Y. Romano, and M. Sesia (2021). Testing for outliers with conformal p-values. *arXiv:2104.08279*, Preprint.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal 52*(6), 708–721.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* ***57***, 289–300.

Bickel, P. J. and E. Levina (2004). Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli 10*(6), 989–1010.

Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics 37*, 1685–1704.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association 106*(496), 1566–1577.

Cai, T., W. Liu, and X. Luo (2011). A constrained $l1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cai, T. and L. Zhang (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(4), 675–705.

Cai, T. T., W. Liu, H. H. Zhou, et al. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics 44*(2), 455–488.

Danziger, S. A., R. Baronio, L. Ho, L. Hall, K. Salmon, G. W. Hatfield, P. Kaiser, and R. H. Lathrop (2009). Predicting positive p53 cancer rescue regions using most informative positive (mip) active learning. *PLoS computational biology 5*(9), e1000498.

Dicker, L. H. and S. D. Zhao (2016). High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika 103*(1), 21–34.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association 104*(487), 1015–1028.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

Franc, V., D. Prusa, and V. Voracek (2021). Optimal strategies for reject option classifiers. *arXiv preprint arXiv:2101.12523*.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association 84*(405), 165–175.

Greenshtein, E. and J. Park (2009). Application of non parametric empirical bayes estimation to high dimensional classification. *Journal of Machine Learning Research 10*(7).

Guan, L. and R. Tibshirani (2022). Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society. Series B, Statistical Methodology 84*(2), 524.

Herbei, R. and M. H. Wegkamp (2006). Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 709–721.

James, W. and C. Stein (1992). Estimation with quadratic loss. In *Breakthroughs in statistics*, pp. 443–460. Springer.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Lei, J. (2014). Classification with confidence. *Biometrika 101*(4), 755–769.

Leung, D. and W. Sun (2021). Zap: z-value adaptive procedures for false discovery rate control with side information. *arXiv preprint arXiv:2108.12623*.

Liu, W. and X. Luo (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of multivariate analysis 135*, 153–162.

Loh, P.-L. and X. L. Tan (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics 12*(1), 1429–1467.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

Rava, B., W. Sun, G. M. James, and X. Tong (2021). A burden shared is a burden halved: A fairness-adjusted approach to classification. *arXiv:2110.05720*, Preprint.

Rigollet, P. and X. Tong (2011). Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 2831–2855.

Scott, C. and R. Nowak (2005). A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory 51*(11), 3806–3819.

Shao, J., Y. Wang, X. Deng, S. Wang, et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statis-*

tics 39(2), 1241–1265.

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research 14*(1), 3385–3418.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* ***102***, 901–912.

Sun, W. and Z. Wei (2011). Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments. *J. Amer. Statist. Assoc.* ***106***, 73–88.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 104–117.

Wang, H., A. Banerjee, C.-J. Hsieh, P. Ravikumar, and I. S. Dhillon (2013). Large scale distributed sparse precision estimation. In *NIPS*, Volume 13, pp. 584–592.

Witten, D. M. and R. Tibshirani (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 615–636.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.

Department of Statistics and Data Science, Fudan University

E-mail: (bgang@fudan.edu)

Department of Statistics, University of Chicago

E-mail: (yuantaos@uchicago.edu)

Center for Data Science, Zhejiang University

E-mail: (wenguans@marshall.usc.edu)