Optimal Feature Selection in High-Dimensional Discriminant Analysis

Mladen Kolar and Han Liu

***Abstract*— We consider the high-dimensional discriminant analysis problem. For this problem, different methods have been proposed and justified by establishing exact convergence rates for the classification risk, as well as the ℓ_2 convergence results to the discriminative rule. However, sharp theoretical analysis for the variable selection performance of these procedures have not been established, even though model interpretation is of fundamental importance in scientific data analysis. This paper bridges the gap by providing sharp sufficient conditions for consistent variable selection using the sparse discriminant analysis. Through careful analysis, we establish rates of convergence that are significantly faster than the best known results and admit an optimal scaling of the sample size n, dimensionality p, and sparsity level s in the high-dimensional setting. Sufficient conditions are complemented by the necessary information theoretic limits on the variable selection problem in the context of high-dimensional discriminant analysis. Exploiting a numerical equivalence result, our method also establish the optimal results for the ROAD estimator and the sparse optimal scoring estimator. Furthermore, we analyze an exhaustive search procedure, whose performance serves as a benchmark, and show that it is variable selection consistent under weaker conditions. Extensive simulations demonstrating the sharpness of the bounds are also provided.**

***Index Terms*— High-dimensional statistics, discriminant analysis, variable selection, optimal rates of convergence.**

I. Introduction

We consider the problem of binary classification with high-dimensional features. More specifically, given n data points, $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, sampled from a joint distribution of $(\mathbf{X}, Y) \in \mathbb{R}^p \times \{1, 2\}$, we want to determine the class label y for a new data point $\mathbf{x} \in \mathbb{R}^p$.

Let $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ be the density functions of $\mathbf{X}$ given $Y = 1$ (class 1) and $Y = 2$ (class 2) respectively, and the prior probabilities $\pi_1 = \mathbb{P}(Y = 1)$, $\pi_2 = \mathbb{P}(Y = 2)$. Classical multivariate analysis theory shows that the Bayes rule classifies a new data point $\mathbf{x}$ to class 2 if and only if

$$\log\left(\frac{p_2(\mathbf{x})}{p_1(\mathbf{x})}\right) + \log\left(\frac{\pi_2}{\pi_1}\right) > 0. \tag{I.1}$$

The Bayes rule usually serves as an oracle benchmark, since, in practical data analysis, the class conditional densities $p_2(\mathbf{x})$ and $p_1(\mathbf{x})$ are unknown and need to be estimated from the data.

Throughout the paper, we assume that the class conditional densities $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are Gaussian. That is, we assume that

$$\begin{aligned} \mathbf{X}|Y = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\ \text{and } \mathbf{X}|Y = 2 &\sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}). \end{aligned} \tag{I.2}$$

This assumption leads us to linear discriminant analysis (LDA) and the Bayes rule in (I.1) becomes

$$g(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) := \begin{cases} 2 & \text{if } \delta(\mathbf{x}) > 0 \\ 1 & \text{otherwise,} \end{cases}$$

where $\delta(\mathbf{x}) = \left(\mathbf{x} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2\right)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \log(\pi_2/\pi_1)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Theoretical properties of the plug-in rule $g(\mathbf{x}; \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}})$, where $(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}})$ are sample estimates of $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, have been well studied when the dimension p is low [1].

In high-dimensions, the standard plug-in rule works poorly and may even fail completely. For example, [2] shows that the classical low dimensional normal-based linear discriminant analysis is asymptotically equivalent to random guessing when the dimension p increases at a rate comparable to the sample size n. To overcome this curse of dimensionality, it is common to impose certain sparsity assumptions on the model and then estimate the high-dimensional discriminant rule using plug-in estimators. The most popular approach is to assume that both $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are sparse. Under this assumption, [22] proposes to use a thresholding procedure to estimate $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ and then plug them into the Bayes rule. In a more extreme case, [8], [23], [28] assume that $\boldsymbol{\Sigma} = \mathbf{I}$ and estimate $\boldsymbol{\mu}$ using a shrinkage method. Another common approach is to assume that $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu}$ are sparse. Under this assumption, [30] proposes the scout method which estimates $\boldsymbol{\Sigma}^{-1}$ using a shrunken estimator. Though these plug-in approaches are simple, they are not appropriate for conducting variable selection in the discriminant analysis setting. As has been elaborated in [1] and [11], for variable selection in high-dimensional discriminant analysis, we need to directly impose sparsity assumptions on the Bayes discriminant direction

Manuscript received October 31, 2013; revised October 14, 2014; accepted November 10, 2014. Date of publication December 18, 2014; date of current version January 16, 2015. This work was supported in part by the Division of Information and Intelligent Systems under Grant IIS-1116730, in part by the National Science Foundation under Grant IIS1408910 and Grant IIS1332109, in part by the National Institutes of Health under Grant R01MH102339, Grant R01GM083084, and Grant R01HG06841, and in part by the IBM Corporation Faculty Research Fund through the University of Chicago Booth School of Business, Chicago, IL, USA. This paper was presented at the 30th International Conference on Machine Learning [13].

M. Kolar is with the University of Chicago Booth School of Business, Chicago, IL 60611 USA (e-mail: mkolar@chicagobooth.edu).

H. Liu is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: hanliu@princeton.edu).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2014.2381241

$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ instead of separately on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. In particular, it is assumed that $\boldsymbol{\beta} = (\boldsymbol{\beta}_T', \mathbf{0}')'$ for $T = \{1, \ldots, s\}$. Their key observation comes from the fact that the Fisher's discriminant rule depends on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ only through the product $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Furthermore, in the high-dimensional setting, it is scientifically meaningful that only a small set of variables are relevant to classification, which is equivalent to the assumption that $\boldsymbol{\beta}$ is sparse. On a simple example of tumor classification, [16] elaborates why it is scientifically more informative to directly impose sparsity assumption on $\boldsymbol{\beta}$ instead of on $\boldsymbol{\mu}$ (For more details, see Section 2 of their paper). In addition, [5] points out that the sparsity assumption on $\boldsymbol{\beta}$ is much weaker than imposing sparsity assumptions $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu}$ separately. A number of authors have also studied classification in this setting [5], [6], [10], [16], [31], [32].

In this paper, we adopt the assumption that $\boldsymbol{\beta}$ is sparse and focus on analyzing the SDA (Sparse Discriminant Analysis) proposed by [16]. This method estimates the discriminant direction $\boldsymbol{\beta}$ (More precisely, they estimate a quantity that is proportional to $\boldsymbol{\beta}$.) and our focus will be on variable selection consistency, that is, whether this method can recover the set T with high probability. In a recent work, [15] proves that the SDA estimator is numerically equivalent to the ROAD estimator [10] and the sparse optimal scoring estimator [6]. By exploiting this result, our theoretical analysis provides a unified theoretical justification for all these three methods.

A. Main Results

Let $n_1 = |\{i : y_i = 1\}|$ and $n_2 = n - n_1$. The SDA estimator is obtained by solving the following least squares optimization problem

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2(n-2)} \sum_{i\in[n]} (z_i - \mathbf{v}'(\mathbf{x}_i - \bar{\mathbf{x}}))^2 + \lambda||\mathbf{v}||_1, \qquad \text{(I.3)}$$

where $[n]$ denotes the set $\{1, \ldots, n\}$, $\bar{\mathbf{x}} = n^{-1}\sum_i \mathbf{x}_i$ and the vector $\mathbf{z} \in \mathbb{R}^n$ encodes the class labels as $z_i = n_2/n$ if $y_i = 1$ and $z_i = -n_1/n$ if $y_i = 2$. Here $\lambda > 0$ is a regularization parameter.

The SDA estimator in (I.3) uses an ℓ_1-norm penalty to estimate a sparse $\mathbf{v}$ and avoid the curse of dimensionality. [16] studied its variable selection property under a different encoding scheme of the response z_i. However, as we show later, different coding schemes do not affect the results (see Appendix C). When the regularization parameter λ is set to zero, the SDA estimator reduces to the classical Fisher's discriminant rule.

The main focus of the paper is to sharply characterize the variable selection performance of the SDA estimator. From a theoretical perspective, unlike the high dimensional regression setting where sharp theoretical results exist for prediction, estimation, and variable selection consistency, most existing theories for high-dimensional discriminant analysis are either on estimation consistency or risk consistency, but not on variable selection consistency (see [5], [10], [22]). [16] provides a variable selection consistency result for the SDA estimator in (I.3). However, as we will show later, their obtained scaling in terms of (n, p, s) is not optimal. Though some theoretical analysis of the ℓ_1-norm penalized M-estimators exists (see [21], [26]), these techniques are not applicable to analyze the estimator given in (I.3). In high-dimensional discriminant analysis the underlying statistical model is different from that of the regression analysis. At a high level, to establish variable selection consistency of the SDA estimator, we characterize the Karush-Kuhn-Tucker (KKT) conditions for the optimization problem in (I.3). Unlike the ℓ_1-norm penalized least squares regression, which directly estimates the regression coefficients, the solution to (I.3) is a quantity that is only proportional to the Bayes rule's direction. To analyze such scaled estimators, we need to resort to different techniques and utilize sophisticated multivariate analysis results to characterize the sampling distributions of the estimated quantities. More specifically, we provide sufficient conditions under which the SDA estimator is variable selection consistent with a significantly improved scaling compared to that obtained by [16]. In addition, we complement these sufficient conditions with information theoretic limitations on recovery of the feature set T. In particular, we provide lower bounds on the sample size and the signal level needed to recover the set of relevant variables by any procedure. We identify the family of problems for which the estimator (I.3) is variable selection optimal. To provide more insights into the problem, we analyze an exhaustive search procedure, which requires weaker conditions to consistently select relevant variables. This estimator, however, is not practical and serves only as a benchmark. The obtained variable selection consistency result also enables us to establish risk consistency for the SDA estimator. In addition, [15] shows that the SDA estimator is numerically equivalent to the ROAD estimator proposed by [10] and [32] and the sparse optimal scoring estimator proposed by [6]. Therefore, the results provided in this paper also apply to those estimators. Some of the main results of this paper are summarized below.

Let $\widehat{\mathbf{v}}^{\mathrm{SDA}}$ denote the minimizer of (I.3). We show that if the sample size

$$n \geq C\left(\max_{a\in N} \sigma_{a|T}\right) \Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log\left((p-s)\log(n)\right), \qquad \text{(I.4)}$$

where C is a fixed constant which does not scale with n, p and s, $\sigma_{a|T} = \sigma_{aa} - \boldsymbol{\Sigma}_{aT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{Ta}$, and $\Lambda_{\min}(\boldsymbol{\Sigma})$ denotes the minimum eigenvalue of $\boldsymbol{\Sigma}$, then the estimated vector $\widehat{\mathbf{v}}^{\mathrm{SDA}}$ has the same sparsity pattern as the true $\boldsymbol{\beta}$, thus establishing variable selection consistency (or sparsistency) for the SDA estimator. This is the first result that proves that consistent variable selection in the discriminant analysis can be done under a similar theoretical scaling as variable selection in the regression setting (in terms of n, p and s). To prove (I.4), we impose conditions that $\min_{j\in T}|\beta_j|$ is not too small and $||\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty \leq 1-\alpha$ with $\alpha \in (0, 1)$, where $N = [p]\backslash T$. The latter one is the irrepresentable condition, which is commonly used in the ℓ_1-norm penalized least squares regression problem [19], [26], [35], [36]. Let $\beta_{\min}$ be the magnitude of the smallest absolute value of the non-zero component of $\boldsymbol{\beta}$. Our analysis of information theoretic limitations reveals that, whenever $n < C_1\beta_{\min}^{-2}\log(p-s)$, no procedure can reliably recover the set T. In particular, under

certain regimes, we establish that the SDA estimator is optimal for the purpose of variable selection. The analysis of the exhaustive search decoder reveals a similar result. However, the exhaustive search decoder does not need the irrepresentable condition to be satisfied by the covariance matrix. Thorough numerical simulations are provided to demonstrate the sharpness of our theoretical results.

In a preliminary work, [13] presented some variable selection consistency results related to the ROAD estimator under the assumption that $\pi_1 = \pi_2 = 1/2$. However, it is hard to directly compare their analysis with that of [16] to understand why an improved scaling is achievable, since the ROAD estimator is the solution to a constrained optimization while the SDA estimator is the solution to an unconstrained optimization. This paper analyzes the SDA estimator and is directly comparable with the result of [16]. As we will discuss later, our analysis attains better scaling due to a more careful characterization of the sampling distributions of several scaled statistics. In contrast, the analysis in [16] hinges on the sup-norm control of the deviation of the sample mean and covariance to their population quantities, which is not sufficient to obtain the optimal rate. Using the numerical equivalence between the SDA and the ROAD estimator, the theoretical results of this paper also apply on the ROAD estimator. In addition, we also study an exhaustive search decoder and information theoretic limits on the variable selection in high-dimensional discriminant analysis. Furthermore, we provide discussions on risk consistency and approximate sparsity, which shed light on future investigations.

The rest of this paper is organized as follows. In the rest of this section, we introduce some more notation. In §II, we study sparsistency of the SDA estimator. An information theoretic lower bound is given in §III. We characterize the behavior of the exhaustive search procedure in §IV. Consequences of our results are discussed in more details in §V. Numerical simulations that illustrate our theoretical findings are given in §VI. We conclude the paper with a discussion and some results on the risk consistency and approximate sparsity in §VII. Technical results and proofs are deferred to the appendix and online supplementary document.

B. Notation

We denote $[n]$ to be the set $\{1, \ldots, n\}$. Let $T \subseteq [p]$ be an index set, we denote $\boldsymbol{\beta}_T$ to be the subvector containing the entries of the vector $\boldsymbol{\beta}$ indexed by the set T, and $\mathbf{X}_T$ denotes the submatrix containing the columns of $\mathbf{X}$ indexed by T. Similarly, we denote $\mathbf{A}_{TT}$ to be the submatrix of $\mathbf{A}$ with rows and columns indexed by T. For a vector $\mathbf{a} \in \mathbb{R}^n$, we denote $\text{supp}(\mathbf{a}) = \{j : a_j \neq 0\}$ to be the support set. We also use $||\mathbf{a}||_q$, $q \in [1, \infty)$, to be the ℓ_q-norm defined as $||\mathbf{a}||_q = (\sum_{i\in[n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $||\mathbf{a}||_0 = |\text{supp}(\mathbf{a})|$ and $||\mathbf{a}||_\infty = \max_{i\in[n]} |a_i|$. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$, we denote $|||\mathbf{A}|||_\infty = \max_{i\in[n]} \sum_{j\in[p]} |a_{ij}|$ the ℓ_∞ operator norm. For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ we denote $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ to be the smallest and largest eigenvalues, respectively. We also represent the quadratic form $||\mathbf{a}||_{\mathbf{A}}^2 = \mathbf{a}'\mathbf{A}\mathbf{a}$ for a symmetric positive definite matrix $\mathbf{A}$. We denote $\mathbf{I}_n$ to be the $n\times n$ identity matrix and $\mathbf{1}_n$ to be the $n \times 1$ vector with all components equal to 1. For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = \mathcal{O}(b_n)$ to denote that $a_n < Cb_n$ for some finite positive constant C. We also denote $a_n = \mathcal{O}(b_n)$ to be $b_n \gtrsim a_n$. If $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$, we denote it to be $a_n \asymp b_n$. The notation $a_n = o(b_n)$ is used to denote that $a_n b_n^{-1} \to 0$.

II. Sparsistency of the SDA Estimator

In this section, we provide sharp sparsistency analysis for the SDA estimator defined in (I.3). Our analysis decomposes into two parts: (i) We first analyze the population version of the SDA estimator in which we assume that $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ are known. The solution to the population problem provides us insights on the variable selection problem and allows us to write down sufficient conditions for consistent variable selection. (ii) We then extend the analysis from the population problem to the sample version of the problem in (I.3). For this, we need to replace $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ by their corresponding sample estimates $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\mu}}_1$, and $\widehat{\boldsymbol{\mu}}_2$. The statement of the main result is provided in §II-B with an outline of the proof in §II-C.

A. Population Version Analysis of the SDA Estimator

We first lay out conditions that characterize the solution to the population version of the SDA optimization problem.

Let $\mathbf{X}_1 \in \mathbb{R}^{n_1\times p}$ be the matrix with rows containing data points from the first class and similarly define $\mathbf{X}_2 \in \mathbb{R}^{n_2\times p}$ to be the matrix with rows containing data points from the second class. We denote $\mathbf{H}_1 = \mathbf{I}_{n_1} - n_1^{-1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}'$ and $\mathbf{H}_2 = \mathbf{I}_{n_2} - n_2^{-1}\mathbf{1}_{n_2}\mathbf{1}_{n_2}'$ to be the centering matrices. We define the following quantities

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_1 &= n_1^{-1}\sum_{i:y_i=1} \mathbf{x}_i = n_1^{-1}\mathbf{X}_1'\mathbf{1}_{n_1},\\
\widehat{\boldsymbol{\mu}}_2 &= n_2^{-1}\sum_{i:y_i=2} \mathbf{x}_i = n_2^{-1}\mathbf{X}_2'\mathbf{1}_{n_2},\\
\widehat{\boldsymbol{\mu}} &= \widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1,\\
\mathbf{S}_1 &= (n_1-1)^{-1}\mathbf{X}_1'\mathbf{H}_1\mathbf{X}_1,\\
\mathbf{S}_2 &= (n_2-1)^{-1}\mathbf{X}_2'\mathbf{H}_2\mathbf{X}_2,\\
\mathbf{S} &= (n-2)^{-1}((n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2).
\end{aligned}$$

With this notation, observe that the optimization problem in (I.3) can be rewritten as

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\mathbf{v}'\left(\mathbf{S} + \frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}'\right)\mathbf{v} - \frac{n_1 n_2}{n(n-2)}\mathbf{v}'\widehat{\boldsymbol{\mu}} + \lambda||\mathbf{v}||_1,$$

where we have dropped terms that do not depend on $\mathbf{v}$. Therefore, we define the population version of the SDA optimization problem as

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\left(\boldsymbol{\Sigma} + \pi_1\pi_2\boldsymbol{\mu}\boldsymbol{\mu}'\right)\mathbf{w} - \pi_1\pi_2\mathbf{w}'\boldsymbol{\mu} + \lambda||\mathbf{w}||_1, \quad \text{(II.1)}$$

Let $\widehat{\mathbf{w}}$ be the solution of (II.1). We are aiming to characterize conditions under which the solution $\widehat{\mathbf{w}}$ recovers the sparsity pattern of $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Recall that $T = \text{supp}(\boldsymbol{\beta}) = \{1, \ldots, s\}$

denotes the true support set and $N = [p]\backslash T$, under the sparsity assumption, we have

$$\boldsymbol{\beta}_T = \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T \quad \text{and} \quad \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T. \tag{II.2}$$

We define $\beta_{\min}$ as

$$\beta_{\min} = \min_{a\in T} |\beta_a|.$$

The following theorem characterizes the solution to the population version of the SDA optimization problem in (II.1).

Theorem 1: Let $\alpha \in (0, 1]$ be a constant and $\widehat{\mathbf{w}}$ be the solution to the problem in (II.1)*. Under the assumptions that*

$$||\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty \le 1-\alpha, \tag{II.3}$$

$$\pi_1\pi_2\frac{1+\lambda||\boldsymbol{\beta}_T||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}\beta_{\min} > \lambda||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty, \tag{II.4}$$

we have $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_T', \mathbf{0}')$ with

$$\widehat{\mathbf{w}}_T = \pi_1\pi_2\frac{1+\lambda||\boldsymbol{\beta}_T||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}\boldsymbol{\beta}_T - \lambda\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T). \tag{II.5}$$

Furthermore, we have $\operatorname{sign}(\widehat{\mathbf{w}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$.

Equations (II.3) and (II.4) provide sufficient conditions under which the solution to (II.1) recovers the true support. The condition in (II.3) takes the same form as the irrepresentable condition commonly used in the ℓ_1-penalized least squares regression problem [19], [26], [35], [36]. Equation (II.4) specifies that the smallest component of $\boldsymbol{\beta}_T$ should not be too small compared to the regularization parameter λ. In particular, let $\lambda = \lambda_0/(1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}})$ for some λ_0. Then (II.4) suggests that $\widehat{\mathbf{w}}_T$ recovers the true support of $\boldsymbol{\beta}$ as long as $\beta_{\min} \ge \lambda_0||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty$. Equation (II.5) provides an explicit form for the solution $\widehat{\mathbf{w}}$, from which we see that the SDA optimization procedure estimates a scaled version of the optimal discriminant direction (when $\lambda = 0$). Whenever $\lambda \neq 0$, $\widehat{\mathbf{w}}$ is a biased estimator. However, such estimation bias does not affect the recovery of the support set T of $\boldsymbol{\beta}$ when λ is small enough.

We present the proof of Theorem 1, as the analysis of the sample version of the SDA estimator will follow the same lines. We start with the Karush-Kuhn-Tucker (KKT) conditions for the optimization problem in (II.1):

$$\left(\boldsymbol{\Sigma} + \pi_1\pi_2\boldsymbol{\mu}\boldsymbol{\mu}'\right)\widehat{\mathbf{w}} - \pi_1\pi_2\boldsymbol{\mu} + \lambda\widehat{\mathbf{z}} = \mathbf{0} \tag{II.6}$$

where $\widehat{\mathbf{z}} \in \partial||\widehat{\mathbf{w}}||_1$ is an element of the subdifferential of $\|\cdot\|_1$.

Let $\widehat{\mathbf{w}}_T$ be defined in (II.5). We need to show that there exists a $\widehat{\mathbf{z}}$ such that the vector $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_T', \mathbf{0}')'$, paired with $\widehat{\mathbf{z}}$, satisfies the KKT conditions and $\operatorname{sign}(\widehat{\mathbf{w}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$. This is achieved in two steps.

Step 1: Define the following oracle optimization problem

$$\min_{\mathbf{w}_T}\ \frac{1}{2}\mathbf{w}_T'\left(\boldsymbol{\Sigma}_{TT} + \pi_1\pi_2\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\right)\mathbf{w}_T - \pi_1\pi_2\mathbf{w}_T'\boldsymbol{\mu}_T + \lambda\mathbf{w}_T'\operatorname{sign}(\boldsymbol{\beta}_T) \tag{II.7}$$

and let $\widetilde{\mathbf{w}}_T$ be the solution to the above optimization problem. In this step we establish that $\operatorname{sign}(\widetilde{\mathbf{w}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$. This is obvious under the conditions of Theorem 1 in the population setting, but will be much more challenging to establish in the sample version of the problem studied in the next section.

Step 2: Verify that $(\widetilde{\mathbf{w}}_T', \mathbf{0}')'$ is the solution to the optimization problem in (II.1) under the assumptions of Theorem 1. The following lemma achieves exactly that.

Lemma 1: Under the conditions of Theorem 1, we have that $\widehat{\mathbf{w}} = (\widetilde{\mathbf{w}}_T', \mathbf{0}')$ is the solution to the problem in (II.1)*, where $\widetilde{\mathbf{w}}_T$ is defined as the minimizer of* (II.7)*.*

Theorem 1 immediately follows from the two steps above.

The above two steps will be used to prove results about the sample version of the SDA estimator as well. Note that, in practice, one cannot form the oracle optimization problem and hence the two steps only provide a constructive way to verify variable selection consistency of the SDA estimator.

The next theorem shows that the irrepresentable condition in (II.3) is almost necessary for sign consistency, even if the population quantities $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are known.

Theorem 2: Let $\widehat{\mathbf{w}}$ be the solution to the problem in (II.1)*. If we have $\operatorname{sign}(\widehat{\mathbf{w}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$, Then, there must be*

$$||\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty \le 1.$$

The proof of this theorem follows similar argument as in the regression settings in [17] and [19].

B. Sample Version Analysis of the SDA Estimator

In this section, we analyze the variable selection performance of the sample version of the SDA estimator $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}^{\text{SDA}}$ defined in (I.3). In particular, we will establish sufficient conditions under which $\widehat{\mathbf{v}}$ correctly recovers the support set of $\boldsymbol{\beta}$ (i.e., we will derive conditions under which $\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}_T', \mathbf{0}')'$ and $\operatorname{sign}(\widehat{\mathbf{v}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$). The proof construction follows the same line of reasoning as the population version analysis. However, proving analogous results in the sample version of the problem is much more challenging and requires careful analysis of the sampling distribution of the scaled functionals of Gaussian random vectors.

The following theorem is the main result that characterizes the variable selection consistency of the SDA estimator.

Theorem 3: We assume that the condition in (II.3) *holds. We denote*

$$A_{\boldsymbol{\beta}} := \left(1 \vee ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right).$$

Choosing $\lambda = \left(1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)^{-1}\lambda_0$ with

$$\lambda_0 = K_{\lambda_0}\sqrt{\pi_1\pi_2\left(\max_{a\in N}\sigma_{a|T}\right)A_{\boldsymbol{\beta}}\frac{\log\left((p-s)\log(n)\right)}{n}}$$

where K_{λ_0} is a sufficiently large constant. Suppose that $\beta_{\min} = \min_{a\in T}|\beta_a|$ satisfies

$$\beta_{\min} \ge K_\beta\left(\sqrt{\left(\max_{a\in T}\left(\boldsymbol{\Sigma}_{TT}^{-1}\right)_{aa}\right)A_{\boldsymbol{\beta}}\frac{\log(s\log(n))}{n}} \bigvee \lambda_0||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty\right) \tag{II.8}$$

for a sufficiently large constant K_β. If

$$n \ge \frac{K\pi_1\pi_2\left(\max_{a\in N}\sigma_{a|T}\right)s\log\left((p-s)\log(n)\right)}{\Lambda_{\min}(\boldsymbol{\Sigma}_{TT})}$$

for some constant K, then $\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}_T', \mathbf{0}')'$ is the solution to the optimization problem in (I.3)*, where*

$$\widehat{\mathbf{v}}_T = \frac{n_1 n_2}{n(n-2)} \frac{1 + \lambda ||\widehat{\boldsymbol{\beta}_T}||_1}{1 + \frac{n_1 n_2}{n(n-2)} ||\widehat{\boldsymbol{\beta}_T}||^2_{\mathbf{S}_{TT}}} \widehat{\boldsymbol{\beta}_T} - \lambda \mathbf{S}_{TT}^{-1} \operatorname{sign}(\widehat{\boldsymbol{\beta}_T}) \tag{II.9}$$

and $\widehat{\boldsymbol{\beta}}_T = \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T$, with probability at least $1 - \mathcal{O}\left(\log^{-1}(n)\right)$. Furthermore, $\operatorname{sign}(\widehat{\mathbf{v}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$.

Theorem 3 is a sample version of Theorem 1 given in the previous section. Compared to the population version result, in addition to the irrepresentable condition and a lower bound on $\beta_{\min}$, we also need the sample size n to be large enough for the SDA procedure to recover the true support set T with high probability.

At the first sight, the conditions of the theorem look complicated. To highlight the main result, we consider a case where $0 < \underline{c} \leq \Lambda_{\min}(\boldsymbol{\Sigma}_{TT})$ and $\left(||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \bigvee ||\boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)||_\infty\right) \leq \bar{C} < \infty$ for some constants $\underline{c}, \bar{C}$. In this case, it is sufficient that the sample size scales as $n \asymp s \log(p-s)$ and $\beta_{\min} \gtrsim s^{-1/2}$. This scaling is of the same order as for the Lasso procedure, where $n \gtrsim s \log(p-s)$ is needed for correct recovery of the relevant variables under the same assumptions (see [26, Th. 3]). In §V, we provide more detailed explanation of this theorem and complement it with the necessary conditions given by the information theoretic limits.

Variable selection consistency of the SDA estimator was studied by [16]. Let $\mathbf{C} = \operatorname{Var}(\mathbf{X})$ denote the marginal covariance matrix (note that, in general, $\mathbf{C} \neq \boldsymbol{\Sigma}$). Under the assumption that $|||\mathbf{C}_{NT}\mathbf{C}_{TT}^{-1}|||_\infty$, $|||\mathbf{C}_{TT}^{-1}|||_\infty$ and $||\boldsymbol{\mu}||_\infty$ are bounded, [16] shows that the following conditions

$$i)\ \lim_{n\to\infty} \frac{s^2 \log p}{n} = 0,$$
$$\text{and} \quad ii)\ \beta_{\min} \gg \sqrt{\frac{s^2 \log(ps)}{n}} \tag{II.10}$$

are sufficient for consistent support recovery of $\boldsymbol{\beta}$. This is suboptimal compared to our results. Inspection of the proof given in [16] reveals that their result hinges on uniform control of the elementwise deviation of $\widehat{\mathbf{C}}$ from $\mathbf{C}$ and $\widehat{\boldsymbol{\mu}}$ from $\boldsymbol{\mu}$. These uniform deviation controls are too rough to establish sharp results given in Theorem 3. In our proofs, we use more sophisticated multivariate analysis tools to control the deviation of $\widehat{\boldsymbol{\beta}_T}$ from $\boldsymbol{\beta}_T$, that is, we focus on analyzing the quantity $\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T$ but instead of studying $\mathbf{S}_{TT}$ and $\widehat{\boldsymbol{\mu}}_T$ separately. The condition $|||\mathbf{C}_{NT}\mathbf{C}_{TT}^{-1}|||_\infty < 1$, given in [16], is equivalent to assuming that $||\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{a}||_\infty < 1$ for all $\mathbf{a} \in \mathbb{R}^s$ such that $||\mathbf{a}||_\infty \leq 1$, since

$$\mathbf{C}_{NT}\mathbf{C}_{TT}^{-1} = \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}, \tag{II.11}$$

as shown in Appendix D. On the other hand, Theorem 3 requires only that $||\Sigma_{NT}\Sigma_{TT}^{-1}\mathbf{a}||_\infty < 1$ holds for $\mathbf{a} = \operatorname{sign}(\boldsymbol{\beta}_T)$. Therefore, our irrepresentable condition is weaker than the one in [16].

A semiparametric extension of the SDA estimator, termed CODA, was studied in [11]. Under the scaling conditions in (II.10), the CODA estimator is able to perform correct variable selection. As we discussed above, these conditions are suboptimal compared to our results. Inspection of the proof in [11] reveals that the analysis of CODA hinges on the uniform control of the elementwise deviation of the sample mean and a rank-based covariance estimator from their population counterparts, which are too crude for establishing sharp scaling results. Furthermore, the proof technique used to establish results of Theorem 3 heavily relies on the assumption that the class conditional distribution of $\mathbf{X}$ is a multivariate Gaussian. This assumption allows us to use results about the distribution of the inverse of a block of the covariance matrix established in [4] (see, for example, proofs of Lemma 3 and Lemma 7 in the Appendix). Estimator of the covariance matrix used in the CODA estimator is based on the non-linear transformation of the Kendall's tau matrix for which we do not have such results. Therefore, improving the results of the CODA estimator would require a different proof strategy.

The proof of Theorem 3 is outlined in the next subsection.

C. Proof of Sparsistency of the SDA Estimator

The proof of Theorem 3 follows the same strategy as the proof of Theorem 1. More specifically, we only need to show that there exists a subdifferential of $\|\cdot\|_1$ such that the solution $\widehat{\mathbf{v}}$ to the optimization problem in (I.3) satisfies the sample version KKT condition (given below in (II.12) and (II.13)) with high probability. For this, we proceed in two steps. In the first step, we assume that the true support set T is known and solve an oracle optimization problem (given below in (II.14)), which exploits the knowledge of T. Let $\widetilde{\mathbf{v}}_T$ be the solution to the oracle optimization problem. In the second step, we show that there exists a dual variable from the subdifferential of $\|\cdot\|_1$ such that the vector $(\widetilde{\mathbf{v}}_T', \mathbf{0}')'$ satisfies the KKT conditions for the original optimization problem given in (I.3). This proves that $\widehat{\mathbf{v}} = (\widetilde{\mathbf{v}}_T', \mathbf{0}')'$ is a global minimizer of the problem in (I.3). Finally, we show that $\widehat{\mathbf{v}}$ is a unique solution to the optimization problem in (I.3) with high probability.

Let $\widehat{T} = \operatorname{supp}(\widehat{\mathbf{v}})$ be the support of a solution $\widehat{\mathbf{v}}$ to the optimization problem in (I.3) and $\widehat{N} = [p] \backslash \widehat{T}$. Any solution to (I.3) needs to satisfy the following Karush-Kuhn-Tucker (KKT) conditions

$$\left(\mathbf{S}_{\widehat{T}\widehat{T}} + \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}}_{\widehat{T}} \widehat{\boldsymbol{\mu}}_{\widehat{T}}'\right) \widehat{\mathbf{v}}_{\widehat{T}} = \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}}_{\widehat{T}} - \lambda \operatorname{sign}(\widehat{\mathbf{v}}_{\widehat{T}}), \tag{II.12}$$

$$\left|\left|\left(\mathbf{S}_{\widehat{N}\widehat{T}} + \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}}_{\widehat{N}} \widehat{\boldsymbol{\mu}}_{\widehat{T}}'\right) \widehat{\mathbf{v}}_{\widehat{T}} - \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}}_{\widehat{N}}\right|\right|_\infty \leq \lambda. \tag{II.13}$$

We construct a solution $\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}_T', \mathbf{0}')'$ to (I.3) and show that it is unique with high probability.

Step 1: We consider the following oracle optimization problem

$$\widetilde{\mathbf{v}}_T = \arg\min_{\mathbf{v} \in \mathbb{R}^s} \frac{1}{2(n-2)} \sum_{i \in [n]} (z_i - \mathbf{v}'(\mathbf{x}_{i,T} - \bar{\mathbf{x}}_T))^2 + \lambda \mathbf{v}' \operatorname{sign}(\boldsymbol{\beta}_T). \tag{II.14}$$

The optimization problem in (II.14) is related to the one in (I.3), however, the solution is calculated only over the

subset T and $||\mathbf{v}_T||_1$ is replaced with $\mathbf{v}_T' \operatorname{sign}(\boldsymbol{\beta}_T)$. Simple algebra gives

$$\widetilde{\mathbf{v}}_T = \frac{n_1 n_2}{n(n-2)} \frac{1 + \lambda \widehat{\boldsymbol{\beta}_T}' \operatorname{sign}(\boldsymbol{\beta}_T)}{1 + \frac{n_1 n_2}{n(n-2)} ||\widehat{\boldsymbol{\beta}_T}||^2_{\mathbf{S}_{TT}}} \widehat{\boldsymbol{\beta}_T} - \lambda \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T). \tag{II.15}$$

Recall that $\widehat{\boldsymbol{\beta}_T} = \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T$. The solution $\widetilde{\mathbf{v}}_T$ is unique, since the matrix $\mathbf{S}_{TT}$ is positive definite with probability 1.

The following result establishes that the solution to the auxiliary oracle optimization problem (II.14) satisfies $\operatorname{sign}(\widetilde{\mathbf{v}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$ with high probability, under the conditions of Theorem 3.

Lemma 2: Under the assumption that the conditions of Theorem 3 are satisfied, $\operatorname{sign}(\widetilde{\mathbf{v}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$ *and* $\operatorname{sign}(\widehat{\boldsymbol{\beta}_T}) = \operatorname{sign}(\boldsymbol{\beta}_T)$ *with probability at least* $1 - \mathcal{O}\left(\log^{-1}(n)\right)$.

The proof Lemma 2 relies on a careful characterization of the deviation of the following quantities $\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T$, $\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\widehat{\boldsymbol{\beta}_T})$, $\mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T$ and $\mathbf{S}_{TT}^{-1} \operatorname{sign}(\widehat{\boldsymbol{\beta}_T})$ from their expected values. Using Lemma 2, we have that $\widetilde{\mathbf{v}}_T$ defined in (II.15) satisfies $\widetilde{\mathbf{v}}_T = \widehat{\mathbf{v}}_T$.

Step 2: The following lemma shows that $\widehat{\mathbf{v}} = (\widetilde{\mathbf{v}}_T', \mathbf{0}')'$ is a solution to (I.3) under the conditions of Theorem 3.

Lemma 3: Assuming that the conditions of Theorem 3 are satisfied, we have that $\widehat{\mathbf{v}} = (\widetilde{\mathbf{v}}_T', \mathbf{0}')'$ *is a solution to* (I.3) *with probability at least* $1 - \mathcal{O}\left(\log^{-1}(n)\right)$.

The proof of Theorem 3 will be complete once we show that $\widehat{\mathbf{v}} = (\widetilde{\mathbf{v}}_T', \mathbf{0}')'$ is the unique solution. We proceed as in [26, Proof of Lemma 1]. Let $\check{\mathbf{v}}$ be another solution to the optimization problem in (I.3) satisfying the KKT condition

$$\left(\mathbf{S} + \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}} \widehat{\boldsymbol{\mu}}'\right) \check{\mathbf{v}} - \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}} + \lambda \widehat{\mathbf{q}} = \mathbf{0}$$

for some subgradient $\widehat{\mathbf{q}} \in \partial ||\check{\mathbf{v}}||_1$. Given the subgradient $\widehat{\mathbf{q}}$, any optimal solution needs to satisfy the complementary slackness condition $\widehat{\mathbf{q}}' \check{\mathbf{v}} = ||\check{\mathbf{v}}||_1$, which holds only if $\check{v}_j = 0$ for all j such that $|\widehat{q}_j| < 1$. In the proof of Lemma 3, we will establish that $|\widehat{q}_j| < 1$ for $j \in N$ (see (A.3)). Therefore, any solution to (I.3) has the same sparsity pattern as $\widehat{\mathbf{v}}$. Uniqueness now follows since $\widetilde{\mathbf{v}}_T$ is the unique solution of (II.14) when constrained on the support set T.

III. Lower Bound

Theorem 3 provides sufficient conditions for the SDA estimator to reliably recover the true set T of nonzero elements of the discriminant direction $\boldsymbol{\beta}$. In this section, we provide results that are of complementary nature. More specifically, we provide necessary conditions that must be satisfied for any procedure to succeed in reliable estimation of the support set T. Thus, we focus on the information theoretic limits in the context of high-dimensional discriminant analysis.

We denote Ψ to be an estimator of the support set T, that is, any measurable function that maps the data $\{\mathbf{x}_i, y_i\}_{i \in [n]}$ to a subset of $\{1, \ldots, p\}$. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ be the problem parameters and Θ be the parameter space. We define the maximum risk, corresponding to the $0/1$ loss, as

$$R(\Psi, \Theta) = \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{P}_{\boldsymbol{\theta}} \left[\Psi(\{\mathbf{x}_i, y_i\}_{i \in [n]}) \neq T(\boldsymbol{\theta})\right]$$

where $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the joint distribution of $\{\mathbf{x}_i, y_i\}_{i \in [n]}$ under the assumption that $\pi_1 = \pi_2 = \frac{1}{2}$, and $T(\boldsymbol{\theta}) = \operatorname{supp}(\boldsymbol{\beta})$ (recall that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$). Let $\mathcal{M}(s, \mathcal{Z})$ be the class of all subsets of the set $\mathcal{Z}$ of cardinality s. We consider the parameter space

$$\Theta(\boldsymbol{\Sigma}, \tau, s) = \bigcup_{\omega \in \mathcal{M}(s, [p])} \left\{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{F}_{\omega, \tau}\right\}$$

with $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \mathcal{F}_{\omega, \tau}$ if

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \in \begin{cases} |\beta_a| \geq \tau & \text{if } a \in \omega, \\ \beta_a = 0 & \text{if } a \notin \omega \end{cases}$$

where $\tau > 0$ determines the signal strength. The minimax risk is defined as

$$\inf_{\Psi} R(\Psi, \Theta(\boldsymbol{\Sigma}, \tau, s)).$$

In what follows we provide a lower bound on the minimax risk. Before stating the result, we introduce the following three quantities that will be used to state Theorem 4

$$\varphi_{\text{close}}(\boldsymbol{\Sigma}) = \min_{T \in \mathcal{M}(s, [p])} \min_{u \in T} \frac{1}{p - s} \sum_{v \in [p] \setminus T} (\Sigma_{uu} + \Sigma_{vv} - 2\Sigma_{uv}), \tag{III.1}$$

$$\varphi_{\text{far}}(\boldsymbol{\Sigma}) = \min_{T \in \mathcal{M}(s, [p])} \frac{1}{\boldsymbol{\Delta}_{\boldsymbol{\Sigma}} \binom{p-s}{s}} \sum_{T' \in \mathcal{M}(s, [p] \setminus T)} \mathbf{1}' \boldsymbol{\Sigma}_{T \cup T', T \cup T'} \mathbf{1}, \tag{III.2}$$

and

$$\tau_{\min} = 2 \cdot \max\left(\sqrt{\frac{\log \binom{p-s}{s}}{n \varphi_{\text{far}}(\boldsymbol{\Sigma})}}, \sqrt{\frac{\log(p - s + 1)}{n \varphi_{\text{close}}(\boldsymbol{\Sigma})}}\right).$$

The first quantity measures the difficulty of distinguishing two close support sets T_1 and T_2 that differ in only one position. The second quantity measures the effect of a large number of support sets that are far from the support set T. The quantity $\tau_{\min}$ is a threshold for the signal strength. Our main result on minimax lower bound is presented in Theorem 4.

Theorem 4: For any $\tau < \tau_{\min}$*, there exists some constant* $C > 0$*, such that*

$$\inf_{\Psi} \sup_{\boldsymbol{\theta} \in \Theta(\boldsymbol{\Sigma}, \tau, s)} \mathbb{P}_{\boldsymbol{\theta}} \left[\Psi(\{\mathbf{x}_i, y_i\}_{i \in [n]}) \neq T(\boldsymbol{\theta})\right] \geq C > 0.$$

Theorem 4 implies that for any estimating procedure, whenever $\tau < \tau_{\min}$, there exists some distribution parametrized by $\boldsymbol{\theta} \in \Theta(\boldsymbol{\Sigma}, \tau, s)$ such that the probability of incorrectly identifying the set $T(\boldsymbol{\theta})$ is strictly bounded away from zero. To better understand the quantities $\varphi_{\text{close}}(\boldsymbol{\Sigma})$ and $\varphi_{\text{far}}(\boldsymbol{\Sigma})$, we consider a special case when $\boldsymbol{\Sigma} = \mathbf{I}$. In this case both quantities simplify a lot and we have $\varphi_{\text{close}}(\mathbf{I}) = 2$ and $\varphi_{\text{far}}(\mathbf{I}) = 2s$. From Theorem 4 and Theorem 3, we see that the SDA estimator is able to recover the true support set T using the optimal number of samples (up to an absolute constant) over the parameter space

$$\Theta(\boldsymbol{\Sigma}, \tau_{\min}, s) \cap \{\boldsymbol{\theta} : ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \leq M\}$$

where M is a fixed constant and $\Lambda_{\min}(\boldsymbol{\Sigma}_{TT})$ is bounded from below. This result will be further illustrated by numerical simulations in §VI.

IV. Exhaustive Search Decoder

In this section, we analyze an exhaustive search procedure, which evaluates every subset T' of size s and outputs the one with the best score. Even though the procedure cannot be implemented in practice, it is a useful benchmark to compare against and it provides deeper theoretical insights into the problem.

For any subset $T' \subset [p]$, we define

$$f(T') = \min_{\mathbf{u}\in\mathbb{R}^{|T'|}} \left\{ \mathbf{u}'\widehat{\mathbf{S}}_{T'T'}\mathbf{u} \ : \ \mathbf{u}'\widehat{\boldsymbol{\mu}}_{T'} = 1 \right\} = \min_{T'\subset[p] \ : \ |T'|=s} \frac{1}{\widehat{\boldsymbol{\mu}}'_{T'}\mathbf{S}^{-1}_{T'T'}\widehat{\boldsymbol{\mu}}'_{T'}}.$$

The exhaustive search procedure outputs the support set $\widehat{T}$ that minimizes $f(T')$ over all subsets T' of size s,

$$\widehat{T} = \operatorname*{argmin}_{T'\subset[p] \ : \ |T'|=s} f(T') = \operatorname*{argmax}_{T'\subset[p] \ : \ |T'|=s} \widehat{\boldsymbol{\mu}}'_{T'}\mathbf{S}^{-1}_{T'T'}\widehat{\boldsymbol{\mu}}_{T'}.$$

Define $g(T') = \widehat{\boldsymbol{\mu}}'_{T'}\mathbf{S}^{-1}_{T'T'}\widehat{\boldsymbol{\mu}}_{T'}$. In order to show that the exhaustive search procedure identifies the correct support set T, we need to show that with high probability $g(T) > g(T')$ for any other set T' of size s. The next result gives sufficient conditions for this to happen. We first introduce some additional notation. Let $A_1 = T \cap T'$, $A_2 = T\backslash T'$ and $A_3 = T'\backslash T$. We define the following quantities

$$a_1(T') = \boldsymbol{\mu}'_{A_1}\boldsymbol{\Sigma}^{-1}_{A_1A_1}\boldsymbol{\mu}_{A_1},$$
$$a_2(T') = \boldsymbol{\mu}'_{A_2|A_1}\boldsymbol{\Sigma}^{-1}_{A_2A_2|A_1}\boldsymbol{\mu}_{A_2|A_1},$$
$$a_3(T') = \boldsymbol{\mu}'_{A_3|A_1}\boldsymbol{\Sigma}^{-1}_{A_3A_3|A_1}\boldsymbol{\mu}_{A_3|A_1},$$

where $\boldsymbol{\mu}_{A_2|A_1} = \boldsymbol{\mu}_{A_2} - \boldsymbol{\Sigma}_{A_2A_1}\boldsymbol{\Sigma}^{-1}_{A_1A_1}\boldsymbol{\mu}_{A_1}$ and $\boldsymbol{\Sigma}_{A_2A_2|A_1} = \boldsymbol{\Sigma}_{A_2A_2} - \boldsymbol{\Sigma}_{A_2A_1}\boldsymbol{\Sigma}^{-1}_{A_1A_1}\boldsymbol{\Sigma}_{A_2A_1}$. The quantities $\boldsymbol{\mu}_{A_3|A_1}$ and $\boldsymbol{\Sigma}_{A_3A_3|A_1}$ are defined similarly.

Theorem 5: Assuming that for all $T' \subseteq [p]$ with $|T'| = s$ and $T' \neq T$ the following holds

$$a_2(T') - \left(1 + C_1\sqrt{\Gamma_{n,p,s,k}}\right) a_3(T') \geq C_2\sqrt{(1 \vee a_1(T'))\, a_2(T')\Gamma_{n,p,s,k}} + C_3\left(1 \vee a_1(T')\right)\Gamma_{n,p,s,k}, \tag{IV.1}$$

where $|T' \cap T| = k$, $\Gamma_{n,p,s,k} = n^{-1}\log\left(\binom{p-s}{s-k}\binom{s}{k}s\log(n)\right)$ and C_1, C_2, C_3 are constants independent of the problem parameters, we have $\mathbb{P}[\widehat{T} \neq T] = \mathcal{O}(\log^{-1}(n))$.

The condition in (IV.1) allows the exhaustive search decoder to distinguish between the sets T and T' with high probability. Note that the Mahalanobis distance decomposes as $g(T) = \widehat{\boldsymbol{\mu}}'_{A_1}\mathbf{S}^{-1}_{A_1A_1}\widehat{\boldsymbol{\mu}}_{A_1} + \widetilde{\boldsymbol{\mu}}'_{A_2|A_1}\mathbf{S}^{-1}_{A_2A_2|A_1}\widetilde{\boldsymbol{\mu}}_{A_2|A_1}$ where $\widetilde{\boldsymbol{\mu}}_{A_2|A_1} = \widehat{\boldsymbol{\mu}}_{A_2} - \mathbf{S}_{A_2A_1}\mathbf{S}^{-1}_{A_1A_1}\widehat{\boldsymbol{\mu}}_{A_1}$ and $\mathbf{S}_{A_2A_2|A_1} = \mathbf{S}_{A_2A_2} - \mathbf{S}_{A_2A_1}\mathbf{S}^{-1}_{A_1A_1}\mathbf{S}_{A_1A_2}$, and similarly $g(T') = \widehat{\boldsymbol{\mu}}'_{A_1}\mathbf{S}^{-1}_{A_1A_1}\widehat{\boldsymbol{\mu}}_{A_1} + \widetilde{\boldsymbol{\mu}}'_{A_3|A_1}\mathbf{S}^{-1}_{A_3A_3|A_1}\widetilde{\boldsymbol{\mu}}_{A_3|A_1}$. Therefore $g(T) > g(T')$ if $\widetilde{\boldsymbol{\mu}}'_{A_2|A_1}\mathbf{S}^{-1}_{A_2A_2|A_1}\widetilde{\boldsymbol{\mu}}_{A_2|A_1} > \widetilde{\boldsymbol{\mu}}'_{A_3|A_1}\mathbf{S}^{-1}_{A_3A_3|A_1}\widetilde{\boldsymbol{\mu}}_{A_3|A_1}$. With infinite amount of data, it would be sufficient that $a_2(T') > a_3(T')$. However, in the finite-sample setting, condition (IV.1) ensures that the separation is big enough. If $\mathbf{X}_T$ and $\mathbf{X}_N$ are independent, then the expression (IV.1) can be simplified by dropping the second term on the left hand side.

Compared to the result of Theorem 3, the exhaustive search procedure does not require the covariance matrix to satisfy the irrepresentable condition given in (II.3). The SDA estimator defined in (I.3) uses the ℓ_1 penalty to find a sparse $\widehat{\mathbf{v}}$. In place of the ℓ_1 penalty one could use the SCAD [9] or the MCP penalty [33]. These nonconvex penalties interpolate between the ℓ_1 and ℓ_0 penalties [18] and do not require as strong assumptions on the covariance matrix $\boldsymbol{\Sigma}$ as the ℓ_1 penalty. However, finding the global solution of a resulting nonconvex objective is challenging. In the regression setting, these penalties allow one to establish variable selection consistency result without the need of irrepresentable condition. However, most of these results are only shown for the global minimizer (see [34] for a recent overview) and it is not clear how this global minimizer can be obtained using polynomial-time algorithm. Alternatively, one can study a particular algorithm and the local solution obtained by this algorithm, however, the analysis of these algorithm crucially depend on the correctness of the underlying regression model (see [27], [29]). Such a regression setting is fundamentally different from the discriminant analysis model we are studying in this paper. The study of nonconvex penalty on SDA estimator is beyond the scope of this paper.

V. Implications of Our Results

In this section, we give some implications of our results. We start with the case when the covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}$. The same implications hold for other covariance matrices that satisfy $\Lambda_{\min}(\boldsymbol{\Sigma}) \geq C > 0$ for some constant C independent of (n, p, s). We first illustrate a regime where the SDA estimator is optimal for the problem of identifying the relevant variables. This is done by comparing the results in Theorem 3 to those of Theorem 4. Next, we point out a regime where there exists a gap between the sufficient and necessary conditions of Theorem 4 for both the exhaustive search decoder and the SDA estimator. Throughout the section, we assume that $s = o(\min(n, p))$.

When $\boldsymbol{\Sigma} = \mathbf{I}$, we have that $\boldsymbol{\beta}_T = \boldsymbol{\mu}_T$. Let $\mu = \min_{a\in T}|\mu_a|$. If

$$\mu \lesssim \sqrt{\frac{\log(p-s)}{n}},$$

then no procedure can reliably recover the support, according to Theorem 4. We will compare this bound with sufficient conditions given in Theorems 3 and 5.

First, we assume that $||\boldsymbol{\mu}_T||_2^2 = C$ for some constant C. If $n \gtrsim s\log(p-s)$, then $\mu \gtrsim \sqrt{\frac{\log(p-s)}{n}}$ is sufficient for the SDA estimator to consistently recover the relevant variables, using Theorem 3. Therefore, in this regime, the sufficient conditions for the SDA estimator to reliably recover the support match the necessary condition.

Next, we investigate the condition in (IV.1), which is sufficient for the exhaustive search procedure to identify the set T. Let $T' \subset [p]$ be a subset of size s. Then, using the notation of Section IV,

$$a_1(T') = ||\boldsymbol{\mu}_{A_1}||_2^2, \ a_2(T') = ||\boldsymbol{\mu}_{A_2}||_2^2, \text{and } a_3(T') = 0.$$

Now, if $|T' \cap T| = s-1$ and T' does not contain the smallest component of $\boldsymbol{\mu}_T$, (IV.1) simplifies to $\mu \gtrsim \sqrt{\frac{\log(p-s)}{n}}$, since $||\boldsymbol{\mu}_{A_1}||_2^2 \leq ||\boldsymbol{\mu}_T||_2^2 = C$. This shows that both the SDA estimator and the exhaustive search procedure can reliably detect signals at the information theoretic limit in the case when the norm of the vector $\boldsymbol{\mu}_T$ is bounded and $\mu \gtrsim s^{-1/2}$. However, when the norm of the vector $\boldsymbol{\mu}_T$ is not bounded by a constant, for example, $\mu = C'$ for some constant C', Theorem 4 gives that at least $n \gtrsim \log(p-s)$ data points are needed, while $n \gtrsim s\log(p-s)$ is sufficient for correct recovery of the support set T. This situation is analogous to the known bounds on the support recovery in the sparse linear regression setting [25].

Next, we show that the largest eigenvalue of a covariance matrix $\boldsymbol{\Sigma}$ can diverge, without affecting the sample size required for successful recovery of the support set T. Let $\boldsymbol{\Sigma} = (1-\gamma)\mathbf{I}_p + \gamma \mathbf{1}_p\mathbf{1}_p'$ for $\gamma \in [0,1)$. We have $\Lambda_{\max}(\boldsymbol{\Sigma}) = 1 + (p-1)\gamma$, which diverges to infinity for any fixed γ as $p \to \infty$. Let $T = [s]$ and set $\boldsymbol{\beta}_T = \beta \mathbf{1}_T$. This gives $\boldsymbol{\mu}_T = \beta(1+\gamma(s-1))\mathbf{1}_T$ and $\boldsymbol{\mu}_N = \gamma\beta s \mathbf{1}_N$. A simple application of the matrix inversion formula gives

$$\boldsymbol{\Sigma}_{TT}^{-1} = (1-\gamma)^{-1}\mathbf{I}_s - \frac{\gamma}{(1-\gamma)(1+\gamma(s-1))}\mathbf{1}_T\mathbf{1}_T'.$$

A lower bound on β is obtained from Theorem 4 as $\beta \geq \sqrt{\frac{2}{1-\gamma}\frac{\log(p-s)}{n}}$. This follows from a simple calculation that establishes $\varphi_{\text{close}}(\boldsymbol{\Sigma}) = 2(1-\gamma)$ and $\varphi_{\text{far}}(\boldsymbol{\Sigma}) = 2s(1-\gamma) + (2s)^2\gamma$.

Sufficient conditions for the SDA estimator follow from Theorem 3. A straightforward calculation shows that

$$\begin{aligned} \sigma_{a|T} &= \frac{(1-\gamma)(1+\gamma s)}{1+\gamma(s-1)}, \\ \Lambda_{\min}(\boldsymbol{\Sigma}) &= 1-\gamma, \\ \text{and } ||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty &= \frac{1}{1+\gamma(s-1)}. \end{aligned}$$

This gives that $\beta \geq K\sqrt{\frac{\log(p-s)}{(1-\gamma)n}}$ (for K large enough) is sufficient for recovering the set T, assuming that $||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2 = \mathcal{O}(1)$. This matches the lower bound, showing that the maximum eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ does not play a role in characterizing the behavior of the SDA estimator.

VI. Simulation Results

In this section, we conduct several simulations to illustrate the finite-sample performance of our results. Theorem 3 describes the sample size needed for the SDA estimator to recover the set of relevant variables. We consider the following three scalings for the size of the set T:

1) fractional power sparsity, where $s = \lceil 2p^{0.45} \rceil$
2) sublinear sparsity, where $s = \lceil 0.4p/\log(0.4p) \rceil$, and
3) linear sparsity, where $s = \lceil 0.4p \rceil$.

For all three scaling regimes, we set the sample size as

$$n = \theta s \log(p)$$

where θ is a control parameter that is varied. We investigate how well can the SDA estimator recovers the true support set T as the control parameter θ varies.

We set $\mathbb{P}[Y=1] = \mathbb{P}[Y=2] = \frac{1}{2}$, $\mathbf{X}|Y=1 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and without loss of generality $\mathbf{X}|Y=2 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We specify the vector $\boldsymbol{\mu}$ by choosing the set T of size $|T| = s$ randomly, and for each $a \in T$ setting μ_a equal to $+1$ or -1 with equal probability, and $\mu_a = 0$ for all components $a \notin T$. We specify the covariance matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{TT} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-s} \end{pmatrix}$$

so that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = (\boldsymbol{\beta}_T', \mathbf{0}')'$. We consider three cases for the block component $\boldsymbol{\Sigma}_{TT}$:

1) identity matrix, where $\boldsymbol{\Sigma}_{TT} = \mathbf{I}_s$,
2) Toeplitz matrix, where $\boldsymbol{\Sigma}_{TT} = [\Sigma_{ab}]_{a,b\in T}$ and $\Sigma_{ab} = \rho^{|a-b|}$ with $\rho = 0.1$, and
3) equal correlation matrix, where $\Sigma_{ab} = \rho$ when $a \neq b$ and $\sigma_{aa} = 1$.

Finally, we set the penalty parameter $\lambda = \lambda_{\text{SDA}}$ as

$$\begin{aligned} \lambda_{\text{SDA}} = 0.3 \times \left(1 + ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2/4\right)^{-1} \\ \times \sqrt{\left(1 \vee ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2\right)\frac{\log(p-s)}{n}} \end{aligned}$$

for all cases. We also tried several different constants and found that our main results on high dimensional scalings are insensitive to the choice of this constant. For this choice of λ, Theorem 3 predicts that the set T will be recovered correctly. For each setting, we report the Hamming distance between the estimated set $\widehat{T}$ and the true set T,

$$h(\widehat{T}, T) = |(\widehat{T} \backslash T) \cup (T \backslash \widehat{T})|,$$

averaged over 200 independent simulation runs.

Figure 1 plots the Hamming distance against the control parameter θ, or the rescaled number of samples. Here the Hamming distance between $\widehat{T}$ and T is calculated by averaging 200 independent simulation runs. There are three subfigures corresponding to different sparsity regimes (fractional power, sublinear and linear sparsity), each of them containing three curves for different problem sizes $p \in \{100, 200, 300\}$. Vertical line indicates a threshold parameter θ at which the set T is correctly recovered. If the parameter is smaller than the threshold value, the recovery is poor. Figure 2 and Figure 3 show results for two other cases, with $\boldsymbol{\Sigma}_{TT}$ being a Toeplitz matrix with parameter $\rho = 0.1$ and the equal correlation matrix with $\rho = 0.1$. To illustrate the effect of correlation, we set $p = 100$ and generate the equal correlation matrices with $\rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$. Results are given in Figure 4.

VII. Discussion

In this paper, we address the problem of variable selection in high-dimensional discriminant analysis problem. The problem of reliable variable selection is important in many scientific areas where simple models are needed to provide insights into complex systems. Existing research has focused primarily on establishing results for prediction consistency, ignoring feature selection. We bridge this gap, by analyzing the variable selection performance of the SDA estimator and an exhaustive search decoder. We establish sufficient conditions required for

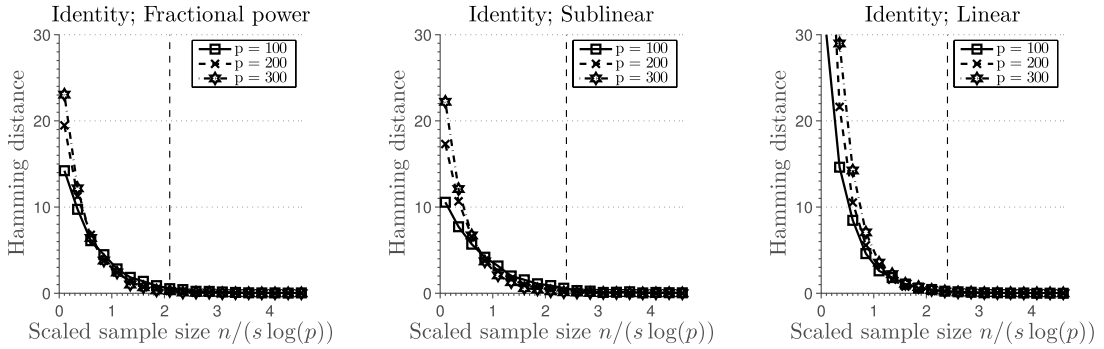

Fig. 1. (The SDA Estimator) Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\widehat{T}$ and T for identity covariance matrix $\mathbf{\Sigma} = \mathbf{I}_p$ (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes $p \in \{100, 200, 300\}$. The first subfigure corresponds to the fractional power sparsity regime, $s = 2p^{0.45}$, the second subfigure corresponds to the sublinear sparsity regime $s = 0.4p/\log(0.4p)$, and the third ssubfigure corresponds to the linear sparsity regime $s = 0.4p$. Vertical lines denote a scaled sample size at which the support set T is recovered correctly.

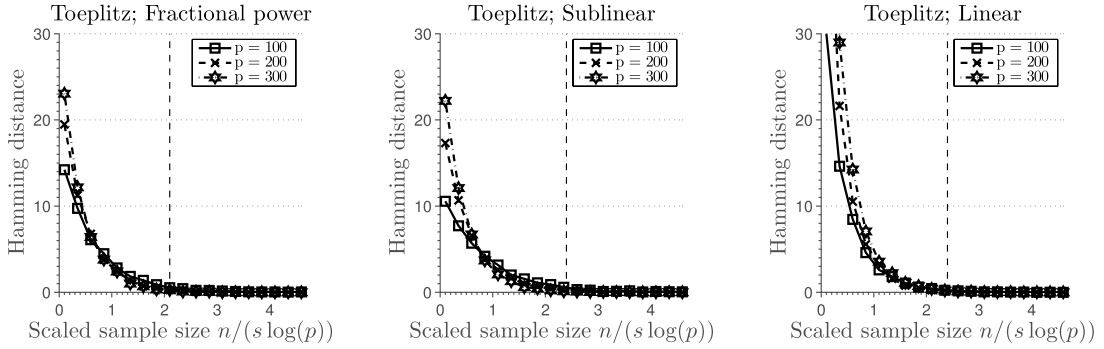

Fig. 2. (The SDA Estimator) Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\widehat{T}$ and T for the Toeplitz covariance matrix $\mathbf{\Sigma}_{TT}$ with $\rho = 0.1$ (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes $p \in \{100, 200, 300\}$. The first subfigure corresponds to the fractional power sparsity regime, $s = 2p^{0.45}$, the second subfigure corresponds to the sublinear sparsity regime $s = 0.4p/\log(0.4p)$, and the third subfiguren corresponds to the linear sparsity regime $s = 0.4p$. Vertical lines denote a scaled sample size at which the support set T is recovered correctly.

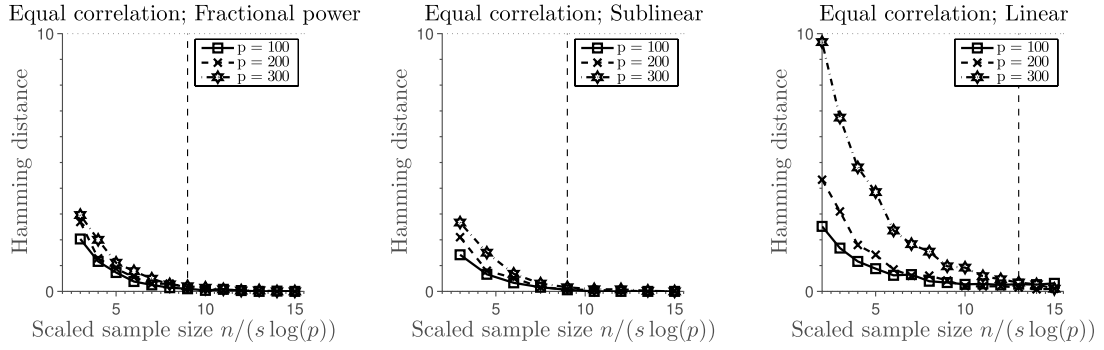

Fig. 3. (The SDA Estimator) Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\widehat{T}$ and T for equal correlation matrix $\mathbf{\Sigma}_{TT}$ with $\rho = 0.1$ (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes $p \in \{100, 200, 300\}$. The first subfigure corresponds to the fractional power sparsity regime, $s = 2p^{0.45}$, the second subfigure corresponds to the sublinear sparsity regime $s = 0.4p/\log(0.4p)$, and the third subfigure corresponds to the linear sparsity regime $s = 0.4p$. Vertical lines denote a scaled sample size at which the support set T is recovered correctly.

successful recovery of the set of relevant variables for these procedures. This analysis is complemented by analyzing the information theoretic limits, which provide necessary conditions for variable selection in discriminant analysis. From these results, we are able to identify the class of problems for which the computationally tractable procedures are optimal. In this section, we discuss some implications and possible extensions of our results.

A. Theoretical Justification of the ROAD and Sparse Optimal Scaling Estimators

In a recent work, [15] shows that the SDA estimator is numerically equivalent to the ROAD estimator proposed by [10] and [32] and the sparse optimal scoring estimator proposed by [6]. More specifically, all these three methods have the same regularization paths up to a constant scaling. This result allows us to apply the theoretical results in this

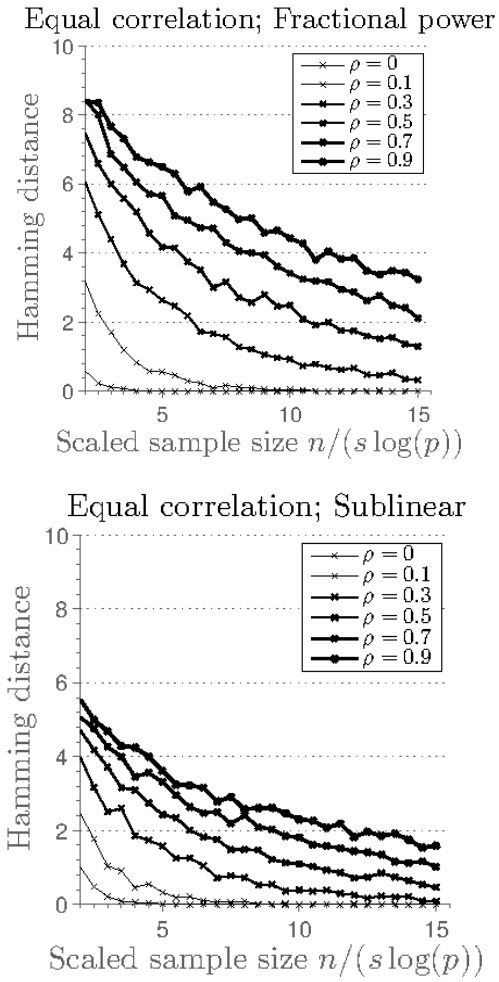

Fig. 4. (The SDA Estimator) Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\widehat{T}$ and T for equal correlation matrix $\mathbf{\Sigma}_{TT}$ with $\rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ (averaged over 200 simulation runs). The ambient dimension is set as $p = 100$. The first subfigure corresponds to the fractional power sparsity regime, $s = 2p^{0.45}$ and the second subfigure corresponds to the sublinear sparsity regime $s = 0.4p/\log(0.4p)$.

paper to simultaneously justify the optimal variable selection performance of the ROAD and sparse optimal scaling estimators. The focus of [10] was on establishing a bound on the misclassification error of the ROAD estimator. Our results provide complimentary insights for the ROAD estimator. From [2, Th. 2], it follows that the ROAD estimator selects the true support consistently under a stringent condition on $\beta_{\min}$, which requires (II.10) to hold. Therefore, our analysis improves the previous result and shows that the ROAD estimator needs only $\beta_{\min} \asymp \sqrt{\log(p-s)/n}$ to hold in order for the true support to be consistently identified.

B. Risk Consistency

The results of Theorem 3 can be used to establish risk consistency of the SDA estimator. Consider the following classification rule

$$\widehat{y}(\mathbf{x}) = \begin{cases} 1 & \text{if } g(\mathbf{x}; \widehat{\mathbf{v}}) = 1 \\ 2 & \text{otherwise} \end{cases}$$

where $g(\mathbf{x}; \widehat{\mathbf{v}}) = I\left[\widehat{\mathbf{v}}'(\mathbf{x} - (\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2)/2) > 0\right]$ with $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}^{\text{SDA}}$. Under the assumption that $\boldsymbol{\beta} = (\boldsymbol{\beta}_T{}', \mathbf{0}')'$, the risk (or the error rate) of the Bayes rule defined in (I.1) is $R_{\text{opt}} = \Phi\left(-\sqrt{\boldsymbol{\mu}_T' \mathbf{\Sigma}_{TT}^{-1} \boldsymbol{\mu}_T}/2\right)$, where Φ is the cumulative distribution function of a standard Normal distribution. We will compare the risk of the SDA estimator against this Bayes risk.

Recall the setting introduced in §I-A, conditioning on the data points $\{\mathbf{x}_i, y_i\}_{i\in[n]}$, the conditional error rate is

$$R(\widehat{\mathbf{w}}) = \frac{1}{2}\sum_{i\in\{1,2\}} \Phi\left(\frac{-\widehat{\mathbf{v}}'(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i) - \widehat{\mathbf{v}}'\widehat{\boldsymbol{\mu}}/2}{\sqrt{\widehat{\mathbf{v}}'\mathbf{\Sigma}\widehat{\mathbf{v}}}}\right).$$

Let $r_n = \lambda||\boldsymbol{\beta}_T||_1$ and $q_n = \text{sign}(\boldsymbol{\beta}_T)'\mathbf{\Sigma}_{TT}\,\text{sign}(\boldsymbol{\beta}_T)$. We have the following result on risk consistency.

Corollary 1: Let $\widehat{\mathbf{v}} = \widehat{\mathbf{v}}^{\text{SDA}}$. We assume that the conditions of Theorem 3 hold with

$$n \asymp K(n)\left(\max_{a\in N}\sigma_{a|T}\right)\Lambda_{\min}^{-1}(\mathbf{\Sigma}_{TT}) \times s\log\left((p-s)\log(n)\right),$$

where $K(n)$ could potentially scale with n, and $||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}} \geq C > 0$. Furthermore, we assume that $r_n \xrightarrow{n\to\infty} 0$. Then

$$R(\widehat{\mathbf{w}}) = \Phi\left(-\frac{||\boldsymbol{\beta}_T||_{\mathbf{\Sigma}_{TT}}\left(1 + \mathcal{O}_P\left(r_n\right)\right)}{2\sqrt{1 + \mathcal{O}_P\left(r_n \vee \frac{\lambda_0^2 q_n}{||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}}\right)}}\right).$$

First, note that

$$||\boldsymbol{\beta}_T||_1/||\boldsymbol{\beta}_T||_{\mathbf{\Sigma}_{TT}} = o\left(\sqrt{K(n)s/\Lambda_{\min}(\mathbf{\Sigma}_{TT})}\right)$$

is sufficient for $r_n \to 0$ as $n \to \infty$. Under the conditions of Theorem 1, we have that $||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}/\left(\lambda_0^2 q_n\right) = \mathcal{O}\left(K(n)s/\left(\Lambda_{\min}(\mathbf{\Sigma}_{TT})q_n\right)\right) = \mathcal{O}\left(K(n)\right)$. Therefore, if $K(n) \xrightarrow{n\to\infty} \infty$ and $K(n) \geq Cs||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}/\left(\Lambda_{\min}(\mathbf{\Sigma}_{TT})||\boldsymbol{\beta}_T||_1^2\right)$ we have

$$R(\widehat{\mathbf{w}}) = \Phi\left(-\frac{||\boldsymbol{\beta}_T||_{\mathbf{\Sigma}_{TT}}}{2}\left(1 + \mathcal{O}_P\left(r_n\right)\right)\right)$$

and $R(\widehat{\mathbf{w}}) - R_{\text{opt}} \to_P 0$. If in addition

$$||\boldsymbol{\beta}_T||_{\mathbf{\Sigma}_{TT}}||\boldsymbol{\beta}_T||_1 = o\left(\sqrt{K(n)s/\Lambda_{\min}(\mathbf{\Sigma}_{TT})}\right),$$

then $R(\widehat{\mathbf{w}})/R_{\text{opt}} \to_P 1$, using [22, Lemma 1].

The above discussion shows that the conditions of Theorem 3 are sufficient for establishing risk consistency. We conjecture that substantially less restrictive conditions are needed to establish risk consistency results. Exploring such weaker conditions is beyond the scope of this paper.

C. Approximate Sparsity

Thus far, we were discussing estimation of discriminant directions that are exactly sparse. However, in many applications it may be the case that the discriminant direction $\boldsymbol{\beta} = (\boldsymbol{\beta}_T{}', \boldsymbol{\beta}_N')' = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ is only approximately sparse, that is, $\boldsymbol{\beta}_N$ is not equal to zero, but is small. In this section, we briefly discuss the issue of variable selection in this context.

In the approximately sparse setting, since $\boldsymbol{\beta}_N \neq \mathbf{0}$, a simple calculation gives

$$\boldsymbol{\beta}_T = \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T - \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{TN}\boldsymbol{\beta}_N \tag{VII.1}$$

and

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T + \left(\boldsymbol{\Sigma}_{NN} - \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{TN}\right)\boldsymbol{\beta}_N. \tag{VII.2}$$

In what follows, we provide conditions under which the solution to the population version of the SDA estimator, given in (II.1), correctly recovers the support of large entries T. Let $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_T', \mathbf{0}')'$ where $\widehat{\mathbf{w}}_T$ is given as

$$\widehat{\mathbf{w}}_T = \pi_1\pi_2\frac{1+\lambda||\widetilde{\boldsymbol{\beta}_T}||_1}{1+\pi_1\pi_2||\widetilde{\boldsymbol{\beta}_T}||_{\boldsymbol{\Sigma}_{TT}}^2}\widetilde{\boldsymbol{\beta}_T} - \lambda\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\widetilde{\boldsymbol{\beta}_T})$$

with $\widetilde{\boldsymbol{\beta}_T} = \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T$. We will show that $\widehat{\mathbf{w}}$ is the solution to (II.1).

We again define $\beta_{\min} = \min_{a\in T}|\beta_a|$. Following a similar argument as the proof of Theorem 1, we have that $\operatorname{sign}(\widehat{\mathbf{w}}_T) = \operatorname{sign}\left(\widetilde{\boldsymbol{\beta}_T}\right)$ holds if $\widetilde{\boldsymbol{\beta}_T}$ satisfies

$$\pi_1\pi_2\frac{1+\lambda||\widetilde{\boldsymbol{\beta}_T}||_1}{1+\pi_1\pi_2||\widetilde{\boldsymbol{\beta}_T}||_{\boldsymbol{\Sigma}_{TT}}^2}\beta_{\min} > \lambda||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}\left(\widetilde{\boldsymbol{\beta}_T}\right)||_\infty. \tag{VII.3}$$

In the approximate sparsity setting, it is reasonable to assume that $\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{TN}\boldsymbol{\beta}_N$ is small compared to $\widetilde{\boldsymbol{\beta}_T}$, which would imply that $\operatorname{sign}(\boldsymbol{\beta}_T) = \operatorname{sign}\left(\widetilde{\boldsymbol{\beta}_T}\right)$ using (VII.1). Therefore, under suitable assumptions we have $\operatorname{sign}(\widehat{\mathbf{w}}_T) = \operatorname{sign}(\boldsymbol{\beta}_T)$. Next, we need conditions under which $\widehat{\mathbf{w}}$ is the solution to (II.1).

Following a similar analysis as in Lemma 1, the optimality condition

$$||\left(\boldsymbol{\Sigma}_{NT} + \pi_1\pi_2\boldsymbol{\mu}_N\boldsymbol{\mu}_T'\right)\widehat{\mathbf{w}}_T - \pi_1\pi_2\boldsymbol{\mu}_N||_\infty \leq \lambda$$

needs to hold. Let $\widehat{\gamma} = \boldsymbol{\Delta}_{\boldsymbol{\Sigma}}\frac{1+\lambda||\widetilde{\boldsymbol{\beta}_T}||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}$. Using (VII.2), the above display becomes

$$\begin{aligned}\|&-\lambda\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)\\ &-\pi_1\pi_2\widehat{\gamma}\left(\boldsymbol{\Sigma}_{NN} - \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{TN}\right)\boldsymbol{\beta}_N\|_\infty < \lambda.\end{aligned}$$

Therefore, using the triangle inequality, the following assumption

$$\pi_1\pi_2\widehat{\gamma}\cdot||\left(\boldsymbol{\Sigma}_{NN} - \boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{TN}\right)\boldsymbol{\beta}_N||_\infty < \alpha\lambda,$$

in addition to (II.3) and (VII.3), is sufficient for $\widehat{\mathbf{w}}$ to recover the set of important variables T.

The above discussion could be made more precise and extended to the sample SDA estimator in (I.3), by following the proof of Theorem 3. This is beyond the scope of the current paper and will be left as a future investigation.

APPENDIX A
PROOFS OF MAIN RESULTS

In this section, we collect proofs of results given in the main text. We will use $C, C_1, C_2, \ldots$ to denote generic constants that do not depend on problem parameters. Their values may change from line to line.

Let

$$\begin{aligned}\mathcal{A} = \mathcal{E}_n \cap \mathcal{E}_1(\log^{-1}(n)) \cap \mathcal{E}_2(\log^{-1}(n))\\ \cap \mathcal{E}_3(\log^{-1}(n)) \cap \mathcal{E}_4(\log^{-1}(n)),\end{aligned} \tag{A.1}$$

where $\mathcal{E}_n$ is defined in (E.1), $\mathcal{E}_1$ in Lemma 4, $\mathcal{E}_2$ in Lemma 5, $\mathcal{E}_3$ in (E.2), and $\mathcal{E}_4$ in (E.3). We have that $\mathbb{P}[\mathcal{A}] \geq 1 - \mathcal{O}\left(\log^{-1}(n)\right)$.

A. Proofs of Results in Section II

Proof of Lemma 1: From the KKT conditions given in (II.6), we have that $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_T', \mathbf{0}')'$ is a solution to the problem in (II.1) if and only if

$$\begin{aligned}&\left(\boldsymbol{\Sigma}_{TT} + \pi_1\pi_2\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\right)\widehat{\mathbf{w}}_T - \pi_1\pi_2\boldsymbol{\mu}_T + \lambda\operatorname{sign}(\widehat{\mathbf{w}}_T) = \mathbf{0}\\ &||\left(\boldsymbol{\Sigma}_{NT} + \pi_1\pi_2\boldsymbol{\mu}_N\boldsymbol{\mu}_T'\right)\widehat{\mathbf{w}}_T - \pi_1\pi_2\boldsymbol{\mu}_N||_\infty \leq \lambda\end{aligned}$$

By construction, $\widehat{\mathbf{w}}_T$ satisfy the first equation. Therefore, we need to show that the second one is also satisfied. Plugging in the explicit form of $\widehat{\mathbf{w}}_T$ into the second equation and using (II.2), after some algebra we obtain that

$$||\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty \leq 1$$

needs to be satisfied. The above display is satisfied with strict inequality under the assumption in (II.3). □

Proof of Lemma 2: Throughout the proof, we will work on the event $\mathcal{A}$ defined in (A.1).

Let $a \in T$ be such that $\widetilde{v}_a > 0$, noting that the case when $\widetilde{v}_a < 0$ can be handled in a similar way. Let

$$\begin{aligned}\delta_1 &= \widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) - \boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T),\\ \delta_2 &= \mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T,\\ \delta_3 &= \mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) - \mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T),\\ \delta_4 &= \widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2,\\ \text{and } \delta_5 &= \frac{n_1n_2}{n(n-2)} - \pi_1\pi_2.\end{aligned}$$

Furthermore, let

$$\widehat{\gamma} = \frac{n_1n_2}{n(n-2)}\frac{1+\lambda\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{1+\frac{n_1n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}$$

and

$$\gamma = \frac{\pi_1\pi_2(1+\lambda||\boldsymbol{\beta}_T||_1)}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}.$$

For sufficiently large n, on the event $\mathcal{A}$, together with Lemma 6, Lemma 10, and Lemma 7, we have that $\widehat{\gamma} \geq \gamma(1-o(1)) > \gamma/2$ and $\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) = \mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)(1+o(1)) \leq \frac{3}{2}\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)$ with probability at least $1-\mathcal{O}\left(\log^{-1}(n)\right)$. Then

$$\begin{aligned}\widetilde{v}_a &\geq \frac{\gamma}{2}(\beta_a+\delta_2) - \frac{3}{2}\lambda\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)\\ &\geq \frac{\pi_1\pi_2(1+\lambda||\boldsymbol{\beta}_T||_1)(\beta_a - |\delta_2|) - 3\lambda_0||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty}{2(1+\pi_1\pi_2||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2)},\end{aligned}$$

so that $\operatorname{sign}(\widetilde{v}_a) = \operatorname{sign}(\beta_a)$ if

$$\pi_1\pi_2(1+\lambda||\boldsymbol{\beta}_T||_1)(\beta_a - |\delta_2|) - 3\lambda_0||\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)||_\infty > 0. \tag{A.2}$$

Lemma 9 gives a bound on $|\delta_2|$, for each fixed $a \in T$, as

$$|\delta_2| \leq C_1\sqrt{\left(\mathbf{\Sigma}_{TT}^{-1}\right)_{aa}\left(1 \vee ||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}\right)\frac{\log(s\log(n))}{n}} + C_2|\beta_a|\sqrt{\frac{\log(s\log(n))}{n}}.$$

Therefore assumption (II.8), with K_β sufficiently large, and a union bound over all $a \in T$ implies (A.2).

Lemma 9 gives $\operatorname{sign}(\widehat{\boldsymbol{\beta}_T}) = \operatorname{sign}(\boldsymbol{\beta}_T)$ with probability $1 - \mathcal{O}(\log^{-1}(n))$. □

Proof of Lemma 3: Throughout the proof, we will work on the event $\mathcal{A}$ defined in (A.1). Using Lemma 2, we have that $\widetilde{\mathbf{v}}_T$ defined in (II.15) satisfies $\widetilde{\mathbf{v}}_T = \widehat{\mathbf{v}}_T$. Therefore, by construction of the oracle optimization problem, the vector $\widehat{\mathbf{v}} = (\widetilde{\mathbf{v}}'_T, \mathbf{0}')'$ satisfies the condition in (II.12).

Therefore, to show that it is a solution to (I.3), we need to show that it also satisfies (II.13). To simplify notation, let

$$\mathbf{C} = \mathbf{S} + \frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}',$$

$$\widehat{\gamma} = \frac{n_1 n_2}{n(n-2)}\frac{1 + \lambda||\widehat{\boldsymbol{\beta}_T}||_1}{1 + \frac{n_1 n_2}{n(n-2)}||\widehat{\boldsymbol{\beta}_T}||^2_{\mathbf{S}_{TT}}},$$

and

$$\gamma = \frac{\pi_1\pi_2(1 + \lambda||\boldsymbol{\beta}_T||_1)}{1 + \pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}}.$$

Recall that $\widetilde{\mathbf{v}}_T = \widehat{\gamma}\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \lambda\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)$ and (II.13) can be written as

$$||\mathbf{C}_{NT}\widetilde{\mathbf{v}}_T - \frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}_N||_\infty \leq \lambda.$$

Let $\mathbf{U} \in \mathbb{R}^{(n-2)\times p}$ be a matrix with each row $\mathbf{u}_i \overset{iid}{\sim} \mathcal{N}(0, \mathbf{\Sigma})$ such that $(n-2)\mathbf{S} = \mathbf{U}'\mathbf{U}$. For $a \in N$, we have

$$\begin{aligned}(n-2)\mathbf{S}_{aT} &= (\mathbf{U}_T\mathbf{\Sigma}_{TT}^{-1}\mathbf{\Sigma}_{Ta} + \mathbf{U}_{a\cdot T})'\mathbf{U}_T \\ &= \mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\mathbf{U}'_T\mathbf{U}_T + \mathbf{U}'_{a\cdot T}\mathbf{U}_T\end{aligned}$$

where $\mathbf{U}_{a\cdot T} \sim \mathcal{N}\left(0, \frac{n_1 n_2}{n(n-2)}\sigma_{a|T}\mathbf{I}_{n-2}\right)$ is independent of $\mathbf{U}_T$, and

$$\widehat{\mu}_a = \mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T + \widehat{\mu}_{a\cdot T}$$

where $\widehat{\mu}_{a\cdot T} \sim \mathcal{N}\left(0, \frac{n}{n_1 n_2}\sigma_{a|T}\right)$ is independent of $\widehat{\boldsymbol{\mu}}_T$. Therefore,

$$\begin{aligned}\mathbf{C}_{aT} &= \mathbf{S}_{aT} + \frac{n_1 n_2}{n(n-2)}\widehat{\mu}_a\widehat{\boldsymbol{\mu}}'_T \\ &= \mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\mathbf{S}_{TT} + (n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T \\ &\quad + \frac{n_1 n_2}{n(n-2)}\widehat{\mu}_a\widehat{\boldsymbol{\mu}}'_T, \\ \mathbf{C}_{aT}\widetilde{\mathbf{v}}_T &= \widehat{\gamma}\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \\ &\quad + \widehat{\gamma}\frac{n_1 n_2}{n(n-2)}\left(\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right)\widehat{\mu}_a \\ &\quad - \lambda\left(\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) + \frac{n_1 n_2}{n(n-2)}||\widehat{\boldsymbol{\beta}}_T||_1 \cdot \widehat{\mu}_a\right) \\ &\quad + (n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\widetilde{\mathbf{v}}_T \\ &= \left(\widehat{\gamma} + \widehat{\gamma}\frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \right. \\ &\quad \left. - \lambda\frac{n_1 n_2}{n(n-2)}||\widehat{\boldsymbol{\beta}}_T||_1\right)\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \\ &\quad - \lambda\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) + (n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\widetilde{\mathbf{v}}_T \\ &\quad + \widehat{\gamma}\frac{n_1 n_2}{n(n-2)}\left(\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right)\widehat{\mu}_{a\cdot T} \\ &\quad - \lambda\frac{n_1 n_2}{n(n-2)}||\widehat{\boldsymbol{\beta}}_T||_1 \cdot \widehat{\mu}_{a\cdot T} \\ &= \frac{n_1 n_2}{n(n-2)}\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \lambda\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) \\ &\quad + (n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\widetilde{\mathbf{v}}_T \\ &\quad + \widehat{\gamma}\frac{n_1 n_2}{n(n-2)}\left(\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right)\widehat{\mu}_{a\cdot T} \\ &\quad - \lambda\frac{n_1 n_2}{n(n-2)}||\widehat{\boldsymbol{\beta}}_T||_1 \cdot \widehat{\mu}_{a\cdot T},\end{aligned}$$

and finally

$$\begin{aligned}&\mathbf{C}_{aT}\widetilde{\mathbf{v}}_T - \frac{n_1 n_2}{n(n-2)}\widehat{\mu}_a \\ &\quad = -\lambda\mathbf{\Sigma}_{aT}\mathbf{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T) + (n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\widetilde{\mathbf{v}}_T \\ &\qquad + \frac{n_1 n_2}{n(n-2)}\left(\widehat{\gamma}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \lambda||\widehat{\boldsymbol{\beta}}_T||_1 - 1\right)\widehat{\mu}_{a\cdot T}.\end{aligned}$$

First, we deal with the term

$$(n-2)^{-1}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\widetilde{\mathbf{v}}_T = \underbrace{\frac{\widehat{\gamma}}{n-2}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}_{T_{1,a}} - \underbrace{\frac{\lambda}{n-2}\mathbf{U}'_{a\cdot T}\mathbf{U}_T\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}_{T_{2,a}}.$$

Conditional on $\{y_i\}_{i\in[n]}$ and $\mathbf{X}_T$, we have that

$$T_{1,a} \sim \mathcal{N}\left(0, \frac{n_1 n_2}{n(n-2)}\sigma_{a|T}\frac{\widehat{\gamma}^2}{n-2}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right)$$

and

$$\max_{a\in N}|T_{1,a}| \leq \sqrt{2\frac{n_1 n_2}{n(n-2)}\left(\max_{a\in N}\sigma_{a|T}\right)\widehat{\gamma}^2\widehat{\boldsymbol{\mu}}'_T\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T} \times\sqrt{\frac{\log\left((p-s)\log(n)\right)}{n-2}}$$

with probability at least $1 - \log^{-1}(n)$. On the event $\mathcal{A}$, we have that

$$\begin{aligned}\max_{a\in N}|T_{1,a}| &\leq (1 + o(1))\sqrt{2\pi_1\pi_2\gamma^2\left(\max_{a\in N}\sigma_{a|T}\right)||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}} \\ &\quad \times\sqrt{\frac{\log\left((p-s)\log(n)\right)}{n}} \\ &= (1 + o(1))\sqrt{2}\pi_1\pi_2(1 + \lambda||\boldsymbol{\beta}_T||_1)\frac{\lambda}{K_{\lambda_0}}.\end{aligned}$$

Since

$$\begin{aligned}||\boldsymbol{\beta}_T||_1 &\leq \sqrt{s}||\boldsymbol{\beta}_T||_2 \\ &= \sqrt{s}||\mathbf{\Sigma}_{TT}^{-1/2}\mathbf{\Sigma}_{TT}^{1/2}\boldsymbol{\beta}_T||_2 \\ &\leq \sqrt{s\Lambda_{\min}^{-1}(\mathbf{\Sigma}_{TT})||\boldsymbol{\beta}_T||^2_{\mathbf{\Sigma}_{TT}}}\end{aligned}$$

and

$$\begin{aligned}\lambda||\boldsymbol{\beta}_T||_1 &= \frac{\lambda_0||\boldsymbol{\beta}_T||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}} \\ &\leq \frac{\lambda_0\sqrt{s\Lambda^{-1}_{\min}(\boldsymbol{\Sigma}_{TT})||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}} \\ &\leq \frac{K_{\lambda_0}}{\sqrt{K}}\frac{\sqrt{\left(1\vee||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}} \\ &\leq \frac{K_{\lambda_0}}{\pi_1\pi_2\sqrt{K}},\end{aligned}$$

we have that

$$\begin{aligned}\max_{a\in N}|T_{1,a}| &\leq (1+o(1))\sqrt{2}\pi_1\pi_2\left(K^{-1}_{\lambda_0}+\left(\pi_1\pi_2\sqrt{K}\right)^{-1}\right)\lambda \\ &< (\alpha/3)\lambda\end{aligned}$$

by taking both K_{λ_0} and K sufficiently large.

Similarly, conditional on $\{y_i\}_{i\in[n]}$ and $\mathbf{X}_T$, we have that $T_{2,a}\sim\mathcal{N}\left(0,\sigma_{T_{2,a}}\right)$, where

$$\sigma_{T_{2,a}} = \frac{n_1n_2}{n(n-2)}\sigma_{a|T}\frac{\lambda^2}{n-2}\operatorname{sign}(\boldsymbol{\beta}_T)'\mathbf{S}^{-1}_{TT}\operatorname{sign}(\boldsymbol{\beta}_T).$$

Therefore, on the event $\mathcal{A}$,

$$\begin{aligned}\max_{a\in N}|T_{2,a}| &\leq (1+o(1))\lambda\sqrt{2\pi_1\pi_2\left(\max_{a\in N}\sigma_{a|T}\right)\Lambda^{-1}_{\min}(\boldsymbol{\Sigma}_{TT})} \\ &\quad\times\sqrt{\frac{s\log\left((p-s)\log(n)\right)}{n}} \\ &\leq (1+o(1))\sqrt{\frac{2}{K}}\lambda < (\alpha/3)\lambda\end{aligned}$$

with probability at least $1-\log^{-1}(n)$ for sufficiently large K.

Next, let

$$T_{3,a} = \frac{n_1n_2}{n(n-2)}\left(\widehat{\gamma}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}^{-1}_{TT}\widehat{\boldsymbol{\mu}}_T-\lambda||\widehat{\boldsymbol{\beta}}_T||_1-1\right)\widehat{\mu}_{a\cdot T}.$$

Simple algebra shows that

$$\widehat{\gamma}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}^{-1}_{TT}\widehat{\boldsymbol{\mu}}_T-\lambda||\widehat{\boldsymbol{\beta}}_T||_1-1 = -\frac{1+\lambda||\widehat{\boldsymbol{\beta}}_T||_1}{1+\frac{n_1n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}^{-1}_{TT}\widehat{\boldsymbol{\mu}}_T}.$$

Therefore conditional on $\{y_i\}_{i\in[n]}$ and $\mathbf{X}_T$, we have that $T_{3,a}\sim\mathcal{N}\left(0,\sigma_{T_{3,a}}\right)$ where

$$\sigma_{T_{3,a}} = \left(\frac{n_1n_2}{n(n-2)}\frac{1+\lambda||\widehat{\boldsymbol{\beta}}_T||_1}{1+\frac{n_1n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}'_T\mathbf{S}^{-1}_{TT}\widehat{\boldsymbol{\mu}}_T}\right)^2\frac{n}{n_1n_2}\sigma_{a|T}$$

On the event $\mathcal{A}$, this gives

$$\begin{aligned}\max_{a\in N}|T_{3,a}| &\leq (1+o(1))\frac{1+\lambda||\boldsymbol{\beta}_T||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}} \\ &\quad\times\sqrt{2\pi_1\pi_2\left(\max_{a\in N}\sigma_{a|T}\right)\frac{\log\left((p-s)\log(n)\right)}{n}} \\ &\leq (1+o(1))\sqrt{2}\frac{1+\lambda||\boldsymbol{\beta}_T||_1}{1+\pi_1\pi_2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}} \\ &\quad\times\frac{1}{K_{\lambda_0}\sqrt{1\vee||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}}\lambda \\ &< (\alpha/3)\lambda\end{aligned}$$

with probability at least $1-\log^{-1}(n)$ when K_{λ_0} and K are chosen sufficiently large.

Piecing all these results together, we have that

$$\max_{a\in N}|\mathbf{C}_{aT}\widetilde{\mathbf{v}}_T-\frac{n_1n_2}{n(n-2)}\widehat{\mu}_a| < 1. \tag{A.3}$$

□

B. Proof of Theorem 4

The theorem will be shown using standard tools described in [24]. First, in order to provide a lower bound on the minimax risk, we will construct a finite subset of $\Theta(\boldsymbol{\Sigma},\tau,s)$, which contains the most difficult instances of the estimation problem so that estimation over the subset is as difficult as estimation over the whole family. Let $\Theta_1\subset\Theta(\boldsymbol{\Sigma},\tau,s)$, be a set with finite number of elements, so that

$$\inf_{\Psi}R(\Psi,\Theta(\boldsymbol{\Sigma},\tau,s)\geq\inf_{\Psi}\max_{\boldsymbol{\theta}\in\Theta_1}\mathbb{P}_{\boldsymbol{\theta}}[\Psi(\{\mathbf{x}_i,y_i\}_{i\in[n]})\neq T(\boldsymbol{\theta})].$$

To further lower bound the right hand side of the display above, we will use [24, Th. 2.5]. Suppose that $\Theta_1=\{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}$ where $T(\boldsymbol{\theta}_a)\neq T(\boldsymbol{\theta}_b)$ and

$$\frac{1}{M}\sum_{a=1}^{M}KL(\mathbb{P}_{\boldsymbol{\theta}_0}|\mathbb{P}_{\boldsymbol{\theta}_a})\leq\alpha\log(M),\quad\alpha\in(0,1/8) \tag{A.4}$$

then

$$\inf_{\Psi}R(\Psi,\Theta(\beta,s)\geq\frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\alpha-\sqrt{\frac{2\alpha}{\log(M)}}\right).$$

Without loss of generality, we will consider $\boldsymbol{\theta}_a=(\boldsymbol{\mu}_a,\mathbf{0},\boldsymbol{\Sigma})$. Denote $\mathbb{P}_{\boldsymbol{\theta}_a}$ the joint distributions of $\{\mathbf{X}_i,Y_i\}_{i\in[n]}$. Under $\mathbb{P}_{\boldsymbol{\theta}_a}$, we have $\mathbb{P}_{\boldsymbol{\theta}_a}(Y_i=1)=\mathbb{P}_{\boldsymbol{\theta}_a}(Y_i=2)=\frac{1}{2}$, $\mathbf{X}_i|Y_i=1\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ and $\mathbf{X}_i|Y_i=2\sim\mathcal{N}(\boldsymbol{\mu}_a,\boldsymbol{\Sigma})$. Denote $f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ the density function of a multivariate Normal distribution. With this we have

$$\begin{aligned}KL(\mathbb{P}_{\boldsymbol{\theta}_0}|\mathbb{P}_{\boldsymbol{\theta}_a}) &= \mathbb{E}_{\boldsymbol{\theta}_0}\log\frac{d\mathbb{P}_{\boldsymbol{\theta}_0}}{d\mathbb{P}_{\boldsymbol{\theta}_a}} \\ &= \mathbb{E}_{\boldsymbol{\theta}_0}\log\frac{\prod_{i\in[n]}d\mathbb{P}_{\boldsymbol{\theta}_0}[\mathbf{X}_i|Y_i]\mathbb{P}_{\boldsymbol{\theta}_0}[Y_i]}{\prod_{i\in[n]}d\mathbb{P}_{\boldsymbol{\theta}_a}[\mathbf{X}_i|Y_i]\mathbb{P}_{\boldsymbol{\theta}_a}[Y_i]} \\ &= \mathbb{E}_{\boldsymbol{\theta}_0}\sum_{i\,:\,y_i=2}\log\frac{f(\mathbf{X}_i;\boldsymbol{\mu}_0,\boldsymbol{\Sigma})}{f(\mathbf{X}_i;\boldsymbol{\mu}_a,\boldsymbol{\Sigma})} \\ &= \frac{\mathbb{E}_{\boldsymbol{\theta}_0}n_2}{2}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_a)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_a) \\ &= \frac{n}{4}(\boldsymbol{\beta}_0-\boldsymbol{\beta}_a)'\boldsymbol{\Sigma}(\boldsymbol{\beta}_0-\boldsymbol{\beta}_a)\end{aligned} \tag{A.5}$$

where $\boldsymbol{\beta}_a=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_a$. We proceed to construct different finite collections for which (A.4) holds.

Consider a collection $\Theta_1=\{\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{p-s}\}$, with $\boldsymbol{\theta}_a=(\boldsymbol{\mu}_a,\mathbf{0})$, that contains instances whose supports differ in only one component. Vectors $\{\boldsymbol{\mu}_a\}_{a=0}^{p-s}$ are constructed indirectly through $\{\boldsymbol{\beta}_a\}_{a=0}^{p-s}$, using the relationship $\boldsymbol{\beta}_a=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_a$. Note that this construction is possible, since $\boldsymbol{\Sigma}$ is a full rank matrix. For every a, all s non-zero elements of the vector $\boldsymbol{\beta}_a$ are equal to τ. Let T be the support and $u(T)$ an element of the support T for which (III.1) is minimized. Set $\boldsymbol{\beta}_0$ so that $\operatorname{supp}(\boldsymbol{\beta}_0)=T$. The remaining $p-s$ parameter vectors

$\{\boldsymbol{\beta}_a\}_{a=1}^{p-s}$ are constructed so that the support of $\boldsymbol{\beta}_a$ contains all $s-1$ element in $T\backslash u(T)$ and then one more element from $[p]\backslash T$. With this, (A.5) gives

$$KL(\mathbb{P}_{\boldsymbol{\theta}_0}|\mathbb{P}_{\boldsymbol{\theta}_a}) = \frac{n\tau^2}{4}(\Sigma_{uu} + \Sigma_{vv} - 2\Sigma_{uv})$$

and (III.1) gives

$$\frac{1}{p-s}\sum_{a=1}^{p-s} KL(\mathbb{P}_{\boldsymbol{\theta}_0}|\mathbb{P}_{\boldsymbol{\theta}_a}) = \frac{n\tau^2}{4}\varphi_{\text{close}}(\boldsymbol{\Sigma}).$$

It follows from the display above that if

$$\tau < \sqrt{\frac{4}{\varphi_{\text{close}}(\boldsymbol{\Sigma})}\frac{\log(p-s+1)}{n}}, \tag{A.6}$$

then (A.4) holds with $\alpha = 1/16$.

Next, we consider another collection $\Theta_2 = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, where $M = \binom{p-s}{s}$, and the Hamming distance between $T(\boldsymbol{\theta}_0)$ and $T(\boldsymbol{\theta}_a)$ is equal to $2s$. As before, $\boldsymbol{\theta}_a = (\boldsymbol{\mu}_a, \mathbf{0})$ and vectors $\{\boldsymbol{\mu}_a\}_{a=0}^{M}$ are constructed so that $\boldsymbol{\beta}_a = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_a$ with s non-zero components equal to τ. Let T be the support set for which the minimum in (III.2) is attained. Set the vector $\boldsymbol{\beta}_0$ so that $\text{supp}(\boldsymbol{\beta}_0) = T$. The remaining vectors $\{\boldsymbol{\beta}_a\}_{a=1}^{M}$ are set so that their support contains s elements from the set $[p]\backslash T$. Now, (A.5) gives

$$KL(\mathbb{P}_{\boldsymbol{\theta}_0}|\mathbb{P}_{\boldsymbol{\theta}_a}) = \frac{n\tau^2}{4}\mathbf{1}'\boldsymbol{\Sigma}_{T(\boldsymbol{\theta}_0)\cup T(\boldsymbol{\theta}_a),T(\boldsymbol{\theta}_0)\cup T(\boldsymbol{\theta}_a)}\mathbf{1}.$$

Using (III.2), if

$$\tau < \sqrt{\frac{4}{\varphi_{\text{far}}(\boldsymbol{\Sigma})}\frac{\log\binom{p-s}{s}}{n}}, \tag{A.7}$$

then (A.4) holds with $\alpha = 1/16$.

Combining (A.6) and (A.7), by taking the larger β between the two, we obtain the result.

C. Proof of Theorem 5

For a fixed T, let $\Delta(T') = f(T) - f(T')$ and $\mathcal{T} = \{T' \subset [p] : |T'| = s, T' \neq T\}$. Then

$$\begin{aligned}\mathbb{P}_T[\widehat{T} \neq T] &= \mathbb{P}_T[\bigcup_{T'\in\mathcal{T}}\{\Delta(T') < 0\}] \\ &\leq \sum_{T'\in\mathcal{T}}\mathbb{P}_T[\Delta(T') < 0].\end{aligned}$$

Partition $\widehat{\boldsymbol{\mu}}_T = (\widehat{\boldsymbol{\mu}}_1', \widehat{\boldsymbol{\mu}}_2')'$, where $\widehat{\boldsymbol{\mu}}_1$ contains the variables in $T \cap T'$, and $\widehat{\boldsymbol{\mu}}_{T'} = (\widehat{\boldsymbol{\mu}}_1', \widehat{\boldsymbol{\mu}}_3')'$. Similarly, we can partition the covariance matrix $\mathbf{S}_{TT}$ and $\mathbf{S}_{T'T'}$. Then,

$$g(T) = \widehat{\boldsymbol{\mu}}_1'\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1 + \widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}$$

where $\widetilde{\boldsymbol{\mu}}_{2|1} = \widehat{\boldsymbol{\mu}}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1$ and $\mathbf{S}_{22|1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ (see [17, Sec. 3.6.2]). Furthermore, we have that

$$\Delta(T') = \widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1} - \widetilde{\boldsymbol{\mu}}_{3|1}'\mathbf{S}_{33|1}^{-1}\widetilde{\boldsymbol{\mu}}_{3|1}.$$

The two terms are correlated, but we will ignore this correlation and use the union bound to lower bound the first term and upper bound the second term. We start with analyzing $\widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}$, noting that the result for the second term will follow in the same way. By [17, Th. 3.4.5], we have that

$$\mathbf{S}_{22|1} \sim \mathcal{W}_{s-|T\cap T'|}\left((n-2)^{-1}\boldsymbol{\Sigma}_{22|1}, n-2-|T\cap T'|\right)$$

and independent of $(\mathbf{S}_{12}, \mathbf{S}_{11}, \widehat{\boldsymbol{\mu}})$. Therefore $\mathbf{S}_{22|1}$ is independent of $\widetilde{\boldsymbol{\mu}}_{2|1}$ and [20, Th. 3.2.12] gives us that

$$(n-2)\frac{\widetilde{\boldsymbol{\mu}}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}}{\widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}} \sim \chi^2_{n-1-s}.$$

As in Lemma 4, we can show that

$$1 - C_1\sqrt{\frac{\log(\eta^{-1})}{n}} \leq \frac{\widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}}{\widetilde{\boldsymbol{\mu}}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}}$$

and

$$1 + C_2\sqrt{\frac{\log(\eta^{-1})}{n}} \geq \frac{\widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}}{\widetilde{\boldsymbol{\mu}}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1}}.$$

For $\widetilde{\boldsymbol{\mu}}_{2|1}$, we have

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_{2|1} &= \widehat{\boldsymbol{\mu}}_2 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1 \\ &= \widehat{\boldsymbol{\mu}}_{2|1} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1 - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1,\end{aligned}$$

where $\widehat{\boldsymbol{\mu}}_{2|1} \sim \mathcal{N}(\boldsymbol{\mu}_{2|1}, \frac{n}{n_1 n_2}\boldsymbol{\Sigma}_{22|1})$, independent of $\widehat{\boldsymbol{\mu}}_1$, and $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$. Conditioning on $\widehat{\boldsymbol{\mu}}_1$ and $\mathbf{S}_1$, we have that

$$\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1 = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1 + \mathbf{Z},$$

where $\mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}, (n-2)^{-1}\widehat{\boldsymbol{\mu}}_1'\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1\boldsymbol{\Sigma}_{22|1}\right)$. Since $\widehat{\boldsymbol{\mu}}_{2|1}$ is independent of $(\mathbf{S}_{12}, \mathbf{S}_{11}, \widehat{\boldsymbol{\mu}}_1)$, we have that

$$\widetilde{\boldsymbol{\mu}}_{2|1}|\widehat{\boldsymbol{\mu}}_1, \mathbf{S}_{11} \sim \mathcal{N}\left(\boldsymbol{\mu}_{2|1}, a\boldsymbol{\Sigma}_{22|1}\right),$$

where $a = \frac{n}{n_1 n_2} + (n-2)^{-1}\widehat{\boldsymbol{\mu}}_1'\mathbf{S}_{11}^{-1}\widehat{\boldsymbol{\mu}}_1$. Then

$$\widetilde{\boldsymbol{\mu}}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1} \mid \widehat{\boldsymbol{\mu}}_1, \mathbf{S}_{11} \sim a\chi^2_{|T\backslash T'|}\left(a^{-1}\boldsymbol{\mu}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\mu}_{2|1}\right).$$

Therefore, conditioned on $(\widehat{\boldsymbol{\mu}}_1, \mathbf{S}_{11})$,

$$\begin{aligned}&\widetilde{\boldsymbol{\mu}}_{2|1}'\mathbf{S}_{22|1}^{-1}\widetilde{\boldsymbol{\mu}}_{2|1} \\ &\quad\geq \left(1 - C_1\sqrt{\frac{\log(\eta^{-1})}{n}}\right) \\ &\qquad\times\Bigg(\left(\boldsymbol{\mu}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\mu}_{2|1} + a|T\backslash T'|\right) \\ &\qquad -2\sqrt{\left(2a\boldsymbol{\mu}_{2|1}'\boldsymbol{\Sigma}_{22|1}^{-1}\boldsymbol{\mu}_{2|1} + a^2|T\backslash T'|\right)\log(\eta^{-1})}\Bigg)\end{aligned}$$

with probability $1-2\eta$. Similarly,

$$\begin{aligned}&\widetilde{\boldsymbol{\mu}}_{3|1}'\mathbf{S}_{33|1}^{-1}\widetilde{\boldsymbol{\mu}}_{3|1} \\ &\quad\leq \left(1 + C_2\sqrt{\frac{\log(\eta^{-1})}{n}}\right) \\ &\qquad\times\Bigg(\left(\boldsymbol{\mu}_{3|1}'\boldsymbol{\Sigma}_{33|1}^{-1}\boldsymbol{\mu}_{3|1} + a|T'\backslash T|\right) \\ &\qquad + 2\sqrt{\left(2a\boldsymbol{\mu}_{3|1}'\boldsymbol{\Sigma}_{33|1}^{-1}\boldsymbol{\mu}_{3|1} + a^2|T'\backslash T|\right)\log(\eta^{-1})} \\ &\qquad + 2a\log(\eta^{-1})\Bigg)\end{aligned}$$

with probability $1-2\eta$. Finally, Lemma 6 gives that $|a| \leq C\left(1 \vee \boldsymbol{\mu}_1' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1\right) n^{-1}$ with probability $1-2\eta$.

Set $\eta_k = \left(\binom{p-s}{s-k}\binom{s}{k} s \log(n)\right)^{-1}$. For any $T' \subset [p]$, where $|T'| = s$ and $|T' \cap T| = k$, we have that

$$\begin{aligned}
&\widetilde{\boldsymbol{\mu}}_{2|1}' \mathbf{S}_{22|1}^{-1} \widetilde{\boldsymbol{\mu}}_{2|1} - \widetilde{\boldsymbol{\mu}}_{3|1}' \mathbf{S}_{33|1}^{-1} \widetilde{\boldsymbol{\mu}}_{3|1} \\
&\quad \geq (1-o(1)) \boldsymbol{\mu}_{2|1}' \boldsymbol{\Sigma}_{22|1}^{-1} \boldsymbol{\mu}_{2|1} - (1+o(1)) \boldsymbol{\mu}_{3|1}' \boldsymbol{\Sigma}_{33|1}^{-1} \boldsymbol{\mu}_{3|1} \\
&\qquad - C\sqrt{\left(1 \vee \boldsymbol{\mu}_1' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1\right) \boldsymbol{\mu}_{2|1}' \boldsymbol{\Sigma}_{22|1}^{-1} \boldsymbol{\mu}_{2|1} \Gamma_{n,p,s,k}} \\
&\qquad - C\left(1 \vee \boldsymbol{\mu}_1' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_1\right) \Gamma_{n,p,s,k}.
\end{aligned}$$

The right hand side in the above display is bounded away from zero with probability

$$1 - \mathcal{O}\left(\left(\binom{p-s}{s-k}\binom{s}{k} s \log(n)\right)^{-1}\right)$$

under the assumptions. Therefore,

$$\begin{aligned}
\mathbb{P}_T[\widehat{T} \neq T] &\leq \sum_{k=0}^{s-1} \sum_{T' \in \mathcal{T} \,:\, |T \cap T'| = k} \mathbb{P}_T[\Delta(T') < 0] \\
&\leq \frac{C}{\log(n)},
\end{aligned}$$

which completes the proof.

Appendix B
Proof of Risk Consistency

In this section, we give a proof of Corollary 1. From Theorem 3 we have that $\widehat{\mathbf{v}} = (\widehat{\mathbf{v}}_T', \mathbf{0}')'$ with $\widehat{\mathbf{v}}_T$ defined in (II.9). Define

$$\widetilde{\mathbf{v}}_T = \frac{n(n-2)}{n_1 n_2}\left(1 + \frac{n_1 n_2}{n(n-2)} \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T\right) \widehat{\mathbf{v}}_T.$$

To obtain a bound on the risk, we need to control

$$\frac{-\widetilde{\mathbf{v}}_T'(\boldsymbol{\mu}_{i,T} - \widehat{\boldsymbol{\mu}}_{i,T}) - \widetilde{\mathbf{v}}_T' \widehat{\boldsymbol{\mu}}_T/2}{\sqrt{\widetilde{\mathbf{v}}_T' \boldsymbol{\Sigma}_{TT} \widetilde{\mathbf{v}}_T}}$$

for $i \in \{1,2\}$. Define the following quantities

$$\begin{aligned}
&\delta_1 = \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - ||\boldsymbol{\beta}_T||_1, \;\; \widetilde{\delta}_1 = \delta_1 / ||\boldsymbol{\beta}_T||_1, \\
&\delta_2 = \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T - ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2, \text{ and } \widetilde{\delta}_2 = \delta_2 / ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2.
\end{aligned}$$

Under the assumptions, we have that

$$\begin{aligned}
\lambda_0 &= \mathcal{O}\left(\sqrt{\frac{\Lambda_{\min}(\boldsymbol{\Sigma}_{TT}) ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}{K(n)s}}\right), \\
r_n &= \mathcal{O}\left(\frac{\lambda_0 ||\boldsymbol{\beta}_T||_1}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}\right) \\
&= \mathcal{O}\left(\frac{||\boldsymbol{\beta}_T||_1}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}} \sqrt{\frac{\Lambda_{\min}(\boldsymbol{\Sigma}_{TT})}{K(n)s}}\right),
\end{aligned}$$

and

$$\widetilde{\delta}_2 = \mathcal{O}_P\left(\sqrt{\frac{\log\log(n)}{n}} \vee \frac{s \vee \log\log(n)}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2 n}\right).$$

The last equation follows from Lemma 6. Note that $\widetilde{\delta}_2 = \mathcal{O}(r_n)$. From Lemma 7, we have that $\widetilde{\delta}_1 = o_p(1)$.

We have

$$\lambda \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) = \lambda ||\boldsymbol{\beta}_T||_1 (1 + \mathcal{O}_P(\widetilde{\delta}_1)) = \mathcal{O}_P(r_n),$$

since Lemma 7 gives $\widetilde{\delta}_1 = o_p(1)$, and

$$\frac{\frac{n(n-2)}{n_1 n_2} + \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T}{1 + \pi_1 \pi_2 ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2} = \mathcal{O}_P(1).$$

Therefore

$$\widetilde{\mathbf{v}}_T = (1 + \mathcal{O}_P(r_n)) \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T - \mathcal{O}_P(1) \lambda_0 \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T).$$

With this, we have

$$\begin{aligned}
\widetilde{\mathbf{v}}_T' \widehat{\boldsymbol{\mu}}_T &= (1 + \mathcal{O}_P(r_n))(1 + \mathcal{O}_P(\widetilde{\delta}_2)) ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2 \\
&\quad - \mathcal{O}_P(1)(1 + \mathcal{O}_P(\widetilde{\delta}_1)) \lambda_0 ||\boldsymbol{\beta}_T||_1 \\
&= ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2 \left[1 + \mathcal{O}_P(r_n) - \mathcal{O}_P(1) \frac{\lambda_0 ||\boldsymbol{\beta}_T||_1}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}\right] \\
&= ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2 [1 + \mathcal{O}_P(r_n)], \qquad \text{(B.1)}
\end{aligned}$$

where the last line follows from

$$\lambda ||\boldsymbol{\beta}_T||_1 \asymp \lambda_0 ||\boldsymbol{\beta}_T||_1 / ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2.$$

Next

$$\begin{aligned}
&(\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T})' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T \\
&\quad \leq ||\mathbf{S}_{TT}^{-1/2}(\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T})||_2 ||\mathbf{S}_{TT}^{-1/2} \widehat{\boldsymbol{\mu}}_T||_2 \\
&\quad \leq (1 + \mathcal{O}_P(\sqrt{s/n})) \Lambda_{\min}^{-1/2}(\boldsymbol{\Sigma}_{TT}) \sqrt{s} \\
&\qquad \times ||\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T}||_\infty ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}} \sqrt{1 + \mathcal{O}_P(\widetilde{\delta}_2)} \\
&\quad = ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}} \mathcal{O}_P\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log\log(n)/n}\right)
\end{aligned}$$

and similarly

$$\begin{aligned}
&(\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T})' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\
&\quad = \sqrt{q_n} \mathcal{O}_P\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log\log(n)/n}\right).
\end{aligned}$$

Combining these two estimates, we have

$$\begin{aligned}
&\left|\widetilde{\mathbf{v}}_T'(\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T})\right| \\
&\quad = ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}} \left(1 - \lambda_0 \sqrt{q_n} / ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}\right) \\
&\qquad \times \mathcal{O}_P\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log\log(n)/n}\right) \\
&\quad = ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}} \mathcal{O}_P\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log\log(n)/n}\right). \qquad \text{(B.2)}
\end{aligned}$$

From (B.1) and (B.2), we have that

$$-\left(\widetilde{\mathbf{v}}_T'(\boldsymbol{\mu}_{1,T} - \widehat{\boldsymbol{\mu}}_{1,T})\right) - \widetilde{\mathbf{v}}_T' \widehat{\boldsymbol{\mu}}_T/2 = -\frac{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}{2}(1 + \mathcal{O}_P(r_n)). \qquad \text{(B.3)}$$

Finally, a simple calculation gives,

$$\begin{aligned}\widetilde{\mathbf{v}}_T'\boldsymbol{\Sigma}_{TT}\widetilde{\mathbf{v}}_T &\le \Lambda_{\max}\left(\mathbf{S}_{TT}^{-1/2}\boldsymbol{\Sigma}_{TT}\mathbf{S}_{TT}^{-1/2}\right)\\ &\quad\times\Big((1+\mathcal{O}_P(r_n))^2\,\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\\ &\qquad+\mathcal{O}_P(1)\lambda_0^2\,\mathrm{sign}(\boldsymbol{\beta}_T)'\mathbf{S}_{TT}^{-1}\,\mathrm{sign}(\boldsymbol{\beta}_T)\\ &\qquad-\mathcal{O}_P(1)\lambda_0\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\,\mathrm{sign}(\boldsymbol{\beta}_T)\Big)\\ &=||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2\\ &\quad\times\left(1+\mathcal{O}_P\left(r_n\vee\widetilde{\delta}_2\vee\frac{\lambda_0^2 q_n}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}\vee\sqrt{\frac{s}{n}}\right)\right)\\ &=||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2\left(1+\mathcal{O}_P\left(r_n\vee\frac{\lambda_0^2 q_n}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}\right)\right).\end{aligned}\tag{B.4}$$

Combining the equation (B.3) and (B.4), we have that

$$\begin{aligned}&\frac{-\widetilde{\mathbf{v}}_T'(\boldsymbol{\mu}_{1,T}-\widehat{\boldsymbol{\mu}}_{1,T})-\widetilde{\mathbf{v}}_T'\widehat{\boldsymbol{\mu}}_T/2}{\sqrt{\widetilde{\mathbf{v}}_T'\boldsymbol{\Sigma}_{TT}\widetilde{\mathbf{v}}_T}}\\ &\quad=-\frac{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}}{2}\frac{(1+\mathcal{O}_P(r_n))}{\sqrt{1+\mathcal{O}_P\left(r_n\vee\frac{\lambda_0^2 q_n}{||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2}\right)}}.\end{aligned}$$

This completes the proof.

Appendix C
Alternative Encoding of the Class Labels

The optimization problem in (I.3) uses a particular scheme to encode class labels in the vector $\mathbf{z}$, though other choices are possible as well. For example, suppose that we choose $z_i = z^{(1)}$ if $y_i = 1$ and $z_i = z^{(2)}$ if $y_i = 2$, with $z^{(1)}$ and $z^{(2)}$ such that $n_1 z^{(1)} + n_2 z^{(2)} = 0$. Let $\widetilde{\mathbf{v}} = (\widetilde{\mathbf{v}}_{\widetilde{T}}', \mathbf{0}')'$ (for some $\widetilde{T} \subset [p]$) be a solution to (I.3) with the alternative coding. The KKT conditions for the vector $\widetilde{\mathbf{v}}$ are

$$\left(\mathbf{S}_{\widetilde{T}\widetilde{T}}+\frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}_{\widetilde{T}}\widehat{\boldsymbol{\mu}}_{\widetilde{T}}'\right)\widetilde{\mathbf{v}}_{\widetilde{T}}=\frac{n_1 z^{(1)}}{n-2}\widehat{\boldsymbol{\mu}}_{\widetilde{T}}-\widetilde{\lambda}\,\mathrm{sign}(\widetilde{\mathbf{v}}_{\widetilde{T}})\tag{C.1}$$

and

$$\left|\left|\left(\mathbf{S}_{\widetilde{N}\widetilde{T}}+\frac{n_1 n_2}{n(n-2)}\widehat{\boldsymbol{\mu}}_{\widetilde{N}}\widehat{\boldsymbol{\mu}}_{\widetilde{T}}'\right)\widetilde{\mathbf{v}}_{\widetilde{T}}-\frac{n_1 z^{(1)}}{n-2}\widehat{\boldsymbol{\mu}}_{\widetilde{N}}\right|\right|_\infty\le\widetilde{\lambda}.\tag{C.2}$$

Now, choosing $\widetilde{\lambda}=\frac{z^{(1)}n}{n_2}\lambda$, we obtain that $\widetilde{\mathbf{v}}_{\widetilde{T}}$, which satisfies (C.1) and (C.2), is proportional to $\widehat{\mathbf{w}}_T$ with $\widetilde{T}=T$ (compare (C.1) and (C.2) to (II.12) and (II.13)). Therefore, the choice of different coding schemes of the response variable z_i does not effect the result.

Appendix D
Derivation of Eq. (II.11)

Let $\bar{\pi}=\pi_1\pi_2$. Recall that $\mathbf{C}=\mathrm{Var}(\mathbf{X})$ and $\mathbf{C}=\boldsymbol{\Sigma}+\bar{\pi}\,\boldsymbol{\mu}\boldsymbol{\mu}'$. Using the Woodbury matrix identity

$$\mathbf{C}_{TT}^{-1}=\boldsymbol{\Sigma}_{TT}^{-1}-\bar{\pi}\frac{\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}}{1+\bar{\pi}\,\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T},$$

which gives us

$$\begin{aligned}\mathbf{C}_{NT}\mathbf{C}_{TT}^{-1}&=\left(\boldsymbol{\Sigma}_{NT}+\bar{\pi}\,\boldsymbol{\mu}_N\boldsymbol{\mu}_T'\right)\left(\boldsymbol{\Sigma}_{TT}^{-1}-\bar{\pi}\frac{\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}}{1+\bar{\pi}\,\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T}\right)\\ &=\left(\boldsymbol{\Sigma}_{NT}+\bar{\pi}\,\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\right)\\ &\quad\times\left(\boldsymbol{\Sigma}_{TT}^{-1}-\bar{\pi}\frac{\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}}{1+\bar{\pi}\,\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T}\right)\\ &=\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}.\end{aligned}$$

This shows the identity (II.11).

Appendix E
Technical Results

We provide some technical lemmas which are useful for proving the main results. Without loss of generality, $\pi_1=\pi_2=1/2$ in model (I.2). Define

$$\mathcal{E}_n=\left\{\frac{n}{4}\le n_1\le\frac{3n}{4}\right\}\cap\left\{\frac{n}{4}\le n_2\le\frac{3n}{4}\right\},\tag{E.1}$$

where n_1, n_2 are defined in §I. Observe that $n_1 \sim \mathrm{Binomial}(n, 1/2)$, which gives $\mathbb{P}[\{n_1 \le n/4\}] \le \exp(-3n/64)$ and $\mathbb{P}[\{n_1 \ge 3n/4\}] \le \exp(-3n/64)$ using standard tail bound for binomial random variable (see [7, p. 130]). Therefore

$$\mathbb{P}[\mathcal{E}_n]\ge 1-4\exp(-3n/64).$$

The analysis is performed by conditioning on $\mathbf{y}$ and, in particular, we will perform analysis on the event $\mathcal{E}_n$. Note that on $\mathcal{E}_n$, $16/9n^{-1} \le n/(n_1 n_2) \le 16n^{-1}$. In our analysis, we do not strive to obtain the sharpest possible constants.

A. Deviation of the Quadratic Scaling Term

In this section, we collect lemmas that will help us deal with bounding the deviation of $\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T$ from $\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T$.

Lemma 4: Define the event

$$\begin{aligned}\mathcal{E}_1(\eta)=&\left\{1-C_1\sqrt{\frac{\log(\eta^{-1})}{n}}\le\frac{\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\right\}\\ &\bigcap\left\{1+C_2\sqrt{\frac{\log(\eta^{-1})}{n}}\ge\frac{\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\right\},\end{aligned}$$

for some constants $C_1, C_2 > 0$. Assume that $s = o(n)$, then $\mathbb{P}[\mathcal{E}_1(\eta)] \ge 1-\eta$ for n sufficiently large.

Proof of Lemma 4: Using [20, Th. 3.2.12]

$$(n-2)\frac{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\sim\chi^2_{n-1-s}$$

(F.2) gives

$$\frac{n-2}{n-1-s}\frac{1}{1+\sqrt{\frac{16\log(\eta^{-1})}{3(n-1-s)}}}\le\frac{\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}$$

and

$$\frac{n-2}{n-1-s}\frac{1}{1-\sqrt{\frac{16\log(\eta^{-1})}{3(n-1-s)}}}\ge\frac{\widehat{\boldsymbol{\mu}}_T'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}$$

with probability at least $1-\eta$. Since $s = o(n)$, the above display becomes

$$1 - C_1\sqrt{\frac{\log(\eta^{-1})}{n}} \leq \frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T} \leq 1 + C_2\sqrt{\frac{\log(\eta^{-1})}{n}}$$

for n sufficiently large. □

Lemma 5: Define the event

$$\begin{aligned}\mathcal{E}_2(\eta) = \Bigg\{ &\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \leq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \\ &+ C_1\left(\frac{s \vee \log(\eta^{-1})}{n} \vee \sqrt{\frac{||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\log(\eta^{-1})}{n}}\right)\Bigg\} \\ \bigcap \Bigg\{ &\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \geq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \\ &- C_2\left(\frac{s}{n} \vee \sqrt{\frac{||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\log(\eta^{-1})}{n}}\right)\Bigg\}.\end{aligned}$$

Assume that $\beta_{\min} \geq cn^{-1/2}$, then $\mathbb{P}[\mathcal{E}_2(\eta)] \geq 1-2\eta$ for n sufficiently large.

Proof of Lemma 5: Recall that

$$\widehat{\boldsymbol{\mu}}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \frac{n}{n_1 n_2}\boldsymbol{\Sigma}_{TT}).$$

Therefore

$$\frac{n_1 n_2}{n}\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \sim \chi^2_s\left(\frac{n_1 n_2}{n}\boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right).$$

Using (F.3), we have that

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T &\leq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} + \frac{ns}{n_1 n_2} \\ &\quad + \frac{2n}{n_1 n_2}\sqrt{\left(s + 2\frac{n_1 n_2}{n}||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)\log(\eta^{-1})} \\ &\quad + \frac{2n}{n_1 n_2}\log(\eta^{-1}) \\ &\leq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} + \frac{16s}{n} \\ &\quad + 32\sqrt{\left(\frac{s}{n^2} + \frac{2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}{n}\right)\log(\eta^{-1})} \\ &\quad + \frac{32\log(\eta^{-1})}{n} \\ &\leq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \\ &\quad + C_1\left(\frac{s}{n} \vee \sqrt{\frac{||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\log(\eta^{-1})}{n}} \vee \frac{\log(\eta^{-1})}{n}\right),\end{aligned}$$

with probability $1-\eta$. The second inequality follows since we are working on the event $\mathcal{E}_n$, and the third inequality follows from the fact that $\beta_{\min} \geq cn^{-1/2}$. A lower bound follows from (F.4),

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T &\geq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} + \frac{ns}{n_1 n_2} \\ &\quad - \frac{2n}{n_1 n_2}\sqrt{\left(s + 2\frac{n_1 n_2}{n}||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)\log(\eta^{-1})} \\ &\geq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} + \frac{16s}{9n} \\ &\quad - 32\sqrt{\left(\frac{s}{n^2} + \frac{2||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}}{n}\right)\log(\eta^{-1})} \\ &\geq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} - C_2\left(\frac{s}{n} \vee \sqrt{\frac{||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\log(\eta^{-1})}{n}}\right)\end{aligned}$$

with probability $1-\eta$. □

Lemma 6: On the event $\mathcal{E}_1(\eta) \cap \mathcal{E}_2(\eta)$ the following holds

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right| \\ &\leq C\left(\left(||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} \vee ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}\right)\sqrt{\frac{\log(\eta^{-1})}{n}} \bigvee \frac{s \vee \log(\eta^{-1})}{n}\right).\end{aligned}$$

Proof of Lemma 6: On the event $\mathcal{E}_1(\eta) \cap \mathcal{E}_2(\eta)$, using Lemma 4 and Lemma 5, we have that

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T &= \frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T \\ &= \frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T \\ &\quad + \frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}\left(\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right) \\ &\leq ||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}} + C_1||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\sqrt{\frac{\log(\eta^{-1})}{n}} \\ &\quad + C_2\left(\frac{s \vee \log(\eta^{-1})}{n} \vee \sqrt{\frac{||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\log(\eta^{-1})}{n}}\right).\end{aligned}$$

A lower bound is obtained in the same way. □

B. Other Results

Let the event $\mathcal{E}_3(\eta)$ be defined as

$$\begin{aligned}\mathcal{E}_3(\eta) = \bigcap_{a\in[s]} \Bigg\{ &\left|\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T - \boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right| \\ &\leq \sqrt{32(\boldsymbol{\Sigma}_{TT}^{-1})_{aa}\frac{\log(s\eta^{-1})}{n}}\Bigg\}. \quad \text{(E.2)}\end{aligned}$$

Since $\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T$ is a multivariate normal with mean $\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T$ and variance $\frac{n}{n_1 n_2}\boldsymbol{\Sigma}_{TT}^{-1}$, we have $\mathbb{P}[\mathcal{E}_3(\eta)] \geq 1-\eta$.

Furthermore, define the event $\mathcal{E}_4(\eta)$ as

$$\begin{aligned}\mathcal{E}_4(\eta) = \Bigg\{ &|(\widehat{\boldsymbol{\mu}}_T - \boldsymbol{\mu}_T)' \boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)| \\ &\leq \sqrt{32\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT})\frac{s\log(\eta^{-1})}{n}}\Bigg\}. \quad \text{(E.3)}\end{aligned}$$

Since $\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)$ follows a Normal distribution with the mean $\boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)$ and variance $\frac{n}{n_1 n_2} \operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)$, we have $\mathbb{P}[\mathcal{E}_4(\eta)] \geq 1 - \eta$.

The next result gives a deviation of $\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)$ from $\boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)$.

Lemma 7: The following inequality

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad \leq C\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \left(1 \vee ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2\right)} \vee ||\beta_T||_1\right) \\ &\quad\quad \times \sqrt{\frac{\log\log(n)}{n}} \end{aligned} \tag{E.4}$$

holds with probability at least $1 - \mathcal{O}(\log^{-1}(n))$.

Proof: Using the triangle inequality

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad \leq \left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad\quad + \left|\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right|. \end{aligned} \tag{E.5}$$

For the first term, we write

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad \leq \operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\ &\quad\quad \times \left|\frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} - \frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}\right| \\ &\quad\quad + \operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\ &\quad\quad \times \left|\frac{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} - 1\right| \\ &\quad\quad \times \left|\frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}\right|. \end{aligned} \tag{E.6}$$

Let

$$\mathbf{G} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T & \widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\ \operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T & \operatorname{sign}(\boldsymbol{\beta}_T) \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \end{pmatrix},$$

and

$$\widehat{\mathbf{G}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T & \widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\ \operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T & \operatorname{sign}(\boldsymbol{\beta}_T) \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \end{pmatrix}.$$

Using [4, Th. 3], we compute the density of $\widehat{z}_a = \widehat{\mathbf{G}}_{12} \widehat{\mathbf{G}}_{22}^{-1}$ conditional on $\widehat{\boldsymbol{\mu}}_T$ and obtain that

$$\sqrt{\frac{n-s}{q}}\left(\frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} - \frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}\right) \mid \widehat{\boldsymbol{\mu}}_T \sim t_{n-s},$$

where

$$\begin{aligned} q &= \frac{\widehat{\boldsymbol{\mu}}_T' \left(\boldsymbol{\Sigma}_{TT}^{-1} - \frac{\boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1}}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}\right) \widehat{\boldsymbol{\mu}}_T}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} \\ &\leq \frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}. \end{aligned}$$

Lemma 14 gives

$$\begin{aligned}&\left|\frac{\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} - \frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}\right| \\ &\quad \leq C \sqrt{\frac{\widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \widehat{\boldsymbol{\mu}}_T}{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} \frac{\log\log(n)}{n}}, \end{aligned}$$

with probability at least $1 - \log^{-1}(n)$. Combining with Lemma 8, Lemma 5, and (E.3), we obtain an upper bound on the RHS of (E.6) as

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \widehat{\boldsymbol{\mu}}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad \leq C\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2} \vee ||\beta_T||_1\right) \\ &\quad\quad \times \sqrt{\frac{\log\log(n)}{n}} \end{aligned} \tag{E.7}$$

with probability at least $1 - \mathcal{O}(\log^{-1}(n))$.

The second term in (E.5) can be bounded using (E.3) with $\eta = \log^{-1}(n)$. Therefore, combining with (E.7), we obtain

$$\begin{aligned}&\left|\widehat{\boldsymbol{\mu}}_T' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \boldsymbol{\mu}_T' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)\right| \\ &\quad \leq C\left(\sqrt{\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \left(1 \vee ||\boldsymbol{\beta}_T||_{\boldsymbol{\Sigma}_{TT}}^2\right)} \vee ||\beta_T||_1\right) \\ &\quad\quad \times \sqrt{\frac{\log\log(n)}{n}} \end{aligned}$$

with probability at least $1 - \mathcal{O}(\log^{-1}(n))$, as desired. □

Lemma 8: There exist constants C_1, C_2, C_3, and C_4 such that each of the following inequalities hold with probability at least $1 - \log^{-1}(n)$:

$$\begin{aligned}&\mathbf{e}_a' \mathbf{S}_{TT}^{-1} \mathbf{e}_a \\ &\quad \leq C_1 \mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \mathbf{e}_a \left(1 + \mathcal{O}\left(\sqrt{\frac{\log(s\log(n))}{n}}\right)\right), \quad \forall a \in T \end{aligned} \tag{E.8}$$

$$\left|\frac{\mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \mathbf{e}_a}{\mathbf{e}_a' \mathbf{S}_{TT}^{-1} \mathbf{e}_a} - 1\right| \leq C_2 \sqrt{\frac{\log(s\log(n))}{n}}, \quad \forall a \in T \tag{E.9}$$

$$\left|\frac{\operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)}{\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)} - 1\right| \leq C_3 \sqrt{\frac{\log\log(n)}{n}}, \tag{E.10}$$

and

$$\begin{aligned}&\operatorname{sign}(\boldsymbol{\beta}_T)' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\ &\quad \leq C_4 \operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \left(1 + \mathcal{O}\left(\sqrt{\frac{\log\log(n)}{n}}\right)\right). \end{aligned} \tag{E.11}$$

Proof: [20, Th. 3.2.12] states that

$$(n-2) \frac{\mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \mathbf{e}_a}{\mathbf{e}_a' \mathbf{S}_{TT}^{-1} \mathbf{e}_a} \sim \chi_{n-s-1}^2.$$

Using Equation (F.2),

$$\left|\frac{n-2}{n-s-1} \frac{\mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \mathbf{e}_a}{\mathbf{e}_a' \mathbf{S}_{TT}^{-1} \mathbf{e}_a} - 1\right| \leq \sqrt{\frac{16 \log(2s\log(n))}{3(n-s-1)}}$$

with probability $1-(2s\log(n))^{-1}$. Rearranging terms in the display above, we have that

$$\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a \le C\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a\left(1+\mathcal{O}\left(\sqrt{\frac{\log(s\log(n))}{n}}\right)\right)$$

and

$$\left|\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-1\right| \le C\sqrt{\frac{\log(s\log(n))}{n}}.$$

A union bound gives (E.8) and (E.9).

Equations (E.10) and (E.11) are shown similarly. □

Lemma 9: There exist constants $C_1, C_2 > 0$ such that the following inequality

$$\begin{aligned}\forall a\in T:&\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\\ &\le C_1\sqrt{\left(\boldsymbol{\Sigma}_{TT}^{-1}\right)_{aa}\left(1\vee||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)\frac{\log(s\log(n))}{n}}\\ &\quad+C_2\left|\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\sqrt{\frac{\log(s\log(n))}{n}}\end{aligned}$$

holds with probability at least $1-\mathcal{O}(\log^{-1}(n))$.

Proof: Using the triangle inequality, we have

$$\begin{aligned}&\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\\ &\quad\le\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right|\\ &\qquad+\left|\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|.\end{aligned}\tag{E.12}$$

For the first term, we write

$$\begin{aligned}&\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right|\\ &\quad\le\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a\left|\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|\\ &\qquad+\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a\left|\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-1\right|\left|\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|.\end{aligned}\tag{E.13}$$

As in the proof of Lemma E.4, we can show that

$$\sqrt{\frac{n-s}{q_a}}\left(\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right)\mid\widehat{\boldsymbol{\mu}}_T\sim t_{n-s},$$

where

$$q_a=\frac{\widehat{\boldsymbol{\mu}}_T'\left(\boldsymbol{\Sigma}_{TT}^{-1}-\frac{\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right)\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\le\frac{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}.$$

Lemma 14 and an application of union bound gives

$$\begin{aligned}&\left|\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|\\ &\quad\le C\sqrt{\frac{\widehat{\boldsymbol{\mu}}_T'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\frac{\log(s\log(n))}{n}},\quad\forall a\in T,\end{aligned}$$

with probability at least $1-\mathcal{O}(\log^{-1}(n))$. Combining Lemma 5, Lemma 8 and Equation (E.2) with $\eta=\log^{-1}(n)$, we can bound the right hand side of (E.13) as

$$\begin{aligned}&\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T\right|\\ &\quad\le C_1\sqrt{\left(\boldsymbol{\Sigma}_{TT}^{-1}\right)_{aa}||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\frac{\log(s\log(n))}{n}}\\ &\qquad+C_2\left|\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\sqrt{\frac{\log(s\log(n))}{n}}\end{aligned}\tag{E.14}$$

with probability at least $1-\mathcal{O}(\log^{-1}(n))$.

The second term in (E.12) is handled by (E.2) with $\eta=\log^{-1}(n)$. Combining with (E.14), we obtain

$$\begin{aligned}&\left|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\widehat{\boldsymbol{\mu}}_T-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\\ &\quad\le C_1\sqrt{\left(\boldsymbol{\Sigma}_{TT}^{-1}\right)_{aa}\left(1\vee||\boldsymbol{\beta}_T||^2_{\boldsymbol{\Sigma}_{TT}}\right)\frac{\log(s\log(n))}{n}}\\ &\qquad+C_2\left|\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\mu}_T\right|\sqrt{\frac{\log(s\log(n))}{n}}\end{aligned}$$

with probability at least $1-\mathcal{O}(\log^{-1}(n))$. This completes the proof. □

Lemma 10: The probability of the event

$$\begin{aligned}\bigcap_{a\in[s]}\Bigg\{&|\mathbf{e}_a'(\mathbf{S}_{TT}^{-1}-\boldsymbol{\Sigma}_{TT}^{-1})\operatorname{sign}(\boldsymbol{\beta}_T)|\\ &\le C\left(\sqrt{(\boldsymbol{\Sigma}_{TT}^{-1})_{aa}\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT})s}\bigvee|\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)|\right)\\ &\quad\times\sqrt{\frac{\log(s\log(n))}{n}}\Bigg\}\end{aligned}$$

is at least $1-2\log^{-1}(n)$ for n sufficiently large.

Proof: Write

$$\begin{aligned}&|\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)-\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)|\\ &\quad\le\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a\left|\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|\\ &\qquad+\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a\left|\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-1\right|\left|\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|.\end{aligned}\tag{E.15}$$

As in the proof of Lemma E.4, we can show that

$$\sqrt{\frac{n-s}{q_a}}\left(\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right)\sim t_{n-s},$$

where

$$q_a=\frac{\operatorname{sign}(\boldsymbol{\beta}_T)'\left(\boldsymbol{\Sigma}_{TT}^{-1}-\frac{\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right)\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}.$$

Therefore,

$$\left|\frac{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\mathbf{S}_{TT}^{-1}\mathbf{e}_a}-\frac{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\operatorname{sign}(\boldsymbol{\beta}_T)}{\mathbf{e}_a'\boldsymbol{\Sigma}_{TT}^{-1}\mathbf{e}_a}\right|\le C\sqrt{q_a\frac{\log(s\log(n))}{n}}\tag{E.16}$$

with probability $1-(s\log(n))^{-1}$. Combining Lemma 8 and (E.16), we can bound the right hand side

of Equation (E.15) as

$$\begin{aligned}
&|\mathbf{e}_a' \mathbf{S}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) - \mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)| \\
&\quad \le C \left((\boldsymbol{\Sigma}_{TT}^{-1})_{aa} \sqrt{q_a} \vee |\mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)| \right) \times \sqrt{\frac{\log(s \log(n))}{n}} \\
&\quad \le C \left(\sqrt{(\boldsymbol{\Sigma}_{TT}^{-1})_{aa}^{-1} \Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s} \vee |\mathbf{e}_a' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)| \right) \\
&\qquad \times \sqrt{\frac{\log(s \log(n))}{n}}
\end{aligned}$$

where the second inequality follows from

$$\begin{aligned}
q_a &\le (\boldsymbol{\Sigma}_{TT}^{-1})_{aa}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T)' \boldsymbol{\Sigma}_{TT}^{-1} \operatorname{sign}(\boldsymbol{\beta}_T) \\
&\le (\boldsymbol{\Sigma}_{TT}^{-1})_{aa}^{-1} \Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s.
\end{aligned}$$

An application of the union bound gives the desired result. □

APPENDIX F
TAIL BOUNDS FOR CERTAIN RANDOM VARIABLES

In this section, we collect useful results on tail bounds of various random quantities used throughout the paper. We start by stating a lower and upper bound on the survival function of the standard normal random variable. Let $Z \sim \mathcal{N}(0, 1)$ be a standard normal random variable. Then for $t > 0$

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} \exp(-t^2/2) \\
&\qquad \le \mathbb{P}(Z > t) \le \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-t^2/2). \qquad \text{(F.1)}
\end{aligned}$$

Next, we collect results concerning tail bounds for central χ^2 random variables.

Lemma 11 [14]: Let $X \sim \chi_d^2$. For all $x \ge 0$,

$$\begin{aligned}
\mathbb{P}[X - d \ge 2\sqrt{dx} + 2x] &\le \exp(-x) \\
\mathbb{P}[X - d \le -2\sqrt{dx}] &\le \exp(-x).
\end{aligned}$$

Lemma 12 [12]: Let $X \sim \chi_d^2$, then

$$\mathbb{P}[|d^{-1}X - 1| \ge x] \le \exp(-\frac{3}{16} d x^2), \quad x \in [0, \frac{1}{2}). \quad \text{(F.2)}$$

The following result provides a tail bound for non-central χ^2 random variable with non-centrality parameter ν.

Lemma 13 [3]: Let $X \sim \chi_d^2(\nu)$, then for all $x > 0$

$$\mathbb{P}[X \ge (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x] \le \exp(-x) \qquad \text{(F.3)}$$

$$\mathbb{P}[X \le (d+\nu) - 2\sqrt{(d+2\nu)x}] \le \exp(-x). \quad \text{(F.4)}$$

The following Lemma gives a tail bound for a t-distributed random variable.

Lemma 14: Let X be a random variable distributed as

$$X \sim \sigma d^{-1/2} t_d,$$

where t_d denotes a t-distribution with d degrees of freedom. Then

$$|X| \le C \sqrt{\sigma^2 d^{-1} \log(4\eta^{-1})}$$

with probability at least $1 - \eta$.

Proof: Let $Y \sim \mathcal{N}(0, 1)$ and $Z \sim \chi_d^2$ be two independent random variables. Then X is equal in distribution to

$$\frac{\sigma d^{-1/2} Y}{\sqrt{d^{-1} Z}}.$$

Using (F.1),

$$|\sigma d^{-1/2} Y| \le \sigma d^{-1/2} \sqrt{\log(4\eta^{-1})}$$

with probability at least $1 - \eta/2$. (F.2) gives

$$d^{-1} X \ge 1 - \sqrt{\frac{16}{3d} \log(2\eta^{-1})}$$

with probability at least $1 - \eta/2$. Therefore, for sufficiently large d,

$$|X| \le C \sqrt{\sigma^2 d^{-1} \log(4\eta^{-1})}.$$

□

REFERENCES

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis* (Wiley Series in Probability and Statistics), 3rd ed. Hoboken, NJ, USA: Wiley, 2003.

[2] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[3] L. Birgé, "An alternative point of view on Lepski's method," in *State of the Art in Probability and Statistics* (IMS Lecture Notes—Monograph Series), vol. 36. Beachwood, OH, USA: Institute of Mathematical Statistics, 2001, pp. 113–133.

[4] T. Bodnar and Y. Okhrin, "Properties of the singular, inverse and generalized inverse partitioned Wishart distributions," *J. Multivariate Anal.*, vol. 99, no. 10, pp. 2389–2405, 2008.

[5] T. Cai, W. Liu, and X. Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *J. Amer. Statist. Assoc.*, vol. 106, no. 494, pp. 594–607, 2011.

[6] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.

[7] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (Applications of Mathematics), vol. 31. New York, NY, USA: Springer-Verlag, 1996.

[8] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.

[9] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[10] J. Fan, Y. Feng, and X. Tong, "A road to classification in high dimensional space: The regularized optimal affine discriminant," *J. Roy. Statist. Soc., Ser. B (Statistical Methodology)*, vol. 74, no. 4, pp. 745–771, 2012.

[11] F. Han, T. Zhao, and H. Liu, "CODA: High dimensional copula discriminant analysis," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 629–671, 2013.

[12] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. Amer. Statist. Assoc.*, vol. 104, no. 486, pp. 682–693, 2009.

[13] M. Kolar and H. Liu, "Feature selection in high-dimensional classification," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 329–337.

[14] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.

[15] Q. Mai and H. Zou, "A note on the connection and equivalence of three sparse linear discriminant analysis methods," *Technometrics*, vol. 55, no. 2, pp. 243–246, 2013.

[16] Q. Mai, H. Zou, and M. Yuan, "A direct approach to sparse discriminant analysis in ultra-high dimensions," *Biometrika*, vol. 99, no. 1, pp. 29–42, 2012.

[17] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Probability and Mathematical Statistics). London, U.K.: Academic, 1979.

[18] R. Mazumder, J. H. Friedman, and T. Hastie, "*SparseNet*: Coordinate descent with nonconvex penalties," *J. Amer. Statist. Assoc.*, vol. 106, no. 495, pp. 1125–1138, 2011.

[19] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.

[20] R. J. Muirhead, *Aspects of Multivariate Statistical Theory* (Wiley Series in Probability and Mathematical Statistics). New York, NY, USA: Wiley, 1982.
[21] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statist. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.
[22] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse linear discriminant analysis by thresholding for high dimensional data," *Ann. Statist.*, vol. 39, no. 2, pp. 1241–1265, 2011.
[23] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids, with applications to DNA microarrays," *Statist. Sci.*, vol. 18, no. 1, pp. 104–117, 2003.
[24] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics). New York, NY, USA: Springer-Verlag, 2009.
[25] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
[26] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[27] L. Wang, Y. Kim, and R. Li, "Calibrating nonconvex penalized regression in ultra-high dimension," *Ann. Statist.*, vol. 41, no. 5, pp. 2505–2536, 2013.
[28] S. Wang and J. Zhu, "Improved centroids estimation for the nearest shrunken centroid classifier," *Bioinformatics*, vol. 23, no. 8, pp. 972–979, 2007.
[29] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Ann. Statist.*, vol. 42, no. 6, pp. 2164–2201, 2014.
[30] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high dimensional problems," *J. Roy. Statist. Soc., Ser. B (Statistical Methodology)*, vol. 71, no. 3, pp. 615–636, 2009.
[31] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *J. Roy. Statist. Soc., Ser. B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.
[32] M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, and X. Lin, "Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection," *Bioinformatics*, vol. 25, no. 9, pp. 1145–1151, 2009.
[33] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.
[34] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statist. Sci.*, vol. 27, no. 4, pp. 576–593, 2012.
[35] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Nov. 2006.
[36] H. Zou, "The adaptive Lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

Mladen Kolar received a Ph.D. degree in Machine Learning from the School of Computer Science at Carnegie Mellon University.

He is currently an Assistant Professor of Econometrics and Statistics at The University of Chicago Booth School of Business. His research interests include statistical machine learning, high-dimensional statistics, and dynamic network modeling. His thesis has received honorable mentions in the ACM SIGKDD Dissertation award.

Han Liu received a joint Ph.D. degree in Machine Learning and Statistics from the Carnegie Mellon University, Pittsburgh, PA, USA in 2011.

He is currently an Assistant Professor of Statistical Machine Learning in the Department of Operations Research and Financial Engineering at Princeton University, Princeton, NJ. He is also an adjunct Professor in the Department of Biostatistics and Department of Computer Science at Johns Hopkins University. He built and is serving as the principal investigator of the Statistical Machine Learning (SMiLe) lab at Princeton University. His research interests include high dimensional semiparametric inference, statistical optimization, Big Data inferential analysis.