

---

# Feature Selection in High-Dimensional Classification

---

Mladen Kolar

MLADENK@CS.CMU.EDU

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15217, USA

Han Liu

HANLIU@PRINCETON.EDU

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

## Abstract

High-dimensional discriminant analysis is of fundamental importance in multivariate statistics. Existing theoretical results sharply characterize different procedures, providing sharp convergence results for the classification risk, as well as the  $\ell_2$  convergence results to the discriminative rule. However, sharp theoretical results for the problem of variable selection have not been established, even though model interpretation is of importance in many scientific domains. In this paper, we bridge this gap by providing sharp sufficient conditions for consistent variable selection using the ROAD estimator (Fan et al., 2010). Our results provide novel theoretical insights for the ROAD estimator. Sufficient conditions are complemented by the necessary information theoretic limits on variable selection in high-dimensional discriminant analysis. This complementary result also establishes optimality of the ROAD estimator for a certain family of problems.

## 1. Introduction

High-dimensional discriminant analysis plays an important role in multivariate statistics and machine learning. In a typical setting, a binary discriminant analysis problem can be formulated as follows: we observe a set of training data  $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$  independently drawn from a joint distribution of  $(\mathbf{X}, Y)$ , where  $\mathbf{X} \in \mathbb{R}^p$  and  $Y \in \{1, 2\}$ . Discriminant analysis aims at classifying the value of  $Y$  given a new data point  $\mathbf{x}$ . Let  $p_1(\mathbf{x})$  and  $p_2(\mathbf{x})$  be the density functions of  $\mathbf{X} | Y = 1$  and  $\mathbf{X} | Y = 2$ , and the prior probabilities

$\pi_1 = \mathbb{P}(Y = 1)$ ,  $\pi_2 = \mathbb{P}(Y = 2)$ . It is well known that the Bayes rule classifies a new data point  $\mathbf{x}$  to the second class if and only if

$$\log p_2(\mathbf{x}) - \log p_1(\mathbf{x}) + \log(\pi_2/\pi_1) > 0. \quad (1.1)$$

One of the most commonly used settings is the conditional Gaussian model, where

$$\mathbf{X}|Y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \quad \text{and} \quad \mathbf{X}|Y = 2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}). \quad (1.2)$$

Let  $\boldsymbol{\mu}_d = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ . The optimal classifier classifies a point to class 2 if and only if

$$(\mathbf{x} - \boldsymbol{\mu}_a)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d + \log(\pi_2/\pi_1) > 0.$$

For the above Gaussian discriminant analysis, Bickel & Levina (2004) show that the classical low dimensional normal-based linear discriminant analysis (LDA) is asymptotically equivalent to random guessing when the dimension  $p$  increases at a rate comparable to the sample size  $n$ . To handle this problem, we generally assume the discriminant direction  $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$  is sparse. In particular, it is assumed that  $\boldsymbol{\beta} = (\boldsymbol{\beta}'_T, \mathbf{0}')'$  for some set<sup>1</sup>  $T \subseteq [p]$ . A number of papers assume  $\boldsymbol{\Sigma} = \mathbf{I}$ , including the nearest shrunken centroids (Tibshirani et al., 2002; Wang & Zhu, 2007) and feature annealed independence rules (Fan & Fan, 2008). More recently, numerous alternative approaches have been proposed by taking more complex covariance matrix structures into consideration (Fan et al., 2010; Shao et al., 2011; Cai & Liu, 2011; Mai et al., 2012).

One particularly interesting proposal is the ROAD estimator (Regularized Optimal Affine Discriminant) due to Fan et al. (2010). Let  $\mathbf{S}$  and  $\hat{\boldsymbol{\mu}}_d$  be empirical estimators of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\mu}_d$ . The ROAD estimator is obtained by minimizing  $\mathbf{v}'\mathbf{S}\mathbf{v}$  with  $\mathbf{v}'\hat{\boldsymbol{\mu}}_d$  restricted to be a constant value, i.e.

$$\min_{\mathbf{v}} \frac{1}{2} \mathbf{v}'\mathbf{S}\mathbf{v} + \lambda \|\mathbf{v}\|_1 \quad \text{subject to } \mathbf{v}'\hat{\boldsymbol{\mu}}_d = 1. \quad (1.3)$$

---

<sup>1</sup>We use  $[p]$  to denote the set  $\{1, \dots, p\}$ .

Here  $\lambda$  is a regularization parameter, when  $\lambda = 0$ , the ROAD estimator reduces to be the classical Fisher's discriminant rule. Later, Cai & Liu (2011) proposed a different version of the sparse LDA, which tries to make  $\mathbf{v}$  close to the Bayes rule's linear term  $\Sigma^{-1}\boldsymbol{\mu}_d$  in the  $\ell_\infty$  norm, i.e.,

$$\min_{\mathbf{v}} \{ \|\mathbf{v}\|_1, \text{ subject to } \|\mathbf{S}\mathbf{v} - \hat{\boldsymbol{\mu}}_d\|_\infty \leq \lambda \}. \quad (1.4)$$

Equation (1.4) turns out to be a linear programming rule highly related to the Dantzig selector (Candes & Tao, 2007; Yuan, 2010; Cai et al., 2011). More recently, Mai et al. (2012) proposed a version of the sparse LDA based on an  $\ell_1$ -norm penalized least square formulation.

To avoid the curse of dimensionality, an  $\ell_1$  penalty is added in all three methods to encourage a sparsity pattern of  $\mathbf{v}$ , and hence nice theoretical properties can be obtained under certain regularity conditions. However, unlike the high dimensional regression settings where sharp theoretical results exist for prediction, estimation, and variable selection consistency, all existing theories for high discriminant analysis are either on estimation consistency or risk consistency, but not on variable selection consistency. The main reason is that analyzing the variable selection consistency of a high dimensional Gaussian discriminant analysis procedure requires us to sharply characterize the sampling distribution and tail behavior of the scaled discriminant direction  $\frac{\mathbf{S}_{TT}^{-1}\hat{\boldsymbol{\mu}}_{d,T}}{\hat{\boldsymbol{\mu}}'_{d,T}\mathbf{S}_{TT}^{-1}\hat{\boldsymbol{\mu}}_{d,T}}$ , which requires more careful theoretical analysis and new proof technique. Mai et al. (2012) provide a variable selection consistency result for their procedure, however, as we will show later, the scaling they obtained is not optimal.

In the current paper, we bridge the theoretical gap in understanding of variable selection in high-dimensional discriminant analysis. We provide a sharp analysis of the variable selection performance of the ROAD estimator. The proof technique is based on the characterization of the Karush-Kuhn-Tucker (KKT) conditions for the constrained optimization problem. Unlike the  $\ell_1$ -norm penalized least squares regression, which directly estimates the regression coefficients, the ROAD estimator is related to the scaled quantity  $\frac{\mathbf{S}_{TT}^{-1}\boldsymbol{\mu}_{d,T}}{\boldsymbol{\mu}'_{d,T}\mathbf{S}_{TT}^{-1}\boldsymbol{\mu}_{d,T}}$ , rather than the Bayes rule's direction  $\mathbf{S}_{TT}^{-1}\boldsymbol{\mu}_{d,T}$ , due to the equality constraint in the optimization problem. To sharply characterize the variable selection consistency, we carefully analyze the tail behavior of this scaled quantity by exploiting sophisticated multivariate analysis results. Sufficient conditions for the variable selection consistency of the ROAD estimator are complemented 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  $T$  by any procedure. Some of the main results of this paper are summarized below.

Let  $T = \{j : \beta_j \neq 0\}$  and  $N = [p] \setminus T$ . Denote  $s = |T|$ . We show that if the sample size

$$n > C \cdot \left( \max_{a \in N} \Sigma_{a|T} \right) \Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log(p - s), \quad (1.5)$$

where  $C$  is a universal constant,  $\Sigma_{a|T} = \Sigma_{aa} - \Sigma_{aT}\boldsymbol{\Sigma}_{TT}^{-1}\boldsymbol{\Sigma}_{Ta}$ , and  $\Lambda_{\min}(\boldsymbol{\Sigma})$  denotes the minimum eigenvalue of  $\boldsymbol{\Sigma}$ , then the estimated vector  $\hat{\boldsymbol{\beta}}$  has the same sparsity pattern as the true  $\boldsymbol{\beta}$ , thus the ROAD is variable selection consistent (or sparsistent). This result suggests that the discriminant analysis has a similar theoretical scaling as the regression setting. To show Eq. (1.5), we need the assumptions that  $\min_{j \in T} |\beta_j|$  is not too small and  $\|\boldsymbol{\Sigma}_{NT}\boldsymbol{\Sigma}_{TT}^{-1}\text{sign}(\boldsymbol{\beta}_T)\|_\infty \leq 1 - \alpha$  with  $\alpha \in (0, 1)$ . The latter assumption is the irrerepresentable condition, which takes a similar form as for the  $\ell_1$ -norm penalized least squares problem. Our analysis of information theoretic limitations reveals that if  $n < C_1\beta_{\min}^{-2} \log(p - s)$ , where  $\beta_{\min}$  is the magnitude of the smallest non-zero component of  $\boldsymbol{\beta}$ , then no procedure can reliably recover the set  $T$ . In particular, for the case where  $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$  has bounded  $\ell_2$  norm and  $\beta_{\min} \asymp s^{-1/2}$ , we establish that the ROAD estimator is optimal for the purpose of variable selection. An illustrative simulation demonstrates sharpness of our results.

The rest of this paper is organized as follows. In the next section, we introduce the notation. In Section 3, we characterize the solution to the population version of the ROAD estimator and outline the proof technique to be used to characterize the solution to the problem in Eq. (1.3). In Section 4, we derive sufficient conditions for the ROAD estimator to be sparsistent. An information theoretic lower bound is given in Section 5. Numerical simulations are provided in Section 6. We conclude the paper with some discussions in Section 7.

## 2. Notation

In this paper we denote  $[n]$  to be the set  $\{1, \dots, n\}$ . For any index set  $T \subseteq [p]$ , we denote  $\boldsymbol{\beta}_T$  to be the subvector containing the components 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  $\mathbf{A}_{TT}$  denotes a submatrix of  $\mathbf{A}$  with rows and columns indexed by  $T$ . For a vector  $\mathbf{a} \in \mathbb{R}^n$ , we denote  $\text{supp}(\mathbf{a}) = \{j :$

$a_j \neq 0\}$  the support set,  $\|\mathbf{a}\|_q$ ,  $q \in [1, \infty)$ , the  $\ell_q$ -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 symmetric matrix  $\mathbf{A} \in \mathbb{R}^{p \times p}$  we denote  $\Lambda_{\min}(\mathbf{A})$  and  $\Lambda_{\max}(\mathbf{A})$  the smallest and largest eigenvalues, respectively. We also use the weighted norm  $\|\mathbf{a}\|_{\mathbf{A}}^2 = \mathbf{a}'\mathbf{A}\mathbf{a}$  for a symmetric matrix  $\mathbf{A}$ . 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  $a_n b_n^{-1} \rightarrow 0$ .

### 3. ROAD: Population Version

In this section, we characterize the solution to the population version of the optimization problem in Eq. (1.3). That is, we characterize the solution  $\widehat{\mathbf{w}}$  to the optimization problem

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}' \Sigma \mathbf{w} + \lambda \|\mathbf{w}\|_1 \quad \text{subject to } \mathbf{w}' \boldsymbol{\mu}_d = 1. \quad (3.1)$$

In particular, we derive conditions under which the vector  $\widehat{\mathbf{w}}$  recovers the sparsity pattern of  $\boldsymbol{\beta}$ . Recall that  $T = \text{supp}(\boldsymbol{\beta})$  and  $N = [p] \setminus T$ . We have the following result. We have the following result.

**Theorem 1.** *Under the assumption that*

$$\frac{1 + \lambda \|\boldsymbol{\beta}_T\|_1}{\|\boldsymbol{\beta}_T\|_{\Sigma_{TT}}^2} \min_{a \in T} |\beta_a| > \lambda \|\Sigma_{TT}^{-1} \text{sign}(\boldsymbol{\beta}_T)\|_\infty \quad (3.2)$$

and that there exists a constant  $\alpha \in (0, 1]$  such that

$$\|\Sigma_{NT} \Sigma_{TT}^{-1} \text{sign}(\boldsymbol{\beta}_T)\|_\infty \leq 1 - \alpha, \quad (3.3)$$

we have  $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}'_T, \mathbf{0}')$  is the solution to the problem in Eq. (3.1), where

$$\widehat{\mathbf{w}}_T = \frac{1 + \lambda \|\boldsymbol{\beta}_T\|_1}{\|\boldsymbol{\beta}_T\|_{\Sigma_{TT}}^2} \boldsymbol{\beta}_T - \lambda \Sigma_{TT}^{-1} \text{sign}(\boldsymbol{\beta}_T). \quad (3.4)$$

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

Theorem 1 provides two conditions under which the solution to Eq. (3.1) recovers the support of  $\boldsymbol{\beta}$ . Eq. (3.2) is a condition on the smallest component of  $\boldsymbol{\beta}_T$ , as well as a condition on the tuning parameter  $\lambda$ . Define  $\beta_{\min}$  as

$$\beta_{\min} = \min_{a \in T} |\beta_a|. \quad (3.5)$$

Let  $\lambda = \lambda_0 \|\boldsymbol{\beta}_T\|_{\Sigma_{TT}}^{-2}$  for some  $\lambda_0$ . Then from Eq. (3.2) we observe that  $\widehat{\mathbf{w}}_T$  recovers the support as long as  $\beta_{\min} \geq \lambda_0 \|\Sigma_{TT}^{-1} \text{sign}(\boldsymbol{\beta}_T)\|_\infty$ . The second condition, given in Eq. (3.3), is related to the irrepressible condition commonly used in the analysis

of Lasso (Zou, 2006; Meinshausen & Bühlmann, 2006; Zhao & Yu, 2007; Wainwright, 2009). Theorem 1 provides an explicit form for the solution  $\widehat{\mathbf{w}}$ . It is clear that the ROAD optimization procedure estimates the scaled discriminant direction  $\|\boldsymbol{\beta}_T\|_{\Sigma_{TT}}^{-2} \boldsymbol{\beta}$ . The estimator  $\widehat{\mathbf{w}}$  is biased when  $\lambda \neq 0$ , but nevertheless it can recover the set  $T$  of non-zero components of  $\boldsymbol{\beta}$ .

Theorem 1 is proven by analyzing the Karush-Kuhn-Tucker (KKT) conditions for the optimization problem in Eq. (3.1). The KKT conditions are given as

$$\Sigma \widehat{\mathbf{w}} + \lambda \widehat{\mathbf{z}} + \widehat{\gamma} \boldsymbol{\mu}_d = \mathbf{0} \quad (3.6)$$

$$\widehat{\mathbf{w}}' \boldsymbol{\mu}_d = 1, \quad (3.7)$$

where  $\widehat{\gamma}$  is the Lagrange multiplier for the constraint and  $\widehat{\mathbf{z}} \in \partial \|\widehat{\mathbf{w}}\|_1$  is an element of the subdifferential. In what follows, we will construct a vector  $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}'_T, \mathbf{0}')$  that satisfies the KKT conditions and  $\text{sign}(\widehat{\mathbf{w}}_T) = \text{sign}(\boldsymbol{\beta}_T)$ .

The following lemma characterizes the vector  $\boldsymbol{\beta}$  and will be useful in analysis that follows.

**Lemma 2.** *Under the model in Eq. (1.2) with  $\boldsymbol{\beta} = \Sigma^{-1} \boldsymbol{\mu}_d = (\boldsymbol{\beta}'_T, \mathbf{0}')$ , we have that*

$$\boldsymbol{\mu}_{d,N} = \Sigma_{NT} \Sigma_{TT}^{-1} \boldsymbol{\mu}_{d,T} \quad (3.8)$$

and

$$\boldsymbol{\beta}_T = \Sigma_{TT}^{-1} \boldsymbol{\mu}_{d,T}. \quad (3.9)$$

The solution  $\widehat{\mathbf{w}}$  to Eq. (3.1) is constructed by considering an oracle optimization problem, where the solution is forced to be non-zero only on the set  $T$ , and then showing that it also satisfies the KKT conditions for the full problem. The following lemma characterizes the solution to the oracle optimization problem.

**Lemma 3.** *Let*

$$\widetilde{\mathbf{w}}_T = \arg \min_{\mathbf{w}_T : \mathbf{w}'_T \boldsymbol{\mu}_{d,T} = 1} \left\{ \frac{1}{2} \mathbf{w}'_T \Sigma_{TT} \mathbf{w}_T + \lambda \mathbf{w}_T \text{sign}(\boldsymbol{\beta}_T) \right\} \quad (3.10)$$

be the oracle optimization problem. Then

$$\widetilde{\mathbf{w}}_T = \frac{1 + \lambda \|\boldsymbol{\beta}_T\|_1}{\|\boldsymbol{\beta}_T\|_{\Sigma_{TT}}^2} \boldsymbol{\beta}_T - \lambda \Sigma_{TT}^{-1} \text{sign}(\boldsymbol{\beta}_T). \quad (3.11)$$

The next lemma shows that the vector  $(\widetilde{\mathbf{w}}'_T, \mathbf{0}')$  is the solution to the unconstrained optimization problem under the assumptions of Theorem 1.

**Lemma 4.** *Assume that conditions of Theorem 1 are satisfied. Then  $\widehat{\mathbf{w}} = (\widetilde{\mathbf{w}}'_T, \mathbf{0}')$  is the solution to the problem in Eq. (3.1), where  $\widetilde{\mathbf{w}}_T$  is defined in Eq. (3.10). Furthermore, we have  $\text{sign}(\widehat{\mathbf{w}}) = \text{sign}(\boldsymbol{\beta})$ .*

Theorem 1 follows directly from Lemma 3 and Lemma 4. In the next section, we will establish sufficient conditions for the ROAD procedure to recover the non-zero components of  $\beta$  when the population quantities in Eq. (3.1) are replaced by their empirical estimates. The proof construction is going to follow the same line of reasoning, however, proving analogous results to Lemma 3 and Lemma 4 in the sample version of the problem is much more challenging.

#### 4. Sparsistency Analysis of ROAD

In this section, we characterize the solution  $\hat{\mathbf{v}}$  to the optimization problem given in Eq. (1.3), which is a sample version of the optimization problem given in Eq. (3.1). We will derive conditions under which  $\hat{\mathbf{v}} = (\hat{\mathbf{v}}_T, \mathbf{0}')$  and  $\text{sign}(\hat{\mathbf{v}}_T) = \text{sign}(\beta_T)$ .

We observe  $n$  independent and identically distributed (iid) data points  $\{\mathbf{x}_i, y_i\}$  from the model in Eq. (1.2) with equal class probabilities, that is, with out loss of generality  $\pi_1 = \pi_2 = \frac{1}{2}$ . Denote  $n_1 = |\{i : y_i = 1\}|$  and  $n_2 = n - n_1$ . Let  $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times p}$  be the matrix with rows containing data points for which the label is one and similarly define  $\mathbf{X}_2 \in \mathbb{R}^{n_2 \times p}$ . Let

$$\mathbf{H}_1 = \mathbf{I}_{n_1} - n_1^{-1} \mathbf{1}_{n_1} \mathbf{1}'_{n_1} \quad \text{and} \quad \mathbf{H}_2 = \mathbf{I}_{n_2} - n_2^{-1} \mathbf{1}_{n_2} \mathbf{1}'_{n_2}$$

be the centering matrices, where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix and  $\mathbf{1}_n$  is the  $n \times 1$  vector with all components equal to 1. We define the following quantities

$$\begin{aligned} \hat{\boldsymbol{\mu}}_1 &= n_1^{-1} \sum_{i:y_i=1} \mathbf{x}_i = n_1^{-1} \mathbf{X}'_1 \mathbf{1}_{n_1}, \\ \hat{\boldsymbol{\mu}}_2 &= n_2^{-1} \sum_{i:y_i=2} \mathbf{x}_i = n_2^{-1} \mathbf{X}'_2 \mathbf{1}_{n_2}, \\ \hat{\boldsymbol{\mu}}_d &= \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1, \quad \hat{\boldsymbol{\mu}}_a = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2, \end{aligned}$$

$$\begin{aligned} \mathbf{S}_1 &= (n_1 - 1)^{-1} \mathbf{X}'_1 \mathbf{H}_1 \mathbf{X}_1, \quad \mathbf{S}_2 = (n_2 - 1)^{-1} \mathbf{X}'_2 \mathbf{H}_2 \mathbf{X}_2, \\ \mathbf{S} &= (n - 2)^{-1} ((n_1 - 1) \mathbf{S}_1 + (n_2 - 1) \mathbf{S}_2). \end{aligned}$$

The matrix  $\mathbf{S} \sim \mathcal{W}_p((n - 2)^{-1} \boldsymbol{\Sigma}, n - 2)$  is the pooled sample covariance matrix, where  $\mathcal{W}_p((n - 2)^{-1} \boldsymbol{\Sigma}, n - 2)$  denotes the Wishart distribution with  $n - 2$  degrees of freedom and the scaling matrix  $(n - 2)^{-1} \boldsymbol{\Sigma}$  (see Theorem 3.4.2 in [Mardia et al., 1980](#)). It is a standard result (see Theorem 3.1.2 in [Muirhead, 1982](#)) that  $\mathbf{S}$  is independent of  $\hat{\boldsymbol{\mu}}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, n_i^{-1} \boldsymbol{\Sigma})$ , ( $i = 1, 2$ ).

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

**Theorem 5.** *We assume that condition in Eq. (3.3) holds. Let the penalty parameter be  $\lambda = \lambda_0 \|\beta_T\|_{\boldsymbol{\Sigma}_{TT}}^{-2}$ ,*

with

$$\lambda_0 = C_0 \sqrt{\left( \max_{a \in N} \Sigma_{a|T} \right) (1 \vee \|\beta_T\|_{\boldsymbol{\Sigma}_{TT}}^2) \frac{\log(p - s)}{n}} \quad (4.1)$$

where  $C_0$  is a sufficiently large constant independent of the problem parameters and  $\Sigma_{a|T} = \Sigma_{aa} - \boldsymbol{\Sigma}_{aT} \boldsymbol{\Sigma}_{TT}^{-1} \boldsymbol{\Sigma}_{Ta}$ . Moreover, we assume that

$$\beta_{\min} \geq K \lambda_0 (\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) + \|\boldsymbol{\Sigma}_{TT}^{-1} \text{sign}(\beta_T)\|_{\infty}) \quad (4.2)$$

for some sufficiently large constant  $K$  independent of the problem parameters. If the sample size

$$n > C_1 \left( \max_{a \in N} \Sigma_{a|T} \right) \Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT}) s \log(p - s), \quad (4.3)$$

where  $C_1$  is a constant independent of the problem parameters, then  $\hat{\mathbf{v}} = (\hat{\mathbf{v}}_T, \mathbf{0}')$ , with

$$\hat{\mathbf{v}}_T = \frac{1 + \lambda \hat{\boldsymbol{\mu}}'_{d,T} \mathbf{S}_{TT}^{-1} \text{sign}(\beta_T)}{\hat{\boldsymbol{\mu}}'_{d,T} \mathbf{S}_{TT}^{-1} \hat{\boldsymbol{\mu}}_{d,T} - \lambda \mathbf{S}_{TT}^{-1} \text{sign}(\beta_T)} \mathbf{S}_{TT}^{-1} \hat{\boldsymbol{\mu}}_{d,T} - \lambda \mathbf{S}_{TT}^{-1} \text{sign}(\beta_T).$$

is the unique solution to the optimization problem in Eq. (1.3) with probability at least  $1 - \mathcal{O}(\log^{-1}(n))$ .

Theorem 5 characterizes the solution  $\hat{\mathbf{v}}$  obtained by solving the ROAD optimization problem in Eq. (1.3). It is a sample version of Theorem 1 given in the previous section. Again, we require a lower bound on  $\beta_{\min}$  in order to distinguish relevant components from the irrelevant ones and an irrerepresentable condition to hold. Theorem 5 provides a lower bound on the sample size  $n$  that is sufficient for the ROAD procedure to identify the set  $T$  with high probability. The sample size scales as  $n \asymp s \log(p - s)$  assuming that there exists a constant  $\underline{c}$  such that  $0 < \underline{c} \leq \Lambda_{\min}(\boldsymbol{\Sigma}_{TT})$ . This scaling is of the same order as for the Lasso procedure, where  $n > 2s \log(p - s)$  is needed for recovery of the relevant variables when  $\boldsymbol{\Sigma} = \mathbf{I}$ .

[Mai et al. \(2012\)](#) analyze an  $\ell_1$ -norm penalized least squares approach for solving the discriminant analysis problem. They require the sample size  $n$  to satisfy

$$\lim_{n \rightarrow \infty} \frac{s^2 \log p}{n} = 0,$$

which is suboptimal compared to our results. At this point, it is not clear how their analysis can be improved, but we conjecture some results developed in our current paper could be useful.

The proof of Theorem 5 is outlined in the next section.

##### 4.1. Proof of Theorem 5

The proof of Theorem 5 parallels the proof of Theorem 1. We construct the solution  $\hat{\mathbf{v}}$  to the optimization problem in Eq. (1.3) that recovers the sparsity

pattern of the vector  $\beta_T$ . To achieve that, we proceed in two steps. In the first step, we consider an oracle optimization problem (defined in Lemma 6), which is minimized at  $\tilde{\mathbf{v}}_T$ . In the second step, we show that the vector  $(\tilde{\mathbf{v}}_T', \mathbf{0}')$  satisfies the KKT conditions for the original optimization problem given in Eq. (1.3), thus showing that  $\hat{\mathbf{v}} = (\tilde{\mathbf{v}}_T', \mathbf{0}')$  is the global minimizer.

The following lemma characterizes the solution to the constrained optimization problem, which is analogous to the population version of the constrained optimization problem given in Eq. (3.10).

**Lemma 6.** *Let*

$$\tilde{\mathbf{v}}_T = \arg \min_{\mathbf{v}_T : \mathbf{v}_T' \hat{\boldsymbol{\mu}}_{d,T} = 1} \left\{ \frac{1}{2} \mathbf{v}_T' \mathbf{S}_{TT} \mathbf{v}_T + \lambda \mathbf{v}_T' \text{sign}(\beta_T) \right\} \quad (4.4)$$

be the oracle optimization problem. Then

$$\hat{\mathbf{v}}_T = \frac{1 + \lambda \hat{\boldsymbol{\mu}}_{d,T}' \mathbf{S}_{TT}^{-1} \text{sign}(\beta_T)}{\hat{\boldsymbol{\mu}}_{d,T}' \mathbf{S}_{TT}^{-1} \hat{\boldsymbol{\mu}}_{d,T}} \mathbf{S}_{TT}^{-1} \hat{\boldsymbol{\mu}}_{d,T} - \lambda \mathbf{S}_{TT}^{-1} \text{sign}(\beta_T). \quad (4.5)$$

Lemma 6 provides an explicit form for the solution of the oracle optimization problem. Note that the solution is unique, since the objective is strongly convex due to the quadratic term as  $\mathbf{S}_{TT}$  is positive definite with probability 1. The next result shows that, under the conditions of Theorem 5, the vector  $(\tilde{\mathbf{v}}_T', \mathbf{0}')$  is the solution to the ROAD optimization problem in Eq. (1.3).

**Lemma 7.** *Let the penalty parameter be  $\lambda = \lambda_0 \|\beta_T\|_{\Sigma_{TT}}^{-2}$ , with*

$$\lambda_0 = C_0 \sqrt{\left( \max_{a \in N} \Sigma_{a|T} \right) (1 \vee \|\beta_T\|_{\Sigma_{TT}}^2) \frac{\log(p-s)}{n}} \quad (4.6)$$

where  $C_0$  is a sufficiently large constant independent of the problem parameters and  $\Sigma_{a|T} = \Sigma_{aa} - \Sigma_{aT} \Sigma_{TT}^{-1} \Sigma_{Ta}$ . Assume that  $\text{sign}(\tilde{\mathbf{v}}_T) = \text{sign}(\beta_T)$  and  $\lambda \|\beta_T\|_1 < C_1$ , for some constant  $C_1$ . If the sample size

$$n > C_2 \left( \max_{a \in N} \Sigma_{a|T} \right) \Lambda_{\min}^{-1}(\Sigma_{TT}) s \log(p-s), \quad (4.7)$$

where  $C_2$  is a constant independent of the problem parameters, then  $\hat{\mathbf{v}} = (\tilde{\mathbf{v}}_T', \mathbf{0}')$  is the solution to the optimization problem in Eq. (1.3) with probability at least  $1 - \mathcal{O}(\log^{-1}(n))$ .

Finally, we need to show that  $\text{sign}(\tilde{\mathbf{v}}_T) = \text{sign}(\beta_T)$ .

**Lemma 8.** *Under the assumptions of Theorem 5,  $\tilde{\mathbf{v}}_T$  defined in Eq. (4.5) recovers the sign pattern of the vector  $\beta_T$  with probability at least  $1 - \mathcal{O}(\log^{-1}(n))$ .*

Under the assumption on  $\beta_{\min}$  in Eq. (4.2), conditions of Lemma 7 are satisfied and  $\text{sign}(\tilde{\mathbf{v}}_T) = \text{sign}(\beta_T)$ . This completes the proof of Theorem 5.

## 5. Lower bound

In this section, we are interested in results of complementary nature to those derived in Theorem 5. Theorem 5 provides sufficient conditions for a particular procedure to recover the support set  $T$  of non-zero elements of  $\beta$ . Here we discuss necessary conditions that must be satisfied for any method to succeed in reliable estimation of the support set.

Let  $\Psi$  be an estimator of  $T$ . We consider the maximum risk, corresponding to 0/1 loss, given as

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

where  $\theta = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$  denotes the problem parameters,  $\mathbb{P}_{\theta}$  denotes the joint law of  $\{\mathbf{x}_i, y_i\}_{i \in [n]}$  assuming  $\pi_1 = \pi_2 = \frac{1}{2}$ ,  $T(\theta) = \text{supp}(\beta)$  (recall that  $\beta = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ ), and  $\Theta$  is a family of parameters. Let  $\mathcal{M}(s, \mathcal{Z})$  be the class of all subsets of the set  $\mathcal{Z}$  of cardinality  $s$ . We consider

$$\begin{aligned} \Theta &= \Theta(\boldsymbol{\Sigma}, \tau, s) \\ &= \bigcup_{\omega \in \mathcal{M}(s, [p])} \left\{ \theta = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) : \begin{array}{l} \beta = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \\ |\beta_a| \geq \tau \text{ if } a \in \omega, \\ \beta_a = 0 \text{ if } a \notin \omega \end{array} \right\}. \end{aligned} \quad (5.1)$$

The minimax risk is defined as the smallest risk over all possible estimators. The main result of this section provides a lower bound on the minimax risk

$$\inf_{\Psi} R(\Psi, \Theta(\boldsymbol{\Sigma}, \tau_{\min}, s)),$$

where  $\tau_{\min} > 0$  determines the signal strength. Before stating the result, we introduce two quantities that will be used to state Theorem 9. We define

$$\begin{aligned} \varphi_{\text{close}}(\boldsymbol{\Sigma}) &= \min_{T \in \mathcal{M}(s, [p])} \min_{u \in T} \frac{1}{p-s} \sum_{v \in [p] \setminus T} (\Sigma_{uu} + \Sigma_{vv} - 2\Sigma_{uv}) \end{aligned} \quad (5.2)$$

and

$$\begin{aligned} \varphi_{\text{far}}(\boldsymbol{\Sigma}) &= \min_{T \in \mathcal{M}(s, [p])} \frac{1}{\binom{p-s}{s}} \sum_{T' \in \mathcal{M}(s, [p] \setminus T)} \mathbf{1}' \boldsymbol{\Sigma}_{T \cup T', T \cup T'} \mathbf{1}. \end{aligned} \quad (5.3)$$

The first quantity will be used to measure the difficulty of distinguishing two close support sets  $T_1$  and  $T_2$  that differ in only one position, while the second quantity measures the effect of a huge number of support sets that are far from the support set  $T$ .



**Theorem 9.** *Let*

$$\tau_{\min} = 2 \max \left( \sqrt{\frac{\log \binom{p-s}{s}}{n\varphi_{\text{far}}(\boldsymbol{\Sigma})}}, \sqrt{\frac{\log(p-s+1)}{n\varphi_{\text{close}}(\boldsymbol{\Sigma})}} \right). \quad (5.4)$$

If  $\tau < \tau_{\min}$ , there exists some constant  $C > 0$ , such that

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

The result can be interpreted in words in the following way: whatever the estimation procedure, when  $\tau < \tau_{\min}$  there exists some distribution indexed by  $\boldsymbol{\theta} \in \Theta(\boldsymbol{\Sigma}, \tau, s)$  such that the probability of incorrectly identifying the set  $T(\boldsymbol{\theta})$  is bounded away from zero.

**Remarks:**

1. Expressions  $\varphi_{\text{close}}(\boldsymbol{\Sigma})$  and  $\varphi_{\text{far}}(\boldsymbol{\Sigma})$  simplify greatly for the case when  $\boldsymbol{\Sigma} = \mathbf{I}$ . In particular, we have  $\varphi_{\text{close}}(\mathbf{I}) = 2$  and  $\varphi_{\text{far}}(\mathbf{I}) = 2s$ .
2. As a consequence of Theorem 9 and Theorem 5, we observe that the ROAD estimator is able to recover the set  $T$  using the optimal number of samples (up to an absolute constant) over the class of problems

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

where  $M$  is a fixed constant and  $\Lambda_{\min}^{-1}(\boldsymbol{\Sigma}_{TT})$  is bounded.

## 6. Simulation Results

In this section, we conduct a few illustrative simulations that show finite sample performance of our results. Theorem 5 describes the sample size needed 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  $s(p) = \lceil 2p^{0.45} \rceil$
2. sublinear sparsity, where  $s(p) = \lceil 0.4p / \log(0.4p) \rceil$ , and
3. linear sparsity, where  $s(p) = \lceil 0.4p \rceil$ .

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

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

where  $\theta$  is the control parameter varied in the interval  $[0.1, 4.5]$  and investigate how well can the ROAD procedure recover the support set  $T$ . 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  $|T| = s(p)$  randomly, and for each  $a \in T$  setting  $\mu_a$  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  $\lambda$  as

$$\lambda = 5 \|\boldsymbol{\beta}_T\|_{\boldsymbol{\Sigma}_{TT}}^{-2} \sqrt{(1 \vee \|\boldsymbol{\beta}_T\|_{\boldsymbol{\Sigma}_{TT}}^2) \frac{\log(p-s)}{n}}$$

for all cases. For this choice of  $\lambda$ , Theorem 5 predicts that the set  $T$  will be recovered correctly. For each setting, we report the Hamming distance between the estimated set  $\hat{T}$  and the true set  $T$ ,

$$h(\hat{T}, T) = |(\hat{T} \setminus T) \cup (T \setminus \hat{T})|,$$

averaged over 200 independent simulation runs.

Figure 1 plots the Hamming distance against the control parameter  $\theta$ , or the rescaled number of samples. Here the Hamming distance between  $\hat{T}$  and  $T$  is calculated by averaging 200 independent simulation runs. There are three subplots corresponding to different sparsity regimes (fractional power, sublinear and linear sparsity), each of them containing three curves for different problem sizes  $p \in \{100, 200, 300\}$ . Note that when the control parameter reaches  $\theta = 3$  almost all the elements of the set  $T$  are recovered, without false positives, while when  $\theta < 3$  the recovery is very poor. Figure 2 and Figure 3 show similar behavior for two other cases, with  $\boldsymbol{\Sigma}_{TT}$  being a Toeplitz matrix with parameter  $\rho = 0.1$  and the equal correlation matrix with  $\rho = 0.1$ .

## 7. 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

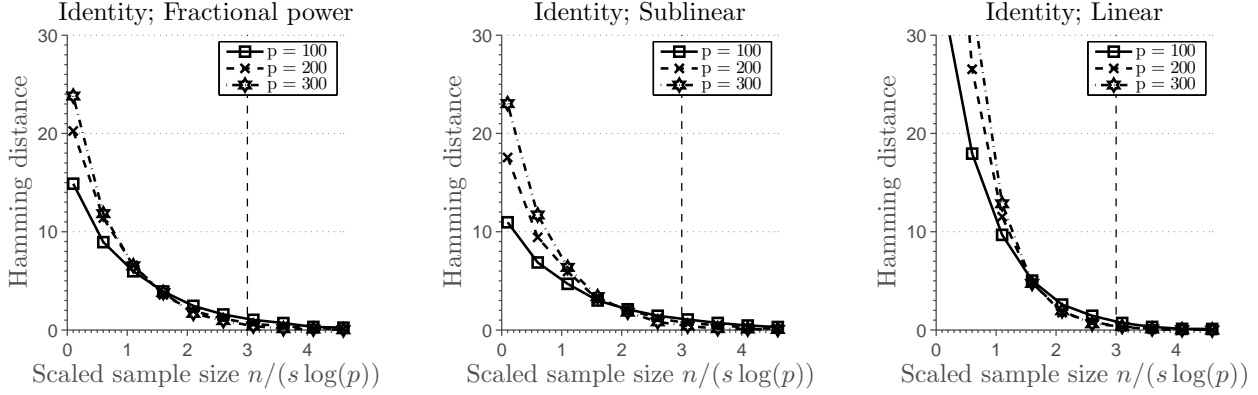


Figure 1. Plots of rescaled sample size  $n/(s \log(p))$  versus the Hamming distance between  $\hat{T}$  and  $T$  for identity covariance matrix  $\Sigma = \mathbf{I}_p$  (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes  $p \in \{100, 200, 300\}$ . The first subplot corresponds to the fractional power sparsity regime,  $s = 2p^{0.45}$ , the second subplot corresponds to the sublinear sparsity regime  $s = 0.4p/\log(0.4p)$ , and the third subplot corresponds to the linear sparsity regime  $s = 0.4p$ . For each scaling regime and problem size, we observe an empirical threshold behavior at  $n = 3s \log(p)$  (vertical line), showing that the result of Theorem 5 is sharp even in the finite sample studies and predicts the correct scaling for the sample size required to recover the set  $T$ .

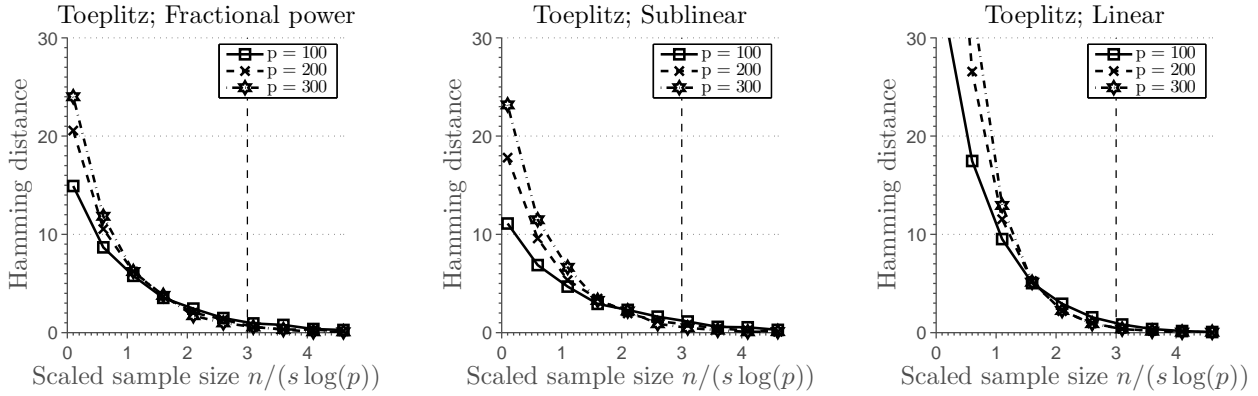


Figure 2. Plots of rescaled sample size  $n/(s \log(p))$  versus the Hamming distance between  $\hat{T}$  and  $T$  for Toeplitz covariance matrix  $\Sigma_{TT}$  with  $\rho = 0.1$  (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes  $p \in \{100, 200, 300\}$ . The first subplot corresponds to the fractional power sparsity regime,  $s = 2p^{0.45}$ , the second subplot corresponds to the sublinear sparsity regime  $s = 0.4p/\log(0.4p)$ , and the third subplot corresponds to the linear sparsity regime  $s = 0.4p$ . For each scaling regime and problem size, we observe an empirical threshold behavior at  $n = 3s \log(p)$  (vertical line), showing that the result of Theorem 5 is sharp even in the finite sample studies and predicts the correct scaling for the sample size required to recover the set  $T$ .

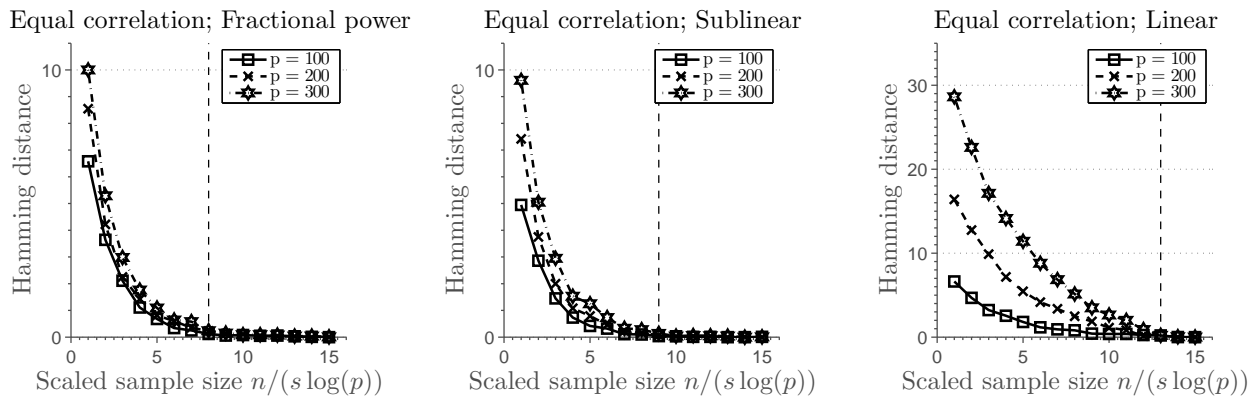


Figure 3. Plots of rescaled sample size  $n/(s \log(p))$  versus the Hamming distance between  $\hat{T}$  and  $T$  for equal correlation matrix  $\Sigma_{TT}$  with  $\rho = 0.1$  (averaged over 200 simulation runs). Each subfigure shows three curves, corresponding to the problem sizes  $p \in \{100, 200, 300\}$ . The first subplot corresponds to the fractional power sparsity regime,  $s = 2p^{0.45}$ , the second subplot corresponds to the sublinear sparsity regime  $s = 0.4p/\log(0.4p)$ , and the third subplot corresponds to the linear sparsity regime  $s = 0.4p$ . For each scaling regime and problem size, we observe an empirical threshold behavior at  $n = 3s \log(p)$  (vertical line), showing that the result of Theorem 5 is sharp even in the finite sample studies and predicts the correct scaling for the sample size required to recover the set  $T$ .

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 variable selection properties of the ROAD procedure and establishing sufficient conditions required for successful recovery of the set of relevant variables. 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 procedure ROAD is optimal.

## Acknowledgments

We thank anonymous reviewers who provided insightful comments that helped improve the paper. HL was supported by NSF Grant III-1116730.

## References

- Bickel, P. J. and Levina, E. Some theory for fisher’s linear discriminant function, ‘naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010, 2004.
- Cai, T. and Liu, W. A direct estimation approach to sparse linear discriminant analysis. *Arxiv preprint arXiv:1107.3442*, 2011.
- Cai, Tony, Liu, Weidong, and Luo, Xi. A constrained  $l_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.
- Candes, Emmanuel and Tao, Terence. The dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35:2313–2351, 2007.
- Fan, J. and Fan, Y. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.
- Fan, J., Feng, Y., and Tong, X. A road to classification in high dimensional space. *Arxiv preprint arXiv:1011.6095*, 2010.
- Mai, Q., Zou, H., and Yuan, M. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 2012.
- Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate analysis*. 1980.
- Meinshausen, N. and Bühlmann, P. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 2006.
- Muirhead, R.J. *Aspects of multivariate statistical theory*, volume 42. Wiley Online Library, 1982.
- Shao, J., Wang, Y., Deng, X., and Wang, S. Sparse linear discriminant analysis by thresholding for high dimensional data. *Arxiv preprint arXiv:1105.3561*, 2011.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567, 2002.



- Wainwright, Martin. Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009.
- Wang, S. and Zhu, J. Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8):972, 2007.
- Yuan, Ming. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.
- Zhao, P. and Yu, B. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.
- Zou, Hui. The adaptive lasso and its oracle properties. *Journal of American Statistical Association*, 101(476):1418–1429, 2006.