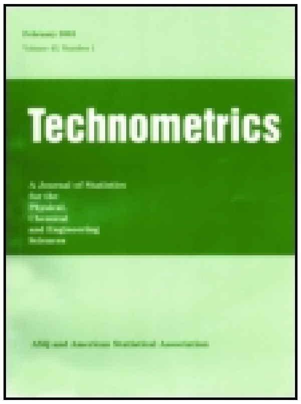

Technometrics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utch20

A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods

Qing Mai [a] & Hui Zou [a]

[a] School of Statistics, University of Minnesota, Minneapolis, Minnesota, 55455, U.S.A.

Accepted author version posted online: 28 Nov 2012.

To cite this article: Qing Mai & Hui Zou (2012): A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods, Technometrics, DOI:10.1080/00401706.2012.746208

To link to this article: http://dx.doi.org/10.1080/00401706.2012.746208

Disclaimer: This is a version of an unedited manuscript that has been accepted for publication. As a service to authors and researchers we are providing this version of the accepted manuscript (AM). Copyediting, typesetting, and review of the resulting proof will be undertaken on this manuscript before final publication of the Version of Record (VoR). During production and pre-press, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to this version also.

A Note On the Connection and Equivalence of Three Sparse Linear Discriminant Analysis Methods

Qing Mai and Hui Zou
School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
maixx034@umn.edu, zouxx019@umn.edu

October 19, 2012

Abstract

In this paper we reveal the connection and equivalence of three sparse linear discriminant analysis methods: the ℓ_1-Fisher's discriminant analysis proposed in Wu et al. (2008), the sparse optimal scoring proposed in Clemmensen et al. (2011) and the direct sparse discriminant analysis proposed in Mai et al. (2012). It is shown that, for any sequence of penalization parameters, the normalized solutions of direct sparse discriminant analysis equal the normalized solutions of the other two methods at different penalization parameters. A prostate cancer dataset is used to demonstrate the theory.

Keywords: Direct sparse discriminant analysis, ℓ_1-Fisher's discriminant analysis, Sparse optimal scoring.

1. INTRODUCTION

Consider a binary classification problem with data $(\mathbf{X}, Y)$ where $\mathbf{X}$ is an $n \times p$ matrix with each row X^i as a p-dimensional predictor, and $Y^i = 1, 2$ is the class label. $(X^i, Y^i)_{i=1}^n$ are independent observations. Denote n_1, n_2 as the within-group sample sizes. Linear discriminant analysis assumes that $X^i \mid Y^i = y \sim N(\mu_y, \Sigma)$. Then, for a new observation X_{new}, the Bayes rule takes the linear form that

$$\hat{Y}_{\mathrm{new}} = 1(X_{\mathrm{new}}^{\mathrm{T}}\beta + \beta_0 > 0) + 1, \tag{1}$$

where $\beta = \Sigma^{-1}(\mu_2 - \mu_1)$ is the discriminant direction. The classical linear discriminant analysis substitutes μ_y, Σ with their sample estimates, $\hat{\mu}_y, \hat{\Sigma}$. Linear discriminant analysis cannot be directly used for high-dimensional classification where p can be much larger than n, because the sample covariance estimator $\hat{\Sigma}$ will be singular. In recent years, significant efforts have been devoted to extending linear discriminant analysis to handle high-dimensional classification. Sparsity is the common theme in these proposals. Sparsity pursuit not only yields more interpretable classifiers but also improves the classification accuracy in the presence of many noise features.

The earliest proposals of sparse linear discriminant analysis are the nearest shrunken centroids classifier (PAM) (Tibshirani et al., 2002) and the features annealed independent rule (FAIR) (Fan and Fan, 2008). These two methods are based on the independence rules that ignore the correlation among features and treat them as if they were independent. The biggest advantage of such classifiers is that they are straightforward to implement. However, because the ignorance of correlation leads to model mis-specification, these methods may select the wrong set of variables and do not achieve the Bayes error rate as n tends to infinity (Mai et al., 2012). In recent years, there has been a sharp rise of interest in developing sparse LDA methods that respect the possible correlation structure between features. An incomplete list includes Trendafilov and Jolliffe (2007); Wu et al. (2008); Clemmensen et al. (2011); Mai et al. (2012); Witten and Tibshirani (2011); Shao et al. (2011); Cai and Liu (2011) and Fan et al. (2012). These methods are generally more reliable than the sparse classifiers based on independence rules. Rigorous theories have been established for the methods in Mai et al. (2012); Shao et al. (2011); Cai and Liu (2011) and Fan et al. (2012). There is no theoretical support for the ℓ_1-Fisher's discriminant analysis (FSDA) by Wu et al. (2008), the penalized classification using Fisher's linear discriminant by Witten and Tibshirani (2011) and the sparse optimal scoring (SOS) by Clemmensen et al. (2011).

In this paper we prove the equivalence of FSDA (Wu et al., 2008), the direct sparse discriminant analysis (DSDA) (Mai et al., 2012), and SOS (Clemmensen et al., 2011). These three sparse discriminant classifiers all have the form as in (1) with different ways to compute the classification coefficient vector β. We start with briefly introducing each proposal. We then present the main theorems concerning the connection and the equivalence between the methods. Finally, a prostate cancer dataset is used to demonstrate the theory. A direct consequence of the theory is that we can directly apply the theoretical results in Mai et al. (2012) to justify SOS and FSDA.

2. REVIEW OF THE THREE METHODS

FSDA solves the following constrained minimization problem

$$\hat{\beta}^{\mathrm{FSDA}}(\lambda) = \arg\min_{\beta} \beta^{\mathrm{T}} \left(\frac{n-2}{n} \hat{\boldsymbol{\Sigma}} \right) \beta + \lambda \|\beta\|_1, \qquad (\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} \beta = 1, \tag{2}$$

with $\lambda \geq 0$ being the ℓ_1 penalization parameter. This proposal is motivated by Fisher's view of linear discriminant analysis: the discriminant direction is obtained by maximizing $\beta^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_b \beta / \beta^{\mathrm{T}} \hat{\boldsymbol{\Sigma}} \beta$, where $\hat{\boldsymbol{\Sigma}}_b = (\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} (\hat{\mu}_2 - \hat{\mu}_1)$. Wu et al. (2008) made an observation that Fisher's problem can be reformulated as minimizing $\beta^{\mathrm{T}} \left((n-2)/n \hat{\boldsymbol{\Sigma}} \right) \beta$ subject to $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} \beta = 1$. They included the ℓ_1 penalty in (2) in order to encourage sparsity in $\hat{\beta}^{\mathrm{FSDA}}(\lambda)$. FSDA was also originally developed as an approach to testing a gene pathway. Wu et al. (2008) developed a solution path algorithm for computing $\hat{\beta}^{\mathrm{FSDA}}(\lambda)$.

Clemmensen et al. (2011) derived a sparse discriminant analysis algorithm by exploiting the connection between linear discriminant analysis and optimal scoring (Hastie et al., 1994). Suppose that the class label has K different values. Define an $n \times 2$ matrix $\mathbf{Y}^{\mathrm{dm}}$ of dummy variables, where $Y^{\mathrm{dm}}_{ik} = 1\{Y^i = k\}$ and θ is a 2-dimensional vector of scores. Let $\tilde{\mathbf{X}}$ denote the centered $\mathbf{X}$, i.e., $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{J}_{n \times n} \mathbf{X}$, where every entry of $\mathbf{J}$ is $1/n$. Then SOS solves for

$$\begin{aligned} \hat{\beta}^{\mathrm{SOS}}(\lambda) &= \arg\min_{\beta, \theta} \{ \|\mathbf{Y}^{\mathrm{dm}} \theta - \tilde{\mathbf{X}} \beta\|^2 + \lambda \|\beta\|_1 \}, \\ &\quad \frac{1}{n} \theta^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}} \mathbf{Y}^{\mathrm{dm}} \theta = 1, \theta^{\mathrm{T}} \mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}} \mathbf{Y}^{\mathrm{dm}} 1 = 0. \end{aligned}$$

SOS can deal with multi-class problems as well as binary problems, by defining a K-dimensional vector θ and an $n\times K$ matrix of dummy variables $\mathbf{Y}^{\mathrm{dm}}$. Clemmensen et al. (2011) used an alternating algorithm to solve SOS. When holding β fixed and optimizing with respect to θ, the problem is reduced to a generalized elastic net regression problem, which can be solved quickly by the algorithm proposed in Zou and Hastie (2005) or by the coordinate descent algorithm as in Friedman et al. (2010).

Mai et al. (2012) developed the direct sparse discriminant analysis by taking advantage of a least squares formulation of linear discriminant analysis. Let $y_i = -n/n_1$ if $Y^i = 1$ and $y_i = n/n_2$ if $Y^i = 2$. Define the solution to DSDA as follows

$$\hat{\beta}^{\mathrm{DSDA}}(\lambda) = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - X^i\beta)^2 + \lambda\|\beta\|_1.$$

Mai et al. (2012) showed that DSDA can recover the support of the Bayes rule and estimate the Bayes classifier direction with an overwhelming probability, even when the dimension grows with the sample size at a non-polynomial rate. DSDA is computationally most efficient among the three methods. One can solve $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$ for all values of λ using the `lars` algorithm (Efron et al., 2004) or solve $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$ for a fine grid values of λ using the coordinate descent algorithm (Friedman et al., 2010).

In what follows we view $\hat{\beta}^{\mathrm{FSDA}}(\lambda)$, $\hat{\beta}^{\mathrm{SOS}}(\lambda)$ and $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$ as functions of λ. We discover a close connection and even an equivalence between these functions.

3. THEORY

We first study the connection between $\hat{\beta}^{\mathrm{FSDA}}(\lambda)$ and $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$. Note that, by definition, $\hat{\beta}^{\mathrm{FSDA}}(\lambda)$ always satisfies the equality constraint in (2). Thus we consider a properly normalized $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$

defined as follows

$$\tilde{\beta}^{\mathrm{DSDA}}(\lambda) = \frac{\hat{\beta}^{\mathrm{DSDA}}(\lambda)}{c_1(\lambda)}, \quad \text{where } c_1(\lambda) = (\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\hat{\beta}^{\mathrm{DSDA}}(\lambda).$$

Theorem 1. *Given any fixed $\lambda > 0$, we have*

$$\tilde{\beta}^{\mathrm{DSDA}}(\lambda) = \hat{\beta}^{\mathrm{FSDA}}(\tilde{\lambda})$$

with $\tilde{\lambda} = \dfrac{\lambda}{n|c_1(\lambda)|}$.

Next we study the equivalence between the sparse optimal scoring and the direct sparse discriminant analysis.

Theorem 2. *Given any $\lambda > 0$, we have*

$$\hat{\beta}^{\mathrm{SOS}}(\lambda) = \sqrt{\hat{\pi}_1\hat{\pi}_2}\hat{\beta}^{\mathrm{DSDA}}\left(\frac{\lambda}{\sqrt{\hat{\pi}_1\hat{\pi}_1}}\right),$$

where $\hat{\pi}_1 = n_1/n, \hat{\pi}_2 = n_2/n$.

Theorems 1 and 2 can be used to provide strong theoretical support to the ℓ_1-Fisher's discriminant analysis and the sparse optimal scoring. Wu et al. (2008) and Clemmensen et al. (2011) provided numerical examples to demonstrate the efficacy of their proposals but there was no theoretical result to explain why their methods work well. In Mai et al. (2012) it has been shown that, under certain regularity conditions, if λ is some properly chosen value λ_n, then $\hat{\beta}^{\mathrm{DSDA}}(\lambda_n)$ consistently recovers the support of the Bayes rule and estimates the Bayes rule coefficient. By Theorems 1 and 2, the ℓ_1-Fisher's discriminant analysis with $\lambda = \lambda_n/(n|c_1(\lambda_n)|)$ and the sparse optimal scoring with $\lambda = \sqrt{\hat{\pi}_1\hat{\pi}_1}\lambda_n$ work as well as the Bayes rule asymptotically.

We would like to make a remark here that the above theorems are established for the binary classification setting. Binary classification has been the center of attention in the modern machine learning literature. For example, both support vector machines and boosting were first proposed for solving binary classification problems. On the other hand, multi-class classification problems can be very different than the binary case. FSDA and DSDA do not have a direct multi-class generalization. SOS was proposed to solve multi-class classification and cover binary classification as a special case. Currently we are not aware of a good multi-class generalization of FSDA or DSDA that would allow us to prove results like Theorems 1 and 2 for the multi-class setting.

4. A NUMERICAL EXAMPLE

Both Theorems 1 and 2 are exact finite sample results that hold for each given dataset. In this section we use the prostate cancer dataset (Singh et al., 2002; Dettling, 2004) to illustrate Theorems 1 and 2. This dataset contains the expression levels of 6033 genes, measured on 50 normal tissues and 52 prostate cancer tumors. We normalized the predictors such that each predictor has zero sample mean and unit sample variance. We took a fine grid of λ values and computed the corresponding $\hat{\beta}^{\mathrm{DSDA}}(\lambda)$. We then computed $\tilde{\lambda}$ from those λs using the formula $\tilde{\lambda} = \lambda/(n|c_1(\lambda)|)$. For each $\tilde{\lambda}$ we computed $\hat{\beta}^{\mathrm{FSDA}}(\tilde{\lambda})$ using the code of Dr. Wu. Figure 1 compares these two solutions and gives a graphical illustration of the equivalence result in Theorem 1. Numeric calculation confirms that the differences between the two panels of Figure 1 are indeed zero. Similarly, Figure 2 demonstrates the equivalence between the sparse optimal scoring and the direct sparse discriminant analysis. We took a fine grid of λ values and computed the corresponding $\hat{\beta}^{\mathrm{SOS}}(\lambda)$ by using the R package sparseLDA (Clemmensen, 2012). For each λ we then computed $\hat{\beta}^{\mathrm{DSDA}}$ at $\lambda/\sqrt{\hat{\pi}_1\hat{\pi}_2}$ and multiplied it by $\sqrt{\hat{\pi}_1\hat{\pi}_2}$. Once again, numeric calculations confirm exact equality of the two curves in Figure 2, demonstrating that Theorem 2 does indeed hold for this example.

5. PROOFS

Sparse optimal scoring works with centered predictors. Both the ℓ_1-Fisher's discriminant analysis and the direct sparse discriminant analysis use an intercept term in their formulation, which allows us to assume that X is centered without loss of generality.

Proof of Theorem 1. First, note the following facts

$$\sum_{k=1}^{2} \hat{\pi}_k \hat{\mu}_k \hat{\mu}_k^{\mathrm{T}} = \hat{\pi}_1 \hat{\pi}_2 (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}},$$

$$y^{\mathrm{T}}\mathbf{X} = 2n(\hat{\mu}_2 - \mu_1)^{\mathrm{T}}, \ \mathbf{X}^{\mathrm{T}}\mathbf{X} = (n-2)\hat{\mathbf{\Sigma}} + n\hat{\mathbf{\Sigma}}_b,$$

where $\hat{\mathbf{\Sigma}}_b = \hat{\pi}_1 \hat{\pi}_2 (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}$. Hence, we can write $\hat{\beta}^{\mathrm{DSDA}} = \arg\min L_3(\beta, \lambda)$, where

$$L_3(\beta, \lambda) = -2n(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\beta + (n-2)\beta^{\mathrm{T}}\hat{\mathbf{\Sigma}}\beta + n\beta^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\beta + \lambda\|\beta\|_1. \tag{3}$$

For notational convenience, write $c_1 = c_1(\lambda)$ and $\tilde{\beta} = \hat{\beta}^{\mathrm{DSDA}}(\lambda)/c_1(\lambda)$. Then $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\tilde{\beta} = 1$. Denote $L_1(\beta, \lambda) = \beta^{\mathrm{T}}[(n-2)/n\hat{\mathbf{\Sigma}}]\beta + \lambda\|\beta\|_1$. Let $\tilde{\lambda} = \lambda/(n|c_1|)$. Now it suffices to check that, for any $\tilde{\beta}'$ such that $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\tilde{\beta}' = 1$, we have

$$L_1(\tilde{\beta}', \tilde{\lambda}) \geq L_1(\tilde{\beta}, \tilde{\lambda}). \tag{4}$$

This is indeed true, because

$$\begin{aligned} L_3(c_1\tilde{\beta}', \lambda) &= -2nc_1(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\tilde{\beta}' + (n-2)c_1^2\tilde{\beta}'^{\mathrm{T}}\hat{\mathbf{\Sigma}}\tilde{\beta}' + nc_1^2\tilde{\beta}'^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\tilde{\beta}' + |c_1|\lambda\|\tilde{\beta}'\|_1, \\ &= -2nc_1 + nc_1^2 + nc_1^2[\tilde{\beta}'^{\mathrm{T}}[(n-2)/n\hat{\mathbf{\Sigma}}]\tilde{\beta}' + \tilde{\lambda}\|\tilde{\beta}'\|_1], \\ &= -2nc_1 + nc_1^2 + nc_1^2 L_1(\tilde{\beta}', \tilde{\lambda}), \end{aligned}$$

which yields

$$L_1(\tilde{\beta}', \tilde{\lambda}) = \frac{1}{nc_1^2}[L_3(c_1\tilde{\beta}', \lambda) + 2nc_1 - nc_1^2]. \tag{5}$$

Similarly,

$$L_1(\tilde{\beta}, \tilde{\lambda}) = \frac{1}{nc_1^2}[L_3(c_1\tilde{\beta}, \lambda) + 2nc_1 - nc_1^2]. \tag{6}$$

Because $\hat{\beta}^{\mathrm{DSDA}}(\lambda) = c_1\tilde{\beta}$ minimizes $L_3(\beta, \lambda)$, we have $L_3(c_1\tilde{\beta}, \lambda) \leq L_3(c_1\tilde{\beta}', \lambda)$. Combine this fact with (5)–(6) and we have (4). □

Proof of Theorem 2. For convenience, write $\hat{\beta}^{\mathrm{SOS}} = \hat{\beta}^{\mathrm{SOS}}(\lambda), \hat{\beta}^{\mathrm{DSDA}} = \hat{\beta}^{\mathrm{DSDA}}(\lambda)$. It is easy to check that, if $(\hat{\theta}, \hat{\beta}^{\mathrm{SOS}})$ is a solution to SOS, then $(-\hat{\theta}, -\hat{\beta}^{\mathrm{SOS}})$ is also a solution. Therefore, we restrict our attention to $\{\beta : (\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\beta > 0\}$. Clemmensen et al. (2011) show that $\hat{\theta}(\beta) = c_2\tilde{\theta}$, where

$$\tilde{\theta} = (\mathbf{I} - 11^{\mathrm{T}}\mathbf{D}_\pi)\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\beta, \mathbf{D}_\pi = \frac{1}{n}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}, c_2 = \frac{1}{\sqrt{\frac{1}{n}\tilde{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}\tilde{\theta}}}.$$

Note that

$$\mathbf{I} - 11^{\mathrm{T}}\mathbf{D}_\pi = \begin{pmatrix} \hat{\pi}_2 & -\hat{\pi}_2 \\ -\hat{\pi}_1 & \hat{\pi}_1 \end{pmatrix}, \mathbf{D}_\pi^{-1}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X} = n(\hat{\mu}_1^{\mathrm{T}}, \hat{\mu}_2^{\mathrm{T}}).$$

Therefore, $\tilde{\theta}^{\mathrm{T}}Y^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\beta = n^2\beta^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\beta$ and $\tilde{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}\tilde{\theta} = n^3\beta^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\beta$. It follows that

$$\hat{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\beta = n\sqrt{\beta^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\beta} = n\sqrt{\hat{\pi}_1\hat{\pi}_2}(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\beta.$$

So $\hat{\beta}^{\mathrm{SOS}} = \arg\min_\beta L_2(\beta, \lambda)$, where

$$L_2(\beta, \lambda) = -2n(\hat{\pi}_1\hat{\pi}_2)^{1/2}(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\beta + \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta + \lambda\|\beta\|_1. \tag{7}$$

Now, for any β, define $\beta' = \beta/\sqrt{\hat{\pi}_1\hat{\pi}_2}$. Compare (7) with (3) and it is easy to see that

$$L_2(\beta, \lambda) = (\hat{\pi}_1\hat{\pi}_2)L_3(\beta', \frac{\lambda}{\sqrt{\hat{\pi}_1\hat{\pi}_2}}). \tag{8}$$

By (8) and the definition of $\hat{\beta}^{\mathrm{DSDA}}$, we have the desired conclusion. □

ACKNOWLEDGEMENT

The authors think the editor, AE and two referees for their helpful comments and suggestions. This work is supported in part by NSF grant DMS-08-46068.

References

Cai, T. and Liu, W. (2011), ‘A direct estimation approach to sparse linear discriminant analysis’, *J. Am. Statist. Assoc.* **106**, 1566–1577.

Clemmensen, L. (2012), *sparseLDA: Sparse Discriminant Analysis*. R package version 0.1-6. **URL:** *http://CRAN.R-project.org/package=sparseLDA*

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011), ‘Sparse discriminant analysis’, *Technometrics* **53**, 406–413.

Dettling, M. (2004), ‘Bagboosting for tumor classification with gene expression data’, *Bioinformatics* **20**, 3583–3593.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), ‘Least angle regression’, *Ann. Statist.* **32**, 407–499.

Fan, J. and Fan, Y. (2008), ‘High dimensional classification using features annealed independence rules’, *Ann. Statist.* **36**, 2605–2637.

Fan, J., Feng, Y. and Tong, X. (2012), ‘A ROAD to classification in high dimensional space’, *J. R. Statist. Soc. B* **74**, 745–771.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), ‘Regularization paths for generalized linear models via coordinate descent’, *J. Stat. Software* **33**, 1–22.

Hastie, T., Tibshirani, R. and Buja, A. (1994), ‘Flexible discriminant analysis by optimal scoring’, *J. Am. Statist. Assoc.* **89**, 1255–1270.

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

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011), ‘Sparse linear discriminant analysis with high dimensional data’, *Ann. Statist.* **39**, 1241–1265.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. and Sellers, W. (2002), ‘Gene expression correlates of clinical prostate cancer behavior’, *Cancer Cell* **1**(2), 203–209.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002), ‘Diagnosis of multiple cancer types by shrunken centroids of gene expression’, *Proceedings of the National Academy of Sciences* **99**, 6567–6572.

Trendafilov, N. T. and Jolliffe, I. T. (2007), ‘DALASS: Variable selection in discriminant analysis via the lasso’, *Comput. Statist. Data Anal.* **51**, 3718–3736.

Witten, D. and Tibshirani, R. (2011), ‘Penalized classification using Fisher’s linear discriminant’, *J. R. Statist. Soc. B* **73**(5), 753–772.

Wu, M., Zhang, L., Wang, Z., Christiani, D. and Lin, X. (2008), ‘Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection’, *Bioinformatics* **25**, 1145–1151.

Zou, H. and Hastie, T. (2005), ‘Regularization and variable selection via the elastic net’, *Journal of Royal Statistical Society, Series B* **67**(2), 301–320.

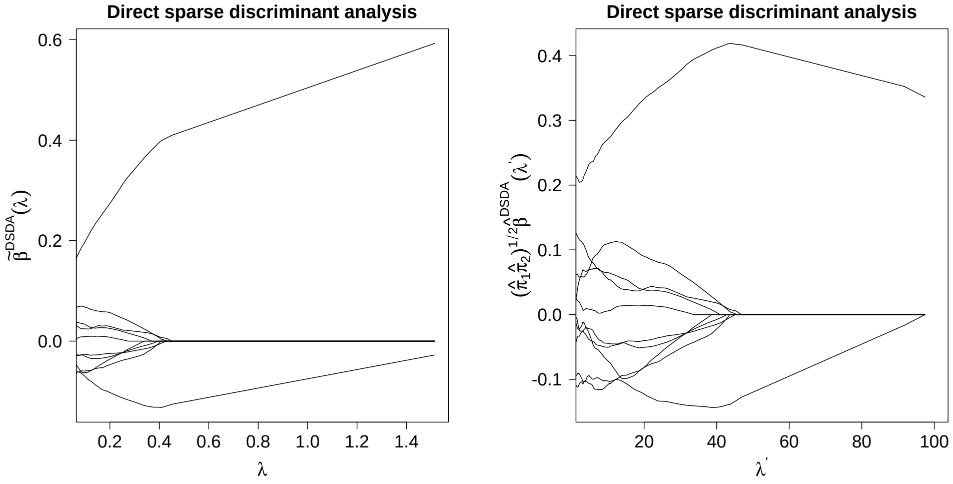

Figure 1: Demonstration of Theorem 1 with the prostate cancer data. We have computed 6033 coefficient curves but only show 10 curves here for ease of presentation. $\tilde{\lambda} = \lambda/(n|c_1(\lambda)|)$.

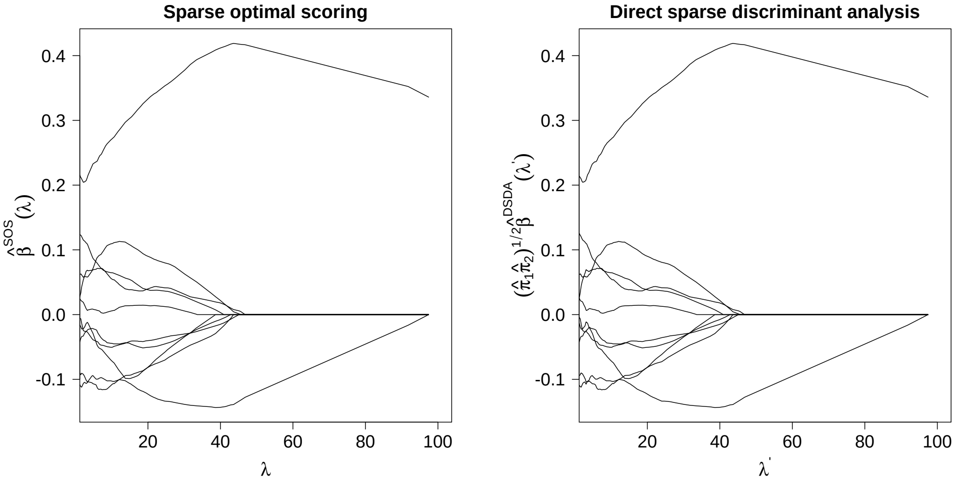

Figure 2: Demonstration of Theorem 2 with the prostate cancer data. We have computed 6033 coefficient curves but only show 10 curves here for ease of presentation. $\lambda' = \lambda/\sqrt{\hat{\pi}_1\hat{\pi}_2}$.