Biometrika (2015), **102**, 1, *pp.* 33–45
Printed in Great Britain

doi: 10.1093/biomet/asu049
Advance Access publication 3 December 2014

Covariance-enhanced discriminant analysis

By PEIRONG XU

Department of Mathematics, Southeast University, Nanjing 211189, China
xupeirong@seu.edu.cn

JI ZHU

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
jizhu@umich.edu

LIXING ZHU

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
lzhu@hkbu.edu.hk

and YI LI

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
yili@umich.edu

Summary

Linear discriminant analysis has been widely used to characterize or separate multiple classes via linear combinations of features. However, the high dimensionality of features from modern biological experiments defies traditional discriminant analysis techniques. Possible interfeature correlations present additional challenges and are often underused in modelling. In this paper, by incorporating possible interfeature correlations, we propose a covariance-enhanced discriminant analysis method that simultaneously and consistently selects informative features and identifies the corresponding discriminable classes. Under mild regularity conditions, we show that the method can achieve consistent parameter estimation and model selection, and can attain an asymptotically optimal misclassification rate. Extensive simulations have verified the utility of the method, which we apply to a renal transplantation trial.

Some key words: Correlation; Graphical lasso; Linear discriminant analysis; Pairwise fusion; Variable selection.

1. Introduction

Rapid technological advances have yielded vast amounts of high-throughput data, such as those arising from microarrays or proteomics, which has brought a high demand for statistical methods that can effectively use such data to make decisions. For example, in the kidney transplant and injury study (Flencher et al., 2004) that motivated this paper, 62 tissue samples were obtained from subjects with four different renal functional types after kidney transplantation. Distinguishing these four types of kidney tissue based on 12 625 gene expression profiles is crucial to balancing, at the molecular level, the need for immunosuppression to prevent transplant rejection while minimizing drug-induced toxicities. Linear discriminant analysis, a popular method in the classical setting where the number of variables is much smaller than the sample size, has been found to perform poorly in the high-dimensional setting because (a) the sample covariance matrix, which is needed in linear discriminant analysis, is singular; and (b) the classification rule involves a linear combination of all the variables, causing difficulty in interpretation and degrading classification performance with many noninformative variables.

To address (a), linear discriminant methods with a variety of penalized versions of covariance matrices have been developed, including the nearest shrunken centroids method assuming a diagonal covariance matrix (Tibshirani et al., 2002), naive Bayes using the diagonal of the sample covariance matrix (Bickel & Levina, 2004), an extension of nearest shrunken centroids with a general covariance matrix (Guo et al., 2007), thresholding of mean effects and the covariance matrix in binary classification (Shao et al., 2011), and a lasso-type classifier (Tibshirani, 1996) based on the estimated product of mean effects and the precision matrix (Cai & Liu, 2011). Other relevant works are Qiao et al. (2008), Witten & Tibshirani (2009), Clemmensen et al. (2011), Witten & Tibshirani (2011), Fan et al. (2012) and some of the references therein.

To address (b), Tibshirani et al. (2002) proposed shrinking the class centroids towards the global centroid,Wang & Zhu (2007) represented the problem as a lasso regression and introduced two new penalties to improve the effectiveness of variable selection, and Guo (2010) used a linear discriminant with pairwise fusion penalties to select informative variables. Theoretical properties are in general elusive for these methods, though some asymptotic results are available for the annealed independence rule proposed by Fan & Fan (2008) and a linear discriminant rule using penalized sparse least squares proposed by Mai et al. (2012). However, both of the latter methods focus on binary classification, and it is not clear how to extend them to the multiple-class case.

In this paper, we propose a covariance-enhanced discriminant analysis method for high-dimensional classification. Our method utilizes the general covariance structure, going beyond the diagonality restriction, in selecting informative variables for linear discriminant analysis. Our method achieves more flexibility than existing methods by allowing a variable to be informative for only a subset of, rather than all, the classes, and it enjoys consistency of parameter estimation and model selection. For binary classification, we show that it achieves the lowest possible asymptotic misclassification rate.

Other authors, including Clemmensen et al. (2011) and Witten & Tibshirani (2011), have also discussed variable selection in the presence of correlation. However, to our knowledge, none of these approaches can identify variables that are specifically informative for discriminating certain classes.

To further illustrate the impact of a nondiagonal covariance matrix on variable selection, Fig. 1 shows a simple binary classification example, wherein the two classes have the same mean in X_2 but different means in X_1. The best classifier would involve both X_1 and X_2, even though the latter does not by itself have any power to separate the two classes. The contribution of X_2 to classification is through its correlation with X_1, which demonstrates the role of using a nondiagonal covariance matrix. As Fig. 1 implies, for the purpose of classification and variable selection, we should consider the differences in the means between each pair of classes as well as possible intervariable correlations.

2. Methodology

2·1. *Model and notation*

Consider a general K-class problem, where Y is the class label taking values in $\{1, \ldots, K\}$ and X is the corresponding p-dimensional vector of predictors. We assume that the population-average probability of class k is $\omega_k = \text{pr}(Y = k) > 0$ for $k = 1, \ldots, K$ and satisfies $\sum_{k=1}^{K} \omega_k = 1$.

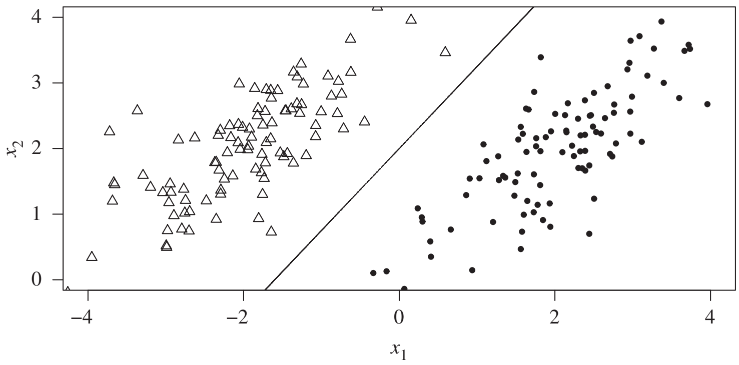

Fig. 1. An illustrative example involving two classes. Even though the two classes have the same mean in X_2, the variable X_2 is informative for classification and variable selection and should not be removed by a variable selection method.

The conditional density of X given class k is modelled by a multivariate Gaussian distribution, $X \mid Y = k \sim N_p(\mu_k, \Sigma)$, where $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^{\mathrm{T}}$ is the class-specific mean vector and Σ is a $p \times p$ positive-definite covariance matrix with (j, j')th element denoted by $\sigma_{jj'}$ $(j, j' = 1, \ldots, p)$. As assumed in linear discriminant analysis, the covariance matrix Σ is constant across different classes; this assumption may be plausible as, for example, gene expressions across disease classes often differ in their means rather than in the covariance structure (Guo et al., 2010).

Let $\omega = (\omega_1, \ldots, \omega_K)^{\mathrm{T}}$ and let Ω be the inverse of Σ with (j, j')th element written as $\Omega_{jj'}$ $(j, j' = 1, \ldots, p)$. Further, let $\mu = (\mu_1^{\mathrm{T}}, \ldots, \mu_K^{\mathrm{T}})^{\mathrm{T}}$ be the vector containing all class means and let $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ be an observation.

Given ω_k, μ_k $(k = 1, \ldots, K)$ and Ω, linear discriminant analysis classifies an observation x to a class, say k^*, that maximizes

$$\mathrm{pr}(Y = k \mid X = x) = c(x)\omega_k \exp\left\{-\frac{1}{2}(x - \mu_k)^{\mathrm{T}}\Omega(x - \mu_k)\right\},$$

where $c(x)$ is a normalizing constant that does not depend on k. For variable selection, we compare classes k and l, where $k, l = 1, \ldots, K$ with $k \neq l$. Specifically, we consider the pairwise difference for $k \neq l$,

$$\begin{aligned}\log \mathrm{pr}(Y = k \mid X = x) - \log \mathrm{pr}(Y = l \mid X = x) = \log \omega_k - \log \omega_l \\ - \frac{1}{2}\sum_{j=1}^{p}\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{kj} + \mu_{lj})(\mu_{kj'} - \mu_{lj'}) \\ + \sum_{j=1}^{p} x_j \left\{\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{kj'} - \mu_{lj'})\right\}.\end{aligned}$$

Hence, a necessary and sufficient condition for variable j to be noninformative in distinguishing classes k and l is that

$$\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{kj'} - \mu_{lj'}) = 0. \tag{1}$$

Further, we note that a sufficient condition leading to (1) is, for $j' = 1, \ldots, p$,

$$\begin{cases} \Omega_{jj'} = 0 \ \text{ or } \ \mu_{kj'} - \mu_{lj'} = 0, & j' \neq j, \\ \mu_{kj} - \mu_{lj} = 0, & j' = j. \end{cases} \tag{2}$$

Since $\Omega_{jj'} = 0$ indicates conditional independence between X_j and $X_{j'}$ given all the other variables, (2) implies that if a variable is conditionally independent of all the variables helpful for discriminating classes k and l, and is itself indistinguishable for classes k and l, then it is noninformative for discriminating classes k and l. Compared with the necessary and sufficient condition (1), the informativeness of features as defined by (2) is more interpretable in practice, as it elucidates why a given variable, say j in (2), is noninformative for discriminating classes k and l in terms of mean and in the presence of correlation. This motivates us to construct a variable selection procedure for selecting informative variables and identifying the distinguishable classes simultaneously.

2·2. *Covariance-enhanced discriminant analysis*

Let (y_i, x_i) be the ith observation $(i = 1, \ldots, n)$ from a K-class problem with known class label y_i and predictor vector x_i. Let $S(\mu) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i = k)(x_i - \mu_k)(x_i - \mu_k)^{\mathrm{T}}$. A natural approach to inference is to maximize the loglikelihood function

$$l_n(\omega, \mu, \Omega) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i = k) \log \omega_k + \frac{1}{2} \log |\Omega| - \frac{1}{2} \mathrm{tr}\{S(\mu)\Omega\};$$

however, with high-dimensional parameters μ and Ω, a direct maximization is not stable. Regularization terms on μ and Ω are needed to enhance stability.

Motivated by condition (2), we propose to regularize the pairwise differences in class centroids for each variable and the off-diagonal elements of the concentration matrix. Specifically, let $p = p_n$ be a function of the sample size n. We maximize

$$Q_n(\omega, \mu, \Omega) = l_n(\omega, \mu, \Omega) - \lambda_{1n} \sum_{j=1}^{p_n} \sum_{1 \leqslant k < l \leqslant K} |\mu_{kj} - \mu_{lj}| - \lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}| \tag{3}$$

subject to

$$\sum_{k=1}^{K} \omega_k = 1, \quad \Omega \succ 0, \tag{4}$$

where $\succ 0$ indicates positive definiteness. The first penalty term in (3) shrinks the pairwise differences in class centroids for each variable, whereas the second penalty term resembles that of the graphical lasso for estimating the concentration matrix (Yuan & Lin, 2007; Friedman et al., 2008). When the tuning parameters, λ_{1n} and λ_{2n}, are large enough, some of the $\mu_{kj} - \mu_{lj}$ and $\Omega_{jj'}$ will be estimated as zero. Further, if for some $k \neq l$ we have

$$\sum_{j'=1}^{p_n} \hat{\Omega}_{jj'}(\hat{\mu}_{kj'} - \hat{\mu}_{lj'}) = 0, \tag{5}$$

then variable j can be considered noninformative for distinguishing classes k and l, though it could still be informative for discriminating other pairs of classes. Moreover, if (5) holds for all pairs (k, l) with $k, l = 1, \ldots, K$ and $k < l$, then variable j is considered to make no contribution to the classification and can be removed from the fitted model.

Remark 1. While the proposed method using (3) and (4) does not directly enforce the structure described by (2), and the double penalization may somewhat bias the results, we choose to use (3) and (4) for two reasons. First, directly using (2) would lead to a complicated nonconvex problem. Second, the second penalty on (3) effectively enforces sparsity on Ω, which seems a reasonable assumption for large precision matrices (see, e.g., Bickel & Levina, 2008; Friedman et al., 2008; Lam & Fan, 2009; Cai et al., 2011; Witten et al., 2011) and can often simplify computation and interpretation.

One natural variant of the proposed method is the doubly l_1-penalized linear discriminant,

$$\max_{\omega,\mu,\Omega} l_n(\omega, \mu, \Omega) - \lambda_{1n} \sum_{j=1}^{p_n} \sum_{k=1}^{K} |\mu_{kj}| - \lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}|, \tag{6}$$

under the constraints $\sum_{k=1}^{K} \omega_k = 1$ and $\Omega \succ 0$. The first penalty term shrinks all class centroids towards zero, the global centroid of the centred data. If all the μ_{kj} $(k = 1, \ldots, K)$ are estimated to be zero, variable j is considered noninformative, in the spirit of the nearest shrunken centroid method (Tibshirani et al., 2003). Criterion (6) can be considered as an improved version of the shrunken centroid method, which assumes that the covariance matrix is diagonal. Further, unlike (3), both (6) and the shrunken centroid method claim a variable as noninformative only when all the μ_{kj} $(k = 1, \ldots, K)$ are estimated as zero, and they do not identify class-specific discriminable variables.

3. Asymptotic properties

Let $\omega = (\omega_{(1)}^{\mathrm{T}}, \omega_K)^{\mathrm{T}}$, where $\omega_{(1)} = (\omega_1, \ldots, \omega_{K-1})^{\mathrm{T}}$ and $\omega_K = 1 - \sum_{k=1}^{K-1} \omega_k$. Let $\omega^* = (\omega_{(1)}^{*\mathrm{T}}, \omega_K^*)^{\mathrm{T}}$, μ^*, Ω^* and Σ^* be the true values of ω, μ, Ω and Σ, respectively. We further define

$$\mathcal{A} = \{(j, l) : \Omega_{jl}^* \neq 0 \text{ for } j, l = 1, \ldots, p_n,\ j \neq l\},$$
$$\mathcal{B} = \{(k, k', j) : \mu_{kj}^* - \mu_{k'j}^* = 0 \text{ for } k, k' = 1, \ldots, K,\ k < k',\ j = 1, \ldots, p_n\};$$

so $\mathcal{A}$ contains the indices of off-diagonal elements in Ω^* which are truly nonzero, and $\mathcal{B}$ contains the indices of class pairs and variables that have zero mean difference.

For a symmetric matrix A, we write $\mathrm{tr}(A)$ for the trace of A, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for the minimum and maximum eigenvalues of A. Define the operator norm and the Frobenius norm by, respectively, $\|A\| = \lambda_{\max}^{1/2}(A^{\mathrm{T}}A)$ and $\|A\|_{\mathrm{F}} = \mathrm{tr}^{1/2}(A^{\mathrm{T}}A)$. We write $|\mathcal{F}|$ for the cardinality and $\mathcal{F}^{\mathrm{c}}$ for the complement of the set $\mathcal{F}$. Let $a_n = |\mathcal{A}|$ and $b_n = K(K-1)p_n/2 - |\mathcal{B}|$; then a_n is the number of nonzero elements among the off-diagonal entries of Ω^*, and b_n is the number of class pairs and variables that have nonzero mean differences. Finally, let $\tau_{ik} = I(Y_i = k)$ and $n_k = \sum_{i=1}^{n} \tau_{ik}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

We assume the following conditions to establish consistency and sparsistency.

Condition 1. There exist positive constants κ_1 and κ_2 such that $\kappa_1 < \lambda_{\min}(\Sigma^*) \leqslant \lambda_{\max}(\Sigma^*) < \kappa_2$ for all n.

Condition 2. There exist positive constants c_1 and c_2 such that $c_1 \leqslant \min_{1\leqslant k\leqslant K} n_k/n \leqslant \max_{1\leqslant k\leqslant K} n_k/n \leqslant c_2$ for all n.

Condition 3. For some $\eta > 0$:

(i) $\lambda_{1n} p_n^{1/2}/(b^*_{\max})^{1/2} \to \infty$, $\lambda_{1n} p_n^{1/2}[b^*_{\max} \log\{K(K-1)p_n/2 - b_n\}]^{-1/2} > 1+\eta$ and $\alpha_n^{\max} = o_{\mathrm{p}}(\lambda_{1n} p_n^{1/2})$;

(ii) $\alpha_n^{\min}/(b^*_{\max})^{1/2} \to \infty$, $\alpha_n^{\min}/(b^*_{\max} \log b_n)^{1/2} > 1+\eta$ and $4\kappa_2 \lambda_{1n} p_n^{1/2}(K-1) < \alpha_n^{\min}$, where $b^*_{\max} = \max_{1 \leqslant j \leqslant p_n} \sigma^*_{jj}$, $\alpha_n^{\max} = \max_{\mathcal{B}} |\sum_{i=1}^n (\tau_{ik'} n_{k'}^{-1} - \tau_{ik} n_k^{-1}) \sum_{l=1}^K \tau_{il} \mu^*_{lj}|$ and $\alpha_n^{\min} = \min_{\mathcal{B}^{\mathrm{c}}} |\sum_{i=1}^n (\tau_{ik'} n_{k'}^{-1} - \tau_{ik} n_k^{-1}) \sum_{l=1}^K \tau_{il} \mu^*_{lj}|$.

Condition 1 bounds the eigenvalues of the covariance matrix Σ^* uniformly, and Condition 2 implies that the K samples are of comparable sizes. Both are commonly used conditions in the high-dimensional setting (Cai & Liu, 2011), which facilitate the proof of consistency. Condition 3 is analogous to the conditions in Theorem 2.3 of Rinaldo (2009), used for proving sparsistency.

Theorem 1. *Under Conditions* 1 *and* 2, *if* $\log p_n/n = O(\lambda_{1n}^2)$, $\log p_n/n = O(\lambda_{2n}^2)$ *and* $(p_n + a_n)(\log p_n)^m/n = O(1)$ *for some* $m > 1$, *then there exists a local maximizer* $(\hat\omega_{(1)}, \hat\mu, \hat\Omega)$ *for the maximization problem* (3)–(4) *such that* $\|\hat\omega_{(1)} - \omega^*_{(1)}\|_2^2 = O_{\mathrm{p}}(n^{-1})$, $\|\hat\mu - \mu^*\|_2^2 = O_{\mathrm{p}}(p_n \log p_n/n)$ *and* $\|\hat\Omega - \Omega^*\|_{\mathrm{F}}^2 = O_{\mathrm{p}}\{(p_n + a_n)\log p_n/n\}$.

Theorem 2. *Under the conditions given in Theorem* 1, *for the local maximizer of* (3)–(4) *satisfying* $\|\hat\omega_{(1)} - \omega^*_{(1)}\|_2^2 = O_{\mathrm{p}}(n^{-1})$, $\|\hat\mu - \mu^*\|_2^2 = O_{\mathrm{p}}(p_n \log p_n/n)$, $\max_{1 \leqslant j \leqslant p_n} \|\hat\mu_{(j)} - \mu^*{}_{(j)}\|_2^2 = O_{\mathrm{p}}(\rho_{n1})$ *for a sequence* $\rho_{n1} \to 0$, $\|\hat\Omega - \Omega^*\|_{\mathrm{F}}^2 = O_{\mathrm{p}}\{(p_n + a_n)\log p_n/n\}$ *and* $\|\hat\Omega - \Omega^*\|^2 = O_{\mathrm{p}}(\rho_{n2})$ *for a sequence* $\rho_{n2} \to 0$, *we have that:*

(i) *if* $\log p_n/n + \rho_{n1} + \rho_{n2} = O(\lambda_{2n}^2)$, *then with probability tending to* 1, $\hat\Omega_{jl} = 0$ *for all* $(j,l) \in \mathcal{A}^{\mathrm{c}}$ *with* $j \neq l$;

(ii) *if Condition* 3 *holds, then* $\lim_{n\to\infty} \mathrm{pr}(\hat{\mathcal{B}} = \mathcal{B}) = 1$, *where* $\hat{\mathcal{B}} = \{(k, k', j) : \hat\mu_{kj} - \hat\mu_{k'j} = 0$ *for* $1 \leqslant k < k' \leqslant K,\ j = 1, \ldots, p_n\}$.

Theorem 1 says that with proper tuning parameters λ_{1n} and λ_{2n}, the covariance-enhanced discriminant analysis estimators are consistent with certain rates of convergence. Theorem 2 shows the sparsistency of $\hat\Omega$ and of the fusion estimator $\hat\mu$, ensuring selection consistency for the true signals among the predictors and identification in accordance with their corresponding discriminable classes.

Theorem 1 indicates that $\hat\mu$ is consistent when $p_n/n = O\{(\log p_n)^{-m}\}$ with some $m > 1$, which seems restrictive. There are at least p_n nonzero elements, each of which can be estimated at best with rate $n^{-1/2}$, so the total squared error is at least of rate p_n/n, and for high dimensionality we pay the price $\log p_n$. The rate decays to zero slowly, which implies that p_n can be comparable to n without violating the results in practice; and what we care about is the mean difference $\delta^*_\mu = \{\mu^*_{kj} - \mu^*_{k'j} : k, k' = 1, \ldots, K,\ k < k',\ j = 1, \ldots, p_n\}$. If δ^*_μ is sparse enough, we expect consistency and sparsistency to hold for $p_n > n$.

Next, we consider the binary classification problem. The following theorem establishes the asymptotic optimality of the proposed method in terms of the misclassification error under certain conditions on the divergence rates of b_n, p_n, a_n and $\Delta^2_{p_n}$, where $\Delta^2_{p_n} = \delta^{*\mathrm{T}}_\mu \Omega^* \delta^*_\mu$.

Theorem 3. *In the binary case,* $K = 2$, *under the conditions given in Theorem* 2 *and assuming that*

$$c_n = \max\left\{\rho_{n2}^{1/2}, \frac{a_n^{1/2}}{\Delta_{p_n} n^{1/2}}, \frac{b_n^{1/2}}{\Delta_{p_n} n^{1/2}}, \frac{b_n^{1/2}\rho_{n1}^{1/2}}{\Delta_{p_n}}\right\} \to 0, \quad n \to \infty, \tag{7}$$

we have that:

(i) *the conditional misclassification rate of the proposed covariance-enhanced discriminant analysis is*

$$R_n = \Phi[-\{1 + O_\mathrm{p}(c_n)\}\Delta_{p_n}/2],$$

where Φ *is the standard normal cumulative distribution function;*

(ii) *if* Δ_{p_n} *is bounded, then the proposed method is asymptotically optimal and*

$$\frac{R_n}{R_{\mathrm{OPT}}} - 1 = O_\mathrm{p}(c_n),$$

where $R_{\mathrm{OPT}} = \Phi(-\Delta_{p_n}/2)$ *denotes the misclassification rate of the optimal classification rule (Anderson, 2003);*

(iii) *if* $\Delta_{p_n} \to \infty$*, then for the proposed method we have* $R_n - R_{\mathrm{OPT}} \to 0$ *in probability;*

(iv) *if* $\Delta_{p_n} \to \infty$ *and* $c_n \Delta_{p_n}^2 \to 0$*, then the proposed method is asymptotically optimal.*

Remark 2. Condition (7) is related to the convergence rate of the estimators $\hat{\mu}$ and $\hat{\Omega}$ and the number of nonzero elements in δ_μ^* and Ω^*. Essentially, it holds under the sparsity assumptions on Ω^* and δ_μ^* and with the existence of consistent estimators of μ^* and Ω^* when the values of nonzero mean differences are bounded. Theorem 3 is important as it discusses the asymptotic optimality in terms of the misclassification error, when $\|\delta_\mu^*\|_2$, the magnitude of mean differences, diverges to infinity at different rates.

4. Implementation and tuning parameter selection

Note that $\hat{\omega}_k = \sum_{i=1}^n I(y_i = k)/n$ $(k = 1, \ldots, K)$, whereas the estimators of μ and Ω can be obtained through an iterative algorithm: we fix μ and estimate Ω; then we fix the estimated Ω and estimate μ; we iterate between these two steps until the algorithm converges. Since the value of the objective function (3) decreases over iterations, convergence is guaranteed.

When μ is fixed, to maximize Q_n with respect to Ω it suffices to maximize

$$Q_1(\Omega) = \log|\Omega| - \mathrm{tr}\{S(\mu)\Omega\} - \frac{1}{2}\lambda_{2n} \sum_{j \neq j'} |\Omega_{jj'}| \qquad (8)$$

over all nonnegative-definite matrices Ω for a known covariance matrix $S(\mu)$, similar to the problem of estimating sparse graphs. Hence, we can apply the graphical lasso algorithm (Friedman et al., 2008) to efficiently solve for Ω.

When Ω is fixed, to maximize Q_n with respect to μ it suffices to minimize

$$n^{-1} \sum_{i=1}^n \sum_{k=1}^K I(y_i = k)(x_i - \mu_k)^\mathrm{T}\Omega(x_i - \mu_k) + \frac{1}{2}\lambda_{1n} \sum_{j=1}^{p_n} \sum_{1 \leqslant k < k' \leqslant K} |\mu_{kj} - \mu_{k'j}|. \qquad (9)$$

It is challenging to directly minimize (9) with respect to μ, due to the fusion penalty. We apply local quadratic approximation (Fan & Li, 2001) to convert the minimization in (9) into a generalized ridge problem. Specifically, we write

$$|\mu_{kj}^{(t+1)} - \mu_{k'j}^{(t+1)}| \approx \frac{(\mu_{kj}^{(t+1)} - \mu_{k'j}^{(t+1)})^2}{2|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|} + \frac{1}{2}|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|,$$

where t is the iteration index used to denote iterations of the local quadratic approximation. Consequently, we only need to consider the objective function

$$Q_2(\mu) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} (x_i - \mu_k)^{\mathrm{T}} \Omega (x_i - \mu_k) + \frac{1}{2} \lambda_{1n} \sum_{j=1}^{p_n} \sum_{1 \leqslant k < k' \leqslant K} \frac{(\mu_{kj} - \mu_{k'j})^2}{2|\mu_{kj}^{(t)} - \mu_{k'j}^{(t)}|}, \quad (10)$$

where $\tau_{ik} = I(y_i = k)$; thus, $\mu^{(t+1)} = \arg\min_{\mu} Q_2(\mu)$.

Overall, the algorithm proceeds as follows.

Step 1. Initialize $\mu^{(0)}$ with some plausible values, and set $s = 1$.

Step 2. For iteration s, apply the graphical lasso algorithm to maximize (8) with μ replaced by $\mu^{(s-1)}$, and obtain $\Omega^{(s)}$.

Step 3. With Ω replaced by $\Omega^{(s)}$, iteratively minimize the generalized ridge criterion (10) until $\sum_{j=1}^{p_n} \sum_{k=1}^{K} |\mu_{kj}^{(t+1)} - \mu_{kj}^{(t)}| / \sum_{j=1}^{p_n} \sum_{k=1}^{K} |\mu_{kj}^{(t)}|$ is small enough to obtain $\mu^{(s)}$.

Step 4. If $|Q_n(\hat{\omega}, \mu^{(s)}, \Omega^{(s)}) - Q_n(\hat{\omega}, \mu^{(s-1)}, \Omega^{(s-1)})|$ is small enough, stop the algorithm; otherwise, set $s \leftarrow s + 1$ and go to Step 2.

In terms of selecting the tuning parameters λ_{1n} and λ_{2n}, we follow the suggestion in Wang et al. (2007) and use a BIC-type criterion:

$$\text{BIC}(\lambda_{1n}, \lambda_{2n}) = -2nl_n(\hat{\omega}, \hat{\mu}, \hat{\Omega}) + (K - 1 + d_{\hat{\mu}} + d_{\hat{\Omega}}) \log(n),$$

where $d_{\hat{\mu}}$ is the number of distinct nonzero elements in $\hat{\mu}$ and $d_{\hat{\Omega}}$ is the number of nonzero elements in $\hat{\Omega}$.

5. Simulation studies

In this section, we assess the finite-sample performance of the proposed method. For comparison, we consider several related methods, including fusion-regularized linear discriminant analysis (Guo, 2010), doubly l_1-penalized linear discriminant analysis as given by (6), sparse discriminant analysis (Clemmensen et al., 2011), l_1-penalized linear discriminant analysis, and fused-penalized linear discriminant analysis (Witten & Tibshirani, 2011). Fusion-regularized linear discriminant analysis is a special case of our method where the covariance matrix is assumed to be diagonal.

Example 1. Consider a three-class scenario with a total of $p = 210$ variables, generated according to the following mechanism: the first ten variables are independent $N(\mu_{kj}, 1)$ for class k, whereas the remaining 200 variables are independent and identically distributed from $N(0, 1)$ for all three classes. Table 1 gives the means of the first ten variables. For example, in class 1, variables 1–5 all have mean 0, and variables 6–10 all have mean 1·5.

Example 2. In this example, the true model is the same as in Example 1, except that the covariance matrix has the AR(1) correlation structure with autocorrelation coefficient 0·6 for variables 1–5 and variables 6–10; variables 1–5 are independent of variables 6–10, and both groups are independent of the remaining 200 variables.

Example 3. In this case, the true model is the same as in Example 1, except that variable 5 has means different from those of variables 1–4 and the correlation structure for variables 1–10

Table 1. *Means of the informative variables in Examples* 1–3

Example	Variables	Class 1	Class 2	Class 3
1 & 2	1–5	0	0	−2·5
	6–10	1·5	−1·5	−1·5
3	1–4	0	0	−2·5
	5	−0·5	2	−2·5
	6–10	1·5	−1·5	−1·5

differs from the structures in Examples 1 and 2. Specifically, the means of variable 5 are −0·5, 2 and −2·5 for the three classes. Variables 1–5 have an interchangeable correlation structure with parameter 0·5. Variables 6–10 are correlated with the same structure but independently of variables 1–5. Table 1 reports the means for the first ten variables.

In each simulation example, only the first ten variables are informative. Moreover, in Examples 1 and 2, a variable is informative for separating a pair of classes if it has unequal means for those two classes. For example, variables 1–5 are informative for separating classes 1 and 3 or classes 2 and 3, but not for separating classes 1 and 2; similarly for variables 6–10. For Example 3, it is less straightforward to identify the informative variables for discriminating classes 1 and 2. For example, variable 1 has equal means for classes 1 and 2, but it contributes to the classification through its correlation with the informative variable 5, as illustrated in Fig. 1. Therefore, unlike in Examples 1 and 2, variables 1–5 are all informative for separating classes 1 and 2.

For each example, we generate 200 datasets, each consisting of $n_1 = n_2 = n_3 = 50$ training and test samples. We then apply each method to the training data and record the misclassification error rate evaluated on the testing data, the proportion of incorrectly removed informative variables, i.e., the false negative rate, the proportion of incorrectly selected noninformative variables, i.e., the false positive rate, and the model size.

Table 2 summarizes the misclassification error rates and variable selection results for the six methods over 200 replications. Overall, the proposed method outperforms the other methods in terms of classification accuracy and has prediction accuracy competitive with smaller models. In terms of variable selection, all methods except for sparse discriminant analysis (Clemmensen et al., 2011) are effective at identifying the informative variables, while the proposed method is more effective at removing noninformative variables. Sparse discriminant analysis has decent classification accuracy overall, but tends to miss important variables.

If a variable is noninformative for discriminating a pair of classes and the corresponding estimated parameters satisfy equation (5), we consider it as correct fusion. Table 3 summarizes the fusion results for all the examples. Each row in the table displays the average proportion of fused variables out of the five for separating the corresponding pair of classes. For example, the first row indicates that for the proposed method, on average 99·5% of the first five variables are fused for classes 1 and 2. Note that 100% is the optimal value except for variables 1–5 in Example 3, because variables 1–5 are informative for separating classes 1 and 2 in Example 3, and thus 0% should be the optimal value for the corresponding row in the table. The methods of Clemmensen et al. (2011) and Witten & Tibshirani (2011) do not provide fusion results for any specific pair of classes, so are not listed in Table 3. The proposed method outperforms the method of Guo (2010) in correctly separating the specific pairs of classes, while doubly l_1-penalized linear discriminant analysis is barely able to fuse any of the first ten variables using the criterion (5), especially when some of variables are correlated. The doubly l_1-penalized method only penalizes the individual μ_{kj}, not the pairwise differences; thus a variable can be fused only if all μ_{kj} $(k = 1, \ldots, K)$ are estimated as zero, but clearly this is not a favourable estimate for the first ten variables as the true class means are different.

Table 2. *Misclassification error rates and variable selection results for Examples* 1–3*: means and standard errors (in parentheses) of various performance measures based on* 200 *replications*

Example	Method	ER (%)	FN (%)	FP (%)	MS
1	Proposed method	0·23 (0·36)	0 (0)	0·29 (0·59)	10·58 (1·18)
	Guo (2010)	0·29 (0·45)	0 (0)	7·66 (6·71)	25·32 (13·41)
	Doubly l_1	2·62 (4·41)	0 (0)	46·86 (9·34)	103·72 (18·69)
	Clemmensen et al. (2011)	0·38 (0·52)	0 (0)	0·47 (0·56)	10·95 (1·12)
	Witten & Tibshirani (2011) l_1	0·29 (0·48)	0 (0)	15·29 (15·79)	40·59 (31·58)
	Witten & Tibshirani (2011) fused	0·29 (0·69)	0 (0)	3·04 (15·81)	16·07 (31·61)
2	Proposed method	3·91 (1·55)	0 (0)	0·53 (0·48)	10·71 (0·87)
	Guo (2010)	11·38 (4·93)	0 (0)	9·22 (8·49)	28·45 (16·97)
	Doubly l_1	7·11 (4·58)	0 (0)	77·64 (13·32)	165·29 (26·64)
	Clemmensen et al. (2011)	4·34 (1·68)	3·40 (5·71)	1·71 (4·08)	13·42 (8·16)
	Witten & Tibshirani (2011) l_1	4·39 (1·63)	0 (0)	33·25 (36·75)	76·50 (73·49)
	Witten & Tibshirani (2011) fused	4·24 (1·66)	0 (0)	14·14 (31·93)	38·28 (63·86)
3	Proposed method	1·87 (1·05)	0 (0)	0·47 (0·53)	10·93 (1·06)
	Guo (2010)	8·11 (2·22)	0 (0)	8·72 (7·03)	27·44 (14·05)
	Doubly l_1	2·43 (1·27)	0 (0)	63·87 (10·99)	137·73 (21·97)
	Clemmensen et al. (2011)	2·00 (1·12)	6·85 (8·12)	1·70 (1·03)	12·72 (1·55)
	Witten & Tibshirani (2011) l_1	2·61 (1·31)	0 (0)	21·80 (32·70)	53·60 (65·40)
	Witten & Tibshirani (2011) fused	3·75 (1·46)	0 (0)	10·92 (29·24)	31·84 (58·48)

ER, misclassification error rate on the test data; FN, false negative rate; FP, false positive rate; MS, model size.

Table 3. *Pairwise class fusion results* (%) *for Examples* 1–3

Example	Variables	Pair	Proposed method	Guo (2010)	Doubly l_1
1	1–5	1,2	99·50	88·10	52·10
			(3·13)	(16·30)	(35·92)
	6–10	2,3	99·40	86·20	0·00
			(3·42)	(16·09)	(0·00)
2	1–5	1,2	97·90	86·30	1·05
			(11·41)	(18·87)	(0·89)
	6–10	2,3	98·10	87·80	0·00
			(10·91)	(18·68)	(0·00)
3	1–5	1,2	0·90	34·40	0·10
			(3·20)	(9·06)	(1·00)
	6–10	2,3	99·80	87·40	0·00
			(2·00)	(19·37)	(0·00)

Each entry in the third column is a pair of indiscriminable classes for the variables in the corresponding row, except for variables 1–5 in Example 3; for instance, the first row indicates that variables 1–5 are noninformative for separating classes 1 and 2. The numbers in the fourth to sixth columns give the proportions of variables in the set (with standard deviations in parentheses) that are identified as noninformative for separating a given pair of classes by each method; the optimal value is 100% in each case except for variables 1–5 in Example 3, where the optimal value should be 0%. All results are averaged over 200 repetitions.

6. Kidney transplant rejection and tissue injury

We consider a kidney transplant rejection and tissue injury dataset (Flencher et al., 2004), which consists of 62 tissue samples from kidney transplant patients, including 17 normal donor kidneys, 19 well-functioning transplants without rejection, 13 kidneys undergoing acute rejection, and 13 transplants with renal dysfunction and without rejection. Each sample is described

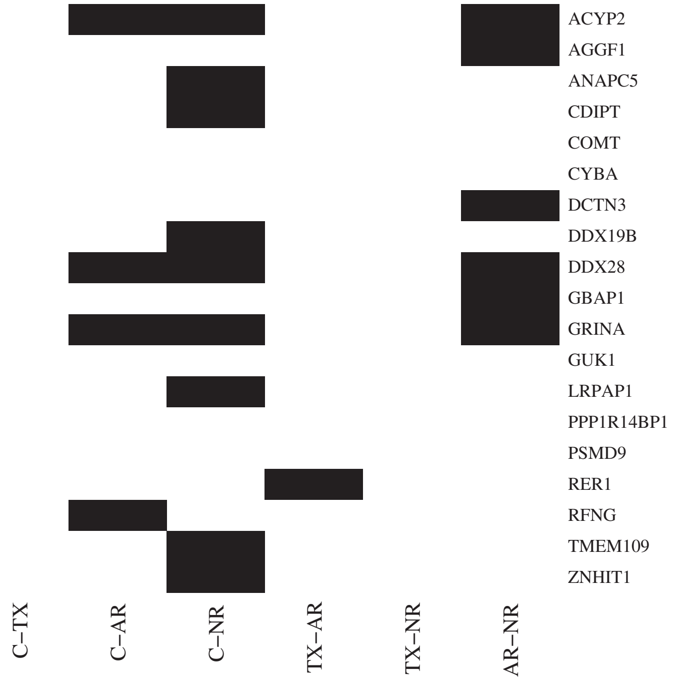

Fig. 2. Pairwise class fusion results for the proposed method with the 19 most informative genes selected based on the kidney transplant rejection and tissue injury dataset. Each row corresponds to a gene, and each column corresponds to a class pair. A dark block indicates that the corresponding gene is noninformative for separating the corresponding pair of classes.

by 12 625 genes from kidney biopsies and peripheral blood lymphocytes. Distinguishing between these four types of kidney tissue is crucial to balancing the need for immunosuppression to prevent rejection while minimizing drug-induced toxicities.

Before applying our method, we preselect a subset of genes according to their variances, since genes with large variability are generally considered to be of potential greatest relevance to biological function (Mar et al., 2011). Similar to Guo et al. (2010), from the 12 625 genes we select the 100 with the largest variances and the 100 with the smallest variances. The selection process does not use any class label information. We then centre these 200 genes before classification.

To assess performance, we randomly split the dataset into training and test datasets in a 2:1 ratio. We estimate and select the genes on the training dataset and evaluate the classification accuracy on the test dataset. This procedure is repeated 100 times. In terms of classification accuracy, the proposed method performs best, while doubly l_1-penalized linear discriminant analysis performs worst; see Fig. S1 in the Supplementary Material.

To assess variable selection, for each gene we count the number of times that it was selected based on 100 random splits. According to this frequency, we choose the 25 most informative genes. There are 19 most informative genes that are selected by all five methods, and besides these 19 common genes, our proposed method selected the following genes as the most informative: HCFC1, PLIN2, LOC646347, IDS, SPAG5 and TIGR(HG4518-HT4921); some of these genes are significantly relevant to renal function. For example, the HCFC1 gene, as a member of the host cell factor family, was reported in Wilson et al. (1995) to be highly expressed in fetal tissues and the adult kidney; the expression of PLIN2 has been shown to be a predictor of cancer-specific survival in clear cell renal carcinoma (Yao et al., 2007); and SPAG5 is highly expressed in normal human kidneys (Chang et al., 2001), while its level of expression is much lower in hypogonadal kidneys (Suzuki et al., 2006).

The proposed method reveals that the 19 genes selected as most informative are not all informative for discriminating every pair of classes. For example, Fig. 2 shows that gene AGGF1, reported to have strong protein expression in blood vessels embedded in kidney tissues (Fan et al., 2009), does not distinguish the acute rejection class from the renal dysfunction without rejection class, while it is informative for the other pairs of classes; gene GRINA, which plays a major role in gentamicin ototoxicity (Leung et al., 2004) and in $1{,}25(\mathrm{OH})_2$ D_3 synthesis (Parisi et al., 2010), does not separate the normal, acute rejection and renal dysfunction without rejection classes; and gene RFNG, which is strongly expressed in the kidney (Challen et al., 2006), does not discriminate the normal class from the acute rejection class. Further, although some of the genes have the same means across different classes, they are informative in classification via correlations with other informative genes. For example, gene AGGF1 discriminates the normal class from the acute rejection and renal dysfunction without rejection classes, even though it has the same mean within these three classes, based on Fig. S2 in the Supplementary Material.

In summary, the proposed method identifies new genes that are relevant to renal function and, by using the underlying covariance structure between genes, elucidates the impact of genes on discriminating particular renal functional classes, which is a crucial step in the development of gene therapy.

ACKNOWLEDGEMENT

The authors thank the editor, associate editor, and two reviewers for their thoughtful comments and helpful suggestions. Xu was supported by the National Natural Science Foundation of Jiangsu Province, China. J. Zhu was supported by the U.S. National Institutes of Health and National Science Foundation. Li was supported by the U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theorems and additional figures for the kidney transplant rejection and tissue injury data example.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 3rd ed.

BICKEL, P. J. & LEVINA, E. (2004). 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.

BICKEL, P. J. & LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

CAI, T. & LIU, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.* **106**, 1566–77.

CAI, T., LIU, W. & LUO, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

CHALLEN, G. A., BERTONCELLO, I., DEANE, J. A., RICARDO, S. D. & LITTLE, M. H. (2006). Kidney side population reveals multilineage potential and renal functional capacity but also cellular heterogeneity. *J. Am. Soc. Nephrol.* **17**, 1896–912.

CHANG, M.-S., HUANG, C.-J., CHEN, M.-L., CHEN, S.-T., FAN, C., CHU, J.-M., LIN, W.-C. & YANG, Y.-C. (2001). Cloning and characterization of hMAP126, a new member of mitotic spindle-associated proteins. *Biochem. Biophys. Res. Commun.* **287**, 116–21.

CLEMMENSEN, L., HASTIE, T. J., WITTEN, D. M. & ERSBØLL, B. (2011). Sparse discriminant analysis. *Technometrics* **53**, 406–13.

FAN, C., OUYANG, P., TIMUR, A. A., HE, P., YOU, S.-A., HU, Y., KE, T., DRISCOLL, D. J., CHEN, Q. & WANG, Q. K. (2009). Novel roles of GATA1 in regulation of angiogenic factor AGGF1 and endothelial cell function. *J. Biol. Chem.* **284**, 23331–43.

Fan, J. & Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–37.

Fan, J., Feng, Y. & Tong, X. (2012). A road to classification in high-dimensional space. *J. R. Statist. Soc.* B **74**, 745–71.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Flencher, S. M., Kurian, S. M., Head, S. M., Sharp, S. M., Whisenant, T. C., Zhang, J., Chismar, J. D., Horvath, S., Mondala, T., Gilmartin, T., Cook, D. J., Kay, S. A., Walker, J. R. & Salomon, D. R. (2004). Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. *Am. J. Transplant* **4**, 1475–89.

Friedman, J. H., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics* **11**, 599–608.

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66**, 793–804.

Guo, Y., Hastie, T. J. & Tibshirani, R. J. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100.

Lam, C. & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–78.

Leung, J. C., Marphis, T., Craver, R. D. & Silverstein, D. M. (2004). Altered NMDA receptor expression in renal toxicity: Protection with a receptor antagonist. *Kidney Int.* **66**, 167–76.

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

Mar, J. C., Matigian, N. A., Mackay-Sim, A., Mellick, G. D., Sue, C. M., Silburn, P. A., McGrath, J. J., Quackenbush, J. & Wells, C. A. (2011). Variance of gene expression identifies altered network constraints in neurological disease. *PLoS Genet.* **7**, article no. e1002207.

Parisi, E., Bozic, M., Ibarz, M., Panizo, S., Valcheva, P., Coll, B., Fernandez, E. & Valdivielso, J. M. (2010). Sustained activation of renal N-methyl-D-aspartate receptors decreases vitamin D synthesis: A possible role for glutamate on the onset of secondary HPT. *Am. J. Physiol. Endocrin. Metab.* **299**, E825–31.

Qiao, Z., Zhou, L. & Huang, J. Z. (2008). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *Int. J. Appl. Math.* **39**, 48–60.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37**, 2922–52.

Shao, J., Wang, Y., Deng, X. & Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39**, 1241–65.

Suzuki, H., Yagi, M. & Suzuki, K. (2006). Duplicated insertion mutation in the microtubule-associated protein Spag5 (astrin/MAP126) and defective proliferation of immature Sertoli cells in rat hypogonadic testes. *Reproduction* **132**, 79–93.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Tibshirani, R. J., Hastie, T. J., Narasimhan, B. & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.* **99**, 6567–72.

Tibshirani, R. J., Hastie, T. J., Narasimhan, B. & Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.* **18**, 104–17.

Wang, H., Li, R. & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–68.

Wang, S. & Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics* **23**, 972–9.

Wilson, A. C., Parrish, J. E., Massa, H. F., Nelson, D. L., Trask, B. J. & Winship, H. (1995). The gene encoding the VP16-accessory protein HCF (HCFC1) resides in human Xq28 and is highly expressed in fetal tissues and the adult kidney. *Genomics* **25**, 462–8.

Witten, D. M., Friedman, J. H. & Simon, N. (2011). New insights and faster computations for the graphical lasso. *J. Comp. Graph. Statist.* **20**, 892–900.

Witten, D. M. & Tibshirani, R. J. (2009). Covariance-regularized regression and classification for high dimensional problems. *J. R. Statist. Soc.* B **71**, 615–36.

Witten, D. M. & Tibshirani, R. J. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc.* B **73**, 753–72.

Yao, M., Huang, Y., Shioi, K., Hattori, K., Murakami, T., Nakaigawa, N., Kishida, T., Nagashima, Y. & Kubota, Y. (2007). Expression of adipose differentiation-related protein: a predictor of cancer-specific survival in clear cell renal carcinoma. *Clin. Cancer Res.* **13**, 152–60.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

[*Received November* 2013. *Revised August* 2014]