Biometrika (2018), **105**, 4, *pp.* 917–930
Printed in Great Britain
doi: 10.1093/biomet/asy047
Advance Access publication 22 October 2018

Integrative linear discriminant analysis with guaranteed error rate improvement

By QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, 3105D McGavran-Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.
quefeng@email.unc.edu

and LEXIN LI

Division of Biostatistics, University of California at Berkeley, 50 University Hall 7360, Berkeley, California 94720, U.S.A.
lexinli@berkeley.edu.

Summary

Multiple types of data measured on a common set of subjects arise in many areas. Numerous empirical studies have found that integrative analysis of such data can result in better statistical performance in terms of prediction and feature selection. However, the advantages of integrative analysis have mostly been demonstrated empirically. In the context of two-class classification, we propose an integrative linear discriminant analysis method and establish a theoretical guarantee that it achieves a smaller classification error than running linear discriminant analysis on each data type individually. We address the issues of outliers and missing values, frequently encountered in integrative analysis, and illustrate our method through simulations and a neuroimaging study of Alzheimer's disease.

Some key words: Bayes error; High-dimensional classification; Integrative analysis; Linear discriminant analysis; Multi-type data; Regularization.

1. Introduction

Integrative analysis is receiving increasing attention in biomedical research. One example is genomics, where gene expression, DNA copy number variation, and DNA methylation data are simultaneously measured from the same biospecimen. Numerous empirical studies have demonstrated that such analysis boosts power for predicting disease progression and identifying important biomarkers (Shen et al., 2013; Liu et al., 2013, 2014; Li et al., 2014; Cai et al., 2016; Richardson et al., 2016). Another example is neuroimaging, where both structural and functional brain imaging scans are acquired for the same study subject. Integrative analysis of multiple types of images improves our understanding of the brain and the diagnosis of neurological disorders (Zhang et al., 2011; Dai et al., 2012; Uludag & Roebroeck, 2014). The advantages of integrative analysis have, however, mostly been established empirically, with little theory on when joint analysis is guaranteed to improve statistical performance. In this paper, we provide a theoretical justification for the benefit of joint analysis in two-class classification.

Linear discriminant analysis is effective for many benchmark datasets (Hand, 2006). Recently, aiming at high-dimensional classification where the number of features exceeds the sample size, there have been a number of proposals of regularized linear discriminant analysis (Wu et al., 2009; Clemmensen et al., 2011; Witten & Tibshirani, 2011; Shao et al., 2011; Cai & Liu, 2012; Fan et al., 2012; Mai et al., 2012; Han et al., 2013). However, all these methods focus on classification using a single data type.

In this article, we develop an integrative linear discriminant analysis method for multi-type data, and we show that it is guaranteed to asymptotically reduce classification error compared with running linear discriminant analysis on a single data type. We first show that the Bayes error for linear discriminant analysis is a decreasing function of the number of variables involved. On the other hand, as it involves more unknown parameters, multi-type classification introduces additional variance and often a larger estimation error. Without proper control, the resulting classifier can perform poorly, and in the extreme case it can be as bad as random guessing (Bickel & Levina, 2004). To address this, we propose a regularized classifier, show that it enjoys rate consistency in that its error rate approaches the Bayes error, and establish a theoretical guarantee that our approach using multi-type data has a smaller classification error than using single-type data. This is achieved by a careful characterization of the trade-off between the additional discriminative information brought by using more variables and the extra estimation error it brings. To the best of our knowledge, this is the first result of its kind in the integrative analysis literature. We also show that our method consistently excludes all non-discriminative features and uniformly consistently estimates the discriminative ones. Most existing regularized linear discriminant analysis solutions obtain either only rate consistency (Shao et al., 2011; Cai & Liu, 2012; Fan et al., 2012) or selection consistency (Mai et al., 2012), but rarely both (Han et al., 2013). We develop a robust version of our method and show that the guaranteed error rate improvement holds under a class of elliptical distributions. Our method can also accommodate blocks of missing values. For the special case where there is only a single data type, our method achieves the same convergence rate as the linear programming discriminant rule of Cai & Liu (2012), but requires less restrictive conditions when the underlying distribution is nonnormal.

2. METHODOLOGY

2·1. *Bayes error*

We consider a binary classification problem, where $Y \in \{0, 1\}$ is a class label with prior distribution $\mathrm{pr}(Y = k) = \pi_k$ $(k = 0, 1)$, $\pi_0 = \pi_1 = 1/2$, and $X \in \mathbb{R}^d$ is the predictor vector. We assume $E(X \mid Y = k) = \mu_k \in \mathbb{R}^d$ and $\mathrm{var}(X \mid Y = k) = \Sigma \in \mathbb{R}^{d\times d}$ $(k = 0, 1)$. If both classes are normally distributed, the Bayes rule classifies a new observation x to class 0 if and only if $\delta^{\mathrm{T}}\Sigma^{-1}(x - \mu) \geqslant 0$, where $\delta = \mu_0 - \mu_1$ and $\mu = (\mu_0 + \mu_1)/2$. The corresponding Bayes error is given by

$$R_d^* = \Phi\left(-\sqrt{\Delta_d}/2\right), \quad \Delta_d = \delta^{\mathrm{T}}\Sigma^{-1}\delta, \tag{1}$$

where $\Phi(x)$ is the standard normal cumulative distribution function and Δ_d, the Mahalanobis distance between the centroids of the two classes, is a normalized Euclidean distance that measures how far the two centroids are apart and quantifies the difficulty of classifying the classes. Clearly, the Bayes error R_d^* is a decreasing function of Δ_d. We now show that Δ_d increases with d, and thus R_d^* is a decreasing function of d. We introduce the following notation. Let $\delta_{1:d-1}$ and δ_d denote the first $d - 1$ and the dth coordinates of δ, let $\sigma_{1:d-1,d}$ denote the first $d - 1$ coordinates

of the dth column of Σ, and let $\Sigma_{1:d-1,1:d-1}$ denote the first $d-1$ rows and columns of Σ. We then have the next proposition.

PROPOSITION 1. *The Mahalanobis distance Δ_d is a nondecreasing function of the dimension d, and it is strictly increasing if $\delta_d \neq \sigma_{1:d-1,d}^{\mathrm{T}} \Sigma_{1:d-1,1:d-1}^{-1} \delta_{1:d-1}$.*

This result shows that, as long as the mean difference of a new variable between the two classes is not a particular linear combination of the mean differences of the variables already included, the distance Δ_d strictly increases, so including more variables in classification leads to a smaller Bayes error. However, using more variables in classification also involves more unknown parameters and thus induces a larger variance. Next we propose an integrative classifier that can effectively control the estimation error and can asymptotically reach the Bayes error.

2·2. *Integrative linear discriminant analysis*

Classical linear discriminant analysis replaces the unknown parameters in the Bayes rule with their maximum likelihood estimators, and classifies a new observation x into class 0 if and only if $\hat{\delta}^{\mathrm{T}} \hat{\Sigma}^{-1}(x - \hat{\mu}) \geqslant 0$, where

$$\begin{aligned} \hat{\mu}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} X_{0i}, \quad \hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \quad \hat{\mu} = (\hat{\mu}_0 + \hat{\mu}_1)/2, \\ \hat{\delta} = \hat{\mu}_0 - \hat{\mu}_1, \quad \hat{\Sigma} = \frac{1}{n} \sum_{k=0}^{1} \sum_{i=1}^{n_k} (X_{ki} - \hat{\mu}_k)(X_{ki} - \hat{\mu}_k)^{\mathrm{T}}. \end{aligned} \tag{2}$$

Next we develop our integrative linear discriminant analysis classifier using multi-type data. From now on, X denotes the vector that concatenates the variables from all data types.

We first consider the scenario in which all M types share a set of p common variables, so that $d = Mp$. By common variables we mean those variables in different data types that are related through some common structure. For instance, in genomics, common variables could represent different genetic measurements corresponding to the same gene; in neuroimaging, they could represent different brain characteristics for the same brain region. The optimal Bayes classification direction $\beta^* = \Sigma^{-1}\delta$ minimizes $\beta^{\mathrm{T}}\Sigma\beta/2 - \delta^{\mathrm{T}}\beta$. Accordingly, to estimate β^*, we replace (δ, Σ) with $(\hat{\delta}, \hat{\Sigma})$ in (2) and solve the regularized problem

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\beta^{\mathrm{T}}\hat{\Sigma}\beta - \hat{\delta}^{\mathrm{T}}\beta + \lambda_n \sum_{j=1}^{p} \|\beta_{S_j}\|_G, \tag{3}$$

where j_m is the index of the jth variable in the mth data type, $S_j = \{j_1, \ldots, j_M\}$ is the set of indices corresponding to its appearance in M data types, and β_{S_j} is the subvector of β with indices in S_j $(j = 1, \ldots, p)$. Although the common variables are shared by all data types, their contributions to classification may vary, resulting in potentially different values of β_{j_m}. In (3),

$$\|\beta_{S_j}\|_G = (1-\alpha)\|\beta_{S_j}\|_1 + \alpha\|\beta_{S_j}\|_2, \quad \alpha \in [0, 1], \tag{4}$$

where $\|\beta_{S_j}\|_1$ and $\|\beta_{S_j}\|_2$ denote the ℓ_1-norm and the Euclidean norm of β_{S_j}. Letting $\hat{\beta}$ denote the minimizer of (3), we build a rule that classifies x into class 0 if and only if $\hat{\beta}^{\mathrm{T}}(x - \hat{\mu}) \geqslant 0$, and call it the integrative linear discriminant analysis classifier.

The regularization term in (4) is a weighted sum of the ℓ_1-norm and the Euclidean norm of β_{S_j}. If $\alpha = 1$, it reduces to a group lasso penalty (Yuan & Lin, 2006), and then either all elements of β_{S_j} appear in the classification rule or none do. In other words, a variable's appearance in M data types acts as a group, and all data types have a common set of variables contributing to the classification rule. If $\alpha = 0$, the term in (4) reduces to an elementwise ℓ_1-penalty. In this case, no group structure is built into the penalty, and different data types can have different sets of variables contributing to classification. The key advantage of (4) is that, by tuning the value of α given the data, it offers a data-adaptive way to balance the two cases. The tuning of α and λ_n is done by crossvalidation. Moreover, when $M = 1$, the penalty in (4) reduces to an ℓ_1 penalty. Therefore our method is applicable to single-type classification as well.

We next consider the more general scenario in which different data types do not share a common set of variables. Suppose there are p_m variables in the mth data type and $d = \sum_{m=1}^{M} p_m$. Let $\mathcal{J}$ denote the set of all unique variables, let $\mathcal{M}$ denote the set of common variables that appear in more than one data type, and let $\mathcal{N}$ denote the set of variables that appear in only one data type. We have $\mathcal{J} = \mathcal{M} \cup \mathcal{N}$. Again let S_j ($j \in \mathcal{J}$) denote the set of indices corresponding to the jth variable's appearance across M data types, i.e., its positions in the concatenated index set $(1, \ldots, d)$. Then, we propose to solve the regularized optimization problem

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\beta^{\mathrm{T}}\hat{\Sigma}\beta - \hat{\delta}^{\mathrm{T}}\beta + \lambda_n \left(\sum_{j \in \mathcal{N}} \|\beta_{S_j}\|_1 + \sum_{j \in \mathcal{M}} \|\beta_{S_j}\|_G \right). \tag{5}$$

That is, for those variables appearing in more than one data type, we continue to impose the penalty in (4), whereas for those variables appearing in a single data type, we only impose the ℓ_1 penalty. For the special case that the same set of variables appears in all types, (5) reduces to (3). When no variable appears in more than one type, (5) reduces to an ℓ_1-regularization problem.

In § 3, we develop a proximal gradient algorithm to solve (5), which includes (3) as a special case. In § § 4–6, we present the results based on (3) for notational simplicity. Parallel results under (5) can be obtained straightforwardly.

3. ESTIMATION

For the objective function in (5), let $L(\beta) = \beta^{\mathrm{T}}\hat{\Sigma}\beta/2 - \hat{\delta}^{\mathrm{T}}\beta$ and $g(\beta) = \lambda_n(\sum_{j \in \mathcal{N}} \|\beta_{S_j}\|_1 + \sum_{j \in \mathcal{M}} \|\beta_{S_j}\|_G)$. Since $L(\beta)$ is differentiable, the convex problem (5) can be solved by a proximal gradient algorithm using a majorization-minimization scheme (Parikh & Boyd, 2014). First, we find a quadratic approximation to $L(\beta)$ centred at $\beta^{(k)}$, the estimate at the kth iteration of the algorithm, that majorizes $L(\beta)$. That is,

$$L(\beta) \leqslant L\{\beta^{(k)}\} + \{\beta - \beta^{(k)}\}^{\mathrm{T}} \nabla L\{\beta^{(k)}\} + \frac{1}{2t}\|\beta - \beta^{(k)}\|_2^2, \tag{6}$$

where t is the step size. Denote the right-hand side of (6) by $Q_t\{\beta, \beta^{(k)}\}$. Then we minimize $Q_t\{\beta, \beta^{(k)}\} + g(\beta)$, which gives the proximal problem

$$\underset{\beta \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{2} \left\| \beta - \left[\beta^{(k)} - t\nabla L\{\beta^{(k)}\} \right] \right\|_2^2 + tg(\beta). \tag{7}$$

Following Friedman et al. (2010) and Simon et al. (2013), one can show that (7) has a closed-form solution

$$\beta^{(k+1)} = \mathrm{prox}_{tg}[\beta^{(k)} - t\nabla L\{\beta^{(k)}\}] = \begin{cases} s[\beta^{(k)}_{S_j} - t\{\hat{\Sigma}\beta^{(k)} - \hat{\delta}\}_{S_j}, \lambda_n t], & j \in \mathcal{N}, \\ z_{S_j}(1 - t\alpha\lambda_n \|z_{S_j}\|_2^{-1})_+, & j \in \mathcal{M}, \end{cases} \tag{8}$$

where $z_{S_j} = s[\beta^{(k)}_{S_j} - t\{\hat{\Sigma}\beta^{(k)} - \hat{\delta}\}_{S_j}, t(1-\alpha)\lambda_n]$, $\{\hat{\Sigma}\beta^{(k)} - \hat{\delta}\}_{S_j}$ denotes the subvector of the gradient $\hat{\Sigma}\beta^{(k)} - \hat{\delta}$ with indices in S_j, $\beta^{(k)}_{S_j}$ is defined similarly, and $s(x, \lambda)$ is the elementwise soft-thresholding operator, whose ith element is defined as $\{s(x,\lambda)\}_i = \mathrm{sgn}(x_i)(|x_i| - \lambda)_+$, with $\mathrm{sgn}(x_i)$ denoting the sign of x_i. The solution given in (8) reveals the effect of different penalties on variables in $\mathcal{M}$ and $\mathcal{N}$. For variables in $\mathcal{N}$, the solution is given by an elementwise soft-thresholding. For variables in $\mathcal{M}$, if $\|z_{S_j}\|_2 < t\alpha\lambda_n$, the ℓ_2-term in the penalty shrinks the entire group to zero. Otherwise, the ℓ_1-term still shrinks some individual elements in that group to zero. Consequently, our method provides a flexible selection of individual variables' effects in different data types. As for the step size, we follow Parikh & Boyd (2014, § 4.2) to perform a backtracking line search. This strategy is commonly used in the proximal gradient method. Moreover, Nesterov (2013) has shown that such a choice of step size yields an $O(1/k)$ convergence rate of computation, meaning that after k iterations the difference between the value of the objective function and its minimum is at most $O(1/k)$. Alternatively, one may also use a fixed step size t, as long as it is smaller than $1/c$, where c is the Lipschitz constant for the gradient $\nabla L(\beta)$ and is equal to the maximum eigenvalue of $\hat{\Sigma}$ in our problem. Finally, we stop iterations when $\|\beta^{(k)} - \beta^{(k-1)}\|_2 \leqslant 10^{-3}$. Our algorithm is summarized as follows:

Algorithm 1. The proximal gradient algorithm.

Initialize β at $\beta^{(0)} \in \mathbb{R}^d$ and t at $t^{(0)} \in \mathbb{R}^+$.
At the kth iteration, let $t = t^{(k-1)}$ and repeat
Let $\beta = \mathrm{prox}_{tg}[\beta^{(k-1)} - t\nabla L\{\beta^{(k-1)}\}]$ as defined in (8).
Break if $L(\beta) \leqslant Q_t\{\beta, \beta^{(k-1)}\}$.
Update $t = 0{\cdot}8t$.
Let $t^{(k)} = t$ and $\beta^{(k)} = \beta$.
Iterate until the stopping criterion is met.

4. Theory

Next we establish the error rate and selection consistency of our classification rule from (3). We assume that both classes follow a normal distribution throughout this section, and relax this assumption in § 5.

We first introduce some notation. For a vector $a \in \mathbb{R}^d$, let $\|a\|_\infty = \max_{1\leqslant j\leqslant d}|a_j|$, $\|a\|_1 = \sum_{j=1}^d |a_j|$ and $\|a\|_2 = (\sum_{j=1}^d a_j^2)^{1/2}$ denote its max, ℓ_1- and Euclidean norms. Recall the definitions of j_m and S_j after (3). Define $\|a\|_{1,2} = \sum_{j=1}^p (\sum_{m=1}^M a_{j_m}^2)^{1/2}$, $\|a\|_{1,G} = \sum_{j=1}^p \|a_{S_j}\|_G$, and $\|a\|_{\infty,2} = \max_{1\leqslant j\leqslant p}\|a_{S_j}\|_2$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{d\times d}$, define $\|A\|_{\max} = \max_{ij}|a_{ij}|$, $\|A\|_\infty = \max_{1\leqslant i\leqslant d}\sum_{j=1}^d |a_{ij}|$ and $\|A\|_{\infty,2} = \max_{1\leqslant j\leqslant p}(\sum_{m=1}^M \|\tilde{a}_{j_m}\|_{\infty,2}^2)^{1/2}$, where $\tilde{a}_{j_m}$ is the j_mth row of A. In addition, let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A. For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists a constant $c > 0$ such that $a_n \leqslant cb_n$, $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \gg b_n$ if $b_n/a_n \to 0$. Moreover, we define sets $\mathcal{A} = \{j_m : \beta^*_{j_m} \neq 0\}$, $\mathcal{B} = \{j_m : \beta^*_{j_m} = 0 \text{ but } \beta^*_{S_j} \neq 0\}$ and $\mathcal{C} = \{j_m : \beta^*_{S_j} = 0\}$, and let $s = |\mathcal{A}|$ be the size of $\mathcal{A}$.

The next lemma establishes the error rate of the integrative linear discriminant analysis rule. Its proof follows that of Shao et al. (2011) and is omitted.

LEMMA 1. *The error rate of the integrative linear discriminant analysis rule is*

$$R_n = \frac{1}{2}\Phi\left\{\frac{(\mu_1 - \hat{\mu})^{\mathrm{T}}\hat{\beta}}{(\hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta})^{1/2}}\right\} + \frac{1}{2}\Phi\left\{\frac{(\hat{\mu} - \mu_0)^{\mathrm{T}}\hat{\beta}}{(\hat{\beta}^{\mathrm{T}}\Sigma\hat{\beta})^{1/2}}\right\}.$$

Next we show that our proposed integrative linear discriminant analysis classifier is error rate consistent. We begin with a set of regularity conditions.

Condition 1. Let $\max_{1\leqslant j\leqslant d}\sigma_{jj} \leqslant c_0$ for some constant $c_0 > 0$, where σ_{jj} are the diagonal elements of Σ.

Condition 2. Let $\Delta_d \geqslant c_0^{-1}$ for some constant $c_0 > 0$, where Δ_d is the Mahalanobis distance as defined in (1).

Condition 3. Let $\nu_n\|\beta^*\|_{1,G} + \varphi_n\|\beta^*\|_{1,G}^2 = o(1)$, where $\nu_n = M\{\Delta_d(\log d)/n\}^{1/2}$ and $\varphi_n = M\{(\log d)/n\}^{1/2}$.

Condition 1 is needed to establish the concentration results. Condition 2 avoids the extreme situation that the Bayes error converges to 1/2 when $\Delta_d \to 0$. Condition 3 requires β^* to be weakly sparse in the sense that $\|\beta^*\|_{1,G}$ is much smaller than n, which further implies that many elements in β^* are small. When $M = 1$, this becomes the weak sparsity condition of Cai & Liu (2012, Theorem 3). In Condition 3, d is allowed to diverge to infinity. Under these conditions, the next theorem establishes the relation between the error rate R_n of our method and the Bayes error R_d^*. The latter is to serve as a reference when we evaluate the error rate of a given classifier.

THEOREM 1. *Assume that Conditions 1–3 hold and* $n_k/n \to \tilde{c}_k$ $(k = 0, 1)$ *for some* $\tilde{c}_k \in (0, 1)$ *as* $n \to \infty$. *If we choose* $\lambda_n = C_0\{M\Delta_d(\log d)/n\}^{1/2}$ *for some large enough constant* C_0, *then there exist positive constants* C_1 *and* C_2 *such that, with probability at least* $1 - C_1 d^{-C_2}$,

$$\frac{R_n}{R_d^*} - 1 = O(\nu_n\|\beta^*\|_{1,G} + \varphi_n\|\beta^*\|_{1,G}^2).$$

Theorem 1 shows that the error rate R_n of our integrative classifier is of the same order as the Bayes error R_d^*. This result is crucial in establishing the guaranteed error rate improvement for our method. When $M = 1$, the convergence rate we obtain is the same as that of the linear programming discriminant rule of Cai & Liu (2012), though the two methods have different formulations. Moreover, we consider a special case where only an elementwise ℓ_1 penalty is imposed in (3). This is equivalent to setting $\alpha = 0$ in the penalty term (4). Following an argument similar to that in Theorem 1, one can show that its error rate $\tilde{R}_n$ satisfies

$$\frac{\tilde{R}_n}{R_d^*} - 1 = O\left(\nu_n\|\beta^*\|_1 + \varphi_n\|\beta^*\|_1^2\right).$$

Therefore, $\tilde{R}_n$ is of the same order as the Bayes error rate. However, when comparing this rate with that of our integrative linear discriminant analysis, we have $\|\beta^*\|_{1,G} \leqslant \|\beta^*\|_1$, since for any vector a, $\|a\|_2 \leqslant \|a\|_1$. Thus, loosely speaking, our method can approach the Bayes error rate

faster than the method only employing an elementwise ℓ_1 penalty, in the sense that the upper bound of the former classifier is tighter than that of the latter. We compare the two methods numerically in § 7.

According to Theorem 1, when more variables are involved in integrative linear discriminant analysis, the difference between the classifier's error rate and the Bayes error can become larger, as $\|\beta^*\|_{1,G}$ and d would increase. This is because more unknown parameters need to be estimated, which induces a larger estimation error. However, if the additional discriminative information brought by the extra variables exceeds the estimation error they bring, the error rate R_n is guaranteed to decrease, as shown by Theorem 2.

Condition 4. Assume $\Delta_d - \Delta_p \geqslant c_1$ for some constant $c_1 > 0$, where $\Delta_{m,p}$ denotes the Mahalanobis distance contributed by variables of the mth data type and $\Delta_p = \max_{1 \leqslant m \leqslant M} \Delta_{m,p}$.

THEOREM 2. *Let R_{1n} and R_{2n} denote the error rates of the linear discriminant analysis classifier using single-type data consisting of p variables and multi-type data consisting of d variables. Assume Conditions 1–4 hold. Then, with probability tending to* 1,

$$\limsup_{n \to \infty} \frac{R_{2n}}{R_{1n}} < 1.$$

Theorem 2 establishes the guaranteed error rate improvement of multi-type analysis through our integrative classifier, and is a key contribution of this article. The increment of the Mahalanobis distance $\Delta_d - \Delta_p$ quantifies the information contained in additional variables. Theorem 1 shows that our method approaches the Bayes error at a rate of $O(\nu_n \|\beta^*\|_{1,G} + \varphi_n \|\beta^*\|_{1,G}^2)$, which eventually becomes $o(1)$ as implied by Condition 3. Condition 4 ensures that the increment $\Delta_d - \Delta_p$ is large enough to surpass this estimation error. Consequently, the error rate of our classifier is guaranteed to decrease.

Theorems 1 and 2 only require β^* to be weakly sparse, in that $\|\beta^*\|_{1,G}$ is much smaller than the sample size n as characterized by Condition 3. Moreover, a consistent linear discriminant analysis rule does not require $\mu_0 - \mu_1$ to be sparse; see also the discussion in Mai et al. (2012). Next, we show that if β^* is exactly sparse, in that many of its components are exactly zero, our method is also selection consistent. That is, it consistently excludes all nondiscriminative features and uniformly consistently estimates the coefficients of the discriminative ones. We first present the required regularity conditions:

Condition 5. $\lambda_{\min}(\Sigma) \geqslant c_0^{-1}$;

Condition 6. $\|\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leqslant c_0$;

Condition 7. $\|\Sigma_{\mathcal{B}\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leqslant (1-\alpha)(1-\epsilon)$, for some $0 < \epsilon < 1$;

Condition 8. $\|\Sigma_{\mathcal{C}\mathcal{A}}\Sigma_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leqslant (1-\epsilon)M^{-1/2}$, for some $0 < \epsilon < 1$.

Conditions 5–8 are standard for variable selection consistency (Zhao & Yu, 2006; Fan & Lv, 2011). Conditions 7 and 8 are irrepresentability conditions required by the ℓ_1 penalty. We need two different irrepresentability conditions, depending on whether or not the zero component belongs to a group whose elements are all zero, as they have different forms in the Karush–Kuhn–Tucker conditions; see the Supplementary Material. These conditions are imposed on the population covariance matrix Σ, while similar conditions hold for $\hat{\Sigma}$ with a high probability.

THEOREM 3. *Assume that Conditions 1 and 5–8 hold. If* $s^2\{(\log d)/n\}^{1/2} = o(1)$, $\min_{j_m \in \mathcal{A}} |\beta^*_{j_m}| \gg \{\Delta_d(\log d)/n\}^{\gamma/2}$ *for some* $0 < \gamma < 1$, *and we choose* $\lambda_n = C_0\{\Delta_d(\log d)/n\}^{\gamma/2}$ *for some large enough constant* C_0, *then there exist positive constants* C_1 *and* C_2 *such that, with probability at least* $1 - C_1 d^{-C_2}$, *the solution* $\hat{\beta}$ *to problem* (3) *satisfies* $\|\hat{\beta}_{\mathcal{A}} - \beta^*_{\mathcal{A}}\|_\infty \lesssim \lambda_n$ *and* $\hat{\mathcal{A}} = \mathcal{A}$.

5. ROBUST INTEGRATIVE LINEAR DISCRIMINANT ANALYSIS

In practice, multi-type data may contain outliers and be nonnormal. We show that the benefit of multi-type classification still holds in those situations.

The classification rule described in § 2·1 remains optimal when X follows an elliptical distribution (Shao et al., 2011). This family includes normal, t, and double exponential distributions. Define $\Psi(t) = \mathrm{pr}\{l^{\mathrm{T}}\Sigma^{-1/2}(X - \mu_k) \leqslant t \mid Y = k\}$ $(k = 0, 1)$, where Y is the class indicator, $l \in \mathbb{R}^d$ is a deterministic vector with unit norm, and $t \in \mathbb{R}$. Analogous to Shao et al. (2011) and Cai & Liu (2012), we assume that $\Psi(t)$ is a continuous cumulative distribution function and its density is symmetric around 0. All elliptical distributions satisfy this assumption. Moreover, we characterize the tail probability of Ψ by assuming that for some constants $0 \leqslant \zeta \leqslant 2$, $w \in \mathbb{R}$, and $c > 0$,

$$0 < \lim_{x \to \infty} \frac{x^w \exp(-cx^\zeta)}{\Psi(-x)} < \infty. \tag{9}$$

We next replace the normality assumption with a moment condition.

Condition 9. Assume that $\max_{0 \leqslant k \leqslant 1, 1 \leqslant j \leqslant d} E\{(X_j - \mu_{kj})^4 \mid Y = k\} < c_0$ for some $c_0 > 0$, where μ_{kj} is the jth component of μ_k.

Some elliptical distributions are heavy-tailed. Accordingly, we propose to replace the usual sample estimators of μ and Σ in (2) with their corresponding robust estimators $\hat{\mu}_{k,\mathrm{robust}} = (\hat{\mu}_{kj})$ and $\hat{\Sigma}_{\mathrm{robust}} = (\hat{\sigma}_{ij})$, where $\hat{\mu}_{kj}$ and $\hat{\sigma}_{ij}$ solve the equations

$$\sum_{\ell=1}^{n_k} \psi\big\{\nu(X_{k,\ell j} - \mu_{kj})\big\} = 0, \quad \sum_{k=0}^{1}\sum_{\ell=1}^{n_k} \psi\big\{\nu(X_{k,\ell i}X_{k,\ell j} - \sigma_{ij})\big\} = 0.$$

In both cases, $\psi(x) = \min\{\max(x, -1), 1\}$ is the Huber score function and $\nu > 0$ is a tuning parameter. We insert the robust estimators $\hat{\mu}_{k,\mathrm{robust}}$ and $\hat{\Sigma}_{\mathrm{robust}}$ into (3) for robust integrative linear discriminant analysis.

For any elliptical distribution, the Bayes error takes the form $\Psi(-\sqrt{\Delta_d}/2)$, which is a decreasing function of d. Under Condition 9, one can show that (Fan et al., 2017, § 4)

$$\|\hat{\mu}_{k,\mathrm{robust}} - \mu_k\|_\infty = O_{\mathrm{p}}\{(\log d)/n\}, \quad \|\hat{\Sigma}_{\mathrm{robust}} - \Sigma\|_{\max} = O_{\mathrm{p}}\{(\log d)/n\}. \tag{10}$$

Such concentration results are exactly the same as those in the Supplementary Material that we establish for integrative discriminant analysis under normality, so a similar result to Theorem 1 also holds for our robust classifier.

THEOREM 4. *Assume that the regularity conditions in Theorem 1 and Condition 9 hold. For any distribution satisfying* (9), *there exist positive constants* C_1 *and* C_2 *such that, with probability at least* $1 - C_1 d^{-C_2}$,

$$\frac{R_n}{R_d^*} - 1 = O(\nu_n \|\beta^*\|_{1,G} + \varphi_n \|\beta^*\|_{1,G}^2).$$

Theorem 4 ensures that for a large class of distributions, our robust integrative classifier continues to enjoy the guaranteed error rate improvement when using multi-type data. The choice of the robust estimators for μ_k and Σ is not unique; see Avella-Medina et al. (2018) for some alternatives. Indeed, any estimators satisfying (10) would ensure the same convergence rate as in Theorem 4. For the special case that $M = 1$, Theorem 4 achieves the same convergence rate as Cai & Liu (2012, Theorem 5), but under less restrictive conditions. For polynomial tail distributions, Cai & Liu (2012) required the existence of higher moments than the fourth moment, and can only handle $d = O(n^\gamma)$ for some $\gamma > 0$. By contrast, our robust method can handle $d = O\{\exp(n^\gamma)\}$ for $0 < \gamma < 1$, because the method of Cai & Liu (2012) was built on the sample mean and covariance estimators.

6. Integrative linear discriminant analysis with missing data

Another issue that frequently arises is that measurements of certain data types for a subset of subjects can be entirely missing. For instance, in genomics, some subjects may have both gene expression and DNA methylation measurements, while others may only have gene expression measurements (Cai et al., 2016). In neuroimaging, one group of subjects may have both structural and functional imaging scans while another may only have structural scans. In such situations, data are missing by blocks. We show that our method can be extended to handle block missing data and the guaranteed error rate improvement still follows.

A simple solution is complete case analysis. That is, we obtain the sample estimators of $\hat{\delta}$ and $\hat{\Sigma}$ using only the subset of subjects with complete observations for all data types. We can establish similar results to Theorem 1, by replacing the sample size n with n_{complete}, the number of subjects with complete observations. Since n_{complete} is only a fraction of n, and is usually very small, the convergence rate in Theorem 1 becomes rather slow.

Alternatively, using a similar idea as G. Yu in a 2016 PhD thesis from the University of North Carolina, we use as many observations as possible to estimate $\hat{\delta}$ and $\hat{\Sigma}$, by considering

$$\hat{\delta}_{\text{effective}} = (\hat{\delta}_j), \quad \hat{\delta}_j = \bar{X}_{0,j} - \bar{X}_{1,j} = \frac{1}{n_{0j}} \sum_{\ell \in S_{0j}} X_{0,\ell j} - \frac{1}{n_{1j}} \sum_{\ell \in S_{1j}} X_{1,\ell j},$$

$$\hat{\Sigma}_{\text{effective}} = (\hat{\sigma}_{ij}), \quad \hat{\sigma}_{ij} = \frac{1}{n_{ij}} \sum_{k=0}^{1} \sum_{\ell \in S_{k,ij}} (X_{k,\ell i} - \bar{X}_{k,i})(X_{k,\ell j} - \bar{X}_{k,j}).$$

Here $X_{0,\ell j}$ and $X_{1,\ell j}$ denote the jth element of the ℓth subject in classes 0 and 1, respectively, $S_{0j} = \{\ell : X_{0,\ell j} \text{ is not missing}\}$, $S_{1j} = \{\ell : X_{1,\ell j} \text{ is not missing}\}$, $S_{k,ij} = \{\ell : \text{both } X_{k,\ell i} \text{ and } X_{k,\ell j} \text{ are not missing}\}$, n_{0j} and n_{1j} are the sizes of S_{0j} and S_{1j}, and n_{ij} is the size of $S_{0,ij} \cup S_{1,ij}$. Due to the unbalanced sample size, $\hat{\Sigma}_{\text{effective}}$ itself may not be positive semidefinite. In that case, we further project it onto the cone of positive-semidefinite matrices by solving $\text{minimize}_{\lambda_{\min}(M) \geq 0} \|M - \hat{\Sigma}_{\text{effective}}\|_{\max}$. This projection can be formulated as a linear programming problem and solved efficiently (Liu et al., 2012). For notational simplicity, we still call the minimizer $\hat{\Sigma}_{\text{effective}}$. Inserting $\hat{\delta}_{\text{effective}}$ and $\hat{\Sigma}_{\text{effective}}$ into (3), we solve (3) in the presence of block missing data. Under the normality assumption, one can show that

$$\|\hat{\delta}_{\text{effective}} - \delta\|_\infty = O_{\text{p}}\left\{(\log d)/n'\right\}, \ \|\hat{\Sigma}_{\text{effective}} - \Sigma\|_{\max} = O_{\text{p}}\left\{(\log d)/n''\right\},$$

where $n' = \min_{1\leqslant j\leqslant d}\min\{n_{0j}, n_{1j}\}$ and $n'' = \min_{1\leqslant i,j\leqslant d} n_{ij}$. Based upon these concentration results and a similar derivation, we can show that Theorem 1 still holds in the presence of block missingness, if we replace ν_n and φ_n by $\nu_{\text{effective}} = M\{\Delta_d(\log d)/n_{\text{effective}}\}^{1/2}$ and $\varphi_{\text{effective}} = M\{(\log d)/n_{\text{effective}}\}^{1/2}$, where $n_{\text{effective}} = \min(n', n'')$ is the effective sample size of the entire problem. In practice, $n_{\text{effective}}$ can be much larger than n_{complete}, so the second solution is more efficient than the complete case solution in handling missing data. Moreover, the normality assumption is not essential. One can combine the two estimators in § 5 and § 6.

7. NUMERICAL STUDIES

7·1. *Simulations*

We conduct simulations to investigate the finite-sample performance, and compare four methods: our proposed integrative linear discriminant analysis classifier with the composite penalty (4) applied to all data types; the integrative classifier with an ℓ_1 penalty only; the linear discriminant classifier applied to each type separately; and a majority vote method, where a sample is classified into the class to which the majority of separate classifiers assign it. All tuning parameters in these methods are chosen by crossvalidation. We evaluate each method in terms of both classification accuracy, measured by misclassification error rate, and variable selection accuracy, measured by sensitivity and specificity. Sensitivity is defined as the proportion of nonzero β^*_{jm}s being estimated as nonzero, and specificity is defined as the proportion of zero β^*_{jm}s being estimated as zero. For the separate classification method, each criterion is taken to be the average across all types.

For each of the two classes, we simulate $n/2$ independent samples of $M = 3$ data types, each of which has p variables: $X \mid Y = 0 \sim N(\mu_0, \Sigma)$, and $X \mid Y = 1 \sim N(\mu_1, \Sigma)$. We set $(n, p) = (50, 100)$ and $(100, 200)$. We first generate β^* and Σ, then set $\mu_1 = 0$ and $\mu_0 = \Sigma\beta^*$, based on the fact that $\delta = \mu_0 - \mu_1 = \Sigma\beta^*$. We consider two scenarios of data generation: different data types contain either a common set of variables, or different sets of variables.

In the first scenario, which we refer to as Example A, all data types share a common set of variables. That is, we generate Mp variables that are grouped into p groups. This mimics the usual multi-type data, for instance in neuroimaging, where M different brain characteristics are measured on the same p brain regions. We then generate β^* as $\beta^*_{jm} = 0{\cdot}8 \times \text{Ber}(\pi)$ $(j = 1, \ldots, 5; m = 1, \ldots, 3)$ and set the rest equal to zero. We consider $\pi = 1$ and $0{\cdot}5$. When $\pi = 1$, the first five features are discriminative in all three data types. When $\pi = 0{\cdot}5$, each of the first five features is discriminative in one data type with probability $0{\cdot}5$. We set $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{d\times d}$, where $d = Mp$, such that $\sigma_{ij} = 1$ for $i = j$, $\sigma_{ij} = 0{\cdot}2$ for $|i - j| = 1$, $\sigma_{ij} = 0{\cdot}1$ for $|i - j| = 2$ and $\sigma_{ij} = 0$ otherwise.

In the second scenario, we let 20% of the variables be shared by all three data types, 40% be shared by two types, and 40% belong to only one type. We further consider two different patterns of discriminative features and refer to them as Examples B and C. In Example B, we set the first two β^*_j to $0{\cdot}3 \times \text{Ber}(\pi)$ and the rest to zero for features appearing in two or three types; and we set the first β^*_j to $0\cdot3 \times \text{Ber}(\pi)$ and the rest to zero for features appearing in one type. In Example C, we set β^* as follows: for type one, we set the first five β^*_j to $0{\cdot}2 \times \text{Ber}(\pi)$ and the rest to zero for features appearing in two or three types; and we set the first β^*_j to $0{\cdot}2 \times \text{Ber}(\pi)$ and the rest to zero for features appearing in one type. For data types two and three, the pattern is the same except that the nonzero coefficients take values of $\beta^*_j = -0{\cdot}3 \times \text{Ber}(\pi)$ and $\beta^*_j = 0{\cdot}1 \times \text{Ber}(\pi)$ respectively. Finally, we set $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0{\cdot}8^{|i-j|}$, which allows all variables to be correlated.

Table 1 reports the average criteria, with standard errors in parentheses, all in percentages, over 100 replications. We report $\Delta_d - \Delta_p$, which quantifies the extra discriminative information from

Table 1. *Classification and variable selection accuracy* (%)

Example A

	$\Delta_d - \Delta_p = 12{\cdot}2$, Bayes error = 2%				$\Delta_d - \Delta_p = 3{\cdot}1$, Bayes error = 11%			
	$n = 50, p = 100, \pi = 1$				$n = 50, p = 100, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	2(3)	3(4)	6(4)	15(7)	11(8)	13(10)	14(10)	20(11)
Sensitivity	81(20)	57(15)	41(11)	41(8)	83(23)	60(25)	33(13)	32(12)
Specificity	94(13)	100(0)	97(1)	98(1)	94(13)	99(2)	98(1)	96(2)
	$n = 100, p = 200, \pi = 1$				$n = 100, p = 200, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	2(2)	3(2)	6(3)	15(6)	11(8)	14(9)	15(7)	24(7)
Sensitivity	90(15)	70(9)	45(10)	43(7)	94(12)	78(22)	34(13)	34(12)
Specificity	97(8)	100(0)	99(0)	99(0)	98(7)	100(0)	99(0)	99(1)

Example B

	$\Delta_d - \Delta_p = 2{\cdot}9$, Bayes error = 10%				$\Delta_d - \Delta_p = 1{\cdot}1$, Bayes error = 22%			
	$n = 50, p = 100, \pi = 1$				$n = 50, p = 100, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	10(3)	12(6)	14(3)	23(5)	24(10)	26(11)	34(17)	39(12)
Sensitivity	63(20)	52(15)	16(10)	16(5)	66(31)	54(33)	3(0)	4(0)
Specificity	96(2)	95(2)	93(2)	91(2)	88(7)	92(3)	98(2)	98(2)
	$n = 100, p = 200, \pi = 1$				$n = 100, p = 200, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	12(6)	14(6)	18(6)	26(6)	26(11)	27(16)	38(19)	40(16)
Sensitivity	45(20)	42(15)	13(10)	18(5)	51(48)	42(52)	3(0)	4(0)
Specificity	85(3)	84(3)	88(4)	82(5)	89(7)	90(3)	97(1)	96(3)

Example C

	$\Delta_d - \Delta_p = 5{\cdot}3$, Bayes error = 8%				$\Delta_d - \Delta_p = 1{\cdot}7$, Bayes error = 21%			
	$n = 50, p = 100, \pi = 1$				$n = 50, p = 100, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	9(4)	15(3)	18(9)	26(7)	22(7)	33(13)	35(10)	38(8)
Sensitivity	61(17)	37(21)	24(10)	26(10)	65(18)	33(32)	13(13)	18(11)
Specificity	92(23)	94(23)	81(8)	79(9)	90(10)	96(7)	89(10)	84(8)
	$n = 100, p = 200, \pi = 1$				$n = 100, p = 200, \pi = 0{\cdot}5$			
	iLDA	iLDA-ℓ_1	m-vote	sLDA	iLDA	iLDA-ℓ_1	m-vote	sLDA
Error rate	9(1)	13(2)	15(5)	25(5)	22(7)	29(11)	34(9)	38(8)
Sensitivity	58(12)	39(14)	22(3)	21(4)	54(19)	43(27)	11(11)	13(8)
Specificity	98(9)	100(1)	90(3)	89(5)	96(8)	97(4)	95(5)	92(5)

iLDA, the integrative linear discriminant analysis classifier with the composite penalty (4); iLDA-ℓ_1, the integrative linear discriminant analysis classifier with the ℓ_1 penalty only; sLDA, the linear discriminant analysis classifier applied to each individual type separately; m-vote, a majority vote based on the class assigned by sLDA.

the additional variables. We also report the Bayes error, which serves as the baseline. In all cases, our method attains a classification error that is close to the Bayes error, and achieves reasonably high sensitivity and specificity. Compared to the separate classifier and the majority vote method, the larger the value of $\Delta_d - \Delta_p$, the more pronounced an improvement our method achieves in classification accuracy. This supports our theoretical findings. Moreover, our method achieves better selection accuracy, with a higher sensitivity and a comparable specificity. Comparing our proposed integrative classifier with the composite penalty with the one using the ℓ_1 penalty only, the former attains both a smaller classification error and better selection accuracy than the latter, by leveraging a more flexible penalty function. When the effect of discriminative variables is large,

such as in Example A, although the method with only the ℓ_1 penalty misses many important variables, its classification error remains small, because the selected variables already enable a reasonable classification. However, when there are many discriminative variables with small effects, such as in Example C, it yields considerably worse classification performance.

In summary, our simulations have shown that the proposed integrative classifier consistently outperforms the alternative classifiers for various scenarios: when all data types share the same or have a different set of variables, when the discriminative variables differ in magnitude and in size, and when the covariance matrix takes different structures. We also consider scenarios where the data follow a heavy-tailed distribution, or when there are missing values. See the Supplementary Material.

7·2. *Positron emission tomography data analysis*

We illustrate our method using a neuroimaging study of Alzheimer's disease, which is characterized by progressive and irreversible impairment of cognitive and memory functions, and is the leading form of dementia in elderly subjects. With the ageing of the worldwide population, it has become imperative to understand, diagnose, and treat this disorder. Positron emission tomography is a nuclear medicine, functional imaging technique widely used in Alzheimer's disease research. It is designed to measure metabolic processes and protein accumulations, by detecting gamma rays emitted indirectly by a positron-emitting radionuclide, or tracer. Different metabolic processes and proteins can be quantified with different tracers: the fluorodeoxyglucose tracer measures glucose metabolism in the brain, the Pittsburgh compound B tracer measures deposition of amyloid-β protein, and the AV-1451 tracer measures accumulation of tau protein.

Our positron emission tomography study includes 35 subjects, each with a glucose metabolism scan, a tau scan, and an amyloid-β scan. Given the amyloid-β measurement, each subject has been classified as amyloid positive or amyloid negative. There are well-validated standard methods for the assessment of amyloid deposition and thresholds for subject classification as amyloid positive or negative (Landau et al., 2013). Moreover, following a common brain atlas that divides the brain into a set of regions of interest, both the glucose metabolism image and the tau image have been summarized by a vector, with each individual entry measuring glucose metabolism or tau accumulation within the same brain region. Brain atlas-based partition has been frequently used in neuroimaging analysis. Without partition, each image scan would take the form of a multi-dimensional array. An integrative classification analysis of array data is of interest, but is beyond the scope of this article. One of the scientific goals of this study is to classify the subjects as amyloid positive or amyloid negative given the glucose metabolism and tau images. Classification of amyloid positivity is a potentially highly useful application in Alzheimer's disease research, and could be readily disseminated in the therapeutic trial arena.

We randomly split the data into three folds, with two folds for training and one for testing. We repeat the splitting 100 times. We compare four methods: our proposed integrative classifier applied to both the glucose metabolism image and the tau image, the integrative classifier using only the ℓ_1 penalty, and the separate classifier applied to only one imaging data type. Table 2 reports the results. Our integrative classifier outperforms the alternatives. We then apply our method to the entire dataset. Figure 1 displays the estimated classification direction imposed on a brain template. The selected top brain regions include the amygdala, the transverse temporal gyrus, the cuneus, and the inferior temporal gyrus. Such findings are generally consistent with the neuroscience literature. For instance, the amygdala has been observed to be severely affected in Alzheimer's disease (Poulin et al., 2011), and has been postulated to be one of the central regions for early amyloid-β changes (Mann et al., 1987). The cuneus is also found to show elevated amyloid-β concentrations that are associated with increased brain atrophy (Tosun et al., 2011).

Table 2. *Misclassification errors* (%) *based on three-fold data splitting*

	Min	1st-Q	Median	3rd-Q	Max	Mean	SD
iLDA	8	17	20	23	32	20	4
iLDA-ℓ_1	13	18	21	27	45	22	6
sLDA glucose	20	24	28	31	46	28	5
sLDA tau	14	20	23	26	46	23	5

iLDA, the integrative linear discriminant analysis classifier with the composite penalty (4); iLDA-ℓ_1, the integrative linear discriminant analysis classifier with the ℓ_1 penalty only; sLDA glucose, the linear discriminant analysis classifier applied to the glucose metabolism image; sLDA tau, the linear discriminant analysis classifier applied to the tau image; Min, minimum; 1st-Q, first quartile; 3rd-Q, third quartile; Max, maximum; SD, standard deviation.

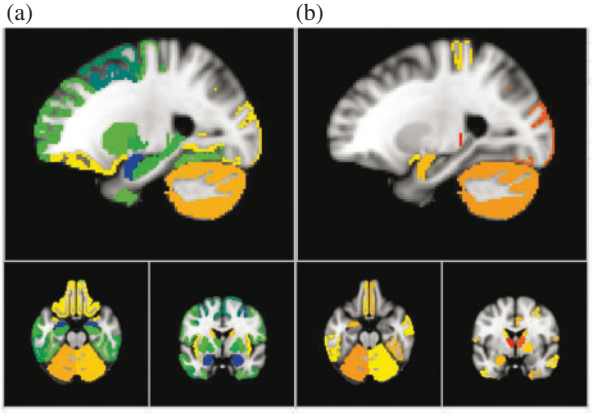

Fig. 1. The discriminative regions of interest selected by our integrative classifier. Panel (a) shows the selected regions corresponding to tau images; (b) shows those corresponding to glucose metabolism images. Each panel contains sagittal, axial, and coronal views. All the estimated coefficients are rescaled to the range -1 to 1. The colour reflects their magnitudes, with red corresponding to 1 and blue to -1. The blue region in (a) is the amygdala.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation and National Institutes of Health. The authors thank Dr William Jagust and Dr Samuel Lockhart for helpful discussions on the positron emission tomography study. The authors also thank the editor, associate editor and two referees for valuable comments and suggestions.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, technical lemmas, and additional simulation results.

References

Avella-Medina, M., Battey, H., Fan, J. & Li, Q. (2018). Robust estimation of high dimensional covariance and precision matrices. *Biometrika* **105**, 271–84.

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.

Cai, T., Cai, T. T. & Zhang, A. (2016). Structured matrix completion with applications to genomic data integration. *J. Am. Statist. Assoc.* **111**, 621–33.

CAI, T. & LIU, W. (2012). A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.* **106**, 1566–77.
CLEMMENSEN, L., HASTIE, T. J., WITTEN, D. M. & ERSBOLL, B. (2011). Sparse discriminant analysis. *Technometrics* **53**, 406–13.
DAI, Z., YAN, C., WANG, Z., WANG, J., XIA, M., LI, K. & HE, Y. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier. *NeuroImage* **59**, 2187–95.
FAN, J., FENG, Y. & TONG, X. (2012). A ROAD to classification in high dimensional space: The regularized optimal affine discriminant. *J. R. Statist. Soc.* B **74**, 745–71.
FAN, J., LI, Q. & WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Statist. Soc.* B **79**, 247–65.
FAN, J. & LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Info. Theory.* **57**, 5467–84.
FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). A note on the group lasso and a sparse group lasso. *arXiv*: 1001.0736.
HAN, F., ZHAO, T. & LIU, H. (2013). CODA: High dimensional copula discriminant analysis. *J. Mach. Learn. Res.* **14**, 629–71.
HAND, D. J. (2006). Classifier technology and the illusion of progress. *Statist. Sci.* **21**, 1–14.
LANDAU, S. M., LU, M., JOSHI, A. D., PONTECORVO, M., MINTUN, M. A., TROJANOWSKI, J. Q., SHAW, L. M., JAGUST, W. J. & THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2013). Comparing positron emission tomography imaging and cerebrospinal fluid measurements of a β-amyloid. *Ann. Neurol.* **74**, 826–36.
LI, Q., WANG, S., HUANG, C. C., YU, M. & SHAO, J. (2014). Meta-analysis based variable selection for gene expression data. *Biometrics* **70**, 872–80.
LIU, H., HAN, F., YUAN, M., LAFFERTY, J. & WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–326.
LIU, J., HUANG, J. & MA, S. (2013). Integrative analysis of multiple cancer genomic datasets under the heterogeneity model. *Statist. Med.* **32**, 3509–21.
LIU, J., HUANG, J., ZHANG, Y., LAN, Q., ROTHMAN, N., ZHENG, T. & MA, S. (2014). Integrative analysis of prognosis data on multiple cancer subtypes. *Biometrics* **70**, 480–8.
MAI, Q., ZOU, H. & YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99**, 29–42.
MANN, D. M. A., TUCKER, C. M. & YATES, P. O. (1987). The topographic distribution of senile plaques and neurofibrillary tangles in the brains of non-demented persons of different ages. *Neuropathol. Appl. Neurobiol.* **13**, 123–39.
NESTEROV, Y. (2013). Gradient methods for minimizing composite functions. *Math. Program.* B **140**, 125–61.
PARIKH, N. & BOYD, S. (2014). Proximal algorithms. *Foundat. Trends Optimiz.* **1**, 123–231.
POULIN, S. P., DAUTOFF, R., MORRIS, J. C., BARRETT, L. F. & DICKERSON, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiat. Res.* **194**, 7–13.
RICHARDSON, S., TSENG, G. C. & SUN, W. (2016). Statistical methods in integrative genomics. *Ann. Rev. Statist. Appl.* **3**, 181–209.
SHAO, J., WANG, Y., DENG, X. & WANG, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39**, 1241–65.
SHEN, R., WANG, S. & MO, Q. (2013). Sparse integrative clustering of multiple omics data sets. *Ann. Appl. Statist.* **7**, 269–94.
SIMON, N., FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2013). A sparse-group lasso. *J. Comp. Graph. Statist.* **22**, 231–45.
TOSUN, D., SCHUFF, N., MATHIS, C. A., JAGUST, W. & WEINER, M. W. (2011). Spatial patterns of brain amyloid-beta burden and atrophy rate associations in mild cognitive impairment. *Brain* **134**, 1077–88.
ULUDAG, K. & ROEBROECK, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage* **102**, 3–10.
WITTEN, D. M. & TIBSHIRANI, R. J. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc.* B **73**, 753–72.
WU, M. C., ZHANG, L., WANG, Z., CHRISTIANI, D. C. & LIN, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics* **25**, 1145–51.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHANG, D., WANG, Y., ZHOU, L., YUAN, H., SHEN, D. & THE ALZHEIMERS DISEASE NEUROIMAGING INITIATIVE (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* **55**, 856–67.
ZHAO, P. & YU, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

[*Received on* 8 *March* 2017. *Editorial decision on* 31 *May* 2018]