TECHNOMETRICS
2020, VOL. 62, NO. 2, 196–205
https://doi.org/10.1080/00401706.2019.1610069

Taylor & Francis
Taylor & Francis Group

Matrix Linear Discriminant Analysis

Wei Hu[a], Weining Shen[a], Hua Zhou[b], and Dehan Kong[c]

[a]Department of Statistics, University of California, Irvine, CA; [b]Department of Biostatistics, University of California, Los Angeles, CA; [c]Department of Statistical Sciences, University of Toronto, Toronto

ABSTRACT

We propose a novel linear discriminant analysis (LDA) approach for the classification of high-dimensional matrix-valued data that commonly arises from imaging studies. Motivated by the equivalence of the conventional LDA and the ordinary least squares, we consider an efficient nuclear norm penalized regression that encourages a low-rank structure. Theoretical properties including a nonasymptotic risk bound and a rank consistency result are established. Simulation studies and an application to electroencephalography data show the superior performance of the proposed method over the existing approaches.

ARTICLE HISTORY
Received September 2018
Revised March 2019

KEYWORDS
Linear discriminant analysis; Low rank; Matrix data; Nuclear norm; Rank consistency; Risk bound.

1. Introduction

Modern technologies have generated a large number of datasets that possess a matrix structure for classification purpose. For example, in neuropsychiatric disease studies, it is often of interest to evaluate the prediction accuracy of prognostic biomarkers by relating two-dimensional imaging predictors, for example, electroencephalography (EEG) and magnetoencephalography, to clinical outcomes such as diagnostic status (Mu and Gage 2011). In this article, we focus on extending one of the most commonly used classification methods, Fisher linear discriminant analysis (LDA) to matrix-valued predictors. Progress has been made in recent years on developing sparse LDA using ℓ_1-regularization (Tibshirani 1996), including Shao et al. (2011), Fan, Feng, and Tong (2012), and Mai, Zou, and Yuan (2012). However, all these methods only deal with vector-valued covariates; and it remains challenging to accommodate the matrix structure. Naively transforming the matrix data into a high-dimensional vector will result in unsatisfactory results for several reasons. First, vectorization destroys the structural information within the matrix such as shapes and spatial correlations. Second, turning a $p \times q$ matrix into a $pq \times 1$ vector generates unmanageably high dimensionality. For example, estimating the population precision matrix for LDA can be troublesome if $pq \gg n$. Third, ℓ_1-regularization does not necessarily work well because the underlying two-dimensional signals are usually approximately low-rank rather than ℓ_0-sparse.

Recently, there are some development of regression methods for matrix data. Chen, Dong, and Chan (2013) invented an adaptive nuclear norm penalization approach for low-rank matrix approximation. Zhou and Li (2014) proposed a class of regularized matrix regression methods based on spectral regularization. Wang and Zhu (2017) developed a generalized scalar-on-image regression model via total variation. Kong et al. (2019) proposed a low-rank linear regression model with high-dimensional matrix response and high-dimensional scalar covariates, while Hu, Kong, and Shen (2019) developed a nonparametric matrix response regression model.

In this article, we propose a new matrix LDA approach by building on the equivalence between the classical LDA and the ordinary least squares. We formulate the binary classification as a nuclear norm penalized least-squares problem, which efficiently exploits the low-rank structure of the two-dimensional discriminant direction matrix. The involved optimization is amenable to the accelerated proximal gradient method. Although our problem is formulated as a penalized regression problem, a fundamental difference is that the covariates $\mathbf{X}_i$ and the residuals ϵ_i are no longer independent in our case. This requires extra effort for developing the risk bound and rank consistency result. The risk bound is explicit in terms of the rank of the image, image size, sample size, and the eigenvalues of the covariance matrix for the image covariates. This result also implies estimation consistency provided the $p \times q$ image satisfies $\max(p, q) = o(n/\log^3 n)$. Under stronger conditions, we show that the rank of the coefficient matrix can be consistently estimated as well. The proof is based on exploiting the spectral norm of random matrices with mixture-of-Gaussian components and extending the results in Bach (2008) to allow diverging matrix dimensions. Finally, we prove that our method enjoys classification error consistency.

It is worth noting that the 2D image classification problem has been studied by Zhong and Suslick (2015), where they proposed a penalized matrix discriminant analysis (PMDA) method that projects the matrix coefficient into row space and column space separately. Those two projections are then estimated iteratively and integrated together for classification. Compared with PMDA, we make the following contributions. First, the rank of the PMDA is set as one because of the separability assumption, while we allow the rank of the direction matrix to take general positive integer values and the rank can then be selected by a data driven procedure. Our rank

CONTACT Dehan Kong *kongdehan@utstat.toronto.edu* Department of Statistical Sciences, University of Toronto, Room 6027G, Sidney Smith Hall, 100 St. George St., Toronto, ON M5S 3G3, Canada.

assumption is more flexible in practice and hence often leads to a lower misclassification error in the numerical studies. Second, our method adopts a direct estimation approach by solving a nuclear norm penalized regression problem, which is computationally much faster compared with PMDA, where the estimation involves an iterative procedure for calculating the inverse of covariance matrices during each iteration. Third, our method can handle the high-dimensional data when image dimensions p and q are much larger than the sample size, which is the case for many applications; while PMDA cannot handle the case when $p + q > n$. Finally, we have provided theoretical guarantee for our estimator when p and q diverge with n. In particular, we have developed an nonasymptotic error bound for the estimated LDA direction, as well as results on rank consistency and classification error consistency. These results are stronger compared with the root-n consistency of the LDA direction in Zhong and Suslick (2015), where both p and q are assumed to be fixed.

2. Method

We first define some useful notations. Let vec$(\cdot)$ be a vectorization operator, which stacks the entries of a matrix into a column vector. The inner product between two matrices of same size is defined as $\langle \mathbf{M}, \mathbf{N} \rangle = \text{tr}(\mathbf{M}^{\mathrm{T}}\mathbf{N}) = \langle \text{vec}(\mathbf{M}), \text{vec}(\mathbf{N}) \rangle$.

Consider a binary classification problem, where $\mathbf{X}$ is a two-dimensional image covariate with dimension $p \times q$ and $G = 1, 2$ denotes the class labels. The LDA assumes that $\text{vec}(\mathbf{X}) \mid G = g \sim N(\mu_g, \Sigma)$, $\text{pr}(G = 1) = \pi_1$, and $\text{pr}(G = 2) = \pi_2$. Suppose we have n subjects with n_1 subjects belonging to class 1 and $n_2 = n - n_1$ subjects to class 2. It is well known that LDA is connected to the linear regression with the class labels as responses (Duda, Hart, and Stork 2012; Mika 2002). When $pq < n$, the classical LDA is equivalent to solving

$$(\hat{\beta}_0^{\text{ols}}, \hat{\mathbf{B}}^{\text{ols}}) = \underset{\beta_0, B}{\arg\min} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \langle \mathbf{X_i}, \mathbf{B} \rangle \Big)^2, \qquad (1)$$

where $\mathbf{X}_i$ is the image covariate from subject i, $\mathbf{B}$ is the coefficient matrix for the image covariate and it represents the direction of the linear discriminant classifier, β_0 is the intercept, and the response $y_i = -n/n_1$ if subject i is in class 1, and $y_i = n/n_2$ if subject i is in class 2. Although this connection gives the exact LDA direction when $pq < n$, it has two potential drawbacks. First, when $pq > n$, the equivalence between Fisher LDA and (1) is lost because of the non-uniqueness of solution. Second, the formulation (1) does not incorporate the 2D image structure when estimating the direction because $\langle \mathbf{X}_i, \mathbf{B} \rangle = \langle \text{vec}(\mathbf{X}_i), \text{vec}(\mathbf{B}) \rangle$. These motivate us to consider a penalized version of (1) as follows

$$(\hat{\beta}_0, \hat{\mathbf{B}}) = \underset{\beta_0, \mathbf{B}}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B} \rangle \Big)^2 + \omega_n \|\mathbf{B}\|_*, \qquad (2)$$

where the nuclear norm $\|\mathbf{B}\|_* = \sum_j \sigma_j(\mathbf{B})$ and $\sigma_j(\mathbf{B})$s are the singular values of the matrix $\mathbf{B}$. The nuclear norm $\|\mathbf{B}\|_*$ plays an important role because it imposes a low rank structure in the estimated direction $\hat{\mathbf{B}}$. An alternative choice is to add a Lasso type penalty, that is, $\omega_n \|\mathbf{B}\|_{1,1} = \omega_n \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}|$, where b_{jk} is the jkth element of $\mathbf{B}$. However, the Lasso type penalty can only identify at most n nonzero components, and for most cases in imaging studies, the signal is usually not that sparse. More importantly, the Lasso type of penalty ignores the matrix structure because it is equivalent to vectorizing the array and applying sparse LDA. Once $\hat{\mathbf{B}}$ from (2) is obtained, a naive classification rule will assign the ith subject to class 2 if $\langle \mathbf{X}_i, \hat{\mathbf{B}} \rangle + \hat{\beta}_0 > 0$. However, it can be shown that the intercept $\hat{\beta}_0$ obtained from (2) is not optimal. Instead, we use the optimal intercept $\tilde{\beta}_0$ that minimizes the training error after obtaining $\hat{\mathbf{B}}$. Mai, Zou, and Yuan (2012) showed that the intercept of LDA actually has a closed form. Their derivations can be easily applied to our case. In particular, if $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} \text{vec}(\hat{\mathbf{B}}) > 0$, then

$$\tilde{\beta}_0 = -(\hat{\mu}_1 + \hat{\mu}_2)^{\mathrm{T}} \text{vec}(\hat{\mathbf{B}})/2 + \text{vec}(\hat{\mathbf{B}})^{\mathrm{T}} \hat{\Sigma} \text{vec}(\hat{\mathbf{B}}) \{(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} \text{vec}(\hat{\mathbf{B}})\}^{-1} \log(n_2/n_1), \qquad (3)$$

where $\hat{\mu}_g$ is the sample mean for subjects in class g and $\hat{\Sigma}$ is the estimated covariance matrix. If $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}} \text{vec}(\hat{\mathbf{B}}) < 0$, we can plug $-\hat{\mathbf{B}}$ into (3) to obtain the optimal intercept $\tilde{\beta}_0$. The optimal classification rule is to assign the ith subject to class 2 if $\langle \mathbf{X}_i, \hat{\mathbf{B}} \rangle + \tilde{\beta}_0 > 0$.

For any fixed ω_n, the optimization problem in (2) can be solved using the accelerated proximal gradient method (Nesterov 1983; Beck and Teboulle 2009). Zhou and Li (2014) studied the algorithm for the nuclear norm regularized matrix regression. As we know, nuclear norm is not differentiable. Fortunately, its subderivative $\partial \|.\|_*$ exists. Therefore (2) has local minima $(\hat{\beta}_0, \hat{\mathbf{B}})$ if and only if $0 \in -\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \epsilon_i + \omega_n \partial \|\hat{\mathbf{B}}\|_*$. Thanks to the convexity of nuclear norm, the local minima is global as well. Based on these facts, singular value thresholding method for nuclear norm regularization was deployed for building blocks of Nesterov's method. Compared with the classical gradient decent method with convergence of $O(t^{-1})$, where t denotes the number of iteration, Nesterov's accelerated gradient decent method achieves convergence rate of $O(t^{-2})$. It differs from traditional algorithms by utilizing the estimators from previous two iterations to generate the next estimator. For computational algorithm, we use the `matrix_sparsereg` function in the Matlab TensorReg Toolbox (*https://hua-zhou.github.io/TensorReg/*) for solving nuclear norm penalized matrix regression. It implements an optimal Nesterov acceleration of the proximal gradient algorithm. Actually, one contribution of our article is to link matrix LDA to regularized matrix regression so that the computational machinery developed for the latter can be applied to matrix LDA problems. For tuning of the ω_n, we adopt the BIC derived by Zhou and Li (2014) under the nuclear norm regularized matrix regression framework.

3. Theory

In this section we discuss the theoretical properties of the proposed regularization estimator. Denote the residuals $\epsilon_i = y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B} \rangle$ and the true coefficient matrix by $\mathbf{B}_0$. By the equivalence between LDA direction and least squares, we know $\text{vec}(\mathbf{B}_0)$ can be written as $c\Sigma^{-1}(\mu_2 - \mu_1)$ for some positive

constant c. Consider the singular value decomposition $\mathbf{B}_0 = \mathbf{U}_0 \mathrm{Diag}(S_0)\mathbf{V}_0^{\mathrm{T}}$ with $\mathbf{U}_0 \in \mathrm{R}^{p\times r}$ and $\mathbf{V}_0 \in \mathrm{R}^{q\times r}$. Let $\mathbf{U}_{0\perp} \in \mathrm{R}^{p\times(p-r)}$ and $\mathbf{V}_{0\perp} \in \mathrm{R}^{q\times(q-r)}$ be (arbitrary) orthogonal complements of $\mathbf{U}_0$ and $\mathbf{V}_0$, respectively. We make the following assumptions.

(A1) We assume that the second-order moment of the covariate $\mathbf{X}$, $\mathrm{E}(\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^{\mathrm{T}})$, denoted by Σ_{xx}, satisfies $\lambda_l \le \lambda_{\min}(\Sigma_{xx}) \le \lambda_{\max}(\Sigma_{xx}) \le \lambda_u$, where $\lambda_{\min}(\Sigma_{xx})$ and $\lambda_{\max}(\Sigma_{xx})$ are the smallest and largest eigenvalues of Σ_{xx}, respectively, and λ_l, λ_u are some positive constants.

(A2) Let $r = \mathrm{rank}(\mathbf{B}_0)$ be the unknown rank of the true coefficient matrix $\mathbf{B}_0$. Define $\Lambda \in \mathrm{R}^{(p-r)\times(q-r)}$ as

$$\mathrm{vec}(\Lambda) = \{(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma^{-1}(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})\}^{-1} \times \{(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma^{-1}(\mathbf{V}_0 \otimes \mathbf{U}_0)\mathrm{vec}(\mathbf{I})\}.$$

We assume its spectral norm $\|\Lambda\|_2 < 1$.

(A3) Assume the quantities ω_n, $\{\min(p,q)\}^{1/2}n^{-1/2}\omega_n^{-1}$, $\min(p,q)n^{-1/2}, \omega_n p^{1/2}q^{1/2}\min(p,q)$ tend to 0 as $n \to \infty$.

(A4) There exists a positive constant C_μ such that $\|\mu_2 - \mu_1\|_2 \le C_\mu(\sqrt{p} + \sqrt{q})$.

Condition (A1) requires bounded eigenvalues for the covariance matrix of the vectored covariate, which is standard in the literature. Condition (A2) is similar with the strict consistency condition in Bach (2008). It is needed to establish rank consistency. This condition extends the classical strong irrepresentable condition in Zhao and Yu (2006), which is commonly used for proving model selection consistency of Lasso. The major difference between our Assumption (A2) and the similar assumption in Bach (2008) is that the number of parameters is fixed in Bach (2008) while in our case the number is diverging with n. Therefore we will need to assume that the regularization parameter ω_n decays slower than the one in Bach (2008). Condition (A3) puts more requirement on the order of p, q, and w_n in order to obtain consistent rank estimation in addition to consistent coefficient estimation. This is expected since rank estimation consistency is usually not implied by parameter estimation consistency. Condition (A4) can be viewed as a sparsity assumption on B_0. Recall the solution (the slope) to classical LDA problem with vector covariates depends on the term $\mu_2 - \mu_1$. This assumption essentially implies that there are at most $O(\max(p,q))$ number of $O(1)$ elements in the true coefficient matrix $\mathbf{B}_0$ given the rank of $\mathbf{B}_0$ is fixed.

Next, we briefly review two important concepts, namely decomposable regularizer and strong convex loss function, proposed by Negahban et al. (2012) and highlight their connection to the risk bound property for our estimator.

Definition 1. A regularizer $R(\cdot)$ is decomposable with respect to a given pair of subspaces (M, N) where $M \subseteq N^{\perp}$ if

$$R(u+v) = R(u) + R(v) \quad \text{for all } u \in M, v \in N.$$

In our setting, $R(\cdot)$ is the nuclear norm. Considering a matrix $\mathbf{B} \in \mathcal{R}^{p\times q}$ to be estimated, we observe that nuclear norm is decomposable given a pair of subspaces

$$M(\mathbf{U},\mathbf{V}) := \{\mathbf{B} \in \mathcal{R}^{p\times q} \mid \mathrm{row}(\mathbf{B}) \subseteq \mathbf{V},\ \mathrm{col}(\mathbf{B}) \subseteq \mathbf{U}\},$$
$$N(\mathbf{U},\mathbf{V}) := \{\mathbf{B} \in \mathcal{R}^{p\times q} \mid \mathrm{row}(\mathbf{B}) \subseteq \mathbf{V}^{\perp},\ \mathrm{col}(\mathbf{B}) \subseteq \mathbf{U}^{\perp}\},$$

where $\mathbf{U}, \mathbf{V}$ represent $\mathbf{B}$'s left and right singular vectors. For any pair of matrices $\mathbf{B}_1 \in M$ and $\mathbf{B}_2 \in N$, the inner product of $\mathbf{B}_1, \mathbf{B}_2$ is 0 due to their mutually orthogonal rows and columns. Hence, we conclude $R(\mathbf{B}_1 + \mathbf{B}_2) = R(\mathbf{B}_1) + R(\mathbf{B}_2)$. Since we assume the true parameter has a low-rank structure, we expect the regularized estimator to have a large value of projection on $M(\mathbf{U},\mathbf{V})$ and a relatively small-valued projection on $N(\mathbf{U},\mathbf{V})$.

When the loss function $L(\hat{\beta}_0, \hat{\mathbf{B}}_{\omega_n})$ defined as $\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \langle \mathbf{X}_i, \hat{\mathbf{B}}_{\omega_n}\rangle\right)^2$ is close to $L(\beta_0, \mathbf{B}_0)$, it is insufficient to claim $\hat{\mathbf{B}}_{\omega_n} - \mathbf{B}_0$ is small if the loss function L is relatively flat. This is why the strong convexity condition is required.

Definition 2. For a given loss function L and norm $\|.\|$, we say L is strong convex with curvature k_L and tolerance function τ_L if

$$\delta L(\Delta, \mathbf{B}_0) \ge k_L\|\Delta\|^2 - \tau_L^2(\mathbf{B}_0), \quad \text{for any } \delta \in \mathcal{C}(M, N; \mathbf{B}_0),$$

where $\mathcal{C}(M, N; \mathbf{B}_0) := \{\Delta \in \mathcal{R}^{p\times q} \mid R(\Delta_N) \le 3R(\Delta_{N^{\perp}}) + 4R(\mathbf{B}_{0N})\}$.

Now we are ready to state the main result on the risk bound for our estimate. The proof is provided in Appendix B.

Theorem 1. Suppose that (A1) and (A4) hold. Let $\hat{\mathbf{B}}$ be the solution to (2). If

$$\omega_n \ge \frac{12(\log n)^{3/2}(C_\mu + \lambda_u^{1/2})(\sqrt{p} + \sqrt{q} + \sqrt{\log n})}{\sqrt{n}},$$

then with probability of at least $1 - Cn^{-1}$ for some constant $C > 0$,

$$\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 + |\hat{\beta} - \beta_0^*|^2 \le 9\frac{\omega_n^2}{\lambda_l}r,$$

where $\beta_0^* = \beta_0 - \pi_2^{-1}\{c - 1 + (\pi_2 - \pi_2^2)(\mathbf{D}^{\mathrm{T}}\Sigma^{-1}\mathbf{D})\}$ and c is some positive constant.

Theorem 1 gives a nonasymptotic risk bound for the proposed estimators. In other words, the results hold for any positive ω_n satisfying the conditions there. However, in order to ensure the consistency of the proposed estimators, we will need the risk bound to go to 0, which requires $\omega_n \to 0$ and $\max(p,q) = o\left(n/(r\log^3 n)\right)$. If the rank of $\mathbf{B}_0$ is fixed, then both p and q can diverge with n at the order of $o(n/\log^3 n)$ and their product $pq > n$. This result is compatible with Theorem 1 in Raskutti and Yuan (2015). Note that the estimated intercept $\hat{\beta}$ converges to β_0^*, which deviates from the truth β_0. This is expected because the solution to OLS is only equivalent with LDA's solution in terms of the slope $\mathbf{B}$, not on β_0. More precisely, for OLS, by taking the derivative of squared loss function with respect to β_0 and set it to 0, we essentially require $E(\epsilon) = 0$. However, this does not hold in our case. Instead we need to shift the residual ϵ by d to balance off the bias in the cross-product term $\mathrm{E}(\epsilon\mathbf{X})$. The proof of the theorem uses Gaussian comparison inequality which allows us to deal with $\mathrm{vec}(\mathbf{X})$ following a general Gaussian distribution instead of standard Gaussian distribution given that the largest singular value of Σ_{xx} is bounded. Based on this connection, we further use

concentration property of spectral norm of Gaussian random matrices.

Next we show that $\hat{\mathbf{B}}$ is rank-consistent under stronger conditions.

Theorem 2. Suppose that (A1)–(A4) hold. Then the estimate $\hat{\mathbf{B}}$ is rank-consistent, that is, $P(\text{rank}(\hat{\mathbf{B}}) = \text{rank}(\mathbf{B}_0)) \to 1$ as $n \to \infty$.

Similar to Lasso, estimation consistency does not guarantee correct rank estimation for matrix regularization. In fact, the assumptions here are stronger than those in Theorem 1. For example, Theorem 1 allows $p + q = o(n/\log^3 n)$ while Theorem 2 requires $\max(p, q) = o\left(n^{1/3}\log^{-3/2} n\right)$ if $\min(p, q) = O(1)$. The proof is based on the arguments in Bach (2008) with modifications to allow diverging p and q.

Remark 1. Although nuclear norm penalized least squares is used to estimate the classification direction, there is a fundamental difference between our theorems and the theoretical results derived for nuclear norm penalized least-squares regression (Bach 2008; Negahban et al. 2012). The previous work assumes that the data obey a linear regression model with covariates-independent additive noise, which is not true in our case. In particular, the covariates $\mathbf{X}_i$ and the residuals ϵ_i are no longer independent in our problem, which brings additional challenges in developing theoretical results.

Next we state a classification error consistency result. To be consistent with the notation in the classification literature, for subject i, we use $Y_i \in \{-1, 1\}$ to denote its true label, $\hat{f}_n(\mathbf{X_i})$ as the classified label for which $\hat{f}_n$ is the classification rule obtained by solving (2), and $l(Y_i, f(\mathbf{X_i})) = I\{Y_i \neq \text{sign}(f(\mathbf{X_i}))\}$ as the 0–1 loss function. Define the risk of $\hat{f}_n$ as $R(\hat{f}_n) = E_{\mathbf{X}} l(Y, \hat{f}_n(\mathbf{X}))$ and the Bayes risk as $R^* = \inf_f R(f)$. In addition, we assume that the true label Y_i given $\mathbf{X_i}$ is determined by the linear classification rule with coefficients β_0^* and $\mathbf{B}_0$. Then the following theorem shows that the proposed classifier achieves the Bayes optimal risk under certain conditions. The proof, given in Appendix B, is based on the general results in Zhang (2004), where the author studied the optimal Bayes error rate using a classifier obtained by minimizing a convex upper bound of the classification error function.

Theorem 3. Assume the same conditions for Theorem 1 hold and $\omega_n \to 0$. Then $R(\hat{f}_n) \to R^*$ as $n \to \infty$.

4. Numerical Results

4.1. Simulation

We conduct simulation studies to evaluate the numerical performance of our proposed method. We compare its performance with that of a few alternatives: "Lasso LDA," which adopts a naive Lasso penalty in LDA without taking into account matrix structure, the regularized matrix logistic regression (Zhou and Li 2014) using nuclear norm and Lasso penalties, denoted by "Logistic Nuclear" and "Logistic Lasso," and the PMDA approach proposed by Zhong and Suslick (2015). We generate $n \in \{100, 200, 500\}$ samples from two classes with weights $(\pi_1, \pi_2) \in \{(0.5, 0.5), (0.75, 0.25)\}$. For each class, we generate predictors from a bivariate normal distribution with means μ_g, $g = 1, 2$, and covariance Σ. We set $\mu_1 = 0$ and $\mu_2 = \Sigma\text{vec}(\mathbf{B}_0)$. The covariance matrix Σ has a 2D autoregressive structure: $\text{cov}(\mathbf{x}_{i_1,j_1}, \mathbf{x}_{i_2,j_2}) = 0.5^{|i_1-i_2|+|j_1-j_2|}$ for $1 \le i_1 \le p$ and $1 \le j_1 \le q$. The true signal $\mathbf{B}_0$ is generated based on a 64-by-64 image. We consider three settings: a cross, a triangle and a butterfly. These pictures are shown in Figure 1(a). In particular, the white color denotes value 0 and black denotes 0.05. We apply each fitted model to an independent test dataset of size 1000 and summarize the misclassification rates based on 1000 Monte Carlo replications. The results are contained in Table 1.

The results show that our method performs much better than "Lasso LDA" and "Logistic Lasso" under all scenarios. This is expected because these two methods ignore the matrix structure. For "Logistic Nuclear," it has similar misclassification rates with our method for balanced data, but does not perform as good as ours for unbalanced data. We have also plotted the estimates using nuclear norm and ℓ_1-norm from one randomly selected Monte Carlo replicate in Figure 1(b,c). It can be seen that the proposed nuclear norm regularization is much better than ℓ_1-regularization in recovering the matrix signal in different shapes. By comparing the recovery of different shapes in Column (b) in Figure 1, we find that our method works better for cross than for triangle and butterfly. This is expected since triangle and butterfly do not have the low-rank structure.

We also compare the performance of our method with that of PMDA proposed by Zhong and Suslick (2015). In Table 1, it can be seen that our proposed method has a lower misclassification rate under all scenarios. This is because we allow flexible values of the rank for the linear discriminant direction $\mathbf{B}$, while in Zhong and Suslick (2015), their assumption is equivalent to assuming $\mathbf{B}$ is of rank 1. In particular, using their notation, for binary case, their direction $\mathbf{B} = \boldsymbol{\beta}_1\boldsymbol{\xi}^{\mathrm{T}}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^q$. Since the true ranks of $\mathbf{B}$ in our simulation studies are all of rank greater than 1, it is not surprising that our method outperforms PMDA. Moreover, PMDA does not apply to the case where $n < p + q$, that is, the sample size is far smaller than the summation of image dimensions. Therefore, their method does not apply to one of our simulation settings $(n, p, q) = (100, 64, 64)$ and we mark their results using $*$ in Table 1. We also compare the computation time between PMDA and our method. In simulation, when $n = 200$ and true signal is a cross, given a fixed regularization parameter, the system running time of PMDA is around 1.5 min whereas the system running time of our method is no more than 13 s. Here, system running time is measured on a Macbook Pro laptop with a 2.9 GHz Intel Core i5. This is because PMDA essentially solves least-square problems with L_1 penalty in each iteration when setting $\omega_1 = 0$ in Algorithm 2 in Zhong and Suslick (2015). Our method is based on the Nesterov optimal gradient method which avoids computing inverse of covariance matrix and hence has a faster convergence rate.

4.2. Real Data Application

We apply our method to an EEG dataset, which is available at *https://archive.ics.uci.edu/ml/datasets/EEG+Database*. The data

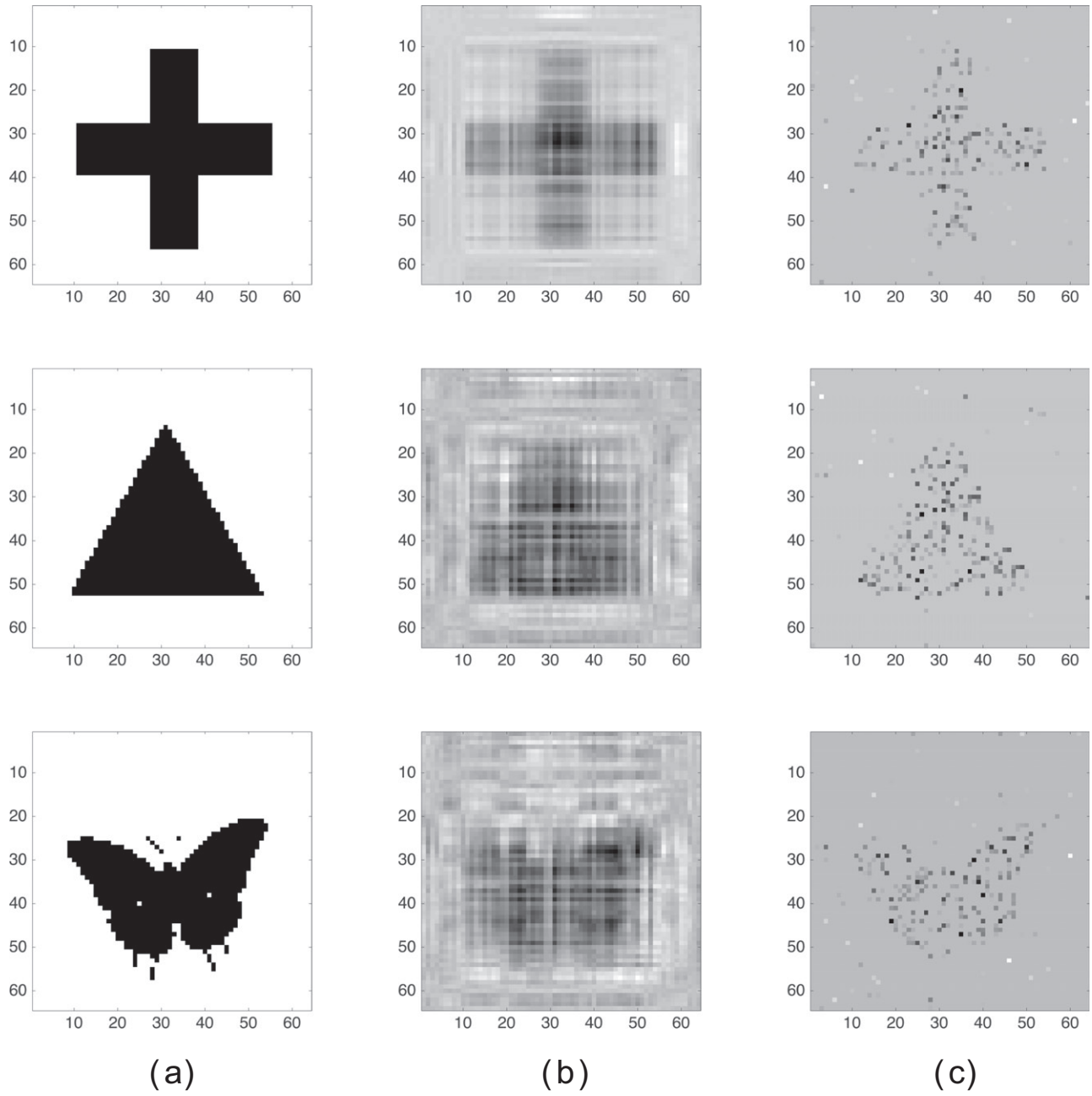

Figure 1. The figures for cross image: (a) original signal; (b) our nuclear regularization estimate; and (c) ℓ_1-regularized estimate.

were collected by the Neurodynamics Laboratory to study the EEG correlates of genetic predisposition to alcoholism. It contained measurements from 64 electrodes placed on each subject's scalps sampled at 256 Hz (3.9-msec epoch) for 1 second. Each subject was exposed to three stimuli: a single stimulus, two matched stimuli, two unmatched stimuli. Among the 122 subjects in the study, 77 were alcoholic individuals and 45 were controls. More details about the study can be found in Zhang et al. (1995). In statistics literature, EEG data have been analyzed using different models, for example, Gao et al. (2019a) considered an unsupervised approach for clustering EEG data, Gao et al. (2019b) and Gao et al. (2018) considered an evolutionary state-space model and graphical model for better understanding brain connectivity, respectively. However, these methods are not directly applicable for classification purpose here.

In our data analysis, for each subject, we use the average of all 120 runs for each subject under single-stimulus condition and use that as the covariate $\mathbf{x}_i$, which is a 256×64 matrix. The classification label is *alcoholic* or not. We randomly divide the dataset into training set of 81 subjects and test set of 41 subjects for 100 times, and each time fit the model on the training set and apply it on the test set to obtain the average mis-classification rate and its standard error. The results for different methods are summarized in Table 2. It can be seen that the proposed method has a significant lower mis-classification rate compared with other methods, which agrees with the simulation findings for the unbalanced data. PMDA does not work here since $p+q > n$ $((n, p, q) = (122, 256, 64))$. We also check the fitted signal matrix and it agrees well with the one obtained by Zhou and Li (2014).

Table 1. Simulation results: misclassification rates (%) and associated standard errors obtained from our method, Lasso LDA, Logistic Nuclear (L-Nuclear), Logistic Lasso (L-Lasso), and PMDA based on 1000 Monte Carlo replications.

Shape	n	(π_1, π_2)	Ours	Lasso LDA	L-Nuclear	L-Lasso	PMDA
Cross	100	(0.5,0.5)	3.65(0.02)	17.81(0.07)	3.70(0.02)	19.51(0.07)	*
	100	(0.75,0.25)	3.32(0.02)	14.89(0.05)	6.62(0.04)	18.84(0.04)	*
	200	(0.5,0.5)	3.22(0.02)	11.69(0.05)	3.26(0.02)	13.39(0.05)	26.93(0.05)
	200	(0.75,0.25)	2.87(0.02)	9.89(0.04)	4.14(0.03)	16.27(0.04)	19.58(0.08)
	500	(0.5,0.5)	3.09(0.02)	6.97(0.03)	3.11(0.02)	8.19(0.04)	25.17(0.04)
	500	(0.75,0.25)	2.62(0.02)	5.81(0.03)	3.59(0.02)	14.91(0.03)	12.05(0.04)
Triangle	100	(0.5,0.5)	3.12(0.02)	15.73(0.06)	3.11(0.02)	17.70(0.07)	*
	100	(0.75,0.25)	2.66(0.02)	13.72(0.05)	6.10(0.04)	17.19(0.04)	*
	200	(0.5,0.5)	2.85(0.02)	9.90(0.04)	2.81(0.02)	11.81(0.04)	30.17(0.08)
	200	(0.75,0.25)	2.43(0.02)	8.72(0.03)	3.62(0.02)	13.40(0.04)	24.63(0.10)
	500	(0.5,0.5)	2.67(0.02)	5.67(0.03)	2.73(0.02)	6.96(0.03)	25.92(0.04)
	500	(0.75,0.25)	2.29(0.01)	4.89(0.02)	2.74(0.02)	9.97(0.03)	14.69(0.05)
Butterfly	100	(0.5,0.5)	3.86(0.02)	17.10(0.06)	4.16(0.02)	18.82(0.07)	*
	100	(0.75,0.25)	3.47(0.02)	14.79(0.05)	7.14(0.04)	17.78(0.04)	*
	200	(0.5,0.5)	3.67(0.02)	11.00(0.04)	3.78(0.02)	12.66(0.05)	29.79(0.07)
	200	(0.75,0.25)	3.26(0.02)	9.80(0.04)	4.50(0.02)	13.93(0.04)	23.83(0.09)
	500	(0.5,0.5)	3.56(0.02)	6.50(0.03)	3.52(0.02)	7.70(0.03)	25.77(0.04)
	500	(0.75,0.25)	3.02(0.02)	5.74(0.03)	3.51(0.02)	10.49(0.03)	14.66(0.05)

Table 2. EEG data analysis: misclassification rates (%) and associated standard errors.

Our method	Lasso LDA	Logistic Nuclear	Logistic Lasso	PMDA
22.20(0.53)	24.12(0.70)	24.44(0.80)	26.24(0.91)	*

In terms of computational efficiency, we measured the computation time among Lasso LDA, Logistic Nuclear, Logistic Lasso, and our method based on one evaluation of the data, that is, partitioning the data into training and test sets, fitting the model on the training set and applying it on the test set. The running time for Lasso LDA, Logistic Nuclear, Logistic Lasso, and our method is 0.67, 1.79, 1.27, and 1.87 s, respectively. The system running time is measured in Matlab R2015b on a Macbook Pro laptop with a 2.9 GHz Intel Core i5.

5. Discussion

In the literature, total variation (TV) regularization has also been commonly used for modeling image data in addition to the proposed nuclear norm regularization. Their focuses are slightly different—the former is on structured sparse pattern and the later is on low-rank pattern. The main reason that we choose to focus on the nuclear norm regularization in this article is because we have found that low rankness is a more reasonable assumption than sparseness assumption in our real data application. In particular, the mis-classification errors of our method are lower than the sparse method (LASSO) in our real data analysis. The TV regularization is an interesting direction to explore as it requires new computational algorithms and theories; and thus we leave this for the future research.

In this article, we only consider the situation where all the image measurements are taking at the same scale, that is, the dimension of the image covariates p and q are equal for every study subject. We believe this is the case for most applications. For the special cases when image dimensions vary across subjects, our method may still be applicable by first resizing the image to the same scale. It will be of future interest to develop flexible statistical methods to handle image data that can be of different sizes in general.

Acknowledgments

The authors would like to thank the Editor, Associate Editor and two reviewers for their constructive comments, which have substantially improved the article.

Funding

Shen's research is partially supported by Simons Foundation Award 512620 and NSF DMS-1509023. Zhou's research is partially supported by NIH grants R01HG006139, R01GM53275 and NSF DMS-1310319. Kong's research is partially supported by the Natural Science and Engineering Research Council of Canada.

Appendix A. Primary Lemmas and Propositions

We start with some useful lemmas in this section. The proof of main theorems are given in Appendix B.

We first restate a singular value thresholding formula in Cai, Candès, and Shen (2010). This result is extremely useful when computing optimal solution of (A.2), by which the important block of Nestorov's algorithm was formed. The proof is based on showing that 0 is one of subgradients of (A.1) at $\hat{\mathbf{B}}$.

Proposition 1. For any $\omega \geq 0$ and a given matrix $\mathbf{B}_0 \in \mathcal{R}^{p\times q}$ with singular value decomposition $U\text{diag}(s)V^{\mathrm{T}}$, the minimizer $\hat{\mathbf{B}}$ of

$$\frac{1}{2}\|\mathbf{B}-\mathbf{B}_0\|_F^2+\omega\|\mathbf{B}\|_* \tag{A.1}$$

has the same singular vectors as B_0 with singular values $(s_i-\omega)_+$.

Next we state a lemma on the risk bound. This result can be viewed as an analog of Theorem 1 in Negahban et al. (2012) under our situation.

Lemma 1. Suppose that (A1) and (A2) hold, and $\omega_n \geq 2\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{X}_i\|_2$. Then any optimal solution $\hat{\mathbf{B}}$ to

$$(\hat{\beta}_0, \hat{\mathbf{B}}) = \arg\min_{\beta_0, \mathbf{B}} \frac{1}{2n}\sum_{i=1}^n \Big(y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B}\rangle\Big)^2 + \omega_n \|\mathbf{B}\|_* \quad \text{(A.2)}$$

satisfies the bound

$$\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 \leq 9\frac{\omega_n^2}{\lambda_l} r.$$

Proof. We apply Theorem 1 in Negahban et al. (2012) to our situation. Observe that the nuclear norm is decomposable, and the squared error loss satisfies $\tau_{\mathcal{L}}(\mathbf{B}_0) = 0$ in that article. Moreover, the dual norm $\mathcal{R}^*$ to the nuclear norm is simply the spectral norm. The curvature constant $\kappa_{\mathcal{L}}$ in the restricted strong convexity (RSC) condition can be chosen as $\lambda_l^{1/2}$ because the squared error loss is used and the Hessian matrix $\mathrm{E}\{\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^{\mathrm{T}}\} = \Sigma_{xx} \geq \lambda_l I$. For a subspace M that contains matrices of the rank at most r, its subspace compatibility constant satisfies

$$\psi(M) = \sup_{\mathbf{U}\in M\setminus\{0\}} \frac{\|\mathbf{U}\|_*}{\|\mathbf{U}\|_F} = \sup_{\mathbf{U}\in M\setminus\{0\}} \frac{\sum_{i=1}^r \sigma_i(\mathbf{U})}{(\sum_{i=1}^r \sigma_i(\mathbf{U})^2)^{1/2}} \leq \sqrt{r},$$

where the last inequality follows by Cauchy–Schwarz inequality. Hence, subspace compatibility constant under the low-rank assumption (A2) is bounded by $\sqrt{r}$. □

Next we state a few commonly used lemmas regarding the concentration property and tail probability inequalities of Gaussian and sub-Gaussian random variable (matrices). Their proofs can be found in standard textbooks, for example, Wainwright (2019).

Lemma 2. (Hoeffding bound) Suppose that the variables $\mathbf{X}_i$, $i = 1, 2, \ldots, n$ are independent and X_i has mean μ_i and sub-Gaussian parameter Σ_i. Then for all $t \geq 0$, we have

$$P\left(\sum_{i=1}^n (\mathbf{X}_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \Sigma_i^2}\right).$$

Lemma 3. Assume $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{p\times q}$ are iid random matrices. Suppose that $\|\mathbf{X}_1\|_2 \leq M$ almost surely, then with probability greater than $1 - \delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i - E\mathbf{X}_1\right\|_2 \leq \frac{6M}{\sqrt{n}}\left(\sqrt{\log\min(p,q)} + \sqrt{\log(1/\delta)}\right).$$

Lemma 4. Let $\mathbf{A}$ be an $p \times q$ matrix whose entries are independent standard normal random variables. Denote $s_{\min}(\mathbf{A})$ and $s_{\max}(\mathbf{A})$ as smallest singular value and largest singular value of $\mathbf{A}$, respectively. Assume $p \geq q$ without loss of generality. Then

$$\sqrt{p} - \sqrt{q} \leq E_{s_{\min}}(\mathbf{A}) \leq E_{s_{\max}}(\mathbf{A}) \leq \sqrt{p} + \sqrt{q}.$$

Lemma 5. Let $\mathbf{Y} \sim N(0, I_{d\times d})$ be a d-dimensional Gaussian random variable. Then for any function F: $\mathcal{R}^d \to \mathcal{R}$ with Lipschitz constant L, that is, $|F(\mathbf{x}) - F(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$, we have

$$P\{|F(\mathbf{Y}) - E(F(\mathbf{Y}))| \geq t\} \leq 2\exp\left(-\frac{t^2}{2L^2}\right),$$

for any $t > 0$.

Lemma 6. (Anderson's comparison inequality (Anderson 1955)) Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean Gaussian random vectors with covariance $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Y}}$, respectively. If $\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{Y}}$ is positive semi-definite then for any convex symmetric set C,

$$P(\mathbf{X} \in C) \leq P(\mathbf{Y} \in C).$$

The following lemma is very useful in establishing rank estimation consistency.

Lemma 7. Assume (A1) and (A2) hold. Let $\hat{\mathbf{B}}$ be a global minimizer of (A.2). If $n^{1/2}\omega_n$ tends to $+\infty$ and ω_n tends to zero, then $\omega_n^{-1}(\hat{\mathbf{B}} - \mathbf{B}_0)$ converges in probability to the unique global minimizer Δ of

$$\min_{\Delta\in\mathrm{R}^{p\times q}} \frac{1}{2}\mathrm{vec}(\Delta)^{\mathrm{T}}\Sigma\mathrm{vec}(\Delta) + \mathrm{tr}\{\mathbf{U}_0^{\mathrm{T}}\Delta\mathbf{V}_0\} + \|\mathbf{U}_{0\perp}^{\mathrm{T}}\Delta\mathbf{V}_{0\perp}\|_*.$$

Moreover, $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n\Delta + O_p\big(\omega_n \min(p,q)n^{-1/2} + \min(p,q)n^{-1/2} + \omega_n^2\min(p,q)^{1/2}n^{-1/2}\big)$.

Proof. We can write $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n\hat{\Delta}$, where $\hat{\Delta}$ is the global minimum of

$$V_n(\Delta) = \frac{1}{2}\mathrm{vec}(\Delta)^{\mathrm{T}}\hat{\Sigma}_{xx}\mathrm{vec}(\Delta) - \omega_n^{-1}\mathrm{tr}\Delta^{\mathrm{T}}\hat{\Sigma}_{\mathbf{X}\epsilon} + \omega_n^{-1} \times (\|\mathbf{B}_0 + \omega_n\Delta\|_* - \|\mathbf{B}_0\|_*),$$

where $\hat{\Sigma}_{xx} = n^{-1}\sum_{i=1}^n \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^{\mathrm{T}}$ and $\hat{\Sigma}_{\mathbf{X}\epsilon} = n^{-1}\sum_{i=1}^n \epsilon_i\mathrm{vec}(\mathbf{X}_i)$. Then $\mathrm{vec}(\Delta)^{\mathrm{T}}\hat{\Sigma}_{xx}\mathrm{vec}(\Delta)/2 - \mathrm{vec}(\Delta)^{\mathrm{T}}\Sigma_{xx}\mathrm{vec}(\Delta)/2$ converges to $\mathrm{vec}(\Delta)^{\mathrm{T}}\mathrm{E}(\hat{\Sigma}_{xx} - \Sigma_{xx})\mathrm{vec}(\Delta)/2$ with probability of 1. Note that $\mathrm{E}\|\hat{\Sigma}_{xx} - \Sigma\|_F^2 = O(n^{-1})$. Denote $\mathrm{vec}(\Delta)_i$ as a_i and $(\hat{\Sigma}_{xx} - \Sigma)_{ij}$ as b_{ij}. Then we have

$$\begin{aligned}
\frac{1}{2}|\mathrm{vec}(\Delta)^{\mathrm{T}}\mathrm{E}(\hat{\Sigma}_{xx} - \Sigma)\mathrm{vec}(\Delta)| &\leq \sum_{i,j=1}^{pq} |a_i a_j \mathrm{E}(b_{ji})| \\
&\leq \Big(\sum_{i,j=1}^{pq} a_i^2 a_j^2 \sum_{i,j=1}^{pq} \mathrm{E}(b_{ij}^2)\Big)^{\frac{1}{2}} \\
&= \sum_{i=1}^{pq} a_i^2 \mathrm{E}\Big(\sum_{i,j=1}^{pq} b_{ij}^2\Big)^{\frac{1}{2}} \\
&= \sum_{i=1}^{pq} a_i^2 \mathrm{E}\|\hat{\Sigma}_{xx} - \Sigma_{xx}\|_F \\
&= \|\Delta\|_F^2 O(n^{-1/2}) \\
&= O\big(\min(p,q)\|\Delta\|_2^2 n^{-1/2}\big).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
|\mathrm{tr}\Delta^{\mathrm{T}}\hat{\Sigma}_{\mathbf{X}\epsilon}| &\leq (\mathrm{tr}\Delta^{\mathrm{T}}\Delta)^{\frac{1}{2}}(\mathrm{tr}\hat{\Sigma}_{\mathbf{X}\epsilon}^{\mathrm{T}}\hat{\Sigma}_{\mathbf{X}\epsilon})^{\frac{1}{2}} \\
&= \|\Delta\|_F O_p(n^{-1/2}) \\
&\leq \min(p,q)^{\frac{1}{2}}\|\Delta\|_2 O_p(n^{-\frac{1}{2}}).
\end{aligned}$$

Therefore

$$\begin{aligned}
V_n(\Delta) &= \frac{1}{2}\mathrm{vec}(\Delta)^{\mathrm{T}}\Sigma\mathrm{vec}(\Delta) + O_p\big(\min(p,q)n^{-1/2}\|\Delta\|_2^2\big) \\
&\quad + O_p\big(\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2}\|\Delta\|_2\big) + \mathrm{tr}(\mathbf{U}_0^{\mathrm{T}}\Delta\mathbf{V}_0) \\
&\quad + \|\mathbf{U}_{0\perp}^{\mathrm{T}}\Delta\mathbf{V}_{0\perp}\|_* + O_p\big(\omega_n p^{1/2}q^{1/2}\min(p,q)\|\Delta\|_2^2\big). \\
&= V(\Delta) + O_p\big(\min(p,q)n^{-1/2}\|\Delta\|_2^2\big) \\
&\quad + O_p\big(\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2}\|\Delta\|_2\big) \\
&\quad + O_p\big(\omega_n p^{1/2}q^{1/2}\min(p,q)\|\Delta\|_2^2\big),
\end{aligned}$$

where $p^{1/2}q^{1/2}$ in the last term comes from the Frobenius norm of any matrix in $\mathcal{R}^{p\times q}$ with bounded entries. Let s_r be the rth largest singular value of B_0, for any $M < s_r/(2\omega_n)$,

$$\begin{aligned}
&\mathrm{E}\sup_{\|\Delta\|_2\le M}|V_n(\Delta)-V(\Delta)| \\
&\quad= O\big(\min(p,q)M^2\mathrm{E}\|\hat{\Sigma}_{xx}-\Sigma\|_F \\
&\qquad + M\min(p,q)^{\frac{1}{2}}\omega_n^{-1}E(\|\hat{\Sigma}_{M\epsilon}\|^2)^{1/2}+\omega_n p^{1/2}q^{1/2}\min(p,q)M^2\big) \\
&\quad= fO\big(\min(p,q)M^2n^{-1/2}+M\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2} \\
&\qquad + \omega_n p^{1/2}q^{1/2}\min(p,q)M^2\big).
\end{aligned}$$

Obviously $V(\Delta)$ achieves its minimum in the bounded ball at $\Delta_0 \neq 0$. Hence, by Markov inequality the probability of the minimum of $V_n(\Delta)$ lying strictly inside the ball $\|\Delta\|_2 < 2\|\Delta_0\|_2$ tends to one and is also the unconstrained minimum. □

The following two lemmas can be viewed as analogs of Proposition 3 and Lemma 11 in Bach (2008). W present them without the proof.

Lemma 8. Let $\mathbf{B}_0 = \mathbf{U}_0\mathrm{Diag}(S_0)\mathbf{V}_0^{\mathrm{T}}$ be the singular value decomposition of $\mathbf{B}_0$. Then the unique global minimizer of

$$\frac{1}{2}\mathrm{vec}(\Delta)^{\mathrm{T}}\Sigma\mathrm{vec}(\Delta)+\mathrm{tr}\mathbf{U}_0^{\mathrm{T}}\Delta\mathbf{V}_0+\|\mathbf{U}_{0\perp}^{\mathrm{T}}\Delta\mathbf{V}_{0\perp}\|_*$$

satisfies $\mathbf{U}_{0\perp}^{\mathrm{T}}\Delta\mathbf{V}_{0\perp}=0$ if and only if

$$\begin{aligned}
&\Big\|\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma^{-1}(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})\}^{-1} \\
&\quad\times\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma^{-1}(\mathbf{V}_0\otimes\mathbf{U}_0)\mathrm{vec}(\mathbf{I})\}\Big\|_2\le 1.
\end{aligned}$$

Furthermore, when $\mathbf{U}_{0\perp}^{\mathrm{T}}\Delta\mathbf{V}_{0\perp}=0$, the solutions has these forms:

$$\begin{aligned}
\mathrm{vec}(\Lambda) &= \{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})\}^{-1} \\
&\quad\times\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\Sigma(\mathbf{V}_0\otimes\mathbf{U}_0)\mathrm{vec}(I)\}, \\
\mathrm{vec}(\Delta) &= -\Sigma^{-1}\mathrm{vec}(\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}}-\mathbf{U}_{0\perp}\Lambda\mathbf{V}_{0\perp}^{\mathrm{T}}). \qquad \text{(A.3)}
\end{aligned}$$

Lemma 9. The matrix $\mathbf{B}$ with singular value decomposition $\mathbf{B} = \mathbf{U}\mathrm{Diag}(\mathbf{S})\mathbf{V}^{\mathrm{T}}$(with strictly positive singular value s) is optimal for the problem in (A.2) if and only if

$$\hat{\Sigma}_{xx}\mathbf{B}-\hat{\Sigma}_{\mathbf{X}y}+\omega_n\mathbf{U}\mathbf{V}^{\mathrm{T}}+N=0,$$

with $\mathbf{U}^{\mathrm{T}}\mathbf{N}=0$, $\mathbf{N}\mathbf{V}=0$ and $\|\mathbf{N}\|_2\le\omega_n$.

Appendix B. Proof of Theorems

Proof of Theorem 1. Throughout the proof, we use C to denote a universal positive constant where its value is not important for the theoretical purpose. In order to apply Lemma 1, we just need to evaluate the term $\|n^{-1}\sum_{i=1}^n\epsilon_i\mathbf{X}_i\|_2$ and then set the tuning parameter w_n to be greater than that quantity. Note that $\epsilon_i = Y_i-\langle\mathbf{X}_i,\mathbf{B}\rangle-\beta_0^*$. Let $\mathbf{X}_i = \pi_1\mathbf{X}_i^{(1)}+\pi_2\mathbf{X}_i^{(2)}$, where $\mathrm{vec}(\mathbf{X}_i^{(g)})\overset{i.i.d.}{\sim}N(\mu_g,\Sigma)$ and $\mu_g\in\mathbb{R}^{pq\times 1}$ for $g=1,2$. Define $\mathbf{X}_i\overset{\text{i.i.d}}{\sim}\mathbf{X}$ and $\epsilon_i\overset{\text{i.i.d}}{\sim}\epsilon$. Observe that

$$\begin{aligned}
\mathrm{vec}\{\mathrm{E}(\epsilon_i\mathbf{X}_i)\} &= \pi_1\mathrm{E}\left\{\left(-\frac{n}{n_1}-\beta_0^*-\langle\mathbf{X}^{(1)},\mathbf{B}_0\rangle\right)\mathrm{vec}(\mathbf{X}^{(1)})\right\} \\
&\quad+\pi_2\mathrm{E}\left\{\left(\frac{n}{n_2}-\beta_0^*-\langle\mathbf{X}^{(2)},\mathbf{B}_0\rangle\right)\mathrm{vec}(\mathbf{X}^{(2)})\right\} \\
&= (\mu_2-\mu_1)-(\pi_1\mu_1+\pi_2\mu_2)\beta_0^* \\
&\quad-\pi_1E\{\mathrm{vec}(\mathbf{X}^{(1)})\mathrm{vec}(\mathbf{X}^{(1)})^{\mathrm{T}}\}\mathrm{vec}(\mathbf{B}_0) \\
&\quad-\pi_2E\{\mathrm{vec}(\mathbf{X}^{(2)})\mathrm{vec}(\mathbf{X}^{(2)})^{\mathrm{T}}\}\mathrm{vec}(\mathbf{B}_0) \\
&= (\mu_2-\mu_1)-(\pi_1\mu_1+\pi_2\mu_2)\beta_0^* \\
&\quad-\pi_1\{\mu_1\mu_1^{\mathrm{T}}+\Sigma\}\mathrm{vec}(\mathbf{B}_0)-\pi_2\{\mu_2\mu_2^{\mathrm{T}}+\Sigma\}\mathrm{vec}(\mathbf{B}_0). \qquad \text{(B.1)}
\end{aligned}$$

Now, to further simplify this result, we reparameterize the mean of two normal populations such that $\mu_1=0$, and $\mu_2=\mathbf{D}$. Then recall by the equivalence between LDA and least-squares solution, we have

$$\begin{aligned}
&\mathrm{vec}(\mathbf{B}) = c\Sigma^{-1}\mathbf{D}, \\
&\beta_0 = -(\pi_1\mu_1+\pi_2\mu_2)^{\mathrm{T}}\mathrm{vec}(\mathbf{B}) = -\pi_2c\mathbf{D}^{\mathrm{T}}\Sigma^{-1}\mathbf{D}, \\
&\beta_0^* = \beta_0-d
\end{aligned}$$

for some positive constants c and d. Then (B.1) can be simplified into

$$\begin{aligned}
&\mathbf{D}-\pi_2\mathbf{D}\beta_0^*-\pi_2\{\mathbf{D}\mathbf{D}^{\mathrm{T}}\}\mathrm{vec}(\mathbf{B})-c\mathbf{D} \\
&=\mathbf{D}-\pi_2\mathbf{D}\beta_0+\pi_2\mathbf{D}\mathbf{D}-\pi_2\{\mathbf{D}\mathbf{D}^{\mathrm{T}}\}\mathrm{vec}(\mathbf{B})-c\mathbf{D} \\
&=\mathbf{D}\{1+\pi_2^2c\mathbf{D}^{\mathrm{T}}\Sigma^{-1}\mathbf{D}+\pi_2d-\pi_2c\mathbf{D}^{\mathrm{T}}\Sigma^{-1}\mathbf{D}-c\} \\
&=0,
\end{aligned}$$

given d is chosen as $\pi_2^{-1}\{c-1+(\pi_2-\pi_2^2)(\mathbf{D}^{\mathrm{T}}\Sigma^{-1}\mathbf{D})\}$.

Next we show that with high probability, $\|\epsilon\mathbf{X}\|_2\le 2\log n(C_\mu+\lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})$. Since ϵ follows a mixture of two normal distributions, ϵ is sub-gaussian with sub-Gaussian parameter denoted by σ, which is a positive constant due to the bounded eigenvalue condition in (A1). By Lemma 2, for sufficiently large n,

$$\begin{aligned}
P(|\epsilon|>2\log n) &\le P(|\epsilon-E(\epsilon)|>\log n) \\
&\le 2\exp\left(-\frac{\log^2 n}{2\sigma^2}\right)\le C\exp(-2\log n)=\frac{C}{n^2}.
\end{aligned}$$

Then we know $|\epsilon|\le 2\log n$ with probability of at least $1-Cn^{-2}$. For $\|\mathbf{X}\|_2$, we first consider its centralized version, that is, $\mathbf{X}\sim N(0,\Sigma)$. Note that we can write the spectral norm of a matrix in the form of a canonical Gaussian process,

$$\|N(0,\Sigma)\|_2=\sup_{\mathbf{A}:\|\mathbf{A}\|_*\le 1}\langle N(0,\Sigma),\mathbf{A}\rangle.$$

This allows us to apply Gaussian comparison inequality (Slepian 1962). Define $\mathbf{Z}\in\mathbb{R}^{p\times q}$ that satisfies $\mathrm{vec}(\mathbf{Z})\sim N(0,\mathbf{I})$. Then by Lemma 6, we have

$$\begin{aligned}
P(\|N(0,\Sigma)\|_2>t_1) &= P\Big(\sup_{A:\|A\|_*\le 1}\langle N(0,\Sigma),\mathbf{A}\rangle>t_1\Big) \\
&\le P\Big(\sup_{\mathbf{A}:\|\mathbf{A}\|_*\le 1}\langle\mathbf{Z},\mathbf{A}\rangle>t_1\lambda_u^{-1/2}\Big) \\
&= P(\|\mathbf{Z}\|_2>t_1\lambda_u^{-1/2}) \qquad \text{(B.2)}
\end{aligned}$$

for any $t_1>0$ because $\Sigma\le\lambda_u\mathbf{I}$ due to (A1). Apply Lemma 5 (or more generally the Tracy-Widow law), we have

$$P(\|\mathbf{Z}\|_2-E\|\mathbf{Z}\|_2>\sqrt{\log n})\le C\exp(-2\log n)=Cn^{-2}$$

for some constant $C>0$. Since $E\|\mathbf{Z}\|_2\le\sqrt{p}+\sqrt{q}$, by Lemma 4, with probability of at least $1-Cn^{-2}$, $\|\mathbf{Z}\|_2\le\sqrt{p}+\sqrt{q}+\sqrt{\log n}$, which leads to $\|N(0,\Sigma)\|_2\le\lambda_u^{1/2}(\sqrt{p}+\sqrt{q}+\sqrt{\log n})$ by (B.2). Therefore with probability of at least $1-Cn^{-2}$,

$$\begin{aligned}
\|\epsilon\mathbf{X}\|_2 &\le (2\log n)\|\mathbf{X}\|_2 \\
&\le 2\log n(\|\mu_1\|_2+\|N(0,\Sigma)\|_2)
\end{aligned}$$

$$\leq 2\log n\left\{C_\mu(\sqrt{p}+\sqrt{q})+\lambda_u^{1/2}(\sqrt{p}+\sqrt{q}+\sqrt{\log n})\right\}$$
$$\leq 2\log n(C_\mu+\lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})$$

using Condition (A4) and since we assume $\mu_2 = 0$ without loss of generality.

Now we apply the standard matrix concentration inequality, (e.g., Lemma 3) with $M = 2\log n(C_\mu+\lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})$ and $\delta = n^{-1}$. Note that $P(\|\mathbf{X}_i\epsilon_i\|_2 \leq M,\ i=1,\ldots,n) = (1-Cn^{-2})^n \geq 1 - Cn^{-1}$ by Bernoulli's inequality. Hence, we obtain that with probability of at least $1 - Cn^{-1}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\epsilon_i - \mathrm{E}(\epsilon\mathbf{X})\right\|_2$$
$$\leq \frac{6M}{\sqrt{n}}\left(\sqrt{\log\min(p,q)}+\sqrt{\log 1/\delta}\right)$$
$$\leq \frac{12(\log n)^{3/2}(C_\mu+\lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})}{\sqrt{n}}.$$

This completes the proof. □

Proof of Theorem 2. By Lemma 7, we obtain $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n\Delta + o_p(\omega_n)$. Since the rank is a lower semi-continuous function, the rank of $\hat{\mathbf{B}}$ is larger than r with probability tending to one by the consistency result, where r is the rank of $\mathbf{B}_0$. Suppose $\hat{\mathbf{B}}$ has singular value decomposition USV^{T} and U_c, V_c are singular vectors corresponding to U and V except the r largest singular values. By Lemma 9, $\hat{\Sigma}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0) - \hat{\Sigma}_{\mathbf{X}\epsilon}$ and $\hat{\mathbf{B}}$ have simultaneous singular value decomposition. Therefore it suffices to show $\|\mathbf{U}_c^{\mathrm{T}}\{\hat{\Sigma}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\Sigma}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\|_2 < \omega_n$ with probability tending to one. Note that

$$\mathbf{U}_c^{\mathrm{T}}\{\hat{\Sigma}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\Sigma}_{\mathbf{X}\epsilon}\}\mathbf{V}_c$$
$$= \mathbf{U}_c^{\mathrm{T}}\{\omega_n\hat{\Sigma}_{xx}\Delta + o_p(\omega_n) - O_p(n^{-1/2})\}\mathbf{V}_c$$
$$= \omega_n\mathbf{U}_c^{\mathrm{T}}(\Sigma\Delta)\mathbf{V}_c + o_p(\omega_n),$$

where $\Sigma\Delta$ is the matrix in $\mathrm{R}^{p\times q}$ satisfying $\mathrm{vec}(\Sigma\Delta) = \Sigma\mathrm{vec}(\Delta)$. Because of the regular consistency and a positive eigengap for $\mathbf{B}_0$, the projection onto the first singular vectors of $\hat{\mathbf{B}}$ converges those of $\mathbf{B}_0$. Hence, the projection on the orthogonal space is also consistent, which means $\mathbf{U}_c\mathbf{U}_c^{\mathrm{T}}$converges to $\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}$ and $\mathbf{V}_c\mathbf{V}_c^{\mathrm{T}}$converges to $\mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}$. Then by Lemma 8, we have

$$\|\mathbf{U}_c^{\mathrm{T}}\{\hat{\Sigma}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\Sigma}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\|_2$$
$$= \|\mathbf{U}_c\mathbf{U}_c^{\mathrm{T}}\{\hat{\Sigma}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\Sigma}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\mathbf{V}_c^{\mathrm{T}}\|_2$$
$$= \omega_n\|\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}(\Sigma\Delta)\mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2 + o_p(\omega_n)$$
$$= \omega_n\|\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}\Sigma\{-\Sigma^{-1}(\mathbf{U}_0\mathbf{V}_0{}^{\mathrm{T}} - \mathbf{U}_{0\perp}\Lambda\mathbf{V}_{0\perp}^{\mathrm{T}})\}$$
$$\times\ \mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2 + o_p(\omega_n)$$
$$= \omega_n\|\mathbf{U}_{0\perp}\Lambda\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2 + o_p(\omega_n)$$
$$= \omega_n\|\Lambda\|_2 + o_p(\omega_n),$$

where the third equality is due to (A.3). Since $\|\Lambda\|_2 < 1$, the last expression is less than ω_n with probability tending to one, which completes the proof. □

Proof of Theorem 3. Based on Corollary 3.1 of Zhang (2004), we have

$$R(\hat{f}_n) \leq R^* + 2c(\epsilon_1+\epsilon_2)^{1/s},$$

where Q is the squared error loss function defined by $Q(f) = E_{\mathbf{X}}\{y - f(\mathbf{X})^2\}$, $\epsilon_1 = \inf_f E_{\mathbf{X}}(2P(Y=1\mid\mathbf{X}) - 1 - f(\mathbf{X}))^2$, ϵ_2 satisfies $Q(\hat{f}_n) \leq \inf_f Q(f) + \epsilon_2$, and c and s can be chosen as $c = 0.5$ and $s = 2$ as explained by the Example 3.1 (for least-squares loss function) in that article. Now note that since $\hat{f}_n$ is determined by the classification coefficient $\hat{\mathbf{B}}$ and $\hat{\beta}_0$ that are both consistent based on Theorem 1. Therefore, ϵ_2 can be chosen arbitrarily close to 0. Also, as we assume the true class label Y given $\mathbf{X}$ is determined by the linear classification rule with β_0^* and $\mathbf{B}_0$, then $\inf_f E_{\mathbf{X}}\{2P(Y=1\mid\mathbf{X}) - 1 - f(\mathbf{X})\}^2 = 0$. Therefore, $\epsilon_1 = 0$. This concludes the proof. □

References

Anderson, T. W. (1955), "The Integral of a Symmetric Unimodal Function Over a Symmetric Convex Set and Some Probability Inequalities," *Proceedings of the American Mathematical Society*, 6, 170–176. [202]

Bach, F. R. (2008), "Consistency of Trace Norm Minimization," *Journal of Machine Learning Research*, 8, 1019–1048. [196,198,199,203]

Beck, A., and Teboulle, M. (2009), "A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202. [197]

Cai, J.-F., Candès, E. J., and Shen, Z. (2010), "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization*, 20, 1956–1982. [201]

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression Via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920. [196]

Duda, R. O., Hart, P. E., and Stork, D. G. (2012), *Pattern Classification*, New York: Wiley. [197]

Fan, J., Feng, Y., and Tong, X. (2012), "A Road to Classification in High Dimensional Space: The Regularized Optimal Affine Discriminant," *Journal of the Royal Statistical Society*, Series B, 74, 745–771. [196]

Gao, X., Shen, W., Hu, J., Fortin, N., Frostig, R., and Ombao, H. (2019a), "Regularized Matrix Data Clustering and Its Application to Image Analysis," arXiv:1808.01749. [200]

Gao, X., Shen, W., Shahbaba, B., Fortin, N., and Ombao, H. (2019b), "Evolutionary State-space Model and Its Application to Time-frequency Analysis of Local Field Potentials," *Statistica Sinica*. [200]

Gao, X., Shen, W., Ting, C.-M., Cramer, S. C., Srinivasan, R., and Ombao, H. (2018), "Modeling Brain Connectivity with Graphical Models on Frequency Domain," arXiv:1810.03279. [200]

Hu, W., Kong, D., and Shen, W. (2019), "Nonparametric Matrix Response Regression with Application to Brain Imaging Data Analysis," arXiv preprint arXiv:1904.00495. [196]

Kong, D., An, B., Zhang, J., and Zhu, H. (2019), "L2RM: Low-rank Linear Regression Models for High-dimensional Matrix Responses," *Journal of the American Statistical Association*, to appear. [196]

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-high Dimensions," *Biometrika*, 99, 29–42. [196,197]

Mika, S. (2002), "Kernel Fisher Discriminant," Ph.D. thesis, University of Technology, Berlin. [197]

Mu, Y., and Gage, F. (2011), "Adult Hippocampal Neurogenesis and Its Role in Alzheimers Disease," *Molecular Neurodegeneration*, 6, 1–85. [196]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [198,199,201,202]

Nesterov, Y. (1983), "A Method of Solving a Convex Programming Problem With Convergence Rate $O(1/k^2)$," in *Soviet Mathematics Doklady*, 27, 372–376. [197]

Raskutti, G. and Yuan, M. (2015), "Convex Regularization for High-Dimensional Tensor Regression," Tech. rep., arXiv:1512.01215. [198]

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011), "Sparse Linear Discriminant Analysis by Thresholding for High Dimensional Data," *The Annals of Statistics*, 39, 1241–1265. [196]

Slepian, D. (1962), "The One-sided Barrier Problem for Gaussian Noise," *Bell System Technical Journal*, 41, 463–501. [203]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [196]

Wainwright, M. J. (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press. [202]

Wang, X., and Zhu, H. (2017), "Generalized Scalar-on-image Regression Models Via Total Variation," *Journal of the American Statistical Association*, 112, 1156–1168. [196]

Zhang, T. (2004), "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *The Annals of Statistics*, 32, 56–85. [199,204]

Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995), "Event Related Potentials During Object Recognition Tasks," *Brain Research Bulletin*, 38, 531–538. [200]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research*, 7, 2541–2563. [198]

Zhong, W., and Suslick, K. S. (2015), "Matrix Discriminant Analysis With Application to Colorimetric Sensor Array Data," *Technometrics*, 57, 524–534. [196,197,199]

Zhou, H., and Li, L. (2014), "Regularized Matrix Regression," *Journal of the Royal Statistical Society*, Series B, 76, 463–483. [196,197,199,200]