Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Sparse subspace linear discriminant analysis

Yanfang Li & Jing Lei

To cite this article: Yanfang Li & Jing Lei (2018) Sparse subspace linear discriminant analysis, Statistics, 52:4, 782-800, DOI: 10.1080/02331888.2018.1469020

To link to this article: https://doi.org/10.1080/02331888.2018.1469020

Published online: 26 Jun 2018.

Submit your article to this journal

Article views: 90

View related articles

View Crossmark data

STATISTICS
2018, VOL. 52, NO. 4, 782–800
https://doi.org/10.1080/02331888.2018.1469020

Sparse subspace linear discriminant analysis

Yanfang Li[a] and Jing Lei[b]

[a]School of Mathematical Sciences, Peking University, Beijing, People's Republic of China; [b]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT
We study high dimensional multigroup classification from a sparse subspace estimation perspective, unifying the linear discriminant analysis (LDA) with other recent developments in high dimensional multivariate analysis using similar tools, such as penalization method. We develop two two-stage sparse LDA models, where in the first stage, convex relaxation is used to convert two classical formulations of LDA to semidefinite programs, and furthermore subspace perspective allows for straightforward regularization and estimation. After the initial convex relaxation, we use a refinement stage to improve the accuracy. For the first model, a penalized quadratic program with group lasso penalty is used for refinement, whereas a sparse version of the power method is used for the second model. We carefully examine the theoretical properties of both methods, alongside with simulations and real data analysis.

ARTICLE HISTORY
Received 2 November 2016
Accepted 7 March 2018

KEYWORDS
Convex relaxation; high dimensional classification; multigroup LDA; sparsity; subspace estimation

1. Introduction

Linear discriminant analysis (LDA) has long been a focus in statistics and machine learning [1–5]. LDA aims to find the best combination of features for classification under a measure of separation. More specifically, for a matrix $X \in \mathbb{R}^{n\times p}$ with p features and n observations, each of which belongs to one of G groups, LDA searches for a d-dimensional direction $V \in \mathbb{R}^{p\times d}$ for some $d < p$ such that XV preserves as much of the discrimination information as possible.

The original measure of separation for LDA, known as *Fisher criterion* [1], is defined in binary classification for the purpose of maximizing the between-class variability with respect to the within-class variability. The most well-known extension of Fisher criterion to multigroup situation [6,7] is

$$V_{\mathrm{f}} = \arg\max \left\{ \mathrm{Tr}\left(\left(V^{\mathrm{T}}\Sigma_w V \right)^{-1} V^{\mathrm{T}}\Sigma_b V \right) : V \in \mathbb{R}^{p\times d} \right\}, \tag{1}$$

where 'Tr(·)' denotes the trace of a matrix, ie, the summation of matrix diagonal entries, Σ_w and Σ_b are the population within-class and between-class covariance matrices (defined in Section 2), and d is the number of directions to consider or the target subspace dimension. The optimization problem (1) can be directly solved by the generalized eigenvalue decomposition of $\Sigma_w^{-1}\Sigma_b$ if Σ_w is invertible. On the other hand, there exists a less well-known but more straightforward formulation for multigroup discriminant analysis [6,8],

$$V_{\mathrm{t}} = \arg\max \left\{ \frac{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right)}{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_w V\right)} : V^{\mathrm{T}}V = I_d \right\}, \tag{2}$$

CONTACT Yanfang Li liyanfang1110@gmail.com

which is called 'trace-ratio' LDA. The numerator and denominator in the objective function of (2) directly reflect sum of squared inter- and intra-class Euclidean distance. But the lack of closed-form solution for $d > 1$ makes it less popular in practice. The first algorithm to obtain the global optimal solution for (2) is proposed in [9], who pointed out that the optimization problem (2) can be converted to an equivalent trace difference problem if the optimal value is given (see Lemma 2.3). In this paper, we will focus on an equivalent problem to (2) with the same form (see Lemma A.7 in the appendix),

$$V_t = \arg\max \left\{ \frac{\mathrm{Tr}\left(V^{\mathrm{T}} \Sigma_b V\right)}{\mathrm{Tr}\left(V^{\mathrm{T}} \Sigma_t V\right)} : V^{\mathrm{T}} V = I_d \right\}, \tag{3}$$

where $\Sigma_t = \Sigma_w + \Sigma_b$ called total covariance matrix. The property of the optimization problem (3) is essentially the same as that of (2), except that the optimal value of (3) lies in $[0, 1]$, which makes it possible to directly use the generalized Envelope Theorem [10] in our theoretical development.

We focus on sparse multigroup LDA, which seeks a few features to span the discriminant directions when dealing with high dimensional data ($p \gg n$) based on (1) and (3). Much effort has been made in the literatures to extend the optimization problem (1) to the high dimensional regime. There are two issues of (1) when p is much larger than n. First, the maximum likelihood estimation for the within-class covariance matrix denoted by $\hat{\Sigma}_w$ is singular. Second, the estimated discriminant directions involve a combination of all features, leading to poor interpretability. Bickel and Levina [11] used the idea of Naive Bayes that assumes all features are independent to estimate a diagonal matrix instead of the singular $\hat{\Sigma}_w$. Fan and Fan [12] extended the independent rule to extract few features for better interpretation. Since $\Sigma_w^{-1}(\mu^{(1)} - \mu^{(2)})$ is the optimal solution for binary classification, where $\mu^{(1)}$ and $\mu^{(2)}$ are the population means of two groups, Shao et al. [13] studied sparse LDA by assuming both Σ_w and $\mu^{(1)} - \mu^{(2)}$ are sparse. More directly, Cai and Liu [14] proposed a Dantzig selector type LDA by assuming that $\Sigma_w^{-1}(\mu^{(1)} - \mu^{(2)})$ is sparse, and Mai et al. [15] developed a lasso type LDA via least squares. For multigroup situations, the common approaches to estimate the discriminant directions are in a sequential fashion: finding each discriminant direction under the constraint that it is orthogonal to all the previous ones. Witten and Tibshirani [16] used ℓ_1 penalized fisher criterion to get sparse discriminant directions. Clemmensen et al. [17] proposed a sparse multigroup LDA from an equivalent way to Fisher criterion called optimal scoring [3,18]. However, sequential estimation suffers from severe error propagation, which is less ideal both computationally and theoretically. To avoid this problem, Gaynanova et al. [19] developed a direct approach for multigroup LDA, which estimates $G{-}1$ discriminant directions simultaneously together with feature selection using group lasso penalty.

In this paper, we extend a recent framework for high dimensional multivariate analysis using sparse subspace estimation to cover the high dimensional LDA problem. We propose two approaches for sparse multigroup LDA based on the formulations (1) and (3), respectively. The estimation procedures for both methods consist of a 'convex relaxation' stage and a 'refinement' stage. The convex relaxation stage aims to convert the optimization problems (1) and (3) to semidefinite programmings (SDPs) with suitable sparsity regularization. The idea behind the convex relaxation is related to the recent development in sparse principal component analysis [20] and sparse canonical correlation analysis [21]. As commonly known in the literature of high dimensional multivariate analysis, refinement stages are further applied to improve the estimation accuracy to reduce the bias caused by sparsity regularization. The refinement stage for the first model uses a quadratic optimization with group lasso penalty. For the second model, a sparse version of the power method is used in the refinement stage. To the best of our knowledge, this is the first attempt to extend the optimization problem (3) to the high dimensional regime.

The rest of the paper is organized as follows. Two sparse multigroup LDA methods together with their theoretical properties are given in Section 2. The algorithms and implementation details are given in Section 3. Section 4 investigates the numerical performance of the proposed methods on both synthetic and real datasets, with comparison to a state-of-the-art method. All technical proofs are given in the appendix.

Notations

Throughout this paper, we assume that the data $(X_i, Y_i)_{i=1}^n$ are generated independently from a common distribution on $\mathbb{R}^p \times \{1, \dots, G\}$. Let $\Sigma_w = \mathrm{Cov}(X \mid Y = y)$ be the common within group covariance, and $\Sigma_b = \sum_{y=1}^G \pi_y(\mu_y - \mu)(\mu_y - \mu)^{\mathrm{T}} = \mathrm{Cov}(\mathbb{E}(X \mid Y))$ be the between group covariance, where $\mu_y = \mathbb{E}(X \mid Y = y)$, $\pi_y = \mathbb{P}(Y = y)$ are the group centre and group proportion, respectively, and $\mu = \mathbb{E}(X) = \sum_{y=1}^G \pi_y \mu_y$ is the marginal mean of X. We use $\hat{\mu}$, $\hat{\mu}_y$, $\hat{\pi}_y$, and $\hat{\Sigma}_b$ to denote the natural sample average and plug-in estimates, and $\hat{\Sigma}_w$ for the standard pooled covariance estimate.

For a general vector $v \in \mathbb{R}^p$, $\|v\|_a$ denotes the ℓ_a norm of v, where a could be any non-negative integer. Especially when a equals to 0, $\|v\|_0$ means the number of non-zero elements of v, differently from the definitions $\|v\|_a = (\sum v_i^a)^{1/a}$ when a is strictly positive.

For a general matrix A, the ith column and the ith row are denoted by A_i and A_{i*} respectively. And four types of matrix norms are defined as follows. $\|A\|_{op}$ is the largest singular value of A. $\|A\|_F = (\mathrm{Tr}(A^{\mathrm{T}}A))^{1/2}$ is the Frobenius norm. $\|A\|_{a,b}$ is the (a, b) norm defined to be the ℓ_b norm of the vector of row-wise ℓ_a norm of A. For any index set $S \subseteq \{1, \dots, p\}^2$, A_S chooses the submatrix corresponding to S by setting all elements in S^c to zero, and $|S|$ counts the number of elements in S, which is called the cardinality of S. Together with a matrix B of compatible dimension, $\langle A, B\rangle = \mathrm{Tr}(A^{\mathrm{T}}B)$ is the inner product of A and B. Furthermore, any calculation of matrices is conventionally defined, such as the inequality $B < A$ denoting that $A - B$ is a positive matrix, $\mathbf{1}$ is an indicator function, which is 0 when the condition is true and $+\infty$ otherwise. Positive constants such as C and c may differ from one situation to another.

2. Method and theory

In this section, we describe our subspace estimation approaches to high dimensional LDA using Fisher criterion and the trace ratio criterion. The formulation (1) will be called 'FLDA' for short because it is the most well-known extension of Fisher criterion. The formulation (3) will be written briefly as 'TRLDA', standing for 'trace-ratio' LDA.

2.1. Sparse FLDA

Convex relaxation stage

To facilitate discussion, we rewrite model (1) to an equivalent form

$$V_{\mathrm{f}} = \arg\max \left\{ \mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right) : V^{\mathrm{T}}\Sigma_w V = I_d \right\}, \tag{4}$$

where for simplicity we take the target number of discriminant directions $d = \mathrm{rank}(\Sigma_b)$ throughout this paper. Extension to other choices of $d < \mathrm{rank}(\Sigma_b)$ is straightforward. By the Lagrange multiplier method, V_{f} consists of eigenvectors corresponding to the non-zero generalized eigenvalues of $\Sigma_w^{-1}\Sigma_b$ if Σ_w is invertible. The optimization problem (4) is non-convex, which makes it difficult to obtain the global solution when sparsity regularization is used to deal with high dimensional data. To overcome this difficulty, convex relaxation is a common and natural strategy. To this end, we further rewrite the objective function of (4) as

$$\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right) = \left\langle \Sigma_b, VV^{\mathrm{T}} \right\rangle,$$

which is linear with respect to $\Pi = VV^{\mathrm{T}}$. Combining this linearity and the fact that the *Fantope* ([22])

$$\mathcal{F}^d = \left\{ U \in \mathbb{R}^{p\times p} : 0 \preceq U \preceq I \quad \text{and} \quad \mathrm{Tr}(U) = d \right\}$$

is the convex hull of all rank-d projection matrices $\{\Pi = LL^{\mathrm{T}} : L \in \mathbb{R}^{p\times d},\ L^{\mathrm{T}}L = I_d\}$, the optimization problem (4), by treating $\Pi = VV^{\mathrm{T}}$ as a single term, can be written as

$$\Pi_{\mathrm{f}} = V_{\mathrm{f}}V_{\mathrm{f}}^{\mathrm{T}} = \arg\max\left\{\langle \Sigma_b, H\rangle : \Sigma_w^{1/2}H\Sigma_w^{1/2} \in \mathcal{F}^d\right\}. \tag{5}$$

In most high dimensional situations, only few features contribute to the discrimination information, which means only few rows of V_{f} are non-zero. The set of non-zero rows of V_{f} and its cardinality are denoted by $S = \{j : \|(V_{\mathrm{f}})_{j,*}\|_2 \neq 0\}$ and $q = |S| \ll p$ respectively. As a consequence, Π_{f} is supported on $J = S \times S$. Though the ℓ_0 norm, which counts the number of non-zero entries, is a natural way to express the sparsity, a high dimensional optimization problem with ℓ_0 penalty or constraint is usually computationally demanding. Instead, the ℓ_1 norm is used to replace the ℓ_0 norm, which leads to more efficiently solvable optimization problems. Then by replacing Σ_w and Σ_b with their maximum likelihood estimations, the ℓ_1 penalized convex relaxation for FLDA can be formulated as

$$\hat{H}_{\mathrm{f}} = \arg\max\left\{\langle \hat{\Sigma}_b, H\rangle - \lambda\,\|H\|_{1,1} : \hat{\Sigma}_w^{1/2}H\hat{\Sigma}_w^{1/2} \in \mathcal{F}^d\right\}, \tag{6}$$

where $\lambda > 0$ is a penalty parameter whose practical choice will be discussed later. Problem (6) can be viewed as the generalized eigenvalue extension of the Fantope-based sparse PCA formulation [20,23]. Gao et al. [21] used a similar formulation for sparse canonical component analysis.

Next we will explore the statistical properties of model (6) under some regularity conditions. In the sparse FLDA problem, the matrix Σ_b is low rank, and hence our setup naturally fits in the framework given in [21]. Though $\Sigma_w \in \mathbb{R}^{p\times p}$ is a high dimensional matrix, only a submatrix needs to be used for classification if the important feature set S is known. Also, similar to the effect of the restricted eigenvalue condition on lasso, here we introduce some sparse eigenvalue conditions on Σ_w.

Definition 2.1: For any integer $1 \leq m \leq p$, the minimum and maximum m-sparse eigenvalue of any matrix $A \in \mathbb{R}^{p\times p}$ are defined,

$$\phi_{\min}^{A}(m) = \min_{\|v\|_0 \leq m} \frac{v^{\mathrm{T}}Av}{\|v\|_2^2} \quad \phi_{\max}^{A}(m) = \max_{\|v\|_0 \leq m} \frac{v^{\mathrm{T}}Av}{\|v\|_2^2}.$$

Then the within-class covariance matrix Σ_w is assumed to be satisfied

$$\phi_{\min}^{\Sigma_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\Sigma_w}(m) > 1/C_w \quad \text{and} \quad \phi_{\max}^{\Sigma_w}(m) < C_w < \infty, \tag{7}$$

for some constant $C_w > 0$ and positive integer m such that $C_m q^2 \leq m \leq q^2 < p$. In addition to the sparse eigenvalue condition, we also have the following two standard conditions on the signal-to-noise ratio and the penalty parameter,

(C1) $n \geq q^2 \log p/C_n\lambda_d$ for some sufficiently small constant $C_n > 0$,
(C2) $\lambda \geq C_\lambda\sqrt{\log p/n}$ for some constant $C_\lambda > 0$,

where λ_d in condition (C1) is the dth largest generalized eigenvalue of $\Sigma_w^{-1}\Sigma_b$. In order to control the entry-wise deviance of the sample covariances from the population covariances, we also need to assume the tail behaviour of X. The standard assumption is sub-Gaussianity (see [24] for a definition in the context of sparse PCA, which is suitable for our purpose here).

Theorem 2.2: *Assume* $(X \mid Y = y)$ *is sub-Gaussian with constant scaling for all* $1 \leq y \leq G$. *Then under conditions* (7), (C1) *and* (C2), *for any* $C' > 0$, *with probability at least* $1 - \exp(-C'(q + \log(ep/q)))$, *we have*

$$\|\hat{H}_{\mathrm{f}} - \Pi_{\mathrm{f}}\|_F \leq C\frac{q\lambda}{\lambda_d},$$

for sufficiently large constant C > 0 corresponding to C'.

Remark 2.1: The proof is adapted from Gao et al. [21] which considered the sparse CCA case. The main difficulty to prove Theorem 2.2 is that the true population parameter $\Pi_{\mathrm{f}} = V_{\mathrm{f}} V_{\mathrm{f}}^{\mathrm{T}}$ is not feasible to the constraint of the optimization problem (6). To solve this thorny problem, we first prove the estimator $\hat{H}_{\mathrm{f}}$ to be consistent to a constructed feasible matrix $\tilde{\Pi} = \tilde{V}_{\mathrm{f}} \tilde{V}_{\mathrm{f}}^{\mathrm{T}}$, where $\tilde{V}_{\mathrm{f}} = V_{\mathrm{f}} (V_{\mathrm{f}}^{\mathrm{T}} \hat{\Sigma}_w V_{\mathrm{f}})^{-1/2}$. Then, from the short distance between $\tilde{\Pi}$ and Π_{f}, we obtain the final ℓ_2 consistency of the estimator $\hat{H}_{\mathrm{f}}$.

Remark 2.2: In this model, we assume that the target dimension d satisfies $d = \mathrm{rank}(\Sigma_b)$. In the most common case, when the group centres $(\mu_y : 1 \le y \le G)$ are in general position, we have $\mathrm{rank}(\Sigma_b) = G - 1$. Otherwise, the rank of Σ_b can be reliably obtained using the scree plot of singular values of $\hat{\Sigma}_b$, which is an accurate estimate of $\mathrm{rank}(\Sigma_b)$ due to its low rank.

Refinement stage

While the ‘projection matrix’ $\Pi_{\mathrm{f}} = V_{\mathrm{f}} V_{\mathrm{f}}^{\mathrm{T}}$ being focused on and estimated in the last convex relaxation stage makes the sparse high dimensional LDA estimate direct and easy, it simultaneously results in loss of accuracy. Now we will place great emphasis on correcting the estimation of V_{f} to improve accuracy. From Lemma A.1 in the appendix, the relationship between Σ_b and Σ_w can be formulated as

$$\Sigma_b = \Sigma_w V_{\mathrm{f}} \Lambda V_{\mathrm{f}}^{\mathrm{T}} \Sigma_w, \tag{8}$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is a diagonal matrix whose diagonal entries are the generalized eigenvalues of $\Sigma_w^{-1} \Sigma_b$, namely $\lambda_1 \ge \cdots \ge \lambda_d > 0$. Let $V_{\mathrm{f}} = U_{\mathrm{f}} D O$ be the (full rank) singular value decomposition of V_{f}. From the fact that $V_{\mathrm{f}}^{\mathrm{T}} \Sigma_w V_{\mathrm{f}} = I_d$, we have $O^{\mathrm{T}} D U_{\mathrm{f}}^{\mathrm{T}} \Sigma_w U_{\mathrm{f}} D O = I$ and hence $U_{\mathrm{f}}^{\mathrm{T}} \Sigma_w U_{\mathrm{f}} = D^{-2}$. Then by right multiplying U_{f} on both sides of (8) and replacing the second V_{f} on the right side of (8) with its singular value decomposition $U_{\mathrm{f}} D O$, we obtain $\Sigma_b U_{\mathrm{f}} = \Sigma_w V_{\mathrm{f}} \Lambda O^{\mathrm{T}} D^{-1}$, from which $V_{\mathrm{f}} \Lambda O^{\mathrm{T}} D^{-1}$ can be formulated as a global solution of the following quadratic optimization problem:

$$V_{\mathrm{f}} \Lambda O^{\mathrm{T}} D^{-1} = \arg\min_{Z \in \mathbb{R}^{p \times d}} \left\{ \frac{1}{2} \mathrm{Tr}(Z^{\mathrm{T}} \Sigma_w Z) - \mathrm{Tr}(U_{\mathrm{f}}^{\mathrm{T}} \Sigma_b Z) \right\}.$$

Let $\hat{U}_{\mathrm{f}}$ be the eigenvectors corresponding to the top d eigenvalues of $\hat{H}_{\mathrm{f}}$. Note that $\hat{U}_{\mathrm{f}}$ is an estimator of U_{f} in the optimization problem above. Then by replacing Σ_w and Σ_b with their estimates and introducing the group lasso penalty, we obtain our refined estimate for sparse FLDA,

$$\hat{V}_1 = \arg\min_{Z \in \mathbb{R}^{p \times d}} \left\{ \frac{1}{2} \mathrm{Tr}(Z^{\mathrm{T}} \hat{\Sigma}_w Z) - \mathrm{Tr}(\hat{U}_{\mathrm{f}}^{\mathrm{T}} \hat{\Sigma}_b Z) + \lambda \|Z\|_{2,1} \right\}, \tag{9}$$

where $\lambda > 0$ is also a penalty parameter differently from the penalty parameter λ used in the convex relaxation model (6). Finally, we rescale $\hat{V}_1$ to satisfy the original constraint condition $V^{\mathrm{T}} \hat{\Sigma}_w V = I_d$,

$$\hat{V}_{\mathrm{f}} = \hat{V}_1 (\hat{V}_1^{\mathrm{T}} \hat{\Sigma}_w \hat{V}_1)^{-1/2}.$$

We note that the $\hat{V}_{\mathrm{f}}$ obtained this way may not approximate V_{f} very well as $\hat{V}_{\mathrm{f}}$ is off by an orthonormal transform. However, if we only need $\hat{V}_{\mathrm{f}}$ for reducing the dimensionality and selecting the relevant variables of X, which will be used for subsequent classification, such a $\hat{V}_{\mathrm{f}}$ is good enough. With discriminant directions $\hat{V}_{\mathrm{f}}$ being estimated, we project the data matrix X onto $\hat{V}_{\mathrm{f}}$ and build a low dimensional classifier based on $X' = X\hat{V}_{\mathrm{f}}$.

2.2. Sparse TRLDA

Convex relaxation stage

Now we consider another variant of LDA based on the 'trace-ratio' formulation. For convenience of computation and proof, we will focus on the equivalent form (3). In this formulation, the target dimension d is no longer restricted to be $\operatorname{rank}(\Sigma_b)$. Compared to (2), the main advantage of (3) is that its optimal value $\eta_t^* \in [0,1]$, which makes the implementation and analysis more straightforward. This form of LDA has received less attention in the literature mainly because that it has no closed-form solution for $d > 1$. Until recently, Guo et al. [9] pointed out that though the solution has no closed form, it can be characterized by the following lemma.

Lemma 2.3 ([9]): *Suppose $\Sigma_t \in \mathbb{R}^{p\times p}$ is positive definite and $\Sigma_b \in \mathbb{R}^{p\times p}$ is positive semidefinite. Then*

$$\eta_t^* = \frac{\operatorname{Tr}\left({V_t}^{\mathrm{T}}\Sigma_b V_t\right)}{\operatorname{Tr}\left({V_t}^{\mathrm{T}}\Sigma_t V_t\right)} = \sup_{V^{\mathrm{T}}V=I_d} \frac{\operatorname{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right)}{\operatorname{Tr}\left(V^{\mathrm{T}}\Sigma_t V\right)}, \tag{10}$$

if and only if

$$\operatorname{Tr}\left({V_t}^{\mathrm{T}}(\Sigma_b - \eta_t^*\Sigma_t)V_t\right) = \sup_{V^{\mathrm{T}}V=I_d} \operatorname{Tr}\left(V^{\mathrm{T}}(\Sigma_b - \eta_t^*\Sigma_t)V\right) = \sup_{V^{\mathrm{T}}V=I_d} \left\langle \Sigma_b - \eta_t^*\Sigma_t, VV^{\mathrm{T}}\right\rangle = 0. \tag{11}$$

Lemma 2.3 shows that the trace ratio LDA can be solved via a trace difference problem if η_t^* is given. Similar to FLDA, it is natural to use convex relaxation together with regularization for high dimensional problems. Using the Fantope formulation of the eigenvalue problem and together with the fact that the objective function of the trace difference problem (11) is linear with respect to VV^{T}, $\Pi_t = V_t V_t^{\mathrm{T}}$ can be formulated as the global optimal solution of the following optimization problem:

$$\Pi_t = V_t {V_t}^{\mathrm{T}} = \arg\max\left\{\left\langle \Sigma_b - \eta_t^*\Sigma_t, H\right\rangle,\ H \in \mathcal{F}^d\right\}, \tag{12}$$

which combined with the *Ky Fan's maximum principle* [25], leads to the cumulative variance explained by the top d eigenvalues of $\Sigma_b - \eta_t^*\Sigma_t$ is zero, ie,

$$\sum_{i=1}^{d} \lambda\left(\Sigma_b - \eta^*\Sigma_t\right) = \left\langle \Sigma_b - \eta_t^*\Sigma_t, \Pi_t\right\rangle = 0. \tag{13}$$

To motivate our sparse estimator, some assumptions are needed to make. (*Identifiability*) First to guarantee the identifiability of Π_t, the eigenvalue gap of $\Sigma_b - \eta^*\Sigma_t$ needs to be strictly positive. (*Sparsity*) Moreover only a few features determine the discriminant directions, or equivalently, the support matrix of Π_t can be denoted by $J = S \times S$, where $S = \{j : \|(V_t)_{j*}\|_2 \neq 0\}$ is the set of decision features. Both assumptions discussed above can be expressed in the following mathematical form.

(S1) $\delta = \lambda_d(\Sigma_b - \eta_t^*\Sigma_t) - \lambda_{d+1}(\Sigma_b - \eta_t^*\Sigma_t) > 0$.
(S2) $\Sigma_b - \eta_t^*\Sigma_t$ has sparse top d principal subspace: $q = |S| \ll q$.

Now replacing Σ_t and Σ_b in the population optimization problem (12) with their sample estimators and adding an ℓ_1 sparsity constraint, we obtain the convex relaxation for TRLDA,

$$\hat{H}_\eta = \arg\max\left\{\left\langle \hat{\Sigma}_b - \eta\hat{\Sigma}_t,\ H\right\rangle : H \in \mathcal{F}^d \text{ and } \|H\|_{1,1} \leq R\right\}. \tag{14}$$

Here we consider a constrained version rather than a penalized version for presentation simplicity. In practice, one can also choose to use the equivalent penalized version. Our estimation of $V_t V_t^{\mathrm{T}}$ consists of two steps: (1) estimate a suitable $\hat{\eta}$, (2) given $\hat{\eta}$, solve the problem (14) with $\eta = \hat{\eta}$.

Suppose $\hat{\eta}$ is obtained, the problem (14) can be solved in several ways. First, by convexity we can use the equivalent penalized version

$$\hat{H}_\eta = \arg\max \left\{ \left\langle \hat{\Sigma}_b - \eta \hat{\Sigma}_t, H \right\rangle - \lambda \|H\|_{1,1} : H \in \mathcal{F}^d \right\}, \tag{15}$$

which is exactly the FPS problem in [20]. The solution path of (15) can be obtained by a standard ADMM algorithm, and one just needs to find the point on the solution path so that the value of the objective function in (14) is zero. Alternatively, one can directly solve (14) using a modified ADMM algorithm, as detailed in Section 3.2.

Now we discuss the choice of η. Theoretically speaking, one can approximate η^* by finding a root of the function

$$\hat{f}(\eta) = \sup_{\substack{H \in \mathcal{F}^d \\ \|H\|_{1,1} \le R}} \langle \hat{\Sigma}_b - \eta \hat{\Sigma}_t, H \rangle. \tag{16}$$

In Section 3.3, we prove that f is monotone with a unique root and provides an efficient Newton–Raphson algorithm for searching this root. In the following, let $\hat{\eta}$ be the estimator. The next theorem ensures the accuracy of $\hat{\eta}$ under the following regularization conditions.

(S3) Σ_t is a positive definite matrix with $\lambda_{\min}(\Sigma_t) > C_t$ for some constant $C_t > 0$.
(S4) $\|\Pi_t\|_{1,1} \le R$ and $R\sqrt{\log p/n} \le C_r$ for some sufficiently small constant $C_r > 0$.

Theorem 2.4: *If data points are (sub)Gaussian distributed sharing the same covariance matrix, there exits an $\hat{\eta}$ such that $\hat{f}(\hat{\eta}) = 0$. Moreover, under conditions (S3) and (S4), with probability greater that $1 - p^{-C}$ for $C > 0$, we have with*

$$\left|\eta^* - \hat{\eta}\right| \le C' \sqrt{\frac{\log p}{n}} R, \tag{17}$$

for some constant $C' > 0$ corresponding to C. As a consequence, letting $W_{\hat{\eta}} = (\Sigma_b - \eta^* \Sigma_t) - (\hat{\Sigma}_b - \hat{\eta} \hat{\Sigma}_t)$, we have

$$\|W_{\hat{\eta}}\|_{\infty,\infty} \le C_0 \sqrt{\frac{\log p}{n}} R, \tag{18}$$

with the same probability for some constant $C_0 > 0$ corresponding to C'.

With $\hat{\eta}$ being estimated and the good property of $\hat{\eta}$ given in Theorem 2.4, a consistent projection matrix $\hat{H}_{\hat{\eta}}$ and a reasonable discriminant directions $\hat{V}_t$ can be leaded to naturally.

Theorem 2.5: *Under the same conditions as in Theorem 2.4, for any constant C > 0 with probability greater than $1 - p^{-C}$, the solution of (14) satisfies*

$$\|\hat{H}_{\hat{\eta}} - \Pi_t\|_F^2 \le \frac{C' R^2}{\delta} \sqrt{\frac{\log p}{n}},$$

for some constant $C' > 0$ corresponding to C. Let $\hat{V}_t$ spans the d-dimensional principal subspace of $\hat{H}_{\hat{\eta}}$ with right rotations, then

$$\|\hat{V}_t - V_t\|_F^2 \le \frac{C' R^2}{\delta} \sqrt{\frac{\log p}{n}}.$$

Remark 2.3: Though Theorems 2.4 and 2.5 point out that $\hat{\eta} \in \mathcal{A} = \{\eta : \hat{f}(\eta) = 0\}$ and the corresponding projection matrix $\hat{H}_{\hat{\eta}}$ make sense under some regularization conditions, practically we suggest treating η as a parameter that can be chosen from cross-validation. There are two reasons for such a choice of η. First, we cannot determine how small $\hat{\eta} - \eta^*$ is because of the propagation of error during estimating $\hat{H}_{\hat{\eta}}$ and $\hat{\eta}$ iteratively. Second, our method still works well when there exists an η such that $\Sigma_b - \eta\Sigma_t$ satisfies assumptions (S1) and (S2), which is less stringent.

Refinement stage

The ℓ_1 regularization inevitably introduces some bias in the estimate. Therefore, a further refinement stage for sparse PCA is proposed in [26], which is a sparse version of the power method. Here we will use this refinement approach to improve the ℓ_1 penalized estimator. More theoretical properties of this refinement approach can be found in [26]. The implementation details can be found in Algorithm 3 in Section 3.4.

3. Algorithms

3.1. An ADMM algorithm for FLDA

Here we solve the convex relaxation step of FLDA (6) using a standard ADMM algorithm (cf. [20,21]). The corresponding scaled augmented Lagrange function is

$$L_\rho(H, Z, U) = 1_{\{Z\in\mathcal{F}^d\}} - \langle\hat{\Sigma}_b, H\rangle + \lambda\,\|H\|_{1,1} + \frac{\rho}{2}\left(\|\hat{\Sigma}_w^{1/2}H\hat{\Sigma}_w^{1/2} - Z + U\|_F^2 - \|U\|_F^2\right),$$

where $\rho > 0$ is the ADMM parameter and U is the scaled dual variable. Then ADMM solves (H, Z, U) iteratively until some convergence criterion is satisfied,

$$H^{k+1} = \arg\min L_\rho(H, Z^k, U^k) \tag{19}$$

$$= \arg\min\left\{\frac{\rho}{2}\left\|\hat{\Sigma}_w^{1/2}H\hat{\Sigma}_w^{1/2} - \left(\frac{1}{\rho}\hat{\Sigma}_w^{-1/2}\hat{\Sigma}_b\hat{\Sigma}_w^{-1/2} + Z^k - U^k\right)\right\|_F^2 + \lambda\|H\|_{1,1}\right\},$$

$$Z^{k+1} = \arg\min L_\rho(H^{k+1}, Z, U^k) = P_{\mathcal{F}^d}\left(\hat{\Sigma}_w^{1/2}H^{k+1}\hat{\Sigma}_w^{1/2} + U^k\right), \tag{20}$$

$$U^{k+1} = U^k + \hat{\Sigma}_w^{1/2}H^{k+1}\hat{\Sigma}_w^{1/2} - Z^{k+1},$$

where in (19), $\hat{\Sigma}_w^{-1/2}$ is the symmetric square root of $\hat{\Sigma}_w$.

Updating H in (19) can be solved via a classical LASSO problem (Lemma 3.1) defined in (21). In order to handle the high dimensionality, Lemma 3.1 is not that efficient due to the Kronecker product. In practice, the proximal gradient descend method directly on (19) is more preferable. For updating Z in (20), Lemma 3.2 can be used.

Lemma 3.1 ([27]): *Let vec be the vectorization operation of any matrix and $\otimes$ be the Kronecker product. Then* vec(H^{k+1}) *is the solution to the Lasso problem*

$$\text{minimize}_x \quad \frac{\rho}{2}\|\Gamma x - b\|_2^2 + \lambda\|x\|_1, \tag{21}$$

where $\Gamma = \hat{\Sigma}_w^{1/2} \otimes \hat{\Sigma}_w^{1/2}$ and $b = vec((1/\rho)\hat{\Sigma}_w^{-1/2}\hat{\Sigma}_b\hat{\Sigma}_w^{-1/2} + Z^k - U^k)$.

Lemma 3.2 ([20]): *If $X = \sum_i \gamma_i u_i u_i^{\mathrm{T}}$ is a spectral decomposition of X, then $P_{\mathcal{F}^d}(X) = \sum_i \gamma_i^+(\theta)u_u u_i^{\mathrm{T}}$, where $\gamma_i^+(\theta) = \min(\max(\gamma_i - \theta, 0), 1)$ and θ satisfies the equation $\sum_i \gamma_i^+(\theta) = d$.*

The refinement formulation (9) is a row-block penalized quadratic optimization problem, so block coordinate descent or proximal gradient descent method could be used. The main steps of sparse FLDA are summarized in Algorithm 1.

Algorithm 1 Sparse FLDA

```
Require: $\hat{\Sigma}_b, \hat{\Sigma}_w, \lambda_c \geq 0, \lambda_r \geq 0, d, \rho > 0, \epsilon > 0$
  $Z^0 \leftarrow 0, U^0 \leftarrow 0$
  repeat
    $H^k \leftarrow$ updata $H^k$ in (19) (with penalty parameter $\lambda_c$)
    $Z^k \leftarrow P_{\mathcal{F}^d}\left(\hat{\Sigma}_w^{1/2} H^k \hat{\Sigma}_w^{1/2} + U^{k-1}\right)$
    $U^k \leftarrow U^{k-1} + \hat{\Sigma}_w^{1/2} H^k \hat{\Sigma}_w^{1/2} - Z^k$
  until max $\left\{\| \hat{\Sigma}_w^{1/2} H^k \hat{\Sigma}_w^{1/2} - Z^k \|_F^2, \rho^2 \|Z^k - Z^{k-1}\|_F^2\right\} \leq d\epsilon^2$
  Compute the eigenvectors $\hat{V}_1$ corresponding to the top $d$ eigenvalues of $H^k$
  Solve
      $\hat{V}_2 = \arg\min_{X \in \mathbb{R}^{p \times d}} \left\{\frac{1}{2}\mathrm{Tr}(X^{\mathrm{T}} \hat{\Sigma}_w X) - \mathrm{Tr}(\hat{V}_1^{\mathrm{T}} \hat{\Sigma}_b X) + \lambda_r \|X\|_{2,1}\right\}$
  return $\hat{V}_{\mathrm{f}} = \hat{V}_2 (\hat{V}_2^{\mathrm{T}} \hat{\Sigma}_w \hat{V}_2)^{-1/2}$
```

3.2. An ADMM algorithm for TRLDA when η is fixed

We first solve problem (14) for a given value of η. The intersection of two convex set constraints $\mathcal{F}^d$ and ℓ_1 ball makes it difficult to solve. However, there exist algorithms for individually projecting on $\mathcal{F}_p^d$ [20] and the ℓ_1 ball [22]. So dealing with optimization problems constrained by either of them is tractable. Therefore, an ADMM algorithm can be used to solve (14). Now we rewrite model (14) in an equivalent way:

$$\begin{aligned} &\text{minimize} && -\left\langle \hat{\Sigma}_b - \eta \hat{\Sigma}_t, H \right\rangle + 1_{\{H \in \mathcal{F}^d\}} + 1_{\{\|Z\|_{1,1} \leq R\}} \\ &\text{subject to} && H - Z = 0 \end{aligned} \tag{22}$$

where $1_{\{x \in A\}}$ is an indicator function. The corresponding scaled augmented Lagrange function of (22) is

$$L_\rho(H, Z, U) = -\left\langle \hat{\Sigma}_b - \eta \hat{\Sigma}_t, H \right\rangle + 1_{\{H \in \mathcal{F}^d\}} + 1_{\{\|Z\|_{1,1} \leq R\}} + \frac{\rho}{2}\left(\|H - Z + U\|_F^2 - \|U\|_F^2\right),$$

where $\rho > 0$ is the ADMM parameter and U is the scaled dual variable. Then ADMM updates (H, Z, U) iteratively until some convergence criterion is satisfied.

$$\begin{aligned} H^{k+1} &= \arg\min L_\rho(H, Z^k, U^k) = P_{\mathcal{F}^d}\left(Z^k - U^k + \frac{1}{\rho}\left(\hat{\Sigma}_b - \eta \hat{\Sigma}_t\right)\right), \\ Z^{k+1} &= \arg\min L_\rho(H^{k+1}, Z, U^k) = P_{R-\ell_1 \mathrm{Ball}}\left(H^{k+1} + U^k\right), \\ U^{k+1} &= U^k + H^{k+1} - Z^{k+1}. \end{aligned} \tag{23}$$

Updating Z in (23) is to project the matrix $H^{k+1} + U^k$ onto a ℓ_1 ball with radius R, which can be realized using the following lemma.

Lemma 3.3 ([28]): *If x is a vector, w is the projection of x on the ℓ_1 ball with radius z, and y is the vector obtained by sorting $|x|$ in a descending order, then the number of non-zero elements in w is*

$$\tau(z, y) = \max\left\{ j \in [n] : y_j - \frac{1}{j}\left(\sum_{r=1}^{j} y_j - z\right) > 0 \right\}$$

and the value of elements in w is $w_i = \operatorname{sign}(x_i)\max\{y_i - \theta, 0\}$, where $\theta = (1/\tau(z, y))(\sum_{i=1}^{\tau(z,y)} y_i - z)$.

Note that if we estimate η from cross-validation and focus on the penalized problem (15), we actually get a regular FPS problem as in [20], which is also solved by an ADMM algorithm.

3.3. A Newton–Raphson algorithm for finding $\hat{\eta}$ in TRLDA

In this section, we give an algorithm to find a root of (16) using the Newton–Raphson method. From the proof of Theorem 2.4, $\hat{f}(\eta)$ is decreasing with $\hat{f}'(\eta) = -\langle \hat{\Sigma}_t, \hat{H}_\eta \rangle$. So the updating step for η in Newton–Raphson is

$$\hat{\eta}^{(k)} = \hat{\eta}^{(k-1)} - \frac{\hat{f}(\hat{\eta}^{(k-1)})}{\hat{f}'(\hat{\eta}^{(k-1)})} = \frac{\left\langle \hat{\Sigma}_b, \hat{H}_{\hat{\eta}^{(k-1)}} \right\rangle}{\left\langle \hat{\Sigma}_t, \hat{H}_{\hat{\eta}^{(k-1)}} \right\rangle} \tag{24}$$

for $k = 1, 2, \ldots$, until $|\hat{f}(\hat{\eta}^{(k)})|$ is smaller than a pre-chosen parameter (say, 10^{-3}, for example). The detailed steps of estimating $\hat{\eta}$ and its corresponding $\hat{H}_{\hat{\eta}}$ are summarized in Algorithm 2.

3.4. Refinement stage for TRLDA

In this section, we give the details of the refinement stage for TRLDA, as proposed in [26], in Algorithm 3.

Remark 3.1: In this refinement procedure, there is a sparsity parameter $\hat{s}$, which can be estimated based on the initialization $\hat{V}_t$ and also can be given in advance from prior knowledge. To estimate $\hat{s}$ from $\hat{V}_t$, a threshold θ_0 can be given and let $\hat{s} = |\{j : \|(\hat{V}_t)_{j*}\|_2 \geq \theta_0\}|$.

Algorithm 2 Sparse TRLDA: Convex relaxation stage

```
Require: $\hat{\Sigma}_b$, $\hat{\Sigma}_t$, $R \geq 0, d \geq 1, \rho > 0, \epsilon > 0$
  $Z^{(0)} \leftarrow 0$, $U^{(0)} \leftarrow 0$, $\eta \leftarrow 0$
  repeat                    ▷ This repeat is optional and cross-validation can be used instead
    $S \leftarrow \hat{\Sigma}_b - \hat{\eta}^{(k-1)}\hat{\Sigma}_t$                    ▷ ADMM for fixed $\hat{\eta}$
    repeat
      $H^{(k)} \leftarrow P_{\mathcal{F}_p^d}\left(Z^{(k-1)} - U^{(k-1)} + \frac{1}{\rho}S\right)$
      $Z^{(k)} \leftarrow P_{R-\ell_1\text{Ball}}\left(H^{(k)} + U^{(k-1)}\right)$
      $U^{(k)} \leftarrow U^{(k-1)} + H^{(k)} - Z^{(k)}$
    until $\max\left\{\|H^{(k)} - Z^{(k)}\|_F^2, \rho^2\|Z^{(k)} - Z^{(k-1)}\|_F^2\right\} \leq d\epsilon^2$
    $\hat{\eta}^{(k)} = \frac{\left\langle \hat{\Sigma}_b, \hat{H}_{\hat{\eta}^{(k-1)}} \right\rangle}{\left\langle \hat{\Sigma}_t, \hat{H}_{\hat{\eta}^{(k-1)}} \right\rangle}$                    ▷ Update $\hat{\eta}$
  until $\left|\left\langle S, H^{(k)} \right\rangle\right| \leq \epsilon$
  Compute the eigenvectors $\hat{V}_t$ corresponding to the top $d$ eigenvalues of $Z^{(k)}$
  return $\eta^{(k)}$ and $\hat{V}_t$
```

Algorithm 3 Sparse TRLDA: Refinement stage

```
Require: $S = \hat{\Sigma}_b - \hat{\eta}\hat{\Sigma}_t$, initialization $\hat{V}_t$ and sparsity parameter $\hat{s}$
  Find a constant $\mu > 0$ such that $S + \mu I$ is a positive definite matrix if $S$ itself is not
  repeat
    $\tilde{U}_t \leftarrow \text{Truncate}(\hat{V}_t, \hat{s})$
    $\hat{U}_t, \hat{R}_t \leftarrow \text{QR}(\tilde{U}_t)$
    $\tilde{V}_t \leftarrow S \cdot \hat{U}_t$
    $\hat{V}_t, \hat{R}_t \leftarrow \text{QR}(\tilde{V}_t)$
  until $\|\hat{U}_t - \hat{U}_t^{\text{pre}}\|_2 \leq \epsilon$
  return $\hat{U} = \hat{U}_t$
```

Remark 3.2: The Truncate step in Algorithm 3 is to select the $\hat{s}$ rows with the top $\hat{s}$ largest ℓ_2 row norms. And the QR decomposition used in Algorithm 3 is the thin QR decomposition, which guarantees that the dimension of $\hat{V}_t$ $(\hat{U}_t)$ is the same as that of $\tilde{V}_t$ $(\tilde{U}_t)$.

4. Numerical examples

In this section, we examine the performance of our two subspace-based sparse LDA methods, compared with the multigroup sparse LDA proposed in [19], which will be called MGSDA for short, using both real and simulated data sets, in which standard LDA is also be compared with. All the sparsity penalty parameters will be chosen using fivefold cross-validation based on regular LDA misclassification rates using the reduced data, and the sparsity level $\hat{s}$ in the refinement stage for TRLDA (Algorithm 3) will be estimated by setting the threshold defined in Remark 3.1 to be $\theta_0 = \max(\omega) \cdot \text{std}(\omega)$, where std stands for standard deviation and $\omega = \{\|\hat{V}_{i*}\|_2 : i \in \{1, \ldots, p\}\}$, where $\hat{V}$ is the estimated discrimination directions from the convex relaxation stage of TRLDA.

4.1. Synthetic data set

Two simulation settings, each of which consisting of four groups of Gaussian random vectors sharing the same covariance matrix, are considered in this section. In both settings, the population means are set to be

$$\mu^{(i)} = (\underbrace{0, \ldots, 0}_{3\times(i-1)}, \underbrace{1.5, \ldots, 1.5}_{3}, \underbrace{0, \ldots, 0}_{p-3\times i}) \in \mathbb{R}^p, \quad i = \{1, 2, 3, 4\}$$

In the first setting (the *Identity* setting), the common within group covariance matrix is set to be $\Sigma_w^{(1)} = I_p$. In the second setting (the *Equal Correlation* setting), we use $\Sigma_w^{(2)} = 0.5I_p + 0.5E_p$, where $E_p \in \mathbb{R}^{p\times p}$ stands for a matrix with all elements equal to 1. Then we generate $n = 100$ independent observations from $\text{N}(\mu^{(i)}, \Sigma_w^{(j)})$ for each $i \in \{1, 2, 3, 4\}$ and $j \in \{1, 2\}$ as training data. To evaluate the performance, an independent testing data of the same size is generated for each combination of (i, j). The dimensions of feature spaces considered here are $p = 100$ and $p = 500$. The experiments are repeated on 100 independently generated pairs of training and testing data sets. The average misclassification errors and the average numbers of selected features are listed in Table 1, with standard deviations in parentheses.

As seen in Table 1, while the three sparse high dimensional methods, ie, FLDA, TRLDA and MGSDA, give comparable classification accuracy, the two sparse subspace-based methods tend to provide more interpretable data reduction. When the covariance matrix is *equal correlation*, TRLDA and MGSDA give more accurate classification than FLDA because of the high noise level in $\hat{\Sigma}_w$, which is used to define the constraint set in FLDA. In both two simulation settings, MGSDA tends to

Table 1. Results for simulation.

Covariance		Identity		EquiCorr	
		$P = 100$	$P = 500$	$P = 100$	$P = 500$
Misclass error (%)	TRUE	8.23 (1.48)	8.35 (1.46)	1.37 (0.59)	1.39 (0.60)
	FLDA	8.71 (1.33)	9.08 (1.61)	1.53 (0.68)	2.18 (2.16)
	TRLDA	8.51 (1.50)	9.10 (1.60)	1.53 (0.67)	1.70 (0.76)
	MGSDA	8.92 (1.55)	9.03 (1.53)	1.61 (0.66)	1.65 (0.64)
No. of features	FLDA	12.02 (0.14)	12.36 (0.81)	11.99 (0.10)	11.84 (0.47)
	TRLDA	12.47 (1.46)	17.36 (8.56)	13.28 (5.22)	16.20 (9.21)
	MGSDA	16.43 (7.98)	17.03 (12.17)	20.95 (16.12)	26.65 (32.97)

select more features than FLDA and TRLDA, but it uses less computation times and has time-saving advantages.

Remark 4.1: Due to the slow convergence of the ADMM and the proximal gradient descent method, in FLDA we used hard thresholding on the rows of $\hat{V}$ after the refinement stage, with a threshold chosen similarly as that used in the refinement stage of TRLDA. In practice, a different threshold may be used.

4.2. Real data analysis

In this section, we apply our methods to the handwritten zip code data. The digital numbers data set from website https://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit consists of 1593 handwritten digits, which are stretched in a rectangular box 16×16 in a gray scale of 256 values, from about 80 persons. Each pixel in this data set was scaled into a boolean (1/0) value using a fixed threshold, and each handwritten digit belongs to $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. For ease of illustration, we choose three groups of digit sets ($\{3, 5, 8\}$, $\{2, 7, 9\}$ and $\{1, 4, 6, 9\}$) to test the performance of our methods. For each set, we randomly split the observations into a training data set (0.8 proportion) and a testing data set (0.2 proportion). Then the sparse LDA methods are applied on the training set and evaluated on the testing set. The procedure is repeated 10 times. The means of misclassification errors and the numbers of selected features with their standard deviations in parentheses are listed in Table 2.

As shown in Table 2, FLDA and MGSDA give more accurate classification than TRLDA in the first two groups of digit sets $\{3, 5, 8\}$ and $\{2, 7, 9\}$, but the result of another set $\{1, 4, 6, 9\}$ is the reverse. From the perspective of dimension reduction, TRLDA tends to select more features than the other two methods. Overall FLDA performs similarly to MGSDA and they both perform differently from TRLDA. A possible reason for such a difference is that the FLDA and MGSDA are based on the 'trace of ratio' formulation of LDA, whereas TRLDA is based on the 'ratio of traces'. These two formulations have different motivating models and can lead to different practical performance. In our experiment, FLDA and MGSDA do better for digit sets $\{3, 5, 8\}$ and $\{2, 7, 9\}$, while TRLDA is more accurate for digit set $\{1, 4, 6, 9\}$, which is the hardest data set for LDA.

Table 2. Result for real data analysis.

Digit set		$\{3, 5, 8\}$	$\{2, 7, 9\}$	$\{1, 4, 6, 9\}$
Misclass error (%)	FLDA	3.85 (1.39)	1.77 (1.21)	4.02 (2.01)
	TRLDA	4.27 (2.75)	2.19 (2.05)	3.70 (1.70)
	MGSDA	3.75 (1.41)	1.77 (1.30)	4.17 (1.44)
No. of features	FLDA	171.50 (29.21)	136.60 (32.38)	190.50 (27.64)
	TRLDA	159.3 (25.07)	163.60 (64.02)	253.10 (1.523)
	MGSDA	120.70 (24.81)	101.40 (32.53)	149.30 (33.65)

5. Discussion

In this paper, we extended the sparse subspace framework to cover high dimensional LDA, the resulting algorithms have good theoretical properties and competitive empirical performance. In particular, the subspace perspective allows us to efficiently use the trace-ratio formulation of LDA with sparsity penalty. In our data examples, the sparse trace-ratio LDA is shown to be a useful complement to the sparse Fisher LDA and other existing sparse LDA methods.

Our sparse subspace formulations of FLDA and TRLDA require solving SDP problems for large matrices. For large-scale problems, the high computational demand of SDP will limit our ability to carefully select tuning parameters using cross-validation. There are several computational tricks that may further improve the performance. First, in the Fantope projection step of the ADMM algorithms, one may apply the projection only on a low-rank approximation of the input matrix. This is based on the observation that most input matrices are nearly low-rank, and the majority of small eigenvalues make little difference in the projection. Second, based on results in [29], one does not have to run the ADMM algorithm very close to convergence. In fact, it suffices to run ADMM so that the current solution is within a certain constant distance from the target. After that, a fast projected gradient descent algorithm will greatly expedite the convergence as the refinement step. Implementation of these ideas, as well as more refined tuning parameter selection for subspace-based sparse LDA, will be pursued in future works.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jing Lei http://orcid.org/0000-0003-3104-9387

References

[1] Fisher RA. The use of multiple measurements in taxonomic problems. Ann human genetics. 1936;7(2):179–188.

[2] Fukunaga K. Introduction to statistical pattern recognition. New York: Academic Press; 2013.

[3] Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. Ann Stat. 1995;23(1):73–102.

[4] Mardia KV, Kent JT, Bibby JM. Multivariate analysis. New York: Academic Press; 1979.

[5] McLachlan GJ. Discriminant analysis and statistical pattern recognition. Vol. 544. Hoboken, NJ: John Wiley & Sons; 2004.

[6] Fukunaga K. Introduction to statistical pattern recognition. New York: Academic Press; 1972.

[7] Johnson RA, Wichern DW. Applied multivariate statistical analysis. Vol. 47. Upper Saddle River, NJ: Prentice Hall Inc; 1992.

[8] Wilks S. Multidimensional statistical scatter. In: Olkin, I, et al., editors. Contributions to probability and statistics: essays in honor of Harold hotelling. Palo Alto, CA: Stanford University Press; 1960. p. 486–503.

[9] Guo YF, Li SJ, Yang JY, et al. A generalized Foley–Sammon transform based on generalized Fisher discriminant criterion and its application to face recognition. Pattern Recognit Lett. 2003;24(1–3):147–158.

[10] Milgrom P, Segal I. Envelope theorems for arbitrary choice sets. Econometrica. 2002;70(2):583–601.

[11] Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli. 2004;10(6):989–1010.

[12] Fan J, Fan Y. High-dimensional classification using features annealed independence rules. Ann Stat. 2008;36(6):2605–2637.

[13] Shao J, Wang Y, Deng X, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. Ann Stat. 2011;39(2):1241–1265.

[14] Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. J Am Stat Assoc. 2011;106(496):1566–1577.

[15] Mai Q, Zou H, Yuan M. A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika. 2012;99(1):29–42.

[16] Witten DM, Tibshirani R. Penalized classification using Fisher's linear discriminant. J R Stat Soc Ser B. 2011;73(5):753–772.

[17] Clemmensen L, Hastie T, Witten D, et al. Sparse discriminant analysis. Technometrics. 2011;53(4):406–413.

[18] Hastie T, Tibshirani R, Buja A. Flexible discriminant analysis by optimal scoring. J Am Stat Assoc. 1994;89(428):1255–1270.

[19] Gaynanova I, Booth JG, Wells MT. Simultaneous sparse estimation of canonical vectors in the $p > > n$ setting. J Am Stat Assoc. 2016;111(514):696–706.

[20] Vu VQ, Cho J, Lei J, et al. Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. In: Burges CJC, Bottou L, Welling M, et al., editors. Advances in neural information processing systems. Lake Tahoe, NV: Curran Associates, Inc.; 2013. p. 2670–2678.

[21] Gao C, Ma Z, Ren Z, et al. Minimax estimation in sparse canonical correlation analysis. Ann Stat. 2015;43(5):2168–2197.

[22] Dattorro J. Convex optimization and Euclidean distance geometry. Vol. 30. Palo Alto, CA: Meboo Publishing; 2005.

[23] d'Aspremont A, El Ghaoui L, Jordan MI, et al. A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. 2007;49(3):434–448.

[24] Lei J, Vu VQ. Sparsistency and agnostic inference in sparse PCA. Ann Statist. 2015;43(1):299–322.

[25] Fan K. On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc. Natl. Acad. Sci. U.S.A. 1949;35(11):652– 655.

[26] Wang Z, Lu H, Liu H. Tighten after relax: minimax-optimal sparse PCA in polynomial time. In: Ghahramani Z, Welling M, Cortes C, et al., editors. Advances in neural information processing systems. Montreal, Canada: Curran Associates, Inc.; 2014. p. 3383–3391.

[27] Gao C, Ma Z, Zhou HH. Sparse CCA: adaptive estimation and computational barriers. Ann Statist. 2017;45(5):2074–2101.

[28] Duchi J, Shalev-Shwartz S, Singer Y, et al. Efficient projections onto the L1-ball for learning in high dimensions. Proceedings of the 25th international conference on Machine learning, ICML; Helsinki, Finland; 2008. p. 272–279.

[29] Chen Y, Wainwright MJ. Fast low-rank estimation by projected gradient descent: general statistical and algorithmic guarantees. 2015. arXiv preprint arXiv:1509.03025.

Appendix

Proof of Theorem 2.2: Define four variables which will be used in this section,

$$\tilde{V}_{\mathrm{f}} = V_{\mathrm{f}}\left({V_{\mathrm{f}}}^{\mathrm{T}}\hat{\Sigma}_w V_{\mathrm{f}}\right)^{-1/2}, \quad \tilde{\Pi} = \tilde{V}_{\mathrm{f}}\tilde{V}_{\mathrm{f}}^{\mathrm{T}},$$
$$\tilde{\Sigma}_b = \hat{\Sigma}_w V_{\mathrm{f}}\Lambda {V_{\mathrm{f}}}^{\mathrm{T}}\hat{\Sigma}_w, \quad \tilde{\Lambda} = ({V_{\mathrm{f}}}^{\mathrm{T}}\hat{\Sigma}_w V_{\mathrm{f}})^{1/2}\Lambda({V_{\mathrm{f}}}^{\mathrm{T}}\hat{\Sigma}_w V_{\mathrm{f}})^{1/2},$$

where Λ is defined in the following lemma. ■

Lemma A.1: *If $d = \mathrm{rank}(\Sigma_b)$, there exists a diagonal matrix Λ such that*

$$\Sigma_b = \Sigma_w V_{\mathrm{f}}\Lambda V_{\mathrm{f}}^{\mathrm{T}}\Sigma_w.$$

Proof: From the definition of V_{f}, $\Sigma_w^{1/2}V_{\mathrm{f}}$ is the optimal solution of the following optimization problem,

$$\max \mathrm{Tr}(U^{\mathrm{T}}\Sigma_w^{-1/2}\Sigma_b\Sigma_w^{-1/2}U) \quad \text{subjectto} : U^{\mathrm{T}}U = I_d.$$

From eigenvalue decomposition and $d = \mathrm{rank}(\Sigma_b)$, there exists a diagonal matrix Λ such that

$$\Sigma_w^{-1/2}\Sigma_b\Sigma_w^{-1/2} = \Sigma_w^{1/2}V_{\mathrm{f}}\Lambda V_{\mathrm{f}}^{\mathrm{T}}\Sigma_w^{1/2}.$$

We complete the proof by rearranging terms. ■

Before giving the main proof for Theorem 2.2, we introduce several technical lemmas.

Lemma A.2: *If $X \in \mathbb{R}^{p\times p}$ has at most r non-zero rows and columns, and $S \in \mathbb{R}^{p\times p}$ is a positive semidefinite matrix, then*

$$\phi_{\min}^S(r)\|X\|_F \le \|S^{1/2}XS^{1/2}\|_F \le \phi_{\max}^S(r)\|X\|_F,$$

where

$$\phi_{\min}^S(r) = \min_{\|u\|_0\le r, u\ne 0}\frac{u^{\mathrm{T}}Su}{u^{\mathrm{T}}u}, \quad \phi_{\max}^S(r) = \max_{\|u\|_0\le r, u\ne 0}\frac{u^{\mathrm{T}}Su}{u^{\mathrm{T}}u}.$$

Proof: Let $W = XS^{1/2}$. For any $i \in \{1, 2, \ldots, p\}$, we have $\|W_i\|_0 \leq r$. Then

$$\|S^{1/2}XS^{1/2}\|_F^2 = \mathrm{Tr}(S^{1/2}X^{\mathrm{T}}SXS^{1/2}) = \sum_{i=1}^{p} W_i^{\mathrm{T}} S W_i \leq \sum_{i=1}^{p} \phi_{\max}^{S}(r)\|W_i\|_2^2$$
$$= \phi_{\max}^{S}(r)\|W\|_F^2 = \phi_{\max}^{S}(r)\mathrm{Tr}(S^{1/2}X^{\mathrm{T}}XS^{1/2}) = \phi_{\max}^{S}(r)\mathrm{Tr}(XSX^{\mathrm{T}}) \leq \phi_{\max}^{S}(r)^2\|X\|_F^2.$$

In the same way, the other inequality can be obtained. ■

Lemma A.3: *If $\Sigma_w \in \mathcal{P}_w$ and under condition* (C1), *there exists a constant $C > 0$ such that*

$$\phi_{\min}^{\hat{\Sigma}_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\hat{\Sigma}_w}(m) > \frac{1}{C} \quad \text{and} \quad \phi_{\max}^{\hat{\Sigma}_w}(m) < C,$$

where C is determined by C_n in condition (C1), *C_m and C_w in matrix family $\mathcal{P}_w$.*

Proof: First, using the Lemma 12 in [21], for any support set $J = S \times S$ with $|S| = u$, we have

$$\|[\hat{\Sigma}_w - \Sigma_w]_J\|_{op}^2 \leq \frac{C}{n}(u\log(ep/u)),$$

with probability greater than $1 - \exp(-C'u\log(ep/u))$. Then from the definition of $\phi_{\max}$ and $\phi_{\min}$, we have

$$\begin{aligned}\phi_{\min}^{\hat{\Sigma}_w}(q+m) &\geq \phi_{\min}^{\Sigma_w}(q+m) + \min_{\|u\|_0 \leq q+m, u\neq 0} \frac{u^{\mathrm{T}}(\hat{\Sigma}_w - \Sigma_w)u}{u^{\mathrm{T}}u}\\ &\geq \phi_{\min}^{\Sigma_w}(q+m) - \max_{\|u\|_0 \leq q+m, u\neq 0} \left|\frac{u^{\mathrm{T}}(\hat{\Sigma}_w - \Sigma_w)u}{u^{\mathrm{T}}u}\right|\\ &\geq \phi_{\min}^{\Sigma_w}(q+m) - C\sqrt{\frac{(q+m)\log p}{n}}\\ \phi_{\max}^{\hat{\Sigma}_w}(m) &\leq \phi_{\max}^{\Sigma_w}(m) + \max_{\|u\|_0 \leq m, u\neq 0} \frac{u^{\mathrm{T}}(\hat{\Sigma}_w - \Sigma_w)u}{u^{\mathrm{T}}u}\\ &\leq \phi_{\max}^{\Sigma_w}(m) + C\sqrt{\frac{m\log p}{n}},\end{aligned} \tag{A1}$$

which leads to

$$\begin{aligned}&\phi_{\min}^{\hat{\Sigma}_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\hat{\Sigma}_w}(m)\\ &\quad\geq \phi_{\min}^{\Sigma_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\Sigma_w}(m) - C\sqrt{\frac{(q+m)\log p}{n}} - 3Cqm^{-1/2}\sqrt{\frac{m\log p}{n}}\\ &\quad\geq \phi_{\min}^{\Sigma_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\Sigma_w}(m) - C'\sqrt{\frac{(q+m)\log p}{n}},\end{aligned} \tag{A2}$$

where C' only depends on C and C_m in $\mathcal{P}_w$. From the fact that $\Sigma_w \in \mathcal{P}_w$ and under condition (C1), we complete the proof from inequalities (A1) and (A2). ■

Lemma A.4 ([27]): *(Curvature) Let $F \in \mathcal{F}^d$ and E such that $E^{\mathrm{T}}E = I_d$. For any matrix $A \in \mathbb{R}^{d\times d}$ and diagonal matrix $D = \mathrm{diag}(D_1, D_2, \ldots, D_d)$ with decreasing order, we have*

$$\left\langle EAE^{\mathrm{T}}, F - EE^{\mathrm{T}}\right\rangle \leq \|A - D\|_F\|F - EE^{\mathrm{T}}\|_F - \frac{D_d}{2}\|F - EE^{\mathrm{T}}\|_F^2.$$

Proof: For $F \in \mathcal{F}^d$ and $E^{\mathrm{T}}E = I_d$, $\|F\|_{\mathrm{op}} \leq 1$ and $\|E_i\|_F = 1$ for any $i = 1, 2, \ldots, d$. So $E_i^{\mathrm{T}}FE_i \leq \|F\|_{\mathrm{op}}\|E_i\|_F^2 \leq 1$.

$$\begin{aligned}\left\langle EAE^{\mathrm{T}}, F - EE^{\mathrm{T}}\right\rangle &= \left\langle D, E^{\mathrm{T}}FE - I\right\rangle + \left\langle A - D, E^{\mathrm{T}}FE - I\right\rangle\\ &\leq \sum_{i=1}^{d} D_i(E_i^{\mathrm{T}}FE_i - 1) + \|A - D\|_F\|E^{\mathrm{T}}FE - I\|_F\\ &\leq D_d\sum_{i=1}^{d}(E_i^{\mathrm{T}}FE_i - 1) + \|A - D\|_F\|F - EE^{\mathrm{T}}\|_F.\end{aligned} \tag{A3}$$

First part of the right-hand side of (A3) can be upper bounded by

$$\begin{aligned}\sum_{i=1}^{d}(E_i^{\mathrm{T}}FE_i-1)&=\sum_{i=1}^{d}E_i^{\mathrm{T}}FE_i-d\\&\le\sum_{i=1}^{d}E_i^{\mathrm{T}}FE_i-\frac{1}{2}\left(\|EE^{\mathrm{T}}\|_F^2+\|F\|_{\mathrm{op}}\|F\|_*\right)\\&\le-\frac{1}{2}\left[\|EE^{\mathrm{T}}\|_F^2+\|F\|_F^2-2\mathrm{Tr}\left(E^{\mathrm{T}}FE\right)\right]\\&=-\frac{1}{2}\|F-EE^{\mathrm{T}}\|_F^2.\end{aligned}\tag{A4}$$

We complete the proof by combining inequalities (A3) and (A4). ■

Lemma A.5: *Under condition* (C1), *with probability greater than* $1-p^{-C'}$, *we have*

$$\left\|\hat{\Sigma}_b-\tilde{\Sigma}_b\right\|_{\infty,\infty}\le C\sqrt{\frac{\log p}{n}},$$

where C is sufficiently large constant only depending on C'.

This Lemma can be proved using Hoeffding's inequality and the proof is omitted here. The following technical lemma is a special case of Lemma 6.1 in [27].

Lemma A.6: *Let* $\epsilon_n=\sqrt{(1/n)(q+\log(ep/q))}$. *Assume* $\epsilon_n^2\le C$ *for some small constant* $C\in(0,1)$. *Then, for any* $C'>0$, *there exists* $C_0>0$ *only depending on* C' *such that*

$$\|\Sigma_w^{1/2}(\tilde{V}_{\mathrm{f}}-V_{\mathrm{f}})\|_{\mathrm{op}}\le C_0\epsilon_n,\qquad\|\tilde{\Lambda}-\Lambda\|_{\mathrm{op}}\le C_0\epsilon_n,$$

with probability greater than $1-\exp(-C'(q+\log(ep/q)))$.

Proof of Theorem 2.2: From the optimality of $\hat{H}$ and feasibility of $\tilde{\Pi}$, we have

$$\langle\hat{\Sigma}_b,\hat{H}\rangle-\lambda\|\hat{H}\|_{1,1}\ge\langle\hat{\Sigma}_b,\tilde{\Pi}\rangle-\lambda\|\tilde{\Pi}\|_{1,1}.$$

By rearranging the terms and using Lemma A.5, with high probability we have

$$\begin{aligned}\langle\tilde{\Sigma}_b,\Delta\rangle&\ge\lambda\|\tilde{\Pi}+\Delta\|_{1,1}-\lambda\|\tilde{\Pi}\|_{1,1}-\langle\hat{\Sigma}_b-\tilde{\Sigma}_b,\Delta\rangle\\&\ge\lambda\|\Delta_{J^c}\|_{1,1}-\lambda\|\Delta_J\|_{1,1}-\|\hat{\Sigma}_b-\tilde{\Sigma}_b\|_{\infty,\infty}\|\Delta\|_{1,1}\\&\ge\frac{\lambda}{2}\|\Delta_{J^c}\|_{1,1}-\frac{3\lambda}{2}\|\Delta_J\|_{1,1},\end{aligned}\tag{A5}$$

where $\Delta=\hat{H}-\tilde{\Pi}$ and J is corresponding to the true support of Π. By application of Lemma A.4, we bound the left-hand side of (A5) by

$$\langle\tilde{\Sigma}_b,\Delta\rangle\le\|\Lambda-\tilde{\Lambda}\|_F\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F-\frac{\lambda_d}{2}\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F^2.\tag{A6}$$

Inequalities (A5) together with (A6) lead to

$$\begin{aligned}\lambda_d\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F^2&\le3\lambda\|\Delta_J\|_{1,1}-\lambda\|\Delta_{J^c}\|_{1,1}+2\|\Lambda-\tilde{\Lambda}\|_F\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F\\&\le3\lambda\|\Delta_J\|_{1,1}+2\|\Lambda-\tilde{\Lambda}\|_F\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F,\end{aligned}$$

from which, using the roots of quadratic equations, we have

$$\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F^2\le\frac{4\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda_d^2}+\frac{6\lambda\|\Delta_J\|_{1,1}}{\lambda_d}\le\frac{4\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda_d^2}+\frac{6q\lambda\|\Delta_J\|_F}{\lambda_d}\tag{A7}$$

and

$$\begin{aligned}\lambda\left\|\Delta_{J^c}\right\|_{1,1}&\le3\lambda\left\|\Delta_J\right\|_{1,1}+2\|\Lambda-\tilde{\Lambda}\|_F\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F-\lambda_d\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F^2\\&\le3\lambda\left\|\Delta_J\right\|_{1,1}+\frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda_d},\end{aligned}\tag{A8}$$

which will be used to give an upper bound of Δ together with inequality (A7). Let J_k be the indices corresponding to the m largest entries in Δ outside of $J \cup \bigcup_{j=1}^{k-1} J_j$. And $\tilde{J} = J \cup J_1$. The size of the last index set J_K is allowed to be less than m. Then from Lemma A.2, we have

$$\begin{aligned}\sum_{k=2}^{K} \|\hat{\Sigma}_w^{1/2}\Delta_{J_k}\hat{\Sigma}_w^{1/2}\|_F &\le \sum_{k=2}^{K}\phi_{\max}^{\hat{\Sigma}_w}(m)\|\Delta_{J_k}\|_F \le \sum_{k=2}^{K}\phi_{\max}^{\hat{\Sigma}_w}(m)m^{1/2}\|\Delta_{J_k}\|_{\infty,\infty}\\ &\le \sum_{k=2}^{K}\phi_{\max}^{\hat{\Sigma}_w}(m)m^{-1/2}\|\Delta_{J_{k-1}}\|_{1,1} \le \phi_{\max}^{\hat{\Sigma}_w}(m)m^{-1/2}\|\Delta_{J^c}\|_{1,1}\\ &\le \phi_{\max}^{\hat{\Sigma}_w}(m)m^{-1/2}\left(3\left\|\Delta_J\right\|_{1,1} + \frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d}\right)\\ &\le \phi_{\max}^{\hat{\Sigma}_w}(m)m^{-1/2}\left(3q\left\|\Delta_{\tilde{J}}\right\|_F + \frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d}\right),\end{aligned} \tag{A9}$$

which leads to

$$\begin{aligned}\|\hat{\Sigma}_w^{1/2}\Delta\hat{\Sigma}_w^{1/2}\|_F &\ge \|\hat{\Sigma}_w^{1/2}\Delta_{\tilde{J}}\hat{\Sigma}_w^{1/2}\|_F - \sum_{k=2}^{K}\|\hat{\Sigma}_w^{1/2}\Delta_{J_k}\hat{\Sigma}_w^{1/2}\|_F\\ &\ge \phi_{\min}^{\hat{\Sigma}_w}(q+m)\|\Delta_{\tilde{J}}\|_F - \phi_{\max}^{\hat{\Sigma}_w}(m)m^{-1/2}\left(3q\left\|\Delta_{\tilde{J}}\right\|_F + \frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d}\right)\\ &= \left(\phi_{\min}^{\hat{\Sigma}_w}(q+m) - 3qm^{-1/2}\phi_{\max}^{\hat{\Sigma}_w}(m)\right)\left\|\Delta_{\tilde{J}}\right\|_F - \frac{\phi_{\max}^{\hat{\Sigma}_w}(m)\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d\sqrt{m}}\\ &= \kappa_1\left\|\Delta_{\tilde{J}}\right\|_F - \frac{\kappa_2\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d\sqrt{m}},\end{aligned}$$

where $\kappa_1 = \phi_{\min}^{\hat{\Sigma}_w}(q+m) - 3\phi_{\max}^{\hat{\Sigma}_w}(m)qm^{-1/2}$ and $\kappa_2 = \phi_{\max}^{\hat{\Sigma}_w}(m)$, and from Lemma A.3, both κ_1 and κ_2 are bounded positive constants with high probability. Together with (A7), there exists a constant $C > 0$ such that

$$\left\|\Delta_{\tilde{J}}\right\|_F \le C\left[\frac{q\lambda}{\lambda_d} + \frac{\|\Lambda-\tilde{\Lambda}\|_F}{\lambda_d} + \frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d\sqrt{m}}\right].$$

From the process of inequality (A9), we have

$$\left\|\Delta_{\tilde{J}^c}\right\|_F \le m^{-1/2}\left(3q\left\|\Delta_{\tilde{J}}\right\|_F + \frac{\|\Lambda-\tilde{\Lambda}\|_F^2}{\lambda\lambda_d}\right).$$

From Lemma A.6 and the fact that $d \le q$, with probability greater than $1-\exp(-C'(q+\log(ep/q)))$, we have

$$\|\Lambda-\tilde{\Lambda}\|_F \le C\sqrt{q}\epsilon_n \le Cq\lambda,$$

for some $C > 0$. Hence,

$$\|\Delta\|_F \le C\frac{q\lambda}{\lambda_d}. \tag{A10}$$

Using Lemma A.6 again, with probability greater than $1-\exp(-C'(q+\log(ep/q)))$, we have

$$\|V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}}\|_{\mathrm{op}} = \|\Sigma_w^{-1/2}\Sigma_w^{1/2}(V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}})\|_{\mathrm{op}} \le \|\Sigma_w^{-1/2}\|_{\mathrm{op}}\|\Sigma_w^{1/2}(V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}})\|_{\mathrm{op}} \le \lambda_{\min}^{-1/2}(\Sigma_w)\epsilon_n,$$

which contributes to

$$\begin{aligned}\|V_{\mathrm{f}}V_{\mathrm{f}}^{\mathrm{T}} - \tilde{V}_{\mathrm{f}}\tilde{V}_{\mathrm{f}}\|_{\mathrm{op}} &\le 2\|(V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}})V_{\mathrm{f}}^{\mathrm{T}}\|_{\mathrm{op}} + \|(V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}})(V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}})^{\mathrm{T}}\|_{\mathrm{op}}\\ &\le 2\|V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}}\|_{\mathrm{op}}\|V_{\mathrm{f}}\|_{\mathrm{op}} + \|V_{\mathrm{f}} - \tilde{V}_{\mathrm{f}}\|_{\mathrm{op}}^2\\ &\le 3\lambda_{\min}^{-1}(\Sigma_w)\epsilon_n \le C\epsilon_n,\end{aligned}$$

the last inequality is from the fact that Σ_w is positive definite. So

$$\|V_{\mathrm{f}}V_{\mathrm{f}}^{\mathrm{T}} - \tilde{V}_{\mathrm{f}}\tilde{V}_{\mathrm{f}}\|_F \le \sqrt{q}\|V_{\mathrm{f}}V_{\mathrm{f}}^{\mathrm{T}} - \tilde{V}_{\mathrm{f}}\tilde{V}_{\mathrm{f}}\|_{\mathrm{op}} \le C\sqrt{q}\epsilon_n \le C'q\lambda. \tag{A11}$$

Combining inequalities (A10) and (A11), with probability at least $1 - \exp(-C'(q + \log(ep/q)))$, we have

$$\|\hat{H} - \Pi\|_F \leq C\frac{q\lambda}{\lambda_d},$$

for sufficiently large constant $C > 0$. ■

Lemma for equivalence between problems (2) and (3)

Lemma A.7: *Same conditions on $\Sigma_w \in \mathbb{R}^{p\times p}$ and $\Sigma_b \in \mathbb{R}^{p\times p}$ are assumed as in Lemma 2.3. Then*

$$\arg\max \left\{ \frac{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right)}{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_w V\right)} : V^{\mathrm{T}}V = I_d \right\} = \arg\max \left\{ \frac{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_b V\right)}{\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_t V\right)} : V^{\mathrm{T}}V = I_d \right\}$$

Proof: Since Σ_w is a positive define matrix, for any $V \in \mathbb{R}^{n\times d}$ satisfying $V^{\mathrm{T}}V = I_d$, we have

$$\mathrm{Tr}\left(V^{\mathrm{T}}\Sigma_w V\right) > 0.$$

Together with the fact that function $g(x) : \mathbb{R}^+ \to [0, 1]$ defined as

$$g(x) = \frac{x}{1 + x},$$

is increasing, we can obtain $X(V) := \mathrm{Tr}(V^{\mathrm{T}}\Sigma_b V)/\mathrm{Tr}(V^{\mathrm{T}}\Sigma_w V)$ attains its maximum iff $g(X(V))$ attains its maximum. ■

Envelope theorem

Theorem A.8 ([10]): *Let X denote the choice set and let the relevant parameter be $t \in [0, 1]$. Letting $f : X \times [0, 1] \to \mathbb{R}$ denote the parameterized objective function, the value function $V(t)$ and the optimal choice correspondence X^* are given by*

$$V(t) = \max_{x\in X} f(x, t),$$
$$X^*(t) = \left\{x \in X, f(x, t) = V(t)\right\}.$$

Suppose the $f(x, \cdot)$ is differentiable for all $x \in X$ and that $X^*(t) \neq \emptyset$ almost everywhere on $[0, 1]$. In addition, suppose also there exist an integrable function $b : [0, 1] \to R_+$ such that $|f_t(x, t)| \leq b(t)$ for all $x \in X$ and almost all $t \in [0, 1]$. Then V is differentiable almost everywhere, and

$$V'(t) = f_t(x^*(t), t)$$
$$V(t) = V(0) + \int_0^t f_t(x^*(s), s)\,\mathrm{d}s.$$

Proof of Theorem 2.4: For simplicity of notations, let

$$f(\eta) = \sup_{\substack{H\in\mathcal{F}_p^d \\ \|H\|_{1,1}\leq R}} \langle \Sigma_b - \eta\Sigma_t, H\rangle, \quad \Sigma^* = \Sigma_b - \eta_t^*\Sigma_t,$$

$$S_\eta = \hat{\Sigma}_b - \eta\hat{\Sigma}_t, \qquad W_\eta = S_\eta - \Sigma^*.$$

First, we prove that $\mathcal{A}$ is not empty. From Theorem A.8, $\hat{f}(\eta)$ is differentiable almost everywhere and its derivative is given by

$$\hat{f}'(\eta) = -\left\langle \hat{\Sigma}_t, \hat{H}_\eta \right\rangle \leq 0,$$

and together with $\hat{f}(0) \geq 0$ and $\hat{f}(1) \leq 0$, we declare that $\mathcal{A}$ is not empty.

Since $\|\Pi_t\|_{1,1} \le R$, we have $f(\eta_t^*) = 0$ and for any $\eta \in [0,1]$,

$$\left|f(\eta) - \hat{f}(\eta)\right| \le \sup_{\substack{H\in\mathcal{F}_p^d \\ \|H\|_{1,1}\le R}} \left|\left\langle\left(\Sigma_b - \hat{\Sigma}_b\right) - \eta\left(\Sigma_t - \hat{\Sigma}_t\right), H\right\rangle\right| \le C\sqrt{\frac{\log p}{n}}R, \tag{A12}$$

where the last inequality can be got from Hoeffding's Inequality and for any $0 \le \eta_1 < \eta_2 \le 1$,

$$\begin{aligned}\hat{f}(\eta_1) - \hat{f}(\eta_2) &\ge - \sup_{\substack{H\in\mathcal{F}_p^d \\ \|H\|_{1,1}\le R}} \left\langle(\eta_1-\eta_2)\hat{\Sigma}_t, H\right\rangle \\ &\ge - \sup_{\substack{H\in\mathcal{F}_p^d \\ \|H\|_{1,1}\le R}} \langle(\eta_1-\eta_2)\Sigma_t, H\rangle - \sup_{\substack{H\in\mathcal{F}_p^d \\ \|H\|_{1,1}\le R}} \left\langle(\eta_1-\eta_2)\left(\hat{\Sigma}_t - \Sigma_t\right), H\right\rangle \\ &\ge (\eta_2-\eta_1)\left[\sum_{i=p-d+1}^{p}\lambda_i(\Sigma_t) - C\sqrt{\frac{\log p}{n}}R\right]. \end{aligned} \tag{A13}$$

Combining (A12) and (A13) and letting

$$\gamma = \frac{1}{\sum_{i=p-d+1}^{p}\lambda_i(\Sigma_t) - C\sqrt{\frac{\log p}{n}}R},$$

we have

$$\left|\eta^* - \hat{\eta}\right| \le \gamma\left|\hat{f}(\eta^*) - \hat{f}(\hat{\eta})\right| = \gamma\left|\hat{f}(\eta^*)\right| = \gamma\left|\hat{f}(\eta^*) - f(\eta^*)\right| \le C\gamma\sqrt{\frac{\log p}{n}}R,$$

as a consequence,

$$\|W_{\hat{\eta}}\|_{\infty,\infty} \le \left\|\Sigma_b - \hat{\Sigma}_b\right\|_{\infty,\infty} + |\hat{\eta} - \eta^*|\,\|\Sigma_t\|_{\infty,\infty} + |\hat{\eta}|\left\|\Sigma_t - \hat{\Sigma}_t\right\|_{\infty,\infty} \le C\sqrt{\frac{\log p}{n}}R. \qquad \blacksquare$$

Proof of Theorem 2.5: From the curvature lemma in [20, Lemma 3.1],

$$\frac{\delta}{2}\|\Delta\|_F^2 \le -\langle\Sigma^*, \Delta\rangle, \tag{A14}$$

where $\Delta = \hat{H}_t - \Pi_t$, then together with the optimality of $\hat{H}$,

$$\langle S_{\hat{\eta}}, \Delta\rangle \ge 0, \tag{A15}$$

with high probability, we have

$$\frac{\delta}{2}\|\Delta\|_F^2 \le \langle W_{\hat{\eta}}, \Delta\rangle \le \|W_{\hat{\eta}}\|_{\infty,\infty}\|\Delta\|_{1,1} \le C\sqrt{\frac{\log p}{n}}R^2, \tag{A16}$$

for some positive constant C. Then with suitable rotation, we have

$$\|\hat{V}_t - V_t\|_F^2 \le \|\Delta\|_F^2 \le \frac{CR^2}{\delta}\sqrt{\frac{\log p}{n}}. \qquad \blacksquare$$