Supplementary Materials: An Efficient Convex Formulation for Reduced-rank Linear Discriminant Analysis in High Dimensions

Jing Zeng, Xin Zhang and Qing Mai

Florida State University

S1 Tuning parameters

The tuning parameters λ_1 and λ_2 are selected by cross-validation. For the given tuning parameter $\lambda = (\lambda_1, \lambda_2)$, let $\widehat{\beta}(\lambda)$ denote the estimator based on the training set in one training-validation set split. We reduce the dimension of the predictors in the training and validation sets by $\widetilde{X}_i^{\text{tr}} = \widehat{\beta}(\lambda)^\top X_i^{\text{tr}}$ and $\widetilde{X}_j^{\text{val}} = \widehat{\beta}(\lambda)^\top X_j^{\text{val}}$, $i = 1, \ldots, n_{\text{tr}}$, $j = 1, \ldots, n_{\text{val}}$, where n_{tr} and n_{val} are the sample sizes of the training and validation sets respectively. Then we fit the classical LDA model to the reduced training set $(\widetilde{X}_i^{\text{tr}}, Y_i^{\text{tr}})$, $i = 1, \ldots, n_{\text{tr}}$, and make the prediction on the reduced validation set $\widetilde{X}_i^{\text{val}}$, $i = 1, \ldots, n_{\text{val}}$. The optimal tuning parameters λ_1 and λ_2 are selected such that the averaged validation criterion over all set splits is maximized.

Three commonly used validation criteria are adopted in LSLDA algorithm, and users choose one to use depending on the structure of data. In

general, the accuracy $(1/n)\sum_{i=1}^{n} I(\widehat{Y}_i = Y_i)$ is a proper choice, where $\widehat{Y}_i$ is the predicted response. However, when the data set is unbalanced, i.e., some classes dominate the others in sample size, the geometric mean is a better choice. The geometric mean is defined as $\prod_{k=1}^{K}(n_{kk}/n_k)^{1/K}$ with n_{kk} denoting the number of correctly classified samples in class k, which measures the balance between the classification performances on majority and minority classes (see Soda 2011, Mirza et al. 2015). In response category combination problems where some classes are indistinguishable, even the Bayes classifier only does the random classification among these indistinguishable classes. In this case, we adopt the validation log-likelihood as the criterion, which is defined as $\sum_{i=1}^{n}\sum_{k=1}^{K} I(Y_i = k)\log \Pr(Y_i = k \mid X_i^{\text{val}})$. This criterion is also used in Price, Geyer and Rothman (2019).

In simulation studies, we use geometric mean in Model (M3), validation log-likelihood in Model (M7), and accuracy in the rest of models. In real data analysis, since all three data sets are balanced, we adopt accuracy as the validation criterion.

Table S1: The means (and the standard errors) of extra criteria in Model (M3). The columns from Err 1 to Err 4 represent the classification error within each class, the column Err represents the overall classification error, and the column G-Mean represents the geometric mean. The results are based on 200 replicates.

Method	Err 1 (%)	Err 2 (%)	Err 3 (%)	Err 4 (%)	G-Mean (%)
Bayes	25.9(0.4)	41.7(0.5)	6.3(0.1)	0.7(0.0)	79.4(0.2)
LSLDA	**29.7(0.6)**	**51.2(0.8)**	**8.7(0.4)**	**1.1(0.1)**	**74.0(0.4)**
PP	100.0(0.0)	100.0(0.0)	59.4(1.1)	0.5(0.0)	0.0(0.0)
SPCALDA	98.4(0.5)	99.9(0.1)	9.8(0.9)	10.7(0.9)	0.8(0.3)
MSDA	**39.8(0.8)**	**67.4(0.8)**	8.9(0.2)	1.6(0.1)	**63.8(0.5)**
SOS($q = K - 1$)	65.5(0.8)	91.0(0.4)	14.0(0.3)	1.6(0.1)	38.2(0.6)
SOS($q = d$)	51.1(1.1)	88.0(0.5)	16.1(0.5)	1.4(0.1)	43.6(0.7)
PLDA($q = K - 1$)	95.1(0.8)	99.3(0.4)	**2.0(0.4)**	1.3(0.1)	1.2(0.6)
PLDA($q = d$)	66.6(1.9)	87.2(1.8)	14.5(1.7)	1.7(0.1)	8.3(1.1)
Logistic	96.7(0.4)	98.7(0.2)	72.8(3.0)	**0.4(0.1)**	6.2(0.8)

S2 Additional numerical results

S2.1 Results in Models (M3) & (M7) and rank selection results

The extra results in Model (M3) is reported in Table S1. The columns from Err 1 to Err 4 represent the classification error within each class. It can be seen that in classes 2, 3 and 4, our proposal outperforms the competitors,

Table S2: The means (and the standard errors) of the KL-divergence, the subspace distance D between $\mathcal{S}_\beta$ and $\mathcal{S}_{\widehat{\beta}}$, the TPR (%) and the FPR (%) on simulated data generated from Model (M7). The results are based on 200 replicates. The standard errors for TPR and FPR are all less than 1.5%, and are thus omitted.

Method	KL-divergence	D	TPR(%)	FPR(%)
LSLDA	**0.100(0.007)**	0.400(0.999)	99.1	1.6
PP	1.929(0.017)	1.540(0.021)	100.0	100.0
SPCALDA	0.946(0.003)	1.652(4.844)	100.0	100.0
MSDA	**0.443(0.012)**	1.278(0.296)	39.6	0.0
SOS($q = K - 1$)	0.819(0.014)	1.319(0.317)	40.5	0.5
SOS($q = d$)	0.449(0.016)	0.816(0.681)	50.2	0.1
Logistic	0.760(0.006)	1.314(0.296)	48.6	1.2

and in class 1, the classification of our proposal is comparable to the best result. Specially, LSLDA performs well on the most difficult classes 1 and 2, while all the competitors make poor predictions in either class 1 or class 2, which results in the overall inaccurate prediction on the unbalanced data set.

The comparison results under Model (M7) is reported in Table S2. Since the classes 2, 3 and 4 are indistinguishable in Model (M7), classification error is unable to measure the performance of classification accurately.

Table S3: The true discriminant ranks d in Models (M1)–(M7) are in the first row. The means (and the standard errors) of the estimated ranks $\widehat{d}$ from LSLDA and SPCALDA on simulated data in each model are in the last two rows. The results are based on 200 replicates.

Model	(M1)	(M2)	(M3)	(M4)	(M5)	(M6)	(M7)
d	2	2	2	2	5	1	1
LSLDA	2.01(0.01)	2.02(0.01)	2.40(0.05)	2.31(0.06)	5.00(0.02)	1.12(0.03)	1.05(0.02)
SPCALDA	6.39(0.42)	7.18(0.40)	1.11(0.07)	11.72(0.40)	10.81(0.36)	10.46(0.38)	5.39(0.39)

Instead, we adopt the Kullback-Leibler (KL) divergence defined as

$$\mathrm{KL}(\widehat{\pi}, \pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \log \left\{ \frac{\widehat{\pi}_k(x_i)}{\pi_k(x_i)} \right\} \widehat{\pi}_k(x_i),$$

where $\widehat{\pi}_k(x_i)$ and $\pi_k(x_i)$ are the estimated and true posterior probability $\Pr(Y = k \mid X = x_i)$. From Table S2, our proposal dominates the competitors in all comparison criteria, which confirms the effectiveness of LSLDA in the response category combination problems.

In Table S3, we report the estimated ranks from LSLDA and SPCALDA under Models (M1)–(M7). It can be seen that LSLDA selects the rank consistently, while SPCALDA always overestimates the rank by a lot, which partially explains why SPCALDA performs poorly in classification under Models (M1)–(M7).

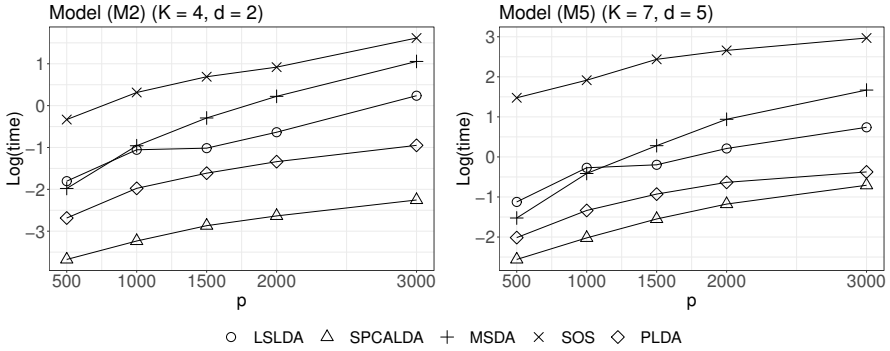

Figure S1: The mean of log computation time (seconds) under models (M2) and (M5). On average, when $p = 3000$, LSLDA, SPCALDA, MSDA, SOS, and PLDA finish one run in (1.267, 0.105, 2.874, 5.029, 0.387) seconds in Model (M2) and in (2.094, 0.493, 5.308, 19.458, 0.688) seconds in Model (M5).

S2.2 Computation time

To gain some insights into the computation efficiency of the LDA-based methods, we record the computation time for each method (LSLDA, SPCALDA, MSDA, SOS, and PLDA) in Models (M2) and (M5) as a demonstration. The true rank is given for SOS and PLDA. For a fair comparison, we record the computation time on the selected tuning parameter for each method.

The simulation procedures in two models are similar, thus we take Model (M2) as an example. In each data replicate, for each class k, where $k = 1, \ldots, 4$, we generate $n_k = 30$ observations X_i, $i = 1, \ldots, n_k$, from

model (M2). For some p, we take the first p variables from each X_i and generate new observations set $\{(X_i^p, Y_i)\}_{i=1}^n$, where X_i^p is the subvector containing the first p variables of X_i. We vary p over $\{500, 1000, 1500, 2000, 3000\}$ and implement each method on each set $\{(X_i^p, Y_i)\}_{i=1}^n$. For each of five methods, we use the same tuning parameter sequences as p varies. For each p, we record the averaged (logarithmic) computation time over 16 replicates for five methods.

The results are displayed in Figure S1. Each line describes the relationship between the (logarithmic) computation time and the dimension p. It can be seen that SPCALDA is the fastest one among the five competitors since the algorithm is basically the eigenvalue decomposition. PLDA is faster than LSLDA while MSDA and SOS are slower than LSLDA as p increases. On average, in either Model (M2) or (M5) and for any p between 500 and 3000, SPCALDA and PLDA finish one run in 0.5 and 0.7 seconds, LSLDA finishes one run in 2 seconds, and SOS and MSDA finish one run in 20 and 5 seconds. The result in Figure S1 also shows the scalability of LSLDA in the high-dimensional settings. Recall that our LSLDA is a low-rank extension of MSDA method. As p gets higher, the computation time of LSLDA increases slower than MSDA. Specifically, in Model (M2), from $p = 500$ to $p = 3000$, the logarithmic computation time of LSLDA

and MSDA increases by 2.042 and 3.035 respectively. And in Model (M5), from $p = 500$ to $p = 3000$, the logarithmic computation time of LSLDA and MSDA increases by 1.863 and 3.194 respectively.

S2.3 Scatterplots on real data

In Section 6, we display the two-dimensional scatterplots of selected classes in data set **grimace**. The colored scatterplots of all data points in three data sets **face94**, **face95** and **grimace** are presented in Fig. S2. Both methods conduct perfect classes separation on **face94** and **grimace**. While on **face95**, the separation among some classes is not clear.

S2.4 Simultaions with $p = 500$

In this section, we conduct another set of simulations with smaller dimension $p = 500$.

We consider seven Models (M1′)–(M7′) which are similar to Models (M1)–(M7) in Section 5. We set the sparsity level $s = 10$ and the dimension $p = 500$. Similar to Model (M3), we generate $10, 10, 50$ and 50 samples for each class separately as the training set for Model (M3′). And for the rest of models, we generate $n_k = 30$ for each class. Let n denote the sample size of the training set. For all models, we generate a separate validation

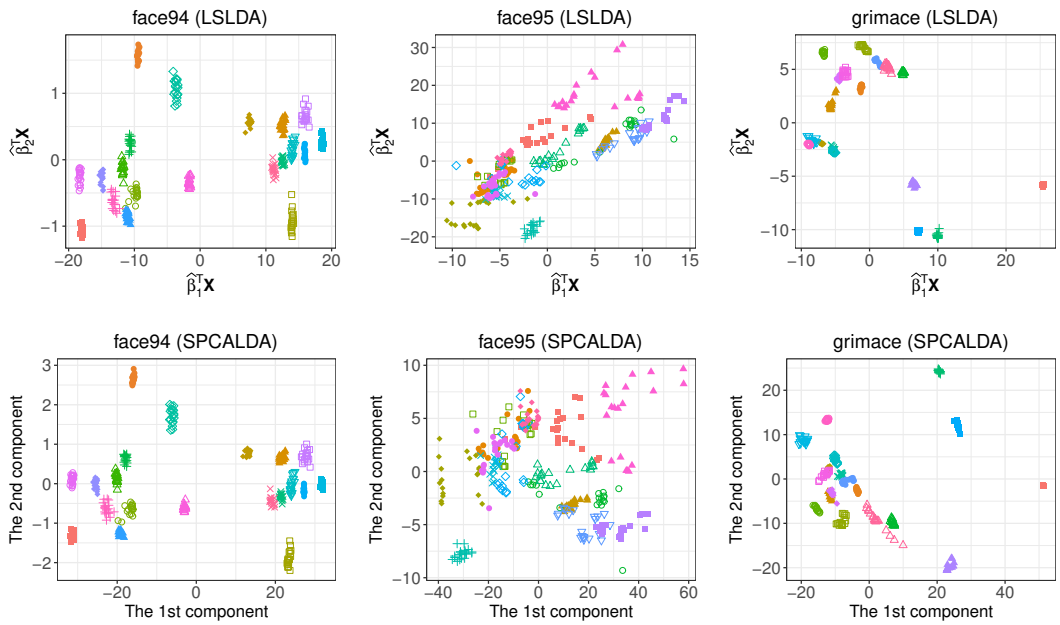

Figure S2: The scatterplots of the first two components from LSLDA (top) and SPCALDA (bottom) on data sets **face94**, **face95** and **grimace**. Each type of points represents a class. The estimated ranks from LSLDA (versus SPCALDA) on three data sets are 4 (versus 4), 6 (versus 18), and 12 (versus 5) respectively.

set of size n to select tuning parameter and a test set of size $5n$ to evaluate the model. For each class k, we generate X from the normal distribution $N(\mu_k, \Sigma)$. Since the discriminant subspace $\mathcal{S} = \Sigma^{-1}\text{span}(\mu_2 - \mu_1, \ldots, \mu_K - \mu_1)$, by fixing $\mu_1 = 0$ and denoting $\theta_k = \Sigma^{-1}\mu_{k+1}$ for $k = 1, \ldots, K-1$, the discriminant basis $\beta \in \mathbb{R}^{p\times d}$ is generated as the top-d left singular vectors of $\theta = (\theta_1, \ldots, \theta_{K-1}) \in \mathbb{R}^{p\times(K-1)}$. The details of each model are listed in the following:

(M1′) (Mild correlation) $K = 4$, θ_{1i} takes the value 0.8 for $i = 1, \ldots, 5$ and 0 otherwise, θ_{2i} takes value 0.8 for $i = 6, \ldots, 10$ and 0 otherwise, and $\theta_3 = \theta_1 + \theta_2$. The covariance matrix Σ has the AR$(0.5, 500)$ structure. The discriminant rank $d = 2$.

(M2′) (Strong correlation) $K = 4$, θ_{1i} takes the value 0.8 for $i = 1, \ldots, 5$ and 0 otherwise, θ_{2i} takes value 0.8 for $i = 6, \ldots, 10$ and 0 otherwise, and $\theta_3 = 1.5\theta_1 + 1.5\theta_2$. The covariance matrix $\Sigma = I_{10} \otimes \Omega$, where Ω has the CS$(0.3, 50)$ structure. The discriminant rank $d = 2$.

(M3′) (Unbalanced data) $K = 4$, θ_{1i} takes the value 0.8 for $i = 1, \ldots, 5$ and 0 otherwise, θ_{2i} takes value 0.8 for $i = 6, \ldots, 10$ and 0 otherwise, and $\theta_3 = 1.5\theta_1 + 1.5\theta_2$. The covariance matrix $\Sigma = I_{10} \otimes \Omega$, where Ω has the CS$(0.3, 50)$ structure. The discriminant rank $d = 2$.

(M4′) (Large K) $K = 7$, $\theta_{1,2i-1} = 2$ and $\theta_{2,2i} = -4$ for $i = 1, \ldots, 5$. For $k = 3, \ldots, K-1$, $\theta_k = (k/2 - 1)(\theta_1 + \theta_2)$. The covariance matrix Σ has the AR$(0.5, 500)$ structure. The discriminant rank $d = 2$.

(M5′) (Near full-rank basis) $K = 7$, θ_{ki} takes the value 2 for $i = 2k - 1, 2k$ and $k = 1, \ldots, 5$, and 0 otherwise, and $\theta_6 = 0.5 \sum_{k=1}^{5} \theta_k$. The covariance matrix Σ has the AR$(0.5, 500)$ structure. The discriminant rank $d = 5$.

(M6′) (Unimodality) $K = 4$, $\theta_2 = 2\theta_1$, $\theta_3 = 3\theta_1$, where θ_{1i} takes the value 1 for $i = 1, \ldots, 5$, the value -1 for $i = 6, \ldots, 10$, and 0 otherwise. The covariance matrix $\Sigma = I_{10} \otimes \Omega$, where Ω has the CS$(0.3, 50)$ structure. The discriminant rank $d = 1$.

(M7′) (Indistinguishable classes) $K = 4$, $\theta_1 = \theta_2 = \theta_3$, where θ_{1i} takes the value 1 for $i = 1, \ldots, 5$, the value -1 for $i = 6, \ldots, 10$, and 0 otherwise. The covariance matrix $\Sigma = I_{10} \otimes \Omega$, where Ω has the CS$(0.3, 50)$ structure. The discriminant rank $d = 1$.

We compare LSLDA with SPCALDA, MSDA, SOS, PLDA and Logistic using different criteria, including the classification error, the subspace estimation error, the TPR, and the FPR. We also consider SOS and PLDA with both the full rank and the true rank. Recall that from Table 1, PP performed poorly. It is thus left out from the comparison here. The comparison criteria over 200 replicates in Models (M1′)–(M6′) are reported in Table S4, and those in Model (M7′) are presented in Table S5. Also, we report the estimated ranks from LSLDA and SPCALDA in Table S6.

From Tables S4 and S5, it can be seen that in all the models (M1′)–(M7′), LSLDA dominates all the other competitors in terms of the classification, the subspace estimation and the variable selection. And from Table S6, we can see that in all model settings, our proposed method se-

Table S4: The means (and the standard errors) of the classification error (%), the subspace distance D, the TPR (%) and the FPR (%) on simulated data generated from Models (M1′)–(M6′). The results are based on 200 replicates. The standard errors for TPR and FPR are all less than 5%, and are thus omitted.

Method	Err(%)	D	TPR(%)	FPR(%)	Err(%)	D	TPR(%)	FPR(%)
		Model (M1′)				Model (M2′)		
Bayes	17.5	–	–	–	14.2	–	–	–
LSLDA	**19.2(0.1)**	0.365(0.008)	100.0	3.4	**16.1(0.1)**	0.357(0.007)	99.8	1.3
SPCALDA	31.3(0.2)	1.129(0.039)	100.0	100.0	32.4(0.1)	1.439(0.034)	100.0	100.0
MSDA	22.3(0.2)	0.791(0.005)	81.2	1.0	**19.9(0.2)**	0.813(0.005)	78.0	1.2
SOS($q = K-1$)	22.6(0.2)	0.641(0.005)	97.7	2.4	33.5(0.2)	0.974(0.003)	71.1	4.8
SOS($q = d$)	**19.5(0.1)**	0.416(0.007)	97.8	1.6	32.3(0.2)	0.834(0.003)	70.8	2.8
PLDA($q = K-1$)	19.6(0.2)	0.329(0.008)	99.7	1.1	31.1(0.1)	0.922(0.010)	100.0	64.6
PLDA($q = d$)	19.5(0.2)	0.327(0.007)	99.7	1.1	31.4(0.1)	0.845(0.006)	100.0	62.8
Logistic	21.5(0.1)	0.780(0.005)	84.0	1.2	22.7(0.2)	0.848(0.004)	78.0	1.7
		Model (M3′)				Model (M4′)		
Bayes	8.6	–	–	–	3.3	–	–	–
LSLDA	**11.0(0.2)**	0.429(1.318)	99.4	4.1	**4.7(0.1)**	0.407(0.005)	100.0	0.3
SPCALDA	17.3(0.2)	1.335(2.848)	100.0	100.0	13.2(0.1)	0.829(0.030)	100.0	100.0
MSDA	**13.5(0.2)**	0.865(0.512)	72.5	2.3	8.5(0.1)	1.176(0.002)	73.0	2.9
SOS($q = K-1$)	19.3(0.1)	0.977(0.239)	69.3	5.1	11.0(0.1)	1.191(0.001)	66.0	7.3
SOS($q = d$)	18.2(0.1)	0.842(0.313)	71.5	3.4	**7.7(0.1)**	0.645(0.001)	71.8	2.9
PLDA($q = K-1$)	21.4(0.3)	0.858(1.046)	100.0	43.5	9.8(0.1)	0.467(0.003)	100.0	0.3
PLDA($q = d$)	22.8(0.3)	0.836(0.640)	100.0	55.9	10.0(0.1)	0.484(0.004)	100.0	16.7
Logistic	27.2(1.2)	0.973(0.274)	58.4	7.1	25.2(0.1)	1.218(0.002)	77.7	2.2
		Model (M5′)				Model (M6′)		
Bayes	10.0	–	–	–	14.0	–	–	–
LSLDA	**11.7(0.1)**	0.262(0.005)	100.0	0.9	**15.3(0.1)**	0.231(0.015)	100.0	2.4
SPCALDA	34.2(0.1)	0.917(0.017)	100.0	100.0	37.0(0.2)	2.563(0.021)	100.0	100.0
MSDA	12.6(0.1)	0.464(0.004)	98.4	0.6	18.5(0.2)	1.060(0.002)	99.0	1.3
SOS($q = K-1$)	14.4(0.1)	0.520(0.004)	100.0	5.7	21.7(0.2)	1.025(0.001)	100.0	5.4
SOS($q = d$)	13.3(0.1)	0.415(0.005)	100.0	5.1	**15.6(0.1)**	0.235(0.006)	99.7	0.9
PLDA($q = K-1$)	19.9(0.2)	0.422(0.004)	99.3	0.2	16.8(0.2)	0.253(0.005)	100.0	0.0
PLDA($q = d$)	19.9(0.3)	0.423(0.005)	99.3	0.7	16.8(0.2)	0.253(0.005)	100.0	0.0
Logistic	**12.1(0.1)**	0.439(0.003)	99.9	1.4	31.7(0.2)	1.104(0.003)	95.3	2.3

lects the rank accurately while SPCALDA overestimates the rank.

Table S5: The means (and the standard errors) of the KL-divergence, the subspace distance D between $\mathcal{S}_\beta$ and $\mathcal{S}_{\widehat{\beta}}$, the TPR (%) and the FPR (%) on simulated data generated from Model (M7′). The results are based on 200 replicates. The standard errors for TPR and FPR are all less than 5%, and are thus omitted.

Method	KL-divergence	D	TPR(%)	FPR(%)
LSLDA	**0.055(0.002)**	0.348(0.012)	100.0	4.3
SPCALDA	0.944(0.003)	1.258(0.036)	100.0	100.0
MSDA	0.301(0.009)	1.218(0.004)	63.1	0.4
SOS($q = K - 1$)	0.548(0.010)	1.250(0.004)	60.7	3.0
SOS($q = d$)	**0.219(0.008)**	0.642(0.008)	80.0	1.3
Logistic	0.583(0.006)	1.227(0.003)	74.2	6.0

Table S6: The true discriminant ranks d in Models (M1′)–(M7′) are in the first row. The means (and the standard errors) of the estimated ranks $\widehat{d}$ from LSLDA and SPCALDA on simulated data in each model are in the last two rows. The results are based on 200 replicates.

Model	(M1′)	(M2′)	(M3′)	(M4′)	(M5′)	(M6′)	(M7′)
d	2	2	2	2	5	1	1
LSLDA	2.12(0.02)	2.00(0.01)	2.01(0.02)	2.00(0.00)	5.17(0.03)	1.12(0.03)	1.10(0.02)
SPCALDA	6.15(0.43)	7.83(0.42)	6.36(0.32)	4.08(0.29)	8.14(0.36)	13.62(0.22)	2.71(0.29)

S3 Alternative algorithm and comparisons

An alternative way to solve the doubly penalized optimization problem (2.8) is the alternating direction method of multipliers algorithm (ADMM; Boyd, Parikh and Chu 2011). We provide the details of our implementation of the ADMM algorithm for the problem (2.8), and then the numerical experiments comparing our implementations of the two algorithms are presented.

Recall that our optimization problem (2.8) is in form of

$$\widehat{B} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\mathrm{tr}(B^\top \widehat{\Sigma} B) - \mathrm{tr}(B^\top \widehat{U}) + \lambda_1 ||B||_{2,1} + \lambda_2 ||B||_\star.$$

By introducing an extra variable $C \in \mathbb{R}^{p\times K}$ and imposing the equality constraint that $B = C$, (2.8) is equivalent to

$$\widehat{B} = \underset{B=C\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\mathrm{tr}(B^\top \widehat{\Sigma} B) - \mathrm{tr}(B^\top \widehat{U}) + \lambda_1 ||B||_{2,1} + \lambda_2 ||C||_\star. \qquad \text{(S3.1)}$$

Following Boyd, Parikh and Chu (2011), the augmented Lagrangian for (S3.1) is

$$\begin{aligned} L_{\lambda_1,\lambda_2,\gamma}(B,C,\omega) =& \frac{1}{2}\mathrm{tr}(B^\top \widehat{\Sigma} B) - \mathrm{tr}(B^\top \widehat{U}) + \lambda_1 ||B||_{2,1} + \lambda_2 ||C||_\star + \mathrm{tr}\{\omega^\top (B-C)\} \\ &+ \frac{\gamma}{2}\|B-C\|_F^2, \end{aligned}$$

where $\omega \in \mathbb{R}^{p\times K}$ is the Lagrange multiplier and $\gamma > 0$ is the ADMM penalty term. We iteratively update $B^{(t)}$, $C^{(t)}$ and $\omega^{(t)}$ as follows for $t = 0, 1, \ldots,$ the sequence of $B^{(t)}$ converges to the global minimizer (Boyd, Parikh and

Chu 2011):

$$B^{(t)} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} L_{\lambda_1,\lambda_2,\gamma}(B, C^{(t-1)}, \omega^{(t-1)}),$$

$$C^{(t)} = \underset{C\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} L_{\lambda_1,\lambda_2,\gamma}(B^{(t)}, C, \omega^{(t-1)}),$$

$$\omega^{(t)} = \omega^{(t-1)} + \gamma(B^{(t)} - C^{(t)}).$$

By some algebra, the updates of $B^{(t)}$ and $C^{(t)}$ can be written as

$$B^{(t)} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\mathrm{tr}\{B^\top(\widehat{\Sigma} + \gamma I_p)B\} - \mathrm{tr}\{B^\top(\widehat{U} - \omega^{(t-1)} + \gamma C^{(t-1)})\} + \lambda_1||B||_{2,1}, \tag{S3.2}$$

$$C^{(t)} = \underset{C\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{\gamma}{2}\|C - (B^{(t)} + \gamma^{-1}\omega^{(t-1)})\|_F^2 + \lambda_2\|C\|_\star. \tag{S3.3}$$

There is no explicit solution for the update of $B^{(t)}$ in (S3.2). We implement a group-wise coordinate descent algorithm proposed by Mai, Yang and Zou (2019) to iteratively solve (S3.2). As a direct result from Corollary 1 in Zhou and Li (2014), the minimizer $C^{(t)}$ from (S3.3) is the truncated $B^{(t)} + \gamma^{-1}\omega^{(t-1)}$, where the singular values of $B^{(t)} + \gamma^{-1}\omega^{(t-1)}$ is soft-thresholded with λ_2/γ. In addition, since an extra tuning parameter γ is introduced, one has to put more effort in parameter tuning. In comparison, for the three-operator splitting algorithm, the major parts of the iteration are two proximal mapping problems, which have explicit solutions and thus can be efficiently solved. Also, there are only two tuning parameters in-

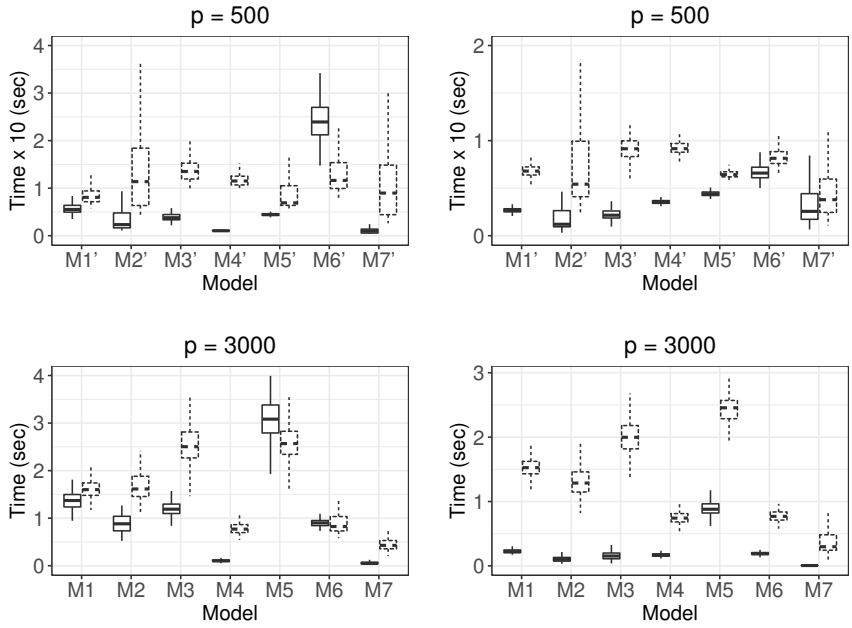

Figure S3: The boxplots of the computation time in Models (M1′)–(M7′) and Models (M1)–(M7): (left) the averaged time (in seconds) over 150 tuning parameters; (right) the averaged time (in seconds) over the tuning parameters resulting in sparse estimators. The solid and dotted boxes correspond to the three-operator splitting algorithm and the ADMM algorithm respectively. The results are based on 200 replicates.

volved, which makes the tuning parameter selection easier.

We compare the computation performances of our implementations of two algorithms under under Models (M1′)–(M7′) ($p = 500$) described in Section S2.4 and Models (M1)–(M7) ($p = 3000$) described in Section 5. In each model, we generate 200 data replicates. In each replicate, two algorithms select the optimal tuning parameters over 150 candidates. We record

the averaged execution time over 150 tuning parameter candidates in each replicate and draw the boxplots based on the results over 200 replicates in the left two plots in Fig. S3. Under all models except for Models (M5) and (M6′), our implementation of three-operator splitting algorithm runs faster than our implementation of ADMM algorithm. Since two algorithms get much slower with the tuning parameters that produce non-sparse estimators, we particularly focus on the tuning parameters that produce sparse estimators with approximately correct variable selection. Since the sparsity level $s = 10$ through Models (M1′)–(M7′) and (M1)–(M7), we define a sparse estimator $\widehat{\beta}$ as the one with $5 \leq \widehat{s} \leq 15$. The averaged execution time over the tuning parameters producing sparse estimators are recorded. The corresponding boxplots are displayed in the right two plots in Fig. S3. It can be seen that our implementation of three-operator splitting algorithm runs much faster than our implementation of ADMM algorithm. The discrepancy is more clear in Models (M1)–(M7) where $p = 3000$. Moreover, since only two tuning parameters are involved, the three-operator splitting algorithm is easier to tune.

S4 Proofs of lemmas in the paper

S4.1 Proof of Lemma 1

The reduced-rank LDA model is connected with the principal fitted component model (Cook and Forzani 2008) via the inverse regression formulation as discussed in Section 2.1:

$$X = \mu + \Sigma\beta\eta\xi_Y + \varepsilon, \quad \varepsilon \sim N(0, \Sigma), \tag{S4.1}$$

where $\xi_Y \in \mathbb{R}^K$ is the indicator functions of Y. By choosing the indicator functions ξ_Y as the fitting functions f_Y, the reduced-rank LDA model is exactly the principal fitted component model. Since $\sum_{k=1}^K \pi_k \beta\eta_k = 0$, there is an intrinsic constraint that $\eta \mathrm{E}(\xi_Y) = 0$ in (S4.1). Following the arguments in deriving the maximum likelihood estimator of central subspace in Section 3.1 in Cook and Forzani (2008), with the constraint that $\eta \mathrm{E}(\xi_Y) = 0$, we can easily obtain the similar partially maximized log-likelihood of (S4.1):

$$L(\Gamma, \Sigma) \simeq -\frac{n}{2}\log|\Sigma| - \frac{n}{2}\mathrm{tr}(\Sigma^{-1}\widehat{\Sigma}) - \frac{n}{2}\sum_{i=d+1}^{p}\lambda_i(\Sigma^{-1}\widehat{\Sigma}_b), \tag{S4.2}$$

where $\widehat{\Sigma} = (1/n)\sum_{k=1}^K \sum_{i=1}^n I(Y_i = k)(X_i - \bar{X}_k)(X_i - \bar{X}_k)^\top$ is the within-class covariance matrix, $\widehat{\Sigma}_b = \sum_{k=1}^K (n_k/n)(\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^\top$ is the between-class covariance matrix and $\lambda_i(\cdot)$ denotes the ith eigenvalue of a matrix. Then by Theorem 3.1 and Lemma A.1 in Cook and Forzani (2008),

the maximal likelihood estimator of $\mathcal{S} = \text{span}(\beta)$ is $\widehat{\Sigma}^{-1/2}\text{span}(\widehat{v}_1, \ldots, \widehat{v}_d)$, where $\widehat{v}_i$ is the ith eigenvector of $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}_b\widehat{\Sigma}^{-1/2}$.

S4.2 Proof for Lemma 2

Let $\widetilde{X}$ and ξ denote the matrices composed of the observations $X_i - \bar{X}$ and ξ_{Y_i}, $i = 1, \ldots, n$, row-wise respectively, the matrix form of the least squares objective function is as follows,

$$\begin{aligned}
&\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \|\widetilde{X} - \xi\widehat{W}B^\top\widehat{\Sigma}\|_F^2 \\
=&\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \text{tr}\{(\widetilde{X} - \xi\widehat{W}B^\top\widehat{\Sigma})^\top(\widetilde{X} - \xi\widehat{W}B^\top\widehat{\Sigma})\} \\
=&\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \text{tr}\{(\xi\widehat{W}B^\top\widehat{\Sigma})^\top(\xi\widehat{W}B^\top\widehat{\Sigma})\} - 2\text{tr}(\widetilde{X}^\top\xi\widehat{W}B^\top\widehat{\Sigma}).
\end{aligned}$$

Recall from Section 2.1 and Section 2.2 that $\widehat{U} = \{(n_1/n)^{1/2}(\bar{X}_1 - \bar{X}), \ldots, (n_K/n)^{1/2}(\bar{X}_K - \bar{X})\}$ and $\widehat{W} = \text{diag}\{(n_1/n)^{-1/2}, \ldots, (n_K/n)^{-1/2}\}$, one can easily show that $\widehat{U} = (1/n)\widetilde{X}^\top\xi\widehat{W}$. Also, we have $\widehat{W}^\top\xi^\top\xi\widehat{W} = nI_K$. Thus, it follows that

$$\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \|\widetilde{X} - \xi\widehat{W}B^\top\widehat{\Sigma}\|_F^2 = \underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \frac{1}{2}\text{tr}(B^\top\widehat{\Sigma}^2B) - \text{tr}(B^\top\widehat{\Sigma}\widehat{U}). \quad \text{(S4.3)}$$

Assume that $\widehat{\Sigma}$ is non-singular, then (S4.3) is equivalent to

$$\underset{B\in\mathbb{R}^{p\times K}}{\text{argmin}} \frac{1}{2}\text{tr}(B^\top\widehat{\Sigma}B) - \text{tr}(B^\top\widehat{U}).$$

S4.3 Proofs for Lemma 3 and Lemma 4

Lemma 3 is the direct result from Lemma 1 in Mai, Yang and Zou (2019) and Corollary 1 in Zhou and Li (2014). Lemma 4 is the direct result from Theorem 1.1 and Theorem 3.1 in Davis and Yin (2017).

S4.4 Proof for Lemma 5

Without loss of generality, we assume that $k' = 1$ and $\mu_2 - \mu_1 \neq 1$. The proof is finished by the contradiction. Suppose there exist at least one $k \in \{2, \ldots, K-1\}$ such that $\mu_{k+1} - \mu_k = \alpha_k(\mu_2 - \mu_1)$ for $\alpha_k \geq 0$ does not hold, and let k_0 denote the smallest such k. Then

$$\mu_{k+1} - \mu_k = \alpha_k(\mu_2 - \mu_1), \quad \alpha_k \geq 0, \quad \forall 1 \leq k < k_0.$$

We first consider the scenario where $\mu_{k_0+1} - \mu_{k_0} = \alpha_{k_0}(\mu_2 - \mu_1)$ with $\alpha_{k_0} < 0$. Since it is assumed that $X \mid (Y = k) \sim N(\mu_k, \Sigma)$,

$$\log \frac{\Pr(Y = k+1 \mid X)}{\Pr(Y = k \mid X)} = \log \frac{\pi_{k+1}}{\pi_k} + (\mu_{k+1} - \mu_k)^\top \Sigma^{-1}(X - \frac{\mu_{k+1} + \mu_k}{2})$$

for $k = 1, \ldots, K-1$. We prove that there exists a point x_0 such that $Y \mid (X = x_0)$ is not unimodal. We select the point $x_0 = -a(\mu_2 - \mu_1)$ with large enough constant $a > 0$ and. Let $\gamma = \Sigma^{-1}(\mu_2 - \mu_1)$. Then for $k < k_0$,

$$\log \frac{\Pr(Y = k+1 \mid X = x_0)}{\Pr(Y = k \mid X = x_0)} = \log \frac{\pi_{k+1}}{\pi_k} - \frac{\alpha_k \gamma^\top (\mu_{k+1} + \mu_k)}{2} - a\alpha_k \gamma^\top \Sigma \gamma. \tag{S4.4}$$

If $\alpha_{k''} > 0$ for some $k'' < k_0$ and a is large enough, then $\log\{\Pr(Y = k''+1 \mid X = x_0)/\Pr(Y = k'' \mid X = x_0)\} < 0$. If $\alpha_{k'''} = 0$ for some $k''' < k_0$, then (S4.4) reduces to

$$\log \frac{\Pr(Y = k''' + 1 \mid X = x_0)}{\Pr(Y = k''' \mid X = x_0)} = \log \frac{\pi_{k'''+1}}{\pi_{k'''}}.$$

In addition, $\mu_{k'''+1} = \mu_{k'''}$. Combined with the assumption that

$$\log \frac{\pi_{k'''+1}}{\pi_k'''} < \frac{(\mu_{k'''+1} - \mu_{k'''})^\top \Sigma^{-1} (\mu_{k'''+1} - \mu_{k'''})}{2} = 0,$$

we have $\log\{\Pr(Y = k''' + 1 \mid X = x_0)/\Pr(Y = k''' \mid X = x_0)\} < 0$. However, for $k = k_0$, since $\alpha_{k_0} < 0$ and $a > 0$ is large enough, we have $\log\{\Pr(Y = k_0 + 1 \mid X = x_0)/\Pr(Y = k_0 \mid X = x_0)\} > 0$. Thus,

$$\Pr(Y = 1 \mid X = x_0) > \ldots > \Pr(Y = k_0 \mid X = x_0), \text{ and}$$

$$\Pr(Y = k_0 \mid X = x_0) < \Pr(Y = k_0 + 1 \mid X = x_0),$$

which contradicts with the unimodality assumption.

Now we consider another scenario where $\mu_{k_0+1} - \mu_{k_0} \notin \text{span}(\mu_2 - \mu_1)$, or equivalently, $\Sigma^{-1}(\mu_{k_0+1} - \mu_{k_0}) \notin \mathcal{S}_\gamma$. We prove that there exists a point x_1 such that $Y \mid (X = x_1)$ is not unimodal. Since $\Sigma^{-1}(\mu_{k_0+1} - \mu_{k_0}) \notin \mathcal{S}_\gamma$, there must exist some $\eta \in \mathcal{S}_\gamma^\perp$ such that $(\mu_{k_0+1} - \mu_{k_0})^\top \Sigma^{-1} \eta > 0$. We select

the point $x_1 = \mu_{k_0-1} + a\eta$ for some large enough $a > 0$. Then

$$\begin{aligned}
&\log \frac{\Pr(Y = k_0 \mid X = x_1)}{\Pr(Y = k_0 - 1 \mid X = x_1)} \\
=& \log \frac{\pi_{k_0}}{\pi_{k_0-1}} + (\mu_{k_0} - \mu_{k_0-1})^\top \Sigma^{-1} \{\mu_{k_0-1} + a\eta - (\mu_{k_0} + \mu_{k_0-1})/2\} \\
=& \log \frac{\pi_{k_0}}{\pi_{k_0-1}} + (\mu_{k_0} - \mu_{k_0-1})^\top \Sigma^{-1} \{\mu_{k_0-1} - (\mu_{k_0} + \mu_{k_0-1})/2\} \qquad \text{(S4.5)} \\
=& \log \frac{\pi_{k_0}}{\pi_{k_0-1}} - \frac{(\mu_{k_0} - \mu_{k_0-1})^\top \Sigma^{-1} (\mu_{k_0} - \mu_{k_0-1})}{2},
\end{aligned}$$

where (S4.5) follows from the fact that $\Sigma^{-1}(\mu_{k_0} - \mu_{k_0-1}) \in \mathcal{S}_\gamma$ and $\eta \in \mathcal{S}_\gamma^\perp$. Combined with the assumption that

$$\log \frac{\pi_{k+1}}{\pi_k} < \frac{(\mu_{k+1} - \mu_k)^\top \Sigma^{-1} (\mu_{k+1} - \mu_k)}{2}, \quad k = 1, \ldots, K-1,$$

we have $\log\{\Pr(Y = k_0 \mid X = x_1)/\Pr(Y = k_0 - 1 \mid X = x_1)\} < 0$. On the other hand,

$$\begin{aligned}
&\log \frac{\Pr(Y = k_0 + 1 \mid X = x_1)}{\Pr(Y = k_0 \mid X = x_1)} \\
=& \log \frac{\pi_{k_0+1}}{\pi_{k_0}} + (\mu_{k_0+1} - \mu_{k_0})^\top \Sigma^{-1} (\mu_{k_0-1} - \frac{\mu_{k_0+1} + \mu_{k_0}}{2}) + a(\mu_{k_0+1} - \mu_{k_0})^\top \Sigma^{-1} \eta.
\end{aligned}$$

Since $a > 0$ is selected large enough and $(\mu_{k_0+1} - \mu_{k_0})^\top \Sigma^{-1} \eta > 0$, we have $\log\{\Pr(Y = k_0 + 1 \mid X = x_1)/\Pr(Y = k_0 \mid X = x_1)\} > 0$. Hence,

$$\begin{aligned}
&\Pr(Y = k_0 - 1 \mid X = x_1) > \Pr(Y = k_0 \mid X = x_1), \text{ and} \\
&\Pr(Y = k_0 \mid X = x_1) < \Pr(Y = k_0 + 1 \mid X = x_1),
\end{aligned}$$

which contradicts with the unimodality assumption. Therefore, we have

$\mu_{k+1} - \mu_k = \alpha_k(\mu_2 - \mu_1)$ for $k = 1, \ldots, K-1$, where constants $\alpha_k \geq 0$ and $\alpha_1 = 1$.

Since $\mu = \sum_{k=1}^{K} \pi_k \mu_k$, it can be easily shown that $\mathcal{S} = \Sigma^{-1}\text{span}\{\mu_1 - \mu, \ldots, \mu_K - \mu\} = \Sigma^{-1}\text{span}\{\mu_2 - \mu_1, \ldots, \mu_K - \mu_{K-1}\}$. Combined with the conclusion that $\mu_{k+1} - \mu_k = \alpha_k(\mu_2 - \mu_1)$ for $k = 1, \ldots, K-1$, then $\mathcal{S} = \Sigma^{-1}\text{span}\{\mu_2 - \mu_1\}$ and the discriminant rank $d = 1$.

S4.5 Proof of Lemma 6

Let $g_X(k)$ denote $\log\{\Pr(Y = k+1 \mid X)/\Pr(Y = k \mid X)\}$, since $\pi_1 = \ldots = \pi_K = 1/K$ and $\mu_k - \mu_{k-1} = \cdots = \mu_2 - \mu_1 \neq 0$, $g_X(k)$ is monotonically decreasing for all X by the fact that

$$g_X(k-1) - g_X(k) = (\mu_2 - \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1) > 0 \quad (k = 2, \ldots, K-1). \tag{S4.6}$$

The monotonically decreasing function $g_X(k)$ are of the same sign for all k, or $g_X(j) > 0$ for $j \leq k$ and $g_X(j) < 0$ for $j > k$ for some k, implying that $Y \mid (X = x)$ is unimodal distributed for any observation $x \in \mathbb{R}^p$.

S4.6 Proof of Lemma 7

Suppose that $\Pr(Y = j \mid X = x) = \Pr(Y = k \mid X = x)$ for some distinct $j, k \in \{1, \ldots, K\}$ and any $x \in \mathbb{R}^p$. Under the LDA model assumption that

$X \mid (Y = k) \sim N(\mu_k, \Sigma)$,

$$\log \frac{\Pr(Y = k \mid X = x)}{\Pr(Y = j \mid X = x)} = (x - \frac{\mu_k + \mu_j}{2})^\top \Sigma^{-1}(\mu_k - \mu_j) + \log \frac{\pi_k}{\pi_j} = 0, \quad \forall x \in \mathbb{R}^p,$$

we have $\Sigma^{-1}(\mu_k - \mu) = \Sigma^{-1}(\mu_j - \mu)$. Recall that $\mathcal{S} = \Sigma^{-1}\text{span}\{\mu_1 - \mu, \ldots, \mu_K - \mu\}$, since $\mu = \sum_{k=1}^K \pi_k \mu_k$, then $d < K - 1$.

S5 Auxiliary lemmas

Some auxiliary lemmas are presented in this section as the preparation for proving Theorem 1 and Corollary 1.

Lemma 1. *For any matrix $A \in \mathbb{R}^{p \times q}$,*

$$\|AA^\top\|_{1,1} \leq \|A\|_{2,1}^2,$$

$$\|A\|_{1,1} \leq \sqrt{q}\|A\|_{2,1},$$

where $\|A\|_{1,1} = \sum_{i=1}^p \sum_{j=1}^q |a_{ij}|$, $\|A\|_{2,1} = \sum_{i=1}^p (\sum_{j=1}^q a_{ij}^2)^{1/2}$.

of Lemma 1. For the first inequality, by Cauchy-Schwartz inequality, we have

$$\|AA^\top\|_{1,1} = \sum_{i=1}^p \sum_{j=1}^p |a_i^\top a_j| \leq \sum_{i=1}^p \sum_{j=1}^p \|a_i\|_2 \|a_j\|_2 = \|A\|_{2,1}^2,$$

where $a_i^\top$ is the ith row vector of matrix A.

For the second inequality, we have

$$\|A\|_{1,1} = \sum_{i=1}^p \|a_i\|_1 \leq \sum_{i=1}^p \sqrt{q}\|a_i\|_2 = \sqrt{q}|A\|_{2,1},$$

and the proof is completed. □

Lemma 2. *Let $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ denote the covariance matrix of random variable $X \in \mathbb{R}^p$, and let $\lambda_{\max}$ denote the maximal eigenvalue of Σ, then $|\sigma_{ij}| \le \lambda_{\max}$ for $i, j = 1, \ldots, p$.*

Lemma 3. *For $Y \in \{1, \ldots, K\}$ and $X \mid (Y = k) \sim N(\mu_k, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ and $\mu_k \in \mathbb{R}^p$, $k = 1, \ldots, K$, assume that*

(i) there exists constant $M > 0$ such that $M \ge \varphi_1(\Sigma) \ge \cdots \ge \varphi_p(\Sigma) \ge 1/M > 0$, where $\varphi_k(\Sigma)$ is the kth largest eigenvalue of Σ, and

(ii) there exists constant $Q > 0$ such that $1/Q \le (\mu_k - \mu_j)^\top \Sigma^{-1}(\mu_k - \mu_j) \le Q$ for $k \ne j$,

then $\max_k\{\|\mu_k\|_2\} \le C$ for some constant $C > 0$.

of Lemma 3. We prove that under (i), if $\max_k\{\|\mu_k\|_2\}$ is not bounded from above, then (ii) is incorrect. Since $\max_k\{\|\mu_k\|_2\}$ is not bounded from above, for any $Q > 0$, there exists $\mu_{k'}$ such that $\|\mu_{k'}\|_2 \ge \|\mu_1\|_2 + \sqrt{MQ}$. Meanwhile,

$$(\mu_{k'} - \mu_1)^\top \Sigma^{-1}(\mu_{k'} - \mu_1) \ge \varphi_1(\Sigma)^{-1}\|\mu_{k'} - \mu_1\|_2^2 \ge M^{-1}(|\|\mu_{k'}\|_2 - \|\mu_1\|_2|)^2 \ge Q.$$

Therefore, $\max_{k\ne j}\{(\mu_k - \mu_j)^\top \Sigma^{-1}(\mu_k - \mu_j)\}$ is not bounded from above. By contradiction, the proof is finished. □

Lemma 4. *Let $X \sim \chi^2_m$ be the chi-squared random variable with m degrees of freedom. For any $0 < t \leq 4m$,*

$$\Pr(|X - m| \geq t) \leq 2\exp\{-t^2/(16m)\}.$$

of Lemma 4. By Lemma 1 in Laurent and Massart (2000), for $X \sim \chi^2_m$ and any $x > 0$,

$$\Pr\{X - m \geq 2(mx)^{1/2} + 2x\} \leq \exp(-x), \quad \Pr\{X - m \leq -2(mx)^{1/2}\} \leq \exp(-x).$$

Thus, $\Pr\{|X - m| \geq 2(mx)^{1/2} + 2x\} \leq 2\exp(-x)$. For any $0 < x \leq m$, $2(mx)^{1/2} + 2x \leq 4(mx)^{1/2}$. Then

$$\Pr\{|X - m| \geq 4(mx)^{1/2}\} \leq \Pr\{|X - m| \geq 2(mx)^{1/2} + 2x\} \leq 2\exp(-x), \quad 0 < x \leq m.$$

The proof is finished by the substitution that $t = 4(mx)^{1/2}$. □

Lemma 5. *(Hoffman and Wielandt 1953) For $n \times n$ Hermitian matrices Z and $\widehat{Z}$, let $\lambda_i(\cdot)$ be the i-th eigenvalue of a matrix, then for $1 \leq p \leq \infty$ we have*

$$\sum_{i=1}^{n} \{\lambda_i(Z) - \lambda_i(\widehat{Z})\}^q \leq \|Z - \widehat{Z}\|^q_{S^q}, \tag{S5.1}$$

where $\|Z - \widehat{Z}\|_{S^q} = [\sum_{i=1}^{n} \{\lambda_i(Z - \widehat{Z})\}^q]^{1/q}$ is the Schatten q-norm of $Z - \widehat{Z}$.

Lemma 6. *For matrices $A, B \in \mathbb{R}^{m \times n}$ $(m \geq n)$, let $\sigma_1(A) \geq \ldots \geq \sigma_n(A) \geq 0$ and $\sigma_1(B) \geq \ldots \geq \sigma_n(B) \geq 0$ be the singular values of A and B respec-*

tively, we have

$$\sum_{i=1}^{n}\{\sigma_i(A)-\sigma_i(B)\}^2 \leq \|A-B\|_F^2. \tag{S5.2}$$

of Lemma 6. According to Hoffman-Wielandt inequality stated in Lemma 5, by taking $q=2$, the Schatten 2-norm becomes the Frobenius norm, and we obtain,

$$\sum_{i=1}^{n}\{\lambda_i(Z)-\lambda_i(\widehat{Z})\}^2 \leq \|Z-\widehat{Z}\|_F^2. \tag{S5.3}$$

Construct the symmetric matrix $\widetilde{A}\in\mathbb{R}^{(m+n)\times(m+n)}$ from A as

$$\widetilde{A}=\begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}, \tag{S5.4}$$

and it can be easily proven that $\widetilde{A}=\widetilde{A}^\top$. Similarly, we obtain the symmetric matrix $\widetilde{B}$ based on B. Let $A=UDV^\top$ denote the singular value decomposition of A for $U\in\mathbb{R}^{m\times m}$, $D\in\mathbb{R}^{m\times n}$ and $V\in\mathbb{R}^{n\times n}$. We split U into two parts in form of $U=(U_1,U_2)$, where $U_1\in\mathbb{R}^{m\times n}$ and $U_2\in\mathbb{R}^{m\times(m-n)}$. And $D=(\Lambda^\top,0^\top)^\top$, where $\Lambda=\mathrm{diag}\{\sigma_1(A),\ldots,\sigma_n(A)\}\in\mathbb{R}^{n\times n}$ and $0\in\mathbb{R}^{(m-n)\times n}$. Then we have the spectral decomposition of $\widetilde{A}$ as $\widetilde{A}=\widetilde{U}\widetilde{D}\widetilde{U}^\top$. Here,

$$\widetilde{U}=\begin{pmatrix} 2^{-1/2}U_1 & -2^{-1/2}U_1 & U_2 \\ 2^{-1/2}V & 2^{-1/2}V & 0_1 \end{pmatrix},\ \widetilde{D}=\mathrm{diag}(\Lambda,-\Lambda,0_2), \tag{S5.5}$$

where $0_1\in\mathbb{R}^{n\times(m-n)}$ and $0_2\in\mathbb{R}^{(m-n)\times(m-n)}$. Thus, the eigenvalues of $\widetilde{A}$

are $\{\sigma_1(A), \ldots, \sigma_n(A), -\sigma_1(A), \ldots, -\sigma_n(A), 0, \ldots, 0\}$. Similarly, we obtain the eigenvalues of $\widetilde{B}$ from the singular values of B.

Since $\sum_{i=1}^{n}\{\lambda_i(\widetilde{A})-\lambda_i(\widetilde{B})\}^2 = 2\sum_{i=1}^{n}\{\sigma_i(A)-\sigma_i(B)\}^2$ and $\|\widetilde{A}-\widetilde{B}\|_F^2 = 2\|A-B\|_F^2$, with (S5.3) we complete the proof. □

Lemma 7. *(Wedin 1972) Let $A, \widehat{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ have singular values $\sigma_1 \geq \ldots \geq \sigma_n$ and $\widehat{\sigma}_1 \geq \ldots \geq \widehat{\sigma}_n$ respectively. And let $U \in \mathbb{R}^{m\times s}$ and $\widehat{U} \in \mathbb{R}^{m\times s}$ denote the matrices composed of the top-s left singular vectors of A and $\widehat{A}$. Define $\sin\Theta(U,\widehat{U}) = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_d) \in \mathbb{R}^{s\times s}$, where $\cos\theta_i$ is the ith singular value of $U^\top\widehat{U}$, and define the eigen-gap as*

$$\delta = \inf\{|\sigma - \widehat{\sigma}| : \sigma \in (-\infty, \sigma_{s+1}], \widehat{\sigma} \in [\widehat{\sigma}_1, \widehat{\sigma}_s]\}. \tag{S5.6}$$

If $\delta > 0$, then

$$\|\sin\Theta(U,\widehat{U})\|_F \leq \frac{\|A-\widehat{A}\|_F}{\delta}. \tag{S5.7}$$

Lemma 8. *Let $X \sim N(\mu, \sigma^2)$, for any $t > 0$, we have*

$$\Pr(X-\mu \geq t) \leq \exp\{-t^2/(2\sigma^2)\}, \quad \Pr(X-\mu \leq -t) \leq \exp\{-t^2/(2\sigma^2)\}.$$

Lemma 9. *(Hoeffding's Lemma) Let X be a random variable such that $X \in [a,b]$ almost surely. Then,*

$$\Pr(|X - \mathrm{E}(X)| \geq \varepsilon) \leq 2\exp\left\{-2\varepsilon^2/(b-a)^2\right\}.$$

S6 Proof of Theorem 1

Recall that Σ is the common covariance matrix and the matrix $U = \{\pi_1^{1/2}(\mu_1 - \mu), \ldots, \pi_K^{1/2}(\mu_K - \mu)\}$. Their sample estimators are $\widehat{\Sigma} = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n} I(Y_i = k)(X_i - \bar{X}_k)(X_i - \bar{X}_k)^\top$ and $\widehat{U} = \{(n_1/n)^{1/2}(\bar{X}_1 - \bar{X}), \ldots, (n_K/n)^{1/2}(\bar{X}_K - \bar{X})\}$. Also, recall that the estimator $\widehat{B} \in \mathbb{R}^{p\times K}$ is the solution of the convex optimization problem as follows,

$$\widehat{B} = \underset{B\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}(B^\top \widehat{\Sigma} B) - \operatorname{tr}(B^\top \widehat{U}) + \lambda_1 \|B\|_{2,1} + \lambda_2 \|B\|_\star. \tag{S6.1}$$

Its population counterpart $B = \Sigma^{-1}U$ is the minimizer of the following function, the population version of the least square objective function in (2.7),

$$L(G) = \frac{1}{2}\operatorname{tr}(G^\top \Sigma G) - \operatorname{tr}(U^\top G). \tag{S6.2}$$

To prove Theorem 1, we need the following two events $H(\varepsilon)$ and $J(\varepsilon)$ for $\varepsilon > 0$:

$$H(\varepsilon) = \{\|\widehat{\Sigma} - \Sigma\|_{\max} \le \varepsilon, \|\widehat{U} - U\|_{\max} \le \varepsilon\},$$

$$J(\varepsilon) = \{\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty \le \varepsilon, |\widehat{\pi}_k - \pi_k| \le \varepsilon, k = 1, \ldots, K\},$$

where $\widehat{\pi}_k = n_k/n$ for $k = 1, \ldots, K$, $\|A\|_{\max} = \max_{i,j}|a_{ij}|$ for a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$ and $\|v\|_\infty = \max_j |v_j|$ for a vector $v \in \mathbb{R}^p$. Recall the three assumptions stated in Section 4, on which Theorem 1 relies on:

(A1) (Bounded eigenvalues) There exists some positive constant M such

that $M \geq \varphi_1(\Sigma) \geq \cdots \geq \varphi_p(\Sigma) \geq 1/M > 0$, where $\varphi_k(\Sigma)$ is the kth eigenvalue of Σ.

(A2) (Bounded prior) There exists constant $T > 0$ such that $T/K \geq \pi_k \geq 1/(TK)$ for all k.

(A3) (Bounded and separable classes) There exists constant $Q > 0$ such that $1/Q \leq (\mu_k - \mu_j)^\top \Sigma^{-1}(\mu_k - \mu_j) \leq Q$ for all $k \neq j$.

The proof of Theorem 1 is split into four parts. In Part I, we establish the concentration result for the estimator $\widehat{B}$. In Part II, we present the proof of rank consistency and the subspace estimation consistency. The consistency result of the classification error is proved in Part III. Parts I–III rely on the events $H(\varepsilon)$ and $J(\varepsilon)$. And in Part IV, we establish the high-probability inequalities for events $H(\varepsilon)$ and $J(\varepsilon)$.

Part I: The concentration result for the estimator $\widehat{B}$

The following lemma establishes the lower bound of the difference between $L(B)$ and $L(G)$.

Lemma 10. *For $B = \Sigma^{-1}U$ and any matrix $G \in \mathbb{R}^{p\times K}$, let $\varphi_{\min}$ denote the minimal eigenvalue of Σ, then*

$$L(G) - L(B) \geq \frac{1}{2}\varphi_{\min}\|G - B\|_F^2.$$

of Lemma 10. Since

$$L(B) = \frac{1}{2}\mathrm{tr}(U^\top \Sigma^{-1} U) - \mathrm{tr}(U^\top \Sigma^{-1} U) = -\frac{1}{2}\mathrm{tr}(B^\top \Sigma B).$$

It follows that

$$\begin{aligned} L(G) - L(B) =& \frac{1}{2}\mathrm{tr}(G^\top \Sigma G) + \frac{1}{2}\mathrm{tr}(B^\top \Sigma B) - \mathrm{tr}(U^\top G) \\ =& \frac{1}{2}\mathrm{tr}(G^\top \Sigma G) + \frac{1}{2}\mathrm{tr}(B^\top \Sigma B) - \mathrm{tr}(B^\top \Sigma G) \\ =& \frac{1}{2}\mathrm{tr}\{(G-B)^\top \Sigma (G-B)\}. \end{aligned} \tag{S6.3}$$

Since $\Sigma - \varphi_{\min} I_p$ is positive semi-definite, we have

$$\mathrm{tr}\{(G-B)^\top (\Sigma - \varphi_{\min} I_p)(G-B)\} \geq 0,$$

and it follows that

$$\mathrm{tr}\{(G-B)^\top \Sigma (G-B)\} \geq \varphi_{\min} \|G-B\|_F^2. \tag{S6.4}$$

The proof is then finished by combining (S6.3) and (S6.4). □

Let $\widehat{L}(G)$ denote the least squares objective function in (2.7),

$$\widehat{L}(G) = \frac{1}{2}\mathrm{tr}(G^\top \widehat{\Sigma} G) - \mathrm{tr}(\widehat{U}^\top G). \tag{S6.5}$$

The next lemma shows the upper bound of the discrepancy between $\widehat{L}(G)$ and $L(G)$ under event $H(\varepsilon)$.

Lemma 11. *Under event $H(\varepsilon)$, for any matrix $G \in \mathbb{R}^{p\times K}$ and the positive constant $\eta \geq \max\{\|G\|_{2,1}, 2K^{1/2}\}$, we have*

$$|\widehat{L}(G) - L(G)| \leq \varepsilon\eta^2.$$

of Lemma 11. Under event $H(\varepsilon)$, we have $\|\widehat{\Sigma}-\Sigma\|_{\max} \leq \varepsilon$, $\|\widehat{U}-U\|_{\max} \leq \varepsilon$. It follows that

$$\begin{aligned}|\widehat{L}(G) - L(G)| =& |\frac{1}{2}\mathrm{tr}\{G^\top(\widehat{\Sigma}-\Sigma)G\} - \mathrm{tr}\{(\widehat{U}-U)^\top G\}| \\ \leq& \frac{1}{2}|\mathrm{tr}\{(\widehat{\Sigma}-\Sigma)GG^\top\}| + |\mathrm{tr}\{(\widehat{U}-U)^\top G\}|.\end{aligned}$$

By definitions of $\|\cdot\|_{\max}$ and $\|\cdot\|_{1,1}$, we have $|\mathrm{tr}(M_1^\top M_2)| \leq \|M_1\|_{\max}\|M_2\|_{1,1}$ for any matrices M_1, $M_2 \in \mathbb{R}^{p\times q}$. Hence,

$$\begin{aligned}|\widehat{L}(G) - L(G)| \leq& \frac{1}{2}\|\widehat{\Sigma}-\Sigma\|_{\max}\|GG^\top\|_{1,1} + \|\widehat{U}-U\|_{\max}\|G\|_{1,1} \\ \leq& \frac{1}{2}\|\widehat{\Sigma}-\Sigma\|_{\max}\|G\|_{2,1}^2 + K^{1/2}\|\widehat{U}-U\|_{\max}\|G\|_{2,1} \\ \leq& \varepsilon\eta^2,\end{aligned}$$

where the second inequality follows from Lemma 1. □

Let $\widehat{L}_\lambda(G)$ denote our targeted penalized objective function in (2.8),

$$\widehat{L}_\lambda(G) = \widehat{L}(G) + \lambda_1\|G\|_{2,1} + \lambda_2\|G\|_*, \tag{S6.6}$$

whose global minimizer is $\widehat{B}$. We define the $L_{2,1}$ norm closed ball with radius r in $\mathbb{R}^{p\times K}$ as

$$\mathcal{G}_r := \{G \in \mathbb{R}^{p\times K} |\ \|G\|_{2,1} \leq r\}. \tag{S6.7}$$

Also recall the constant

$$\tau = \max\{\|B\|_{2,1} + \|B\|_\star, 2K^{1/2}\} \tag{S6.8}$$

defined in Section 4.

The following lemma shows that the global minimizer $\widehat{B}$ is bounded from above in $L_{2,1}$ norm.

Lemma 12. *Under event $H(\varepsilon)$, for any $N > 1$, $\lambda_1 > \varepsilon\tau(1+N^2)/(N-1)$, $0 < \lambda_2 \leq \lambda_1$ and the constant τ defined in* (S6.8)*, the global minimizer $\widehat{B} = \operatorname{argmin}_{B\in\mathbb{R}^{p\times K}} \widehat{L}_\lambda(B)$ is attainable in the interior of closed ball $\mathcal{G}_{N\tau}$, i.e.,*

$$\|\widehat{B}\|_{2,1} < N\tau. \tag{S6.9}$$

of Lemma 12. Since $\widehat{L}_\lambda(G)$ is convex, any local minimizer is also the global minimizer. Then it suffices to prove that there exists a local minimizer in the interior of the closed ball $\mathcal{G}_{N\tau}$. In fact, since $\widehat{L}_\lambda(G)$ is continuous and the closed ball $\mathcal{G}_{N\tau}$ is compact, the minimizer $\widehat{B}_{N\tau} = \operatorname{argmin}_{G\in\mathcal{G}_{N\tau}} \widehat{L}_\lambda(G)$ is always attainable. Thus, we just need to show that $\widehat{B}_{N\tau}$ does not reside on the boundary of the ball $\mathcal{G}_{N\tau}$. We prove it by contradiction.

Assume that $\widehat{B}_{N\tau}$ is on the boundary of the ball $\mathcal{G}_{N\tau}$, i.e. $\|\widehat{B}_{N\tau}\|_{2,1} = N\tau$. By the definition of τ, since $N > 1$, we have $\tau \geq \max\{\|B\|_{2,1}, 2K^{1/2}\}$ and $N\tau \geq \max\{\|\widehat{B}_{N\tau}\|_{2,1}, 2K^{1/2}\}$. Then according to Lemma 11, under

event $H(\varepsilon)$, we have $|\widehat{L}(B)-L(B)| \leq \varepsilon\tau^2$ and $|\widehat{L}(\widehat{B}_{N\tau})-L(\widehat{B}_{N\tau})| \leq \varepsilon N^2\tau^2$. It follows that

$$\begin{aligned}&\widehat{L}(\widehat{B}_{N\tau}) + \lambda_1\|\widehat{B}_{N\tau}\|_{2,1} + \lambda_2\|\widehat{B}_{N\tau}\|_* \geq \widehat{L}(\widehat{B}_{N\tau}) + \lambda_1 N\tau \\ \geq &L(\widehat{B}_{N\tau}) - \varepsilon N^2\tau^2 + \lambda_1 N\tau \geq L(B) - \varepsilon N^2\tau^2 + \lambda_1 N\tau, \end{aligned} \tag{S6.10}$$

where the last inequality follows from the optimality of B in the function $L(G)$. Since $\|B\|_{2,1} \in \mathcal{G}_{N\tau}$, then according to the optimality of $\widehat{B}_{N\tau}$, we have

$$\widehat{L}(\widehat{B}_{N\tau}) + \lambda_1\|\widehat{B}_{N\tau}\|_{2,1} + \lambda_2\|\widehat{B}_{N\tau}\|_* \leq \widehat{L}(B) + \lambda_1\|B\|_{2,1} + \lambda_2\|B\|_*. \tag{S6.11}$$

Meanwhile, since $|\widehat{L}(B)-L(B)| \leq \varepsilon\tau^2$, $0 < \lambda_2 \leq \lambda_1$ and $\|B\|_{2,1}+\|B\|_* \leq \tau$, we obtain

$$\widehat{L}(B) + \lambda_1\|B\|_{2,1} + \lambda_2\|B\|_* \leq L(B) + \varepsilon\tau^2 + \lambda_1\tau. \tag{S6.12}$$

Combining (S6.10)(S6.11)(S6.12), we have

$$L(B) - \varepsilon N^2\tau^2 + \lambda_1 N\tau \leq L(B) + \varepsilon\tau^2 + \lambda_1\tau. \tag{S6.13}$$

On the other hand, since we select $\lambda_1 > \varepsilon\tau(1+N^2)/(N-1)$, then

$$L(B) - \varepsilon N^2\tau^2 + \lambda_1 N\tau > L(B) + \varepsilon\tau^2 + \lambda_1\tau,$$

which contradicts with (S6.13). Therefore, $\widehat{B}_{N\tau}$ is in the interior of the closed ball $\mathcal{G}_{N\tau}$ and is thus the global minimizer $\widehat{B}$. □

The next lemma presents the concentration result of $\widehat{B}$ under mild conditions.

Lemma 13. *Under event $H(\varepsilon)$, for $\lambda_1 \geq \lambda_2 > 0$ and $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$, we have*

$$\|\widehat{B}\|_{2,1} < 2\tau \quad \textit{and} \quad \|\widehat{B} - B\|_F \leq \tau(22\varepsilon/\varphi_{\min})^{1/2}.$$

of Lemma 13. By Lemma 10, we have

$$\|\widehat{B} - B\|_F^2 \leq \frac{2}{\varphi_{\min}}\{L(\widehat{B}) - L(B)\} \tag{S6.14}$$

$$= \frac{2}{\varphi_{\min}}\left[\{L(\widehat{B}) - \widehat{L}(\widehat{B})\} + \{\widehat{L}(\widehat{B}) - \widehat{L}(B)\} + \{\widehat{L}(B) - L(B)\}\right]. \tag{S6.15}$$

According to Lemma 12, under event $H(\varepsilon)$, by taking $N = 2$, $\lambda_1 > 5\varepsilon\tau$ and $0 < \lambda_2 \leq \lambda_1$, the global minimizer $\|\widehat{B}\|_{2,1} < 2\tau$. To bound λ_1 at the order $O(\varepsilon\tau)$, we further take $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$. Since $\tau \geq \max\{\|B\|_{2,1}, 2K^{1/2}\}$ and $2\tau \geq \max\{\|\widehat{B}\|_{2,1}, 2K^{1/2}\}$, by Lemma 11,

$$L(\widehat{B}) - \widehat{L}(\widehat{B}) \leq 4\varepsilon\tau^2, \quad \widehat{L}(B) - L(B) \leq \varepsilon\tau^2. \tag{S6.16}$$

Based on the optimality of $\widehat{B}$, we have

$$\widehat{L}(\widehat{B}) + \lambda_1\|\widehat{B}\|_{2,1} + \lambda_2\|\widehat{B}\|_* \leq \widehat{L}(B) + \lambda_1\|B\|_{2,1} + \lambda_2\|B\|_*,$$

implying that

$$\widehat{L}(\widehat{B}) - \widehat{L}(B) \leq \lambda_1\|B\|_{2,1} + \lambda_2\|B\|_* \leq \lambda_1\tau \leq 6\varepsilon\tau^2. \tag{S6.17}$$

We complete the proof by combining (S6.15)(S6.16)(S6.17). □

Part II: The rank consistency and the subspace estimation consistency

Recall that the estimated rank $\widehat{d} = \sum_i I\{\sigma_i(\widehat{B}) \geq \delta\}$, where $\delta > 0$ is the thresholding value. Also, $\widehat{\beta}$ is composed of the top-$\widehat{d}$ left singular vectors of $\widehat{B}$ and the basis matrix β is referred to the matrix composed of the top-d left singular vectors of B. The following lemma establishes the consistency results for the estimated rank $\widehat{d}$ and the estimated subspace $\mathcal{S}_{\widehat{\beta}}$.

Lemma 14. *Under event $H(\varepsilon)$, where $0 < \varepsilon \leq \varphi_{\min}\sigma_{\min}^2/(198\tau^2)$, for $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$, $0 < \lambda_2 \leq \lambda_1$, $\tau(22\varepsilon/\varphi_{\min})^{1/2} < \delta \leq 2\tau(22\varepsilon/\varphi_{\min})^{1/2}$, we have $\widehat{d} = d$ and $D^2(\beta, \widehat{\beta}) \leq C\varepsilon\tau^2$ for some positive constant C.*

of Lemma 14. According to Lemma 6,

$$\sum_{i=1}^{\min\{p,K\}} \{\sigma_i(\widehat{B}) - \sigma_i(B)\}^2 \leq \|\widehat{B} - B\|_F^2.$$

Specially, $|\sigma_i(\widehat{B}) - \sigma_i(B)| \leq \|\widehat{B} - B\|_F$ for $i = 1, \ldots, \min\{p, K\}$. Thus, by Lemma 13,

$$|\sigma_i(\widehat{B}) - \sigma_i(B)| \leq \|\widehat{B} - B\|_F \leq \tau(22\varepsilon/\varphi_{\min})^{1/2}, \quad i = 1, \ldots, \min\{p, K\}. \tag{S6.18}$$

Recall that $\sigma_{\min} = \sigma_d(B)$ is the smallest non-zero singular value of B, we choose $0 < \varepsilon \leq \varphi_{\min}\sigma_{\min}^2/(198\tau^2)$. Then for $1 \leq i \leq d$ where $\sigma_i(B) \geq \sigma_{\min}$, we have $\sigma_i(\widehat{B}) \geq \sigma_{\min} - \tau(22\varepsilon/\varphi_{\min})^{1/2} \geq 2\tau(22\varepsilon/\varphi_{\min})^{1/2}$. And for $d < i \leq \min\{p, K\}$ where $\sigma_i(B) = 0$, $\sigma_i(\widehat{B}) \leq \tau(22\varepsilon/\varphi_{\min})^{1/2}$. Since we take $\tau(22\varepsilon/\varphi_{\min})^{1/2} < \delta \leq 2\tau(22\varepsilon/\varphi_{\min})^{1/2}$, then $\widehat{d} = d$.

Since $\widehat{d} = d$, then β and $\widehat{\beta}$ have equal ranks. We have known that $\sigma_i(\widehat{B}) \geq \sigma_{\min} - \tau(22\varepsilon/\varphi_{\min})^{1/2}$ for $1 \leq i \leq d$ and $\tau(22\varepsilon/\varphi_{\min})^{1/2} \leq (1/3)\sigma_{\min}$, then $\sigma_d(\widehat{B}) \geq (2/3)\sigma_{\min}$. Moreover, since $\sigma_{d+1}(B) = 0$, then the eigen-gap

$$|\sigma_d(\widehat{B}) - \sigma_{d+1}(B)| \geq (2/3)\sigma_{\min}.$$

By Wedin's sin θ theorem stated in Lemma 7, we have

$$\|\sin\Theta(\beta, \widehat{\beta})\|_F^2 \leq \frac{\|B - \widehat{B}\|_F^2}{|\sigma_d(\widehat{B}) - \sigma_{d+1}(B)|^2} \leq \frac{198\varepsilon\tau^2}{4\varphi_{\min}\sigma_{\min}^2},$$

where $\sin\Theta(\beta, \widehat{\beta}) = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_d)$ and $\cos\theta_i$ is the i-th singular value of matrix $\widehat{\beta}^\top\beta$. Moreover, since

$$\begin{aligned}
\|\sin\Theta(\beta, \widehat{\beta})\|_F^2 &= \sum_{i=1}^{d}(1 - \cos^2\theta_i) = d - \sum_{i=1}^{d}\cos^2\theta_i \\
&= \frac{1}{2}\mathrm{tr}\{\beta\beta^\top + \widehat{\beta}\widehat{\beta}^\top - 2\widehat{\beta}^\top\beta(\widehat{\beta}^\top\beta)^\top\} \\
&= \frac{1}{2}\|\beta\beta^\top - \widehat{\beta}\widehat{\beta}^\top\|_F^2 = dD^2(\beta, \widehat{\beta}),
\end{aligned}$$

then we have

$$D^2(\beta, \widehat{\beta}) = \|\sin\Theta(\beta, \widehat{\beta})\|_F^2/d \leq C\varepsilon\tau^2,$$

for some positive constant C, which finishes the proof. □

Part III: The classification error consistency

For the new observation $(X^\star, Y^\star)$, recall that the Bayes error $R = \Pr(\phi(X^\star) \neq Y^\star)$. The empirical classification error $R_n = \Pr(\widehat{\phi}(X^\star) \neq Y^\star \mid \widehat{\phi})$, the events $H(\varepsilon)$ and $J(\varepsilon)$ are defined conditioning on the training data. When there is no ambiguity, we omit the conditioning on training data in $R_n = \Pr(\widehat{\phi}(X^\star) \neq Y^\star)$. Under events $H(\varepsilon)$ and $J(\varepsilon)$, we establish the classification error consistency in the following lemma.

Lemma 15. *Under Assumptions (A1)(A2) and event $H(\varepsilon) \cap J(\varepsilon)$, where $\varepsilon \leq \min_k\{\pi_k/2\}$ and $\varepsilon\tau^2 \leq C$, for $\lambda_1 \geq \lambda_2 > 0$ and $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$, we have*

$$|R_n - R| \leq C(\varepsilon\tau^2)^{1/3}$$

for some constants $C, C' > 0$.

of Lemma 15. Recall that $B = \Sigma^{-1}\{\pi_1^{1/2}(\mu_1 - \mu), \ldots, \pi_k^{1/2}(\mu_K - \mu)\}$. Let B_k and $\widehat{B}_k$ denote the kth column of the matrices B and $\widehat{B}$, the Bayes rule can be written as $\phi(X^\star) = \text{argmax}_{k=1,\ldots,K}\, l_k(X^\star)$, where

$$l_k(X^\star) = \{X^\star - (\mu_k + \mu)/2\}^\top B_k \pi_k^{-1/2} + \log \pi_k, \tag{S6.19}$$

and the classification rule of our method is $\widehat{\phi}(X^\star) = \text{argmax}_{k=1,\ldots,K}\, \widehat{l}_k(X^\star)$, where

$$\widehat{l}_k(X^\star) = \{X^\star - (\bar{X}_k + \bar{X})/2\}^\top \widehat{B}_k \widehat{\pi}_k^{-1/2} + \log \widehat{\pi}_k. \tag{S6.20}$$

For notational convenience, let θ_k denote $B_k \pi_k^{-1/2} = \Sigma^{-1}(\mu_k - \mu)$, and $\widehat{\theta}_k$ denote $\widehat{B}_k \widehat{\pi}_k^{-1/2}$.

For any $\varepsilon_0 > 0$, we bound $|R_n - R|$ as follows, where the probability $\Pr(\cdot)$ is conditioning on the training data so that $\widehat{\phi}$ is non-stochastic.

$$\begin{aligned}
|R_n - R| =& |\Pr\{\widehat{\phi}(X^\star) \neq Y^\star, \phi(X^\star) \neq \widehat{\phi}(X^\star)\} - \Pr\{\phi(X^\star) \neq Y^\star, \phi(X^\star) \neq \widehat{\phi}(X^\star)\}| \\
\leq& \Pr\{\phi(X^\star) \neq \widehat{\phi}(X^\star)\} \\
\leq& 1 - \Pr(|l_k(X^\star) - l_{k'}(X^\star)| > \varepsilon_0, |\widehat{l}_k(X^\star) - l_k(X^\star)| < \varepsilon_0/2, k, k' = 1, \ldots, K) \\
\leq& \Pr\left(\bigcup_{k=1}^K \bigcup_{k'=1}^K \{|l_k(X^\star) - l_{k'}(X^\star)| \leq \varepsilon_0\}\right) + \Pr\left(\bigcup_{k=1}^K \{|\widehat{l}_k(X^\star) - l_k(X^\star)| \geq \varepsilon_0/2\}\right) \\
\equiv& L_8 + L_9.
\end{aligned} \tag{S6.21}$$

Since $\{l_k(X^\star) - l_{k'}(X^\star)\} \mid (Y^\star = k'') \sim N\{u(k, k', k''), (\theta_k - \theta_{k'})^\top \Sigma (\theta_k - \theta_{k'})\}$, where $u(k, k', k'') = \mu_{k''}^\top (\theta_k - \theta_{k'}) - (\mu_k + \mu)^\top \theta_k/2 + (\mu_{k'} + \mu)^\top \theta_{k'}/2 + \log(\pi_k/\pi_{k'})$, the probability $\Pr(|l_k(X^\star) - l_{k'}(X^\star)| \leq \varepsilon_0)$ is bounded as fol-

lows,

$$\begin{aligned}\Pr(|l_k(X^\star) - l_{k'}(X^\star)| \le \varepsilon_0) &= \mathrm{E}\{\Pr(|l_k(X^\star) - l_{k'}(X^\star)| \le \varepsilon_0 \mid Y^\star)\} \\ &\le \sum_{k''=1}^{K} \pi_{k''} \frac{2\varepsilon_0}{\{2\pi(\theta_k - \theta_{k'})^\top \Sigma(\theta_k - \theta_{k'})\}^{1/2}} \\ &\le C\varepsilon_0,\end{aligned}$$

where the first inequality is based on the fact that for a normal variable $Z \sim N(\nu, \sigma^2)$, its density function $f(z) \le (2\pi\sigma^2)^{-1/2}$, and the second inequality follows from Assumption (A3). Thus, it follows that

$$L_8 \le C\varepsilon_0. \tag{S6.22}$$

On the other hand, conditioning on the training data, $\{\widehat{l}_k(X^\star) - l_k(X^\star)\} \mid (Y^\star = k') \sim N\{w(k,k'), (\widehat{\theta}_k - \theta_k)^\top \Sigma(\widehat{\theta}_k - \theta_k)\}$, where $w(k,k') = \mu_{k'}^\top(\widehat{\theta}_k - \theta_k) - (\bar{X}_k + \bar{X})^\top \widehat{\theta}_k/2 + (\mu_k + \mu)^\top \theta_k/2 + \log(\widehat{\pi}_k/\pi_k)$. Since we are conditioning on the training data, $w(k,k')$ and $\widehat{\theta}_k$ are non-stochastic. By selecting $\varepsilon_0/2 > |w(k,k')|$ and using Chebyshev's inequality, we have

$$\begin{aligned}&\Pr(|\widehat{l}_k(X^\star) - l_k(X^\star)| \ge \varepsilon_0/2) \\ =&\mathrm{E}\{\Pr(|\widehat{l}_k(X^\star) - l_k(X^\star)| \ge \varepsilon_0/2 \mid Y^\star)\} \\ \le&\sum_{k'=1}^{K} \pi_{k'} \Pr\left(\left|\widehat{l}_k(X^\star) - l_k(X^\star) - w(k,k')\right| \ge \varepsilon_0/2 - |w(k,k')| \mid Y^\star = k'\right) \\ \le&\sum_{k'=1}^{K} \pi_{k'} \frac{(\widehat{\theta}_k - \theta_k)^\top \Sigma(\widehat{\theta}_k - \theta_k)}{(\varepsilon_0/2 - |w(k,k')|)^2}.\end{aligned} \tag{S6.23}$$

To further bound (S6.23), we need the upper bounds for $\|\widehat{\theta}_k - \theta_k\|_1$ and $\|\widehat{\theta}_k - \theta_k\|_2$. For $\|\widehat{\theta}_k - \theta_k\|_1$, we have

$$\begin{aligned}\|\widehat{\theta}_k - \theta_k\|_1 =& \|\widehat{B}_k\widehat{\pi}_k^{-1/2} - B_k\pi_k^{-1/2}\|_1 \\ \leq& \|(\widehat{B}_k - B_k)(\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2})\|_1 + \|(\widehat{B}_k - B_k)\pi_k^{-1/2}\|_1 + \|B_k(\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2})\|_1 \\ =& |\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2}|\|\widehat{B}_k - B_k\|_1 + \pi_k^{-1/2}\|\widehat{B}_k - B_k\|_1 + |\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2}|\|B_k\|_1.\end{aligned}$$

According to Lemma 13, under event $H(\varepsilon)$, for $\lambda_1 \geq \lambda_2 > 0$ and $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$, the global minimizer $\|\widehat{B}\|_{2,1} < 2\tau$. Since $\|\widehat{B}_k\|_1 \leq \|\widehat{B}\|_{2,1}$, we have $\|\widehat{B}_k\|_1 < 2\tau$. Similarly, $\|B_k\|_1 \leq \|B\|_{2,1} \leq \tau$. Thus, $\|\widehat{B}_k - B_k\|_1 < 3\tau$. Additionally, under event $J(\varepsilon)$ and Assumption (A2), since $\varepsilon \leq \min_k\{\pi_k/2\}$, we have

$$\begin{aligned}|\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2}| =& \frac{|\widehat{\pi}_k^{1/2} - \pi_k^{1/2}|}{(\widehat{\pi}_k\pi_k)^{1/2}} \leq \frac{|\widehat{\pi}_k - \pi_k|}{(\widehat{\pi}_k\pi_k)^{1/2}(\widehat{\pi}_k^{1/2} + \pi_k^{1/2})} \\ \leq& \frac{\varepsilon}{\{(\pi_k/2)\pi_k\}^{1/2}\{(\pi_k/2)^{1/2} + \pi_k^{1/2}\}} \leq C\varepsilon.\end{aligned}$$

Thus, it follows that

$$\|\widehat{\theta}_k - \theta_k\|_1 \leq C\varepsilon\tau + C\tau + C\varepsilon\tau \leq C\tau.$$

By Lemma 13, under event $H(\varepsilon)$, for $\lambda_1 \geq \lambda_2 > 0$ and $5\varepsilon\tau < \lambda_1 \leq 6\varepsilon\tau$, $\|\widehat{B} - B\|_F \leq \tau(22\varepsilon/\varphi_{\min})^{1/2}$. Since $\|\widehat{B}_k - B_k\|_2 \leq \|\widehat{B} - B\|_F$ and $\|B_k\|_2 \leq$

$\|B\|_F \leq \|B\|_\star \leq \tau$, then for $\|\widehat{\theta}_k - \theta_k\|_2$ we have

$$\begin{aligned}\|\widehat{\theta}_k - \theta_k\|_2 =& \|\widehat{B}_k\widehat{\pi}_k^{-1/2} - B_k\pi_k^{-1/2}\|_2 \\ =& |\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2}|\|\widehat{B}_k - B_k\|_2 + \pi_k^{-1/2}\|\widehat{B}_k - B_k\|_2 + |\widehat{\pi}_k^{-1/2} - \pi_k^{-1/2}|\|B_k\|_2 \\ \leq& C\varepsilon\tau(22\varepsilon/\varphi_{\min})^{1/2} + C\tau(22\varepsilon/\varphi_{\min})^{1/2} + C\varepsilon\tau \\ \leq& C\varepsilon^{1/2}\tau.\end{aligned}$$

Therefore, by Assumption (A1), we obtain

$$(\widehat{\theta}_k - \theta_k)^\top\Sigma(\widehat{\theta}_k - \theta_k) \leq \varphi_1(\Sigma)\|\widehat{\theta}_k - \theta_k\|_2^2 \leq C\varepsilon\tau^2, \tag{S6.24}$$

where $\varphi_1(\Sigma)$ is the largest eigenvalue of Σ. And for $w(k, k')$,

$$\begin{aligned}|w(k,k')| \leq& \|\mu_{k'}\|_2\|\widehat{\theta}_k - \theta_k\|_2 + (1/2)\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty\|\widehat{\theta}_k - \theta_k\|_1 \\ &+ (1/2)\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty\|\theta_k\|_1 + (1/2)\|\mu_k + \mu\|_2\|\widehat{\theta}_k - \theta_k\|_2 \\ &+ |\log\widehat{\pi}_k - \log\pi_k|.\end{aligned}$$

By Lemma 3, under Assumptions (A1) and (A3), $\max_k\{\|\mu_k\|_2\} \leq C$ for some constant $C > 0$. Since $\|B_k\|_1 \leq \|B\|_{2,1} \leq \tau$, then $\|\theta_k\|_1 = \pi_k^{-1/2}\|B_k\|_1 \leq C\tau$. And $\|\widehat{\theta}_k - \theta_k\|_\infty \leq \|\widehat{\theta}_k - \theta_k\|_2 \leq C\varepsilon^{1/2}\tau$. By mean value theorem, $|\log\widehat{\pi}_k - \log\pi_k| \leq \max\{\pi_k^{-1}, \widehat{\pi}_k^{-1}\}|\widehat{\pi}_k - \pi_k|$. Since under event $J(\varepsilon)$ and $\varepsilon \leq \min_k\{\pi_k/2\}$, we have $\widehat{\pi}_k \geq \pi_k/2$, then $|\log\widehat{\pi}_k - \log\pi_k| \leq C\varepsilon$. In addition, under event $J(\varepsilon)$, we have $\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty \leq \varepsilon$. Collecting

all the pieces, we obtain

$$|w(k,k')| \leq C\varepsilon^{1/2}\tau + C\varepsilon\tau + C\varepsilon\tau + C\varepsilon^{1/2}\tau + C\varepsilon \leq C_1\varepsilon^{1/2}\tau. \quad \text{(S6.25)}$$

Since $\varepsilon\tau^2 \leq C$ for some constant C, we select $\varepsilon_0 = C_2\varepsilon^{1/3}\tau^{2/3} \geq 4C_1\varepsilon^{1/2}\tau$ by selecting large enough constant $C_2 > 0$, where C_1 is from (S6.25). Then $\varepsilon_0/2 \geq |w(k,k')|$. Combining (S6.23)(S6.24)(S6.25), it follows that

$$\Pr(|\widehat{l}_k(X^\star) - l_k(X^\star)| \geq \varepsilon_0/2) \leq C(\varepsilon\tau^2)^{1/3}$$

and

$$L_9 \leq C(\varepsilon\tau^2)^{1/3}. \quad \text{(S6.26)}$$

Combining (S6.21)(S6.22)(S6.26), the proof is finished. □

Part IV: High-probability result

Lemma 16 and Lemma 17 to be presented in the following provide the high-probability results for events $H(\varepsilon)$ and $J(\varepsilon)$, based on which we have

$$\Pr(H(\varepsilon) \cap J(\varepsilon)) \geq 1 - Cp^2\exp(-C'n\varepsilon^2) \quad \text{(S6.27)}$$

for $0 < \varepsilon \leq C'$, where C and C' are some positive constants. In Lemma 14 and Lemma 15, since τ is allowed to diverge with n, the value of ε in Theorem 1 should satisfy $0 < \varepsilon \leq C\tau^{-2}$ for some constant $C > 0$. Thus, with Lemma 14 and Lemma 15, the proof of Theorem 1 is finished.

Now we derive the high-probability results for $H(\varepsilon)$ and $J(\varepsilon)$. For notational simplicity, we use C and C' to denote some generic positive constants that could vary from line to line. Since we assume that $\sigma_{\min}, \varphi_{\min}, d$ and K are fixed, these constants will be absorbed into constants C and C' when necessary.

Lemma 16. *Under Assumptions (A1)–(A3), for any $\varepsilon > 0$ and the event $H(\varepsilon)$ defined as*

$$H(\varepsilon) = \{\|\widehat{\Sigma} - \Sigma\|_{\max} \leq \varepsilon, \|\widehat{U} - U\|_{\max} \leq \varepsilon\},$$

there exist positive constants C and C' such that the event $H(\varepsilon)$ holds with probability at least $1 - Cp^2 \exp(-C'n\varepsilon^2)$ for $0 < \varepsilon \leq C'$.

of Lemma 16. For each entry $\widehat{\sigma}_{ij}$ and σ_{ij} in $\widehat{\Sigma}$ and Σ, the concentration inequality $\Pr(|\widehat{\sigma}_{ij} - \sigma_{ij}| \geq \varepsilon)$ directly follows from (S2.20) in Proposition 1 in Supplementary Materials of Mai, Yang and Zou (2019) by replacing the scaling factor $(n - K)^{-1}$ with n^{-1} in the definition of $\widehat{\sigma}_{ij}$ and σ_{ij}. Since K is assumed fixed in our proof, the results remain the same after the replacement of scaling factor. Thus, we have

$$\Pr(|\widehat{\sigma}_{ij} - \sigma_{ij}| \geq \varepsilon) \leq C \exp(-C'n\varepsilon^2), \quad 0 < \varepsilon \leq C'.$$

Then,

$$\Pr(\|\widehat{\Sigma} - \Sigma\|_{\max} \geq \varepsilon) \leq Cp^2 \exp(-C'n\varepsilon^2), \quad 0 < \varepsilon \leq C'.$$

To obtain the upper bound of $\Pr(\|\widehat{U} - U\|_{\max} \geq \varepsilon)$, we first derive the upper bound for the terms $\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon\}$ and $\Pr\{\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon\}$ respectively. The term $\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon\}$ is decomposed as

$$\begin{aligned}\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon\} =& \Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon, n_k \geq \pi_k n/2\} \\ &+ \Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon, n_k < \pi_k n/2\}.\end{aligned} \tag{S6.28}$$

The first term in (S6.28) is bounded as

$$\begin{aligned}&\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon, n_k \geq \pi_k n/2\} \\ =&\Pr\{|n_k/n - \pi_k| \geq \varepsilon |(n_k/n)^{1/2} + \pi_k^{1/2}|, n_k \geq \pi_k n/2\} \\ \leq&\Pr\{|n_k/n - \pi_k| \geq \varepsilon \pi_k^{1/2}(1 + 2^{-1/2})\} \\ \leq&2\exp(-Cn\varepsilon^2),\end{aligned}$$

where the last inequality follows from Lemma 9 and Assumption (A2). The second term in (S6.28) is bounded as

$$\begin{aligned}&\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon, n_k < \pi_k n/2\} \leq \Pr(n_k < \pi_k n/2) \\ =&\Pr(n_k - \pi_k n < -\pi_k n/2) \leq \exp(-Cn).\end{aligned}$$

Thus,

$$\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon\} \leq C\exp(-C'n\varepsilon^2). \tag{S6.29}$$

The term $\Pr\{\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon\}$ is bounded as follows,

$$\begin{aligned}
&\Pr\{\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon\} \\
\leq &\Pr[\|(1 - n_k/n)\bar{X}_k - (1 - \pi_k)\mu_k - \sum_{k' \neq k} \{(n_{k'}/n)\bar{X}_{k'} - \pi_{k'}\mu_{k'}\}\|_\infty \geq \varepsilon] \\
\leq &\Pr\{\|(1 - n_k/n)\bar{X}_k - (1 - \pi_k)\mu_k\|_\infty \geq \varepsilon/K\} + \sum_{k' \neq k} \Pr\{\|(n_{k'}/n)\bar{X}_{k'} - \pi_{k'}\mu_{k'}\|_\infty \geq \varepsilon/K\}
\end{aligned} \tag{S6.30}$$

We only derive the upper bound for the first term in (S6.30), while the derivation for the rest of terms is similar.

$$\begin{aligned}
&\Pr\{\|(1 - n_k/n)\bar{X}_k - (1 - \pi_k)\mu_k\|_\infty \geq \varepsilon/K\} \\
\leq &\Pr\{\|(1 - n_k/n)\bar{X}_k - (1 - n_k/n)\mu_k\|_\infty \geq \varepsilon/(2K)\} \\
&+ \Pr\{\|(1 - n_k/n)\mu_k - (1 - \pi_k)\mu_k\|_\infty \geq \varepsilon/(2K)\} \\
\equiv &L_6 + L_7
\end{aligned} \tag{S6.31}$$

Since $1 - n_k/n \leq 1$, then $L_6 \leq \Pr\{\|\bar{X}_k - \mu_k\|_\infty \geq \varepsilon/(2K)\}$. By (S2.19) in Proposition 1 in Supplementary Materials of Mai, Yang and Zou (2019), we directly have the concentration inequality for each entry $\bar{X}_{kj}$, $j = 1, \ldots, p$, as follows,

$$\Pr\{|\bar{X}_{kj} - \mu_{kj}| \geq \varepsilon/(2K)\} \leq C \exp(-C' n\varepsilon^2), \quad \varepsilon > 0.$$

Then,

$$L_6 \leq \Pr\{\|\bar{X}_k - \mu_k\|_\infty \geq \varepsilon/(2K)\} \leq Cp \exp(-C' n\varepsilon^2), \quad \varepsilon > 0. \tag{S6.32}$$

By Lemma 3, under Assumptions (A1) and (A3), $\max_k\{\|\mu_k\|_2\}$ is bounded from above, so is $\max_k\{\|\mu_k\|_\infty\}$. Then by Lemma 9,

$$L_7 \leq \Pr\{|n_k/n - \pi_k| \geq \varepsilon/(2CK)\} \leq C\exp(-C'n\varepsilon^2). \tag{S6.33}$$

Combining (S6.31)(S6.32)(S6.33), we obtain

$$\Pr\{\|(1-n_k/n)\bar{X}_k - (1-\pi_k)\mu_k\|_\infty \geq \varepsilon/K\} \leq Cp\exp(-C'n\varepsilon^2).$$

Thus, by (S6.30),

$$\Pr\{\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon\} \leq Cp\exp(-C'n\varepsilon^2). \tag{S6.34}$$

Now, we are ready to obtain the upper bound of $\Pr(\|\widehat{U} - U\|_{\max} \geq \varepsilon)$. Recall that $\widehat{U} = \{(n_1/n)^{1/2}(\bar{X}_1 - \bar{X}), \ldots, (n_K/n)^{1/2}(\bar{X}_K - \bar{X})\}$ and $U = \{\pi_1^{1/2}(\mu_1 - \mu), \ldots, \pi_K^{1/2}(\mu_K - \mu)\}$, then

$$\Pr(\|\widehat{U} - U\|_{\max} \geq \varepsilon) \leq \sum_{k=1}^{K} \Pr(\|(n_k/n)^{1/2}(\bar{X}_k - \bar{X}) - \pi_k^{1/2}(\mu_k - \mu)\|_\infty \geq \varepsilon)$$

Each term in the right-hand side is bounded as

$$\begin{aligned}
&\Pr(\|(n_k/n)^{1/2}(\bar{X}_k - \bar{X}) - \pi_k^{1/2}(\mu_k - \mu)\|_\infty \geq \varepsilon) \\
\leq &\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}|\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon/3)\} \\
&+ \Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}|\|\mu_k - \mu\|_\infty \geq \varepsilon/3\} \\
&+ \Pr\{\pi_k^{1/2}\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon/3\}
\end{aligned}$$

Under Assumptions (A1) and (A3), by Lemma 3, $\max_k\{\|\mu_k\|_2\}$ is bounded from above, so is $\max_k\{\|\mu_k\|_\infty\}$. In addition, by the upper bounds of $\Pr\{|(n_k/n)^{1/2} - \pi_k^{1/2}| \geq \varepsilon\}$ and $\Pr\{\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty \geq \varepsilon\}$ and Assumption (A2), it follows that

$$\Pr(\|\widehat{U} - U\|_{\max} \geq \varepsilon) \leq Cp\exp(-C'n\varepsilon^2). \tag{S6.35}$$

Hence, the proof is finished. □

Lemma 17. *Under Assumptions (A1)–(A3), for any $\varepsilon > 0$ and the event $J(\varepsilon)$ defined as*

$$J(\varepsilon) = \{\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty \leq \varepsilon, |n_k/n - \pi_k| \leq \varepsilon, k = 1, \ldots, K\},$$

there exist positive constants C and C' such that the event $J(\varepsilon)$ holds with probability at least $1 - Cp\exp(-C'n\varepsilon^2)$.

of Lemma 17. By Lemma 9, we have

$$\Pr(|n_k/n - \pi_k| > \varepsilon) \leq C\exp(-C'n\varepsilon^2). \tag{S6.36}$$

The derivation of $\Pr(\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty > \varepsilon)$ is similar to that of $\Pr(\|(\bar{X}_k - \bar{X}) - (\mu_k - \mu)\|_\infty > \varepsilon)$ given in the proof of Lemma 16. Similar to (S6.34), we have

$$\Pr\{\|(\bar{X}_k + \bar{X}) - (\mu_k + \mu)\|_\infty > \varepsilon\} \leq Cp\exp(-C'n\varepsilon^2). \tag{S6.37}$$

The proof is completed by combining (S6.36) and (S6.37). □

S7 Proof of Corollary 1

Take $\varepsilon = C_1(\log p/n)^{1/2}$ for large enough constant $C_1 > 0$ such that $C'C_1 > 2$. Since $\log p = o(n\tau^{-4})$, then $\varepsilon = o(\tau^{-2})$. According to Theorem 1, by plugging in $\varepsilon = C_1(\log p/n)^{1/2}$, if $5C_1\tau(\log p/n)^{1/2} < \lambda_1 \leq 6C_1\tau(\log p/n)^{1/2}$, $0 < \lambda_2 \leq \lambda_1$ and $C_2\tau(\log p/n)^{1/4} < \delta \leq 2C_2\tau(\log p/n)^{1/4}$ where $C_2 = (22C_1/\varphi_{\min})^{1/2}$, we have $\widehat{d} = d$, $D^2(\beta, \widehat{\beta}) \leq C'\varepsilon\tau^2$ and $|R_n - R| \leq C'(\varepsilon\tau^2)^{1/3}$ with probability at least $1 - C\exp\{-(C'C_1 - 2)\log p\}$.

Let $n, p \to \infty$, we have $\widehat{d} = d$, $D^2(\beta, \widehat{\beta}) \to 0$ and $|R_n - R| \to 0$ with probability tending to one, which finishes the proof.

References

Boyd, S., Parikh, N. and Chu, E. (2011), *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc.

Cook, R. D. and Forzani, L. (2008), 'Principal fitted components for dimension reduction in regression', *Statistical Science* **23**(4), 485–501.

Davis, D. and Yin, W. (2017), 'A three-operator splitting scheme and its optimization applications', *Set-valued and variational analysis* **25**(4), 829–858.

Hoffman, A. J. and Wielandt, H. W. (1953), 'The variation of the spectrum of a normal matrix', *Duke Math. J.* **20**(1), 37–39.

URL: *https://doi.org/10.1215/S0012-7094-53-02004-3*

Laurent, B. and Massart, P. (2000), 'Adaptive estimation of a quadratic functional by model selection', *Annals of Statistics* pp. 1302–1338.

Mai, Q., Yang, Y. and Zou, H. (2019), 'Multiclass sparse discriminant analysis', *Statistica Sinica* **29**(1), 97–111.

Mirza, B., Lin, Z., Cao, J. and Lai, X. (2015), Voting based weighted online sequential extreme learning machine for imbalance multi-class classification, *in* '2015 IEEE International Symposium on Circuits and Systems (ISCAS)', IEEE, pp. 565–568.

Price, B. S., Geyer, C. J. and Rothman, A. J. (2019), 'Automatic response category combination in multinomial logistic regression', *Journal of Computational and Graphical Statistics* **28**(3), 758–766.

Soda, P. (2011), 'A multi-objective optimisation approach for class imbalance learning', *Pattern Recognition* **44**(8), 1801–1810.

Wedin, P.-Å. (1972), 'Perturbation bounds in connection with singular value decomposition', *BIT Numerical Mathematics* **12**(1), 99–111.

Zhou, H. and Li, L. (2014), 'Regularized matrix regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2), 463–483.