A New Reduced-Rank Linear Discriminant Analysis Method and Its Applications

Yue Selena Niu, Ning Hao, and Bin Dong

University of Arizona and Peking University

March 27, 2017

Supplementary Material

This document contains supplementary materials for paper "A New Reduced-Rank Linear Discriminant Analysis Method and Its Applications".

S1 Technical proofs

Proof of Proposition 1. Recall that, by our convention, the data have been centered, $\hat{\boldsymbol{\mu}} = n^{-1}\sum_{k=1}^{K} n_k \hat{\boldsymbol{\mu}}_k = 0$, so $\mathbf{B} = n^{-1}\sum_{k=1}^{K} n_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^\top$. Note that $\mathbf{B}$ is semi-positive definite.

For a special case $\mathbf{W} = \mathbf{I}$, $\{\mathbf{v}_k\}_{k=1}^r$ are just eigenvectors of $\mathbf{B}$ corresponding to positive eigenvalues. For any vector $\mathbf{u} \perp \hat{\mathbf{C}}$, we have

$$
\begin{aligned}
& \mathbf{u} \perp \hat{\boldsymbol{\mu}}_k, \quad k = 1, 2, ..., K \\
\Leftrightarrow \quad & \mathbf{u}^\top \mathbf{B} \mathbf{u} = \frac{1}{n}\sum_{k=1}^{K} n_k \mathbf{u}^\top \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^\top \mathbf{u} = \frac{1}{n}\sum_{k=1}^{K} n_k (\hat{\boldsymbol{\mu}}_k^\top \mathbf{u})^2 = 0 \\
\Leftrightarrow \quad & \mathbf{u} \text{ belongs to the eigen-space of } \mathbf{B} \text{ corresponding to eigenvalue } 0 \\
\Leftrightarrow \quad & \mathbf{u} \perp \operatorname{span}\{\mathbf{v}_k\}_{k=1}^r.
\end{aligned}
$$

That is, $\hat{\mathbf{C}}$ and $\text{span}\{\mathbf{v}_k\}_{k=1}^r$ have the same orthogonal complement. Hence they are the same linear subspace and have the same dimension.

For arbitrary nonsingular $\mathbf{W}$, we may transform the data by linear operator $\mathbf{W}^{-1/2}$. That is, define $\tilde{\mathbf{X}}_i = \mathbf{W}^{-1/2}\mathbf{X}_i$, $1 \le i \le n$. It is easy to see that the statistics after transformation satisfy $\tilde{\mathbf{W}} = \mathbf{I}$, $\tilde{\mathbf{B}} = \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, $\tilde{\hat{\boldsymbol{\mu}}}_k = \mathbf{W}^{-1/2}\hat{\boldsymbol{\mu}}_k$, $\tilde{\mathbf{C}} = \mathbf{W}^{-1/2}\hat{\mathbf{C}}$, $\tilde{\mathbf{v}}_k = \mathbf{W}^{1/2}\mathbf{v}_k$ (no negative sign on the power). By the argument above, we have $\tilde{\mathbf{C}} = \text{span}\{\tilde{\mathbf{v}}_k\}_{k=1}^r$, so $\mathbf{W}^{-1}\hat{\mathbf{C}} = \mathbf{W}^{-1/2}\tilde{\mathbf{C}} = \text{span}\{\mathbf{W}^{-1/2}\tilde{\mathbf{v}}_k\}_{k=1}^r = \text{span}\{\mathbf{v}_k\}_{k=1}^r$.

In fact, the proof goes through if $\mathbf{W}$ is replaced by an arbitrary nonsingular equivariant covariance estimator. Hence we have the following corollary.

Corollary 1 *The conclusion of Proposition 1 still holds if* $\mathbf{W}$ *is replaced by any nonsingular equivariant within-class covariance estimate. In particular, replacing* $\mathbf{W}$ *by its diagonal part* $\hat{\mathbf{D}}_w$*, we can view diagonal LDA as a dimension reduction tool.*

Proof of Theorem 1. We show a proof for a large family described in Remark 5 $\boldsymbol{\Sigma}_{\boldsymbol{\rho}} = \boldsymbol{\Sigma}_w + \sum_{k=1}^K \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$, where $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)^\top$ with $\rho_k > 0$ for all k. Theorem 1 can be obtained as a special case because the family $\{\boldsymbol{\Sigma}_\gamma\}_{\gamma>0}$ is included in the larger one.

Let us fix an arbitrary $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)^\top$ with all positive entries, and $\mathbf{U}_O^\top \boldsymbol{\Sigma}_{\boldsymbol{\rho}} \mathbf{U}_O = \mathbf{D}_O$. By the spiked condition, we can write

$$\boldsymbol{\Sigma}_w = \lambda_p \mathbf{I} + \sum_{i=1}^s (\lambda_i - \lambda_p) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top,$$

where $\{\xi_i\}_{i=1}^s$ are eigenvectors to eigenvalues larger than λ_p. For $1 \le k < \ell \le K$, we have

$$\begin{aligned}
& \boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \\
= & \left(\lambda_p \mathbf{I} + \sum_{i=1}^s (\lambda_i - \lambda_p) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right)^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \\
= & \left(\lambda_p^{-1} \mathbf{I} - \sum_{i=1}^s \frac{\lambda_i - \lambda_p}{\lambda_p \lambda_i} \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top\right) (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \\
= & \lambda_p^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) - \sum_{i=1}^s \left[\frac{\lambda_i - \lambda_p}{\lambda_p \lambda_i} \boldsymbol{\xi}_i^\top (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)\right] \boldsymbol{\xi}_i \\
\in & \text{span}\{\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell, \boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_s\}.
\end{aligned} \tag{S1.1}$$

Moreover,

$$\boldsymbol{\Sigma}_{\boldsymbol{\rho}} = \lambda_p \mathbf{I} + \sum_{i=1}^{s} (\lambda_i - \lambda_p) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^\top + \sum_{k=1}^{K} \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top . \tag{S1.2}$$

If $p > s + K - 1$, the dimension of linear subspace $\mathbf{S} = \text{span}\left\{\{\boldsymbol{\xi}_i\}_{i=1}^s, \{\boldsymbol{\mu}_k\}_{k=1}^K\right\}$ is at most $s+K-1$ because of our convention $\sum_{k=1}^K \pi_k \boldsymbol{\mu}_k = 0$. On one hand, by (S1.1), $\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \in \mathbf{S}$. On other the hand, the eigenspace of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}}$ corresponding to eigenvalue λ_p is orthogonal to $\mathbf{S}$ by (S1.2). Therefore, columns of $\mathbf{U}_{O2}$ are orthogonal to $\mathbf{S}$, and hence to $\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$ for all k, ℓ.

Proof of Theorem 2. The proof follows the proof of Theorem 1 by noticing that $\boldsymbol{\mu}_k = \sum_{t=1}^{R_k} \pi_{kt} \boldsymbol{\mu}_{kt}$, and $\text{span}\{\boldsymbol{\mu}_k\}_{k=1}^K \subset \text{span}\{\boldsymbol{\mu}_{kt} : 1 \le k \le K; 1 \le t \le R_k\}$.

Proof of Lemma 1.

$$\begin{aligned}
\mathbf{T}_\gamma &= \mathbf{W} + \gamma \mathbf{B} \\
&= \frac{1}{n}\left(\sum_{k=1}^K \sum_{i \in C_k} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_k)^\top + \sum_{k=1}^K \gamma n_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^\top\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{i} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}_{Y_i})(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_{Y_i})^\top + \sum_{k=1}^K \gamma n_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^\top\right) \\
&= \frac{1}{n} \mathbf{A}_\gamma^\top \mathbf{A}_\gamma
\end{aligned}$$

Lemma 2 In the context of formula (2.1), let $\boldsymbol{\beta}_{k,\ell} = \boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$ and $\mathbf{H} \subset \mathbb{R}^p$ is arbitrary linear subspace such as $\boldsymbol{\beta}_{k,\ell} \in \mathbf{H}$. Let $\mathbf{P}_{\mathbf{H}}$ be the projection operator from $\mathbb{R}^p$ to $\mathbf{H}$. Then the normal vector to the optimal discriminant boundary separating groups k and ℓ using information from only the projected data $\mathbf{P}_{\mathbf{H}}(\mathbf{X})$ is the same as $\boldsymbol{\beta}_{k,\ell}$.

The conclusion below (2.1) follows Lemma 2 with the choice $\mathbf{H} = \boldsymbol{\Sigma}_w^{-1}\mathbf{C}$.

Proof of Lemma 2. Let $\{\mathbf{h}_j\}_{j=1}^p$ be an orthonormal basis for $\mathbb{R}^p$, and $\mathbf{H} = \text{span}\{\mathbf{h}_j\}_{j=1}^q$, $\mathbf{G} = \text{span}\{\mathbf{h}_j\}_{j=q+1}^p$. By abuse of notation, we also use $\mathbf{H}$ and $\mathbf{G}$ to denote $q \times p$ matrix $(\mathbf{h}_1, ..., \mathbf{h}_q)^\top$ and $(p-q) \times p$ matrix $(\mathbf{h}_{q+1}, ..., \mathbf{h}_p)^\top$, respectively. Let $\mathbf{F} = (\mathbf{H}^\top, \mathbf{G}^\top)^\top$ be an orthogonal matrix. Let $\tilde{\mathbf{X}} = \mathbf{F}\mathbf{X}$. Then $(\tilde{\mathbf{X}} | Y = k) \sim \mathcal{N}(\mathbf{F}\boldsymbol{\mu}_k, \mathbf{F}\boldsymbol{\Sigma}_w \mathbf{F}^\top)$.

Now we work on an equivalent model $(\tilde{\mathbf{X}}, Y)$, where the projection $\mathbf{P}_{\mathbf{H}}$ is simply a projection to the first q coordinates. In this equivalent model, it is sufficient to show that the optimal discriminant boundaries obtained from whole data $\tilde{\mathbf{X}}$ and the projected data are exactly the same.

First, using the whole data $\tilde{\mathbf{X}}$, the normal vector to the optimal discriminant boundary separating groups k and ℓ is

$$\tilde{\boldsymbol{\beta}}_{k,\ell} = \left(\mathbf{F}\boldsymbol{\Sigma}_w\mathbf{F}^\top\right)^{-1}(\mathbf{F}\boldsymbol{\mu}_k - \mathbf{F}\boldsymbol{\mu}_\ell) = \mathbf{F}\boldsymbol{\Sigma}_w^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) = \mathbf{F}\boldsymbol{\beta}_{k,\ell}. \tag{S1.3}$$

Note that the condition $\boldsymbol{\beta}_{k,\ell} \in \mathbf{H}$ implies $\mathbf{F}\boldsymbol{\beta}_{k,\ell} = \binom{\mathbf{H}\boldsymbol{\beta}_{k,\ell}}{\mathbf{G}\boldsymbol{\beta}_{k,\ell}} = \binom{\mathbf{H}\boldsymbol{\beta}_{k,\ell}}{\mathbf{0}}$. That is, $\tilde{\boldsymbol{\beta}}_{k,\ell}$ is a sparse vector supported in its first q coordinates. By (S1.3), we have

$$\mathbf{F}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) = \left(\mathbf{F}\boldsymbol{\Sigma}_w\mathbf{F}^\top\right)\mathbf{F}\boldsymbol{\beta}_{k,\ell},$$

which implies

$$\begin{pmatrix}\mathbf{H}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)\\ \mathbf{G}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)\end{pmatrix} = \begin{pmatrix}\mathbf{H}\boldsymbol{\Sigma}_w\mathbf{H}^\top & \mathbf{G}\boldsymbol{\Sigma}_w\mathbf{H}^\top\\ \mathbf{H}\boldsymbol{\Sigma}_w\mathbf{G}^\top & \mathbf{G}\boldsymbol{\Sigma}_w\mathbf{G}^\top\end{pmatrix}\begin{pmatrix}\mathbf{H}\boldsymbol{\beta}_{k,\ell}\\ \mathbf{0}\end{pmatrix}.$$

Comparing the top q rows of both sides, we have $\mathbf{H}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) = (\mathbf{H}\boldsymbol{\Sigma}_w\mathbf{H}^\top)\mathbf{H}\boldsymbol{\beta}_{k,\ell}$. So

$$\mathbf{H}\boldsymbol{\beta}_{k,\ell} = (\mathbf{H}\boldsymbol{\Sigma}_w\mathbf{H}^\top)^{-1}\mathbf{H}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell). \tag{S1.4}$$

To summarise, $\tilde{\boldsymbol{\beta}}_{k,\ell}$ is a sparse vector with its first q coordinates defined as in (S1.4).

Second, we consider the projected data. Write $\tilde{\mathbf{X}} = \binom{\mathbf{HX}}{\mathbf{GX}} = \binom{\tilde{\mathbf{X}}_1}{\tilde{\mathbf{X}}_2}$, where $\tilde{\mathbf{X}}_1 | Y = k \sim \mathcal{N}(\mathbf{H}\boldsymbol{\mu}_k, \mathbf{H}\boldsymbol{\Sigma}_w\mathbf{H}^\top)$. Using information from the projected data $\tilde{\mathbf{X}}_1$ only, we find the normal vector to the optimal discriminant boundary is $(\mathbf{H}\boldsymbol{\Sigma}_w\mathbf{H}^\top)^{-1}\mathbf{H}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$ which is the same as $\mathbf{H}\boldsymbol{\beta}_{k,\ell}$ by (S1.4). Therefore, we lose no information to retain $\tilde{\boldsymbol{\beta}}_{k,\ell}$ using projected data $\tilde{\mathbf{X}}_1$ instead of whole data $\tilde{\mathbf{X}}$.