Supplementary Materials for "Efficient integration of sufficient dimension reduction and prediction in discriminant analysis"

Xin Zhang and Qing Mai

S1 Additional numerical results

S1.1 Sensitivity of λ

As suggested in Section 5, for the ENDS estimator based on (5.1), we use $\lambda = 1$ if the LDA classifier is applied and select λ based on cross-validation from the set of $\{0, 0.1, \ldots, 1\}$ if the QDA classifier is applied. In our experience, the ENDS estimator is not sensitive to the choice of the tuning parameter λ. We use model (Q2) in the simulation (Section 8.1) of the paper for illustration.

Recall from Section 8.1, in this model, we considered the QDA model with $K = 4$ classes and equal number of observations per class, $n_1 = \cdots = n_K = 75$. Basis matrix for the envelope $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$ is first generated randomly and orthogonalized, then $\mathbf{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ is obtained correspondingly so that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ is an orthogonal matrix. The envelope dimension $u = 2$ and the original predictor dimension $p = 15$. Symmetric and positive definite matrices $\mathbf{\Omega}_k \in \mathbb{R}^{u\times u}$, $k = 1, \ldots, K$, and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ are generated as $\mathbf{A}\mathbf{A}^T/\|\mathbf{A}\mathbf{A}^T\|_F$ with $\mathbf{A}$ being a random matrix with the same dimension as $\mathbf{\Omega}_k$ or $\mathbf{\Omega}_0$, and have independent Uniform$(0, 1)$ elements. Within each class k, we generate data from the QDA model, $\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$, $k = 1, \ldots K$, where $\boldsymbol{\mu}_k = \mathbf{\Gamma}\boldsymbol{\eta}_k$ for some randomly generated $\boldsymbol{\eta}_k \in \mathbb{R}^{u\times 1}$ with $N(0, 1)$ elements. We then let $\mathbf{\Sigma}_k^* = \exp(-k) \cdot \mathbf{\Gamma}\mathbf{\Omega}_k\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and standardize it to get $\mathbf{\Sigma}_k = 25 \cdot \mathbf{\Sigma}_k^*/\|\mathbf{\Sigma}_k^*\|_F$, where the scaling constant 25 makes the Bayes' classification error about 0.22.

We varied λ from 0 to 1, and calculated the classification error rate and the principal angle between the estimated and the true subspace spanned by $\widehat{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$, respectively. For each λ value, we repeated 20 times for independent training and testing sets, and also included the results of

$p = 15, K = 4$	Bayes	LDA	ENDS-LDA	QDA	ENDS-QDA	S.E.$\leq$
$n_k = 10$	22.0	48.8	46.7	54.8	38.4	0.5
$n_k = 15$	22.0	46.8	43.8	55.9	34.6	0.4

Table 1: Small n_k simulation setting. The QDA uses pseudoinverse of sample covariance due to insufficient sample size. The last column is the largest standard error among all four estimators.

SIR, SAVE and DR as references. From Figure A1, we confirm that the ENDS estimator is very insensitive to the tuning parameter λ for either the basis estimation or the final classification.

S1.2 ENDS-QDA with small sample size

Using the same data generating procedure as in the previous section, we now consider the scenarios where $15 = n_k = p$ and $10 = n_k < p$. Table 1 summarizes the classification results from LDA, QDA, ENDS-LDA and ENDS-QDA. Recall that, for ENDS-LDA, we use $\lambda = 1$ and the group-specific covariance estimators all equal to the pooled sample covariance: $\mathbf{S}_k(\lambda) = \mathbf{S}$. For ENDS-QDA, by letting $\lambda = 0.1$, we impose a small amount of regularization just to avoid singularity in $\mathbf{S}_k(\lambda) = 0.9\mathbf{S}_k + 0.1\mathbf{S}$. As we discussed at the end of Section 5, if we do not impose such regularization, the symmetric matrix $\mathbf{\Gamma}^T\mathbf{S}_k\mathbf{\Gamma}$ from (5.1) can be singular during iterations because the rank of sample covariance $\mathbf{S}_k$ is at most $(n_k - 1)$. From Table 1, we see that QDA is worse than LDA in terms of classification accuracy because of $n_k \leq p$, but ENDS-QDA with a small amount of regularization is substantially better than LDA and ENDS-LDA.

S2 A technical lemma and its proof

Lemma 1. *Under the QDA model* (4.1)*,* $\mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}} = \mathcal{L} \cup \mathcal{Q}$ *and, moreover,* $\mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}$*.*

Proof. First, we show that $\mathcal{B} \cup \mathcal{Q}$ is the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. From Cook and Forzani (2009, Theorem 1), the central subspace under the QDA model is the smallest subspace $\mathcal{S}$ such that (i) $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) \in \mathcal{S}$ for all $k = 1, \ldots, K$; and (ii) $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_1^{-1} = \ldots = \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_K^{-1}$. Then condition (i) is equivalent to $\mathcal{B} \subseteq \mathcal{S}$ and condition (ii) is equivalent to $\mathcal{Q} \subseteq \mathcal{S}$. Therefore, the smallest such $\mathcal{S}$ is simply $\mathcal{B} \cup \mathcal{Q}$ and is the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

Next, we show that $\mathcal{L} \cup \mathcal{Q}$ is the central discriminant subspace $\mathcal{S}_{D(Y|\mathbf{X})}$. By combining the QDA Bayes' rule 4.3 with the definitions of $\mathcal{L}$ and $\mathcal{Q}$, it is clear that $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{L}\cup\mathcal{Q}}\mathbf{X})$. Hence

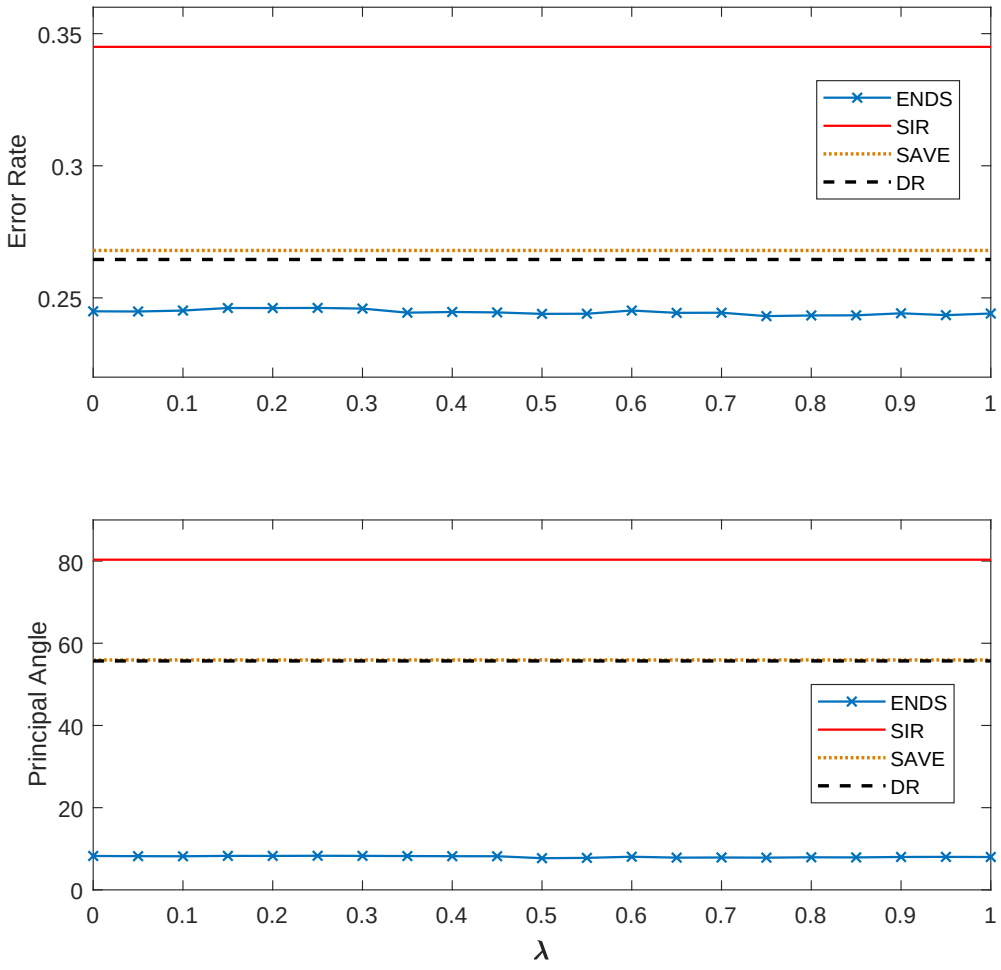

Figure A1: ENDS estimator with different tuning parameter λ. The averaged classification error rate (top plot) and the principal angle between the estimated and the true subspaces (bottom plot).

$\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{L} \cup \mathcal{Q}$. Then we prove by contradiction for the equality $\mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{L} \cup \mathcal{Q}$. Suppose that $\mathcal{S}_{D(Y|\mathbf{X})} \subset \mathcal{L} \cup \mathcal{Q}$. Then there exists some vector $\mathbf{v} \notin \mathcal{S}_{D(Y|\mathbf{X})}$ and $\mathbf{v} \in \mathcal{L} \cup \mathcal{Q}$. We can vary the value of the variable $\mathbf{v}_k^T \mathbf{Q}_{\mathcal{S}_{D(Y|\mathbf{X})}} \mathbf{X}$ without any effect on $\phi(\mathbf{P}_{\mathcal{S}_{D(Y|\mathbf{X})}} \mathbf{X}) = \phi(\mathbf{X})$. We can re-write the QDA Bayes' rule in the quadratic form of $\phi(\mathbf{X}) = \arg\max_k \{a_k + \mathbf{b}_k^T \mathbf{X} + \mathbf{X}^T \mathbf{C}_k \mathbf{X}\}$ for some $\mathbf{b}_k \in \mathbb{R}^p$ and $\mathbf{C}_k \in \mathbb{R}^{p \times p}$ such that $\text{span}(\mathbf{b}_k) \subseteq \mathcal{L} \cup \mathcal{Q}$ and $\text{span}(\mathbf{C}_k) \subseteq \mathcal{L} \cup \mathcal{Q}$. Then we notice that $\mathbf{v}^T \mathbf{Q}_{\mathcal{S}_{D(Y|\mathbf{X})}} \mathbf{X}$ is stochastic (because $\mathbf{v} \notin \mathcal{S}_{D(Y|\mathbf{X})}$) and could still change $\phi(\mathbf{X})$ since the value of $a_k + \mathbf{b}_k^T \mathbf{X} + \mathbf{X}^T \mathbf{C}_k \mathbf{X}$ can vary from $-\infty$ to $+\infty$.

Finally, we show that $\mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}$. By definition, the union of two subspace of $\mathbb{R}^p$ is $\mathcal{S}_1 \cup \mathcal{S}_2 = \{\mathbf{a} + \mathbf{b} \mid \mathbf{a} \in \mathcal{S}_1, \mathbf{b} \in \mathcal{S}_2\}$. For vectors in $\mathcal{L}$, we re-write $\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 = (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1}) \boldsymbol{\mu}_k + \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, where the first term is in $\mathcal{Q}$. Therefore,

$$\mathcal{L} \cup \mathcal{Q} = \text{span}\{\boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \mid k = 2, \ldots, K\} \cup \mathcal{Q}.$$

Similarly, we re-write $\boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) + (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}^{-1})(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, where the first term on the right hand side of the equation is in $\mathcal{B} = \text{span}\{\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \mid k = 2, \ldots, K\}$. Therefore, to prove $\mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}$, it suffices to show that $\text{span}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}^{-1}) \subseteq \mathcal{Q}$. From Cook and Forzani (2009, Proposition 1), we have the following equivalence for any subspace $\mathcal{S} \subseteq \mathbb{R}^p$,

$$\mathbf{Q}_{\mathcal{S}} \boldsymbol{\Sigma}_1^{-1} = \ldots = \mathbf{Q}_{\mathcal{S}} \boldsymbol{\Sigma}_K^{-1} \iff \mathbf{Q}_{\mathcal{S}} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}^{-1}) = 0, \ \forall k.$$

This implies that, for $\mathcal{Q}$, we have $\mathbf{Q}_{\mathcal{Q}} (\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}^{-1}) = 0$ and hence $\text{span}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}^{-1}) \subseteq \mathcal{Q}$.

□

S3 Proof for Proposition 1

By definition, ENDS will contain $\mathcal{S}_{D(Y|\mathbf{X})}$ because of the first condition in (2.1). The second condition in (2.1) is equivalent to stating that $\mathcal{S}$ is a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Together, we have that ENDS is the intersection of all reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathcal{S}_{D(Y|\mathbf{X})}$, which is the $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of the CDS, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})})$ (cf. Definition 1).

The existence and uniqueness of ENDS is thus guaranteed by the existence of CDS. When CDS exists, the ENDS $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})})$ is the intersection of all reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathcal{S}_{D(Y|\mathbf{X})}$ by Definition 1. Trivially, $\mathbb{R}^p$ is a reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that contains $\mathcal{S}_{D(Y|\mathbf{X})}$. Moreover, if any two subspaces $\mathcal{S}_1 \subseteq \mathbb{R}^p$ and $\mathcal{S}_2 \subseteq \mathbb{R}^p$ both are reducing subspace of $\boldsymbol{\Sigma}_{\mathbf{X}}$ that

contains $\mathcal{S}_{D(Y|\mathbf{X})}$, then we need to show that their intersection is still a reducing subspace of $\mathbf{\Sigma_X}$ that contains $\mathcal{S}_{D(Y|\mathbf{X})}$. This implies the uniqueness and existence of the ENDS $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{S}_{D(Y|\mathbf{X})})$.

First, it is easy to see that the intersection $\mathcal{S}_1 \cap \mathcal{S}_2$ still contains $\mathcal{S}_{D(Y|\mathbf{X})}$. To see that it is still a reducing subspace of $\mathbf{\Sigma_X}$, we note that $\mathcal{S}_1$ is a reducing subspace of $\mathbf{\Sigma_X}$ implies that $\mathbf{P}_{\mathcal{S}_1}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{S}_1} = 0$. Similarly, we have $\mathbf{P}_{\mathcal{S}_2}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{S}_2} = 0$. Therefore,

$$\begin{aligned}
\mathbf{P}_{\mathcal{S}_1\cap\mathcal{S}_2}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{S}_1\cap\mathcal{S}_2} &= \mathbf{P}_{\mathcal{S}_1}\mathbf{P}_{\mathcal{S}_2}\mathbf{\Sigma_X}(\mathbf{I} - \mathbf{P}_{\mathcal{S}_1}\mathbf{P}_{\mathcal{S}_2}) \\
&= \mathbf{P}_{\mathcal{S}_1}\mathbf{P}_{\mathcal{S}_2}\mathbf{\Sigma_X}(\mathbf{Q}_{\mathcal{S}_1} + \mathbf{P}_{\mathcal{S}_1} - \mathbf{P}_{\mathcal{S}_1}\mathbf{P}_{\mathcal{S}_2}) \\
&= \mathbf{P}_{\mathcal{S}_1}\mathbf{P}_{\mathcal{S}_2}\mathbf{\Sigma_X}(\mathbf{Q}_{\mathcal{S}_1} + \mathbf{P}_{\mathcal{S}_1}\mathbf{Q}_{\mathcal{S}_2}) \\
&= \mathbf{P}_{\mathcal{S}_2}(\mathbf{P}_{\mathcal{S}_1}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{S}_1}) + \mathbf{P}_{\mathcal{S}_1}(\mathbf{P}_{\mathcal{S}_2}\mathbf{\Sigma_X}\mathbf{Q}_{\mathcal{S}_2})\mathbf{P}_{\mathcal{S}_1} \\
&= 0,
\end{aligned}$$

which implies that $\mathcal{S}_1 \cap \mathcal{S}_2$ is a reducing subspace of $\mathbf{\Sigma_X}$. This completes the proof of existence and uniqueness of the ENDS (provided that the CDS exists).

S4 Proof for Proposition 2 and ENDS-LDA parameterization

Under the LDA model assumption (3.1), we prove Proposition 2 by showing in the following that, (2.2)$\Longrightarrow$(2.1); (2.1)$\Longrightarrow$(3.4); and (3.4)$\Longrightarrow$(2.2).

(1) The strong envelope pursuit (2.2) implies the weak envelope pursuit (2.1).

By the definition of the Bayes' rule, the first statement in (2.2), $\Pr(Y = k \mid \mathbf{X}) = \Pr(Y = k \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$ for all $k = 1, \ldots, K$, implies that $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$. The second statement in (2.2), $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X} \mid (Y = k)$ for all $k = 1, \ldots, K$, implies that $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X}$, which further implies $\mathrm{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0$.

(2) The weak envelope pursuit (2.1) implies the envelope parameterization (3.4).

First, we show that (2.1) implies $\mathcal{B} \subseteq \mathcal{S}$ in (3.4). Recall from (3.3) that the Bayes' rule under the LDA model (3.1) is $\phi(\mathbf{X}) = \arg\max_k[\log(\pi_k/\pi_1) + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}\{\mathbf{X} - (\boldsymbol{\mu}_k + \boldsymbol{\mu}_1)/2\}]$. From this LDA Bayes' rule, we can see that $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{B}}\mathbf{X})$ by definitions of $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$ and $\boldsymbol{\beta}_k = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, $k = 2, \ldots, K$. Moreover, $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$ implies that $\mathcal{B} \subseteq \mathcal{S}$. To see this, we prove by contradiction. Suppose that for some $\boldsymbol{\beta}_k \notin \mathcal{S}$, then we can vary the value of the univariate variable $\boldsymbol{\beta}_k^T\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ without any effect on $\phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$. But, by re-writing the Bayes' rule

in the form of $\phi(\mathbf{X}) = \arg\max_k\{\alpha_k + \boldsymbol{\beta}_k^T\mathbf{X}\}$, we notice that $\boldsymbol{\beta}_k^T\mathbf{Q}_\mathcal{S}\mathbf{X}$ could still change $\phi(\mathbf{X})$ since the value of $\alpha_k + \boldsymbol{\beta}_k^T\mathbf{X}$ can vary from $-\infty$ to $+\infty$. Therefore, $\phi(\mathbf{X}) = \phi(\mathbf{P}_\mathcal{S}\mathbf{X})$ implies that $\mathcal{B} \subseteq \mathcal{S}$.

Next, we need to show $\boldsymbol{\Sigma} = \mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{S} + \mathbf{Q}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S}$. More specifically, we want to derive $\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S} = 0$ from $\mathrm{cov}(\mathbf{P}_\mathcal{S}\mathbf{X}, \mathbf{Q}_\mathcal{S}\mathbf{X}) = \mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_\mathcal{S} = 0$. Notice that,

$$\begin{aligned}
\boldsymbol{\Sigma}_\mathbf{X} &= \mathrm{cov}(\mathbf{X}) = \mathrm{cov}\{\mathrm{E}(\mathbf{X} \mid Y)\} + \mathrm{E}\{\mathrm{cov}(\mathbf{X} \mid Y)\} \\
&= \sum_k \pi_k(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})^T + \boldsymbol{\Sigma} \\
&= \boldsymbol{\Sigma}\mathbf{P}_\mathcal{B}\mathbf{M}_1\mathbf{P}_\mathcal{B}\boldsymbol{\Sigma} + \boldsymbol{\Sigma},
\end{aligned}$$

where the last equality follows from introducing $\mathbf{M}_1 = \boldsymbol{\Sigma}^{-1}\sum_k \pi_k(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})^T\boldsymbol{\Sigma}^{-1} \geq 0$ and hence $\mathrm{span}(\mathbf{M}_1) = \mathcal{B}$ by the definition of $\mathcal{B}$. Since we have shown that (2.1) implies $\mathcal{B} \subseteq \mathcal{S}$. We can further write $\boldsymbol{\Sigma}_\mathbf{X} = \boldsymbol{\Sigma}\mathbf{P}_\mathcal{S}\mathbf{M}_1\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}$, and thus

$$\begin{aligned}
\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_\mathcal{S} &= \mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{S}\mathbf{M}_1\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S} + \mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S} \\
&= (\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{S}\mathbf{M}_1 + \mathbf{I}_p)\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S}.
\end{aligned}$$

Finally, because $\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{P}_\mathcal{S}\mathbf{M}_1 + \mathbf{I}_p > 0$ in the above equation, we have $\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}_\mathbf{X}\mathbf{Q}_\mathcal{S} = 0$ implies $\mathbf{P}_\mathcal{S}\boldsymbol{\Sigma}\mathbf{Q}_\mathcal{S} = 0$.

(3) The envelope parameterization (3.4) implies the strong envelope pursuit (2.2).

From the LDA Bayes' rule (3.3), we know that $\Pr(Y = k \mid \mathbf{X}) = \Pr(Y = k \mid \mathbf{P}_\mathcal{B}\mathbf{X})$, $k = 1, \ldots, K$. Therefore, $\mathcal{B} \subseteq \mathcal{S}$ from (3.4) immediately implies the first statement in (2.2). The second statement in (2.2) follows directly from the normal distribution and common within class covariance.

Finally, since we have $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_\boldsymbol{\Sigma}(\mathcal{B})$. Therefore, $\mathrm{span}(\boldsymbol{\beta}_k) \subseteq \mathcal{B} \subseteq \mathrm{span}(\boldsymbol{\Gamma})$ and we have (3.5). Next, we show $\mathrm{span}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$ for (3.6). By definition of $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, $k = 2, \ldots, K$, we can see that $\mathrm{span}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = \mathrm{span}(\boldsymbol{\Sigma}\boldsymbol{\beta}_k) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$, where the last inequality can be derived directly from plugging in (3.5). Notice that $\sum_{j=1}^K \pi_j = 1$, we can write $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = \sum_{j=1}^K \pi_j(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j) = \sum_{j=1}^K \pi_j\{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) - (\boldsymbol{\mu}_j - \boldsymbol{\mu}_1)\}$ and have $\mathrm{span}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$ and $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = \boldsymbol{\Gamma}\boldsymbol{\alpha}_k$. Finally, the intrinsic constraint $\sum_{j=1}^K \pi_j\boldsymbol{\alpha}_j = 0$ is from $\sum_{j=1}^K \pi_j(\boldsymbol{\mu}_j - \overline{\boldsymbol{\mu}}) = 0$.

S5 Derivations of MLEs for ENDS-LDA

In the LDA model with envelope parameterization (3.6), the conditional log-likelihood of $\mathbf{X}$ given Y can be written conveniently as

$$\begin{aligned} & L_n(\overline{\boldsymbol{\mu}}, \{\boldsymbol{\alpha}_k\}_{k=1}^K, \boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) \\ = & -\frac{n}{2}\left(\log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0|\right) \\ - & \frac{1}{2}\left\{\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\boldsymbol{\mu}})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\boldsymbol{\mu}})\right\}, \quad \text{(A1)} \end{aligned}$$

which is to be maximized over the intrinsic mean constraint that $\sum_{j=1}^K \widehat{\pi}_j\boldsymbol{\alpha}_j = 0$ where $\widehat{\pi}_j = n_j/n$ is the MLE for π_j. The log-determinant terms are derived from the fact that $\log|\boldsymbol{\Sigma}| = \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| = \log|\boldsymbol{\Omega}| + \log|\boldsymbol{\Omega}_0|$.

MLE for $\overline{\boldsymbol{\mu}}$. By taking derivative with respect to $\overline{\boldsymbol{\mu}}$ in (A1) and set it to zero, we have the following equality,

$$0 = \sum_{k=1}^K\sum_{i=1}^{n_k}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\left\{\overline{\boldsymbol{\mu}} - (\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k)\right\},$$

which lead to the MLE for $\overline{\boldsymbol{\mu}}$:

$$\widehat{\overline{\boldsymbol{\mu}}} = n^{-1}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\alpha}}_k) = n^{-1}\sum_{k=1}^K\sum_{i=1}^{n_k}\mathbf{X}_{i(k)} - \widehat{\boldsymbol{\Gamma}}\sum_{k=1}^K\widehat{\pi}_k\widehat{\boldsymbol{\alpha}}_k = \overline{\mathbf{X}},$$

where the last equality holds because $\sum_{k=1}^K \widehat{\pi}_k\widehat{\boldsymbol{\alpha}}_k = 0$ from the intrinsic mean constraint.

MLE for $\boldsymbol{\alpha}_k$. Replacing $\overline{\boldsymbol{\mu}}$ by $\overline{\mathbf{X}}$, the second part of the conditional log-likelihood in (A1) is now partially maximized and can be simplified as follows,

$$\begin{aligned} & -\frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\mathbf{X}})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\mathbf{X}}) \\ = & -\frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\}^T\boldsymbol{\Omega}^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\} \\ - & \frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}})^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}). \quad \text{(A2)} \end{aligned}$$

Then because $\boldsymbol{\alpha}_k$ only appears in the above quadratic form, we now need to maximize $\mathrm{F}_k(\boldsymbol{\alpha}_k) \equiv -\frac{1}{2}\sum_{i=1}^{n_k}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\}^T\boldsymbol{\Omega}^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\}$ for $k = 1, \ldots, K$, under the constraint

that $\sum_{j=1}^{K} \widehat{\pi}_j \boldsymbol{\alpha}_j = 0$. If we ignore the constraint, then the maximizer for each $\mathrm{F}_k(\boldsymbol{\alpha}_k)$ will be $\boldsymbol{\Gamma}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})$, which automatically satisfies the constraint. Therefore, the maximum likelihood estimators for $\boldsymbol{\alpha}_k$ under the constraint that $\sum_{j=1}^{K} \widehat{\pi}_j \boldsymbol{\alpha}_j = 0$ is simply $\widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\Gamma}}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})$, where $\widehat{\boldsymbol{\Gamma}}$ is the MLE for $\boldsymbol{\Gamma}$.

MLE for $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Substitute MLEs for $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\alpha}_k$'s into the log-likelihood function, the partially maximized log-likelihood is $L_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ as shown in the following,

$$\begin{aligned}
-\frac{2}{n} L_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &\simeq \log|\boldsymbol{\Omega}| + n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)\}^T \boldsymbol{\Omega}^{-1} \{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)\} \\
&+ \log|\boldsymbol{\Omega}_0| + n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{X}_{i(k)} - \overline{\mathbf{X}})^T \boldsymbol{\Gamma}_0 \boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}_0^T (\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) \\
&= \log|\boldsymbol{\Omega}| + \mathrm{tr}\left[\boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}^T \left\{ n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)^T \right\} \boldsymbol{\Gamma} \right] \\
&+ \log|\boldsymbol{\Omega}_0| + \mathrm{tr}\left[\boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}_0^T \left\{ n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T \right\} \boldsymbol{\Gamma}_0 \right] \\
&= \log|\boldsymbol{\Omega}| + \mathrm{tr}(\boldsymbol{\Omega}^{-1} \boldsymbol{\Gamma}^T \mathbf{S} \boldsymbol{\Gamma}) + \log|\boldsymbol{\Omega}_0| + \mathrm{tr}(\boldsymbol{\Omega}_0^{-1} \boldsymbol{\Gamma}_0^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}_0),
\end{aligned}$$

where $\mathbf{S} = n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)^T$ and $\mathbf{S}_{\mathbf{X}} = n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$ are the sample covariance matrices for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathrm{cov}(\mathbf{X})$. Then the MLEs for $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^T \mathbf{S} \widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T \mathbf{S}_{\mathbf{X}} \widehat{\boldsymbol{\Gamma}}_0$ by noticing that $\mathbf{B} = \arg\min_{\mathbf{A}} \{\log|\mathbf{A}| + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})\}$ for optimization over positive definite matrices.

MLE for $\boldsymbol{\Gamma}$. Finally, the partially maximized log-likelihood satisfies

$$\begin{aligned}
-\frac{2}{n} L_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) &\simeq \log|\boldsymbol{\Gamma}^T \mathbf{S} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}_0| \\
&\simeq \log|\boldsymbol{\Gamma}^T \mathbf{S} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}|,
\end{aligned}$$

which implies that the MLE for $\boldsymbol{\Gamma}$ is obtained by minimizing $\log|\boldsymbol{\Gamma}^T \mathbf{S} \boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}|$ over the semi-orthogonal constraint that $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_u$, so that the optimization is essentially over the Grassmann manifold $\mathcal{G}(p, u)$.

MLE for $\boldsymbol{\beta}_k$. From the definition, $\boldsymbol{\beta}_k = \boldsymbol{\Gamma} \boldsymbol{\eta}_k$, we have $\boldsymbol{\eta}_k = \boldsymbol{\Gamma}^T \boldsymbol{\beta}_k = \boldsymbol{\Gamma}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$. By plugging in the MLEs of $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}$ and $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$ to the parameterizations in Proposition 2, we have $\widehat{\boldsymbol{\eta}}_k = (\widehat{\boldsymbol{\Gamma}}^T \mathbf{S} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ and hence $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\Gamma}} \widehat{\boldsymbol{\eta}}_k = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S})} \mathbf{S}^{-1} (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$, $k = 2, \ldots, K$.

S6 Proof for Proposition 3 and ENDS-QDA parameterization

Similar to the proof for Proposition 2, now under the QDA model (4.1), we prove this Proposition by showing in the following that, (2.2)$\Longrightarrow$(2.1); (2.1)$\Longrightarrow$(4.5); (4.5) $\Longrightarrow$ (4.6); and (4.6) $\Longrightarrow$ (2.2).

(1) The strong envelope pursuit (2.2) implies the weak envelope pursuit (2.1).

By the definition of the Bayes' rule, the first statement in (2.2), $\Pr(Y = k \mid \mathbf{X}) = \Pr(Y = k \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$ for all $k = 1, \ldots, K$, implies that $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$. The second statement in (2.2), $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X} \mid (Y = k)$ for all $k = 1, \ldots, K$, implies that $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X}$, which further implies $\mathrm{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0$.

(2) The weak envelope pursuit (2.1) implies the envelope parameterization (4.5).

First, we show that $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$ in (2.1) implies $\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}$ in (4.5). From the definition of the central discriminant subspace, $\phi(\mathbf{X}) = \phi(\mathbf{P}_{\mathcal{S}}\mathbf{X})$ implies that $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{S}$. From Lemma 1, we have $\mathcal{B} \cup \mathcal{Q} = \mathcal{S}_{D(Y|\mathbf{X})}$ under the QDA model. Therefore, $\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}$.

Next, it is easy to have $\mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$ from $\mathrm{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0$ and $\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}$. From the proof of Proposition 2, we know that $\mathrm{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0$ and $\mathcal{B} \subseteq \mathcal{S}$ implies $\mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$.

(3) The envelope parameterization (4.5) implies the envelope parameterization (4.6).

Assuming the QDA model and $\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}$, we need to show that $\mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$ implies $\mathbf{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{Q}_{\mathcal{S}}$ for all $k = 1, \ldots, K$. Let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{p \times p}$ be an orthogonal basis matrix such that $\mathrm{span}(\mathbf{\Gamma}) = \mathcal{S}$. From Cook and Forzani (2009, Proposition 1), we have the following equality because of $\mathcal{Q} \subseteq \mathcal{S}$,

$$\mathbf{\Sigma}_k = \mathbf{\Sigma} + \mathbf{P}^T_{\mathbf{\Gamma}(\mathbf{\Sigma})}(\mathbf{\Sigma}_k - \mathbf{\Sigma})\mathbf{P}_{\mathbf{\Gamma}(\mathbf{\Sigma})}, \text{ for all } k = 1, \ldots, K, \tag{A3}$$

where $\mathbf{P}_{\mathbf{\Gamma}(\mathbf{\Sigma})} = \mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{\Sigma}$ is the projection matrix onto $\mathrm{span}(\mathbf{\Gamma})$ with $\mathbf{\Sigma}$ inner product. Because $\mathbf{\Sigma} = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$, we can write $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$ and $\mathbf{P}_{\mathbf{\Gamma}(\mathbf{\Sigma})} = \mathbf{P}_{\mathbf{\Gamma}} = \mathbf{P}_{\mathcal{S}}$. Therefore, (A3) become $\mathbf{\Sigma}_k = \mathbf{\Sigma} + \mathbf{P}_{\mathcal{S}}(\mathbf{\Sigma}_k - \mathbf{\Sigma})\mathbf{P}_{\mathcal{S}}$, which implies that $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k = \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}$ for all k. Hence, $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} = \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{P}_{\mathcal{S}} = 0$ since $\mathcal{S}$ is a reducing subspace for $\mathbf{\Sigma}$. The reducing subspace statement $\mathbf{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{Q}_{\mathcal{S}}$ is equivalent to $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} = 0$.

(4) The envelope parameterization (4.6) implies the strong envelope pursuit (2.2).

The first statement, $\Pr(Y = k \mid \mathbf{X}) = \Pr(Y = k \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$ in (2.2) is because $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}$ (from Lemma (1)) implies that $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}$. The second statement, $\mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X} \mid (Y = k)$ is

equivalent to $\boldsymbol{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}_k\mathbf{Q}_{\mathcal{S}}$ under the QDA model assumption.

S7 Derivations of MLEs for ENDS-QDA

In the QDA model with envelope parameterization (4.7), the conditional log-likelihood of $\mathbf{X}$ given Y can be written conveniently as

$$\begin{aligned} & L_n(\overline{\boldsymbol{\mu}}, \{\boldsymbol{\alpha}_k\}_{k=1}^K, \boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_k\}_{k=1}^K, \boldsymbol{\Omega}_0) \\ = & \sum_{k=1}^K (-\frac{n_k}{2}) \left(\log|\boldsymbol{\Omega}_k| + \log|\boldsymbol{\Omega}_0|\right) \\ - & \frac{1}{2}\left\{\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\boldsymbol{\mu}})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\boldsymbol{\mu}})\right\}, \quad \text{(A4)} \end{aligned}$$

which is to be maximized over the intrinsic mean constraint that $\sum_{j=1}^K \widehat{\pi}_j\boldsymbol{\alpha}_j = 0$ where $\widehat{\pi}_j = n_j/n$ is the MLE for π_j. The log-determinant terms are derived from the fact that $\log|\boldsymbol{\Sigma}_k| = \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}_k\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| = \log|\boldsymbol{\Omega}_k| + \log|\boldsymbol{\Omega}_0|$.

MLE for $\overline{\boldsymbol{\mu}}$. By taking derivative with respect to $\overline{\boldsymbol{\mu}}$ in (A4) and set it to zero, we have the following equality,

$$0 = \sum_{k=1}^K\sum_{i=1}^{n_k}(\boldsymbol{\Gamma}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\left\{\overline{\boldsymbol{\mu}} - (\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k)\right\},$$

which lead to the MLE for $\overline{\boldsymbol{\mu}}$:

$$\widehat{\overline{\boldsymbol{\mu}}} = n^{-1}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\alpha}}_k) = n^{-1}\sum_{k=1}^K\sum_{i=1}^{n_k}\mathbf{X}_{i(k)} - \widehat{\boldsymbol{\Gamma}}\sum_{k=1}^K\widehat{\pi}_k\widehat{\boldsymbol{\alpha}}_k = \overline{\mathbf{X}},$$

where the last equality holds because $\sum_{k=1}^K \widehat{\pi}_k\widehat{\boldsymbol{\alpha}}_k = 0$ from the intrinsic mean constraint.

MLE for $\boldsymbol{\alpha}_k$. Replacing $\overline{\boldsymbol{\mu}}$ by $\overline{\mathbf{X}}$, the second part of the conditional log-likelihood in (A4) is now partially maximized and can be simplified as follows,

$$\begin{aligned} & -\frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\mathbf{X}})^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)(\mathbf{X}_{i(k)} - \boldsymbol{\Gamma}\boldsymbol{\alpha}_k - \overline{\mathbf{X}}) \\ = & -\frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\}^T\boldsymbol{\Omega}_k^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) - \boldsymbol{\alpha}_k\} \\ - & \frac{1}{2}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}})^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}). \quad \text{(A5)} \end{aligned}$$

Then because $\boldsymbol{\alpha}_k$ only appears in the above quadratic form, we now need to maximize $\mathrm{F}_k(\boldsymbol{\alpha}_k) \equiv -\frac{1}{2}\sum_{i=1}^{n_k}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)}-\overline{\mathbf{X}})-\boldsymbol{\alpha}_k\}^T\boldsymbol{\Omega}_k^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)}-\overline{\mathbf{X}})-\boldsymbol{\alpha}_k\}$ for $k = 1, \ldots, K$, under the constraint that $\sum_{j=1}^K \widehat{\pi}_j\boldsymbol{\alpha}_j = 0$. If we ignore the constraint, then the maximizer for each $\mathrm{F}_k(\boldsymbol{\alpha}_k)$ will be $\boldsymbol{\Gamma}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})$, which automatically satisfies the constraint. Therefore, the maximum likelihood estimators for $\boldsymbol{\alpha}_k$ under the constraint that $\sum_{j=1}^K \widehat{\pi}_j\boldsymbol{\alpha}_j = 0$ is simply $\widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\Gamma}}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})$, where $\widehat{\boldsymbol{\Gamma}}$ is the MLE for $\boldsymbol{\Gamma}$.

MLE for $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_0$. Substitute MLEs for $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\alpha}_k$'s into the log-likelihood function, the partially maximized log-likelihood is $L_n(\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ as shown in the following,

$$\begin{aligned}
-\frac{2}{n}L_n(\boldsymbol{\Gamma}, \{\boldsymbol{\Omega}_k\}_{k=1}^K, \boldsymbol{\Omega}_0) &\simeq n^{-1}\sum_{k=1}^K\left[n_k\log|\boldsymbol{\Omega}_k| + \sum_{i=1}^{n_k}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)\}^T\boldsymbol{\Omega}_k^{-1}\{\boldsymbol{\Gamma}^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)\}\right] \\
&+ \log|\boldsymbol{\Omega}_0| + n^{-1}\sum_{k=1}^K\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}})^T\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}) \\
&= n^{-1}\sum_{k=1}^K\left(n_k\log|\boldsymbol{\Omega}_k| + \mathrm{tr}\left[\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma}^T\left\{\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)^T\right\}\boldsymbol{\Gamma}\right]\right) \\
&+ \log|\boldsymbol{\Omega}_0| + \mathrm{tr}\left[\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\left\{n^{-1}\sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T\right\}\boldsymbol{\Gamma}_0\right] \\
&= \sum_{k=1}^K\frac{n_k}{n}\{\log|\boldsymbol{\Omega}_k| + \mathrm{tr}(\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma}^T\mathbf{S}_k\boldsymbol{\Gamma})\} + \log|\boldsymbol{\Omega}_0| + \mathrm{tr}(\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0),
\end{aligned}$$

where $\mathbf{S}_k = n_k^{-1}\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)^T$ and $\mathbf{S}_\mathbf{X} = n^{-1}\sum_{i=1}^n(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$ are the sample covariance matrices for $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}_\mathbf{X} = \mathrm{cov}(\mathbf{X})$. Then the MLEs for $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_0$ are $\widehat{\boldsymbol{\Omega}}_k = \widehat{\boldsymbol{\Gamma}}^T\mathbf{S}_k\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T\mathbf{S}_\mathbf{X}\widehat{\boldsymbol{\Gamma}}_0$ by noticing that $\mathbf{B} = \arg\min_\mathbf{A}\{\log|\mathbf{A}| + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})\}$ for optimization over positive definite matrices.

MLE for $\boldsymbol{\Gamma}$. Finally, the partially maximized log-likelihood satisfies

$$\begin{aligned}
-\frac{2}{n}L_n(\boldsymbol{\Gamma}) &\simeq \sum_{k=1}^K\frac{n_k}{n}\log|\boldsymbol{\Gamma}^T\mathbf{S}_k\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}_0^T\mathbf{S}_\mathbf{X}\boldsymbol{\Gamma}_0| \\
&\simeq \sum_{k=1}^K\frac{n_k}{n}\log|\boldsymbol{\Gamma}^T\mathbf{S}_k\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}^{-1}\boldsymbol{\Gamma}|,
\end{aligned}$$

which implies that the MLE for $\boldsymbol{\Gamma}$ is obtained by minimizing $\sum_{k=1}^K\frac{n_k}{n}\log|\boldsymbol{\Gamma}^T\mathbf{S}\boldsymbol{\Gamma}| + \log|\boldsymbol{\Gamma}^T\mathbf{S}_\mathbf{X}^{-1}\boldsymbol{\Gamma}|$ over the semi-orthogonal constraint that $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$, so that the optimization is essentially over the Grassmann manifold $\mathcal{G}(p, u)$.

MLE for $\boldsymbol{\beta}_k$, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Sigma}$. The conclusion follows by plugging in the MLEs of $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_0$ to the parameterizations in Proposition 3.

S8 Proof for Proposition 4 and Equation (7.3)

The parameters of the original LDA model are $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}$. We can also re-parameterize them as $\overline{\boldsymbol{\mu}}, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K$ and $\boldsymbol{\Sigma}$. For all methods, the estimator of $\overline{\boldsymbol{\mu}}$ is simply the sample mean of $\mathbf{X}$, which will be asymptotically independent of all other estimators. We thus focus on the asymptotic properties of the discriminant directions $\boldsymbol{\beta}_k$, $k = 2, \ldots, K$.

The parameters involved in the coordinate representation of the ENDS-LDA model are $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Omega}_0$ and $\boldsymbol{\eta}_k$, $k = 2, \ldots, K$. We focus on the asymptotic properties of the estimable functions $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) = \boldsymbol{\Gamma}\boldsymbol{\eta}_k$, $k = 2, \ldots, K$, and the covariance $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. Specifically, we study the asymptotic covariances of the following parameters $\boldsymbol{\phi}$ and estimable functions $\mathbf{h}$,

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma}) \\ \boldsymbol{\eta}_2 \\ \vdots \\ \boldsymbol{\eta}_K \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_K \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}. \tag{A6}$$

Since $\mathbf{h}$ is a vector of length $(K-1)p + p(p+1)/2$ and $\boldsymbol{\phi}$ is a vector of length $pu - u^2 + (K-1)u + u(u+1)/2 + (p-u)(p-u+1)/2 = (K-1)u + p(p+1)/2$, the ENDS reduces the total number of parameters by $(K-1)(p-u)$.

We let $\widetilde{\mathbf{h}} = (\widetilde{\boldsymbol{\beta}}_2^T, \ldots, \widetilde{\boldsymbol{\beta}}_K^T, \mathrm{vech}^T(\mathbf{S}))^T$ be the standard MLE of the LDA model parameters, where recall that $\widetilde{\boldsymbol{\beta}}_k = \mathbf{S}^{-1}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$.

The asymptotic covariance matrix $\mathbf{C}_1$ of $\widetilde{\boldsymbol{\psi}}_1 = (\overline{\mathbf{X}}_1^T, \ldots, \overline{\mathbf{X}}_K^T, \mathrm{vech}^T(\mathbf{S}))^T$, the MLE of $\boldsymbol{\psi}_1 = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T, \mathrm{vech}^T(\boldsymbol{\Sigma}))$, can be derived straightforwardly as

$$\mathbf{C}_1 = \mathbf{J}_1^{-1}, \quad \mathbf{J}_1 = \mathrm{diag}\left\{\pi_1\boldsymbol{\Sigma}^{-1}, \ldots, \pi_K\boldsymbol{\Sigma}^{-1}, \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right\}, \tag{A7}$$

where $\mathbf{J}_1$ is the Fisher information matrix of parameter vector $\boldsymbol{\psi}_1$.

Next, the asymptotic covariance matrix $\mathbf{C}_2$ of $\widetilde{\boldsymbol{\psi}}_2 \equiv (\overline{\mathbf{X}}_2^T - \overline{\mathbf{X}}_1^T, \ldots, \overline{\mathbf{X}}_K^T - \overline{\mathbf{X}}_1^T, \mathrm{vech}^T(\mathbf{S}))^T$,

the MLE of $\boldsymbol{\psi}_2 = (\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_1^T \ldots, \boldsymbol{\mu}_K^T - \boldsymbol{\mu}_1^T, \mathrm{vech}^T(\boldsymbol{\Sigma}))^T$ is thus

$$\mathbf{C}_2 = \mathbf{L}_1 \mathbf{C}_1 \mathbf{L}_1^T = \begin{pmatrix} (\pi_2^{-1} + \pi_1^{-1})\boldsymbol{\Sigma} & \cdots & \pi_1^{-1}\boldsymbol{\Sigma} & 0 \\ \vdots & \ddots & \vdots & 0 \\ \pi_1^{-1}\boldsymbol{\Sigma} & \cdots & (\pi_K^{-1} + \pi_1^{-1})\boldsymbol{\Sigma} & 0 \\ 0 & 0 & 0 & \left\{\frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right\}^{-1} \end{pmatrix}, \tag{A8}$$

where the matrix $\mathbf{L}_1$ is the linear map between the estimators: $(\overline{\mathbf{X}}_2^T - \overline{\mathbf{X}}_1^T, \ldots, \overline{\mathbf{X}}_K^T - \overline{\mathbf{X}}_1^T, \mathrm{vech}^T(\mathbf{S}))^T = \mathbf{L}_1 \cdot (\overline{\mathbf{X}}_1^T, \ldots, \overline{\mathbf{X}}_K^T, \mathrm{vech}^T(\mathbf{S}))^T$. Specifically,

$$\mathbf{L}_1 = \mathrm{diag}\left\{\left(-\mathbf{1}_{(K-1)}, \mathbf{I}_{(K-1)}\right) \otimes \mathbf{I}_p,\ \mathbf{I}_{p(p+1)/2}\right\}, \tag{A9}$$

We can write $\mathbf{C}_2$ more compactly as

$$\mathbf{C}_2 = \mathrm{diag}\left[\mathbf{A} \otimes \boldsymbol{\Sigma},\ \left\{\frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right\}^{-1}\right], \tag{A10}$$

where $\mathbf{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ is a symmetric matrix has diagonal elements $(\pi_k^{-1} + \pi_1^{-1})$ and off-diagonal π_1^{-1}.

An alternative way of reaching these results is by calculating the Fisher information matrix of $\boldsymbol{\psi}_3 = (\overline{\boldsymbol{\mu}}^T, \boldsymbol{\mu}_2^T - \boldsymbol{\mu}_1^T \ldots, \boldsymbol{\mu}_K^T - \boldsymbol{\mu}_1^T, \mathrm{vech}^T(\boldsymbol{\Sigma}))^T = (\overline{\boldsymbol{\mu}}^T, \boldsymbol{\psi}_2^T)^T$. It is easy to see that there is a full-rank square matrix to transform between $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_3$. Let $\boldsymbol{\psi}_1 = \mathbf{L}_3\boldsymbol{\psi}_3$, then $\mathbf{L}_3 = \partial\boldsymbol{\psi}_1/\partial\boldsymbol{\psi}_3$ and $\mathbf{J}_3 = \mathbf{L}_3^T\mathbf{J}_1\mathbf{L}_3$. Straightforward calculation leads to block diagonal structure as $\mathbf{J}_3 = \mathrm{diag}\left\{\boldsymbol{\Sigma}^{-1}, \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}^{-1}, \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right\}$, which also verifies that $\overline{\mathbf{X}}$ is asymptotically independent of $(\overline{\mathbf{X}}_2^T - \overline{\mathbf{X}}_1^T, \ldots, \overline{\mathbf{X}}_K^T - \overline{\mathbf{X}}_1^T, \mathrm{vech}^T(\mathbf{S}))^T$. Then write $\mathbf{J}_3 = \mathrm{diag}\{\boldsymbol{\Sigma}^{-1}, \mathbf{J}_2\}$, we have the Fisher information matrix for $\boldsymbol{\psi}_2$,

$$\mathbf{J}_2 = \mathbf{C}_2^{-1} = \mathrm{diag}\left\{\mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}^{-1}, \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p\right\}. \tag{A11}$$

We can write $\mathbf{A} = \mathrm{diag}(\pi_2^{-1}, \ldots, \pi_K^{-1}) + \pi_1^{-1}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T$, then apply the Woodbury matrix identity, we have

$$\begin{aligned} \mathbf{A}^{-1} &= \mathrm{diag}(\pi_2, \ldots, \pi_K) \\ &- \mathrm{diag}(\pi_2, \ldots, \pi_K)\mathbf{1}_{K-1}(\pi_1 + \mathbf{1}_{K-1}^T\mathrm{diag}(\pi_2, \ldots, \pi_K)\mathbf{1}_{K-1})^{-1}\mathbf{1}_{K-1}^T\mathrm{diag}(\pi_2, \ldots, \pi_K) \\ &= \mathrm{diag}(\pi_2, \ldots, \pi_K) - \mathrm{diag}(\pi_2, \ldots, \pi_K)\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T\mathrm{diag}(\pi_2, \ldots, \pi_K) \end{aligned}$$

that is, $[\mathbf{A}^{-1}]_{ij}$ is $\pi_j(1 - \pi_j)$ for $i = j$ and is $-\pi_i\pi_j$ for $i \neq j$.

Finally, the asymptotic covariance matrix $\mathbf{C}_h$ of $\widetilde{\mathbf{h}}$ can be calculated from the one-to-one transformation between $\boldsymbol{\psi}_2$ and $\mathbf{h}$. Specifically, we noticed the simple linear transformation that $\boldsymbol{\mu}_k - \boldsymbol{\mu}_1 = \boldsymbol{\Sigma}\boldsymbol{\beta}_k$, and thus,

$$\frac{\partial(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial \boldsymbol{\beta}_k} = \boldsymbol{\Sigma} \tag{A12}$$

$$\frac{\partial(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)}{\partial \mathrm{vech}(\boldsymbol{\Sigma})} = (\boldsymbol{\beta}_k^T \otimes \mathbf{I}_p)\frac{\partial \mathrm{vec}(\boldsymbol{\Sigma})}{\partial \mathrm{vech}(\boldsymbol{\Sigma})} = (\boldsymbol{\beta}_k^T \otimes \mathbf{I}_p)\mathbf{E}_p. \tag{A13}$$

where $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ is the expansion matrix such that $\mathrm{vec}(\boldsymbol{\Sigma}) = \mathbf{E}_p \mathrm{vech}(\boldsymbol{\Sigma})$. Therefore,

$$\mathbf{C}_h = \mathbf{J}_h^{-1} = (\mathbf{G}^T \mathbf{J}_2 \mathbf{G})^{-1}, \quad \text{where } \mathbf{G} = \frac{\partial \boldsymbol{\psi}_2}{\partial \mathbf{h}} = \begin{pmatrix} \mathbf{I}_{K-1} \otimes \boldsymbol{\Sigma} & (\mathbf{B}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ 0 & \mathbf{I}_{p(p+1)/2} \end{pmatrix}, \tag{A14}$$

where $\mathbf{B} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p \times (K-1)}$.

Combining (A11) and (A14), we have

$$\begin{aligned}
\mathbf{J}_h &= \mathbf{G}^T \mathbf{J}_2 \mathbf{G} \\
&= \begin{pmatrix} \mathbf{I}_{K-1} \otimes \boldsymbol{\Sigma} & 0 \\ \mathbf{E}_p^T(\mathbf{B} \otimes \mathbf{I}_p) & \mathbf{I}_{p(p+1)/2} \end{pmatrix} \mathrm{diag}\left\{ \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma}^{-1}, \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \right\} \begin{pmatrix} \mathbf{I}_{K-1} \otimes \boldsymbol{\Sigma} & (\mathbf{B}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ 0 & \mathbf{I}_{p(p+1)/2} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{K-1} \otimes \boldsymbol{\Sigma} & 0 \\ \mathbf{E}_p^T(\mathbf{B} \otimes \mathbf{I}_p) & \mathbf{I}_{p(p+1)/2} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} \otimes \mathbf{I}_p & (\mathbf{A}^{-1}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{A}^{-1} \otimes \boldsymbol{\Sigma} & (\mathbf{A}^{-1}\mathbf{B}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{E}_p^T(\mathbf{B}\mathbf{A}^{-1} \otimes \mathbf{I}_p) & \mathbf{E}_p^T(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p + \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix},
\end{aligned}$$

whose inverse can be obtained using the block matrix inversion formula,

$$\mathbf{C}_h = \mathbf{J}_h^{-1} = \begin{pmatrix} \mathbf{C}_B & \mathbf{C}_{B\Sigma} \\ \mathbf{C}_{B\Sigma}^T & \mathbf{C}_\Sigma \end{pmatrix}, \tag{A15}$$

where the top left block is of primary interests. We calculate the asymptotic covariance of $\mathrm{vec}(\mathbf{B}) = (\boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_K^T)^T$, i.e. $\mathbf{C}_B \in \mathbb{R}^{p(K-1) \times p(K-1)}$, in the following,

$$\begin{aligned}
\mathbf{C}_B &= \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} + (\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \left\{ \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \right\}^{-1} \mathbf{E}_p^T(\mathbf{B} \otimes \boldsymbol{\Sigma}^{-1}) \\
&= \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} + (\mathbf{B}^T \otimes \boldsymbol{\Sigma}^{-1}) 2\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathbf{P}_{\mathbf{E}_p}(\mathbf{B} \otimes \boldsymbol{\Sigma}^{-1}) \\
&= \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} + 2(\mathbf{B}^T\boldsymbol{\Sigma} \otimes \mathbf{I}_p)\mathbf{P}_{\mathbf{E}_p}(\mathbf{B} \otimes \boldsymbol{\Sigma}^{-1}) \\
&= \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} + (\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B} \otimes \boldsymbol{\Sigma}^{-1}) + (\mathbf{B}^T \otimes \mathbf{B})\mathbf{K}_{p,(k-1)}.
\end{aligned}$$

Finally, for binary case, $\mathbf{B} = \boldsymbol{\beta} \in \mathbb{R}^p$ is a vector itself. We have $\mathbf{K}_{p,1} = \mathbf{I}_p$ since $\mathrm{vec}(\boldsymbol{\beta}) = \mathrm{vec}(\boldsymbol{\beta}^T)$. The conclusion follows by plugging in $\boldsymbol{\beta} = (\Delta, 0, \ldots, 0)^T$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$ and noticing $\mathbf{C}_B$ will coincide with Efron's (1975) results.

S9 Proof for Proposition 5

Recall that, $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \boldsymbol{\Gamma}\boldsymbol{\eta}$ under the envelope LDA model, therefore $\boldsymbol{\mu}_{Diff} \equiv (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\eta}$ under the envelope parameterization. We define the following parameters and estimable functions,

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta} \\ \mathrm{vec}(\boldsymbol{\Gamma}) \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \end{pmatrix}, \ \mathbf{h} = \mathbf{h}(\boldsymbol{\phi}) = \begin{pmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \ \mathbf{g} = \mathbf{g}(\boldsymbol{\phi}) = \begin{pmatrix} \boldsymbol{\mu}_{Diff} \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix}. \tag{A16}$$

In the proof of Proposition 4, we have already obtained the Fisher information of $\mathbf{h}$ and $\mathbf{g}$ under the classical LDA model,

$$\mathbf{J}_{\mathbf{g}} = \begin{pmatrix} \pi_1\pi_2\boldsymbol{\Sigma}^{-1} & 0 \\ 0 & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix}; \tag{A17}$$

$$\begin{aligned} \mathbf{J}_{\mathbf{h}} &= \begin{pmatrix} \pi_1\pi_2\boldsymbol{\Sigma} & (\pi_1\pi_2\boldsymbol{\beta}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{E}_p^T(\pi_1\pi_2\boldsymbol{\beta} \otimes \mathbf{I}_p) & \mathbf{E}_p^T(\pi_1\pi_2\boldsymbol{\beta}\boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p + \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix} \\ &= \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\Sigma} & (\boldsymbol{\beta}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{E}_p^T(\boldsymbol{\beta} \otimes \mathbf{I}_p) & \mathbf{E}_p^T\left\{\left(\boldsymbol{\beta}\boldsymbol{\beta}^T + (2\pi_1\pi_2)^{-1}\boldsymbol{\Sigma}^{-1}\right) \otimes \boldsymbol{\Sigma}^{-1}\right\}\mathbf{E}_p \end{pmatrix} \end{aligned}$$

The Fisher information of $\boldsymbol{\phi}$ is hence, $\mathbf{J}_{\boldsymbol{\phi}} = \mathbf{G}^T\mathbf{J}_{\mathbf{g}}\mathbf{G}$, where $\mathbf{G}$ is

$$\mathbf{G} = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \boldsymbol{\Gamma}\boldsymbol{\Omega} & \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \mathbf{I}_p & (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma})\mathbf{E}_u & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \tag{A18}$$

The asymptotic covariance of envelope estimator $\widehat{\mathbf{h}}$ is

$$\mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T = \mathbf{H}\mathbf{J}_{\boldsymbol{\phi}}^{\dagger}\mathbf{H}^T, \tag{A19}$$

where

$$\mathbf{H} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\phi}} = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_p - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \tag{A20}$$

Following Cook et al. (2010), we have the following transformation,

$$\mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_1)^{\dagger}\mathbf{H}_1^T, \tag{A21}$$

where $\mathbf{H}_1$ satisfies $\mathrm{span}(\mathbf{H}_1) = \mathrm{span}(\mathbf{H})$ and is given in the following,

$$\begin{aligned}\mathbf{H}_1 &= \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & 0 & 0 \\ 0 & 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_p(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix} \\ &\equiv \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} \end{pmatrix}.\end{aligned}$$

We next calculate each block in $\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1$ as follows,

$$\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1 = \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\Omega} & 0 & (\boldsymbol{\eta}^T \otimes \mathbf{I}_u)\mathbf{E}_u & 0 \\ 0 & \mathbf{H}_{12}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{12} & 0 & 0 \\ \mathbf{E}_u^T(\boldsymbol{\eta} \otimes \mathbf{I}_u) & 0 & \mathbf{E}_u^T\{(\boldsymbol{\eta}\boldsymbol{\eta}^T + (2\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1}) \otimes \boldsymbol{\Omega}^{-1}\}\mathbf{E}_u & 0 \\ 0 & 0 & 0 & \mathbf{H}_{14}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{14} \end{pmatrix}. \tag{A22}$$

Since the top blocks of $\mathbf{H}_{13}$ and $\mathbf{H}_{14}$ are both zero, they have no contribution to the asymptotic covariance of $\widehat{\boldsymbol{\beta}}$ (top block of $\mathbf{h}$). We hence have,

$$\mathbf{V}_{\boldsymbol{\beta}} = \boldsymbol{\Gamma}\{(\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1)^{-1}\}_{11}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \mathbf{I}_p)(\mathbf{H}_{12}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{12})^{-1}(\boldsymbol{\eta} \otimes \mathbf{I}_p), \tag{A23}$$

where $\{(\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1)^{-1}\}_{11}$ is the top left $p \times p$ block of $(\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1)^{-1}$.

The first term in (A23) can be calculated as,

$$\begin{aligned}\{(\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1)^{-1}\}_{11} &= (\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1} \\ &+ (\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1}(\boldsymbol{\eta}^T \otimes \mathbf{I}_u)\mathbf{E}_u\{(2\pi_1\pi_2)^{-1}\mathbf{E}_u^T(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_u\}^{-1}\mathbf{E}_u^T(\boldsymbol{\eta} \otimes \mathbf{I}_u)\boldsymbol{\Omega}^{-1} \\ &= (\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1} \\ &+ (\pi_1\pi_2)^{-1}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_u\{(2\pi_1\pi_2)^{-1}\mathbf{E}_u^T(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Omega}^{-1})\mathbf{E}_u\}^{-1}\mathbf{E}_u^T(\boldsymbol{\eta} \otimes \boldsymbol{\Omega}^{-1}) \\ &= (\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\Omega}^{-1})2\mathbf{P}_{\mathbf{E}_u}(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\boldsymbol{\eta} \otimes \boldsymbol{\Omega}^{-1}) \\ &= (\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1} + (\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\boldsymbol{\Omega}^{-1} + (\boldsymbol{\eta}^T \otimes \boldsymbol{\eta}),\end{aligned}$$

therefore,

$$\begin{aligned}\boldsymbol{\Gamma}\{(\mathbf{H}_1^T\mathbf{J}_\mathbf{h}\mathbf{H}_1)^{-1}\}_{11}\boldsymbol{\Gamma}^T &= (\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + (\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}(\boldsymbol{\eta}^T \otimes \boldsymbol{\eta})\boldsymbol{\Gamma}^T \\ &= (\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + (\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\beta}^T \otimes \boldsymbol{\beta},\end{aligned}$$

where the last equality is because of $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\eta}^T\boldsymbol{\Gamma}^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\boldsymbol{\Gamma}\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta}$.

The second term in (A23) $\mathbf{H}_{12}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{12}$ can be calculated as,

$$
\begin{aligned}
& \mathbf{J}_\mathbf{h}\mathbf{H}_{12} \\
= \; & \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\Sigma} & (\boldsymbol{\beta}^T \otimes \mathbf{I}_p)\mathbf{E}_p \\ \mathbf{E}_p^T(\boldsymbol{\beta} \otimes \mathbf{I}_p) & \mathbf{E}_p^T\left\{\left(\boldsymbol{\beta}\boldsymbol{\beta}^T + (2\pi_1\pi_2)^{-1}\boldsymbol{\Sigma}^{-1}\right) \otimes \boldsymbol{\Sigma}^{-1}\right\}\mathbf{E}_p \end{pmatrix} \times \\
\times \; & \begin{pmatrix} \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 \\ 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \end{pmatrix} \\
= \; & \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0 + \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0 \\ \mathbf{E}_p^T(\boldsymbol{\beta}\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \mathbf{E}_p^T\left\{\left(\boldsymbol{\beta}\boldsymbol{\beta}^T + (2\pi_1\pi_2)^{-1}\boldsymbol{\Sigma}^{-1}\right) \otimes \boldsymbol{\Sigma}^{-1}\right\} 2\mathbf{E}_p\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \end{pmatrix} \\
= \; & \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 \\ \mathbf{E}_p^T(\boldsymbol{\beta}\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \mathbf{E}_p^T\left\{\left(\boldsymbol{\beta}\boldsymbol{\beta}^T + (2\pi_1\pi_2)^{-1}\boldsymbol{\Sigma}^{-1}\right) \otimes \boldsymbol{\Sigma}^{-1}\right\} 2\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \end{pmatrix},
\end{aligned}
$$

where the bottom block of the matrix can be simplified as follows. Notice that (Corollary D1 and D2 in Cook et al. (2010)) $2(\mathbf{A} \otimes \mathbf{B})\mathbf{P}_\mathbf{E}(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} + (\mathbf{AD} \otimes \mathbf{BC})\mathbf{K}$ where $\mathbf{E}$ and $\mathbf{K}$ are the expansion matrix and the communication matrix with correct dimensions, and also that $\mathbf{E}^T(\mathbf{A} \otimes \mathbf{A})\mathbf{P}_\mathbf{E} = \mathbf{E}^T(\mathbf{A} \otimes \mathbf{A})$. We have the following derivations,

$$
\begin{aligned}
\mathbf{E}_p^T(\boldsymbol{\beta}\boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})2\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) &= \mathbf{E}_p^T(\boldsymbol{\beta}\boldsymbol{\beta}^T \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \\
&= \mathbf{E}_p^T(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) \\
\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})2\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) &= 2\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \\
&= 2\mathbf{E}_p^T(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0)
\end{aligned}
$$

Therefore, we continue the calculation for $\mathbf{H}_{12}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{12}$ as

$$
\begin{aligned}
& \mathbf{H}_{12}^T\mathbf{J}_\mathbf{h}\mathbf{H}_{12} \\
= \; & \mathbf{H}_{12}^T\pi_1\pi_2 \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 \\ \mathbf{E}_p^T(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) + \mathbf{E}_p^T(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0) \\ +(2\pi_1\pi_2)^{-1}2\mathbf{E}_p^T(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0) \end{pmatrix} \\
= \; & \mathbf{H}_{12}^T \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 \\ \mathbf{E}_p^T(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}) + (\pi_1\pi_2)^{-1}\mathbf{E}_p^T(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0) \end{pmatrix} \\
= \; & \pi_1\pi_2 \begin{pmatrix} \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 \\ 2\mathbf{C}_p(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) \end{pmatrix}^T \times \\
\times \; & \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 \\ \mathbf{E}_p^T(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}) + (\pi_1\pi_2)^{-1}\mathbf{E}_p^T(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1} - \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0) \end{pmatrix}.
\end{aligned}
$$

Direct calculation shows,

$$
\begin{aligned}
\mathbf{H}_{12}^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_{12} &= \pi_1\pi_2(\boldsymbol{\eta}\otimes\boldsymbol{\Gamma}_0^T)(\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Gamma}_0)\\
&+ \pi_1\pi_2 2(\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\mathbf{P}_{\mathbf{E}_p}\times\\
&\times \left\{(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1})+(\pi_1\pi_2)^{-1}(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}_0)\right\}\\
&= \pi_1\pi_2(\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\mathbf{I}_{p-u})+\pi_1\pi_2 2(\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1})\\
&+ 2(\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\mathbf{P}_{\mathbf{E}_p}(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}_0)\\
&= \pi_1\pi_2(\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\mathbf{I}_{p-u})+\pi_1\pi_2(\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)(\boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1})\\
&+ (\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}_0)\\
&= \pi_1\pi_2(\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\mathbf{I}_{p-u})+\pi_1\pi_2(\boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\mathbf{I}_{p-u})\\
&+ (\boldsymbol{\Omega}\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}_0^T-\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)(\boldsymbol{\Gamma}\otimes\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}-\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Gamma}_0)\\
&= \pi_1\pi_2(\boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1})+\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-\mathbf{I}_u\otimes\mathbf{I}_{p-u}-\mathbf{I}_u\otimes\mathbf{I}_{p-u}+\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0\\
&= \pi_1\pi_2(\boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1})+\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-2(\mathbf{I}_u\otimes\mathbf{I}_{p-u})+\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0.
\end{aligned}
$$

Finally, we have the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ as

$$
\begin{aligned}
\mathbf{V}_{\boldsymbol{\beta}} &= \boldsymbol{\Gamma}\{(\mathbf{H}_1^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_1)^{-1}\}_{11}\boldsymbol{\Gamma}^T+(\boldsymbol{\eta}^T\otimes\mathbf{I}_p)(\mathbf{H}_{12}^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_{12})^{-1}(\boldsymbol{\eta}\otimes\mathbf{I}_p),\\
&= (\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}\\
&+ (\boldsymbol{\eta}^T\otimes\mathbf{I}_p)(\mathbf{H}_{12}^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_{12})^{\dagger}(\boldsymbol{\eta}\otimes\mathbf{I}_p),
\end{aligned}
$$

where $\mathbf{H}_{12}^T\mathbf{J}_{\mathbf{h}}\mathbf{H}_{12}=\pi_1\pi_2(\boldsymbol{\Omega}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1})+\boldsymbol{\Omega}\otimes\boldsymbol{\Omega}_0^{-1}-2(\mathbf{I}_u\otimes\mathbf{I}_{p-u})+\boldsymbol{\Omega}^{-1}\otimes\boldsymbol{\Omega}_0$.

Finally, we identify the first part is indeed $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})=(\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+(\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+\boldsymbol{\beta}^T\otimes\boldsymbol{\beta}$. Since $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})=\mathrm{avar}(\sqrt{n}\boldsymbol{\Gamma}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})=\boldsymbol{\Gamma}\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^T$, we need to find the asymptotic distribution of $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$, which is the LDA direction of the reduced variables $\boldsymbol{\Gamma}^T\mathbf{X}$. Notice that, given $\boldsymbol{\Gamma}$, we can let $\mathbf{Z}=\boldsymbol{\Gamma}^T\mathbf{X}\in\mathbb{R}^u$ and have the LDA model of $\mathbf{Z}\mid(Y=k)\sim N(\boldsymbol{\nu}_k,\boldsymbol{\Omega})$, where the discriminant directions are exactly $\boldsymbol{\eta}_k=\boldsymbol{\Omega}^{-1}(\boldsymbol{\nu}_k-\boldsymbol{\nu}_1)$. Hence, the asymptotic property of $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ can be easily obtained by replacing $\boldsymbol{\Sigma}\to\boldsymbol{\Omega}$, $\boldsymbol{\beta}\to\boldsymbol{\eta}$ in the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}$. Specifically, $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})=(\pi_1\pi_2)^{-1}\boldsymbol{\Omega}^{-1}+(\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\boldsymbol{\Omega}^{-1}+\boldsymbol{\eta}^T\otimes\boldsymbol{\eta}$. Therefore, $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})=\boldsymbol{\Gamma}\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})\boldsymbol{\Gamma}^T=(\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+(\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta})\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T+\boldsymbol{\eta}^T\boldsymbol{\Gamma}^T\otimes\boldsymbol{\Gamma}\boldsymbol{\eta}$. The conclusion follows by noticing that $\boldsymbol{\beta}=\boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}=\boldsymbol{\eta}^T\boldsymbol{\Omega}\boldsymbol{\eta}$.

S10 Proof for Propositions 6, 7 and 8

First, for LDA model, the $\sqrt{n}$-consistency of the envelope estimator from maximizing the likelihood-based objective function $L_n(\{\boldsymbol{\mu}_k\}_{k=1}^K, \boldsymbol{\Sigma})$ relies on Shapiro's (1986; Proposition 4.1) results on the asymptotics of over-parameterized structural models. The proof is parallel to the proof of Proposition 4 in Cook and Zhang (2015) for simultaneous envelope and is thus ommitted. Similarly, for QDA model, the likelihood-based objective function $L_n(\{\boldsymbol{\mu}_k\}_{k=1}^K, \{\boldsymbol{\Sigma}_k\}_{k=1}^K)$ satisfies conditions to apply Shapiro's results. Thus we have the similar results. Under normality, the estimators becomes MLE, and thus $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{h}}) = \boldsymbol{\Xi} = \mathbf{J}_{\mathbf{h}}^{-1}$ and $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T$ is always less than or equals to $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{h}})$ because any subspace, including $\mathrm{span}(\mathbf{J}_{\mathbf{h}}^{1/2}\mathbf{H})$, is a reducing subspace of $\mathbf{J}_{\mathbf{h}}^{1/2}\boldsymbol{\Xi}\mathbf{J}_{\mathbf{h}}^{1/2} = \mathbf{I}_p$.

S11 Implementation Details

Recall that the ENDS objective function is

$$\mathrm{F}(\boldsymbol{\Gamma}) = \log|\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}| + \sum_{k=1}^{K}\frac{n_k}{n}\log|\boldsymbol{\Gamma}^T\mathbf{S}_k(\lambda)\boldsymbol{\Gamma}|, \tag{A24}$$

which is to be minimized subject to constraint $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$, and includes the ENDS-LDA and ENDS-QDA objective functions as its special cases. We hence discribe the implementation of algorithms for solving this optimization problem of $\mathrm{F}(\boldsymbol{\Gamma})$. The derivative can be calculated straightforwardly as,

$$\frac{d\mathrm{F}(\boldsymbol{\Gamma})}{d\boldsymbol{\Gamma}} = 2\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\boldsymbol{\Gamma})^{-1} + 2\sum_{k=1}^{K}\frac{n_k}{n}\mathbf{S}_k(\lambda)\boldsymbol{\Gamma}\{\boldsymbol{\Gamma}^T\mathbf{S}_k(\lambda)\boldsymbol{\Gamma}\}^{-1}. \tag{A25}$$

Then based on the explicit forms of the objective function and its derivatives, we can use any off-the-shelf methods to find the optimizer $\widehat{\boldsymbol{\Gamma}}$. To preserve the geometric constraint, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$, we choose to implement the estimation using the sg_min package in Matlab that is a generic computing package for Stiefel and Grassmann manifolds optimization (see Edelman et al. (1998) for more backgrounds). Other more recent methods for optimization with orthogonality constraints (for example, Wen and Yin (2013)) can also be straightforwardly incorporated into our implementation.

When u is bigger than one, the optimization can be expensive and requires good initial value to avoid local minima. Therefore, we adopt the sequential 1D algorithm (Cook and Zhang, 2016) to

Algorithm 1 ENDS optimization based on the 1D algorithm from Cook and Zhang (2016).

Let $\mathbf{g}_j \in \mathbb{R}^p$, $j = 1, \ldots, u$, be the sequential directions obtained. Let $\mathbf{G}_j = (\mathbf{g}_1, \ldots, \mathbf{g}_j)$, let $(\mathbf{G}_j, \mathbf{G}_{0j})$ be an orthogonal basis for $\mathbb{R}^p$ and set initial value $\mathbf{g}_0 = \mathbf{G}_0 = 0$.
For $j = 0, \ldots, u-1$, repeat Step 1 and 2 in the following.

1. Define $\mathbf{A}_j = \mathbf{G}_{0j}^T \mathbf{S}_{\mathbf{X}} \mathbf{G}_{0j}$, $\mathbf{B}_{jk} = \mathbf{G}_{0j}^T \mathbf{S}_k(\lambda) \mathbf{G}_{0j}$, and the objective function for $\mathbf{w} \in \mathbb{R}^{p-j+1}$

$$\phi_j(\mathbf{w}) = \log |\mathbf{w}^T \mathbf{A}_j^{-1} \mathbf{w}| + \sum_{k=1}^{K} \frac{n_k}{n} \log |\mathbf{w}^T \mathbf{B}_{jk} \mathbf{w}|. \tag{A26}$$

2. Solve $\mathbf{w}_{j+1} = \arg\min \phi_j(\mathbf{w})$ subject to $\mathbf{w}^T \mathbf{w} = 1$, then the $(j+1)$-th envelope direction is $\mathbf{g}_{j+1} = \mathbf{G}_{0j} \mathbf{w}_{j+1}$.

The resulting ENDS basis estimator is $\widehat{\mathbf{\Gamma}} = \mathbf{G}_u$.

speed up the computation. The 1D algorithm and its variations (Cook and Zhang, 2017) provides a fast and reliable way to obtain envelope subspaces one direction at a time. It breaks down the full optimization of $\mathrm{F}(\mathbf{\Gamma})$ into u sequential optimizations as summarized in Algorithm 1.

References

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208. S2, S6

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300. S11, 1

Cook, R. D. and Zhang, X. (2017). Fast envelope algorithms. *Statistica Sinica (just-accepted)*. S11

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353. S11

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434. S11