Journal of Machine Learning Research 27 (2026) 1-54 Submitted 1/25; Revised 11/25; Published 5/26

Transfer Learning via Regularized Random-effects Linear Discriminant Analysis

Hongzhe Zhang HONGZHE.ZHANG@PENNMEDICINE.UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

Arnab Auddy AUDDY.1@OSU.EDU
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Hongzhe Li HONGZHE@UPENN.EDU
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Ji Zhu

Abstract

Linear discriminant analysis is a widely used method for classification. However, the high dimensionality of predictors combined with small sample sizes often results in large classification errors. To address this challenge, it is crucial to leverage data from related source models to enhance the classification performance of a target model. This paper proposes a transfer learning approach via regularized random-effects linear discriminant analysis, where the discriminant direction is estimated as a weighted combination of ridge estimates obtained from both the target and source models. Multiple strategies for determining these weights are introduced and evaluated, including one that minimizes the estimation risk of the discriminant vector and another that minimizes the classification error. Utilizing results from random matrix theory, we explicitly derive the asymptotic values of these weights and the associated classification error rates in the high-dimensional setting, where the aspect ratio $\gamma := p/n$ as $p, n \to \infty$, with p representing the predictor dimension and n the sample size. Extensive numerical studies, including simulations, the analysis of the proteomics-based cardiovascular disease risk classification and the lipid traits classification problem with genotype data, demonstrate the effectiveness of the proposed approach.

Keywords: Covariate shift; Estimation risk; High dimension; Prediction risk; Random matrix theory; Regularization.

1. Introduction

Large and diverse data sets are ubiquitous in modern applications, including those in genomics and medical decisions. It is of significant interest to integrate different data sets to obtain more accurate parameter estimates or to make a more accurate prediction or classification of an outcome. The success of a supervised statistical learning method relies on the availability of training data. When data is scarce, choosing an appropriately

©2026 Hongzhe Zhang, Arnab Auddy and Hongzhe Li.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0031.html.

flexible model becomes critical to achieving optimal prediction accuracy. This is a classic illustration of the well-known "bias-variance tradeoff". When building a prediction model for a target population, many auxiliary source data sets may exist and provide additional information for building the model for the target population. Modern techniques in the field of transfer learning (Pan and Yang, 2009; Weiss et al., 2016) aim to exploit these additional information. Given a target problem to solve, transfer learning (Torrey and Shavlik, 2010) aims at transferring the knowledge from different but related samples or studies to improve the learning performance of the target problem. In biomedical studies, some clinical or biological outcomes are hard to obtain due to ethical or cost issues, in which case transfer learning can be leveraged to boost the prediction and estimation performance by effectively utilizing information from related studies. Other relevant approaches include meta-learning (Peng, 2020; Huisman et al., 2021), domain adaptation (Redko et al., 2020; Sun et al., 2015), and, more recently, continual learning (De Lange et al., 2021).

In the high dimensional setting, Li et al. (2022, 2023) developed transfer learning methods for sparse high dimensional regressions and demonstrated that one can improve predictions of gene expression levels using data across different tissues. Such sparse models work well when the true models are sparse and the sample sizes are large. However, there are settings where sparse model assumption may not be valid. In genetics, estimating polygenetic risk scores (PRSs) using genome-wide genotype data (Mak et al., 2017; Torkamani et al., 2018) is an active area of research. Such PRSs can be used in risk stratification, or can be treated as risk factors in population health studies. However, due to very large number of genetic variants but relatively small sample sizes, building a PRS model that accurately predicts the PRS scores is challenging. Alternatively, ridge regression, which does not require the sparseness assumption, but can handle the linkage disequilibrium among the genetic variants, provides a viable method for PRS prediction. Zhang and Li (2023) studied the estimation and prediction of random coefficient ridge regression in the setting of transfer learning, where in addition to observations from the target model, source samples from different but possibly related regression models are available. The informativeness of the source model to the target model can be quantified by the correlation between the regression coefficients.

The method of Zhang and Li (2023) is developed for continuous outcomes. In this paper, we propose a transfer learning framework for regularized linear discriminant analysis (RDA) (Friedman, 1989), which we term TL-RDA. Our model assumes a random classification weights setup, where the means of the covariates between two classes differ by a random quantity, δ, with zero mean and constant variance. The auxiliary and target populations are related through the correlation structure of δ. The TL-RDA framework aims to estimate the Bayes optimal predictor by combining naive RDA estimates from both the auxiliary and target populations through a weighted summation. The weights are designed to minimize the distance between the TL-RDA estimator and the Bayes optimal direction in a high-dimensional setting, where the number of features p grows proportionally to the sample size n in all populations. We derive the explicit asymptotic error rate for TL-RDA and show that it achieves the lowest error rate among all estimators based on weighted summations, including naive RDA using only target population data.

Several other LDA–style transfer learning methods have been proposed, but differ in the way of incorporating the source discriminant information. Yang and Gao (2013) derive the multi-view discriminant vectors through a single generalized-eigenvalue problem that

maximizes the inter-view correlation while reducing domain discrepancy. Wang et al. (2020) imposes a group-sparse Bayesian prior on discriminant weights to select and linearly combine the informative source domains. Han et al. (2020); Liu et al. (2023) reformulate the classical LDA to learn explicitly the domain-invariant projections. They also augment Fisher's between-/within-class scatter ratio with a distribution-alignment regularizer that forces the projected source and target means and covariances to match. However, none of these methods has been analyzed in the high-dimensional random effects regime, and none aggregates source information by explicitly minimizing the target-domain prediction risk; which are two aspects that are central to our proposed TL-RDA.

A separate stream of work chooses to fuse or shrink parameters across populations with different objectives. In multi-task learning frameworks, discriminant directions or regression coefficients are encouraged to be close by adding fused penalties (Price and al., 2014; Okazaki and Aoyagi, 2024; Lozano and Swirszcz, 2012). These methods minimize a joint loss over all tasks, seeking a model that performs reasonably well simultaneously across domains. Komárek and Lesaffre (2010) proposes a related random-effects LDA that pools information through hierarchical priors from a Bayesian perspective. In the domain-adaptation literature, hypothesis-weighting techniques combine pre-trained source classifiers by shrinking the weights toward a simplex that balances average source risk (Mansour et al., 2009). Closer in spirit to our work, Gu et al. (2024) moves a high-dimensional ridge estimate toward multiple source coefficients along geodesic paths, with step-sizes chosen to minimize an asymptotic proxy for the target prediction error. Their analysis, however, is confined to linear regression and assumes access only to fitted source coefficients. A complementary direction is to bypass parameter fusion and instead learn latent features whose label–relationship is invariant across environments (Peters et al., 2016; Arjovsky et al., 2019). Such invariance-seeking objectives inevitably trade off target-specific accuracy, while TL-RDA tailors an optimized discriminant vector for the single target population.

The remainder of the paper is organized as follows. We first provide a detailed problem setup and outline the proposed TL-RDA approach in Section 2. We then introduce various types of weights used to integrate auxiliary information, followed by the technical assumptions required for analyzing the estimator in the context of random matrix theory in Section 3. In Section 4, we present the analysis of TL-RDA in the proportional regime, providing explicit asymptotic expressions for the different weighting schemes, along with their corresponding error rates. Section 5 offers interpretations of these weights and guidance on how users can select the most appropriate weights for their applications. In the first five sections, we have assumed all populations share the same population covariance matrix. Section 6 extends the TL-RDA to a heterogeneous population covariance matrix set up. In Section 7, we evaluate the proposed methods on two tasks: (i) cardiovascular-disease classification using proteomics data from the Chronic Renal Insufficiency Cohort (CRIC), and (ii) binary lipid traits prediction using gene-expression data from the Penn Medicine BioBank (PMBB). Section 8 offers a brief discussion.

2. Transfer Learning via Regularized Discriminant Analysis

We consider the setting of two-class LDA in the setting of transfer learning, where we have data observed from both the target model, indexd by K, and $K-1$ source models, indexed

by $k = 1, \cdots, K-1$. We assume that all models, $k = 1, \cdots, K$, follow the classic two-class Gaussian mixture model. More specifically, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$,

$$y_k \in \{-1, +1\} \quad \mathbb{P}(y_k = \pm 1) = \pi_{\pm 1} \quad (X_k)_i | y_k \sim N(\mu_{y,k}, \Sigma). \tag{1}$$

Here $(X_k)_i, i = 1, \cdots, n_k$ is a $p-$dimensional vector, and for ease of notation, we write $\mathbf{X}_k = ((X_k)_1 (X_k)_2 \ldots (X_k)_{n_k})^\top \in \mathbb{R}^{n_k \times p}$ as a $n_k \times p$ dimensional matrix. For simplicity of notation, we assume that the mixing proportions $\pi_{\pm 1}$ remain the same across populations. In fact, without much loss of generality, we mainly discuss the simpler case $\pi_{-1} = \pi_{+1} = 1/2$. The more general case can be managed in a similar manner without significant technical difficulty (Appendix C). Moreover, in most of the paper, we assume the population covariance matrix are the same for all K populations, and are denoted by Σ. This assumption is relaxed in Section 6 where we allow each population to have their own covariance matrix Σ_k for $k = 1, \ldots, K$.

Under this set up, the Bayes optimal prediction direction for the target population K (Anderson, 1958) is given by,

$$d_{Bayes} := \Sigma^{-1} \delta_K \quad \text{where} \quad \bar{\mu}_K = \frac{\mu_{+1,K} + \mu_{-1,K}}{2} \quad \text{for} \quad \mu_{\pm 1,K} = \bar{\mu}_K \pm \delta_K,$$

and $\delta_K = (\mu_{+1,K} - \mu_{-1,K})/2$. The Bayes prediction for a testing data point x_0 from population K is

$$\widehat{y}_{Bayes}(x_0) = \text{sign}\left[d_{Bayes}^\top (x_0 - \bar{\mu}_K) \right].$$

The regularized discriminant analysis (RDA) classifier uses an empirical version of the unkown Bayes direction. The RDA classifier for population k is a linear classifier

$$\widehat{y}_{RDA,k}(x_0) = \text{sign}\left(\widehat{d}_k^\top x_0 + \widehat{b}_k \right),$$

where we use the plug-in estimates for the population parameters Σ_k and δ_k as follows:

$$\begin{aligned}
\widehat{d}_k &= (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \widehat{\delta}_k \\
\widehat{b}_k &= -\widehat{\delta}_k^\top (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\mu}_{-1,k} + \widehat{\mu}_{+1,k})/2 \\
\widehat{\mu}_{\pm 1,k} &= \frac{2}{n_k} \sum_{i:(y_k)_i = \pm 1} (X_k)_i \quad \widehat{\delta}_k = \frac{\widehat{\mu}_{+1,k} - \widehat{\mu}_{-1,k}}{2} \\
\widehat{\Sigma}_k &= \frac{1}{n_k - 2} \sum_{i=1}^{n_k} [(X_k)_i - \widehat{\mu}_{(y_k)_i,k}][(X_k)_i - \widehat{\mu}_{(y_k)_i,k}]^\top.
\end{aligned}$$

Here $\widehat{\Sigma}_k$ is the usual sample covariance matrix, while $\widehat{\mu}_{\pm 1,k}, \widehat{\delta}_k$ are simple estimators and the population-level counterpart $\mu_{\pm 1,k}, \delta_k$. The sample covariance matrix is penalized on its diagonal to overcome the overwhelming variances in this estimation when p grows proportionally with n. As the penalization parameter λ_k goes to zero or infinity, this classifier recover the Fisher's discriminant analysis or the naive Bayes method (Bickel and Levina, 2004).

To utilize the underlying relatedness of the RDA problems in K populations, for a vector of weights $\mathbf{w} \in \mathbb{R}^K$, we now define a classifier based on a weighted linear combination of the

population specific discriminator vectors $\widehat{d}_k$s as follows:

$$\widehat{d}(\mathbf{w}) = \sum_{k=1}^{K} w_k \widehat{d}_k.$$

Note that taking the co-ordinate basis vectors as weights, i.e., $\mathbf{w} = \mathbf{e}_k \in \mathbb{R}^K$, we end up with the population specific discriminant directions $\widehat{d}(\mathbf{e}_k) = \widehat{d}_k$ for $k = 1, \ldots, K$. Now using the weighted combination discriminant vector we define the *transfer-learning* (TL) classifier

$$\widehat{y}_{\mathbf{w}}(x_0) = \text{sign}\left(\widehat{d}(\mathbf{w})^\top x_0 + \widehat{b}_K\right) \tag{2}$$

where $x_0 \in \mathbb{R}^p$ is a test point in the target population. We call this TL-RDA. We use the regular intercept term $\widehat{b}_K$ as all intercept terms are asymptotically zero in the regime considered.

We formulate two criteria to be optimized over $\mathbf{w}$. Let us recall the Bayes classification direction d_{Bayes}. Now for a testing data pair (x_0, y_0) sampled from population K, we consider the two errors:

$$\left\| d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k \right\|_2^2, \tag{3}$$

$$\mathbb{E}_{x_0}\left[(d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k)^\top x_0 \right]^2. \tag{4}$$

The first criteria compares the transfer-learning classifier (TL-RDA) with the Bayes optimal classifier in terms of the estimation error, while the second criteria compares the two classifiers in terms of their prediction errors. In later sections, we present explicit solutions for optimal $\mathbf{w}$ that optimizes the two criteria above. Though none of (3) or (4) is explicitly related to prediction error rate, we show the optimal weight that minimizes (4) also minimizes the prediction error rate in the target population. We will also show the weights that minimizes these two criteria, are in fact related to one another.

In addition, we assume that the population covariance matrix are the same for all K populations (assumption 1). With this in mind, a natural way to further exploit mutual information across populations is to use a pooled sample covariance for all discriminant directions $\widehat{d}_k$. We write $\widehat{\Sigma}_P$ to denote the pooled sample covariance matrix

$$\widehat{\Sigma}_P = \sum_{k=1}^{K} \sum_{i=1}^{n_k} [(X_k)_i - \widehat{\mu}_{y_i,k}][(X_k)_i - \widehat{\mu}_{y_i,k}]^\top / \sum_{k=1}^{K} (n_k - 2).$$

We then define the pooled classification weights and pooled transfer learning regularized discriminant analysis (TLP-RDA) estimator as

$$\widehat{y}^P_{\mathbf{w}}(x_0) = \text{sign}\left(\widehat{d}^P(\mathbf{w})^\top x_0 + \widehat{b}_K\right) \tag{5}$$

where the direction estimates $\widehat{d}_k^P$ are now estimated using the pooled covariance matrix as follows

$$\widehat{d}_k^P := (\widehat{\Sigma}_P + \lambda_k \mathbb{I}_p)^{-1} \widehat{\delta}_k \quad \text{and} \quad \widehat{d}^P(\mathbf{w}) := \sum_{k=1}^K w_k \widehat{d}_k^P.$$

Once again using (3) and (4), we can optimize $\mathbf{w}$ in $\widehat{y}_{\mathbf{w}}^P(x_0)$ with respect to the two criteria above as well, and we will discuss how to choose between $\widehat{y}_{\mathbf{w}}^P(x_0)$ and $\widehat{y}_{\mathbf{w}}(x_0)$.

This form of TL estimator (2) has been used in Zhang and Li (2023) and Dobriban and Sheng (2020) for regression methods. Helm et al. (2024) also considers a TL estimator that aggregates auxiliary information by weighted summations of individual discriminant directions. However, their aggregation is less adaptive in the form $\widehat{d}(\mathbf{w})^* = a\,\widehat{d}_K + (1-a)\sum_{k=1}^{K-1} \widehat{d}_k/(K-1)$ where a is a constant. In addition, the final weight a minimizes only empirical quantities and does not guarantee to minimize population-level criteria such as (3) or (4).

3. Random-effects RDA and the Assumptions on Population Parameters

To find the weights that minimize the estimation or prediction risk defined as 3 and 4, we consider random-effects RDA where the population means are random and are potentially correlated across different models. In addition, we also make the random matrix assumptions on the covariance matrix, which allows us to apply the recent advances in random matrix theory to derive the limiting values of the weights and the corresponding classification errors.

3.1 Classification Weights

A key parameter in the two-class classification problem is the classification weights δ_k that separate the two classes. In this paper, we consider a random classification weights set up formalized below.

Assumption 1 (Random Classification Weights) *The following conditions hold for populations $k = 1, \cdots, K$.*

1. *The class-specific population mean vectors $\mu_{-1,k}, \mu_{+1,k} \in \mathbb{R}^p$ are randomly generated as $\mu_{\pm 1,k} = \bar{\mu}_k \pm \delta_k$, where each δ_k has i.i.d. coordinates with*

$$\mathbb{E}((\delta_k)_i) = 0 \quad \mathrm{Var}((\delta_k)_i) = \alpha_k^2/p \quad \mathbb{E}(|(\delta_k)_i|^{4+\eta}) \leq \frac{C}{p^{2+\eta}}$$

   for fixed $C, \eta > 0$.

2. *$\bar{\mu}_k$ are either fixed or randomly distributed independent of $\delta_k, \mathbf{X}_k, y_k$. The second moment of $\bar{\mu}_k$ are bounded almost surely such that $\limsup_{p\to\infty} \|\bar{\mu}_k\|/p^{1/2-\xi} \leq C$ for some constants $\xi, C > 0$.*

The parameter α_k^2 here plays the role of signal strength. This key assumption asserts that all coordinates of δ_k have the same zero mean and diminishing variance, therefore, all coordinates of an observation play equally important roles on determining the response class. This assumption also used by Dobriban and Wager (2018), is standard in the large

n, large p regime. In addition, the bounded moment assumptions allow use to circumvent difficulties arise when minimizing criteria (3) and (4) which depends on unknown quantities such as δ_k, Σ^{-1} by invoking Lemma A.5. The performance of the TL estimators clearly depends on how the K populations are related. We further pose the correlated classification weight assumption below.

Assumption 2 (Correlated Classification Weights) *For $k = 1, \cdots, K$ the vectors δ_k are correlated across populations such that $\mathrm{Corr}(\delta_k, \delta_{k'}) = \rho_{kk'}\mathbb{I}_p, k \neq k'$.*

For most of the paper, we assume that the parameters α_k^2 and $\rho_{k,k'}$ are known constants. We discuss consistent estimators for these parameters in Appendix B.3.

3.2 Random Matrix Assumption and Related Results

The sample covariance matrix is an important part of the transfer learning estimator. We consider the Marchenko-Pastur type sample covariance matrices as in several relevant works (Dobriban and Wager, 2018; Zhao et al., 2023; Zhang and Li, 2023). For a symmetric matrix, we can characterize its spectral distribution by a cumulative distribution function that places equal point mass on its eigenvalues. Our asymptotic analysis is based on the convergences of spectral distributions of sample covariance matrices, for which the following assumptions are required.

Assumption 3 (RMT assumption) *For $k = 1, \cdots, K$, the design matrix $\mathbf{X}_k \in \mathbb{R}^{n_k \times p}$ is generated as*

$$\mathbf{X}_k = \left(\mu_{(y_k)_1,k}\, \mu_{(y_k)_2,k}\, \cdots\, \mu_{(y_k)_{n_k},k}\right)^\top + \mathbf{Z}_k \Sigma^{1/2}$$

for a matrix $\mathbf{Z}_k \in \mathbb{R}^{n_k \times p}$ with i.i.d. entries coming from an infinite array. The entries $(Z_k)_{ij}$ of $\mathbf{Z}_k$ satisfy the moment conditions: $\mathbb{E}[(Z_k)_{ij}] = 0$ and $\mathbb{E}[(Z_k)_{ij}^2] = 1$.

1. *The population covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ is deterministic. The observations have unit variance, i.e., $\Sigma_{jj} = 1$ for $j = 1, \cdots, p$.*
2. *The eigenvalues of Σ are uniformly bounded from above and away from zero with constants independent of the dimension p.*
3. *The sequence of spectral distributions $T := T_{\Sigma,p}$ of $\Sigma := \Sigma_p$ converges weakly to a limiting distribution H supported on $[0, \infty)$, called the population spectral distribution (PSD).*

For sample covariance $\widehat{\Sigma} := (\mathbf{X} - \mathbb{E}\mathbf{X})^\top(\mathbf{X} - \mathbb{E}\mathbf{X})/n$ with $\mathbf{X} \in \mathbb{R}^{n\times p}$ generated following the Assumption 3, its empirical spectral distribution (ESD) is the probability measure

$$F^{\widehat{\Sigma}}(\omega) \;=\; \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}\{\lambda_i(\widehat{\Sigma}) \leq \omega\}, \omega \in \mathbb{R},$$

where $\lambda_i(\widehat{\Sigma})$ is the i^{th} eigenvalue of $\widehat{\Sigma}$. The Marchenko-Pastur theorem (Marchenko and Pastur, 1967) claims that $F^{\widehat{\Sigma}}$ converges weakly (in distribution) to a limiting distribution $F_\gamma := F_\gamma(H)$ supported on $[0, \infty)$ with probability 1, where γ is the aspect ratio defined as the limit of $p/n \to \gamma$ as $p, n \to \infty$.

Furthermore, for any distribution G supported on $[0, \infty)$, we define its Stieltjes transform as

$$m_G(z) := \int_{l=0}^{\infty} \frac{dG(l)}{l - z}, \ z \in \mathbb{C} \setminus \mathbb{R}^+. \tag{6}$$

The ESD of sample covariance matrix is uniquely determined by a fixed-point equation for its Stieltjes transform. The limit of the Stieltjes transform for the ESD is given by:

$$m_{F^{\widehat{\Sigma}}}(z) = \mathrm{tr}\{(\widehat{\Sigma} - z\mathbb{I}_p)^{-1}/p\} \to_{a.s.} m_{F_\gamma}(z). \tag{7}$$

This can be understood by noting that the Stieltjes transform is a smooth, continuous functional of the ESD and the ESD converges weakly to F_γ.

We continue to define the Stieltjes transform of the limiting spectral distribution of $\overline{\Sigma} := (\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})^\top/p$ as $v_{F_\gamma}(z)$, called companion Stieltjes transform. For all $z \in \mathbb{C} \setminus \mathbb{R}^+$, the Stieltjes transform $v_{F_\gamma}(z)$ is related to $m_{F_\gamma}(z)$ by

$$\gamma \left[m_{F_\gamma}(z) + \frac{1}{z} \right] = v_{F_\gamma}(z) + \frac{1}{z} \tag{8}$$

where γ is the limiting aspect ratio. In addition, we denote by $m'_F(-\lambda)$ the derivative of the Stieltjes transform $m_F(z)$ evaluated at $z = -\lambda$, where

$$m'_{F_\gamma}(z) = \int_{l=0}^{\infty} \frac{dG(l)}{(l - z)^2} \quad v'_{F_\gamma}(z) = \gamma \left(m'_{F_\gamma}(z) - \frac{1}{z^2} \right) + \frac{1}{z^2}. \tag{9}$$

In terms of the empirical quantities, similarly we have

$$m'_{F^{\widehat{\Sigma}}}(z) = \mathrm{tr}\{(\widehat{\Sigma} - z\mathbb{I}_p)^{-2}/p\} \to_{a.s.} m'_{F_\gamma}(z). \text{ (Bai and Silverstein, 2010)}$$

These convergences form the bases on which we develop the limiting error rate and the limiting optimal weights according to criteria (3) and (4).

4. Asymptotic Analysis of Weights and Classification Errors

In this section, we present the expressions for the limiting prediction erorrs and the optimal weights. All formulae are compared with simulated data, and in Appendix B, they are demonstrated to be accurate even under small data sizes.

4.1 Classification Error

Under the two-class Gaussian classification model (Assumption 1), the expected test error of the linear classifier TL-RDA under weight $\mathbf{w}$ in target population K can be written as

$$Err(\mathbf{w}) = \pi_{K,-} \Phi \left(\frac{(\widehat{d}(\mathbf{w}))^\top \mu_{-1,K} + \widehat{b}_K}{\sqrt{(\widehat{d}(\mathbf{w}))^\top \Sigma (\widehat{d}(\mathbf{w}))}} \right) + \pi_{K,+} \Phi \left(- \frac{(\widehat{d}(\mathbf{w}))^\top \mu_{1,K} + \widehat{b}_K}{\sqrt{(\widehat{d}(\mathbf{w}))^\top \Sigma (\widehat{d}(\mathbf{w}))}} \right)$$

$$\widehat{d}(\mathbf{w}) := \sum_{k=1}^{K} w_k \widehat{d}_k$$

where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution and $\widehat{d}(\mathbf{w})$ is the transfer learning discriminating vector. A simple proof of this formula is given in the Appendix. In this paper, we assume the balanced class such that $\pi_+ = \pi_-$. The proof techniques generalize immediately to unbalanced cases, which is discussed briefly in Appendix C. The limiting form of $Err(\mathbf{w})$ is given by Theorem 4.1 below. Two technical assumptions are required so Z_k and the spectrum of Σ are well-behaved.

Assumption 4 (Bounded Moment) *Assume for each natural number p, the entries of Z written by $(Z_k)_{ij}$ has uniformly bounded p-th moment. That is, there are constants C_p such that*

$$\mathbb{E}|(Z_k)_{ij}|^p \leq C_p.$$

Assumption 5 (Anisotropic Local Laws) *Define the cumulative distribution function of H as $F(x) = \sum_{i=1}^{p} I(\lambda_i \leq x)/p$, where $I(\cdot)$ is the indicator function. Recall $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of Σ. For a arbitrarily small positive constant $\tau > 0$, we assume*

$$F(\tau) \leq 1 - \tau.$$

We are now in position to state the first main result of this section. The following theorem describes the asymptotic classification error for the regularized transfer learning classifier defined in (2). The result holds under the aforementioned model assumptions.

Theorem 4.1 (Asymptotic Classification Error for TL-RDA) *Suppose that assumptions 1-2 as well as 4 and 5 hold. Then for a fixed $K \geq 2$, as $n_k, p \to \infty, p/n_k \to \gamma_k \in (0, \infty]$ for $1 \leq k \leq K$, we have that the limiting form of $Err(\mathbf{w})$ for a given weight vector $\mathbf{w} \in \mathbb{R}^K$ is given by:*

$$\Phi\left(-\frac{\mathbf{u}^\top \mathbf{w}}{\sqrt{\mathbf{w}^\top \mathcal{A} \mathbf{w}}}\right) \tag{10}$$

where the elements of $\mathbf{u} \in \mathbb{R}^K$ and $\mathcal{A} \in \mathbb{R}^{K \times K}$ are as follows:

$$u_k = \rho_{kK} \alpha_k \alpha_K m_{F_{\gamma_k}}(-\lambda_k) \quad \text{for } k = 1, \ldots, K,$$

and

$$\mathcal{A}_{kk'} = \begin{cases} \alpha_k^2 \left[\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\gamma_k [\lambda_k v_{F_{\gamma_k}}(-\lambda_k)]^2}\right] + \frac{v'_{F_{\gamma_k}}(-\lambda_k) - v^2_{F_{\gamma_k}}(-\lambda_k)}{\lambda_k^2 v^4_{F_{\gamma_k}}(-\lambda_k)} & \text{if } k = k', \\ \rho_{kk'} \alpha_k \alpha_{k'} \mathcal{M}_{kk'} & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. Here $m_{F_{\gamma_k}}(\cdot)$, $v_{F_{\gamma_k}}(\cdot)$ and $v'_{F_{\gamma_k}}(\cdot)$ are the Stieltjes transform, the companion Stieltjes transform and its derivative respectively, as defined in (6)-(9). Moreover for $k, k' = 1, \cdots, K$ let $\mathcal{M}_{kk'}$ be the constants defined as the following limiting quantities:

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} \Sigma]/p \to_{a.s.} \mathcal{M}_{kk'}.$$

Remark 4.2 *The limits of the off-diagonal terms in $\mathcal{A}$ depends on the interplay of sample covariances in separated populations. Similar terms also appear later in the paper when finding the optimal weight vector (see Theorems 4.5, 4.10). Depending on the specific scenario, the term $\mathcal{M}_{kk'}$ has explicit expressions, some of which are presented separately in Lemma A.8.*

Remark 4.3 *When $K = 1$, writing λ_1, γ_1, α_1, $m_{F_{\gamma_1}}(-\lambda_1)$, $v_{F_{\gamma_1}}(-\lambda_1)$, and $v'_{F_{\gamma_1}}(-\lambda_1)$ as $\lambda, \gamma, \alpha, m, v$ and v' respectively yields*

$$\mathcal{A}_{11} = \frac{1}{\lambda^2 v^2}\left[\frac{\alpha^2(v - \lambda v')}{\gamma} + \frac{v'}{v^2} - 1\right],$$

which implies that the limiting misclassification error is given by

$$\mathrm{Err}(w) \to \Phi\left(\frac{\alpha^2 m}{\sqrt{\mathcal{A}_{11}}}\right) = \Phi\left(-\alpha^2 \lambda m v \left\{\frac{\alpha^2(v - \lambda v')}{\gamma} + \frac{v'}{v^2} - 1\right\}^{-1/2}\right).$$

This matches exactly with Theorem 3.1 of Dobriban and Wager (2018).

We now consider the classification error of the transfer learning classifier, when using the pooled covariance matrix. Let us replace $\widehat{d}(\mathbf{w})$ with $\widehat{d}_P(\mathbf{w})$ in $Err(\mathbf{w})$ and define this new function of $\mathbf{w}$ as $Err_P(\mathbf{w})$. This gives us the classification error for TLP-RDA given a weight vector $\mathbf{w}$.

Corollary 4.4 (Asymptotic Classification Error for TLP-RDA) *Under the same set up as Theorem 4.1, we further define $p/\sum_{k=1}^K n_k \to \bar{\gamma}$. Then assuming $\lambda_1 = \cdots = \lambda_K = \lambda$, we have*

$$Err_P(\mathbf{w}) = \Phi\left(-\frac{\mathbf{u}_P^\top \mathbf{w}}{\sqrt{\mathbf{w}^\top \mathcal{A}_P \mathbf{w}}}\right)$$

where the elements of $\mathbf{u}_P \in \mathbb{R}^K$ and $\mathcal{A}_P \in \mathbb{R}^{K\times K}$ are as follows:

$$(u_P)_k = \rho_{kK}\alpha_k\alpha_K m_{F_{\bar{\gamma}}}(-\lambda) \quad \text{for } k = 1, \ldots, K,$$

and

$$(\mathcal{A}_P)_{kk'} = \begin{cases} \alpha_k^2\left[\frac{v_{F_{\bar{\gamma}}}(-\lambda) - \lambda v'_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}[\lambda v_{F_{\bar{\gamma}}}(-\lambda)]^2}\right] + \gamma_k\left[\frac{v'_{F_{\bar{\gamma}}}(-\lambda_k) - v^2_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}\lambda^2 v^4_{F_{\bar{\gamma}}}(-\lambda)}\right] & \text{if } k = k', \\ \rho_{kk'}\alpha_k\alpha_{k'}\left[\frac{v_{F_{\bar{\gamma}}}(-\lambda) - \lambda v'_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}[\lambda v_{F_{\bar{\gamma}}}(-\lambda)]^2}\right] & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. Here $m_{F_{\bar{\gamma}}}(\cdot)$, $v_{F_{\bar{\gamma}}}(\cdot)$ and $v'_{F_{\bar{\gamma}}}(\cdot)$ are the Stieltjes transform, the companion Stieltjes transform and its derivative respectively, as defined in (6)-(9).

In this corollary, we make the assumption that all studies have the same degree of penalization, which brings us the simplified cross terms in $\mathcal{A}_P$ and a more amenable expression for $\mathcal{A}_P$ as a whole. This assumption might not be reasonable when the signal strengths α_k^2 are vastly different, as lighter penalization may be given to population with stronger signal strength. In this case, one can use the general formula in Theorem 4.1.

4.2 Minimum Estimation Risk Weight

We firstly present the way to minimize the coordinate-wise estimation error of $\widehat{d}(W)$ with respect to d_{Bayes}.

Theorem 4.5 (Asymptotic Estimation Error Minimization for TL-RDA) *Suppose that assumptions 1-2 as well as 4 and 5 hold. Then for a fixed $K \geq 2$, as $n_k, p \to \infty, p/n_k \to \gamma_k \in (0, \infty]$ for $1 \leq k \leq K$, the weight for minimizing the error in estimating the Bayes optimal discriminator d_{Bayes} is given by:*

$$\mathbf{w}^E := \arg\min_{\mathbf{w}} \left\| d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k \right\|_2^2 = (\mathcal{A}^E + \mathcal{R}^E)^{-1} \mathbf{u}^E,$$

where the elements of $\mathbf{u}^E \in \mathbb{R}^K$, $\mathcal{A}^E \in \mathbb{R}^{K \times K}$, and $\mathcal{R}^E \in \mathbb{R}^{K \times K}$ are:

$$(u^E)_k = \rho_{kK} \alpha_k \alpha_K \left[\frac{1}{\lambda_k} \mathbb{E}(T^{-1}) - m_{F_{\gamma_k}}(-\lambda_k)^2 \right] \quad \text{for } k = 1, \ldots, K,$$

$$\mathcal{A}^E_{kk'} = \begin{cases} \alpha_k^2 m'_{F_{\gamma_k}}(-\lambda_k) & \text{if } k = k', \\ \rho_{kk'} \alpha_k \alpha_{k'} \mathcal{E}_{kk'} & \text{otherwise,} \end{cases}$$

and

$$\mathcal{R}^E_{kk'} = \begin{cases} \frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)^2} & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. Here $m_{F_{\gamma_k}}(\cdot)$, $v_{F_{\gamma_k}}(\cdot)$ and $v'_{F_{\gamma_k}}(\cdot)$ are the Stieltjes transform, the companion Stieltjes transform and its derivative respectively, as defined in (6)-(9), while T is the limiting spectral distribution of the population covariance matrix Σ. Moreover for $k, k' = 1, \cdots, K$, let $\mathcal{E}_{kk'}$ be the constants defined as the following limiting quantities:

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p \to_{a.s.} \mathcal{E}_{kk'}.$$

Just like the expressions for the asymptotic classification rate in Theorem 4.1, more explicit forms of $\mathcal{E}_{kk'}$ are available if γ_k and λ_k are the same across populations. We call $\mathbf{w}^E$ the minimum estimation weight, as it minimizes the ℓ_2 error in estimating the Bayes optimal discriminating vector d_{Bayes}. In order to better illustrate the effect of transfer learning, we consider next a simpler situation where all the sources are homogeneous, in that $\gamma_1 = \gamma_2 = \cdots = \gamma_K$ and the corresponding regularization parameters are also the same, i.e., $\lambda_1 = \lambda_2 = \cdots = \lambda_K$. We further assume equal correlation among all sources, i.e., $\rho_{kk'} = \rho$ whenever $k \neq k'$.

Corollary 4.6 (Estimation Error for Homogeneous Sources) *Under the setup of Theorem 4.5, suppose $\gamma_1 = \gamma_2 = \cdots = \gamma_K =: \gamma$, $\lambda_1 = \lambda_2 = \cdots = \lambda_K =: \lambda$, and $\rho_{kk'} = \rho$ whenever $k \neq k'$, $1 \leq k, k' \leq K$. Then the weight vector $\mathbf{w}^E$ that minimizes the asymptotic error in estimating d^{Bayes} is given by:*

$$\mathbf{w}^E = \alpha_K \left[\frac{1}{\lambda} \mathbb{E}(T^{-1}) - m_{F_\gamma}(-\lambda)^2 \right] \mathrm{vec}\left(\frac{\{\rho - \xi + (1-\rho) I(k = K)\} \alpha_k}{t_{m,\varepsilon,\lambda} \alpha_k^2 + t_{v,\lambda}} : 1 \leq k \leq K \right)$$

where $\varepsilon = \mathcal{E}_{12}$, $t_{\rho,\varepsilon,\lambda} := -\rho\varepsilon + m'_{F_\gamma}(-\lambda)$, $t_{v,\lambda} := \frac{v_{F_\gamma}(-\lambda) - \lambda v'_{F_\gamma}(-\lambda)}{\lambda v_{F_\gamma}(-\lambda)^2}$ *and*

$$\xi = \frac{\rho\varepsilon \sum_{k=1}^{K} \{\rho + (1-\rho) I(k=K)\} \frac{\alpha_k^2}{t_{m,\varepsilon,\lambda}\alpha_k^2 + t_{v,\lambda}}}{1 + \rho\varepsilon \sum_{k=1}^{K} \frac{\alpha_k^2}{t_{m,\varepsilon,\lambda}\alpha_k^2 + t_{v,\lambda}}}.$$

The above corollary shows that when all sources have equal sample sizes, and equal correlation, the optimal weight is proportional to $\alpha_k \alpha_K / (t_{m,\varepsilon,\lambda}\alpha_k^2 + t_{v,\lambda})$. In particular, if $\alpha_k \to 0$ for some source population, then the corresponding weight for that source increases at the rate of $1/\alpha_k$. This is intuitive since in the homogeneous setting of equal sample sizes, the signal strength is completely determined by the inverse of the variance of δ_k, which is precisely $1/\alpha_k$. For practical usage of the optimal transfer weights, we now turn to the estimation of the above quantities.

Remark 4.7 (Estimating optimal weights) *The Marchenko-Pastur law, along with the equations* (6), (7), (8), *and* (9) *imply that the sample Stieltjes transform* $\mathrm{tr}[(\widehat{\Sigma}_k - z\mathbb{I}_p)^{-1}]$ *can be suitably used to estimate the functions* $m_{F_{\gamma_k}}(\cdot)$, $v_{F_{\gamma_k}}(\cdot)$, *and* $m'_{F_{\gamma_k}}(\cdot)$ *reliably. We remind the reader that the estimation of* α_k^2 *and* $\rho_{kk'}$ *is tackled in Appendix B.3. Finally, we note that the expression of* $\mathbf{u}^E$ *in all the above cases involves the expectation of* T^{-1}, *and we discuss a consistent estimator for* $\mathbb{E}(T^{-1})$ *in the following proposition, the proof of which is immediate from Corollary 4.2 of Dobriban and Sheng (2021).*

The following proposition holds under Assumptions 3.

Proposition 4.8 *When* $\gamma_k < 1$, *we have* $\mathrm{tr}(\widehat{\Sigma}_k^{-1})/p \to_{a.s.} \mathbb{E}(T^{-1})/(1-\gamma_k)$.

Since in our developments so far we have assumed all populations to share the same covariance matrix Σ, one can always use the pooled sample covariance in the estimation scheme above. That means we can reliably estimate $\mathbb{E}(T^{-1})$ as long as $\bar{\gamma} < 1$. We then present the optimal estimation for TLP-RDA, i.e., the transfer learning discriminant analysis done using the pooled covariance matrix.

Corollary 4.9 (Asymptotic Estimation Error Minimization for TLP-RDA) *Under the same set up as Theorem 4.5, assuming* $\lambda_1 = \cdots = \lambda_K = \lambda$, *we have*

$$\mathbf{w}_P^E := \arg\min_{\mathbf{w}} \left\| d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k^P \right\|_2^2 = (\mathcal{A}_P^E + \mathcal{R}_P^E)^{-1} \mathbf{u}_P^E,$$

where the elements of $\mathbf{u}_P^E \in \mathbb{R}^K$, $\mathcal{A}_P^E \in \mathbb{R}^{K\times K}$, *and* $\mathcal{R}_P^E \in \mathbb{R}^{K\times K}$ *are:*

$$(u_P^E)_k = \rho_{kK}\alpha_k\alpha_K \left[\frac{1}{\lambda}\mathbb{E}(T^{-1}) - m_{F_{\bar{\gamma}}}(-\lambda)^2 \right] \quad \text{for } k = 1, \ldots, K,$$

while

$$(\mathcal{A}_P^E)_{kk'} = \rho_{kk'}\alpha_k\alpha_{k'} m'_{F_{\bar{\gamma}}}(-\lambda)$$

and

$$(\mathcal{R}_P^E)_{kk'} = \begin{cases} \gamma_k \frac{v_{F_{\bar{\gamma}}}(-\lambda) - \lambda v'_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}\lambda v_{F_{\bar{\gamma}}}(-\lambda)^2} & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$.

Once again we can follow the remark and proposition above for estimating the weight that minimizes the estimation error when using the pooled covariance matrix.

4.3 Minimum Prediction Risk Weight

In the high dimensional regime considered in this paper, a weight minimizing the estimation error does not translate into maximizing the classification score $\widehat{d}(\mathbf{w})^\top x_0$. We can as well compute the optimal prediction weight by directly minimizing the difference between the TL-RDA classification score and the Bayes classification score. This is precisely the objective of the next theorem.

Theorem 4.10 (Asymptotic Prediction Error Minimization for TL-RDA) *Suppose that Assumptions 1-2 as well as 4 and 5 hold. Then for a fixed* $K \geq 2$, *as* $n_k, p \to \infty, p/n_k \to \gamma_k \in (0, \infty]$ *for* $1 \leq k \leq K$, *the weight for minimizing the excess risk, i.e., the error in predicting the class at a random test point* x_0, *when compared to the Bayes optimal discriminator* d_{Bayes}, *is given by:*

$$\mathbf{w}^P := \arg\min_{\mathbf{w}} \mathbb{E}_{x_0} \left[(d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k)^\top x_0 \right]^2 = (\mathcal{A}^P + \mathcal{R}^P)^{-1} \mathbf{u}^P,$$

where the elements of $\mathbf{u}^P \in \mathbb{R}^K$, $\mathcal{A}^P \in \mathbb{R}^{K \times K}$, *and* $\mathcal{R}^P \in \mathbb{R}^{K \times K}$ *are:*

$$(u^P)_k = \rho_{kK} \alpha_k \alpha_K m_{F_{\gamma_k}}(-\lambda_k) \quad \text{for } k = 1, \ldots, K,$$

$$\mathcal{A}^P_{kk'} = \begin{cases} \alpha_k^2 \left[\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\gamma_k [\lambda_k v_{F_{\gamma_k}}(-\lambda_k)]^2} \right] & \text{if } k = k', \\ \rho_{kk'} \alpha_k \alpha_{k'} \mathcal{M}_{kk'} & \text{otherwise,} \end{cases}$$

and

$$\mathcal{R}^E_{kk'} = \begin{cases} \frac{v'_k(-\lambda_k) - v_k^2(-\lambda_k)}{\lambda_k^2 v_k^4(-\lambda_k)} & \text{if } k = k' \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. *Here* $m_{F_{\gamma_k}}(\cdot)$, $v_{F_{\gamma_k}}(\cdot)$ *and* $v'_{F_{\gamma_k}}(\cdot)$ *are the Stieltjes transform, the companion Stieltjes transform and its derivative respectively, as defined in* (6)-(9). *Moreover for* $k, k' = 1, \cdots, K$, *let* $\mathcal{M}_{kk'}$ *be the constants defined as the following limiting quantities:*

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} \Sigma]/p \to_{a.s.} \mathcal{M}_{kk'}.$$

As in the case of estimation error weights, we discuss the issue of estimating the optimal weight that minimizes prediction error in this setting. Note that unlike the estimation error minimizing weight $\mathcal{W}^E$, the prediction risk minimizing weight $\mathcal{W}^P$, involves no population spectral distribution. It is thus estimable in all cases, including the case when $\gamma > 1$. On the other hand, estimating $\mathcal{M}_{kk'}$ requires more care, especially in the case $k = K$ or $k' = K$. The following proposition describes the estimation in this case.

Proposition 4.11 *We have the following consistent estimators $\widehat{\mathcal{M}}_{kk'}$ for $\mathcal{M}_{kk'}$ separately for three cases:*

$$\widehat{\mathcal{M}}_{kk'} = \begin{cases} \operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\widehat{\Sigma}_K]/p & \text{if } k \neq k', k \neq K, k' \neq K \\ \operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-2}\widehat{\Sigma}_K]/p & \text{if } k = k' \neq K \\ \frac{1}{px_p}\operatorname{tr}[(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}] - \frac{\lambda_K}{px_p}\operatorname{tr}[(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}] & \text{if } k = K,\ k' \neq K \end{cases}$$

where $x_p = x(\gamma_K, \lambda_K)$ is the solution to the equation:

$$1 - x_p = \gamma_K \left[1 - \lambda_K \int (x_p t + \lambda_K)^{-1} dH_K(t)\right]$$

Then

$$\mathcal{M}_{kk'} - \widehat{\mathcal{M}}_{kk'} \to_{a.s.} 0.$$

In order to better illustrate the effect of transfer learning, we consider next a simpler situation where all the sources are homogeneous, in that $\gamma_1 = \gamma_2 = \cdots = \gamma_K$ and the corresponding regularization parameters are also the same, i.e., $\lambda_1 = \lambda_2 = \cdots = \lambda_K$. We further assume equal correlation among all sources, i.e., $\rho_{kk'} = \rho$ whenever $k \neq k'$.

Corollary 4.12 (Prediction Error for Homogeneous Sources) *Under the earlier setup of Theorem 4.5, suppose $\gamma_1 = \gamma_2 = \cdots = \gamma_K =: \gamma$, $\lambda_1 = \lambda_2 = \cdots = \lambda_K =: \lambda$, and $\rho_{kk'} = \rho$ whenever $k \neq k'$, $1 \leq k, k' \leq K$. Then the weight vector $\mathbf{w}^P$ that minimizes the asymptotic excess risk is given by:*

$$\mathbf{w}^P = \alpha_K m_{F_\gamma}(-\lambda)\operatorname{vec}\left(\frac{\{\rho - \xi^P + (1-\rho)I(k=K)\}\alpha_k}{t^P_{v,\rho,\lambda}\alpha_k^2 + t^P_{v,\lambda}} : 1 \leq k \leq K\right)$$

where $m = \mathcal{M}_{12}$, $t^P_{v,\rho,\lambda} := -m\rho + \frac{v_{F_\gamma}(-\lambda) - \lambda v'_{F_\gamma}(-\lambda)}{\gamma[\lambda v_{F_\gamma}(-\lambda)]^2}$, $t^P_{v,\lambda} := \frac{v'(-\lambda) - v^2(-\lambda)}{\lambda^2 v^4(-\lambda)}$ and

$$\xi^P = \frac{m\rho \sum_{k=1}^K \{\rho + (1-\rho)I(k=K)\}\frac{\alpha_k^2}{t^P_{v,\rho,\lambda}\alpha_k^2 + t^P_{v,\lambda}}}{1 + m\rho \sum_{k=1}^K \frac{\alpha_k^2}{t^P_{v,\rho,\lambda}\alpha_k^2 + t^P_{v,\lambda}}}.$$

The above corollary shows that when all sources have equal sample sizes, and equal correlation, the optimal weight is proportional to $\alpha_k \alpha_K/(t^P_{v,\rho,\lambda}\alpha_k^2 + t^P_{v,\lambda})$. Thus similar to Corollary 4.6, if $\alpha_k \to 0$ for some source population, then the corresponding weight for that source increases at the rate of $1/\alpha_k$. This is intuitive since in the homogeneous setting of equal sample sizes, the signal strength is completely determined by the inverse of the variance of δ_k, which is precisely $1/\alpha_k$. Finally, since our assumption so far posits the same variance Σ for all populations, we also describe the estimation based on the pooled sample covariance matrix. This is given in the following corollary for the pooled covariance based classifier TLP-RDA.

Corollary 4.13 (Asymptotic Prediction Error Minimization for TLP-RDA) *Under the same set up as theorem 4.10, assume $\lambda_1 = \cdots = \lambda_K = \lambda$, we have*

$$\mathbf{w}_P^P := \arg\min_{\mathbf{w}} \mathbb{E}_{x_0} \left[(d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k^P)^\top x_0 \right]^2 = (\mathcal{A}_P^P + \mathcal{R}_P^P)^{-1} \mathbf{u}_P^P$$

where the elements of $\mathbf{u}_P^P \in \mathbb{R}^K$, $\mathcal{A}_P^P \in \mathbb{R}^{K\times K}$, and $\mathcal{R}_P^P \in \mathbb{R}^{K\times K}$ are:

$$(u_P^P)_k = \rho_{kK}\alpha_k\alpha_K m_{F_{\bar{\gamma}}}(-\lambda) \quad \text{for } k = 1, \ldots, K,$$

while

$$(\mathcal{A}_P^P)_{kk'} = \rho_{kk'}\alpha_k\alpha_{k'} \left[\frac{v_{F_{\bar{\gamma}}}(-\lambda) - \lambda v'_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}[\lambda v_{F_{\bar{\gamma}}}(-\lambda)]^2} \right]$$

and

$$(\mathcal{R}_P^P)_{kk'} = \begin{cases} \gamma_k \frac{v'_{F_{\bar{\gamma}}}(-\lambda_k) - v^2_{F_{\bar{\gamma}}}(-\lambda)}{\bar{\gamma}\lambda^2 v^4_{F_{\bar{\gamma}}}(-\lambda)} & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$.

The optimal prediction weights $\mathbf{w}_P^E, \mathbf{w}_P^P$ dominate $\mathbf{w}^E, \mathbf{w}^P$ respectively in classification error, meaning

$$Err(\mathbf{w}^E) \geq Err(\mathbf{w}^P), Err_P(\mathbf{w}_P^E) \geq Err_P(\mathbf{w}_P^P).$$

Note that a priori it is not immediate that the weight which minimizes the difference in prediction error from the Bayes optimal direction, also minimizes the classification error. In particular, the accuracy of classification depends on the sign of $\widehat{d}(\mathbf{w})^\top x_0$, and not the actual value. However, we now show that in addition to optimizing the score $\widehat{d}(\mathbf{w})^\top x_0$, the prediction weights $\mathbf{w}^P$ and $\mathbf{w}_P^P$ also minimize the misclassification error. This is formalized by the following proposition.

Proposition 4.14 *The optimal prediction weight minimizes the testing data classification error:*

$$\mathbf{w}^P = \arg\min_{\mathbf{w}} \mathbb{E}_{x_0,y_0}[I(\text{sign}[(\widehat{d}(\mathbf{w}))^\top x_0] \neq y_0)],$$
$$\mathbf{w}_P^P = \arg\min_{\mathbf{w}} \mathbb{E}_{x_0,y_0}[I(\text{sign}[(\widehat{d}_P(\mathbf{w}))^\top x_0] \neq y_0)].$$

We conclude this section by validating the existence of the optimal weights. Indeed, the solution to all four types of weights implicitly assume that the matrices $\mathcal{A}^E + \mathcal{R}^E, \mathcal{A}^P + \mathcal{R}^P, \mathcal{A}_P^E + \mathcal{R}_P^E, \mathcal{A}_P^P + \mathcal{R}_P^P$ are invertible. We prove this is always the case in the following proposition.

Proposition 4.15 (Existence of Optimal Weights) *The matrices $\mathcal{A}^E+\mathcal{R}^E, \mathcal{A}^P+\mathcal{R}^P, \mathcal{A}_P^E+\mathcal{R}_P^E, \mathcal{A}_P^P+\mathcal{R}_P^P$ are invertible, and hence the limiting optimal weights $\mathbf{w}^E, \mathbf{w}^P, \mathbf{w}_P^E, \mathbf{w}_P^P$ exist.*

We compare the error rates of the four transfer learning estimators under several different scenarios in Appendix B.

4.4 Geometric Interpretation

We now provide some geometric interpretations on the optimal weights derived in the previous subsections. Let us define the following discriminant directions:

$$d_{est} := \widehat{d}(\mathbf{w}_E), \quad d_{err} := \widehat{d}(\mathbf{w}_P).$$

Also recall the Bayes discriminant direction d_{Bayes}. These discriminant directions are visualized in Figure 1. The blue plane is the space spanned by linear combinations of the local discriminant directions $\{\widehat{d}_k\}$, and the top black line stands for the Bayes direction, which is not necessarily in the linear span of $\{\widehat{d}_k\}$. Since both the TL-RDA directions d_{est} and d_{err} are linear combinations of $\widehat{d}_k$, they both lie on the blue plane and are denoted by colored lines. The direction obtained by minimizing the error in estimating d_{Bayes} is given by d_{est}, which is the projection of d_{Bayes} onto the blue plane, denoting span$\{\widehat{d}_k\}$. The criterion 3 suggests that d_{est} is the minimizer of the OLS loss when fitting d_{Bayes} with linear combinations of d_k.

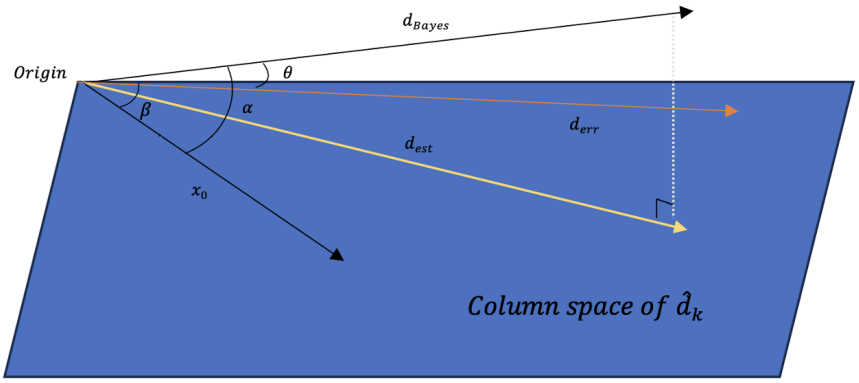

Figure 1: Geometric interpretations of optimal weights

Denoting the angle between d_{Bayes} and $\widehat{d}(\mathbf{w})$ as θ, we now show that the cosine of θ is directly related to $Err(\mathbf{w})$ accounting for the scaling Σ. Let us define the scaled inner product $\langle a, b\rangle_\Sigma = a^\top \Sigma b$ and the scaled cos angle $\cos_\Sigma(a, b) = \langle a, b\rangle_\Sigma / \sqrt{\langle a, a\rangle_\Sigma \langle b, b\rangle_\Sigma}$. Then we have

$$\cos\theta := \cos_\Sigma \angle(\widehat{d}(\mathbf{w}), d_{Bayes}) = \widehat{d}(\mathbf{w})^\top \delta_K / \sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w}) \delta_K^\top \Sigma^{-1} \delta_K} = \frac{\Theta(\mathbf{w})}{\Theta_{Bayes}}$$

$$\Theta(\mathbf{w}) := \frac{\widehat{d}(\mathbf{w})^\top \delta_K}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}}, \quad \Theta_{Bayes} := \sqrt{\delta_K^\top \Sigma^{-1} \delta_K}$$

Recall that $\Phi(-\Theta(\mathbf{w}))$ is the classification error rate of TL-RDA (Theorem 4.1) and $\Phi(-\Theta_{Bayes})$ is the Bayes error rate. This implies that

$$\cos\theta = \frac{\Theta(\mathbf{w})}{\Theta_{Bayes}} = \frac{\Phi^{-1}(Err(\mathbf{w}))}{\Phi^{-1}(Err_{Bayes})}.$$

Since $\Phi^{-1}(\cdot)$ is a monotonically increasing function, it is clear that $\cos(\theta)$ is close to one, i.e., θ is close to zero, if and only if $Err(\mathbf{w})$ is close to the Bayes error Err_{Bayes}. Therefore, the size of θ directly quantifies the inefficiency of TL-RDA relative to d_{Bayes} in terms of classification error, and $\mathbf{w}^P$ simultaneously minimizes the classification error and $\cos\theta$ (Proposition 4.14).

Finally, let us also consider an observation x_0 pointing in a random direction, and denote the angle between d_{err} and x_0 as β and the angle between d_{Bayes} and x_0 as α. Straightforwardly, d_{err} minimizes the difference between β and α as $\mathbf{w}^P$ minimizes the difference between the inner products $d_{err}^\top x_0$ and $d_{Bayes}^\top x_0$.

5. Robustness and Weight Selection

In this section, we present some guidance on how one can choose between different weighting schemes. As we have repeatedly illustrated in the previous section, the d_{est} is dominated by d_{err} in terms of classification error on unseen data. This claim, however, holds only when the test data distribution is as given by Assumption 1. The first goal of this section is to demonstrate that d_{est} outperforms d_{err} considerably in terms of classification error, when there is a distribution shift in the covariates. The second goal is to compare between the individual vs. pooled covariance estimators: i.e., TL-RDA and TLP-RDA. We will prove, at least in special cases, that the optimal TLP-RDA outperforms the optimal TL-RDA when the aspect ratio is large enough.

5.1 Robustness of Optimal Estimation Weight

In this section we demonstrate the robustness of d_{est} to covariate shifts in test data. This robustness is intuitive as we know the optimal estimation weight only attempts to minimize the difference in TL discriminant direction ($\widehat{d}(\mathbf{w})$) and the Bayes discriminant direction d_{Bayes}. In fact, we can show the weights obtained by minimizing criteria (3) are equivalent to a conservative solution to the problem of minimizing criteria (4) when test data distribution is unknown. A similar argument is in Zhang and Li (2023), we summarize it in the following proposition.

Proposition 5.1 (Robustness of Estimation Weight) *For the class of covariate distributions given by* $\mathcal{P} := \{P : x \sim P, \mathbb{E}_P(\|x\|_2) \leq c\}$*, we have:*

$$\arg\min_{\mathbf{w}} \left\| d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k \right\|^2 = \arg\min_{\mathbf{w}} \max_{x_0 \in \mathcal{P}} \mathbb{E}_{x_0} \left[\left(d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k \right)^\top x_0 \right]^2 .$$

Proposition 5.1 claims that, when we only know the target data comes from a distribution with bounded expected norm, the safest option to minimize prediction error criteria (4) is to focus only on the estimation error criteria (3). We demonstrate the usefulness of this in Figure 2. Here $p = 150$ and $n_k = 250, \cdots, 160$ for study 1 to study K. The pairwise correlations across studies are fixed at 0.5. We use the same Toeplitz covariance matrix for all training data, however, the eigenvalues of the test data x_0 covariance matrix are modified to decay much faster. The testing accuracy of the three methods are shown on the y axis as λ changes. One can see the optimal estimation weight consistently outperforms others, suggesting the robustness and conservative nature of the optimal estimation weight boost the performance of TL-RDA when there is a change in testing data distribution.

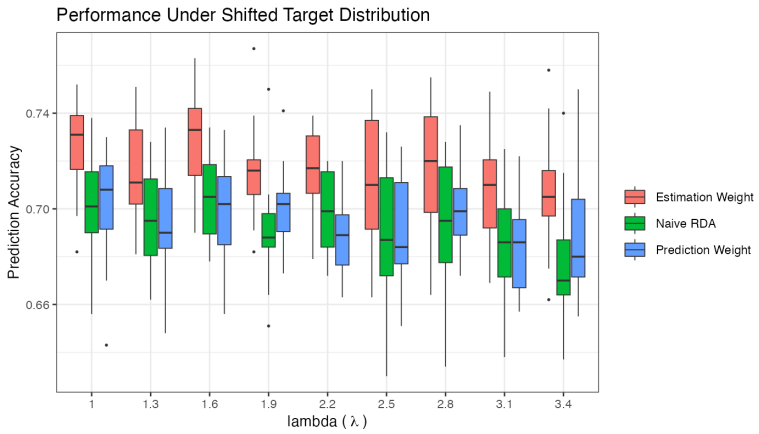

Figure 2: TL-RDA with the optimal estimation weight outperforms regular RDA and TL-RDA with the optimal prediction weight when the target distribution changes.

5.2 Pooled Sample Covariance and Individual Sample Covariance Matrix

For TLP-RDA, all discriminant directions uses the same covariance estimate $\widehat{\Sigma}_P$. Although $\widehat{\Sigma}_P$ is a better covariance estimate than all $\widehat{\Sigma}_k$, $\widehat{d}_k^P$ are inevitably similar and the columns space of $\widehat{d}_k^P$ would be less informational. This essentially becomes a bias-variance trade off. When γ_k is small, the estimates $\widehat{d}_k$ are reliable already. In this case, individual covariance matrices bring more variances to this column space, therefore, increase the quality of final TL-RDA estimator. When γ_k are large, direction estimate based on $\widehat{\Sigma}_k$ are no longer reliable and one should consider the more stable $\widehat{\Sigma}_P$. We will formalize this statement in this section also.

Proposition 5.2 (TLP-RDA out performs TL-RDA when γ is large) *Assume* $\Sigma = \mathbb{I}_p$, $\gamma_1 = \cdots = \gamma_K = \gamma$, $\lambda_1, \cdots, \lambda_K = \lambda = r\left(\gamma - \frac{1}{r+1}\right)$ *for some fixed* $r > (1-\gamma)_+/\gamma$. *For the pooled covariance matrix, we choose* $\lambda' = r'\left(\gamma/K - \frac{1}{r'+1}\right)$ *for some* $r' > (K-\gamma)_+/\gamma$.

1. *When* $\rho = 1$, $Err(\mathbf{w}^P) \geq Err_P(\mathbf{w}_P^P)$, *if and only if*

$$\gamma^2[(1+r')^2 - K(1+r)^2] \geq K\left([\gamma(1+r)^2 - 1]\sum_k \alpha_k^2\right).$$

2. *When* $\rho = 0$, $Err(\mathbf{w}^P) \geq Err_P(\mathbf{w}_P^P)$, *if and only if*

$$\gamma[(1+r')^2 - (1+r)^2] \geq K - 1.$$

We also numerically demonstrate this phenomenon by plugging values into the limiting expressions under a more general set up (Figure 3). We again use $\rho = 0.5$ while changing the other parameters, including the number of auxiliary studies K, the decay rate of the eigenvalues of Σ and the signal strength α. We can see that TLP-RDA can outperform TL-RDA in all cases as γ_K grows, as the error rates of TLP-RDA decrease at slower rates.

The transition point γ^*, defined as the γ_K when TLP-RDA outperforms TL-RDA, differs in different scenarios. One can see that it decreases when the eigenvalues of Σ decrease slower, and interestingly, also when the number of auxiliary population decreases.

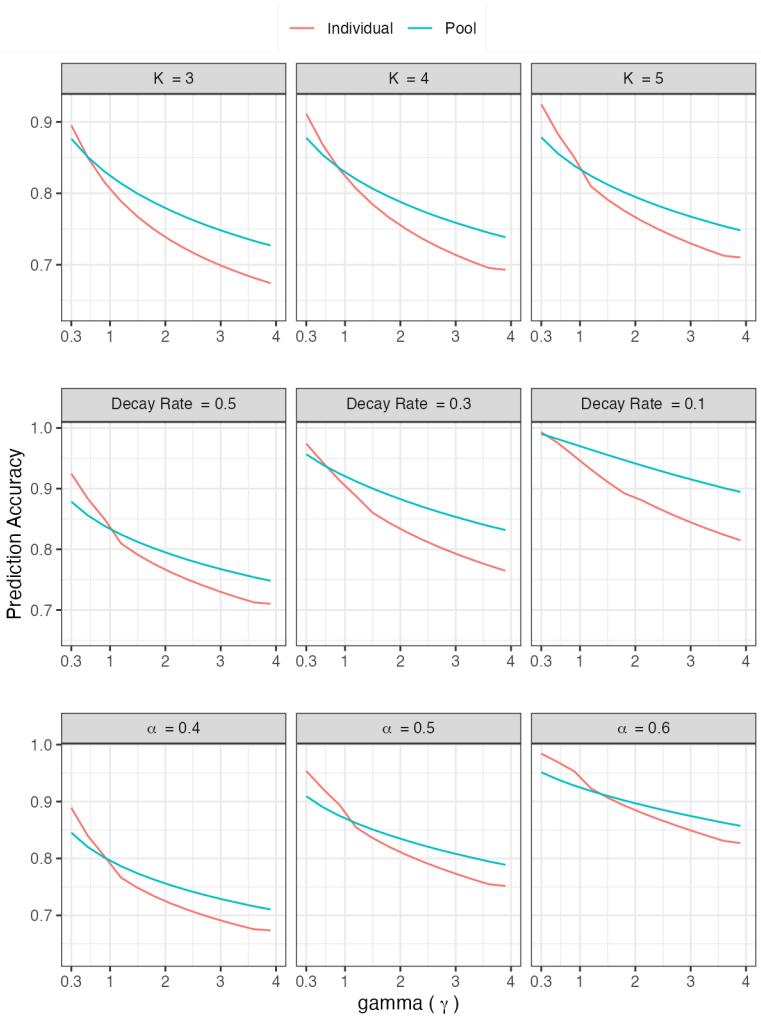

Figure 3: TLP-RDA outperforms TL-RDA when γ is large under general setups.

6. Heterogeneous Population Covariance Matrix

The previous discussions of TL-RDA have been restricted to the case where observed covariates X_k in all populations $k = 1, \cdots, K$ share the same covariance matrix Σ. In this section we extend TL-RDA to accommodate the scenario when covariance matrices are different. We call this generalization beyond identical matrices as transfer-learning-heterogeneous (TLH)-RDA. We firstly formalize this set up.

Assumption 6 (Heterogeneous Two-class Gaussian) *We assume all populations $k = 1, \cdots, K$ follow the classic two-class Gaussian mixture model. More specifically, for $i =$*

$1, \dots, n_k$ *and* $k = 1, \dots, K$,

$$(y_k)_i \in \{-1, +1\} \quad \mathbb{P}((y_k)_i = \pm 1) = \pi_{\pm 1} \quad (X_k)_i | (y_k)_i \sim N(\mu_{(y_k)_i}, \Sigma_k) \tag{11}$$

Note that this is identical to the original set up except each population has a different population covariance matrix Σ_k. In addition, we assume the assumption 3 holds for all covariance matrices Σ_k.

Assumption 7 (Heterogeneous RMT assumption) *For $k = 1, \cdots, K$, the design matrix $\mathbf{X}_k \in \mathbb{R}^{n_k \times p}$ is generated as*

$$\mathbf{X}_k = \left(\mu_{(y_k)_1,k}\, \mu_{(y_k)_2,k}\, \cdots\, \mu_{(y_k)_{n_k},k}\right)^\top + \mathbf{Z}_k \Sigma_k^{1/2}$$

for a matrix $\mathbf{Z}_k \in \mathbb{R}^{n_k \times p}$ with i.i.d. entries coming from an infinite array. The entries $(Z_k)_{ij}$ of $\mathbf{Z}_k$ satisfy the moment conditions:

$$\mathbb{E}[(Z_k)_{ij}] = 0,\ \mathbb{E}[(Z_k)_{ij}^2] = 1 \text{ and } \mathbb{E}[(Z_k)_{ij}^4] \le C.$$

1. *The population covariance matrix $\Sigma_k \in \mathbb{R}^{p \times p}$ is deterministic. The observations have unit variance, i.e., $(\Sigma_k)_{jj} = 1$ for $j = 1, \cdots, p$.*
2. *The eigenvalues of Σ_k are uniformly bounded from above and away from zero with constants independent of the dimension p.*
3. *The sequence of spectral distributions $T_k := (T_k)_{\Sigma,p}$ of $\Sigma_k := (\Sigma_k)_p$ converges weakly to a limiting distribution H_k supported on $[0, \infty)$, called the population spectral distribution (PSD).*

We can then derive the optimal prediction and estimation weights in a manner identical to Theorems 4.5 and 4.10. Since the Bayes optimal discriminant direction is given by $\Sigma_K^{-1}\delta_K$, and as before we aim to leverage the related observations in each source through a weighted linear combination of their discriminant directions $\widehat{d}_k$, for $k = 1, \dots, K$.

Theorem 6.1 (Asymptotic Estimation Error Minimization for TLH-RDA) *Suppose that assumptions 1, 2, 4, 5 as well as 6 and 7 hold. Then for a fixed $K \ge 2$, as $n_k, p \to \infty, p/n_k \to \gamma_k \in (0, \infty]$ for $1 \le k \le K$, the weight for minimizing the error in estimating the Bayes optimal discriminator d_{Bayes} is given by:*

$$\mathbf{w}_H^E := \arg\min_{\mathbf{w}} \left\| d_{Bayes} - \sum_{k=1}^K w_k \widehat{d}_k \right\|_2^2 = (\mathcal{A}_H^E + \mathcal{R}_H^E)^{-1} \mathbf{u}_H^E,$$

where the elements of $\mathbf{u}_H^E \in \mathbb{R}^K$, $\mathcal{A}_H^E \in \mathbb{R}^{K \times K}$, and $\mathcal{R}_H^E \in \mathbb{R}^{K \times K}$ are:

$$(u_H^E)_k = \begin{cases} \rho_{kK} \alpha_k \alpha_K \operatorname{tr}(\Sigma_K^{-1}(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}) & \text{if } k = 1, \dots, K-1, \\ \alpha_K^2 \left[\frac{1}{\lambda_k} \mathbb{E}(T_K^{-1}) - m_{F_{\gamma_k}}(-\lambda_k)^2 \right] & \text{if } k = K. \end{cases}$$

$$(\mathcal{A}_H^E)_{kk'} = \begin{cases} \alpha_k^2 m'_{F_{\gamma_k}}(-\lambda_k) & \text{if } k = k', \\ \rho_{kk'} \alpha_k \alpha_{k'} \mathcal{U}_{kk'} & \text{otherwise,} \end{cases}$$

and

$$(\mathcal{R}_H^E)_{kk'} = \begin{cases} \frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)^2} & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. *Moreover for* $k, k' = 1, \cdots, K$, *let* $\mathcal{U}_{kk'}$ *be the constants defined as the following limiting quantities:*

$$\operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p \to_{a.s.} \mathcal{U}_{kk'}. \tag{12}$$

Since the quantities on the left hand side of (12) are exactly known in terms of sample quantities, we do not seek the exact limits of the traces of the cross sample covariance terms $\mathcal{U}_{kk'}$. Instead, we advocate directly using the known quantities

$$\operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p.$$

The situation changes however for estimating $(u_H^E)_k$ when $k \neq K$, since they depend on the unknown target population covariance Σ_K. This is guaranteed by the almost sure convergence of a suitable sample based quantity, as recorded in the following proposition.

Proposition 6.2

$$\operatorname{tr}(\Sigma_K^{-1}(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}) - (1 - \gamma_K) \operatorname{tr}(\widehat{\Sigma}_K^{-1}(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}) \to_{a.s.} 0$$

as $n_k, p \to 0; p/n_k \to \gamma_k$ *for* $k = 1, \cdots, K-1$.

The proof of this proposition is immediate from Corollary 4.2 of Dobriban and Sheng (2021). We next move on to the optimal prediction weights for TLH-RDA.

Theorem 6.3 (Asymptotic Prediction Error Minimization for TLH-RDA) *Suppose that assumptions 1, 2 as well as 6 and 7 hold. Then for a fixed* $K \geq 2$, *as* $n_k, p \to \infty, p/n_k \to \gamma_k \in (0, \infty]$ *for* $1 \leq k \leq K$, *the weight for minimizing the excess risk, i.e., the error in predicting the class at a random test point* x_0, *when compared to the Bayes optimal discriminator* d_{Bayes}, *is given by:*

$$\mathbf{w}_H^P := \arg\min_{\mathbf{w}} \mathbb{E}_{x_0} \left[(d_{Bayes} - \sum_{k=1}^{K} w_k \widehat{d}_k)^\top x_0 \right]^2 = (\mathcal{A}_H^P + \mathcal{R}_H^P)^{-1} \mathbf{u}_H^P,$$

where the elements of $\mathbf{u}^P \in \mathbb{R}^K$, $\mathcal{A}^P \in \mathbb{R}^{K \times K}$, *and* $\mathcal{R}^P \in \mathbb{R}^{K \times K}$ *are:*

$$(u_H^P)_k = \rho_{kK} \alpha_k \alpha_K m_{F_{\gamma_k}}(-\lambda_k) \quad \text{for } k = 1, \ldots, K,$$

$$(\mathcal{A}_H^P)_{kk'} = \begin{cases} \alpha_k^2 \left[\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\gamma_k [\lambda_k v_{F_{\gamma_k}}(-\lambda_k)]^2} \right] & \text{if } k = k' = K, \\ \rho_{kk'} \alpha_k \alpha_{k'} \mathcal{Y}_{kk'} & \text{otherwise,} \end{cases}$$

and

$$(\mathcal{R}_H^E)_{kk'} = \begin{cases} \frac{v'_k(-\lambda_k) - v_k^2(-\lambda_k)}{\lambda_k^2 v_k^4(-\lambda_k)} & \text{if } k = k' \\ 0 & \text{otherwise,} \end{cases}$$

for $k, k' = 1, \cdots, K$. Moreover for $k, k' = 1, \cdots, K$, let $\mathcal{Y}_{kk'}$ be the constants defined as the following limiting quantities:

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\Sigma_K]/p \to_{a.s.} \mathcal{Y}_{kk'}.$$

In order to utilize the above weights in a practical scenario, we need to replace all unknown quantities by their estimates. As before, the Marcenko Pastur law is crucial in estimating the parameters related to the spectral distributions of Σ_k, for $k = 1, \ldots, K$. Most of the details are straightforward and identical in estimation procedure as the rest of the paper. It remains to provide consistent estimators for $\mathcal{Y}_{kk'}$, which we describe in the following proposition.

Proposition 6.4 *We have the following consistent estimators $\widehat{\mathcal{Y}}_{kk'}$ for $\mathcal{Y}_{kk'}$ separately for three cases:*

$$\widehat{\mathcal{Y}}_{kk'} = \begin{cases} \mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\widehat{\Sigma}_K]/p & \text{if } k \neq k', k \neq K, k' \neq K \\ \mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-2}\widehat{\Sigma}_K]/p & \text{if } k = k' \neq K \\ \frac{1}{px_p}\mathrm{tr}[(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}] - \frac{\lambda_K}{px_p}\mathrm{tr}[(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}] & \text{if } k = K,\ k' \neq K \end{cases}$$

where $x_p = x(\gamma_K, \lambda_K)$ is the solution to the equation:

$$1 - x_p = \gamma_K \left[1 - \lambda_K \int (x_p t + \lambda_K)^{-1} dH_K(t)\right]$$

Then

$$\mathcal{Y}_{kk'} - \widehat{\mathcal{Y}}_{kk'} \to_{a.s.} 0.$$

7. Real Data

7.1 Proteomics-based Prediction of 10-year Cardiovascular Disease Risk

We utilize the data set from the Chronic Renal Insufficiency Cohort (CRIC) study to evaluate the performance of the proposed transfer learning methods. The dataset comprises observations from $2,182$ subjects who had not had any cardiovascular disease at the baseline, with $4,830$ protein measurements collected. The number of subjects are $272, 280, 333, 326, 409, 180, 382$ respectively for seven university sites. Our goal is to build a classification model for 10-year cardiovascular disease risk based on the baseline plasma proteomic data. We hypothesize that while the mechanisms by which proteins predict events are related across these sites, they are not identical. Consequently, we sequentially designate each site as the target population, with the remaining sites serving as auxiliary populations, aiming to enhance prediction accuracy using TL-RDA or TLP-RDA.

For each target population, approximately 20% of the patients were set aside as a testing dataset, while the remaining 80% were used for training. We ensured that the proportion of events was consistent between the training and testing datasets ($\sim$7%). A univariate filtering procedure was applied to the training dataset using two-sample t-tests, and we evaluated the benefits of transfer learning using the top 500, 1000, and 1500 proteins. All predictors are centered and scaled within each sites, using training dataset information only. We

considered all four variants of transfer learning RDA: TL-RDA with optimal estimation or prediction weights and TLP-RDA with optimal estimation/prediction weights. The "optimal weights" were estimated as instructions given by remark 4.7. The tuning parameters λ for each model were selected independently by four-fold cross-validations over a length 9 logarithmically spaced grid ranges from 0.01 to 100. The domain specific penalization is obtained by $\lambda_k = \lambda\gamma_k/\alpha_k^2$. We refit each model on full training data with chosen λ and test them on the held-out testing dataset. All transfer learning weights are estimated based on empirical estimators provided in the theorems.

The testing AUCs were then compared with those obtained from competing methods, namely naive RDA and pool-RDA. As implied by their names, naive RDA is fitted using only the target population data, while pool-RDA is fitted on data pooled from all populations. These procedures were repeated 25 times, and we report the average AUCs and standard deviations of the AUCs in Figure 4.

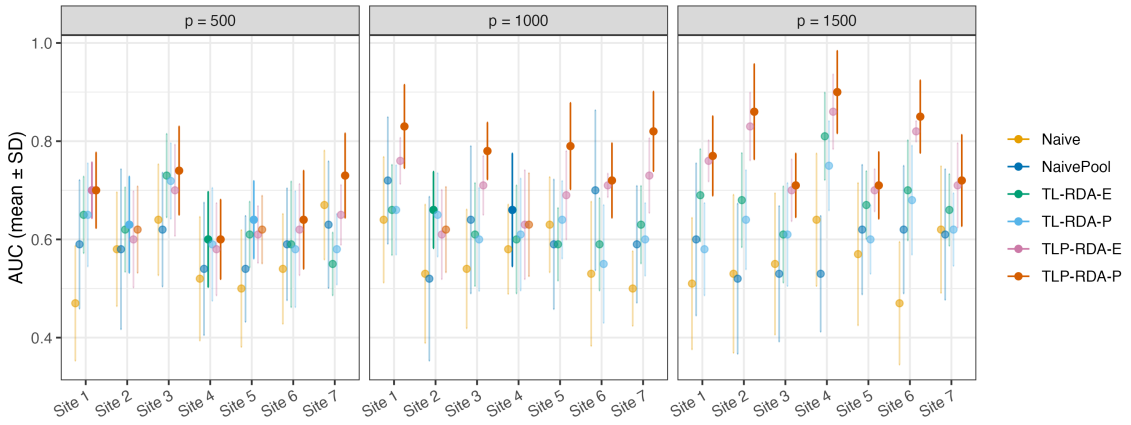

Figure 4: AUC ± SD on testing data for 10-year cardiovascular disease risk for each of the 7 clinical sites in the CRIC cohort using 500, 1000 and 1500 proteins, respectively. TL-RDA-E, TL-RDA-P, TLP-RDA-E and TLP-RDA-P denote TL-RDA with optimal estimation/prediction weights and TLP-RDA with optimal estimation/prediction weights, respectively.

It is evident that the transfer learning RDA methods outperform the others in the vast majority of cases. This demonstrates that not only are TL-RDAs capable of borrowing information to improve prediction accuracy, but they also do so more efficiently than simply pooling the data. Notably, TLP-RDA with optimal prediction weights outperforms the other methods most frequently. The amount of improvement brought by transfer learning also seemingly increase as we include more proteins in the model. Additionally, we observe that methods based on pooled sample covariance tend to perform better when the number of features p is larger, consistent with Proposition 5.2.

7.2 Lipid traits classification using genotype data

We also evaluate the proposed transfer-learning estimators using a data set from Penn Medicine BioBank, which contains both genome-wide genotype data and the electronic health record–derived phenotypes. We focus on three lipid traits with substantial genetic signal, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and triglycerides (TRI), and dichotomize each outcome using guideline-based clinical cutoffs (Grundy et al., 2019, 2005; NCEP, 2002). To emulate settings where related auxiliary traits are available, we analyze each trait in turn as the target and use the remaining two traits as sources. The analytic cohort contains approximately 12000 individuals, partitioned into three trait-specific datasets. For each target analysis, feature screening is performed solely on the target training split via univariate association tests, and the top 1000, 2000, and 3000 nucleotide polymorphism (SNPs) (ranked by significance) are retained as the predictors.

All regularization hyperparameters were tuned exactly as in the proteomics experiment, using the same grid of candidate values. Figure 5 reports averaged AUCs and their standard deviations on held-out test sets for the proposed methods and the baselines. The statistics are based on 25 independent train / test splits and cross-validations. Across most configurations, Trans-Ridge improves upon target-only ridge regression. In contrast to the proteomics based prediction across different clinical sites in previous Section, the TL-RDA methods outperform the TLP-RDA variants based on pooled sample-covariance estimators. This ordering is again consistent with Proposition 5.2, which predicts an advantage for TL-RDA in the regime of small γ_k, the regime relevant to the present data.

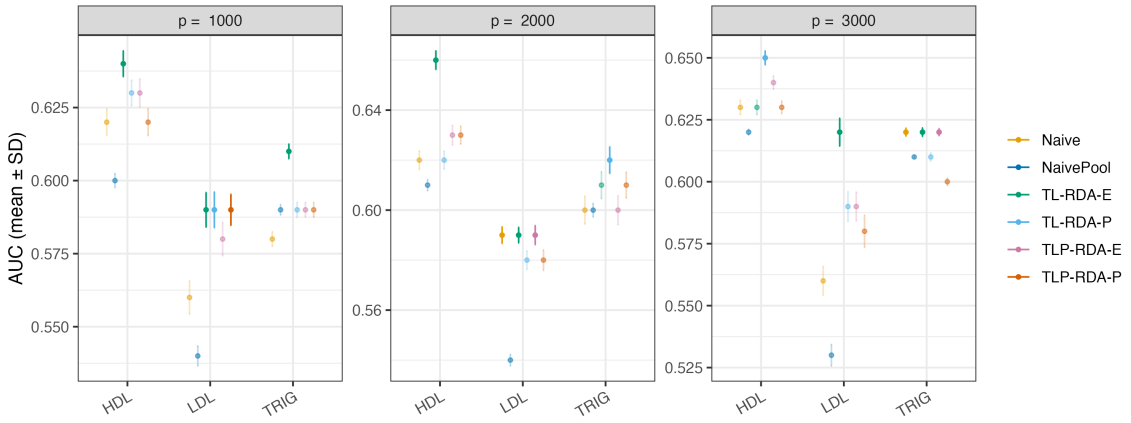

Figure 5: AUC ± SD on testing data for HDL/LDL/Trig using 1000, 2000 and 3000 SNPs, respectively. TL-RDA-E, TL-RDA-P, TLP-RDA-E and TLP-RDA-P denote TL-RDA with optimal estimation/prediction weights and TLP-RDA with optimal estimation/prediction weights, respectively.

8. Discussion

We have developed methods of transfer learning for classification using regularized random effects linear discriminant analysis. In this approach, the discriminant direction is estimated

through a weighted combination of the regularized estimates of discriminant directions derived from both the target and source models. By leveraging results from random matrix theory, we have demonstrated that the weights, as well as the classification error rate, can be accurately estimated using the data. Our findings highlight that the optimal choice of weights depends critically on the underlying true models and the presence or absence of distributional shifts between the testing and training data in the target model.

Through comprehensive empirical evaluations, we have illustrated the practical utility of the proposed methods. Specifically, we applied the transfer learning approach to predict 10-year cardiovascular disease risk using high-dimensional protein expression data and to the lipid trait prediction problem with genotype data. The results demonstrate that incorporating information from related source datasets substantially improved classification performance compared to target-only models or models based on pooled data. This improvement underscores the importance of accounting for shared information across datasets while accommodating potential distributional differences. In addition, the theoretical guidance for selecting among the four candidate weighting schemes is supported by the empirical evidence. In the high-dimensional proteomics application, pooled-covariance weighting performs the best, consistent with its variance-reduction advantage when signals are diffuse. For the genotype data, where target and source data includes correlated but different outcomes, weights optimized for estimation typically outperform those optimized for prediction. When the requisite assumptions are uncertain, we recommend selecting among the four schemes using cross-validation.

In terms of computational scalability, TL-RDA has the same leading complexity as running the K source RDA models. To compute RDA directions with naive algorithms, forming the covariances costs $O(Kn_kp^2)$ and the $K+1$ matrix inversions cost $O(Kp^3)$. Solving and estimating the optimal weights are equivalent to at most two linear systems and one eigen-problem of size $K \times K$. This $O(K^3)$ overhead is negligible because $K \ll p$ in our applications. In addition, all heavy computations are performed locally within each domain. Each domain returns only its RDA direction and a few scalar trace statistics required by Theorem 4.5 and Theorem 4.10. The central node only needs to combine these $O(Kp)$ and $O(K)$ summaries to obtain the weight vector. Moreover, because each inversion has the ridge form $(\widehat{\Sigma}_k + \lambda I_p)^{-1}$, we can apply Woodbury and perform a Cholesky factorization on an $n_k \times n_k$ system, thereby eliminating the cubic dependence on p and reducing per-domain cost to $O(n_k^3 + n_kp)$ when $n_k \ll p$.

This paper focuses on two-class LDA models. A promising direction for future research is extending these methods to multi-class classification within the framework of transfer learning. Other interesting avenues include exploring this problem under privacy or communication constraints, as has been recently studied for other classification methods, or analyzing LDA trained with mixed in-distribution and out-of-distribution samples (see, e.g., Auddy et al., 2024; De Silva et al., 2023). Also, we have assumed the number of source domains K is small compared to number of observations and predictors (n_k, p). If instead K is very large, source selection becomes necessary. As noted in Section 4.4, the optimal weight vectors solves a least-squares problem in which the Bayes discriminant direction is regressed on the collection of domain-specific discriminant directions. When the number of sources K grows to the same order as n and p, this becomes a high-dimensional regression and variable-selection problem. A simple heuristic is to retain only those sources whose discriminant

directions have the largest absolute correlations with the target estimate; however, this ignores cross-source dependencies and may discard complementary information. A natural extension is to add a ridge-type weight penalization $\|\mathbf{w}\|_2^2$ on criterion 3 or criterion 4. The resulting estimator can still be computed in closed form via the machinery of Theorems 4.5 and Theorems 4.10. Developing principled, sparsity-inducing penalties or screening rules that account for inter-source correlations and analyzing their theoretical guarantees, remains an interesting direction for future research.

Acknowledgments

This research was supported by NIH grants R01GM129781 and U01HG013841.

References

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press, 2010.

T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

A. Auddy, T. T. Cai, and A. Chakraborty. Minimax and adaptive transfer learning for nonparametric classification under distributed differential privacy constraints. *arXiv preprint arXiv:2406.20088*, 2024.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

P. J. Bickel and E. Levina. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

M. Capitaine and M. Casalis. Asymptotic freeness by generalized moments for gaussian and wishart matrices. application to beta random matrices. *Indiana University mathematics journal*, pages 397–431, 2004.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

A. De Silva, R. Ramesh, C. Priebe, P. Chaudhari, and J. T. Vogelstein. The value of out-of-distribution data. In *International Conference on Machine Learning*, pages 7366–7389. PMLR, 2023.

E. Dobriban and Y. Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 419:918–943, 2021.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

S. M. Grundy, J. I. Cleeman, S. R. Daniels, K. A. Donato, R. H. Eckel, B. A. Franklin, D. J. Gordon, R. M. Krauss, P. J. Savage, S. C. Smith, J. A. Spertus, and F. Costa. Diagnosis and management of the metabolic syndrome: An american heart association/national heart, lung, and blood institute scientific statement. *Circulation*, 112(17):2735–2752, 2005. doi: 10.1161/CIRCULATIONAHA.105.169404.

S. M. Grundy, N. J. Stone, A. L. Bailey, and et al. 2018 aha/acc/aacvpr/aapa/abc/acpm/ada/ags/apha/aspc/nla/pcna guideline on the management of blood cholesterol. *Circulation*, 139(25):e1082–e1143, 2019. doi: 10.1161/CIR.0000000000000625.

T. Gu, Y. Han, and R. Duan. Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society, Series B*, 84(1):149–173, 2024. doi: 10.1093/jrsssb/qkad045.

N. Han, J. Wu, X. Fang, J. Wen, S. Zhan, S. Xie, and X. Li. Transferable linear discriminant analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5630–5638, 2020. doi: 10.1109/TNNLS.2020.2973202.

H. Helm, A. De Silva, J. T. Vogelstein, C. E. Priebe, and W. Yang. Approximately optimal domain adaptation with fisher's linear discriminant. *Mathematics*, 12(5):746, 2024.

M. Huisman, J. N. Van Rijn, and A. Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021.

A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017.

A. Komárek and E. Lesaffre. Discriminant analysis using a multivariate linear mixed model. *Biostatistics*, 2010.

O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

S. Li, T. T. Cai, and H. Li. Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):149–173, 2022.

S. Li, L. Zhang, T. T. Cai, and H. Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119(546):1274–1285, 2023.

G. Liu, X. Li, and W. Liu. Indirect category data transfer learning algorithm using regularization discrimination. *Big Data*, 11(1):59–70, 2023. doi: 10.1089/big.2022.0156.

A. C. Lozano and G. Swirszcz. Multi-level lasso for sparse multi-task regression. *ICML*, 2012.

T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41(6):469–480, 2017.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2009.

V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

NCEP. Third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III) final report. *Circulation*, 106(25):3143–3421, 2002. doi: 10.1161/circ.106.25.3143.

A. Okazaki and M. Aoyagi. Multi-task learning regression via convex clustering. *Statistical Papers*, 2024.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

H. Peng. A comprehensive overview and survey of recent advances in meta-learning. *arXiv preprint arXiv:2004.11149*, 2020.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

B. S. Price and al. Fusion penalties in statistical learning. *arXiv:1405.xxxx*, 2014.

I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*, 2020.

V. I. Serdobolskii. *Multiparametric statistics*. Elsevier, 2007.

Y. Sheng and E. Dobriban. One-shot distributed ridge regression in high dimensions. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8763–8772. PMLR, 13–18 Jul 2020.

J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

S. Sun, H. Shi, and Y. Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015.

A. Torkamani, N. E. Wineinger, and E. J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19:581–590, 2018.

L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

J. Wang, T. Yu, and Z. Huang. Integrated transfer learning based on group sparse bayesian linear discriminant analysis for error-related potentials detection. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 645–650, 2020. doi: 10.1109/ICAICA50127.2020.9182754.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

P. Yang and W. Gao. Multi-view discriminant transfer learning. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1340–1346, 2013.

H. Zhang and H. Li. Transfer learning with random coefficient ridge regression. *arXiv preprint arXiv:2306.15915*, 2023.

B. Zhao, F. Zou, and H. Zhu. Cross-trait prediction accuracy of summary statistics in genome-wide association studies. *Biometrics*, 79(2):841–853, 2023.

Appendix A. Appendix A

A.1 Proofs of Propositions

Proof [Proof of Proposition 4.8] Denote E as the random variable distributed according to the empirical spectral distribution F_γ. When $\gamma < 1$, E is supported on a compact set bounded away from 0 (Bai and Silverstein, 2010). Therefore, we can take the limit of $m_{F_\gamma}(-\lambda)$ as $\lambda \to 0$. Recall the Marchenko–Pastur equation

$$m_{F_\gamma}(z) = \int_{t=0}^{\infty} \frac{dH(t)}{t(1-\gamma-\gamma z m_{F_\gamma}(z)) - z}$$

We find $m_{F_\gamma}(0) = \int 1/[t(1-\gamma)]dH(t)$ or equivalently $\mathrm{tr}(\widehat{\Sigma}^{-1})/p \to_{a.s.} \mathbb{E}(T^{-1})/(1-\gamma)$. ■

Proof [Proof of Proposition 4.11] The proof is identical to the proof of Proposition 6.4 which considers the more general case of unequal covariance matrices Σ_k. ■

Proof [Proof of Proposition 4.14] Recall from Theorem 10 that the limiting prediction error is in the form of

$$\Phi\left(-\frac{\mathbf{w}^\top \mathbf{u}}{\sqrt{\mathbf{w}^\top A \mathbf{w}}}\right)$$

Any weight $\mathbf{w}$ with negative linear term $\mathbf{w}^\top \mathbf{u}$ has higher prediction error than the weight with positive linear term. Without loss of generality, we assume $\mathbf{w}^\top v$ to be positive. Then we know

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{u}}{\sqrt{\mathbf{w}^\top \mathcal{A} \mathbf{w}}} = \arg\max \frac{\mathbf{w}^\top \mathbf{u}\mathbf{u}^\top \mathbf{w}}{\mathbf{w}^\top \mathcal{A} \mathbf{w}}$$

this becomes the generalized Rayleigh quotient problem and the solution is given by the first unit eigenvector of $\mathcal{A}^{-1}\mathbf{u}\mathbf{u}^\top$, which is a rank one matrix, and hence

$$\mathbf{w}^* = \frac{1}{\|\mathcal{A}^{-1}\mathbf{u}\|}\mathcal{A}^{-1}\mathbf{u}, \text{ which implies, } Err_{opt} := \Phi\left(-\sqrt{\mathbf{u}^\top \mathcal{A}^{-1}\mathbf{u}}\right).$$

Note the standardization factor in $\mathbf{w}$ is not necessary due to the form of $Err(\mathbf{w})$. The proof finishes as one recognizes the $\mathcal{A} = \mathcal{A}^P + \mathcal{R}^P$ and $\mathbf{u} = \mathbf{u}^P$. We can prove the statement for $\mathbf{w}_P^P$ in the same manner. ■

Proof [Proof of Proposition 4.15] We only need to prove $\mathcal{A}^E + \mathcal{R}^E, \mathcal{A}^P + \mathcal{R}^P, \mathcal{A}_P^E + \mathcal{R}_P^E, \mathcal{A}_P^P + \mathcal{R}_P^P$ are positive definite, therefore, invertible. Looking at the individual sample covariance matrix cases, taking individual prediction weight as an example. We will firstly show that $\mathcal{A}^P$ is positive semi-definite. Note that

$$\begin{aligned}
(\mathcal{A}^P)_{kk'} &= \alpha_k \alpha_{k'} \rho_{kk'} \mathrm{tr}[\Sigma(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p \\
&= \alpha_k \alpha_{k'} \rho_{kk'} \mathrm{tr}[\mathbf{M}_k^\top \mathbf{M}_{k'}]/p \\
&= \alpha_k \alpha_{k'} \rho_{kk'} \mathbf{v}_{M_k}^\top \mathbf{v}_{M_{k'}}/p
\end{aligned}$$

where $\mathbf{M}_k = \Sigma^{1/2}(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}$ and $\mathbf{v}_{M_k} = \text{vec}(\mathbf{M}_k) \in \mathbb{R}^{K^2}$. For any $\mathbf{x} \in \mathbb{R}^K$, $\mathbf{x} \neq \mathbf{0}$, we have

$$\begin{aligned}\mathbf{x}^\top \mathcal{A}^P \mathbf{x} &= \sum_{k,k'} (\alpha_k x_k)(\alpha_{k'} x_{k'}) \rho_{kk'} \mathbf{v}_{M_k}^\top \mathbf{v}_{M_{k'}} / p \\ &= \mathbf{y}^\top \mathcal{A}^P \mathbf{y} / p\end{aligned}$$

where $\mathbf{y} = ((\alpha_i x_i))_i \in \mathbb{R}^K$. Since $\alpha_i > 0$ for all i, we have $\mathbf{y} \neq \mathbf{0}$ if and only if $\mathbf{x} \neq \mathbf{0}$. It is thus enough to show that $\mathcal{A}^{(3)}$, defined through

$$(\overline{\mathcal{A}})_{ij} = \rho_{ij} \mathbf{v}_{M_i}^\top \mathbf{v}_{M_j}$$

is positive-semi definite. To this end, note that

$$\begin{aligned}\mathbf{y}^\top \overline{\mathcal{A}} \mathbf{y} &= \sum_{i,j} \rho_{ij} (y_i \mathbf{v}_{M_i})^\top (y_j \mathbf{v}_{M_j}) \\ &= \sum_{i,j} \rho_{ij} (\mathbf{c}_i)^\top \mathbf{c}_j = \text{tr}(\Sigma_\delta \mathbf{C}^\top \mathbf{C}) \\ &= \text{tr}(\mathbf{C} \Sigma_\delta^{1/2} \Sigma_\delta^{1/2} \mathbf{C}^\top) \\ &= \|\boldsymbol{\rho}^{1/2} \mathbf{C}^\top\|_{\text{F}}^2 \geq 0.\end{aligned}$$

Here $\mathbf{C} \in \mathbb{R}^{K^2 \times K}$ matrix with columns $\mathbf{c}_i = y_i \mathbf{v}_{M_i}$, for $i = 1, \ldots, K$. This shows that $\mathcal{A}^{(1)}$ is positive semi-definite since for any $\mathbf{x} \in \mathbb{R}^K$, $\mathbf{x} \neq \mathbf{0}$ we have

$$\mathbf{x}^\top \mathcal{A}^P \mathbf{x} = \mathbf{y}^\top \overline{\mathcal{A}} \mathbf{y} \geq 0. \tag{13}$$

Now note that $\mathcal{A}^{(2)}$ is a diagonal matrix with non-negative elements, this finishes the proof of the positive definiteness of $\mathcal{A}^P + \mathcal{A}^R$. To prove the case of estimation weight, one can simply define $\mathbf{M}_k = (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}$.

Looking at the pooled sample covariance matrix cases, taking pooled prediction weight as an example, we would have

$$\begin{aligned}\mathcal{A}_P^P &= c^A \Sigma_{\alpha\delta} \\ \Sigma_{\alpha\delta} &= mat\,[\rho_{kk'} \alpha_k \alpha_{k'}] \\ c^A &= \text{tr}[(\widehat{\Sigma}_P + \lambda I_p)^{-2} \Sigma] / p\end{aligned}$$

Since $\mathcal{R}^P$ has non-negative diagonal entries and $c^A > 0$, we only need to prove the positive semi-definiteness of $\Sigma_{\alpha\delta}$. Consider the quadratic form $x^\top \Sigma_{\alpha\delta} x$:

$$x^\top \Sigma_{\alpha\delta} x = \sum_{i=1}^K \sum_{j=1}^K \alpha_i x_i \alpha_j (\Sigma_{\alpha\delta})_{ij} x_j = \left(\sum_{i=1}^K \alpha_i x_i \right) \left(\sum_{j=1}^K \alpha_j \Sigma_{ij} x_j \right)$$

Define $y_i = \alpha_i x_i$. Then we have:

$$x^\top \Sigma_{\alpha\delta} x = \sum_{i=1}^K y_i \sum_{j=1}^K (\Sigma_\delta)_{ij} y_j = y^\top \Sigma y$$

Since Σ_δ is positive semi-definite, $\mathcal{A}_P^P + \mathcal{R}_P^P$ will be positive definite. We can prove the positive definiteness of $\mathcal{A}_P^E + \mathcal{R}_P^E$ in exactly the same way. ■

Proof [Proof of Proposition 5.1] Since the maximum value of the dot product of two vectors is achieved when the two vectors are aligned, the maximization of the min-max problem is solved when x_0 is in the same direction as the error with $\mathbb{E}_P(\|x_0\|_2) = c$. Since x_0 is independent of the estimation error vector $(\sum_{k=1}^K w_k \widehat{d}_k - d_{Bayes})^\top$, we have the solution of the max problem as

$$x_0^* = c\frac{d_{Bayes} - \sum_{k=1}^K w_k \widehat{d}_k}{\|d_{Bayes} - \sum_{k=1}^K w_k \widehat{d}_k\|} \text{ with probability 1.}$$

This leads to the estimation risk minimization problem with criteria (3). ■

Proof [Proof of Proposition 5.2] For simplicity, we assume $\Sigma = \mathbb{I}_p$, in which case the optimal esimation weight coincides with the optimal prediction weights. In addition, we have the following simplifications

Corollary A.1 (Simplification of Limiting Terms for Individual Sample Covariance Matrix) *Recall the form of limiting prediction error given a weight vector* $\mathbf{w}$ (10). *We have the term* $\mathcal{A}$ *simplified as following. For* $k = 1, \cdots, K$,

$$\mathcal{A}_{kk} = \alpha_k^2 m_k'(-\lambda_k) + \gamma_k m_k'(-\lambda_k)$$

else when $k \neq k'; k, k' = 1, \cdots, K$

$$\mathcal{A}_{kk'} = \rho_{kk'}\alpha_k\alpha_{k'} m_k(-\lambda_k) m_{k'}(-\lambda_{k'})$$

and we have $\mathbf{u}^E = vec[\rho_{kK}\alpha_k\alpha_K m_{F_{\gamma_k}}(-\lambda_k)]$, $\mathcal{A}_{kk'}^E + \mathcal{R}^E = \mathcal{A}$.

Corollary A.2 (Simplification of Limiting Terms for Pooled Sample Covariance Matrix) *Define the weighted covariance matrix as* $\Sigma_\delta = mat[\rho_{kk'}\alpha_k\alpha_{k'}], \rho_{kk} = 1$. *We have the term* $\mathcal{A}$ *simplified as:*

$$\mathcal{A}_P = \Sigma_\delta m'_{F_{\gamma/K}}(-\lambda) + \gamma_k m'_{F_{\gamma/K}}(-\lambda)\mathbb{I}_K$$

and $\mathbf{u}_P^E = vec[\rho_{kK}\alpha_k\alpha_K m_{F_{\gamma/K}}(-\lambda)]$, $\mathcal{A}_P^E + \mathcal{R}_P^E = \mathcal{A}_P$.

We can now write the limiting error of using individual sample covariance matrix and pooled covariance matrix as

$$Err_{ind} = \Phi\left(-\sqrt{(\mathbf{u}^E)^\top \mathcal{A}^{-1}\mathbf{u}^E}\right) \text{ and } Err_{pool} = \Phi\left(-\sqrt{(\mathbf{u}_P^E)^\top \mathcal{A}_P^{-1}\mathbf{u}_P^E}\right).$$

Recall that we have the explicit expressions

$$m_{F_\gamma}(-\lambda) = \frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\gamma\lambda}}{2\gamma\lambda} \tag{14}$$

$$m'_{F_\gamma}(-\lambda) = \frac{m^2_{F_\gamma}(-\lambda)[1+\gamma m_{F_\gamma}(-\lambda)]}{1+\gamma\lambda m^2_{F_\gamma}(-\lambda)}. \tag{15}$$

For some $r > (1-\gamma)_+/\gamma + c$ (for some small constant c to be chosen later), we choose λ as follows:

$$\begin{aligned} 1-\gamma+\lambda = r\gamma - \lambda/r \iff (1+1/r)\lambda &= (r+1)\gamma - 1 \\ \iff \lambda_* &= r\left(\gamma - \frac{1}{r+1}\right) \end{aligned}$$

The following proposition contains a specific choice of r and some inequalities that will be useful for the rest of this proof. The proof of this proposition is immediate and hence omitted.

Proposition A.3 *The following statements hold for $r > (1-\gamma)_+/\gamma + c$ (for some constant $c > 0$):*

1. $m_\gamma(-\lambda_*) = \frac{1}{r\gamma}$
2. $\frac{m'_{F_\gamma}(-\lambda_*)}{m^2_{F_\gamma}(-\lambda_*)} = \frac{\gamma(1+r)^2}{\gamma(1+r)^2-1}$
3. $\Delta := \frac{m'_{F_\gamma}(-\lambda_*)}{m^2_{F_\gamma}(-\lambda_*)} - 1 = \frac{1}{\gamma(1+r)^2-1} \le \frac{1}{\gamma-1}$

We now look at the quadratic term in Err_{pool} and assume $\rho_{kk'} = \rho$. Then by Corollary A.2 we have

$$\begin{aligned} \mathcal{A}_P &= \Sigma_\delta m'_{F_{\gamma/K}}(-\lambda) + m'_{F_{\gamma/K}}(-\lambda)\mathrm{diag}(\{\gamma_k : 1 \le k \le K\}) \\ &= m'_{F_{\gamma/K}}(-\lambda)\left[\rho\widetilde{\boldsymbol{\alpha}}\widetilde{\boldsymbol{\alpha}}^\top + \mathrm{diag}(\{(1-\rho)\alpha_k^2 + \gamma_k : 1 \le k \le K\})\right] \\ &=: m'_{F_{\gamma/K}}(-\lambda)\left[\rho\widetilde{\boldsymbol{\alpha}}\widetilde{\boldsymbol{\alpha}}^\top + D_{\rho,\widetilde{\boldsymbol{\gamma}}}\right] \end{aligned}$$

and

$$\mathbf{u}_P^E = vec[\rho_{kK}\alpha_K\alpha_k m_{F_{\gamma/K}}(-\lambda)] = m_{F_{\gamma/K}}(-\lambda)\alpha_K[\rho\widetilde{\boldsymbol{\alpha}} + (1-\rho)\alpha_K\mathbf{e}_K].$$

By the Sherman Morrison formula, we have:

$$\mathcal{A}_P^{-1} = \frac{1}{m'_{F_{\gamma/K}}(-\lambda)}\left[D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}} - \frac{\rho}{1+\rho\sum_{k=1}^K \frac{\alpha_k^2}{(1-\rho)\alpha_k^2+\gamma_k}} D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\widetilde{\boldsymbol{\alpha}}\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\right].$$

Consequently:

$$\begin{aligned} \widetilde{\boldsymbol{\alpha}}^\top\mathcal{A}_P^{-1}\widetilde{\boldsymbol{\alpha}} &= \frac{1}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \frac{\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\widetilde{\boldsymbol{\alpha}}}{1+\rho\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\widetilde{\boldsymbol{\alpha}}} \\ \widetilde{\boldsymbol{\alpha}}^\top\mathcal{A}_P^{-1}\mathbf{e}_K &= \frac{1}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \frac{\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\mathbf{e}_K}{1+\rho\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\widetilde{\boldsymbol{\alpha}}} = \frac{\alpha_K}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \frac{\mathbf{e}_K^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\mathbf{e}_K}{1+\rho\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{\rho,\widetilde{\boldsymbol{\gamma}}}\widetilde{\boldsymbol{\alpha}}} \end{aligned}$$

$$\mathbf{e}_K^\top \mathcal{A}_P^{-1} \mathbf{e}_K = \frac{1}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \left[\mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K - \frac{\alpha_K^2}{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}} (\mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K)^2 \right]$$
$$= \frac{\mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \left[\frac{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}} - \rho \alpha_K^2 \mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K}{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}} \right]$$

which, upon writing $\widetilde{\boldsymbol{\alpha}}_{/K} = (\alpha_1\, \alpha_2\, \ldots\, \alpha_{K-1}\, 0)^\top$, implies that:

$$\begin{aligned}
&(\mathbf{u}_P^E)^\top \mathcal{A}_P^{-1} \mathbf{u}_P^E \\
&= \frac{\alpha_K^2 \cdot m_{F_{\gamma/K}}^2(-\lambda)}{m'_{F_{\gamma/K}}(-\lambda)} \left[\rho^2 \widetilde{\boldsymbol{\alpha}}^\top \mathcal{A}_P^{-1} \widetilde{\boldsymbol{\alpha}} + 2\rho(1-\rho)\alpha_K \widetilde{\boldsymbol{\alpha}}^\top \mathcal{A}_P^{-1} \mathbf{e}_K + (1-\rho)^2 \alpha_K^2 \mathbf{e}_K^\top \mathcal{A}_P^{-1} \mathbf{e}_K \right] \\
&= \frac{m_{F_{\gamma/K}}^2(-\lambda)}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \frac{\alpha_K^2}{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}} \times \\
&\quad \times \left[\rho^2 \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}} + (1-\rho)\alpha_K^2 \mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K [2\rho + 1 - \rho + \rho(1-\rho) \widetilde{\boldsymbol{\alpha}}_{/K}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}_{/K}] \right] \\
&= \frac{m_{F_{\gamma/K}}^2(-\lambda)}{m'_{F_{\gamma/K}}(-\lambda)} \cdot \frac{\alpha_K^2}{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}} \times \\
&\quad \times \left[\alpha_K^2 \mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K + \rho^2 \widetilde{\boldsymbol{\alpha}}_{/K}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}_{/K} + \rho(1-\rho)^2 (\alpha_K^2 \mathbf{e}_K^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \mathbf{e}_K) \widetilde{\boldsymbol{\alpha}}_{/K}^\top D_{\rho,\widetilde{\boldsymbol{\gamma}}}^{-1} \widetilde{\boldsymbol{\alpha}}_{/K} \right] \qquad (16)
\end{aligned}$$

Let us now look at the quadratic term of the individual directions, once again assuming $\rho_{kk'} = \rho$. We make another simplifying assumption at this point:
Assumption: $\gamma_k = \gamma$ and $\lambda_k = \lambda$ for $k = 1, \ldots, K$. Then the population specific covariance matrix based quadratic form reduces through the following calculation.

By Corollary A.1 we now have

$$\begin{aligned}
\mathcal{A} &= m_{F_\gamma}^2(-\lambda) \left[\rho \widetilde{\boldsymbol{\alpha}} \widetilde{\boldsymbol{\alpha}}^\top + \operatorname{diag} \left(\left\{ \left(\frac{m'_{F_\gamma}(-\lambda)}{m_{F_\gamma}^2(-\lambda)} - \rho \right) \alpha_k^2 + \frac{m'_{F_\gamma}(-\lambda)\gamma}{m_{F_\gamma}^2(-\lambda)} : 1 \le k \le K \right\} \right) \right] \\
&=: m_{F_\gamma}^2(-\lambda) \left[\rho \widetilde{\boldsymbol{\alpha}} \widetilde{\boldsymbol{\alpha}}^\top + D_{\rho,\gamma,ind} \right]
\end{aligned}$$

and

$$\mathbf{u}^E = vec[\rho_{kK} \alpha_K \alpha_k m_{F_\gamma}(-\lambda)] = m_{F_\gamma}(-\lambda) \alpha_K [\rho \widetilde{\boldsymbol{\alpha}} + (1-\rho)\alpha_K \mathbf{e}_K].$$

Then following the exact steps leading to (16) we arrive at:

$$\begin{aligned}
&(\mathbf{u}^E)^\top \mathcal{A}^{-1} \mathbf{u}^E \\
&= \frac{\alpha_K^2}{1+\rho \widetilde{\boldsymbol{\alpha}}^\top D_{\rho,\gamma,ind}^{-1} \widetilde{\boldsymbol{\alpha}}} \times \\
&\quad \times \left[\alpha_K^2 \mathbf{e}_K^\top D_{\rho,\gamma,ind}^{-1} \mathbf{e}_K + \rho^2 \widetilde{\boldsymbol{\alpha}}_{/K}^\top D_{\rho,\gamma,ind}^{-1} \widetilde{\boldsymbol{\alpha}}_{/K} + \rho(1-\rho)^2 (\alpha_K^2 \mathbf{e}_K^\top D_{\rho,\gamma,ind}^{-1} \mathbf{e}_K) \widetilde{\boldsymbol{\alpha}}_{/K}^\top D_{\rho,\gamma,ind}^{-1} \widetilde{\boldsymbol{\alpha}}_{/K} \right]. \qquad (17)
\end{aligned}$$

Now comparing Equations (16) and (17) we can determine whether the pooled or the individual covariance matrix builds a better estimator. We consider two special cases below:

Case 1 ($\rho = 1$): In this case we have from (16) and (17) that:

$$(\mathbf{u}_P^E)^\top \mathcal{A}_P^{-1} \mathbf{u}_P^E = \frac{m^2_{F_{\gamma/K}}(-\lambda')\alpha_K^2}{m'_{F_{\gamma/K}}(-\lambda')} \cdot \frac{\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}}{1 + \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}}$$

and

$$(\mathbf{u}^E)^\top \mathcal{A}^{-1} \mathbf{u}^E = \frac{\alpha_K^2 (\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\gamma,ind} \widetilde{\boldsymbol{\alpha}})}{1 + \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\gamma,ind} \widetilde{\boldsymbol{\alpha}}}.$$

Note that by definition of $D_{\rho,\gamma,ind}$ we have

$$\begin{aligned}
\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\gamma,ind} \widetilde{\boldsymbol{\alpha}} &= \sum_{k=1}^K \frac{\alpha_k^2}{\left(\frac{m'_{F_\gamma}(-\lambda)}{m^2_{F_\gamma}(-\lambda)} - \rho\right)\alpha_k^2 + \frac{m'_{F_\gamma}(-\lambda)\gamma}{m^2_{F_\gamma}(-\lambda)}} \\
&\leq \frac{m^2_{F_\gamma}(-\lambda)}{m'_{F_\gamma}(-\lambda)} \sum_{k=1}^K \frac{\alpha_k^2}{\gamma} = \frac{m^2_{F_\gamma}(-\lambda)}{m'_{F_\gamma}(-\lambda)} \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}} \\
&= \frac{\gamma(1+r)^2 - 1}{\gamma(1+r)^2} \times \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}
\end{aligned}$$

where the last line follows from our choice of λ. Since $f(t) = \frac{t}{1-t}$ is an increasing function of t it follows that:

$$\frac{\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\gamma,ind} \widetilde{\boldsymbol{\alpha}}}{1 + \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\gamma,ind} \widetilde{\boldsymbol{\alpha}}} \leq \frac{[\gamma(1+r)^2 - 1]\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}}{\gamma(1+r)^2 + [\gamma(1+r)^2 - 1]\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}} = \frac{[\gamma(1+r)^2 - 1]\sum_k \alpha_k^2}{\gamma^2(1+r)^2 + [\gamma(1+r)^2 - 1]\sum_k \alpha_k^2}.$$

On the other hand,

$$\frac{m^2_{F_{\gamma/K}}(-\lambda')}{m'_{F_{\gamma/K}}(-\lambda')} \cdot \frac{\widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}}{1 + \widetilde{\boldsymbol{\alpha}}^\top D^{-1}_{1,\widetilde{\boldsymbol{\gamma}}} \widetilde{\boldsymbol{\alpha}}} = \frac{\gamma(1+r')^2 - K}{\gamma(1+r')^2} \times \frac{\sum_k \alpha^2}{\gamma + \sum_k \alpha_k^2}.$$

Thus the pooled covariance matrix performs better when

$$\big(\gamma(1+r')^2 - K\big)\left(\gamma^2(1+r)^2 + [\gamma(1+r)^2 - 1]\sum_k \alpha_k^2\right) \geq \gamma(1+r')^2[\gamma(1+r)^2 - 1](\gamma + \sum_k \alpha_k^2)$$

which happens when

$$\begin{aligned}
\big(\gamma(1+r')^2 - K\big)\left([\gamma(1+r)^2 - 1]\sum_k \alpha_k^2\right) &- K\gamma^2(1+r)^2 \\
&\geq \gamma(1+r')^2[\gamma(1+r)^2 - 1](\sum_k \alpha_k^2) - \gamma^2(1+r')^2 \\
\iff \gamma^2[(1+r')^2 - K(1+r)^2] &\geq K\left([\gamma(1+r)^2 - 1]\sum_k \alpha_k^2\right).
\end{aligned}$$

i.e., when $\gamma \geq \gamma_*$ where γ_* is the largest value of γ for which equality holds in the above inequality. This follows since the coefficient of γ^2 on the LHS is positive, by our choice of r, r'.

Case 2 ($\rho = 0$): Once again using (16) and (17) we get that in this case:

$$(\mathbf{u}_P^E)^\top \mathcal{A}_P^{-1} \mathbf{u}_P^E = \frac{m^2_{F_{\gamma/K}}(-\lambda')\alpha_K^2}{m'_{F_{\gamma/K}}(-\lambda')} \cdot \frac{\alpha_K^2}{\alpha_K^2 + \gamma} = \frac{\gamma(1+r')^2 - K}{\gamma} \cdot \frac{\alpha_K^4}{\alpha_K^2 + \gamma}$$

and

$$(\mathbf{u}^E)^\top \mathcal{A}^{-1} \mathbf{u}^E = \frac{m^2_{F_\gamma}(-\lambda)\alpha_K^2}{m'_{F_\gamma}(-\lambda)} \cdot \frac{\alpha_K^2}{\alpha_K^2 + \gamma} = \frac{\gamma(1+r)^2 - 1}{\gamma} \cdot \frac{\alpha_K^4}{\alpha_K^2 + \gamma}$$

where we use the definitions of $D_{\rho,\widetilde{\gamma}}$ and $D_{\rho,\gamma,ind}$ for $\rho = 0$. Thus the pooled covariance matrix leads to the better estimator when

$$\gamma[(1+r')^2 - (1+r)^2] > K - 1.$$

■

Proof [Proof of Proposition 6.4] Let us first discuss the estimation of $\mathcal{Y}_{Kk'}$ for $k' \neq K$. By Theorem 1 of Serdobolskii (2007) we have the deterministic equivalence

$$(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1} \asymp (x_p \Sigma_K + \lambda_K \mathbb{I}_p)^{-1}$$

for $x_p = x(\gamma_K, \lambda_K)$ as specified earlier. Thus

$$\begin{aligned}
& \operatorname{tr}[(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}\Sigma_K]/p \\
= \; & \operatorname{tr}[(x_p \Sigma_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}\Sigma_K]/p + \Omega_n \\
= \; & \frac{1}{p x_p} \operatorname{tr}[(x_p \Sigma_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}(x_p \Sigma_K + \lambda_K \mathbb{I}_p - \lambda_K \mathbb{I}_p)] + \Omega_n \\
= \; & \frac{1}{p x_p} \operatorname{tr}[(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}] - \frac{\lambda_K}{p x_p} \operatorname{tr}[(x_p \Sigma_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}] + \Omega_n \\
= \; & \frac{1}{p x_p} \operatorname{tr}[(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}] - \frac{\lambda_K}{p x_p} \operatorname{tr}[(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}] + \Omega_n
\end{aligned}$$

for a sequence $\Omega_n \to_{a.s.} 0$ as p, $n_k \to \infty$ with $p/n_k = \gamma_k$ for $k = 1, \ldots, K$. Here the first and last equalities follow by the deterministic equivalent for $(\widehat{\Sigma}_K + \lambda_K \mathbb{I}_p)^{-1}$ quoted above. For estimating $\mathcal{Y}_{kk'}$ where $k \neq K$ and $k' \neq K$, the result follows by the definition of $\widehat{\mathcal{Y}}_{kk'}$ and then using Lemma A.4 along with the independence of $\mathbf{X}_k$ for $k = 1, \ldots, K$. ■

A.2 Technical Lemmas and Their Proofs

Lemma A.4 *For a deterministic matrix $B \in \mathbb{R}^{p \times p}$ with operator norm $\|B\| \leq c$, the sample and population covariance matrices, $\widehat{\Sigma}_k$ and Σ_k, for $k = 1, \cdots, K$; satisfy:*

$$\operatorname{tr}(B(\widehat{\Sigma}_k - \Sigma_k))/p \to_{a.s.} 0.$$

Proof [Proof of Lemma A.4] Denoting the rows of $\mathbf{X}_k$ and $\mathbf{Z}_k$ by $X_{k,i}$ and $Z_{k,i} \in \mathbb{R}^p$ for $i = 1, \ldots, n_k$, we have $X_{k,i} - \mathbb{E}(X_{k,i}) = \Sigma_k^{1/2} Z_{k,i}$, and hence

$$\begin{aligned}
\operatorname{tr}(B(\widehat{\Sigma}_k - \Sigma_k))/p &= \frac{1}{pn_k} \sum_{i=1}^{n_k} \operatorname{tr}(B((X_{k,i} - \mathbb{E}(X_{k,i}))(X_{k,i} - \mathbb{E}(X_{k,i}))^\top - \Sigma_k)) \\
&= \frac{1}{pn_k} \sum_{i=1}^{n_k} \left(Z_{k,i}^\top \Sigma_k^{1/2} B \Sigma_k^{1/2} Z_{k,i} - \mathbb{E}[Z_{k,i}^\top \Sigma_k^{1/2} B \Sigma_k^{1/2} Z_{k,i}] \right) \\
&= \frac{1}{pn_k} \sum_{i=1}^{n_k} T_{k,i}
\end{aligned}$$

where for $i = 1, \ldots, n_k$,

$$T_{k,i} := Z_{k,i}^\top \Sigma_k^{1/2} B \Sigma_k^{1/2} Z_{k,i} - \mathbb{E}[Z_{k,i}^\top \Sigma_k^{1/2} B \Sigma_k^{1/2} Z_{k,i}]$$

are i.i.d. random variables with $\mathbb{E}T_{k,i} = 0$. By assumption 7, we have

$$\mathbb{E}[Z_{k,i}^\top \Sigma_k^{1/2} B \Sigma_k^{1/2} Z_{k,i}] = \operatorname{tr}(\Sigma_k^{1/2} B \Sigma_k^{1/2}).$$

We now compute the variance, writing $B' = \Sigma_k^{1/2} B \Sigma_k^{1/2}$ and using moment assumptions on Z from assumption 7:

$$\begin{aligned}
&\operatorname{Var}(T_{k,i}) \\
&= \operatorname{Var}\left(\sum_{l_1, l_2} B'_{l_1 l_2} Z_{l_1} Z_{l_2} \right) \\
&= \sum_{l_1, l_2, l_3, l_4} \mathbb{E}(B'_{l_1 l_2} B'_{l_3 l_4} Z_{l_1} Z_{l_2} Z_{l_3} Z_{l_4}) - (\operatorname{tr}(B'))^2 \\
&= \sum_{l_1=1}^{p} \sum_{l_3=1}^{p} (B'_{l_1 l_1})(B'_{l_3 l_3}) + \sum_{l_1=1}^{p} \sum_{l_2=1}^{p} (B'_{l_1 l_2})^2 + \sum_{l_1=1}^{p} \sum_{l_2=1}^{p} (B'_{l_1 l_2})(B'_{l_2 l_1}) + \sum_{l_1=1}^{p} (B'_{l_1 l_1})^2 (\mathbb{E}Z_{l_1}^4 - 1) - (\operatorname{tr}(B'))^2 \\
&= 2\|B'\|_{\mathrm{F}}^2 + \sum_{l_1=1}^{p} (B'_{l_1 l_1})^2 (\mathbb{E}Z_{l_1}^4 - 1) \leq Cp
\end{aligned}$$

for a constant $C > 0$. The last line follows since

$$\|B'\|_{\mathrm{F}}^2 \leq p \|\Sigma_k^{1/2} B \Sigma_k^{1/2}\|^2 \leq Cp$$

due to our assumption on $\|B\|$. Thus by Chebyshev inequality, for any $t > 0$, we have

$$\mathbb{P}\left(|*| \frac{1}{pn_k} \sum_{i=1}^{n_k} T_{k,i} > t \right) \leq \frac{1}{p^2 n_k t^2} \operatorname{Var}(T_{k,i}) \leq \frac{C}{pn_k t^2} = \frac{C}{\gamma_k n_k^2 t^2},$$

from where the almost sure convergence follows via Borel-Cantelli lemma. ■

Lemma A.5 *Under assumptions 3-2, consider a matrix $\mathbf{Z} \in \mathbb{R}^{p\times p}$ whose entries Z_{ij} have finite $(8+\epsilon)$-th moment for some $\epsilon > 0$. Suppose further that $\mathbf{Z}$ is independent of δ_k, $k, k' = 1, \cdots, K; k \neq k'$, we have as $p \to \infty$*

$$\delta_k^\top A \delta_{k'} - \rho_{kk'}\alpha_k\alpha_{k'} \operatorname{tr}(\mathbf{Z})/p \to_{a.s.} 0$$
$$\delta_k^\top A \delta_k - \alpha_k^2 \operatorname{tr}(\mathbf{Z})/p \to_{a.s.} 0$$

Proof [Proof of Lemma A.5] When $k = k'$, this lemma is the same as Lemma C.3 in Dobriban and Wager (2018) and theorem 2 in Sheng and Dobriban (2020). When $k \neq k'$, the same results still holds trivially under the bounded moments condition of A. This result has already been used by theorem 3.1 and theorem 4.1 of Zhao et al. (2023). ■

Lemma A.6 *Under the assumption 3, recall the definition of sample covariance matrix $\widehat{\Sigma} = (\mathbf{X} - \mathbf{1}_n\overline{X}^\top)^\top(\mathbf{X} - \mathbf{1}_n\overline{X}^\top)/n$ and its companion $\underline{\widehat{\Sigma}} = (\mathbf{X} - \mathbf{1}_n\overline{X}^\top)(\mathbf{X} - \mathbf{1}_n\overline{X}^\top)^\top/p$; we have*

$$\begin{aligned}
\operatorname{tr}[(\widehat{\Sigma} + \lambda\mathbb{I}_p)^{-1}]/p &\to_{a.s.} m_{F_\gamma}(-\lambda)\\
\operatorname{tr}[(\underline{\widehat{\Sigma}} + \lambda\mathbb{I}_p)^{-1}]/n &\to_{a.s.} v_{F_\gamma}(-\lambda)\\
\operatorname{tr}[(\widehat{\Sigma} + \lambda\mathbb{I}_p)^{-2}]/p &\to_{a.s.} m'_{F_\gamma}(-\lambda)\\
\operatorname{tr}[(\underline{\widehat{\Sigma}} + \lambda\mathbb{I}_p)^{-2}]/n &\to_{a.s.} v'_{F_\gamma}(-\lambda)\\
\operatorname{tr}[(\widehat{\Sigma} + \lambda\mathbb{I}_p)^{-1}\Sigma]/p &\to_{a.s.} \frac{1}{\gamma}\left(\frac{1}{\lambda v_{F_\gamma}(-\lambda)} - 1\right)\\
\operatorname{tr}[(\widehat{\Sigma} + \lambda\mathbb{I}_p)^{-2}\Sigma]/p &\to_{a.s.} \frac{1}{\gamma}\left(\frac{v_{F_\gamma}(-\lambda) - \lambda v'_{F_\gamma}(-\lambda)}{[\lambda v_{F_\gamma}(-\lambda)]^2}\right)\\
\operatorname{tr}[(\widehat{\Sigma} + \lambda\mathbb{I}_p)^{-2}\Sigma^2]/p &\to_{a.s.} \frac{1}{\gamma}\left(\frac{v'_{F_\gamma}(-\lambda) - v^2_{F_\gamma}(-\lambda)}{\lambda^2 v^4_{F_\gamma}(-\lambda)}\right).
\end{aligned}$$

Proof [Proof of Lemma A.6] The first four convergence statements follow from Marchenko and Pastur (1967) and Silverstein (1995). The convergence of last three trace terms are from Lemma 2 of Ledoit and Péché (2011), Lemma 2.2 of Dobriban and Wager (2018) and Lemma 3.11 of Dobriban and Wager (2018). ■

Lemma A.7 *Assume $n_1, \cdots, n_K, p \to \infty$, $p/n_k \to \gamma_k$ for $k = 1, \cdots, K$ and assumption 3. We have*

$$E_{kk'} := \operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k\mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}]/p \to_{a.s.} \mathcal{E}_{kk'}.$$

(1). Assume $n_1 = \cdots = n_K = n$, so $\gamma_1 = \cdots = \gamma_K = \gamma$, and use $\lambda_1 = \cdots = \lambda_K = \lambda$. Thus for all k, $m_{F_{\gamma_k}}(\lambda_k) = m_{F_\gamma}(\lambda)$. We have for $k \neq k'$

$$\mathcal{E}_{kk'} = \frac{(1-\gamma)m'_{F_\gamma}(-\lambda) + 2\gamma\lambda m_{F_\gamma}(-\lambda)m'_{F_\gamma}(-\lambda) - \gamma m_{F_\gamma}(-\lambda)^2}{1 - \gamma + \gamma\lambda^2 m'_{F_\gamma}(-\lambda)}.$$

(2). Under the assumption $\Sigma = \mathbb{I}_p$,

$$\mathcal{E}_{kk'} = m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'}).$$

(3). Under the Assumptions 4 and 5, we have

$$\begin{aligned}\mathcal{E}_{kk'} &= \frac{1}{\lambda_k \lambda_{k'}} \left\{ \lambda_k m_{F_{\gamma_k}}(-\lambda_k) + \lambda_{k'} m_{F_{\gamma_{k'}}}(-\lambda_{k'}) + \frac{\lambda_k m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'})}{(m_{F_{\gamma_k}}(-\lambda_k) - m_{F_{\gamma_{k'}}}(-\lambda_{k'}))} \right. \\ &\quad \left. - \frac{\lambda_{k'} m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'})}{(m_{F_{\gamma_k}}(-\lambda_k) - m_{F_{\gamma_{k'}}}(-\lambda_{k'}))} \right\}.\end{aligned}$$

Proof [Proof of Lemma A.7] In the first case, when $n_1 = \cdots = n_K = n$ so $\gamma_1 = \cdots = \gamma_K = \gamma$ and use $\lambda_1 = \cdots = \lambda_K = \lambda$, the limits of term $E_{kk'}$ has been found in the proof theorem 3 in Sheng and Dobriban (2020). When $\Sigma = \mathbb{I}_p$, the term $E_{kk'}$ boils down to

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p = \mathrm{tr}[(\mathbf{Z}_k^\top \mathbf{Z}_k / n_k + \lambda_k \mathbb{I}_p)^{-1} (\mathbf{Z}_{k'}^\top \mathbf{Z}_{k'} / n_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p]$$

Note we always have

$$E_{kk'} \to_{a.s.} \mathbb{E}_H \frac{1}{(x_k T + \lambda_k)(x_{k'} T + \lambda_{k'})}$$

Recall H is the limiting population spectral distribution, and x_k is the fixed point solution to

$$1 - x_k = \gamma_k \left[1 - \lambda_k \int \frac{1}{x_k t + \lambda_k} dH(t) \right] \tag{18}$$

When $\Sigma = \mathbb{I}_p$, H only has a point mass on 1 so the expectation decomposes and

$$E_{kk'} \to_{a.s.} m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'})$$

As an alternative proof when we have Assumption 4; we know $\mathbf{Z}_k$ will be asymptotically free from any bounded constant matrices, this is a standard result, ex. theorems 5.4.5 in Anderson et al. (2010). Further, we know sample covariances of the form $\mathbf{Z}_k^\top \mathbf{Z}_k / n_k$ is asymptotically free from $\mathbf{Z}_{k'}^\top \mathbf{Z}_{k'} / n_{k'}$ [see Capitaine and Casalis (2004)]. Two arguments combined suggests that $(\mathbf{Z}_k^\top \mathbf{Z}_k / n_k + \lambda_k \mathbb{I}_p)^{-1}$ is asymptotically free from $(\mathbf{Z}_{k'}^\top \mathbf{Z}_{k'} / n_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}$ therefore,

$$E_{kk'} - m_{F_{\gamma_k}}(-\lambda_k) m_{F_{\gamma_{k'}}}(-\lambda_{k'}) \to_{a.s.} 0$$

For the third case, a slight generalization of Corollary 3.9 of Knowles and Yin (2017) tells us

$$\ell_1^\top \left[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} - \frac{1}{\lambda_k (1 + m_{F_{\gamma_k}}(-\lambda) \Sigma)} \right] \ell_2 \to_{a.s.} 0$$

where ℓ_1, ℓ_2 can be any continuous random vectors independent from $(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}$. We can decompose $E_{kk'}$ by

$$\underbrace{\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1}]/p}_{E_{kk'}} - \frac{1}{\lambda_k} \sum_{i=1}^{p} \ell_{1,i}^\top (\mathbb{I}_p + m_{F_{\gamma_k}}(-\lambda) \Sigma)^{-1} \ell_{2,i} / p \to_{a.s.} 0$$

where $\ell_{1,i}$ is e_i with 1 in its i^{th} entry and 0 else where; and $\ell_{2,i} := (\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}e_i$. From now on, simplify the notation by using $m_k := m_{F_{\gamma_k}}(-\lambda_k)$ and $m_{k'} := m_{F_{\gamma_{k'}}}(-\lambda_{k'})$. Perform the similar trick to $\ell_{2,i}$, we have

$$E_{kk'} - \frac{1}{\lambda_k\lambda_{k'}}\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}(\mathbb{I}_p + m_{k'}\Sigma)^{-1})/p \to_{a.s.} 0$$

In addition

$$\begin{aligned}
&\frac{1}{\lambda_k\lambda_{k'}}\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}(\mathbb{I}_p + m_{k'}\Sigma)^{-1})/p \\
=\ &\frac{1}{\lambda_k\lambda_{k'}}[1 - m_k\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}\Sigma)/p - m_{k'}\operatorname{tr}((\mathbb{I}_p + m_{k'}\Sigma)^{-1}\Sigma)/p \\
&+ m_k m_{k'}\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}\Sigma(\mathbb{I}_p + m_{k'}\Sigma)^{-1}\Sigma)]/p
\end{aligned}$$

where we used the matrix identity

$$(\mathbb{I}_p + m_{F_{\gamma_k}}(-\lambda_k)\Sigma)^{-1} = \mathbb{I}_p - m_{F_{\gamma_k}}(-\lambda_k)(\mathbb{I}_p + m_k(-\lambda_k)\Sigma)^{-1}\Sigma$$

Each of the terms can be expressed in empirical quantities by

$$\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}\Sigma)/p - \mathbb{E}_H\frac{1}{m_k(1+tm_k)} + \mathbb{E}_H\frac{1}{m_k(1+tm_k)} - \frac{1}{m_k}[1 - \lambda_k m_k] \to_{a.s.} 0$$

With the same techniques, we get

$$\begin{aligned}
&\operatorname{tr}((\mathbb{I}_p + m_k\Sigma)^{-1}\Sigma(\mathbb{I}_p + m_{k'}\Sigma)^{-1}\Sigma)/p \\
\to_{a.s.}&\frac{\lambda_k m_k}{m_k(m_k - m_{k'})} - \frac{\lambda_{k'}m_{k'}}{m_{k'}(m_k - m_{k'})} + \frac{1}{m_k m_{k'}}
\end{aligned}$$

Substitute these expressions back into the expressions for $E_{kk'}$ finishes the proof. ■

Lemma A.8 *With assumption 3, and under $n_k, p \to \infty$, $p/n_k \to \gamma_k$, we have the following convergence results*

$$\operatorname{tr}[(\widehat{\Sigma}_k + \lambda_k\mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\Sigma]/p \to_{a.s.} \mathcal{M}_{kk'}.$$

For $k \neq k'; k, k' \in \{1, \cdots, K\}$, we have:

1. *Assume $n_1 = \cdots = n_K = n$, so $\gamma_1 = \cdots = \gamma_K = \gamma$, and use $\lambda_1 = \cdots = \lambda_K = \lambda$.*

$$\mathcal{M}_{kk'} = \frac{m_{F_\gamma}(-\lambda) - \lambda m'_{F_\gamma}(-\lambda)}{1 - \gamma + \gamma\lambda^2 m'_{F_\gamma}(-\lambda)}.$$

2. *Under the assumption $\Sigma = \mathbb{I}_p$,*

$$\mathcal{M}_{kk'} = m_{F_{\gamma_k}}(-\lambda_k)m_{F_{\gamma_{k'}}}(-\lambda_{k'}).$$

3. Under Assumptions 4 and 5,

$$\mathcal{M}_{kk'} = \frac{\lambda_k m_{F_{\gamma_k}}(-\lambda_k) - \lambda_{k'} m_{F_{\gamma_{k'}}}(-\lambda_{k'})}{\lambda_k \lambda_{k'}(m_{F_{\gamma_{k'}}}(-\lambda_{k'}) - m_{F_{\gamma_k}}(-\lambda_k))}.$$

Proof [Proof of Lemma A.8] The proof is similar to the proof of lemma A.7. In the first case where $n_1 = \cdots = n_K = n$ so $\gamma_1 = \cdots = \gamma_K = \gamma$ and use $\lambda_1 = \cdots = \lambda_K = \lambda$, we have

$$\mathrm{tr}[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\Sigma]/p \to_{a.s.} \int \frac{t}{(xt+\lambda)^2} dH(t)$$

When γ is equal across all populations, we will use the shorter notation $m := m_{F_\gamma}(-\lambda), m' = m'_{F_\gamma}(-\lambda)$ in this supplement. By definitions, we have

$$\int \frac{1}{xt+\lambda} dH(t) := m \text{ and } \int \frac{x't+1}{(xt+\lambda)^2} dH(t) := m'.$$

Here $x := x_k$ is the solution to the fixed point equation in (18), and x' is the derivative of x with respect to λ. Then

$$m' = \int \frac{(xt+\lambda-\lambda)\frac{x'}{x}+1}{(xt+\lambda)^2} dH(t) = \frac{x'}{x} m + (1 - \frac{\lambda x'}{x}) \int \frac{1}{(xt+\lambda)^2} dH(t)$$

$$\int \frac{1}{(xt+\lambda)^2} dH(t) = \frac{xm' - x'm}{x - \lambda x'}$$

So the functional of interest is

$$\begin{aligned}
\int \frac{t}{(xt+\lambda)^2} dH(t) &= \frac{\int \frac{x't+1}{(xt+\lambda)^2} dH(t) - \int \frac{1}{(xt+\lambda)^2} dH(t)}{x'} \\
&= \frac{m' - \frac{xm'-x'm}{x-\lambda x'}}{x'} = \frac{\frac{m'x - \lambda m'x' - xm' + x'm}{x-\lambda x'}}{x'} \\
&= \frac{m - \lambda m'}{x - \lambda x'} = \frac{m - \lambda m'}{1 - \gamma + \gamma \lambda^2 m'}.
\end{aligned}$$

For the second case, when $\Sigma = \mathbb{I}_p$, $\mathcal{M}_{kk'} = \mathcal{E}_{kk'}$. So the proof follows from lemma A.7. For the third case, pulling the trick with results from Knowles and Yin (2017) again gives

$$\mathrm{tr}(\Sigma(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1})/p \to_{a.s.} \frac{1}{\lambda_k \lambda_{k'}} \mathbb{E}_H \frac{t}{(1+tm_k)(1+tm_{k'})}$$

which can be consistently estimated by

$$\frac{1}{\lambda_k \lambda_{k'}} \frac{\lambda_k m_k - \lambda_{k'} m_{k'}}{m_{k'} - m_k}$$

as claimed by the lemma. ■

A.3 Proofs of Theorems

Proof [Proof of Theorem 4.1] Firstly, when $\pi_- = \pi_+$, we have

$$Err(\mathbf{w}) = \Phi\left(\frac{(\widehat{d}(\mathbf{w}))^\top \mu_{-1} + \widehat{b}_K}{\sqrt{(\widehat{d}(\mathbf{w}))^\top \Sigma(\widehat{d}(\mathbf{w}))}}\right)$$

Under Assumptions 1 and 2; by Lemma 3.7 in Dobriban and Wager (2018), we know $\widehat{b} \to_{a.s.} 0$. By Lemma 3.8 in the same paper, we can use the almost sure limit $(\widehat{d}(\mathbf{w}))^\top \mu_{-1} \to_{a.s.} (\widehat{d}(\mathbf{w}))^\top \delta_K$ and arrive at

$$Err(\mathbf{w}) \to_{a.s.} \Phi\left(\frac{(\widehat{d}(\mathbf{w}))^\top \delta_K}{\sqrt{(\widehat{d}(\mathbf{w}))^\top \Sigma(\widehat{d}(\mathbf{w}))}}\right).$$

Observe that

$$(\widehat{d}(\mathbf{w}))^\top \delta_K = \mathbf{w}^\top \widehat{\mathbf{u}}, \quad \widehat{\mathbf{u}} = vec\left[\widehat{\delta}_k (\widehat{\Sigma}_k + \lambda_k)^{-1} \delta_K\right]$$

$$(\widehat{d}(\mathbf{w}))^\top \Sigma(\widehat{d}(\mathbf{w})) = \mathbf{w}^\top \widehat{\mathcal{A}} \mathbf{w}, \quad \widehat{\mathcal{A}} = mat\left[\widehat{\delta}_k^\top (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \Sigma (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} \widehat{\delta}_{k'}\right]$$

Here $vec[\cdot], mat[\cdot]$ are the vector operator and matrix operator respectively. As argued in the proof of Lemma 3.7 in Dobriban and Wager (2018), one can decompose $\widehat{\delta}_k$ in the k^{th} population as

$$\widehat{\delta}_k = \delta_k + \frac{1}{\sqrt{n_k}} \Sigma^{1/2} \widetilde{\mathbf{Z}}_k$$

where $\widetilde{\mathbf{Z}}_k \in \mathbb{R}^p$ are standard normal random vectors independent of $\mathbf{X}_{k'}$ conditionally on $\mu_{\pm 1,k'}, \delta_{k'}$ for all $k' = 1, \cdots, K$. Thus, Lemma A.5 and Lemma A.6 gives us

$$\widehat{\mathbf{u}} \to_{a.s.} \rho_{kK} \alpha_k \alpha_K \; vec\left[m_{F_{\gamma_k}}(-\lambda_k)\right]$$

We decompose $\widehat{\mathcal{A}}$ into three parts for further analysis, such that

$$\widehat{\mathcal{A}}_{kk'} = \widehat{\delta}_k^\top (\widehat{\Sigma}_k + \lambda_k)^{-1} \Sigma (\widehat{\Sigma}_{k'} + \lambda_{k'})^{-1} \widehat{\delta}_{k'} = \widetilde{A}_{kk'} + 2\widetilde{B}_{kk'} + \widetilde{C}_{kk'}$$

where

$$\begin{aligned}
M^{(kk')} &:= (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \Sigma (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} \\
\widetilde{A}_{kk'} &:= \delta_k^\top M^{(kk')} \delta_{k'} \\
\widetilde{B}_{kk'} &:= \delta_k^\top M^{(kk')} (\widehat{\delta}_{k'} - \delta_{k'}) \\
\widetilde{C}_{kk'} &:= (\widehat{\delta}_k - \delta_k)^\top M^{(kk')} (\widehat{\delta}_{k'}^\top - \delta_{k'})
\end{aligned}$$

Firstly, one can show $\widetilde{B}_{kk'} \to_{a.s.} 0$ with the same techniques used to prove $\widehat{b} \to_{a.s.} 0$. For the off-diagonal terms in $\widetilde{A}$, one can again invoke Lemma A.5 and Lemma A.6 to show

$$\widetilde{A}_{kk'} - \rho_{kk'} \alpha_k \alpha_{k'} \operatorname{tr}(M^{(kk')})/p \to_{a.s.} 0.$$

Next, the limit of $\mathrm{tr}(M^{(kk')})/p$ is provided by Lemma A.8. For the diagonal terms $\widetilde{A}_{kk}$, one can simply read off the limit of $\mathrm{tr}(M^{(kk)})/p$ from Lemma A.6. Finally, we have the following equality for $\widetilde{C}$ based on decomposing $\widehat{\delta}_k$:

$$\widetilde{C}_{kk'} = \frac{1}{n_k}\widetilde{Z}_k^\top \Sigma^{1/2} M^{(kk')} \Sigma^{1/2} \widetilde{Z}_{k'}$$

We know $\widetilde{C}_{kk'} \to_{a.s.} 0$ for $k \neq k'$ based on Lemma A.5 due to the independence between $\widetilde{Z}_k, \widetilde{Z}_{k'}$. For the diagonal terms of $\widetilde{C}$, we have

$$\widetilde{C}_{kk} - \gamma_k \,\mathrm{tr}(\Sigma M^{(kk)})/p \to_{a.s.} 0$$

and the limit of the trace of $\Sigma M^{(kk)}$ can again be read off from Lemma A.6. ■

Proof [Proof of Theorem 4.5] Recall the decomposition $\widehat{\delta}_k = \delta_k + \frac{1}{\sqrt{n_k}}\Sigma^{1/2}\widetilde{\mathbf{Z}}_k$, we then have from the objective in (3), that

$$\left\| \sum_{k=1}^K w_k (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \widehat{\delta}_k - \Sigma^{-1}\delta_K \right\|_2^2 = \mathbf{w}^\top \left(\widehat{\mathcal{A}}^E + \widehat{\mathcal{R}}^E \right) \mathbf{w} - 2(\widehat{\mathbf{u}}^E)^\top \mathbf{w} + \|\delta_K^\top \Sigma^{-1}\|_2^2$$

where

$$\begin{aligned}
\widehat{\mathcal{A}}^E &= mat\left[\delta_k^\top (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} \delta_{k'} \right] \\
\widehat{\mathcal{R}}^E &= mat\left[(\widehat{\delta}_k - \delta_k)^\top (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} (\widehat{\Sigma}_{k'} + \lambda_{k'} \mathbb{I}_p)^{-1} (\widehat{\delta}_{k'} - \delta_{k'}) \right] \\
\widehat{\mathbf{u}}^E &= \delta_K^\top \Sigma^{-1} vec\left[(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \delta_{k'} \right]
\end{aligned}$$

Taking the derivative with respect to $\mathbf{w}$, we get the finite-sample expression for optimal estimation weight

$$\widehat{\mathbf{w}}^E = \left(\widehat{\mathcal{A}}^E + \widehat{\mathcal{R}}^E \right)^{-1} \widehat{\mathbf{u}}^E$$

Taking $n, p \to \infty; p/n_k \to \gamma_k$, we can get the asymptotic expressions for each of the three terms above. Consider the linear term $\widehat{\mathbf{u}}^E$ first. By Lemma A.5

$$\widehat{\mathbf{u}}^E \to_{a.s.} \mathbf{u}^E = vec\left[\rho_{kK} \alpha_k \alpha_K \,\mathrm{tr}[\Sigma^{-1}(\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}]/p \right]$$

Again, the *anisotropic local law* tells us that:

$$\mathrm{tr}(\Sigma^{-1}(\widehat{\Sigma}_{c,k} + \lambda_k \mathbb{I}_p)^{-1})/p \to_{a.s.} \frac{1}{\lambda_k}\mathbb{E}_H\left[\frac{1}{t(1 + t m_k(-\lambda_k))} \right] = \frac{1}{\lambda_k}\mathbb{E}_H\left[\frac{1}{t} - \frac{m_k(-\lambda_k)}{1 + t m_k(-\lambda_k)} \right]$$

$$\begin{aligned}
\mathrm{tr}(\Sigma^{-1}(\widehat{\Sigma}_{c,k} + \lambda_k \mathbb{I}_p)^{-1})/p \to\ & \frac{1}{\lambda_k}\mathbb{E}_H\left[\frac{1}{t[1 + t m_k(-\lambda_k)]} \right] \\
&= \frac{1}{\lambda_k}\mathbb{E}_H\left[\frac{1}{t} - \frac{m_k(-\lambda_k)}{1 + t m_k(-\lambda_k)} \right]
\end{aligned}$$

$$\to \frac{1}{\lambda_k}\operatorname{tr}(\Sigma^{-1})/p - m_k(-\lambda_k)\operatorname{tr}[(\widehat{\Sigma}_{c,k}+\lambda_k\mathbb{I}_p)^{-1})]/p.$$

For diagonal terms of $\widehat{\mathcal{A}}^E$, we have

$$\widehat{\mathcal{A}}^{\mathbb{E}}_{kk} \to_{a.s.} \alpha_k^2 m_{F_{\gamma_k}}(-\lambda_k)$$

For off diagonal terms of $\widehat{\mathcal{A}}^E$, we have

$$\widehat{\mathcal{A}}^{\mathbb{E}}_{kk'} \to_{a.s.} \rho_{kk'}\alpha_k\alpha_{k'}\mathcal{E}_{kk'}$$

where $\mathcal{E}_{kk'}$ can be found in Lemma A.7. Lastly, we firstly observe that the asymptotic off-diagonal terms of $\widehat{\mathcal{R}}^E$ are zero as $\widetilde{\mathbf{Z}}_k$ and $\widetilde{\mathbf{Z}}_{k'}$ are independent (see Lemma A.5). For diagonal terms of R^E, we have

$$\widehat{\mathcal{R}}^E_{kk} = \widetilde{\mathbf{Z}}_k^\top\Sigma^{1/2}(\widehat{\Sigma}_k+\lambda_k\mathbb{I}_p)^{-1}(\widehat{\Sigma}_k+\lambda_k\mathbb{I}_p)^{-1}\Sigma^{1/2}\widetilde{\mathbf{Z}}_k \to_{a.s.} \frac{v_{F_{\gamma_k}}(-\lambda_k)-\lambda_k v'_{F_{\gamma_k}}(-\lambda_k)}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)^2}$$

Gathering the asymptotic expressions together, along with the uniqueness of the asymptotic minimizer, we finish the proof. ■

Proof [Proof of Theorem 4.10] The proof is similar to the proof of Theorem 4.5. From the objective function in (4), we have

$$\mathbb{E}_{x_0}\left\|[\sum_{k=1}^K w_k(\widehat{\Sigma}_c+\lambda_k\mathbb{I}_p)^{-1}\widehat{\delta}_k - \Sigma^{-1}\delta_K]^\top x_0\right\|_2^2 = \mathbf{w}^\top[\widehat{\mathcal{A}}^P+\widehat{\mathcal{R}}^P]\mathbf{w} - 2\mathbf{w}^\top\widehat{\mathbf{u}}^P + \|\delta_K^\top\Sigma^{-1/2}\|_2^2$$

where

$$\begin{aligned}
\widehat{\mathcal{A}}^P &= mat\left[\delta_k^\top(\widehat{\Sigma}_k+\lambda_k\mathbb{I}_p)^{-1}\Sigma(\widehat{\Sigma}_{k'}+\lambda_{k'}\mathbb{I}_p)^{-1}\delta_{k'}\right]\\
\widehat{\mathcal{R}}^P &= mat\left[(\widehat{\delta}_k-\delta_k)^\top(\widehat{\Sigma}_k+\lambda_k\mathbb{I}_p)^{-1}\Sigma(\widehat{\Sigma}_{k'}+\lambda_{k'}\mathbb{I}_p)^{-1}(\widehat{\delta}_{k'}-\delta_{k'})\right]\\
\widehat{\mathbf{u}}^P &= vec\left[\delta_K^\top(\widehat{\Sigma}_k+\lambda_k\mathbb{I}_p)^{-1}\delta_{k'}\right]
\end{aligned}$$

We can again take the derivative with respect to $\mathbf{w}$ and set it to zero, we get

$$\widehat{\mathbf{w}}^P = \left(\widehat{\mathcal{A}}^P+\widehat{\mathcal{R}}^P\right)^{-1}\widehat{\mathbf{u}}^P$$

Taking $n,p\to\infty; p/n_k\to\gamma_k$, we can get the expression for each of the three terms above. Consider the linear term $\widehat{\mathbf{u}}^P$, by Lemma A.5

$$\widehat{\mathbf{u}}^P \to_{a.s.} vec\left[\rho_{kK}\alpha_k\alpha_K m_{F_{\gamma_k}}(-\lambda_k)\right].$$

For diagonal terms of $\widehat{\mathcal{A}}^P$, we have

$$\widehat{\mathcal{A}}^P_{kk} \to_{a.s.} \frac{\alpha_k^2}{\gamma_k}\frac{v_{F_{\gamma_k}}(-\lambda_k)-\lambda v'_{F_{\gamma_k}}(-\lambda_k)}{[\lambda v_{F_{\gamma_k}}(-\lambda_k)]^2}$$

For off diagonal terms of $\widehat{\mathcal{A}}^P$, we have

$$\widehat{\mathcal{A}}^P_{kk'} \to_{a.s.} \rho_{kk'}\alpha_k\alpha_{k'}\mathcal{M}_{kk'}.$$

The off-diagonal terms of $\widehat{\mathcal{R}}^P$ again converges to zero, and the diagonals are

$$\widehat{\mathcal{R}}^P_{kk} = \widetilde{\mathbf{Z}}_k^\top \Sigma\Sigma^{1/2}(\widehat{\Sigma}_k + \lambda_k\mathbb{I}_p)^{-1}(\widehat{\Sigma}_k + \lambda_k\mathbb{I}_p)^{-1}\Sigma^{1/2}\widetilde{\mathbf{Z}}_k \to_{a.s.} \frac{v_k'(-\lambda_k) - v_k^2(-\lambda_k)}{\lambda_k^2 v_k^4(-\lambda_k)}.$$

Gathering the asymptotic expressions together, along with the uniqueness of the asymptotic minimizer, we finish the proof. ■

Proof [Proof of Theorem 6.1] The proof is identical to that of Theorem 4.5 with minimal changes due to the differing covariance matrices Σ_k. ■

Proof [Proof of Theorem 6.3] The proof is identical to that of Theorem 4.10 with minimal changes due to the differing covariance matrices Σ_k. ■

A.4 Proofs of Corollaries

Proof [Proof of Corollary A.1] We follow the proof of Theorem 4.10 for the special case $\Sigma = \mathbb{I}_p$. Consequently,

$$\widehat{\mathcal{A}}^P_{kk'} = \delta_k^\top(\widehat{\Sigma}_k + \lambda_k\mathbb{I}_p)^{-1}\Sigma(\widehat{\Sigma}_{k'} + \lambda_{k'}\mathbb{I}_p)^{-1}\delta_{k'} \to_{a.s.} \begin{cases} \alpha_k^2 m_k'(-\lambda_k) & \text{if } k = k' \\ \rho_{kk'}\alpha_k\alpha_{k'}m_k(-\lambda_k)m_{k'}(-\lambda_{k'}) & \text{otherwise} \end{cases}$$

using the limits from Lemma A.6. The rest of the expressions follow similarly. ■

Appendix B. Numerical Experiments

B.1 Validation of Limiting Error Formulae

In this subsection, we use synthetic data to verify the accuracy of the theoretical formula in Theorems 4.1, 4.5, 4.10. We use a Toeplitz-type covariance matrix, $n_1, \cdots, n_K = 150, 140, 130, 120, 110, 100$ and $p = 150$. We use the same signal strength for all population $\alpha^2 = 0.5$ and the same weight correlation $\rho = 0.5$. The prediction error is based on 2000 simulated testing data points. We then vary λ_k from 0.3 to 10. For each λ_k, the simulation is repeated 50 times. One can refer to figure 6 and figure 7 for the comparison between simulated testing error rate (boxplots) and theoretical error rate (red line) for TL-RDA and TL-RDA respectively. We can see that the limiting formulae are very accurate across the λ_k considered.

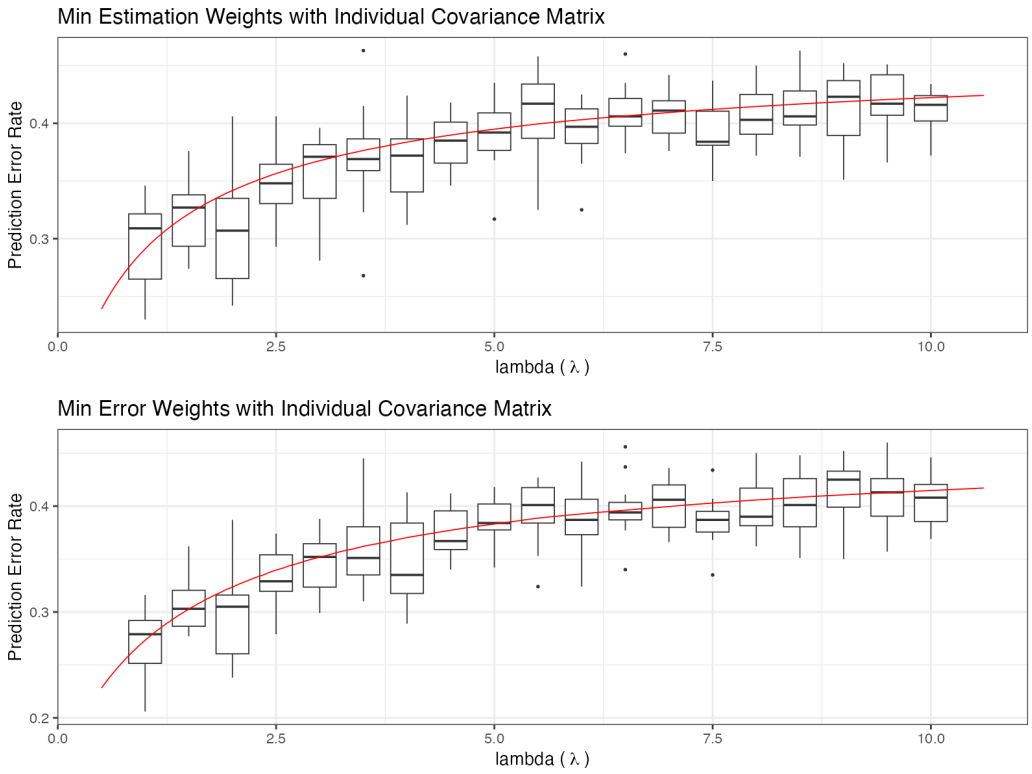

Figure 6: Individual sample covariance matrices

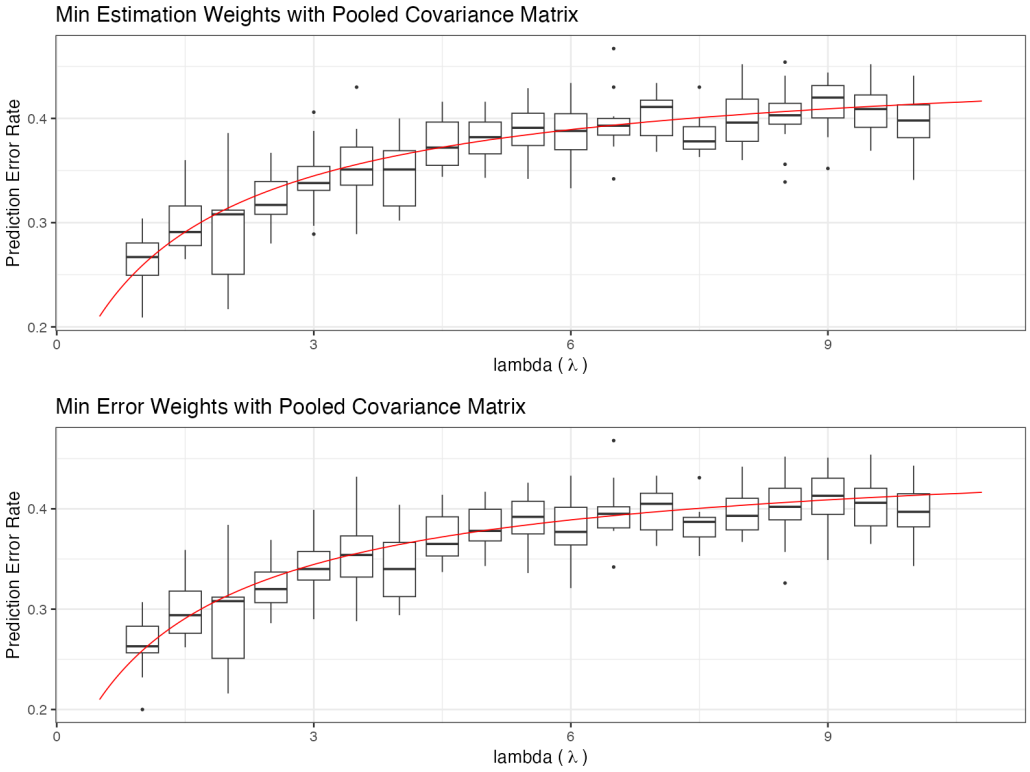

Figure 7: Pooled sample covariance matrices

B.2 Theoretical Error comparison

Here we plot the limiting errors of the naive RDA, TL-RDA with estimation weight and TL-RDA with prediction weight under different scenarios. Note the naive RDA is fitted on target population data only. One can refer to figure 8 and figure 9 for comparisons of TL-RDA and TLP-RDA respectively. We vary all parameters including correlation ρ, aspect ratio γ, signal strength α^2, number of populations K and population covariance matrix eigenvalues Σ. One can directly observe that TL estimators beat naive RDA in every set up considered, and not surprisingly, the prediction weight (minError) beats estimation weight (MinEst) in all set ups. Another interesting fact is estimation weight is usually not too much worse than prediction weight, especially in the TLP-RDA case. Lastly, we observe that the prediction errors turn to increase when γ_ks increase, decreases when α^2 increases, decrease when the number of populations increases. In addition, the error rate decreases when the eigenvalue of Σ decreases faster, that is, when the correlation strength between the features get stronger.

B.3 Estimators for Hyperparameters

We have assumed throughout this paper that $\alpha_k^2, \rho_{kk'}$ are known constants. We prove that the usual moment estimators, listed below, are consistent after debiasing.

$$\widehat{\alpha}_k^2 = \sum_{i=1}^{p} [(\widehat{\mu}_{+1,k})_i - (\widehat{\mu}_{-1,k})_i]^2/4 - \frac{1}{n}\text{trace}(\widehat{\Sigma}).$$

$$\widehat{\rho}_{kk'} = \sum_{i=1}^{p} [(\widehat{\mu}_{+1,k})_i - (\widehat{\mu}_{-1,k})_i][(\widehat{\mu}_{+1,k'})_i - (\widehat{\mu}_{-1,k'})_i]/4.$$

In order to prove the consistency of the estimators above, it is sufficient to prove proposition B.1 and proposition B.2.

Proposition B.1 (Variance of site mean) *For rows $X_1, \ldots, X_n \in \mathbb{R}^p$ in X generated according to set up 1, assumption 1 and assumption 3 for a single class. Define the coordinate-wise sample means $\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ and*

$$\widehat{\alpha}^2 := \sum_{j=1}^{p} \widehat{\mu}_j^2 - \frac{1}{n}\text{trace}(\widehat{\Sigma}).$$

Then

$$\mathbb{E}(\widehat{\alpha}^2) = \alpha^2.$$

Moreover, as $p, n \to \infty$ with $p/n \to \gamma$,

$$\widehat{\alpha}^2 \overset{\text{in probability}}{\longrightarrow} \alpha^2.$$

Proof Write $U = \widehat{\mu} - \mu$, so $U \mid \mu \sim \mathsf{N}(0, \Sigma/n)$ and $U \perp\!\!\!\perp \mu$. Then $\widehat{\mu} = \mu + U$, and for each j,

$$\mathbb{E}(\widehat{\mu}_j^2) = \text{Var}(\widehat{\mu}_j) = \mathbb{E}\big[\,\text{Var}(\widehat{\mu}_j \mid \mu)\big] + \text{Var}\left(\mathbb{E}[\widehat{\mu}_j \mid \mu]\right) = \frac{\Sigma_{jj}}{n} + \frac{\alpha^2}{p}.$$

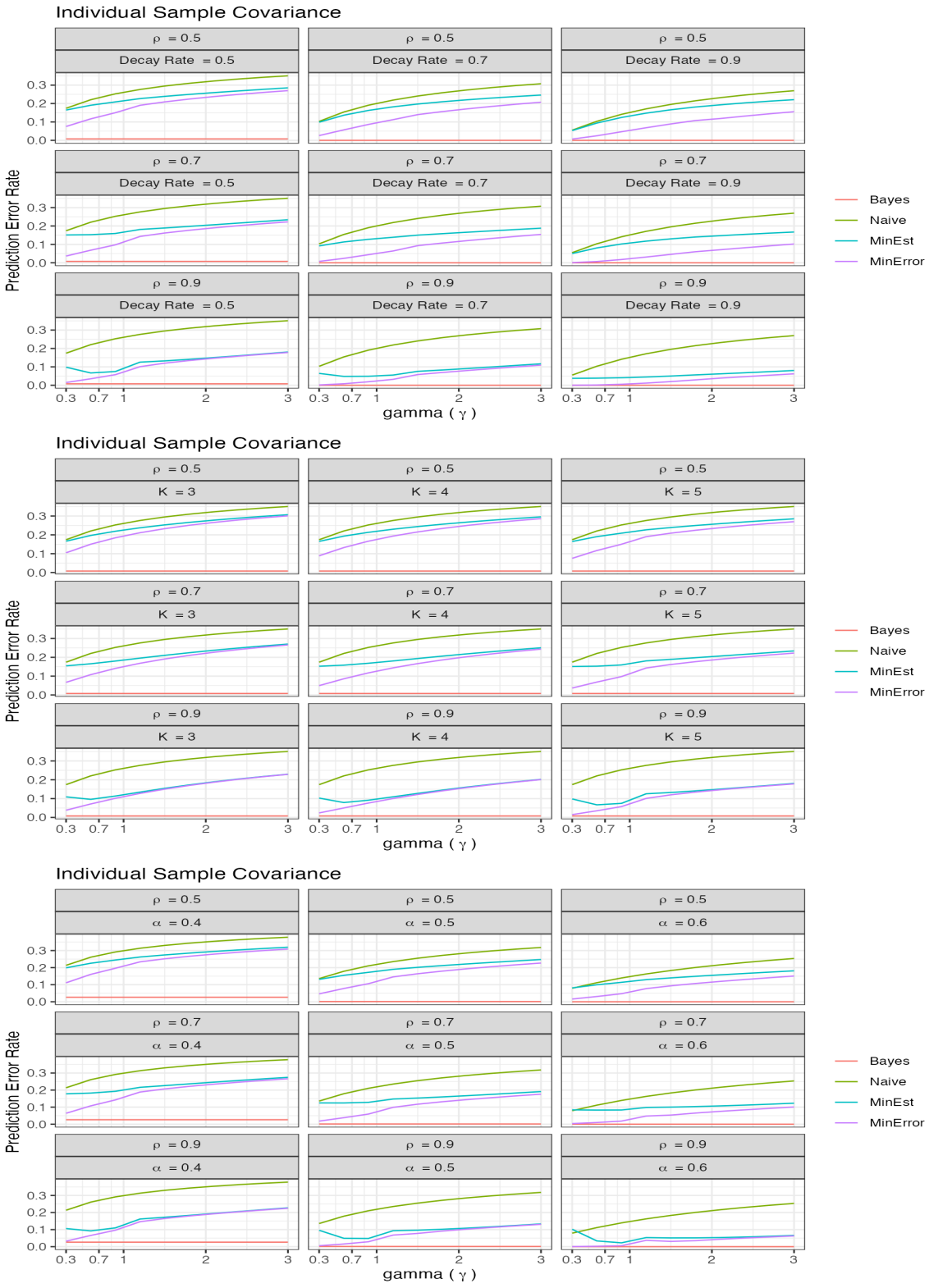

Figure 8: Individual sample covariance matrices

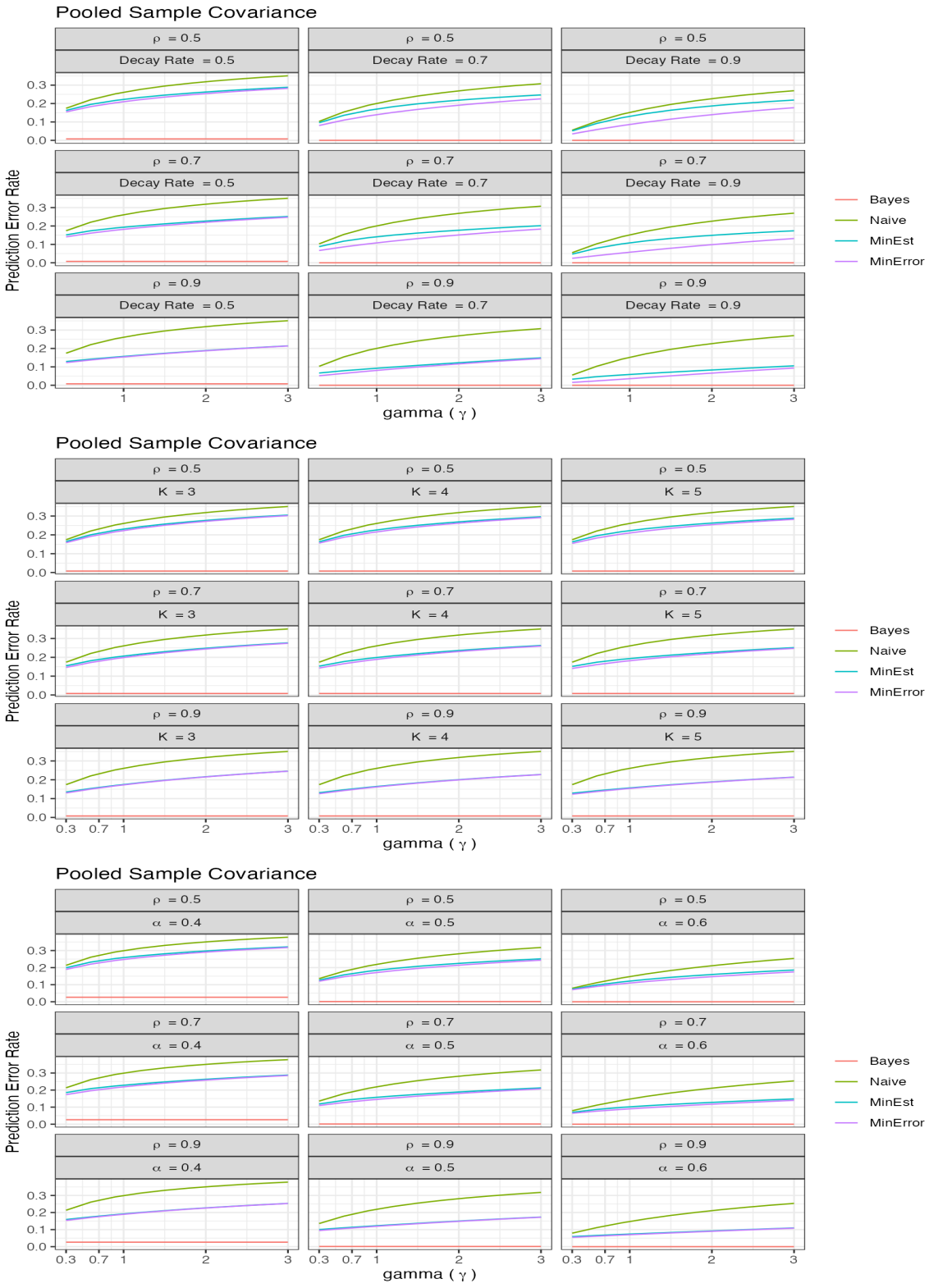

Figure 9: Pooled covariance matrices

Summing over j yields $\mathbb{E}(\widetilde{\alpha}^2) = \alpha^2$, where

$$\widetilde{\alpha}^2 = \sum_{j=1}^{p} \widehat{\mu}_j^2 - \operatorname{tr}(\Sigma)/n.$$

For the variance, expand

$$\operatorname{Var}(\widetilde{\alpha}^2) = \sum_{j=1}^{p} \operatorname{Var}(\widehat{\mu}_j^2) \; + \; 2 \sum_{1 \le j < k \le p} \operatorname{Cov}(\widehat{\mu}_j^2, \widehat{\mu}_k^2).$$

Since $\widehat{\mu}$ is mean-zero and jointly Gaussian conditional on μ and the μ's are independent with mean zero, one can either compute directly via $U = \widehat{\mu} - \mu$ and independence, or invoke standard Gaussian fourth-moment identities (Isserlis/Wick), to obtain unconditionally

$$\operatorname{Var}(\widehat{\mu}_j^2) = 2\left(\frac{\Sigma_{jj}}{n} + \frac{\alpha^2}{p}\right)^2, \qquad \operatorname{Cov}(\widehat{\mu}_j^2, \widehat{\mu}_k^2) = \frac{2}{n^2}\Sigma_{jk}^2, \quad j \neq k.$$

Therefore,

$$\begin{aligned}
\operatorname{Var}(\widetilde{\alpha}^2) &= 2\sum_{j=1}^{p}\Big(\frac{\Sigma_{jj}}{n} + \frac{\alpha^2}{p}\Big)^2 + 4 \sum_{1 \le j < k \le p}\Big(\frac{\Sigma_{jk}}{n}\Big)^2 \\
&= 2\sum_{j=1}^{p}\Big(\frac{\Sigma_{jj}^2}{n^2} + \frac{2\alpha^2\,\Sigma_{jj}}{pn} + \frac{\alpha^4}{p^2}\Big) + \frac{2}{n^2}\sum_{j \neq k}\Sigma_{jk}^2 \\
&= \frac{2}{n^2}\|\Sigma\|_{\mathrm{F}}^2 + \frac{4\alpha^2}{pn}\operatorname{tr}\Sigma + \frac{2\alpha^4}{p}.
\end{aligned}$$

Specifically, assumption 3 upper bounds the eigenvalue of Σ, say with a constant C. Then as $p, n \to \infty$, $p/n \to \gamma$,

$$\frac{2}{n^2}\|\Sigma\|_F^2 \;\le\; \frac{2pC^2}{n^2} \;=\; O\!\left(\frac{1}{n}\right).$$

the second and third terms are $O((1/p) \cdot p/n) = O(p^{-1})$, , and $O(p^{-1})$ respectively. Hence $\operatorname{Var}(\widetilde{\alpha}^2) \to 0$, and Chebyshev yields $\widetilde{\alpha}^2$ converges in probability to α^2. To finish the proof we show that $\widehat{\alpha}^2 - \widetilde{\alpha}^2$ converges to zero in probability. Indeed

$$\widehat{\alpha}^2 - \widetilde{\alpha}^2 = \; -\frac{1}{n}\sum_{j=1}^{p}(\widehat{\sigma}_{jj} - \sigma_{jj})$$

For any fixed $j = 1, \ldots, p$, by standard results for $X_{ij} \overset{indep}{\sim} N(0, \sigma_{jj})$ for $i = 1, \ldots, n$ we have that $\widehat{\sigma}_{jj} - \sigma_{jj}$ converges to zero in probability. Averaging over j shows that $(\widehat{\alpha}^2 - \widetilde{\alpha}^2)$ also converges to zero and this finishes the proof. ■

Proposition B.2 (Correlation of site means) *Fix two sites k, k' with sample sizes $n_k, n_{k'}$ and dimension p. For $s \in \{k, k'\}$, let rows $X_1^{(s)}, \dots, X_{n_s}^{(s)} \in \mathbb{R}^p$ be generated under set up 1, assumption 1, assumption 2, assumption 3 for a single class, and the scalars $\alpha_k^2, \alpha_{k'}^2$ are known. Let $\widehat{\mu}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} X_i^{(s)}$ be the sample mean for site s, and define*

$$\widehat{\rho}_{kk'} \;:=\; \frac{1}{\widehat{\alpha}_k \widehat{\alpha}_{k'}}\, \widehat{\mu}_k^\top \widehat{\mu}_{k'}.$$

where $\widehat{\alpha}_k$ are as defined in Proposition B.1. If as $p \to \infty$ we have $p/n_s \to \gamma_s \in (0, \infty)$,

$$\widehat{\rho}_{kk'} \overset{\text{in probability}}{\longrightarrow} \rho_{kk'}.$$

Proof Let us define

$$\widetilde{\rho}_{kk'} \;:=\; \frac{1}{\alpha_k \alpha_{k'}}\, \widehat{\mu}_k^\top \widehat{\mu}_{k'}$$

by replacing $\widehat{\alpha}_k$ with their in probability limits α_k in $\widehat{\rho}_{kk'}$. Write $\widehat{\mu}_s = \mu_s + \varepsilon_s$ with $\varepsilon_s := \widehat{\mu}_s - \mu_s$. Conditional on μ_s, $\varepsilon_s \sim \mathsf{N}(0, \Sigma/n_s)$ and $\varepsilon_k, \varepsilon_{k'}$ are independent of $(\mu_k, \mu_{k'})$ and of each other (across sites). Using $\mathbb{E}[\varepsilon_s] = 0$ and cross-site independence of the noises,

$$\mathbb{E}\big[\widehat{\mu}_k^\top \widehat{\mu}_{k'}\big] = \mathbb{E}\big[\mu_k^\top \mu_{k'}\big] = \operatorname{tr}\big(\operatorname{Cov}(\mu_k, \mu_{k'})\big) = \alpha_k \alpha_{k'}\, \rho_{kk'}.$$

Dividing by $\alpha_k \alpha_{k'}$ yields $\mathbb{E}[\widetilde{\rho}_{kk'}] = \rho_{kk'}$. Expand

$$\widehat{\mu}_k^\top \widehat{\mu}_{k'} - \mathbb{E}[\widehat{\mu}_k^\top \widehat{\mu}_{k'}] = \underbrace{\big(\mu_k^\top \mu_{k'} - \mathbb{E}[\mu_k^\top \mu_{k'}]\big)}_{A} + \underbrace{\mu_k^\top \varepsilon_{k'}}_{B} + \underbrace{\mu_{k'}^\top \varepsilon_k}_{C} + \underbrace{\varepsilon_k^\top \varepsilon_{k'}}_{D}.$$

The four terms are mean-zero and pairwise uncorrelated (by independence across sites and Gaussian odd-moment cancellations), so variances add. A standard Gaussian fourth-moment computation gives

$$\operatorname{Var}(A) \;=\; \frac{\alpha_k^2 \alpha_{k'}^2}{p}\,(1 + \rho_{kk'}^2), \qquad \operatorname{Var}(B) \;=\; \frac{\alpha_k^2}{p\, n_{k'}}\,\operatorname{tr}(\Sigma), \qquad \operatorname{Var}(C) \;=\; \frac{\alpha_{k'}^2}{p\, n_k}\,\operatorname{tr}(\Sigma),$$

and, since $\varepsilon_k \perp \varepsilon_{k'}$ with covariances Σ/n_k and $\Sigma/n_{k'}$,

$$\operatorname{Var}(D) \;=\; \frac{1}{n_k n_{k'}}\,\operatorname{tr}(\Sigma^2).$$

Therefore

$$\operatorname{Var}(\widetilde{\rho}_{kk'}) = \frac{1}{\alpha_k^2 \alpha_{k'}^2}\Big\{\frac{\alpha_k^2 \alpha_{k'}^2}{p}(1 + \rho_{kk'}^2) + \frac{\alpha_k^2}{p\, n_{k'}}\operatorname{tr}(\Sigma) + \frac{\alpha_{k'}^2}{p\, n_k}\operatorname{tr}(\Sigma) + \frac{1}{n_k n_{k'}}\operatorname{tr}(\Sigma^2)\Big\}.$$

Again, assumption 3 upper bounds the eigenvalue of Σ, say with a constant C. Then as $p, n_s \to \infty$, $p/n_s \to \gamma_s$,

$$\frac{1}{n_k n_{k'}}\,\operatorname{tr}(\Sigma^2) \;\le\; \frac{pC^2}{n_k n_{k'}} \;=\; O\!\left(\frac{1}{\min\{n_k, n_{k'}\}}\right)$$

and $\operatorname{tr}(\Sigma)/n_s = (\operatorname{tr}\Sigma/p)\cdot(p/n_s) = O(1)$, so each bracketed term is $o(1)$. Hence $\operatorname{Var}(\widetilde{\rho}_{kk'}) = O(p^{-1}) \to 0$, and Chebyshev yields $\widetilde{\rho}_{kk'} \overset{\text{in probability}}{\longrightarrow} \rho_{kk'}$. To complete the proof we use the consistency of $\widehat{\alpha}_k$ for α_k from Proposition B.1 and the continuous mapping theorem. ■

Appendix C. Limiting Error and Optimal Weight under Unequal Sampling

Recall the error rate under Assumption 1 is

$$Err(\mathbf{w}) = \pi_- \Phi\left(\frac{\widehat{d}(\mathbf{w})^\top \mu_{-1} + \widehat{b}_K}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}}\right) + \pi_+ \Phi\left(-\frac{\widehat{d}(\mathbf{w})^\top \mu_1 + \widehat{b}_K}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}}\right)$$

Now the stochastic representations become

$$\widehat{\delta}_k = \delta_k + \Sigma^{1/2}\left[\frac{\widetilde{\mathbf{Z}}_{k,+1}}{\sqrt{n_{k,+1}}} - \frac{\widetilde{\mathbf{Z}}_{k,-1}}{\sqrt{n_{k,-1}}}\right]$$

$$\widehat{\mu}_k = \bar{\mu}_k + \Sigma^{1/2}\left[\frac{\widetilde{\mathbf{Z}}_{k,+1}}{\sqrt{n_{k,+1}}} + \frac{\widetilde{\mathbf{Z}}_{k,-1}}{\sqrt{n_{k,-1}}}\right]$$

The intercept terms are no longer zeros and

$$\begin{aligned}
\widehat{b}_k =& -\widehat{\delta}_k^\top (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1}(\widehat{\mu}_{-1,k} + \widehat{\mu}_{+1,k})/2 \\
=& \frac{1}{4n_{k,-1}} \widetilde{\mathbf{Z}}_{k,-1}^\top \Sigma^{1/2} (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \Sigma^{1/2} \widetilde{\mathbf{Z}}_{k,-1} - \frac{1}{4n_{k,+1}} \widetilde{\mathbf{Z}}_{k,+1}^\top \Sigma^{1/2} (\widehat{\Sigma}_k + \lambda_k \mathbb{I}_p)^{-1} \Sigma^{1/2} \widetilde{\mathbf{Z}}_{k,+1} \\
\to_{a.s.} & \frac{\gamma_{k,-1} - \gamma_{k,+1}}{4\gamma_k} \left(\frac{1}{\lambda_k v_{F_{\gamma_k}}(-\lambda_k)} - 1\right)
\end{aligned}$$

Where we use $p/n_{k,\pm 1} \to \gamma_{k,\pm 1}$. For the numerator we have

$$\widehat{d}(\mathbf{w})^\top \mu_{\pm 1} = \mathbf{w}^\top vec\left[(\bar{\mu}_k \pm \widehat{\delta}_k)^\top (\widehat{\Sigma}_k + \lambda_k)^{-1} \delta_K\right] \to_{a.s.} \mathbf{w}^\top vec\left[\pm \rho_{kK} \alpha_k \alpha_K m_{F_{\gamma_k}}(-\lambda_k)\right]$$

For the denominator term,

$$\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w}) = \mathbf{w}^\top mat\left[\widehat{\delta}_k^\top (\widehat{\Sigma}_k + \lambda_k)^{-1} \Sigma (\widehat{\Sigma}_{k'} + \lambda_{k'})^{-1} \widehat{\delta}_{k'}\right] \mathbf{w} := \mathbf{w}^\top S \mathbf{w}.$$

For $k \neq k'$

$$S_{kk'} \to_{a.s.} \alpha_k^2 \frac{v_{F_{\gamma_k}}(-\lambda) - \lambda v'_{F_{\gamma_k}}(-\lambda)}{\gamma_k [\lambda v_{F_{\gamma_k}}(-\lambda)]^2} + \frac{\gamma_{k,-1} + \gamma_{k,+1}}{4} \frac{v'_{F_{\gamma_k}}(-\lambda) - v^2_{F_{\gamma_k}}(-\lambda)}{\lambda^2 v^4_{F_{\gamma_k}}(-\lambda)}$$

while

$$S_{kk} \to_{a.s.} \rho_{kk'} \alpha_k \alpha_{k'} \frac{\mathrm{tr}(M^{kk'})/p}{\gamma_k} + \frac{\gamma_{k,-1} + \gamma_{k,+1}}{4} \frac{v'_{F_{\gamma_k}}(-\lambda) - v^2_{F_{\gamma_k}}(-\lambda)}{\lambda^2 v^4_{F_{\gamma_k}}(-\lambda)}$$

for $1 \leq k, k' \leq K$. Again the limit of $\mathrm{tr}(M^{kk'})/p$ can be found in Lemma A.8.

As for the optimal weights, we know the Bayes optimal prediction is

$$d_{Bayes}(x_0) = \delta_K^\top \Sigma^{-1} x_0 + \delta_K^\top \Sigma^{-1} \bar{\mu}_K + \log\left(\frac{\pi_{K,+}}{\pi_{K,-}}\right).$$

Note δ_K and $\bar{\mu}_K$ are independent by Assumption 1, therefore by Lemma A.5 the second term is zero. In this case, we recommend an estimator in the form of $\sum_{k=1}^{K} \widehat{d}_k + \log(\pi_{K,+}/\pi_{K,-})$ whose intercept term is consistent with the one of the Bayes direction. Further, one can still obtain the optimal weights that minimize criteria (3) and (4). However, the proposition 4.14 no longer holds and the prediction weight is not guaranteed to minimize the testing data error rate. The error weight on testing data has the following expression.

$$\pi_{K,-}\Phi\left(\frac{\widehat{d}(\mathbf{w})^\top \mu_{K,-1}}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}} + \frac{\widehat{b}_K}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}}\right) + \pi_{K,+}\Phi\left(-\frac{\widehat{d}(\mathbf{w})^\top \mu_{K,1}}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}} - \frac{\widehat{b}_K}{\sqrt{\widehat{d}(\mathbf{w})^\top \Sigma \widehat{d}(\mathbf{w})}}\right)$$

The difficulty arises as $\widehat{b}_K$ is now non-zero, and merely maximizing the first fraction, as the optimal prediction weight does, is no longer sufficient. When K is low, one can potentially search for a local or even global optimal point for $\mathbf{w}$ with numeric methods.

Expected Prediction Error

For any linear classifier $\text{sign}(\mathbf{w}^\top x_0 + b)$, we can write its prediction error rate as followed

$$\begin{aligned}
&\mathbb{P}(\text{sign}(\mathbf{w}^\top x_0 + b) \neq y_0) \\
&= \pi_- \mathbb{P}(\text{sign}(\mathbf{w}^\top x_0 + b) \neq -1 | y_0 = -1) + \pi_+ \mathbb{P}(\text{sign}(\mathbf{w}^\top x_0 + b) \neq 1 | y_0 = 1)
\end{aligned}$$

The expressions for the two parts above are symmetric, so we can only look the first part.

$$\begin{aligned}
&\mathbb{P}(\text{sign}(\mathbf{w}^\top x_0 + b) \neq -1 | y_0 = -1) \\
&= \mathbb{P}(\mathbf{w}^\top \mu_{-1} + \mathbf{w}^\top \epsilon + b > 0) \\
&= \mathbb{P}(\mathbf{w}^\top \epsilon \leq \mathbf{w}^\top \mu_{-1} + b) = \Phi\left(\frac{\mathbf{w}^\top \mu_{-1} + b}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}\right)
\end{aligned}$$