



*Research article***Transfer learning for sparse linear discriminant analysis based on mean-variance similarity****Ping Wang¹, Xiaoyu Ma², Huanlin Li³ and Lu Lin^{3,*}**¹ Shandong Management University, Jinan, China² National University of Defense Technology, Changsha, China³ Shandong University, Jinan, China* **Correspondence:** Email: linsdu_2021@163.com, linlu@sdu.edu.cn.

Abstract: We study transfer learning for high-dimensional linear discriminant analysis (LDA) with auxiliary data. To describe the differences between the target and source domains, we introduce a mean–variance similarity condition based on class-specific means and covariance structures. We first derive an exact source-domain moment relation and use it to construct a transfer learning estimator, called Trans-LPD. For a fixed set of auxiliary studies, we establish estimation and misclassification-risk bounds, which show the effects of the enlarged sample size and transfer bias. A data-driven adaptive procedure is then proposed to rank the auxiliary studies and select the final classifier using a target validation sample. The adaptive Trans-LPD classifier is shown to be asymptotically no worse than the target-only method. Simulation studies and an application to TCGA gene-expression data are also presented to illustrate the numerical performance of the proposed method.

Keywords: transfer learning; linear programming discriminant; mean-variance similarity; high-dimensional data; moment equation

Mathematics Subject Classification: 62H30, 62R07

1. Introduction

The era of big data has produced vast sources of information, some of which are closely related. This raises the question of how to use such relationships to improve statistical inference. Transfer learning is a promising method for utilizing knowledge from related data sources to improve inference in target datasets [1]. It allows information to be transferred across domains, even when the domains differ substantially [2,3]. Experience in one domain can help learning in another. For example, playing the electric organ may make it easier to learn the piano [4], and knowledge of Latin can facilitate the

learning of Romance languages such as French. Transfer learning has been applied in medicine and biology [5–10], natural language processing [11], and recommender systems [12].

These developments have also received considerable attention in statistics. Cai and Wei [13] investigated minimax and adaptive methods for nonparametric classification in the context of transfer learning. Li et al. [14–17] considered related problems including estimation and prediction in high-dimensional linear models, proxy data inference, federated transfer learning in precision medicine, and Gaussian graphical models. Related work has also covered high-dimensional linear regression [18], high-dimensional generalized linear models [2, 19, 20], functional linear regression [21], and nonparametric models [3], among others.

Discriminant analysis is a classical tool for classification. Among its many forms, Linear Discriminant Analysis (LDA) has been widely used in applications such as face recognition and microarray gene expression analysis. For high-dimensional linear discriminant problems, existing studies have mainly focused on regularization methods and optimization theory. Regularization-based methods are developed in [22–27]. Studies from an optimization perspective include [28–30]. Less work has considered LDA in a transfer learning setting. Safo et al. [31] proposed Sparse Integrative Discriminant Analysis (SIDA), which combines association analysis and classification for multiview data. However, transfer learning aims to borrow information across related domains, whereas multiview learning focuses on integrating multiple views or modalities of the same subjects or task. This difference motivates us to study LDA under a transfer learning framework.

Our proposed approach is inspired by the innovative work of Li et al. [15]. We consider two p -dimensional normal distributions, $N(\mu_1, \Sigma)$ (denoted as class 1) and $N(\mu_2, \Sigma)$ (denoted as class 2), which share a common covariance matrix Σ . Let Z be a new observation randomly drawn from one of the above classes with equal prior probability. The task of discriminant analysis is to determine from which class Z has been drawn. When μ_1 , μ_2 , and Σ are known, Fisher's linear discriminant function is given by:

$$\psi(Z) = \mathbb{1} \{ (Z - \mu)^\top \Omega \delta \geq 0 \}, \quad (1.1)$$

where $\mu = (\mu_1 + \mu_2)/2$, $\delta = \mu_1 - \mu_2$, and $\Omega = \Sigma^{-1}$ denotes the precision matrix of the target domain throughout this paper. According to Fisher's linear discriminant rule, if $\psi(Z) = 1$, then Z is classified into class 1; otherwise, it is assigned to class 2. Under the Gaussian model with a common covariance matrix and equal class priors, Fisher's linear discriminant rule is the Bayes optimal rule and attains the minimum misclassification rate.

In practice, the parameters μ_1 , μ_2 , and Σ are unknown and need to be estimated from observed data. Let $\{x_k; 1 \leq k \leq n_1\}$ and $\{y_k; 1 \leq k \leq n_2\}$ be independent and identically distributed (i.i.d.) random samples drawn from $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, respectively. In the low-dimensional setting, the classical estimators for μ_1 and μ_2 are the sample means $\bar{x}$ and $\bar{y}$, respectively. The pooled sample covariance matrix $\hat{\Sigma}_n$ is used to estimate Σ , where the total sample size is $n = n_1 + n_2$. Plugging these estimates into (1.1) yields $\hat{\psi}(Z)$, the empirical version of $\psi(Z)$. Theoretical properties of $\hat{\psi}(Z)$ are well-studied in the fixed- p setting, see, e.g., [32]. In the high-dimensional setting, consistent estimation usually requires some structural conditions on Ω (or Σ) and δ . Two types of sparsity assumptions are commonly used in the literature. The first assumes that both Ω (or Σ) and δ are sparse; see [22–24, 33–35]. A weaker assumption was proposed by Cai and Liu [25], where only the product $\Omega\delta$ is assumed to be sparse. Under this assumption, they developed the Linear Programming Discriminant (LPD) method. Instead of estimating Ω and δ separately, the LPD method directly estimates the sparse discriminant

direction $\mathbf{\Omega}\delta$, denoted by β , through a constrained ℓ_1 minimization problem:

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 \text{ subject to } \left| \widehat{\Sigma}_n \beta - (\bar{x} - \bar{y}) \right|_{\infty} \leq \lambda_n \right\}, \quad (1.2)$$

where $\widehat{\Sigma}_n$, $\bar{x}$, and $\bar{y}$ denote sample estimates of Σ , μ_1 , and μ_2 , respectively, while λ_n is a tuning parameter. The estimator $\hat{\beta}$ can be computed by solving a linear programming problem. With the estimator, we classify Z into class 1 if and only if

$$(Z - \hat{\mu})^\top \hat{\beta} \geq 0, \quad (1.3)$$

where $\hat{\mu} = (\bar{x} + \bar{y})/2$.

In many modern applications, estimating LDA parameters from a single study can be difficult. The number of covariates p is often large, while the sample size n is relatively small. This may lead to large estimation uncertainty and poor classification performance. At the same time, data from several related studies may be available. These studies may share similar distributional structures with the target study. This provides an opportunity to borrow information across studies. We therefore develop a transfer learning approach for high-dimensional LDA.

1.1. Model set-up

Suppose we observe i.i.d. samples $\{x_i \in \mathbb{R}^p, i = 1, 2, \dots, n_1\}$ and $\{y_i \in \mathbb{R}^p, i = 1, 2, \dots, n_2\}$ generated from the target models $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, respectively. Following the linear programming formulation in (1.2), we focus on estimating $\beta = \mathbf{\Omega}\delta$ in the high-dimensional setting, where the dimension p may exceed the sample size $n = n_1 + n_2$, and $\mathbf{\Omega}\delta$ is assumed to be sparse. Here, sparsity includes both vector and matrix sparsity, as discussed in Cai and Liu [25]. Specifically, δ is called k_1 -sparse if it has at most k_1 nonzero entries, while $\mathbf{\Omega}$ is sparse if each row or column has at most k_2 nonzero entries. In this case, β is at most $k_1 k_2$ -sparse, and we directly impose the sparsity condition on β .

In transfer learning, data are available not only from the target study but also from auxiliary studies. For $k = 1, \dots, K$, let $x_i^{(k)} \in \mathbb{R}^p, i = 1, \dots, n_1^{(k)}$, and $y_i^{(k)} \in \mathbb{R}^p, i = 1, \dots, n_2^{(k)}$, be independent observations from $N(\mu_1^{(k)}, \Sigma^{(k)})$ and $N(\mu_2^{(k)}, \Sigma^{(k)})$, respectively. Set

$$\delta^{(k)} = \mu_1^{(k)} - \mu_2^{(k)}, \quad \mathbf{\Omega}^{(k)} = (\Sigma^{(k)})^{-1}, \quad k = 1, \dots, K.$$

The goal of transfer learning for LDA is to use auxiliary data to improve the estimation of the discriminant direction β and reduce misclassification. To do this, one needs to characterize the similarity between the target study and the auxiliary studies.

In transfer learning, measuring domain distance is essential. Common methods include Distribution Distance Measures, Feature-based Distance Measures, and Model Parameter Distance, among others [36, 37]. For linear discriminant analysis, if the data come from a normal population, the mean and variance characterize the distribution. We therefore consider the differences in means and variances separately. To quantify the difference between the source domain mean $\mu_s^{(k)}$ and the target domain mean μ_s , we use the Mahalanobis distance with a common precision matrix $\mathbf{\Omega} = \Sigma^{-1}$:

$$\mathcal{D}_M^s(\mu_s^{(k)}, \mu_s) = (\mu_s - \mu_s^{(k)})^\top \mathbf{\Omega} (\mu_s - \mu_s^{(k)}), \quad s = 1, 2. \quad (1.4)$$

A small value of $\mathcal{D}_M^s(\mu_s^{(k)}, \mu_s)$ indicates that the class- s mean in the auxiliary study is close to that in the target study after being standardized by the target covariance structure.

Mean similarity alone is not sufficient when covariance structures differ across studies. We also consider the discrepancy between the target and auxiliary covariance structures. Following Li et al. [15], we define $\mathcal{D}_q(\Omega^{(k)}, \Omega)$ using the row-wise and column-wise ℓ_q norms of $\Delta^{(k)} = \Sigma^{(k)}\Omega - I_p$. Since $\Delta^{(k)}$ is generally non-symmetric, both row and column deviations are considered.

$$\begin{aligned} \mathcal{A}_q(h_M, h_q) \\ = \{k \in \mathcal{K} : \mathcal{D}_M^1(\mu_1^{(k)}, \mu_1) \vee \mathcal{D}_M^2(\mu_2^{(k)}, \mu_2) \leq h_M \text{ and } \mathcal{D}_q(\Omega^{(k)}, \Omega) \leq h_q\}, \end{aligned}$$

where $\mathcal{K} = 1, \dots, K$.

For convenience, define

$$\eta_0 := \sqrt{h_M} + h_q, \quad h := \eta_0^2 = (\sqrt{h_M} + h_q)^2. \quad (1.5)$$

The set $\mathcal{A}_q(h_M, h_q)$ is defined at the population level and is not assumed to be known in practice. In Section 2, the auxiliary sources are selected using a data-driven procedure.

1.2. Article structure and notations

Using the similarity measure above, we develop a transfer learning method, called Trans-LPD, to estimate the target discriminant direction in LDA. This method is motivated by the LPD approach of Cai and Liu [25] for the single-study setting. The rest of the paper is organized as follows. Section 2 presents the Trans-LPD method and its theoretical properties, including adaptive source selection and convergence rates. Section 3 presents the simulation studies, and Section 4 gives a real-data application. Section 5 concludes the paper. Proofs and additional materials are provided in the Appendix.

We use the following notation. For a vector $x = (x_1, \dots, x_p)^T$, we denote ℓ_1 -norm as $\|x\|_1 = \sum_{i=1}^p |x_i|$ and ℓ_∞ -norm as $\|x\|_\infty = \max_{i \leq p} |x_i|$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p \times p}$, we define $\|A\|_{\max} = \max_{i,j} |a_{ij}|$, $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_{1,1} = \sum_{i,j} |a_{ij}|$. When needed, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ denotes the maximum absolute row-sum norm, and $\|A\|_2$ denotes the spectral norm. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ for $n \geq 1$ if there exists a constant C such that $|a_n| \leq C|b_n|$. We write $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$, and we write $a_n \asymp b_n$ if there are positive constants c and C such that $c \leq a_n/b_n \leq C$ for all $n \geq 1$.

2. Transfer learning for linear discriminant analysis

We first derive the population moment equations and then construct the Trans-LPD estimator using a data-driven source-selection procedure.

2.1. Rationale and moment equations

The LPD method can be formulated through the moment equation for the discriminant direction. Since $\beta = \Omega\delta$, we have:

$$\Sigma\beta - \delta = 0. \quad (2.1)$$

The LPD method estimates the discriminant direction by solving an empirical version of (2.1) with sparsity constraints. For the k th auxiliary study, we define $\Delta^{(k)} = \Sigma^{(k)}\Omega - I_p$. Then

$$\Sigma^{(k)}\beta - (I_p + \Delta^{(k)})\delta = 0.$$

Subtracting the auxiliary mean difference $\delta^{(k)}$ yields

$$\begin{aligned}\Sigma^{(k)}\beta - \delta^{(k)} &= (\Sigma^{(k)} - \Sigma)\beta + (\delta - \delta^{(k)}) \\ &= \Delta^{(k)}\delta + (\delta - \delta^{(k)}) =: r^{(k)}.\end{aligned}\quad (2.2)$$

Thus, $r^{(k)}$ measures the source-domain discrepancy at the target discriminant direction.

For any source subset $S \subseteq \{1, \dots, K\}$, let $n_S = \sum_{k \in S} n^{(k)}$ and $N = n_T + n_S$, where n_T denotes the target training sample size. Define the population pooled moments

$$\Sigma_P^S = \frac{n_T}{N}\Sigma + \sum_{k \in S} \frac{n^{(k)}}{N}\Sigma^{(k)}, \quad \delta_P^S = \frac{n_T}{N}\delta + \sum_{k \in S} \frac{n^{(k)}}{N}\delta^{(k)}.\quad (2.3)$$

Combining (2.1) and (2.2) gives the exact pooled relation

$$\Sigma_P^S\beta - \delta_P^S = r_S, \quad r_S := \sum_{k \in S} \frac{n^{(k)}}{N}r^{(k)}.\quad (2.4)$$

In practice, the auxiliary studies may not always be sufficiently similar to the target study, and incorporating non-informative sources may lead to negative transfer [38]. We therefore use a data-driven procedure to rank the auxiliary studies and select the final estimator using a target validation sample.

2.2. Trans-LPD algorithm

2.2.1. Data processing and notations

For the target data, we split each class into two disjoint subsets. Let I_x be a subset of $\{1, \dots, n_1\}$ with size $|I_x| = \lfloor n_1/2 \rfloor$, and let I_y be a subset of $\{1, \dots, n_2\}$ with size $|I_y| = \lfloor n_2/2 \rfloor$. Denote the complements of I_x and I_y by V_x and V_y , respectively. Define

$$\bar{x}_T = \frac{1}{|I_x|} \sum_{i \in I_x} x_i, \quad \bar{y}_T = \frac{1}{|I_y|} \sum_{i \in I_y} y_i, \quad \widehat{\delta}_T = \bar{x}_T - \bar{y}_T.$$

The corresponding sample covariance matrices are

$$\widehat{\Sigma}_{x,T} = \frac{1}{|I_x|} \sum_{i \in I_x} (x_i - \bar{x}_T)(x_i - \bar{x}_T)^\top, \quad \widehat{\Sigma}_{y,T} = \frac{1}{|I_y|} \sum_{i \in I_y} (y_i - \bar{y}_T)(y_i - \bar{y}_T)^\top,$$

and

$$\widehat{\Sigma}_T = \frac{1}{n_T} \left(|I_x| \widehat{\Sigma}_{x,T} + |I_y| \widehat{\Sigma}_{y,T} \right), \quad n_T = |I_x| + |I_y|.$$

For the auxiliary data, $\{x_i^{(k)} \in \mathbb{R}^p, i = 1, 2, \dots, n_1^{(k)}\}$ and $\{y_i^{(k)} \in \mathbb{R}^p, i = 1, 2, \dots, n_2^{(k)}\}$ are independently drawn from the k th auxiliary study with distributions $N(\mu_1^{(k)}, \Sigma^{(k)})$ (class 1^(k)) and $N(\mu_2^{(k)}, \Sigma^{(k)})$ (class

$2^{(k)}$), respectively, for $k = 1, \dots, K$. Define $\bar{x}^{(k)} = \frac{1}{n_1^{(k)}} \sum_{i=1}^{n_1^{(k)}} x_i^{(k)}$, $\bar{y}^{(k)} = \frac{1}{n_2^{(k)}} \sum_{i=1}^{n_2^{(k)}} y_i^{(k)}$, and $\widehat{\delta}^{(k)} = \bar{x}^{(k)} - \bar{y}^{(k)}$. The sample covariance matrices are

$$\widehat{\Sigma}_{x^{(k)}} = \frac{1}{n_1^{(k)}} \sum_{i=1}^{n_1^{(k)}} (x_i^{(k)} - \bar{x}^{(k)})(x_i^{(k)} - \bar{x}^{(k)})^\top$$

and

$$\widehat{\Sigma}_{y^{(k)}} = \frac{1}{n_2^{(k)}} \sum_{i=1}^{n_2^{(k)}} (y_i^{(k)} - \bar{y}^{(k)})(y_i^{(k)} - \bar{y}^{(k)})^\top.$$

Set

$$\widehat{\Sigma}^{(k)} = \frac{1}{n^{(k)}} \left(n_1^{(k)} \widehat{\Sigma}_{x^{(k)}} + n_2^{(k)} \widehat{\Sigma}_{y^{(k)}} \right),$$

where $n^{(k)} = n_1^{(k)} + n_2^{(k)}$. For a source subset $S \subseteq \{1, \dots, K\}$, let $n_S = \sum_{k \in S} n^{(k)}$, $N = n_T + n_S$. The corresponding empirical pooled moments are defined as

$$\widehat{\Sigma}_P^S = \frac{n_T \widehat{\Sigma}_T + \sum_{k \in S} n^{(k)} \widehat{\Sigma}^{(k)}}{N}, \quad \widehat{\delta}_P^S = \frac{n_T \widehat{\delta}_T + \sum_{k \in S} n^{(k)} \widehat{\delta}^{(k)}}{N}. \quad (2.5)$$

2.2.2. The steps of the Trans-LPD algorithm

We estimate the target discriminant direction β using both the target data and auxiliary data. The Trans-LPD procedure has four main steps: Computing a target-only LPD estimator, ranking the auxiliary studies, constructing Trans-LPD candidates, and selecting the final estimator using the independent target validation sample.

Step 1: Estimation based on the target data.

Using the target training data, we first compute the target-only LPD estimator:

$$\widehat{\beta}^{(L)} = \arg \min_{b \in \mathbb{R}^p} \|b\|_1 \quad \text{subject to} \quad \|\widehat{\Sigma}_T b - \widehat{\delta}_T\|_\infty \leq \lambda_L. \quad (2.6)$$

Step 2: Data-driven source ranking.

Following the data-driven source-ranking idea of Li et al. [15], we define a discrepancy score based on the LDA moment equation. For the k th auxiliary study, let

$$\widehat{\tau}_k = \left\| \widehat{\Sigma}^{(k)} \widehat{\beta}^{(L)} - \widehat{\delta}^{(k)} \right\|_\infty, \quad k = 1, \dots, K. \quad (2.7)$$

Let π be a permutation of $\{1, \dots, K\}$ such that $\widehat{\tau}_{\pi(1)} \leq \widehat{\tau}_{\pi(2)} \leq \dots \leq \widehat{\tau}_{\pi(K)}$. The candidate source sets are then defined by

$$S_0 = \emptyset, \quad S_m = \{\pi(1), \dots, \pi(m)\}, \quad m = 1, \dots, K. \quad (2.8)$$

Thus, the source-selection procedure does not require the population similarity set $\mathcal{A}_q(h_M, h_q)$ or the unknown thresholds h_M and h_q as inputs.

Step 3: Estimation using the target and auxiliary data.

For each candidate source set S_m , let $n_m = \sum_{k \in S_m} n^{(k)}$, $N_m = n_T + n_m$. Using $\widehat{\Sigma}_P^{S_m}$ and $\widehat{\delta}_P^{S_m}$ defined above, the corresponding Trans-LPD candidate is

$$\widehat{\beta}^{S_m}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \|b\|_1 \quad \text{subject to} \quad \left\| \widehat{\Sigma}_P^{S_m} b - \widehat{\delta}_P^{S_m} \right\|_\infty \leq \lambda. \quad (2.9)$$

For each S_m , define

$$\widehat{\eta}_m = \frac{1}{N_m} \sum_{k \in S_m} n^{(k)} \widehat{\tau}_k, \quad \widehat{\eta}_0 = 0.$$

Let $C_\lambda = \{2^{-2}, 2^{-1}, 1, 2, 4, 8, 16\}$. For $m = 1, \dots, K$, define

$$\Lambda_m = \left\{ a \left(\sqrt{\frac{\log(pK)}{N_m}} + \widehat{\eta}_m \right) : a \in C_\lambda \right\}. \quad (2.10)$$

Step 4: Validation-based selection.

Under equal class priors, for each (m, λ) with $\lambda \in \Lambda_m$, define

$$\widehat{\psi}_{m,\lambda}(z) = \mathbf{1} \left\{ (z - \widehat{\mu}_T)^\top \widehat{\beta}^{S_m}(\lambda) \geq 0 \right\}.$$

Its balanced misclassification error on the independent target validation sample is

$$\widehat{R}_V(m, \lambda) = \frac{1}{2} \left[\sum_{i \in V_x} \mathbf{1} \{ \widehat{\psi}_{m,\lambda}(x_i) = 0 \} + \sum_{i \in V_y} \mathbf{1} \{ \widehat{\psi}_{m,\lambda}(y_i) = 1 \} \right].$$

Let $\mathcal{M} = \{(m, \lambda) : m = 0, \dots, K, \lambda \in \Lambda_m\}$. The final candidate is selected by

$$(\widehat{m}, \widehat{\lambda}) \in \arg \min_{(m,\lambda) \in \mathcal{M}} \widehat{R}_V(m, \lambda), \quad \widehat{S} = S_{\widehat{m}}, \quad \widehat{\beta} = \widehat{\beta}^{S_{\widehat{m}}}(\widehat{\lambda}). \quad (2.11)$$

Since $S_0 = \emptyset$ and $\Lambda_0 = \{\lambda_L\}$, the candidate family contains the target-only LPD estimator. This provides a safeguard against negative transfer.

Remark 2. The optimization in (2.9) directly regularizes the target discriminant direction β , whose sparsity is assumed in the theoretical analysis. The source discrepancy scores are used only to construct candidate source sets, while the independent target validation sample determines the final choice among the target-only and transfer candidates.

Both the target-only LPD estimator and the Trans-LPD candidates are obtained by solving a p -dimensional linear programming problem. The source path contains only $K + 1$ nested candidate sets, and each tuning-parameter grid is finite. Once $\widehat{\beta}$ is obtained, a new observation Z is classified using the rule defined above.

2.3. Asymptotic properties of Trans-LPD

In this subsection, we study the theoretical properties of Trans-LPD. We first consider the estimation and classification performance for a fixed set of auxiliary studies and then study the final adaptive procedure.

When μ_1, μ_2 , and Σ are known, Fisher's rule (see, e.g., [32]) is optimal. Its misclassification rate is

$$R = 1 - \Phi \left(\frac{\sqrt{\Delta_R}}{2} \right), \quad \Delta_R = \delta^\top \Omega \delta, \quad (2.12)$$

where $\delta = \mu_1 - \mu_2$ and $\Omega = \Sigma^{-1}$ is the target-domain precision matrix.

Let $\widehat{\beta}$ be the final Trans-LPD estimator defined in (2.11), and let $\widehat{\mu}_T = \frac{\bar{x}_T + \bar{y}_T}{2}$. Following Cai and Liu [25], the conditional misclassification rate of the resulting classifier is

$$R_n = 1 - \frac{1}{2} \Phi \left(-\frac{(\widehat{\mu}_T - \mu_1)^\top \widehat{\beta}}{(\widehat{\beta}^\top \Sigma \widehat{\beta})^{1/2}} \right) - \frac{1}{2} \Phi \left(\frac{(\widehat{\mu}_T - \mu_2)^\top \widehat{\beta}}{(\widehat{\beta}^\top \Sigma \widehat{\beta})^{1/2}} \right). \quad (2.13)$$

We compare R_n with the oracle risk R and the misclassification rate of the target-only LPD classifier. We first state the conditions used in the theoretical analysis.

(C1). The Gaussian model stated in Section 1 holds. The target training samples satisfy $|I_x| \asymp |I_y|$, and, for every auxiliary study, $n_1^{(k)} \asymp n_2^{(k)}$, $k = 1, \dots, K$. There exist constants $0 < C_1 < C_2 < \infty$ such that $C_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_2$ and $C_1 \leq \lambda_{\min}(\Sigma^{(k)}) \leq \lambda_{\max}(\Sigma^{(k)}) \leq C_2$, $k = 1, \dots, K$. Moreover, $0 < c_R \leq \Delta_R \leq C_R < \infty$.

(C2). The target discriminant direction is sparse and satisfies $\|\beta\|_0 \leq s$. We assume $n_T \geq Cs \log p$, $\min_{1 \leq k \leq K} n^{(k)} \geq C \log(pK)$. For a fixed source set S , let $n_S = \sum_{k \in S} n^{(k)}$, $N = n_T + n_S$, and assume $N \geq Cs \log(pK)$. Under Conditions C1–C2, $\|\beta\|_2 \leq C$, $\|\beta\|_1 \leq C\sqrt{s}$. For a fixed source set S , define $\eta_S = \|r_S\|_\infty$, where r_S is given in (2.4). If $S \subseteq \mathcal{A}_q(h_M, h_q)$, the mean–variance similarity conditions imply

$$\eta_S \leq C \frac{n_S}{N} (\sqrt{h_M} + h_q) = C \frac{n_S}{N} \eta_0. \quad (2.14)$$

Thus, the discrepancy between the pooled estimating equation and the target moment equation is controlled by the mean–variance similarity condition. The detailed derivation of (2.14) is given in the Appendix.

Lemma 2.1. Suppose Conditions C1–C2 hold. For a fixed source set S , if

$$\lambda_S \geq C \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\},$$

then, with probability tending to one, $\|\widehat{\Sigma}_p^S \beta - \widehat{\delta}_p^S\|_\infty \leq \lambda_S$. Moreover,

$$\|\widehat{\beta}^S - \beta\|_2 \leq C \sqrt{s} \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\}.$$

Lemma 2.1 shows that the estimation error contains two parts: the sampling error and the bias introduced by differences between the target and auxiliary studies.

Theorem 2.1 (Classification performance under transfer learning). Suppose Conditions C1–C2 hold. For a fixed source set S , choose $\lambda_S \asymp \sqrt{\frac{\log(pK)}{N}} + \eta_S$. Let R_n^S denote the conditional misclassification rate of the corresponding Trans-LPD classifier. Then

$$\left| \frac{R_n^S}{R} - 1 \right| = O_p \left(\sqrt{\frac{s \log(pK)}{N}} + \sqrt{s} \eta_S + \sqrt{\frac{s \log p}{n_T}} \right). \quad (2.15)$$

In particular, if

$$\sqrt{\frac{s \log(pK)}{N}} + \sqrt{s} \eta_S + \sqrt{\frac{s \log p}{n_T}} = o(1),$$

then

$$\frac{R_n^S}{R} \longrightarrow 1$$

in probability.

Theorem 2.1 shows how the source sample size and transfer bias affect classification. The first term in (2.15) decreases as the total sample size increases, while the second term is due to the discrepancy between the target and auxiliary studies. Thus, auxiliary data are useful when the reduction in sampling error is large enough relative to the transfer bias.

Theorem 2.2 (Adaptive Trans-LPD and negative transfer). *Let R_n be the conditional misclassification rate of the final Trans-LPD classifier defined in (2.13), and let $R_n^{(L)}$ denote the conditional misclassification rate of the target-only LPD classifier. Let $n_{V,\min} = \min\{|V_x|, |V_y|\}$. For each $(m, \lambda) \in \mathcal{M}$, let $R_n(m, \lambda)$ denote the conditional misclassification rate of the corresponding candidate classifier. Conditional on the target training samples and all auxiliary samples, for any $0 < \alpha < 1$, with probability at least $1 - \alpha$,*

$$R_n \leq \min_{(m, \lambda) \in \mathcal{M}} R_n(m, \lambda) + 2 \sqrt{\frac{\log(4|\mathcal{M}|/\alpha)}{2n_{V,\min}}}. \quad (2.16)$$

Since $S_0 = \emptyset$, $\widehat{\beta}^{S_0} = \widehat{\beta}^{(L)}$, and $\Lambda_0 = \{\lambda_L\}$, the target-only LPD classifier belongs to the candidate family. Hence,

$$R_n \leq R_n^{(L)} + 2 \sqrt{\frac{\log(4|\mathcal{M}|/\alpha)}{2n_{V,\min}}}. \quad (2.17)$$

Consequently, if $\frac{\log |\mathcal{M}|}{n_{V,\min}} \longrightarrow 0$, then

$$R_n \leq R_n^{(L)} + o_p(1). \quad (2.18)$$

Theorem 2.2 shows that the final Trans-LPD classifier is asymptotically no worse than the target-only LPD classifier. If the candidate family contains an informative source set S^* , Theorems 2.1 and 2.2 further give

$$R_n \leq \min \{R_n^{(L)}, R_n^{S^*}\} + O_p \left(\sqrt{\frac{\log |\mathcal{M}|}{n_{V,\min}}} \right).$$

Thus, the adaptive procedure can use informative auxiliary studies while keeping the target-only LPD classifier in the candidate family as a safeguard against negative transfer.

3. Numerical studies

We evaluate the numerical performance of the Trans-LPD using simulated data. The LPD optimization problems are solved by the Alternating Direction Method of Multipliers (ADMM) algorithm, which is computationally efficient and less sensitive to initial values than the interior point method. Under the Trans-LPD procedure, the target-only estimator and the pooled transfer candidates

are obtained from LPD-type optimization problems. The auxiliary studies are ranked using the data-driven procedure in Section 2.2, and the final candidate is selected using the independent target validation sample.

3.1. Implementation of LPD with the ADMM algorithm

The ADMM, introduced by Glowinski and Marrocco [39] (see also [40]), can be used to solve the LPD optimization problem. Zhang et al. [41] and Wang and Yuan [42] showed that the ADMM provides an efficient way to solve the Dantzig selector.

To handle both the target-only LPD estimator and the pooled Trans-LPD candidates, we consider the following generic LPD problem:

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{subject to} \quad \|\widehat{\Sigma}\beta - \widehat{\delta}\|_{\infty} \leq \lambda. \quad (3.1)$$

For the target-only estimator, $(\widehat{\Sigma}, \widehat{\delta}, \lambda) = (\widehat{\Sigma}_T, \widehat{\delta}_T, \lambda_L)$. For a Trans-LPD candidate associated with S_m , $(\widehat{\Sigma}, \widehat{\delta}, \lambda) = (\widehat{\Sigma}_P^{S_m}, \widehat{\delta}_P^{S_m}, \lambda)$.

To apply the ADMM, problem (3.1) can be reformulated as

$$\min_{\beta \in \mathbb{R}^p, x \in \mathcal{X}} \left\{ \|\beta\|_1 : \widehat{\Sigma}\beta - \widehat{\delta} = x \right\}, \quad (3.2)$$

where

$$\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_{\infty} \leq \lambda\},$$

and $x \in \mathbb{R}^p$ is an auxiliary variable.

The augmented Lagrangian for (3.2) is defined as

$$L_A(\beta, x, z) = \|\beta\|_1 - \langle z, \widehat{\Sigma}\beta - \widehat{\delta} - x \rangle + \frac{\gamma}{2} \left\| \widehat{\Sigma}\beta - \widehat{\delta} - x \right\|_2^2, \quad (3.3)$$

where $z \in \mathbb{R}^p$ is the Lagrange multiplier, $\gamma > 0$ is a penalty parameter, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^p$.

The ADMM iterations are

$$\begin{cases} \beta^{k+1} = \arg \min_{\beta \in \mathbb{R}^p} L_A(\beta, x^k, z^k), \\ x^{k+1} = \arg \min_{x \in \mathcal{X}} L_A(\beta^{k+1}, x, z^k), \\ z^{k+1} = z^k - \gamma (\widehat{\Sigma}\beta^{k+1} - \widehat{\delta} - x^{k+1}). \end{cases} \quad (3.4)$$

It x subproblem in (3.4) has the closed-form solution

$$\begin{aligned} x^{k+1} &= \arg \min_{\|x\|_{\infty} \leq \lambda} \left\| x - \left(\widehat{\Sigma}\beta^{k+1} - \widehat{\delta} + \frac{z^k}{\gamma} \right) \right\|_2^2 \\ &= \Pi_{\mathcal{X}} \left(\widehat{\Sigma}\beta^{k+1} - \widehat{\delta} + \frac{z^k}{\gamma} \right), \end{aligned} \quad (3.5)$$

where $\Pi_{\mathcal{X}}$ denotes the componentwise projection onto $[-\lambda, \lambda]$.

Up to an additive constant, the β -subproblem can be written as

$$\arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 + \frac{\gamma}{2} \left\| \widehat{\Sigma}\beta - \widehat{\delta} - x^k + \frac{z^k}{\gamma} \right\|_2^2 \right\}. \quad (3.6)$$

The β -subproblem in (3.4) has no closed-form solution unless $\widehat{\Sigma}$ is diagonal. We therefore linearize the quadratic term and obtain the approximate subproblem

$$\arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 + \frac{\gamma}{2} \left(2(v^k)^\top (\beta - \beta^k) + \frac{\mu}{2} \|\beta - \beta^k\|_2^2 \right) \right\}, \quad (3.7)$$

where

$$v^k = \widehat{\Sigma}^\top \left(\widehat{\Sigma}\beta^k - \widehat{\delta} - x^k + \frac{z^k}{\gamma} \right)$$

is the gradient of the quadratic term at $\beta = \beta^k$, and $\mu > 0$ is a proximal parameter.

Rearranging (3.7) gives

$$\arg \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 + \frac{\gamma\mu}{4} \left\| \beta - \left(\beta^k - \frac{2}{\mu} v^k \right) \right\|_2^2 \right\},$$

whose closed-form solution is

$$\beta^{k+1} = \text{shrinkage} \left(\beta^k - \frac{2}{\mu} v^k, \frac{2}{\gamma\mu} \right), \quad (3.8)$$

where the shrinkage operator is applied componentwise and is defined by

$$\text{shrinkage}(u, \tau) = \text{sign}(u)(|u| - \tau)_+.$$

We use (3.8) to update in the ADMM iterations.

Algorithm 1. Linearized ADMM solver for the LPD optimization

Input: $\widehat{\Sigma}$, $\widehat{\delta}$, λ , $\gamma > 0$, $\mu > 0$, tolerance tol , and the maximum number of iterations $M_{\max}$.

Initialization: Set $(\beta, x, z) = (\beta^0, x^0, z^0)$, $k = 0$.

Repeat Steps 1–4 until the stopping criterion is satisfied.

Step 1. Compute β^{k+1} according to (3.8).

Step 2. Compute x^{k+1} according to (3.5).

Step 3. Update

$$z^{k+1} = z^k - \gamma \left(\widehat{\Sigma}\beta^{k+1} - \widehat{\delta} - x^{k+1} \right).$$

Step 4. Update $k \leftarrow k + 1$.

The algorithm stops when

$$\left\| \widehat{\Sigma}\beta^k - \widehat{\delta} - x^k \right\|_2 \leq \text{tol}$$

and

$$\|\beta^k - \beta^{k-1}\|_2 \leq \text{tol},$$

or when $k = M_{\max}$.

Output: $\widehat{\beta} = \beta^k$.

We use the following settings for γ and μ in Algorithm 1. The ADMM iteration starts $((\beta^0, x^0, z^0))$ and continues until the stopping criterion is met. Following Wang et al. [42], we set the initial value of γ as 0.1 and multiply it by 1.5 after every five iterations, with an upper bound of 100. For μ , we set $\mu^0 = 3\rho(\widehat{\Sigma}^\top \widehat{\Sigma})$, where $\rho(\cdot)$ denotes the spectral radius. It is then updated by

$$\mu^{k+1} = \begin{cases} \mu^k, & \text{if } \text{mod}(k, 5) \neq 0, \\ \max\{0.75\mu^k, 0.15\rho(\widehat{\Sigma}^\top \widehat{\Sigma})\}, & \text{if } \text{mod}(k, 5) = 0, \quad k \leq 50, \\ \min\{1.15\mu^k, 15\rho(\widehat{\Sigma}^\top \widehat{\Sigma})\}, & \text{if } \text{mod}(k, 5) = 0, \quad k > 50. \end{cases}$$

The same ADMM solver is used for the target-only LPD estimator and every pooled Trans-LPD candidate. Thus, no separate numerical solver is needed for the transfer candidates.

3.2. Simulation settings

In this subsection, we describe the simulation settings and compare four methods. The first is Target-LPD, which uses only the target training sample. The second is Pooled-LPD, which combines the target and all auxiliary studies without source selection. The third is Oracle Trans-LPD, which uses the population informative source set and is included only as a benchmark. The fourth is Adaptive Trans-LPD, the data-driven procedure described in Section 2.2.

We set $n_1 = n_2 \in \{10, 20, 30, 50, 100\}$, $p \in \{200, 400, 600\}$, with $n_1^{(k)} = n_2^{(k)} = 200, k = 1, \dots, K$, where $K \in \{2, 3, 4, 5\}$. The target class means are set as $\mu_1 = 0$, $\mu_2 = (2, \dots, 2, 0, \dots, 0)^\top$, with $s_0 = 10$ nonzero entries. We consider the following three covariance structures:

- **Model 1.** $\Sigma = (\sigma_{ij})_{p \times p}^{-1}$, where $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.8^{|i-j|} \mathbf{1}\{|i-j| \leq 2\}$, $i \neq j$.
- **Model 2.** $\Sigma = \Omega^{-1}$, where Ω is a block diagonal matrix with block size 4 and each block is a Toeplitz matrix $\text{Toeplitz}(1.0, 0.5, 0.4, 0.2)$.
- **Model 3.** $\Sigma = (\sigma_{ij})_{p \times p}^{-1}$, where $\sigma_{ii} = 1$ and, for $i \neq j$, σ_{ij} is independently drawn from $U(0, 0.6)$. The matrix is symmetrized and, if necessary, adjusted to be positive definite.

Let $S^* \subseteq \{1, \dots, K\}$ denote the population informative source set in the simulation design. The set S^* is used only for data generation and for constructing Oracle Trans-LPD, and is not available to Adaptive Trans-LPD. For $k \in S^*$, each entry of $\Delta^{(k)}$ is set to zero with a probability of 0.9 and is independently drawn from $U(-r/p, r/p)$ with a probability of 0.1, where $r \in \{10, 20, 30\}$. The auxiliary covariance matrix is constructed from $\Delta^{(k)} = \Sigma^{(k)}\Omega - I_p$ and, when necessary, adjusted to be positive definite. After symmetrization, positive definiteness is ensured by replacing eigenvalues smaller than 0.05 by 0.05. For $k \notin S^*$, a larger covariance discrepancy is introduced by setting $\Sigma_{ij}^{(k)} = \mathbf{1}(i = j) + d_{ij}^{(k)}$, where $d_{ij}^{(k)}$ is zero with a probability of 0.4 and nonzero with a probability of 0.6. For the auxiliary class means, the mean perturbations are generated sparsely from $U(-r/p, r/p)$ for $k \in S^*$, whereas

larger Bernoulli-type perturbations are used for $k \notin S^*$. For each Monte Carlo replication, each target class is randomly divided into a training subset and a validation subset, with

$$|I_x| = \left\lfloor \frac{n_1}{2} \right\rfloor, \quad |I_y| = \left\lfloor \frac{n_2}{2} \right\rfloor, \quad V_x = I_x^c, \quad V_y = I_y^c.$$

The target training sample is used to estimate the target-only pilot $\widehat{\beta}^{(L)}$ and construct the transfer candidates, while the validation sample is used only to select the final candidate. For each auxiliary study, we compute

$$\widehat{\tau}_k = \left\| \widehat{\Sigma}^{(k)} \widehat{\beta}^{(L)} - \widehat{\delta}^{(k)} \right\|_{\infty}, \quad k = 1, \dots, K,$$

and rank the auxiliary studies so that $\widehat{\tau}_{\pi(1)} \leq \dots \leq \widehat{\tau}_{\pi(K)}$. The nested candidate source sets are $S_0 = \emptyset$, $S_m = \{\pi(1), \dots, \pi(m)\}$, $m = 1, \dots, K$. For each S_m , let

$$n_m = \sum_{k \in S_m} n^{(k)}, \quad N_m = n_T + n_m,$$

and

$$\widehat{\eta}_m = \frac{1}{N_m} \sum_{k \in S_m} n^{(k)} \widehat{\tau}_k, \quad \widehat{\eta}_0 = 0.$$

The Trans-LPD estimator is obtained by solving

$$\widehat{\beta}^{S_m}(\lambda) = \arg \min_{b \in \mathbb{R}^p} \|b\|_1 \quad \text{subject to} \quad \left\| \widehat{\Sigma}_P^{S_m} b - \widehat{\delta}_P^{S_m} \right\|_{\infty} \leq \lambda.$$

For each m , the tuning parameter is selected from

$$\Lambda_m = \left\{ a \left(\sqrt{\frac{\log(pK)}{N_m}} + \widehat{\eta}_m \right) : a \in C_{\lambda} \right\},$$

where the same finite multiplier grid C_{λ} is used in all Monte Carlo replications. For $m = 0$, we set $S_0 = \emptyset$ and $\Lambda_0 = \{\lambda_L\}$, so the corresponding candidate is the target-only LPD estimator. The final pair $(\widehat{m}, \widehat{\lambda})$ is selected by minimizing the balanced misclassification error on the independent target validation sample, and the final Adaptive Trans-LPD estimator is

$$\widehat{\beta} = \widehat{\beta}^{S_{\widehat{m}}}(\widehat{\lambda}).$$

The procedure is repeated 100 times for each simulation setting. The simulation results are presented below.

3.3. Simulation results

Figure 1 illustrates the average misclassification rates of the four methods for $K = 2$ under three dimensions and three covariance models. In this setting, the auxiliary data contain one informative and one non-informative study, and the target sample size varies from 10 to 100. The transfer methods have lower misclassification rates than Target-LPD in most settings. Adaptive Trans-LPD performs particularly well under Models 2 and 3. As the target sample size increases, the misclassification

rates of most methods decrease, with some finite-sample fluctuations. Pooled-LPD has similar performance to Adaptive Trans-LPD in several settings, but shows greater variation because it uses the non-informative auxiliary study without source selection. Oracle Trans-LPD uses the prespecified informative source set and serves as a benchmark. In some settings, Adaptive Trans-LPD has a misclassification rate comparable to or lower than that of Oracle Trans-LPD. This can occur because the validation procedure may select a smaller source subset or the target-only candidate when including additional auxiliary studies does not improve validation performance.

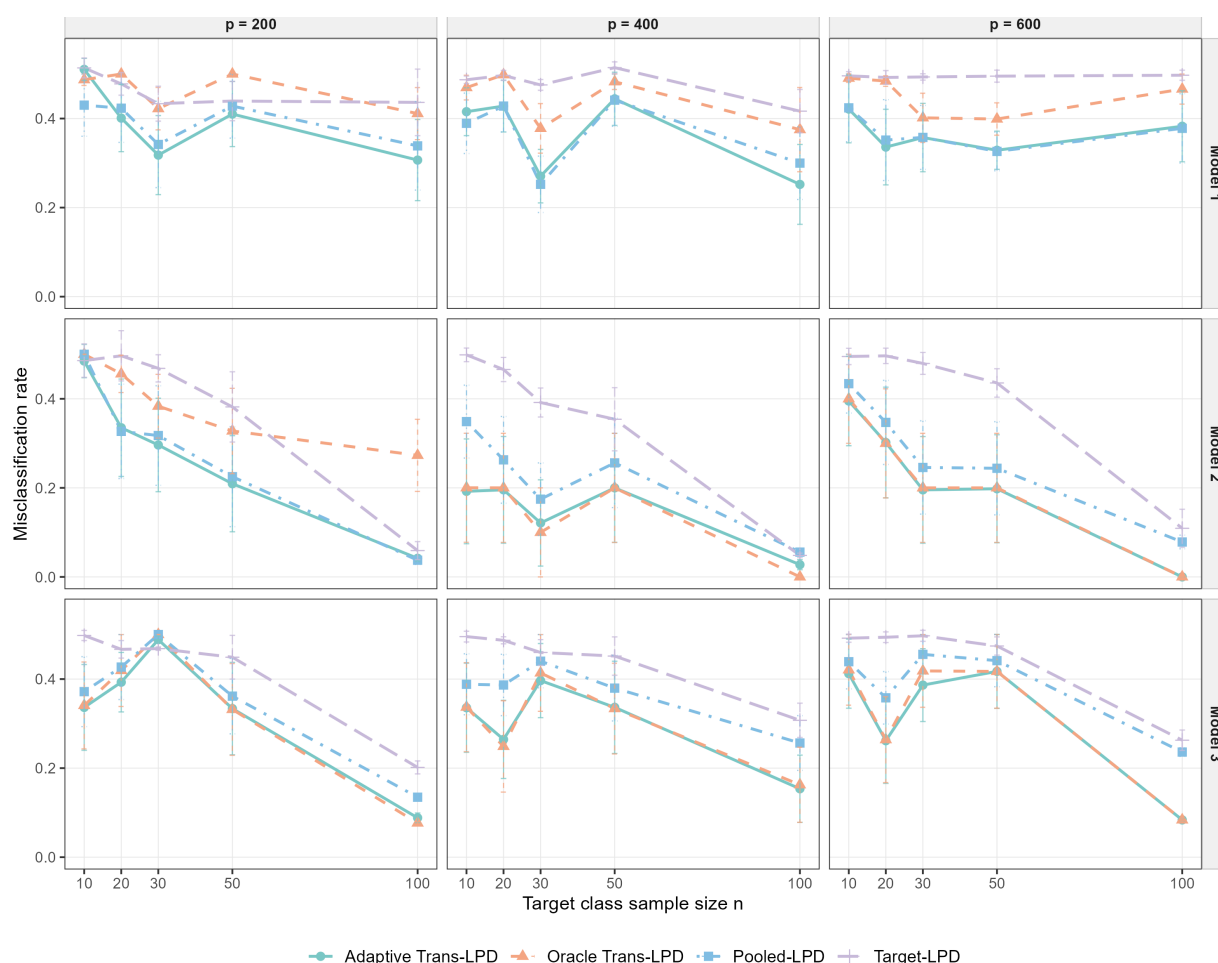


Figure 1. Average misclassification rates of Target-LPD, Pooled-LPD, Oracle Trans-LPD, and Adaptive Trans-LPD under $K = 2$ for different target sample sizes, dimensions, and covariance models.

Figure 2 presents the classification errors of the four methods when the target class sample size is fixed at $n = 10$ and the number of auxiliary studies increases from $K = 2$ to $K = 5$. In this experiment, only one non-informative auxiliary study is included, so the proportion of informative auxiliary studies increases with K . The performance of Target-LPD changes little with K because it does not use auxiliary data. The three transfer methods, however, vary as the number of auxiliary studies changes. Adaptive Trans-LPD has a lower misclassification rate than Target-LPD in a number of settings, especially under Models 2 and 3. Pooled-LPD performs similarly to Adaptive Trans-LPD

in some cases, but its error can increase when the non-informative study has a stronger effect. By ranking the auxiliary studies and using the independent target validation sample for final selection, Adaptive Trans-LPD can avoid using this source in some cases. The misclassification rates are not monotone in K for several combinations of p and the covariance model. This suggests that adding auxiliary studies does not always improve classification, since the result also depends on the difference between the auxiliary and target distributions.

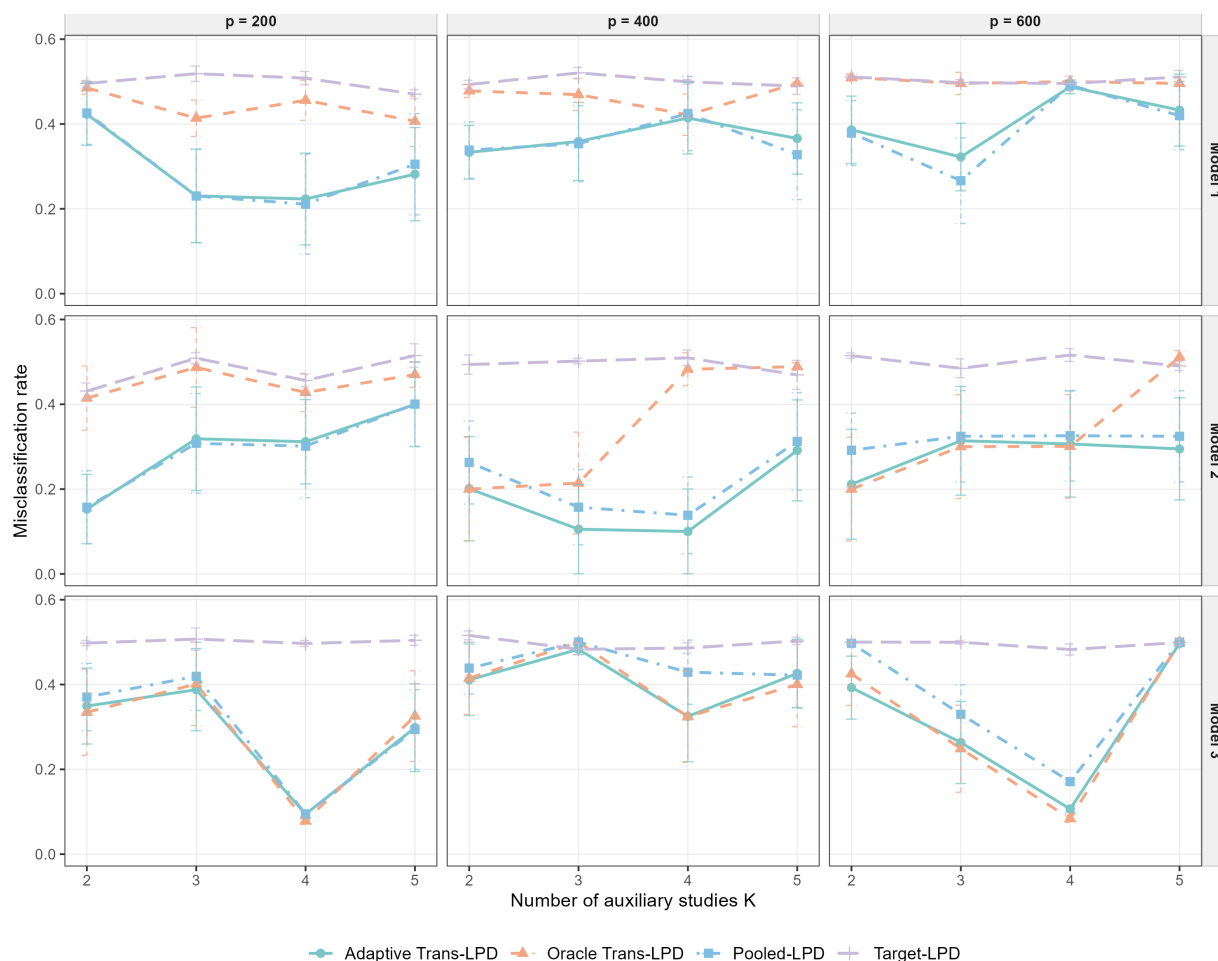


Figure 2. Average misclassification rates of Target-LPD, Pooled-LPD, Oracle Trans-LPD, and Adaptive Trans-LPD as the number of auxiliary studies K varies from 2 to 5, with the target class sample size fixed at $n = 10$.

Figure 3 reports the corresponding results when the target class sample size is fixed at $n = 50$. Compared with Figure 2, the larger target sample gives a more reliable target-domain estimator. Target-LPD therefore provides a stronger baseline, although auxiliary information can still improve classification in many settings. Adaptive Trans-LPD has a lower misclassification rate than Target-LPD in most settings and performs similarly to Pooled-LPD in many cases. Under Models 1 and 2, the adaptive procedure performs better in several settings, where source selection avoids pooling all auxiliary studies. For Model 3, the relative performance varies more with K and p . This indicates that the effect of auxiliary data depends on both the amount of source data and the differences between the

target and auxiliary covariance structures. These results are consistent with the variance–bias trade-off discussed in the theoretical analysis.

The simulation results show that Adaptive Trans-LPD can use auxiliary information while retaining the target-only candidate as a safeguard. When the auxiliary studies are informative, incorporating them can reduce the misclassification rate by increasing the effective sample size. When some auxiliary studies differ from the target domain, the data-driven source ranking and validation step can reduce the effect of directly pooling all auxiliary data. Thus, the benefit of auxiliary information depends on an appropriate balance between variance reduction and transfer bias.

We also examine the accuracy of the source ranking. Table 1 reports the results over 100 replications with $p = 400$. The exact ranking recovery (ERR) rate is the proportion of replications in which all informative sources are ranked ahead of the non-informative source. The pairwise ranking accuracy (PRA) is the average proportion of informative–non-informative source pairs that are correctly ordered. Sources are ranked by $\widehat{\tau}_k$ in ascending order.

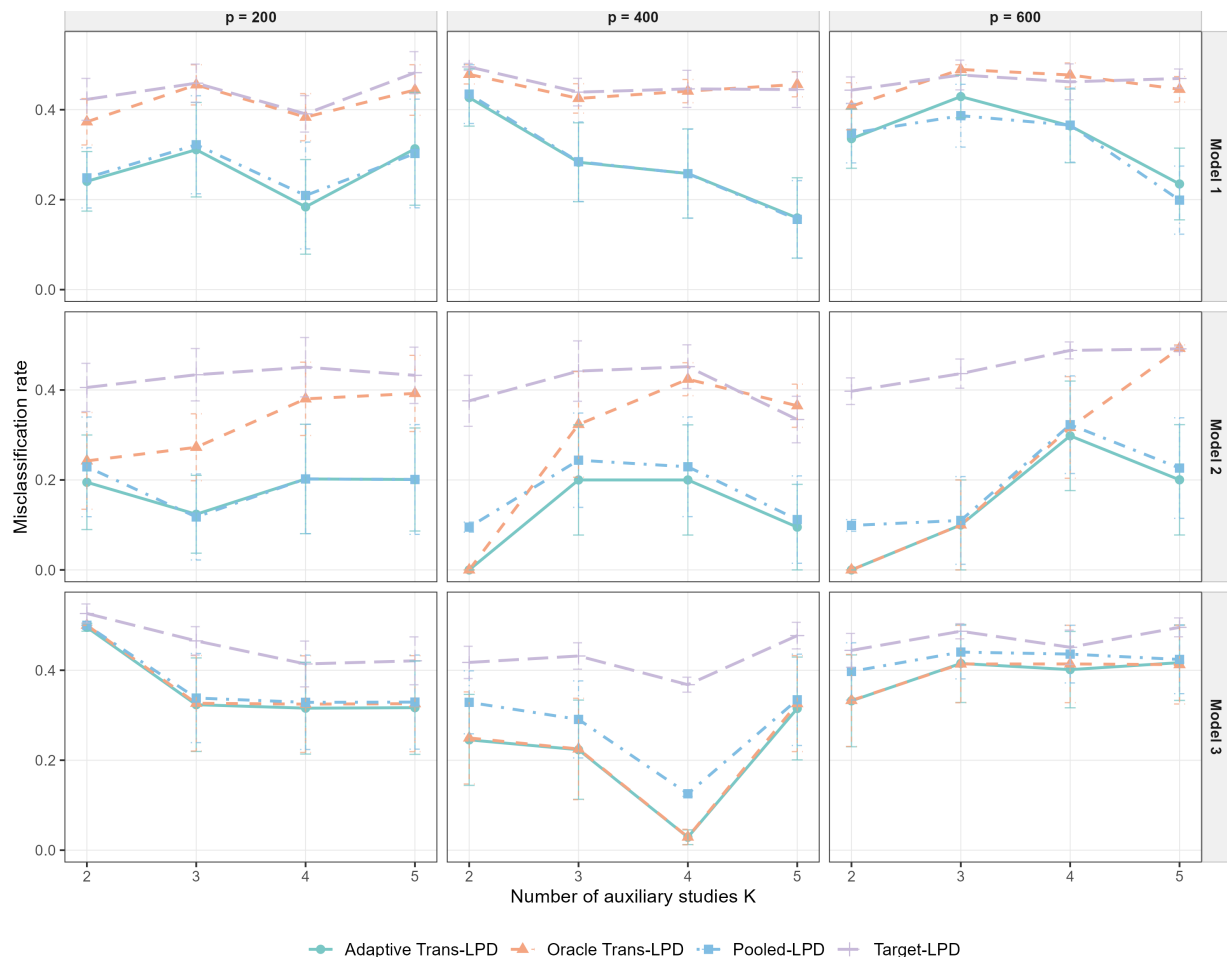


Figure 3. Average misclassification rates of Target-LPD, Pooled-LPD, Oracle Trans-LPD, and Adaptive Trans-LPD as the number of auxiliary studies K varies from 2 to 5, with the target class sample size fixed at $n = 50$.

Table 1. Source-ranking accuracy (%) over 100 replications with $p = 400$. Each entry gives the ERR / PRA.

n	K	Model 1	Model 2	Model 3
10	2	90.0 / 90.0	100.0 / 100.0	68.0 / 68.0
	3	83.0 / 86.0	100.0 / 100.0	72.0 / 76.0
	4	76.0 / 83.7	100.0 / 100.0	63.0 / 72.7
	5	77.0 / 87.0	100.0 / 100.0	67.0 / 77.5
50	2	89.0 / 89.0	100.0 / 100.0	85.0 / 85.0
	3	89.0 / 93.0	98.0 / 99.0	77.0 / 80.5
	4	92.0 / 95.3	98.0 / 99.3	70.0 / 77.3
	5	86.0 / 90.3	99.0 / 99.0	71.0 / 81.0

Both measures are close to 100% under Model 2. The accuracy is lower under Model 3. When n increases from 10 to 50, both measures improve in most settings under Model 1 and in all settings under Model 3.

4. Application

In this section, we apply Adaptive Trans-LPD to the cholangiocarcinoma (CCA) gene expression dataset from The Cancer Genome Atlas (TCGA), which is publicly available at the UCSC Xena data portall (<https://xenabrowser.net/datapages/>). The gene expression data are provided in the HiSeqV2 format (Illumina HiSeq RNA-seq, $\log_2(\text{norm_count} + 1)$ transformed), and the clinical information is provided in the `clinicalMatrix` format. We use these data to classify tumor versus normal tissue samples based on high-dimensional mRNA expression profiles. Cholangiocarcinoma is a rare malignancy arising from the biliary epithelium, and the TCGA-CHOL cohort contains only 45 samples in total ($n_T = 45$), with 36 tumor samples (case) and 9 adjacent normal samples (control). Each sample is originally characterized by 20,530 gene expression features. This is a typical high-dimensional, small-sample-size problem ($p \gg n$), under which classical LDA fails due to the singularity of the sample covariance matrix. We use the TCGA-LIHC (hepatocellular carcinoma) dataset as the source domain, which contains $n_S = 423$ samples (373 tumor and 50 normal). LIHC and CCA are related liver and biliary tract cancers and share several biological characteristics, which provides a basis for using an auxiliary source.

We treat TCGA-CHOL as the target domain and TCGA-LIHC as the source domain. We first retain the genes measured in both cohorts, remove genes with missing values or zero variance, and then perform variance-based feature screening to retain the top 2000 most variable genes.

For each replication, the target data are randomly split into a training set (70%) and an independent testing set (30%). The training set is further divided into fitting and validation subsets. The fitting subset is used to construct the target-only estimator and the transfer candidate, while the validation subset is used only for final candidate selection. The full source dataset is used to assist the estimation of the discriminant direction. The testing set is used only to evaluate the final classification performance.

We compare three methods: Target-LPD, which uses only the target data; Pooled-LPD, which combines the target and source data without source selection; and our proposed Adaptive Trans-LPD. Since only one auxiliary study is available in this application, the candidate source sets reduce to $S_0 = \emptyset$ and $S_1 = \{\text{LIHC}\}$. Therefore, Adaptive Trans-LPD selects between the target-only candidate and the candidate using TCGA-LIHC according to the independent target validation sample.

Finally, we evaluate the methods on the testing data using the misclassification error rate and the area under the ROC curve (AUC). A higher AUC and a smaller misclassification error indicate better classification performance. The procedure is repeated 100 times. Table 2 reports the mean AUC, mean misclassification error rate, and the standard deviation (SD) of the error rate. The corresponding ROC curves and boxplots are shown in Figures 4 and 5, respectively.

As shown in Table 2, and Figures 4 and 5, Target-LPD has an AUC of 0.498, a misclassification error of 0.502, and an SD of 0.247. This is consistent with the difficulty of a high-dimensional discriminant direction from the small TCGA-CHOL target sample alone. Pooled-LPD performs much better, with an AUC of = 1.000 and a misclassification error of = 0.000. These results suggest that, under the selected gene representation, the TCGA-LIHC data provide useful information for the TCGA-CHOL classification task.

Adaptive Trans-LPD also performs better than Target-LPD, with AUC = 0.850 and error rate = 0.287. Its performance, however, is lower than that of Pooled-LPD in this application. One reason is the validation step used for candidate selection. Since the TCGA-CHOL cohort is small, the validation subset contains only a limited number of observations, and the validation error can vary across random splits. In our experiment, Adaptive Trans-LPD selects the target-only candidate in about half of the replications and the candidate using TCGA-LIHC in the remaining replications. Its average performance therefore lies between Target-LPD and Pooled-LPD.

The better performance of Pooled-LPD in this application suggests that TCGA-LIHC provides an useful auxiliary source for TCGA-CHOL. This result should not be interpreted as evidence that pooling all auxiliary studies is generally preferable. Adaptive Trans-LPD does not assume in advance that auxiliary study is informative. When the usefulness of the source domain is uncertain, the validation-based procedure retains the target-only candidate as a safeguard against negative transfer. In this application, the results also show that adaptive source selection may lose some efficiency when the target sample available for validation is very small.

Table 2. Mean AUC, mean misclassification error rate, and standard deviation (SD) of the error rate over 100 replications on the TCGA-CHOL dataset.

Method	Target-LPD	Pooled-LPD	Adaptive Trans-LPD
AUC	0.498	1.000	0.850
Error rate	0.502	0.000	0.287
SD	0.247	0.000	0.152

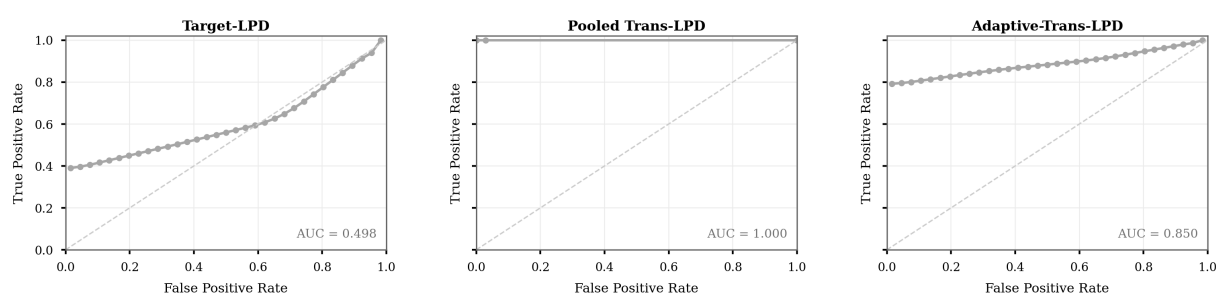


Figure 4. ROC curves of Target-LPD, Pooled-LPD, and Adaptive-Trans-LPD on the TCGA-CHOL testing data.

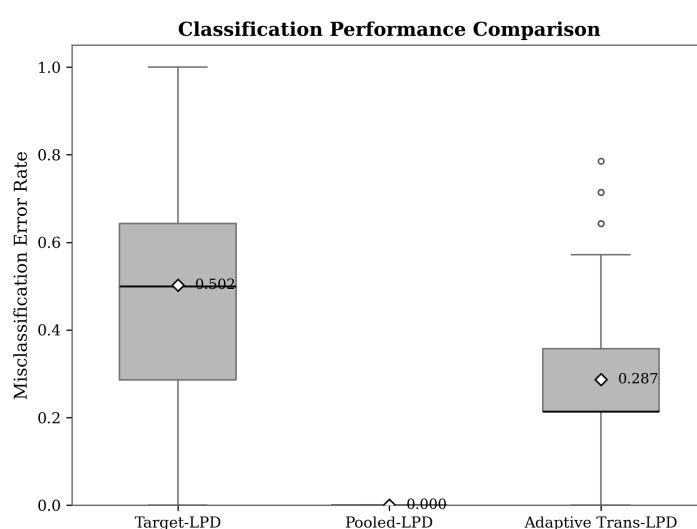


Figure 5. Boxplots of the misclassification error rates of Target-LPD, Pooled-LPD, and Adaptive-Trans-LPD over 100 replications on the TCGA-CHOL dataset.

5. Conclusions

This paper studied high-dimensional linear discriminant analysis with auxiliary data. We assumed that the divergence matrices between the target model and the informative auxiliary models were relatively sparse and that the Mahalanobis distance between the corresponding means was small. Based on the mean–variance similarity structure, we developed the Trans-LPD method to estimate the target discriminant direction. In particular, the exact source-domain moment discrepancy was decomposed into a covariance component $\Delta^{(k)}\delta$ and a mean component $\delta - \delta^{(k)}$, which gave a direct description of the transfer bias from auxiliary studies. The theoretical results showed that incorporating informative auxiliary data can reduce the sampling error by increasing the effective sample size, while the resulting classification performance also depended on the magnitude of the transfer bias.

An important practical issue in transfer learning was the selection of auxiliary studies. Rather than assuming that the informative auxiliary set was known, we constructed a data-driven source-ranking rule based on the empirical source-domain moment discrepancy. A nested sequence of transfer candidates was then constructed, and the final estimator was selected using an independent

target validation sample. Since the target-only LPD estimator was always included as a candidate, the resulting procedure reduced the risk of negative transfer up to the validation-selection error. For computation, we used the ADMM to solve the LPD optimization problems, instead of the interior point method used by Cai and Liu [25].

Several directions remain for future research. While this paper focuses on linear discriminant analysis for two-class problems, the method could also be considered for multi-class classification. Merchante et al. [43] studied penalized Optimal Scoring [44] with a group-Lasso penalty. By reformulating classification within a regression framework, their method provides a possible way to introduce transfer learning into multi-class discriminant analysis. Extending Trans-LPD in this direction is left for future work. Another direction is to study more efficient source-selection and validation strategies when the target sample size is extremely small or when a large number of auxiliary studies are available. Additionally, Lin and Li [2] considered transfer learning from a different perspective. It would also be useful to study how such ideas can be incorporated into high-dimensional discriminant analysis.

Appendix

Throughout the proofs, $C, C_1, C_2, \dots$ denote positive constants that may vary from line to line. For a fixed source set $S \subseteq \{1, \dots, K\}$, recall that

$$n_S = \sum_{k \in S} n^{(k)}, \quad N = n_T + n_S, \quad n_T = |I_x| + |I_y|.$$

The corresponding population pooled quantities are

$$\Sigma_P^S = \frac{n_T}{N} \Sigma + \sum_{k \in S} \frac{n^{(k)}}{N} \Sigma^{(k)}$$

and

$$\delta_P^S = \frac{n_T}{N} \delta + \sum_{k \in S} \frac{n^{(k)}}{N} \delta^{(k)}.$$

By (2.4),

$$\Sigma_P^S \beta - \delta_P^S = r_S, \quad r_S = \sum_{k \in S} \frac{n^{(k)}}{N} r^{(k)},$$

and

$$\eta_S = \|r_S\|_\infty.$$

Appendix A. Control of the transfer bias

We first prove (2.14). For $k \in \mathcal{A}_q(h_M, h_q)$, let

$$d_j^{(k)} = \mu_j^{(k)} - \mu_j, \quad j = 1, 2.$$

By the mean similarity condition,

$$(d_j^{(k)})^\top \Omega d_j^{(k)} \leq h_M.$$

Since $\lambda_{\max}(\Sigma) \leq C_2$,

$$\|d_j^{(k)}\|_2^2 \leq C_2 (d_j^{(k)})^\top \Omega d_j^{(k)} \leq C_2 h_M.$$

Hence,

$$\|\delta^{(k)} - \delta\|_\infty \leq \|d_1^{(k)}\|_2 + \|d_2^{(k)}\|_2 \leq C \sqrt{h_M}.$$

We next consider the covariance discrepancy. Recall that

$$\Delta^{(k)} = \Sigma^{(k)} \Omega - I_p.$$

For $0 < q \leq 1$, $D_q(\Omega^{(k)}, \Omega) \leq h_q$ implies

$$\max_i \|\Delta_{i,\cdot}^{(k)}\|_1 \leq \max_i \|\Delta_{i,\cdot}^{(k)}\|_q \leq h_q.$$

For $q = 0$, each row of $\Delta^{(k)}$ contains at most h_q nonzero entries. Moreover, Condition C1 gives

$$\|\Delta^{(k)}\|_{\max} \leq \|\Sigma^{(k)} \Omega\|_2 + 1 \leq C,$$

and therefore

$$\max_i \sum_j |\Delta_{ij}^{(k)}| \leq C h_q.$$

Since

$$\Delta_R = \delta^\top \Omega \delta \leq C_R.$$

Condition C1 implies

$$\|\delta\|_2 \leq C.$$

Thus, for $0 \leq q \leq 1$,

$$\|\Delta^{(k)} \delta\|_\infty \leq C h_q.$$

Using

$$r^{(k)} = \Delta^{(k)} \delta + (\delta - \delta^{(k)}),$$

we obtain

$$\|r^{(k)}\|_\infty \leq C \left(\sqrt{h_M} + h_q \right) = C \eta_0.$$

Consequently,

$$\begin{aligned} \eta_S &= \left\| \sum_{k \in S} \frac{n^{(k)}}{N} r^{(k)} \right\|_\infty \\ &\leq \sum_{k \in S} \frac{n^{(k)}}{N} \|r^{(k)}\|_\infty \\ &\leq C \frac{n_S}{N} \eta_0. \end{aligned}$$

This proves (2.14).

Appendix B. Proof of Lemma 2.1

Proof. By the exact pooled moment equation,

$$\Sigma_P^S \beta - \delta_P^S = r_S.$$

Therefore,

$$\widehat{\Sigma}_P^S \beta - \widehat{\delta}_P^S = (\widehat{\Sigma}_P^S - \Sigma_P^S) \beta - (\widehat{\delta}_P^S - \delta_P^S) + r_S.$$

It follows that

$$\begin{aligned} \left\| \widehat{\Sigma}_P^S \beta - \widehat{\delta}_P^S \right\|_{\infty} &\leq \left\| (\widehat{\Sigma}_P^S - \Sigma_P^S) \beta \right\|_{\infty} \\ &\quad + \left\| \widehat{\delta}_P^S - \delta_P^S \right\|_{\infty} + \eta_S. \end{aligned}$$

Under Condition C1, Gaussian concentration and a union bound give

$$\left\| (\widehat{\Sigma}_P^S - \Sigma_P^S) \beta \right\|_{\infty} = O_p \left(\sqrt{\frac{\log(pK)}{N}} \right)$$

and

$$\left\| \widehat{\delta}_P^S - \delta_P^S \right\|_{\infty} = O_p \left(\sqrt{\frac{\log(pK)}{N}} \right).$$

Hence,

$$\left\| \widehat{\Sigma}_P^S \beta - \widehat{\delta}_P^S \right\|_{\infty} \leq C \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\}$$

with probability tending to one. Thus, for

$$\lambda_S \geq C \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\},$$

β belongs to the feasible set.

Let

$$u = \widehat{\beta}^S - \beta, \quad \mathcal{I} = \text{supp}(\beta).$$

Since $\widehat{\beta}^S$ minimizes the ℓ_1 norm and β is feasible,

$$\|\widehat{\beta}^S\|_1 \leq \|\beta\|_1.$$

Hence,

$$\|u_{\mathcal{I}^c}\|_1 \leq \|u_{\mathcal{I}}\|_1,$$

and therefore

$$\|u\|_1 \leq 2\|u_{\mathcal{I}}\|_1 \leq 2\sqrt{s}\|u\|_2.$$

Since both $\widehat{\beta}^S$ and β are feasible,

$$\|\widehat{\Sigma}_P^S u\|_{\infty} \leq 2\lambda_S.$$

Under Conditions C1–C2, standard sparse quadratic concentration for Gaussian observations gives, with probability tending to one,

$$v^\top \widehat{\Sigma}_p^S v \geq c \|v\|_2^2$$

uniformly over vectors satisfying

$$\|v_{I^c}\|_1 \leq \|v_I\|_1.$$

Indeed, $\lambda_{\min}(\widehat{\Sigma}_p^S) \geq C_1$, while $N \geq Cs \log(pK)$ guarantees the corresponding restricted eigenvalue condition for $\widehat{\Sigma}_p^S$.

Consequently,

$$\begin{aligned} c \|u\|_2^2 &\leq u^\top \widehat{\Sigma}_p^S u \\ &\leq \|u\|_1 \|\widehat{\Sigma}_p^S u\|_\infty \\ &\leq 4 \sqrt{s} \lambda_S \|u\|_2. \end{aligned}$$

Thus,

$$\|\widehat{\beta}^S - \beta\|_2 \leq C \sqrt{s} \lambda_S.$$

Substituting the order of λ_S yields

$$\|\widehat{\beta}^S - \beta\|_2 \leq C \sqrt{s} \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\}.$$

This proves the lemma. □

Appendix C. Proof of Theorem 2.1

Proof. For a fixed source set S , let

$$u = \widehat{\beta}^S - \beta, \quad \mu = \frac{\mu_1 + \mu_2}{2},$$

and write

$$b_S = \sqrt{s} \left\{ \sqrt{\frac{\log(pK)}{N}} + \eta_S \right\}.$$

By Lemma 2.1,

$$\|u\|_2 = O_p(b_S).$$

Recall that

$$\delta = \Sigma \beta$$

and

$$\Delta_R = \delta^\top \Omega \delta = \beta^\top \Sigma \beta.$$

Since $\|\delta\|_2 \leq C$ under Condition C1,

$$\begin{aligned} \delta^\top \widehat{\beta}^S - \Delta_R &= \delta^\top u \\ &= O_p(b_S). \end{aligned}$$

Therefore,

$$\delta^\top \widehat{\beta}^S = \Delta_R + O_p(b_S). \tag{5.1}$$

For the denominator,

$$\begin{aligned}(\widehat{\beta}^S)^\top \Sigma \widehat{\beta}^S - \Delta_R &= 2\beta^\top \Sigma u + u^\top \Sigma u \\ &= 2\delta^\top u + u^\top \Sigma u.\end{aligned}$$

Since $\lambda_{\max}(\Sigma) \leq C_2$ and $b_S = o(1)$ under the conditions of the theorem,

$$u^\top \Sigma u \leq C_2 \|u\|_2^2 = O_p(b_S^2) = o_p(b_S).$$

Hence,

$$(\widehat{\beta}^S)^\top \Sigma \widehat{\beta}^S = \Delta_R + O_p(b_S). \quad (5.2)$$

We next consider estimation of the target midpoint. Since

$$\widehat{\mu}_T = \frac{\bar{x}_T + \bar{y}_T}{2}, \quad \mu = \frac{\mu_1 + \mu_2}{2},$$

Gaussian concentration gives

$$\|\widehat{\mu}_T - \mu\|_\infty = O_p\left(\sqrt{\frac{\log p}{n_T}}\right).$$

Moreover, since β is feasible,

$$\|\widehat{\beta}^S\|_1 \leq \|\beta\|_1.$$

Conditions C1–C2 imply

$$\|\beta\|_1 \leq \sqrt{s}\|\beta\|_2 \leq C\sqrt{s}.$$

Therefore,

$$\left|(\widehat{\mu}_T - \mu)^\top \widehat{\beta}^S\right| = O_p\left(\sqrt{\frac{s \log p}{n_T}}\right). \quad (5.3)$$

Let R_n^S denote the conditional misclassification rate of the classifier based on $\widehat{\beta}^S$. Define

$$A_1^S = \frac{(\mu_1 - \widehat{\mu}_T)^\top \widehat{\beta}^S}{\{(\widehat{\beta}^S)^\top \Sigma \widehat{\beta}^S\}^{1/2}}$$

and

$$A_2^S = \frac{(\widehat{\mu}_T - \mu_2)^\top \widehat{\beta}^S}{\{(\widehat{\beta}^S)^\top \Sigma \widehat{\beta}^S\}^{1/2}}.$$

Then

$$R_n^S = 1 - \frac{1}{2}\Phi(A_1^S) - \frac{1}{2}\Phi(A_2^S).$$

Since

$$\mu_1 - \mu = \frac{\delta}{2}, \quad \mu - \mu_2 = \frac{\delta}{2},$$

Eqs (5.1)–(5.3) imply

$$A_j^S = \frac{\sqrt{\Delta_R}}{2} + O_p(a_S), \quad j = 1, 2,$$

where

$$a_S = \sqrt{\frac{s \log(pK)}{N}} + \sqrt{s}\eta_S + \sqrt{\frac{s \log p}{n_T}}.$$

Since Φ is Lipschitz continuous,

$$|R_n^S - R| = O_p(a_S).$$

Furthermore, Condition C1 gives

$$0 < c_R \leq \Delta_R \leq C_R < \infty,$$

so

$$R = 1 - \Phi\left(\frac{\sqrt{\Delta_R}}{2}\right)$$

is bounded away from zero. Therefore,

$$\left|\frac{R_n^S}{R} - 1\right| = O_p(a_S),$$

that is,

$$\left|\frac{R_n^S}{R} - 1\right| = O_p\left(\sqrt{\frac{s \log(pK)}{N}} + \sqrt{s}\eta_S + \sqrt{\frac{s \log p}{n_T}}\right).$$

If

$$\sqrt{\frac{s \log(pK)}{N}} + \sqrt{s}\eta_S + \sqrt{\frac{s \log p}{n_T}} = o(1),$$

then

$$\frac{R_n^S}{R} \rightarrow 1$$

in probability. This proves Theorem 2.1. $\square$

Appendix D. Proof of Theorem 2.2

Proof. Condition on the target training observations indexed by I_x and I_y and on all auxiliary samples. Then all candidate classifiers

$$\{\widehat{\psi}_{m,\lambda} : (m, \lambda) \in \mathcal{M}\}$$

are fixed and are independent of the validation samples indexed by V_x and V_y .

For each $(m, \lambda) \in \mathcal{M}$, let

$$R_n(m, \lambda)$$

denote its conditional misclassification rate in the target population.

Write the class-specific validation errors as

$$\widehat{R}_{V,1}(m, \lambda) = \frac{1}{|V_x|} \sum_{i \in V_x} \mathbf{1}\{\widehat{\psi}_{m,\lambda}(x_i) = 0\}$$

and

$$\widehat{R}_{V,2}(m, \lambda) = \frac{1}{|V_y|} \sum_{i \in V_y} \mathbf{1}\{\widehat{\psi}_{m,\lambda}(y_i) = 1\}.$$

Thus,

$$\widehat{R}_V(m, \lambda) = \frac{1}{2} \left\{ \widehat{R}_{V,1}(m, \lambda) + \widehat{R}_{V,2}(m, \lambda) \right\}.$$

By Hoeffding's inequality and a union bound over the two classes and all $(m, \lambda) \in \mathcal{M}$, for every $0 < \alpha < 1$, with probability at least $1 - \alpha$,

$$\sup_{(m, \lambda) \in \mathcal{M}} \left| \widehat{R}_V(m, \lambda) - R_n(m, \lambda) \right| \leq \varepsilon_V,$$

where

$$\varepsilon_V = \sqrt{\frac{\log(4|\mathcal{M}|/\alpha)}{2n_{V,\min}}}, \quad n_{V,\min} = \min\{|V_x|, |V_y|\}.$$

By the definition of the final estimator,

$$(\widehat{m}, \widehat{\lambda}) \in \arg \min_{(m, \lambda) \in \mathcal{M}} \widehat{R}_V(m, \lambda).$$

Hence, for every $(m, \lambda) \in \mathcal{M}$,

$$\begin{aligned} R_n &= R_n(\widehat{m}, \widehat{\lambda}) \\ &\leq \widehat{R}_V(\widehat{m}, \widehat{\lambda}) + \varepsilon_V \\ &\leq \widehat{R}_V(m, \lambda) + \varepsilon_V \\ &\leq R_n(m, \lambda) + 2\varepsilon_V. \end{aligned}$$

Taking the minimum over $\mathcal{M}$ gives

$$R_n \leq \min_{(m, \lambda) \in \mathcal{M}} R_n(m, \lambda) + 2\sqrt{\frac{\log(4|\mathcal{M}|/\alpha)}{2n_{V,\min}}},$$

which proves (2.16).

For $m = 0$,

$$S_0 = \emptyset, \quad \widehat{\beta}^{S_0} = \widehat{\beta}^{(L)}, \quad \Lambda_0 = \{\lambda_L\}.$$

Therefore, the target-only LPD classifier belongs to the candidate family. Denoting its conditional misclassification rate by $R_n^{(L)}$, we have

$$R_n \leq R_n^{(L)} + 2\sqrt{\frac{\log(4|\mathcal{M}|/\alpha)}{2n_{V,\min}}}.$$

This proves (2.17).

Moreover,

$$\sup_{(m, \lambda) \in \mathcal{M}} \left| \widehat{R}_V(m, \lambda) - R_n(m, \lambda) \right| = O_p \left(\sqrt{\frac{\log |\mathcal{M}|}{n_{V,\min}}} \right).$$

Thus, if

$$\frac{\log |\mathcal{M}|}{n_{V,\min}} \longrightarrow 0,$$

then

$$R_n \leq R_n^{(L)} + o_p(1).$$

This proves (2.18) and completes the proof of Theorem 2.2. $\square$

Appendix E. Accuracy of the source-ranking procedure

For completeness, we give a sufficient condition for the data-driven source ranking to recover a separated low-discrepancy source set.

Recall that

$$\tau_k = \|\Sigma^{(k)}\beta - \delta^{(k)}\|_\infty = \|r^{(k)}\|_\infty.$$

From the definition of $\widehat{\tau}_k$,

$$\begin{aligned} |\widehat{\tau}_k - \tau_k| &\leq \left\| (\widehat{\Sigma}^{(k)} - \Sigma^{(k)})\beta \right\|_\infty \\ &\quad + \left\| \widehat{\Sigma}^{(k)}(\widehat{\beta}^{(L)} - \beta) \right\|_\infty \\ &\quad + \left\| \widehat{\delta}^{(k)} - \delta^{(k)} \right\|_\infty. \end{aligned}$$

Under Conditions C1–C2,

$$\max_{1 \leq k \leq K} \left\| (\widehat{\Sigma}^{(k)} - \Sigma^{(k)})\beta \right\|_\infty = O_p \left(\max_{1 \leq k \leq K} \sqrt{\frac{\log(pK)}{n^{(k)}}} \right)$$

and

$$\max_{1 \leq k \leq K} \left\| \widehat{\delta}^{(k)} - \delta^{(k)} \right\|_\infty = O_p \left(\max_{1 \leq k \leq K} \sqrt{\frac{\log(pK)}{n^{(k)}}} \right).$$

For the target-only LPD estimator,

$$\|\widehat{\beta}^{(L)} - \beta\|_1 = O_p \left(s \sqrt{\frac{\log p}{n_T}} \right).$$

Since

$$\max_{1 \leq k \leq K} \|\widehat{\Sigma}^{(k)}\|_{\max} = O_p(1),$$

we obtain

$$\max_{1 \leq k \leq K} |\widehat{\tau}_k - \tau_k| = O_p(\varepsilon_n),$$

where

$$\varepsilon_n = \max_{1 \leq k \leq K} \sqrt{\frac{\log(pK)}{n^{(k)}}} + s \sqrt{\frac{\log p}{n_T}}.$$

Let $S^\star$ be a benchmark low-discrepancy source set. If

$$\max_{k \in S^\star} \tau_k + 2\varepsilon_n < \min_{k \notin S^\star} \tau_k,$$

then, with probability tending to one,

$$\widehat{\tau}_k < \widehat{\tau}_l, \quad k \in S^\star, \quad l \notin S^\star.$$

Therefore, if $m^\star = |S^\star|$,

$$S_{m^\star} = S^\star$$

with probability tending to one.

Author contributions

All authors contributed to the conception and design of the study, the analysis and interpretation of the results, and the preparation and revision of the manuscript. All authors have read and approved the final version.

Use of Generative-AI tools declaration

During the preparation of this manuscript, the authors used Artificial Intelligence (AI) to assist with language editing and improving the clarity of the text. The authors reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.

Acknowledgments

Xiaoyu Ma's research was supported by the National Natural Science Foundation of China (No. 12401397) and the Natural Science Foundation of Hunan Province (No. 2026JJ60117). Lu Lin's research was supported by the National Natural Science Foundation of China (No. 12471281) and the National Statistical Science Research Project (No. 2022LD03).

Conflict of interest

The authors declare no conflicts of interest.

References

1. L. Torrey, J. Shavlik, *Transfer learning*, In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 2010, 242–264. <https://doi.org/10.4018/978-1-60566-766-9.ch011>
2. L. Lin, W. Li, A Correlation-ratio transfer learning and variational Stein's paradox, *arXiv preprint*, 2022, arXiv:2206.06086. <https://doi.org/10.48550/arXiv.2206.06086>
3. L. Lin, W. Li, Source-function weighted-transfer learning for nonparametric regression with seemingly similar sources, *arXiv preprint*, 2023, arXiv:2302.11222. <https://doi.org/10.48550/arXiv.2302.11222>
4. S. Pan, Q. Yang, A survey on transfer learning, *IEEE T. Knowl. Data Eng.*, **22** (2010), 1345–1359. <https://doi.org/10.1109/TKDE.2009.191>
5. S. Mei, W. Fei, S. Zhou, Gene ontology based transfer learning for protein subcellular localization, *BMC Bioinform.*, **12** (2011), 44. <https://doi.org/10.1186/1471-2105-12-44>
6. H. C. Shin, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, *IEEE T. Med. Imaging*, **35** (2016), 1285–1298. <https://doi.org/10.1109/TMI.2016.2528162>
7. T. Turki, Z. Wei, J. Wang, Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, **5** (2017), 7381–7393. <https://doi.org/10.1109/ACCESS.2017.2696523>

8. Y. Sun, Y. Hu, Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases, *Adv. Genet.*, **93** (2016), 147–190. <https://doi.org/10.1016/bs.adgen.2015.11.004>
9. Y. Hu, M. Li, Q. Lu, H. Weng, J. Wang, S. M. Zekavat, et al., A statistical framework for cross-tissue transcriptomewide association analysis, *Nat. Genet.*, **51** (2019), 568–576. <https://doi.org/10.1038/s41588-019-0345-7>
10. Z. Huang, Y. Song, Q. Zhang, Individualized treatment rules based on adaptive transfer-dragonnet, *Stat. Comput.*, **35** (2025), 165. <https://doi.org/10.1007/s11222-025-10704-9>
11. H. Daumé III, *Frustratingly easy domain adaptation*, In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, 256–263.
12. W. Pan, Q. Yang, Transfer learning in heterogeneous collaborative filtering domains, *Artif. Intell.*, **197** (2013), 39–55. <https://doi.org/10.1016/j.artint.2013.01.003>
13. T. Cai, H. Wei, Transfer learning for nonparametric classification: Minimax rate and adaptive classifier, *Ann. Stat.*, **49** (2021), 100–128. <https://doi.org/10.1214/20-AOS1949>
14. S. Li, T. Cai, R. Duan, Targeting underrepresented populations in precision medicine: A federated transfer learning approach, *arXiv preprint*, 2021. <https://doi.org/10.48550/arXiv.2108.12112>
15. S. Li, T. Cai, H. Li, Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality, *J. Roy. Stat. Soc. Ser. B*, **84** (2022), 149–173. <https://doi.org/10.1111/rssb.12479>
16. S. Li, T. Cai, H. Li, Estimation and inference with proxy data and its genetic applications, *arXiv preprint*, 2022. <https://doi.org/10.48550/arXiv.2201.03727>
17. S. Li, T. Cai, H. Li, Transfer learning in large-scale graphical models with false discovery rate control, *J. Am. Stat. Assoc.*, **118** (2023), 2171–2183. <https://doi.org/10.1080/01621459.2022.2044333>
18. T. Gu, Y. Han, R. Duan, Robust angle-based transfer learning in high dimensions, *J. Roy. Stat. Soc. Ser.*, **87** (2025), 723–745. <https://doi.org/10.1093/jrsssb/qkae111>
19. Y. Tian, Y. Feng, Transfer learning under high-dimensional generalized linear models, *J. Am. Stat. Assoc.*, **118** (2022), 2684–2697. <https://doi.org/10.1080/01621459.2022.2071278>
20. F. Sun, Q. Zhang, Robust transfer learning of high-dimensional generalized linear model, *Physica A*, **618** (2023), 128674. <https://doi.org/10.1016/j.physa.2023.128674>
21. H. Lin, M. Reimherr, On transfer learning in functional linear regression, *arXiv preprint*, 2023. <https://doi.org/10.48550/arXiv.2206.04277>
22. M. Wu, L. Zhang, Z. Wang, D. Christiani, X. Lin, Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection, *Bioinformatics*, **25** (2009), 1145–1151. <https://doi.org/10.1093/bioinformatics/btp019>
23. D. Witten, R. Tibshirani, Covariance-regularized regression and classification for high dimensional problems, *J. Roy. Stat. Soc. Ser. B*, **71** (2009), 615–636. <https://doi.org/10.1111/j.1467-9868.2009.00699.x>
24. J. Shao, Y. Wang, X. Deng, S. Wang, Sparse linear discriminant analysis with high dimensional data, *Ann. Stat.*, **39** (2011), 1241–1265. <https://doi.org/10.1214/10-AOS870>

25. T. Cai, W. Liu, A direct estimation approach to sparse linear discriminant analysis, *J. Am. Stat. Assoc.*, **106** (2011), 1566–1577. <https://doi.org/10.1198/jasa.2011.tm11199>
26. Q. Mai, H. Zou, M. Yuan, A direct approach to sparse discriminant analysis in ultrahigh dimensions, *Biometrika*, **99** (2012), 29–42. <https://doi.org/10.1093/biomet/asr066>
27. J. Fan, Y. Feng, X. Tong, A road to classification in high dimensional space: the regularized optimal affine discriminant, *J. Roy. Stat. Soc. Ser. B*, **74** (2012), 745–771. <https://doi.org/10.1111/j.1467-9868.2012.01029.x>
28. M. Azizyan, A. Singh, L. Wasserman, *Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation*, In: Proc. Advances in Neural Information Processing Systems, Lake Tahoe, Dec. 5th-10th, 2013, 2139–2147.
29. T. Li, X. Yi, C. Carmanis, P. Ravikumar, *Minimax Gaussian classification & clustering*, In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017, 1–9.
30. T. Cai, L. Zhang, High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data, *J. Roy. Stat. Soc. Ser. B*, **81** (2019), 675–705. <https://doi.org/10.1111/rssb.12326>
31. S. Safo, E. Min, L. Haine, Sparse linear discriminant analysis for multiview structured data, *Biometrics*, **78** (2022), 612–623. <https://doi.org/10.1111/biom.13458>
32. T. Anderson, *An introduction to multivariate statistical analysis*, 3 Eds., New Jersey: Wiley-Interscience, 2003, 1566–1568.
33. Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, **8** (2007), 86–100. <https://doi.org/10.1093/biostatistics/kxj035>
34. A. Rothman, P. Bickel, E. Levina, J. Zhu, Sparse permutation invariant covariance estimation, *Electron. J. Stat.*, **2** (2008), 494–515. <https://doi.org/10.1214/08-EJS176>
35. P. Hall, D. M. Titterton, J. H. Xue, Median-based classifiers for high-dimensional data, *J. Am. Stat. Assoc.*, **104** (2009), 1597–1608. <https://doi.org/10.1198/jasa.2009.tm08107>
36. J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, B. Schölkopf, *Correcting sample selection bias by unlabeled data*, In: Proceedings of NIPS, 2006, 601–608. <https://doi.org/10.7551/mitpress/7503.003.0080>
37. B. Quanz, J. Huan, *Large margin transductive transfer learning*, In: Proceedings of CIKM, 2009, 1327–1336. <https://doi.org/10.1145/1645953.1646121>
38. S. Hanneke, S. Kpotufe, A no-free-lunch theorem for multitask learning, *arXiv preprint*, 2020. <https://doi.org/10.48550/arXiv.2006.15785>
39. R. Glowinski, A. Marrocco, Sur l’D. Gabay and B. Mercier approximation, par éléments finis d’ordre un, et la résolution, par pénalisation-dualité, d’une classe de problèmes de Dirichlet non linéaires, *Rev. Française D’automatique, Informatique, Recherche Opérationnelle*, **2** (1975), 41–76.
40. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. Appl.*, **2** (1976), 17–40. [https://doi.org/10.1016/0898-1221\(76\)90003-1](https://doi.org/10.1016/0898-1221(76)90003-1)

41. Z. Lu, T. Pong, Y. Zhang, An alternating direction method for finding Dantzig selectors, *Comput. Stat. Data Anal.*, **56** (2012), 4037–4046. <https://doi.org/10.1016/j.csda.2012.04.019>
42. X. Wang, X. Yuan, The linearized alternating direction method of multipliers for Dantzig selector, *SIAM J. Sci. Comput.*, **34** (2012), 2792–2811. <https://doi.org/10.1137/110833543>
43. L. Merchante, Y. Grandvalet, G. Govaert, *An efficient approach to sparse linear discriminant analysis*, In: Proceedings of the 29th International Conference on Machine Learning, 2012.
44. T. Hastie, A. Buja, R. Tibshirani, Penalized discriminant analysis, *Ann. Stat.*, **23** (1995), 73–102. <https://doi.org/10.1214/aos/1176324456>



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)