Statistica Sinica Preprint No: SS-2025-0232	
Title	Transfer Learning for Ridge Regression with Random Coefficients
Manuscript ID	SS-2025-0232
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0232
Complete List of Authors	Hongzhe Zhang and Hongzhe Li
Corresponding Authors	Hongzhe Li
E-mails	hongzhe@upenn.edu
Notice: Accepted author version.	

TRANSFER LEARNING FOR RIDGE REGRESSION WITH RANDOM COEFFICIENTS

Hongzhe Zhang, and Hongzhe Li

Department of Biostatistics, Epidemiology and Informatics

University of Pennsylvania

Abstract: Ridge regression with random coefficients provides a flexible approach for modeling many small but nonzero effects in high-dimensional data. We embed this framework in transfer learning by leveraging source samples from related regression models: the informativeness of each source is captured via the correlation between its coefficients and those of the target. We propose two weighted estimators—one minimizing estimation risk and the other minimizing prediction risk—each formed as an optimal blend of target and source ridge estimates. Under the high-dimensional regime $p/n \to \gamma$, where p is the number of the predictors and n is the sample size, random matrix theory yields closed-form limits for these optimal weights and their associated risks. Through simulations and applications to lipid-trait and colorectal-cancer microbiome prediction, our methods consistently outperform both target-only and pooled-data ridge regression.

Key words and phrases: Covariate shift; estimation risk; prediction risk; random matrix theory

1. Introduction

Large, diverse datasets, common in genomics and medical research, invite integration across studies to improve estimation and prediction. Transfer learning (Torrey and Shavlik, 2010) does exactly that, borrowing strength from related but distinct datasets when target outcomes are scarce or costly. It has found success in protein localization (Mei et al., 2011), imaging diagnosis (Shin et al., 2016), drug-sensitivity prediction (Turki et al., 2017), multi-omics integration (Hu et al.,

2019; Wang et al., 2019), NLP (Daumé III, 2007), and recommendation systems (Pan and Yang, 2013).

Most high-dimensional transfer methods assume sparsity—e.g. Li et al. (2022) use sparse regressions to predict gene expression across tissues. But in settings like polygenic risk–score (PRS) estimation—where thousands of weakly associated variants exceed sample sizes—sparse penalties exclude too many loci. Ridge regression, which accommodates linkage disequilibrium without enforcing sparsity, has outperformed Lasso for PRS (Marotta et al., 2021) and proven effective in microbiome and metabolomic phenotype prediction (Rothschild et al., 2022; Faquih et al., 2023).

In this paper, we consider the problem of transfer learning for ridge regressions for outcome predictions using high dimensional predictors. Specially, the target model is characterized by regression coefficients β_K,

$$\underbrace{y_K}_{\text{Target outcome}} = \beta_{K1}\underbrace{x_{K1}}_{\text{Var 1}} + \beta_{K2}\underbrace{x_{K2}}_{\text{Var 2}} + \cdots + \beta_{Kp}\underbrace{x_{Kp}}_{\text{Var p}} + \epsilon_K \tag{1.1}$$

where $\beta_K = (\beta_{Kj}, j = 1, \cdots, p)$ is the vector of regression coefficients, and $x_K = (x_{K1}, \cdots, x_{Kp})$ is the predictor vector. In the setting of ridge transfer learning, we assume that we additionally have data from $K-1$ source models given as

$$\underbrace{y_k}_{\text{Source outcomes}} = \beta_{k1}\underbrace{x_{k1}}_{\text{Var 1}} + \beta_{k2}\underbrace{x_{k2}}_{\text{Var 2}} + \cdots + \beta_{kp}\underbrace{x_{kp}}_{\text{Var p}} + \epsilon_k, \tag{1.2}$$

for $k = 1, \ldots, K-1$, let $\beta_k = (\beta_{k1}, \ldots, \beta_{kp})$ denote the coefficient vector for study k and $x_k = (x_{k1}, \ldots, x_{kp})$ the predictor vector. Across studies $k = 1, \ldots, K$ we observe n_k i.i.d. pairs $\{(Y_{ki}, X_{ki})\}_{i=1}^{n_k}$, stacked as $Y_k = (Y_{k1}, \ldots, Y_{kn_k})^\top$ and $X_k = [X_{k1}^\top; \ldots; X_{kn_k}^\top] \in \mathbb{R}^{n_k \times p}$. Here Y_{ki} and X_{ki} are the outcome and predictor for subject i in study k. In PRS applications, $X_{k,ij}$ is the

standardized score of variant j for individual i in study k. Futher, we adopt a random-coefficients (RRC) perspective for both the target (1.1) and the sources (1.2): each coefficient vector β_k is mean-zero with coordinatewise variance α_k^2/p. This induces a cross-population similarity parameter $\rho_{kk'}$ for $(\beta_k, \beta_{k'})$ that quantifies transferability across studies and serves as the key quantity of our weighting scheme introduced later.

In genetic applications, $(\alpha_k^2, \rho_{kk'})$ align with SNP-heritability h_k^2 and genetic correlation $\varphi_{kk'}$ (Zhou et al., 2020), defined by

$$h_k^2 := \frac{\beta_k^\top \Sigma \beta_k}{\beta_k^\top \Sigma \beta_k + \sigma_k^2}, \qquad \varphi_{kk'} := \frac{\beta_k^\top \Sigma \beta_{k'}}{\sqrt{\beta_k^\top \Sigma \beta_k \, \beta_{k'}^\top \Sigma \beta_{k'}}}.$$

Under standard high-dimensional scaling, $h_k^2 \to \alpha_k^2 \sigma_k^2$ and $\varphi_{kk'} \to \rho_{kk'}$, where σ_k^2 is the variance of the error terms ϵ_k. Consequently, $(\alpha_k^2, \rho_{kk'})$ can be estimated consistently with established tools (e.g., Lee et al., 2012). We treat them as known inputs in our theory. In other settings, closely analogous estimates arise from simple Bayesian hierarchical regressions (see Supplement Section F for an example). This correspondence provides both a principled motivation for our modeling choices and a practical advantage of the proposed method: the quantities that determine knowledge transfer are measurable from data and readily interpretable.

Ridge regressions under the RRC assumptions have been studied by Dobriban and Wager (2018), Sheng and Dobriban (2020) and Zhao and Zhu (2019). Dobriban and Wager (2018) was the first to use the random matrix theory (RMT) results on trace of functions of sample and population covariances to find the limiting prediction risk of ridge regression. Sheng and Dobriban (2020) extended ridge regression to a distributed learning setting, and Zhao and Zhu (2019) studied the in-sample and out-of-sample R^2 of several ridge-type estimators using the same RMT results. A difference is that their results are stated mostly in terms of moments of population spectral distri-

butions. Among these works, Zhao and Zhu (2019) and Sheng and Dobriban (2020) aim to obtain a better estimator for β by integrating information from multiple sources of data. Both papers use a weighted sum of the one sample ridge estimators as an aggregated estimator:

$$\hat{\beta} = \sum_{k=1}^{K} W_k \hat{\beta}_k, \tag{1.3}$$

where $\hat{\beta}_k = (X_k^T X_k + \lambda_k I)^{-1} X_k^T Y_k$ is the ridge estimator based on data from the k^{th} study (Sheng and Dobriban, 2020) with tuning parameter λ_k. They study the aggregated ridge estimator that optimizes a limiting estimation risk by assuming a common set of linear coefficients β in all populations. Zhao and Zhu (2019) proposes an aggregated marginal estimator that ignores the correlation structure of the design matrix to optimize the out-of-sample R^2. Although they still assume a common β in all training populations, their objective is to explain the most variance in an "out-of-sample" population with different coefficients. For the purpose of PRS, the idea of using linear combinations of PRSs across different populations is considered by Márquez-Luna et al. (2017). Zhao et al. (2022) takes a more algorithmic approach to fine-tune an existing PRS with a gradient descent.

Building on the linear combination in (1.3), our goal is to optimize performance in the Kth (target) study. We propse to determine the weight vector W by two criteria: (i) the limiting weight that minimizes estimation risk and (ii) the limiting weight that minimizes prediction risk. We refer to the resulting estimator as Trans-Ridge. We further show that minimizing estimation risk coincides with minimizing a form of prediction risk that is robust to covariate shift between the training and test covariate distributions. The cross-study coefficient correlation ρ_{kK} governs transferability: it determines how much each source contributes to the final estimator. We derive explicit expressions for the limiting risks and the associated optimal weights, which are consistently estimable

from data under general, study-heterogeneous covariance structures.

From a technical standpoint, our analysis hinges on the interplay among the spectra of sample covariance matrices drawn from different studies under anisotropic population covariances. The key weight and risk expressions reduce to normalized traces of cross-population resolvent products, which intrinsically couple heterogeneous aspect ratios (p/n_k) and ridge parameters (λ_k) across studies. Such objects fall outside classical single-sample or homogeneous-design random-coefficient ridge analyses (Dobriban and Wager, 2018; Sheng and Dobriban, 2020; Zhao and Zhu, 2019). We therefore derive new almost-sure limits for these cross-population resolvent traces—stated as Lemmas **??** and **??**—which underpin our main results and are of independent interest for multi-cohort inference.

The rest of the paper is organized as follows. Section 2 introduces the modeling assumptions and spectral notations used throughout. Section 3 develops the estimation–risk criterion for the aggregated estimator and derives the corresponding optimal weight. We then show how the optimal limiting estimation weight adapts to the strength of the regression coefficient correlation in Section 4. We discuss the prediction risk in Section 5, where the optimal weight that minimizes the prediction risk is proposed. In Section 6, we give a brief discussion on the robustness of the transfer ridge estimator under the optimal estimation weight with respective to distributional shift. Finally, in Section 7, we present applications of the proposed method to the lipid phenotype predictions using genetic data and the colorectal cancer risk prediction with microbiome data.

The codes to implement the proposed method is available at `https://github.com/hongzhezzzz/Trans-Ridge`.

2. Modeling Framework and Spectral Preliminaries

In this section we formalize the data-generating setup and state the modeling assumptions used throughout. We then collect the spectral regime and random matrix theory (RMT) notations needed to state the random matrix limits that underpin our limiting risks and optimal transfer weights.

2.1 Observation Under Random Coefficients

Recall we observe K studies (Y_k, X_k) each with sample sizes n_k and common feature dimension p. We impose the following normalization assumption

Assumption 1 (Standardized Observations)**.** For each study $k = 1, \ldots, K$ and observation $i = 1, \ldots, n_k$,

$$E[(Y_k)_i] = 0, \quad \mathrm{Var}[(Y_k)_i] = 1, \qquad X_k = Z_k \, \Sigma^{1/2},$$

where $\Sigma \in \mathbb{R}^{p \times p}$ is a deterministic positive semidefinite covariance with $\Sigma_{jj} = 1$, and the entries of Z_k satisfy $E[(Z_k)_{ij}] = 0$ and $\mathrm{Var}[(Z_k)_{ij}] = 1$.

We work under the RRC framework that encodes study-level signal strength and cross-study similarity.

Assumption 2 (Random Coefficients Regression)**.** For $k = 1, \ldots, K$, the regression coefficients satisfy $E(\beta_k) = 0$ and $\mathrm{Var}\big(\sqrt{p}\,\beta_k\big) = \alpha_k^2 \, \sigma_k^2 \, I_p$, with noise variance $\mathrm{Var}(\epsilon_k) = \sigma_k^2$.

Assumption 3 (Correlated Random Coefficients)**.** For $k \neq k'$,

$$\mathrm{Cov}\big(\sqrt{p}\,\beta_k, \sqrt{p}\,\beta_{k'}\big) = \rho_{kk'} \, \alpha_k \alpha_{k'} \sigma_k \sigma_{k'} \, I_p,$$

where $\rho_{kk'} \in [0, 1]$ quantifies cross-study similarity.

For simplicity of exposition, we restrict to nonnegative correlations, $\rho_{kk'} \in [0, 1]$. All results extend symmetrically to negative correlations. For example, reparametrizing a source outcome y_k to $-y_k$ flips the sign of its cross-study correlations without otherwise altering the setup.

2.2 Spectral Regime and Notations

In this study, we adopt the proportional high-dimensional regime that yields Marchenko–Pastur limits for sample covariances, which is formally defined below

Assumption 4 (RMT assumption)**.** As $p \to \infty$ and $n_k \to \infty$ for all k,

1. $p/n_k \to \gamma_k \in (0, \infty)$ for $k = 1, \ldots, K$;
2. Define F_Σ as the spectral distribution for a $p-$dimensional covariance matrix Σ that places equal point mass on the eigenvalues of Σ. F_Σ converges weakly to a limiting population spectral distribution H (PSD) on $[0, \infty)$.

Under Assumption 4, the empirical spectral distribution (ESD) of the sample covariance

$$\hat{\Sigma}_k := X_k^\top X_k / n_k$$

converges almost surely to a deterministic Marchenko–Pastur limit F_{γ_k} for aspect ratio γ_k supported on $[0, \infty)$.

We continue to define the Stieltjes and its companion transforms. They let us encode the empirical spectral distributions through resolvents so that our risks and optimal weights admit explicit formula. For a probability measure G on $[0, \infty)$, define the Stieltjes transform

$$m_G(z) := \int \frac{dG(t)}{t - z}, \qquad z \in \mathbb{C} \setminus \mathbb{R}_+,$$

and write $m'_G(z)$ for its derivative. Let $v_G(z)$ denote the companion Stieltjes transform, related via

$$\gamma\big(m_G(z) + z^{-1}\big) = v_G(z) + z^{-1}.$$

Then, for each study k,

$$p^{-1}\operatorname{tr}\left((\hat{\Sigma}_k - zI)^{-1}\right) \xrightarrow{a.s.} m_{F_{\gamma_k}}(z), \qquad p^{-1}\operatorname{tr}\left((\hat{\Sigma}_k - zI)^{-2}\right) \xrightarrow{a.s.} m'_{F_{\gamma_k}}(z),$$

uniformly on compact subsets of $\mathbb{C} \setminus \mathbb{R}_+$; see Silverstein (1995). We will also use $v_{F_{\gamma_k}}(z)$ and $v'_{F_{\gamma_k}}(z)$ via the identity above.

Finally, we impose uniform moment bounds (assumption 5) and exclude excessive spectral mass near zero (assumption 6)to ensure resolvents stability and negligible fluctuations for normalized traces of cross-population resolvent products. In certain structured cases, these two conditions can be waived, see lemma **??** and lemma **??** for details.

Assumption 5 (Bounded moments)**.** For each integer $\ell \geq 1$, there exists $C_\ell < \infty$ such that $E\,|(Z_k)_{ij}|^\ell \leq C_\ell$ for all i, j, k.

Assumption 6 (Anisotropic local laws)**.** Let $F(x) = p^{-1}\sum_{i=1}^{p} \mathbb{I}(\lambda_i \leq x)$ denote the cdf of H, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of Σ. For an arbitrary $\tau > 0$, assume $F(\tau) \leq 1 - \tau$.

With these ingredients in place, we now introduce the proposed Trans-Ridge estimator and characterize its optimal weights and risks in the proportional regime.

3. Optimal Weight that Minimizes Estimation Risk And Its Limiting Value

3.1 Optimal weight that minimizes the estimation risk

Define the regular ridge estimator for population $k, k = 1, \cdots, K$ as

$$\hat{\beta}_k = (X_k^T X_k + n_k \lambda_k I_p)^{-1} X_k^T Y_k = (\hat{\Sigma}_k + \lambda_k I_p)^{-1} X_k^T Y_k / n_k$$

Here we hide the dependence of $\hat{\beta}_k$ on tuning parameter λ_k. We propose the Trans-Ridge estimator of β_K as a sum of the study-specific ridge estimators, including $\hat{\beta}_K$ but weighted by W (1.3). We first consider choosing the weight W that minimizes the estimation risk, defined as

$$M(W) = E_{\epsilon_1, \cdots, \epsilon_K} \left(|| \sum_{k=1}^{K} W_k \hat{\beta}_k - \beta_K ||^2 | X_1, \cdots, X_K \right),$$

where $W = (W_1, \cdots, W_K)^T$, and the expectation is over the randomness in $(\epsilon_1, \cdots, \epsilon_K)$. We denote the optimal estimation weight that minimizes the $M(W)$ as W_E^*. We emphasize that both $M(W)$ and W_E^* are conditioned on the random training data $(X_1, \cdots, X_K)$, and are therefore random variables. The following theorem states the expressions for W_E^* as well as $M(W_E^*)$.

Theorem 1 (Finite Sample Optimal Weight and Estimation Risk)**.** *Define the following quantities*

$$
\begin{aligned}
Q_k &= (\hat{\Sigma}_k + \lambda_k I_p)^{-1}\hat{\Sigma}_k, \\
B &= [Q_1\beta_1, \cdots, Q_K\beta_K], \\
V &= B^T\beta_K = \textit{vec}[\beta_k^T Q_k \beta_K], \\
A &= B^T B = \textit{mat}[\beta_k^T Q_k Q_{k'} \beta_{k'}], \\
R &= \textit{diag}[n_k^{-1}\sigma_k^2 \operatorname{Tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-2}\hat{\Sigma}_k\}].
\end{aligned}
$$

The optimal estimation weight and the corresponding estimation risk are

$$
\begin{aligned}
W_E^* &= (B^T B + R)^{-1} B^T \beta_K = (A + R)^{-1} V, \\
M(W_E^*) &= \beta_K^T \{I_p - B(B^T B + R)^{-1} B^T\}\beta_K = \beta_K^T \beta_K - V^T (A + R)^{-1} V.
\end{aligned}
$$

The results in Theorem 1 holds under finite samples, but as we shall see, their limiting forms are more useful as they depend only on α_k^2 and σ_k^2 instead of the unknown linear coefficients β_k.

3.2 Asymptotic behavior of the weight when $p, n \to \infty$

In this subsection we provide the limiting values for optimal weight and estimation risk in terms of the limiting ESD of $\hat{\Sigma}_k$ in the proportional regime. We present the results only in terms of the empirical quantities that can be easily estimated from the observations. In particular, our results do not involve the PSD H of Σ. Most of the terms in the limits of estimation risk (Theorem 2) can be easily expressed given the prior work (Dobriban and Wager, 2018). The off-diagonal terms in

matrix A involve the trace of the product of sample covariance matrices from different studies,

$$E_{kk'} := \mathrm{tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}/p\}.$$

This term requires novel treatment, and its limiting value of $\mathcal{E}_{kk'}$ is presented as Lemma **??** in Supplemental Materials. Together with this Lemma, we express the limiting optimal estimation weight and limiting estimation risk in the following theorem.

Theorem 2 (Asymptotic Optimal Estimation Weights and Estimation Risk)**.** *Given assumptions 1, 2, 3, and 4 hold, and all eigenvalues of* Σ *are uniformly bounded away from zero and infinity, the limiting optimal estimation weight and risk satisfy*

$$W_E^* \ \rightarrow_{a.s.} \ \mathcal{W}_E^* = (\mathcal{A} + \mathcal{R})^{-1}\mathcal{V}$$
$$M(W_E^*) \ \rightarrow_{a.s.} \ \mathcal{M}(\mathcal{W}_E^*) = \alpha_K^2\sigma_K^2 \ - \ \mathcal{V}^\top(\mathcal{A} + \mathcal{R})^{-1}\mathcal{V}$$

Here $\mathcal{V}$*,* $\mathcal{A}$*, and* $\mathcal{R}$ *are the almost-sure limits of* V*,* A*, and* R*, respectively, and are provided in Appendix A.*

Remark 1. The assumption (1) that all studies have the same population covariance matrix facilitates the derivation of the theoretical limits of the cross-population term $E_{kk'}$. However, this assumption is not needed in real applications. Instead of the theoretical limits, one can use the plug-in estimates with sample covariance matrices to estimate the optimal weight.

A practical, step-by-step procedure to estimate and deploy the optimal estimation weight vector is provided in Appendix B. We show the theoretical and empirical estimation risks with 50 independent simulations in Figure 1 under three different setups, each corresponding to a set up in

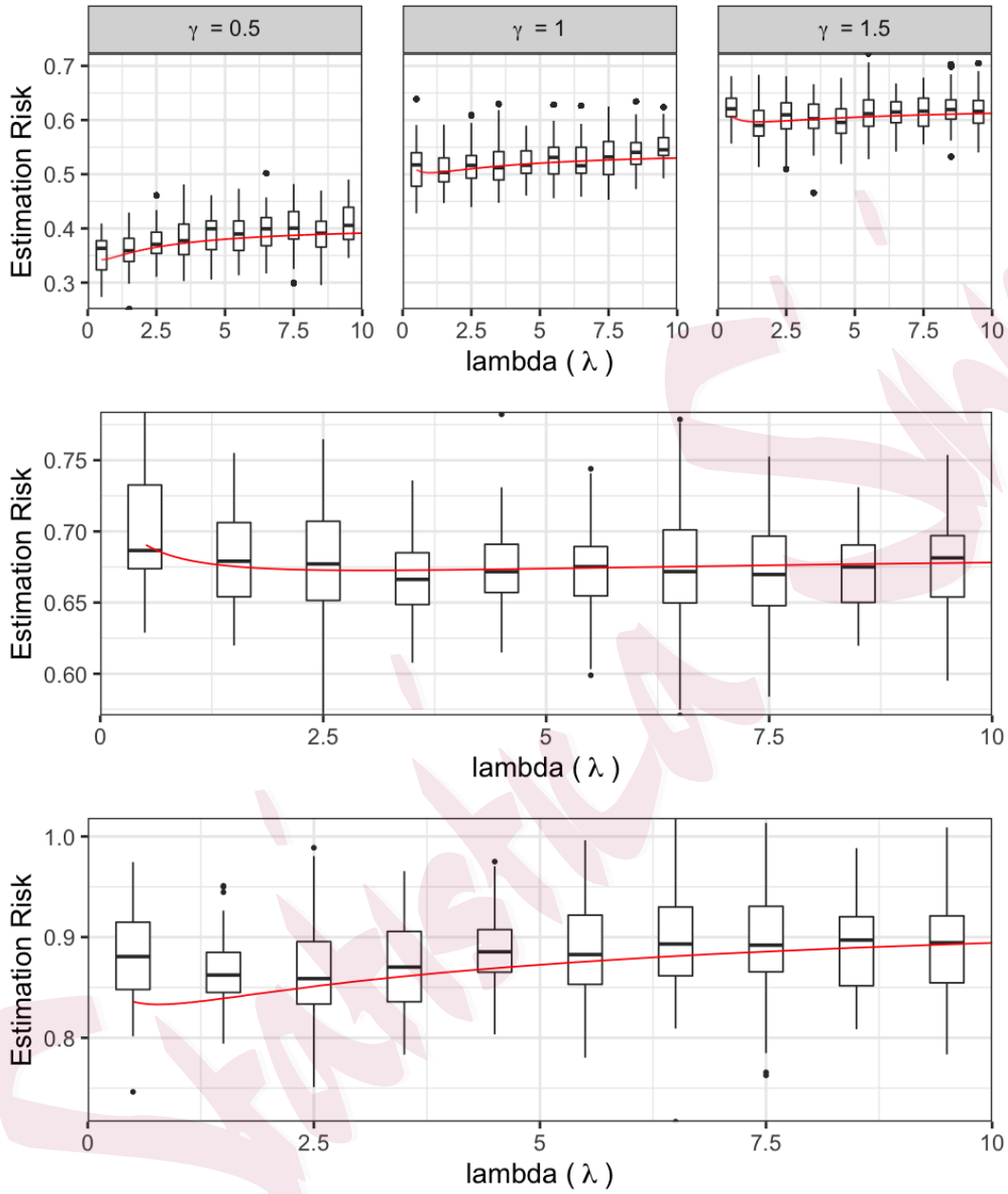

Figure 1: Empirical estimation errors (boxplots) and theoretical risk (solid red) for the ridge estimator as a function of the tuning parameter λ, with aspect ratio $\gamma = p/n$. Top: equal-sample setting with $n = 500$ per study (facets indicate γ). Middle: identity covariance with unequal sample sizes, $p = 750$, target size $n_K = 750$, and source sizes $n_k \in \{700, \ldots, 250\}$. Bottom: the same unequal-sample configuration under a Toeplitz covariance.

Lemma **??**. We fix the random coefficient correlations between all the studies as $\rho = 0.5$. In the top plot, $p = 500$ and the number of observations $n = p/\gamma$. We fix n to be the same for target and five source studies, but we vary γ from 0.5 to 1.5. For the middle and the bottom plots, we set the number of observations in all six studies to decrease from 750 to 250, where the target study has 750 observations and $p = 750$. In addition, the population covariance matrices of the top and the bottom plots share a Toeplitz structure, while the middle plot has an identity covariance matrix. We observe that the limiting estimation risk are very accurate in all three scenarios across a wide range of λ values.

4. Adaptivity of the Weight $\mathcal{W}_E^*$ to the Informativeness of the Source Data

4.1 Adaptivity of $\mathcal{W}_E^*$ to the regression coefficient correlations

The regression coefficient correlations $\rho_{kk'}$ play a crucial role determining the weights assigned to target and source studies. We show in the this section that $\mathcal{W}_E^*$ has the expected adaptability properties. Firstly we present in Corollary 1 below the limiting behavior of $\mathcal{W}_E^*$ when all values of $\rho_{kk'}, k, k' \in \{1, \cdots, K\}, k \neq k'$ approach the boundary values, 0 or 1.

Corollary 1 ($\mathcal{W}_E^*$ Under Boundary ρ_{ij})**.** *Under general Σ, divide the study indices into $\mathcal{I}_1, \mathcal{I}_2$, where $\mathcal{I}_1 = \{k : k \in \{1, \cdots, K-1\}, \rho_{kK} = 0\}$ and $\mathcal{I}_2 = \{1, \cdots, K\} \setminus \mathcal{I}_1$. For any $k \in \mathcal{I}_1$, if $\rho_{kk'} = 0$ for all $k' \in \mathcal{I}_2$, then $(\mathcal{W}_E^*)_k = 0$. In the special case, when $\rho_{kk'} \to 0, k, k' \in \{1, \cdots, K\}, i \neq j$, $\lambda_K = \gamma_K/\alpha_K^2$, we have $W_E^* \to (0, \cdots, 1)^T$. When $\alpha_1^2 = \cdots = \alpha_K^2 = \alpha^2$, $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma^2$ and $\gamma_1 = \cdots = \gamma_K = \gamma$, under $\rho_{kk'} \to 1$ for $k, k' \in \{1, \cdots, K\}$ and $k \neq k'$, all studies receive the same weight with $(\mathcal{W}_E^*)_1 = \cdots = (\mathcal{W}_E^*)_K$.*

Corollary 1 states that when all correlations are zero, the limiting weights assigned to all source ridge estimators are zero, and the limiting weight assigned to the target ridge estimator is one.

In this case the proposed transfer learning estimator becomes identical to the target-only ridge estimator using only the target data, which should be the best that we can do when there is no relevant information in source populations. When all correlations are one, the limiting weights for all ridge estimators are the same, which is expected since the information in all studies is essentially the same.

We also show how the weight $\mathcal{W}_E^*$ changes as a function of ρ. Under $\Sigma = I_p$ and equicorrelation $\rho_{kk'} = \rho$ for $k \neq k'$, define, for any source k, the weight ratio

$$r(\rho) := \frac{(\mathcal{W}_E^*)_K(\rho)}{(\mathcal{W}_E^*)_k(\rho)}.$$

Corollary **??** in the Supplemental Materials shows that $r'(\rho) < 0$ for $\rho \in (0, 1)$, i.e., as ρ increases the weights assigned to the sources increase relative to the target weight. Moreover, since $r(\rho) \geq 1$ for all ρ (Corollary 1), the target weight $(\mathcal{W}_E^*)_K$ is at least as large as each individual source weight $(\mathcal{W}_E^*)_k$ for every ρ.

4.2 Limiting estimation risk under the optimal weight

We next investigate the behavior of the limiting estimation risk using the optimal weight $\mathcal{M}_K^*(\mathcal{W}_E^*)$ under the setting when $\Sigma = I_p$. For simplicity, we assume $\rho_{kk'} = \rho$ for $k, k' = 1, \cdots, K$ and $k \neq k'$,$\lambda_1 = \cdots, \lambda_K = \lambda$, $\alpha_1^2 = \cdots = \alpha_K^2 = \alpha^2$, and without loss of generality, $\sigma_1^2 = \cdots = \sigma_K^2 = 1$. Under these settings, Corollary **??** in Supplemental Materials presents an explicit expression of $\mathcal{M}_K^*(\mathcal{W}_E^*)$. We observe that the limiting estimation risk reduces to the risk of the ridge regression fitted with only the target data when $\rho \to 0$, and it reduces to the risk of the distributed learning ridge (Sheng and Dobriban, 2020) when $\rho \to 1$.

A simple analysis of the upper bound given in Corollary **??** in the Supplemental Materials re-

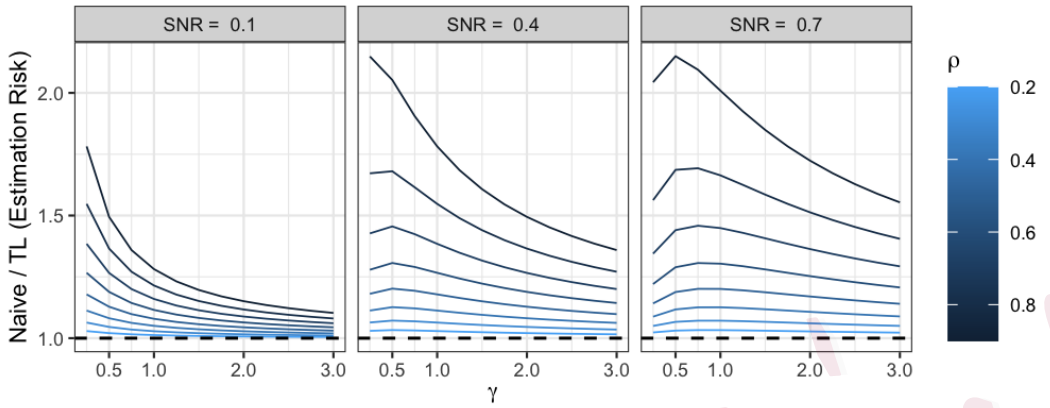

Figure 2: Ratio of the estimation risk of one sample ridge regression limiting over the risk of Trans-Ridge as a function of the strength of correlation ρ_{iK}, denoted by the colors, and the aspect ratio $\gamma = p/n$ as $n, p \to \infty$.

veals that it strictly decreases with every ρ_{iK} for $i \in \{1, \cdots, K-1\}$ under very mild conditions on the magnitude of ρ_{iK}. The upper bound in this corollary arises from bounding the eigenvalues of the correlation matrix of linear coefficients in each study via $a^* = \sup_{i\in\{1,\cdots,K\}} \sum_{k=1,k\neq i}^{K} |\rho_{ik}|$. To demonstrate this point, Figure 2 shows the ratio of the limiting risk of the target-only ridge estimator under optimal tuning and $\mathcal{M}^*(\mathcal{W}_E^*)$ as a function of γ and ρ. The ratio monotonically increases as ρ increases, suggesting greater improvement when coefficient correlations are larger. We also observe that if SNR is small, the improvement brought by Trans-Ridge estimator monotonically increases as γ gets larger. This monotone relationship does not hold when SNR is large enough, instead it rises first, but then goes down.

5. Optimal Weight that Minimizes the Prediction Risk

5.1 Optimal weight that minimizes the prediction risk

An alternative to estimating the weight vector is by minimizing the prediction risk, defined as

$$r(W) = E_{\epsilon_1,\cdots,\epsilon_K,x_0,\epsilon_0}\left((y_0 - (\sum_{k=1}^{K} W_k\hat{\beta}_k)^T x_0)^2 | X_1,\cdots,X_K\right),$$

where (x_0, y_0) is an observation sampled from the target population such that $y_0 = x_0^T\beta_K + \epsilon_0$. The expectation in the risk is taken over the independent random test example (x_0, y_0), and over the randomness in β_K and $(\epsilon_1,\cdots,\epsilon_K)$. Similar to the estimation risk, we can find the weight that minimizes the prediction risk, and denote this optimal prediction weight as W_P^*. We hide the dependence of the prediction risk on the error terms and design matrices from now on for simplicity. The following theorem gives the finite-sample expressions for optimal prediction weight and the corresponding prediction risk.

Theorem 3 (Finite sample optimal weight for prediction)**.** *Define the following matrices*

$$B = \left(E(\hat{\beta}_1) \quad \cdots \quad E(\hat{\beta}_K)\right) = \left(Q_1\beta_1 \quad \cdots \quad Q_K\beta_K\right),$$

$$D = B^T\Sigma\beta_K = \textit{vec}[\beta_k^T Q_k\Sigma\beta_K],$$

$$C = \textit{mat}[E_\epsilon(\beta_k^T Q_k\Sigma Q_{k'}\beta_{k'})],$$

$$F = \textit{diag}[\sigma_k^2\,\mathrm{Tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma\}/n_k].$$

The optimal prediction weight and the corresponding prediction risk are

$$W_P^* = (C + F)^{-1} D,$$
$$r(W_P^*) = \sigma_K^2 + \beta_K^T \Sigma \beta_K - D^T (C + F)^{-1} D.$$

Similarly to the estimation risk, the limits of D, F and diagonal entries of C can be expressed in empirical terms with existing techniques (Dobriban and Wager, 2018). As before, the off-diagonal entries of C, like A_{ij} in Theorem 1, is an interplay between the spectrum of $\hat{\Sigma}_k$ and $\hat{\Sigma}_{k'}$. Further, since the prediction error is averaged over the randomness of x_0, it also involves the spectrum of population covariance matrix Σ through

$$P_{kk'} := \mathrm{tr}\{\Sigma(\hat{\Sigma}_k + \lambda_k)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'})^{-1}\}/p.$$

We present its limits $\mathcal{P}_{kk'}$ such that $P_{kk'} \to_{a.s.} \mathcal{P}_{kk'}$ in lemma **??** in Supplemental Materials, and show it can be broken into pieces of functions of $m_{F_{\gamma_K}}(-\lambda_k)$. Applying this lemma leads to the limiting prediction risk and optimal prediction weight given in Theorem 4 below.

Theorem 4 (Asymptotic Optimal Prediction Weights and Prediction Risk)**.** *Given assumptions 1, 2, 3, and 4 hold, and all eigenvalues of* Σ *are uniformly bounded away from zero and infinity, the limiting optimal prediction weight and risk satisfy*

$$W_P^* \to_{a.s.} \mathcal{W}_P^* = (\mathcal{C} + \mathcal{F})^{-1} \mathcal{D}$$
$$r(W_P^*) \to_{a.s.} \mathfrak{r}(\mathcal{W}_P^*) = \sigma_K^2 + \alpha_K^2 \sigma_K^2 - \mathcal{D}^T (\mathcal{C} + \mathcal{F})^{-1} \mathcal{D}.$$

Here $\mathcal{D}$, $\mathcal{C}$, *and* $\mathcal{F}$ *are the almost-sure limits of* D, C, *and* F, *respectively, and are provided in*

Appendix A.

Remark 2. The assumption (1) that all studies have the same population covariance matrix is again not needed. Suppose the observations $(X_k)_i$ has population covariance Σ_k that meets assumption (4) for all these studies $k = 1, \cdots, K$. This facilitates the consistent estimations for the limits of the cross-population terms, which would appear in Theorem 4 under a heterogeneous covariance set up,

$$P_{kk'} = \mathrm{tr}\{\Sigma_K(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}\}/p,$$
$$d_{kK} := \mathrm{tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\Sigma_K\}/p.$$

The remaining terms in Theorem 4 can be evaluated in the same way. These two terms appear in the optimal prediction weight ($C_{kk'}$, D_k respectively) and involve population covariance matrix Σ_K. One can prove Σ_k is deterministically equivalent (Hachem et al., 2007) to $\hat{\Sigma}_k$ for $k = 1, \cdots, K$ in the sense that for an independent matrix $B \in R^{p\times p}$ with bounded operator norm, we have $\mathrm{tr}(B(\hat{\Sigma}_k - \Sigma_k))/p \to_{a.s.} 0$. This leads to

$$\hat{P}_{kk'} := \mathrm{tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}\hat{\Sigma}_K\}/p \to_{a.s.} P_{kk'},$$
$$\hat{d}_{kK} := \mathrm{tr}\{(\hat{\Sigma}_k + \lambda_k I_p)^{-1}\hat{\Sigma}_K\}/p \to_{a.s.} d_{kK}.$$

For $k = K$, one can decompose $\hat{\mathcal{P}}_{kk'}$ into the terms from study K and study k and evaluate them separately (see proofs of lemma **??**).

See Appendix B for a step-by-step pipeline to estimate and apply the optimal estimation weights. Figure 3 depicts the limiting prediction risk across the three scenarios stipulated by

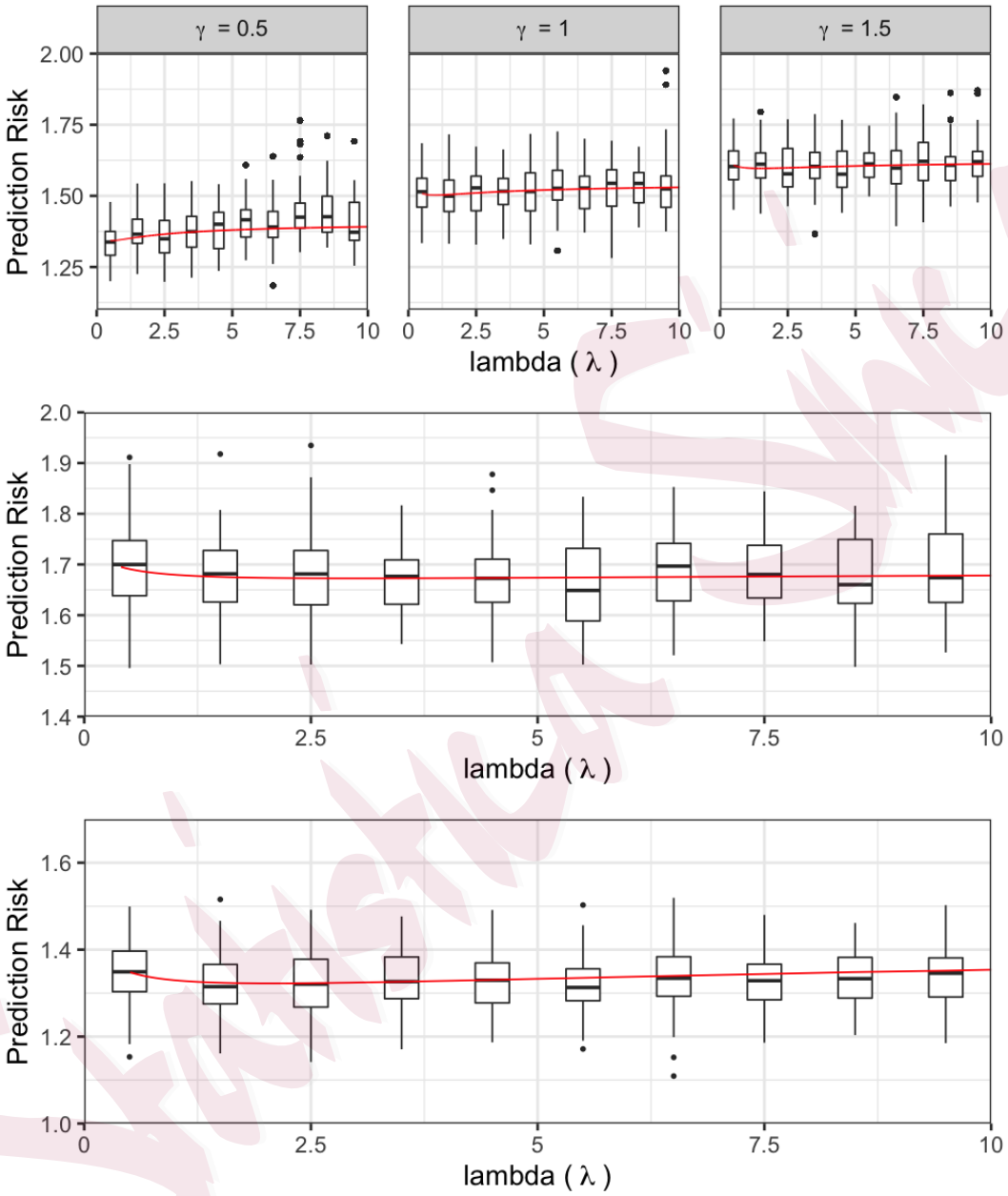

Figure 3: Comparison between theoretical prediction risk and averaged prediction errors. Top: all data sources have the same sample size $n = 750, p = 750$. Middle: The second scenario where each study have different sample sizes under identity population covariance matrix, $n_K = 750$, the source sample sizes range from 700 to 250. Bottom: The same setup as the middle plot, but with a Toeplitz population covariance matrix.

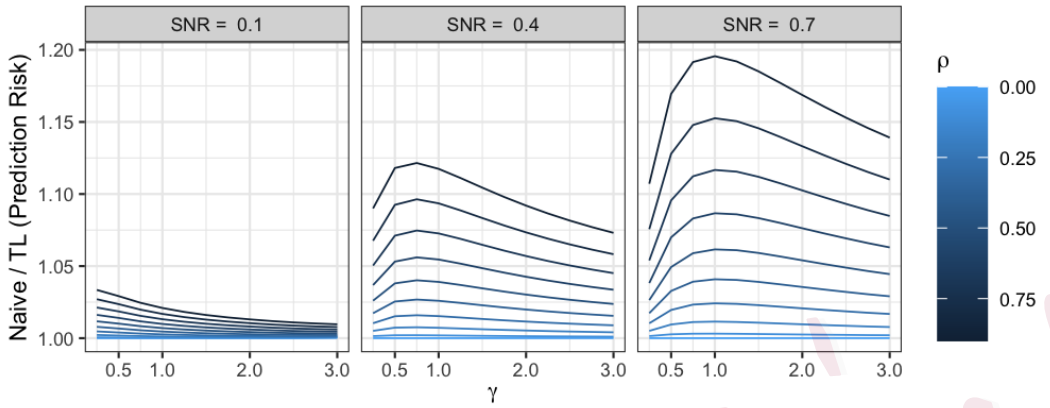

Figure 4: Ratio of the limiting prediction risk of one sample ridge regression over the risk of Trans-Ridge as a function of the strength of correlation ρ_{iK} denoted by the colors, and the aspect ratio $\gamma = p/n$.

Lemma **??**, and the simulation designs match those in Figure 1. The plots show that the limiting prediction risks agree well with the empirically averaged prediction errors in all scenarios.

Figure 4 shows plots of the ratio of the limiting prediction risk from Trans-Ridge and ridge regression using only the target data. First, the improvement on the limiting prediction risk is not as large as the improvement of the estimation risk. Even when the study correlation is as high as 0.9, the magnitude of improvement is at most 20%. Second, like the limiting estimation risk, when no source study is informative ($\rho = 0$), the limiting risk of Trans-Ridge estimator with prediction weight is the same as the target-only ridge estimator. A similar result to Corollary 1 for optimal prediction weight can be easily derived as the expressions of $\mathcal{W}_P^*$ and $\mathcal{W}_E^*$ share the same form. Third, larger ρ leads to a larger improvement in prediction from Trans-Ridge compared with the target-only ridge estimator. Finally, the regime of γ where Trans-Ridge leads to the largest improvement increases with SNR. When SNR is large, the largest improvement is observed for some particular value of γ.

6. Robustness of the Weight $\mathcal{W}_E^*$ to Covariate Shift

Trans-Ridge with the weights that minimize the prediction risk result in smaller prediction error on a testing data point x_0 if the test data are from the same distribution as the target data. However, under the covariate shift, the optimal estimation weight is expected to be more robust to covariate shift since the weight is chosen to minimize the error of estimating the linear coefficients. We demonstrate that minimizing estimation risk is equivalent to minimizing prediction risk computed under a covariate-shift–robust loss, yielding the same optimal weight.

Consider the scenario when the testing data x_0 in the target population has a different distribution from the training data specified by the assumption (1). We can no longer reliably estimate the prediction risk and the optimal prediction weight, as the distribution of x_0 is unknown. We can instead employ a "min-max" strategy to find the optimal weight. Define a family of distributions $\mathcal{P} := \{P : x \sim P, E_P(||x||_2) \leq c\}$ and consider the following problem

$$\min_{W} \max_{x_0 \in \mathcal{P}} E_{\epsilon, x_0}(||(\sum_{k=1}^{K} W_k \hat{\beta}_k - \beta_K)^T x_0||^2). \tag{6.1}$$

The solution of (6.1) finds a weight vector that minimizes the worst case prediction risk when the testing data x_0 has a different distribution. This type of distributional robust problems have been considered by Duchi and Namkoong (2021) for learning models with uniform performance. Let $e(W) := \sum_{k=1}^{K} W_k \hat{\beta}_k - \beta_K$ be the "estimation difference". For any fixed W, the inner maximization in (6.1) over distributions P satisfying $E_P \|x_0\|_2 = c$ reduces to

$$\sup_{P:\, E_P \|x_0\|_2 = c} E_P\big[(x_0^\top e(W))^2\big].$$

By Cauchy–Schwarz, $(x_0^\top e(W))^2 \le \|x_0\|_2^2 \, \|e(W)\|_2^2$, and the supremum under the moment budget $E_P\|x_0\|_2 = c$ is attained by concentrating x_0 in the direction of $e(W)$ with fixed norm c. Hence

$$\sup_{P:\, E_P\|x_0\|_2 = c} E_P\big[(x_0^\top e(W))^2\big] = c^2 \, \|e(W)\|_2^2, \quad \text{with maximizer } \; x_0^* = c\,\frac{e(W)}{\|e(W)\|_2} \text{ a.s.}$$

Crucially, the factor c^2 is multiplicative and does not depend on W. Therefore the outer minimization is invariant to c and reduces to the estimation–risk problem

$$\min_W \; E_\epsilon \, \|e(W)\|_2^2.$$

Equivalently, the estimation-risk minimization yields the same optimal weight as the covariate-shift–robust prediction criterion with no knowledge of c required. This finding is also consistent with existing literatures. Shimodaira (2000); Ge et al. (2023) show that the target-risk–optimal parameter is invariant to shifts in the marginal covariate distribution under correctly specified parametric models. The presented robustness is consistent with this view: the estimator and its weights depend on hyperparameters and cross-moment quantities, not on the law of the test covariates.

To demonstrate this robustness, Figure 5 shows the simulation results based on 50 replications. Here we choose $p = 750$ and $n_k = 500$ for all 5 source studies. The pairwise correlations of the regression coefficients across the studies are fixed at 0.5. We use a Toeplitz covariance matrix in this demonstration. The prediction error using the optimal prediction weight when testing data x_0 has the same distribution as the training data points, is shown by the light blue boxplots, resulting in the smallest prediction errors when all the assumptions hold.

We then consider two types of covariate shifts. In the first subplot, there is a mean-shift of x_0 so that all the coordinates of x_0 has a mean 1 but with the same population covariance matrix.

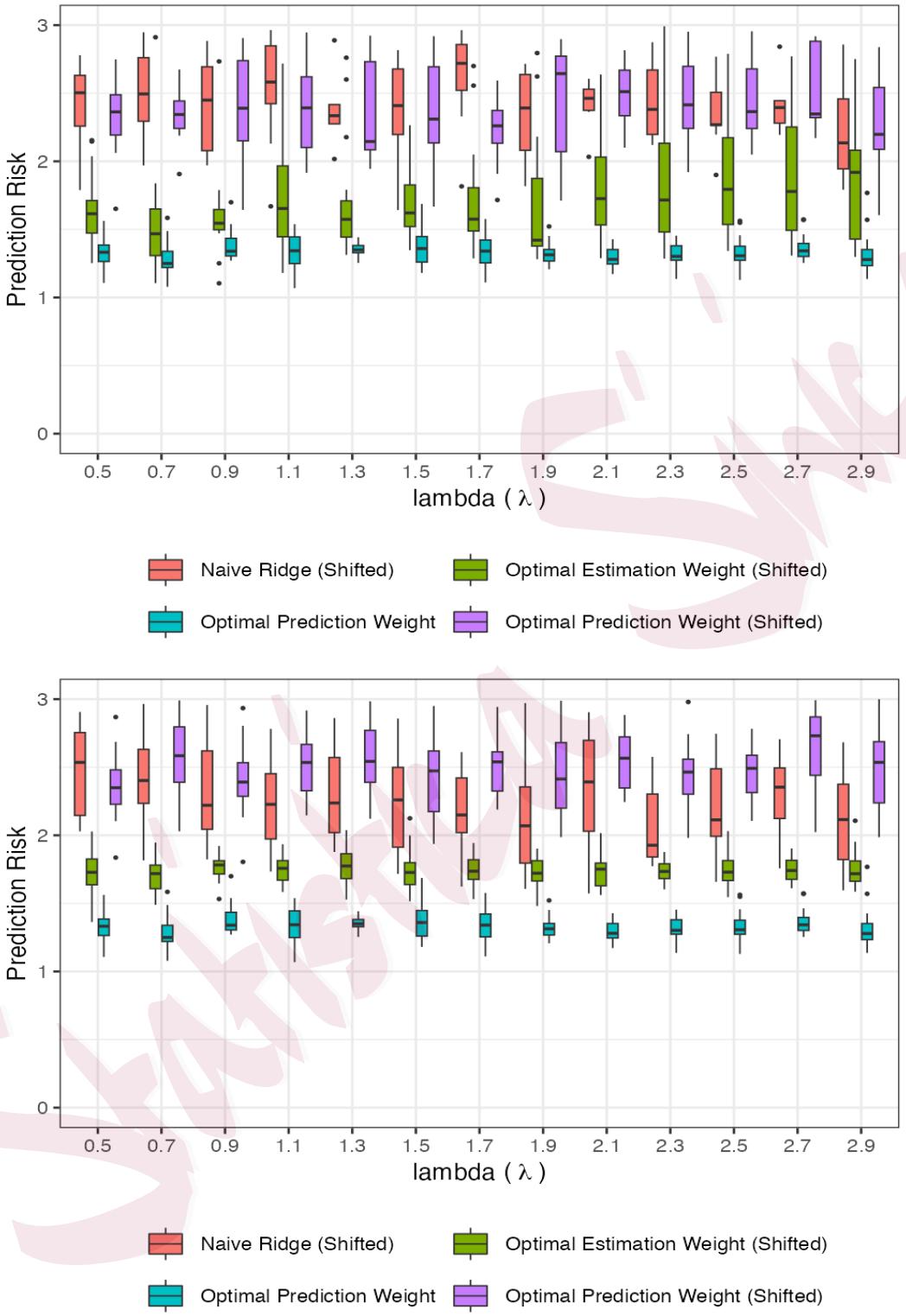

Figure 5: The top plots shows the performances of different models when there is a mean-shift of x_0 while the bottom shows the performances when there is a covariance-shift. In both examples, the Trans-Ridge estimator with optimal estimation weight outperforms the naive ridge estimator, showing the effectiveness of Trans-Ridge when there is a covariate shift.

Table 1: Analysis of lipid trait data set. The testing data mean square error is shown, where each trait is a target, and the other two traits are source traits. The null prediction errors is 1 for all traits as they were standardized. The number of SNPS used are $p = 2000, 3000, 4000$ and 5000. OptEst: Trans-Ridge with estimation-based weight; OptPred: Trans-Ridge with prediction-based weight; Naive: ridge regression using only target data; MTAG: method of Turley et al. (2018).

	OptEst	OptPred	Naive	Trans-Lasso	Lasso	MTAG
p				HDL		
2000	0.885	0.903	0.960	**0.870**	0.917	0.873
3000	0.878	0.905	1.039	**0.869**	0.917	0.869
4000	**0.854**	0.867	0.986	0.867	0.917	0.873
5000	**0.832**	0.835	0.951	0.863	0.917	0.873
				LDL		
2000	0.959	**0.948**	0.960	0.965	1.003	0.982
3000	**0.935**	0.947	0.948	0.962	1.004	0.984
4000	**0.933**	0.934	0.935	0.962	1.004	0.983
5000	**0.915**	0.922	0.923	0.962	1.004	0.986
				TRI		
2000	0.915	**0.906**	0.914	0.914	0.987	0.953
3000	0.917	**0.910**	0.917	0.951	0.990	0.953
4000	**0.891**	0.904	0.905	0.952	0.990	0.954
5000	0.924	**0.909**	0.934	0.990	0.956	0.954

In the second subplot, we weaken the correlation strength among the coordinates of x_0. In both types of the covariate shifts, the performance of Trans-Ridge with the optimal prediction weight drops significantly (purple boxplots). However Trans-Ridge with optimal estimation weight has a much smaller prediction error, and importantly, its performance is better than the target-only ridge estimator. These results show that the proposed Trans-Ridge estimator with the optimal estimation weight is still useful when the testing data has a shifted distribution.

7. Application to Real Data Sets

7.1 Lipid traits prediction using genotype data

We apply the proposed transfer learning ridge estimator using Penn medicine biobank (PMBB), a large biobank of genomic data linked to electronic health record phenotype data. Three genetically linked lipid phenotypes that are generally predictable by SNPs are considered, including high density lipoprotein cholesterol (HDL), low density lipoprotein cholesterol (LDL) and triglycerides (TRI). We consider when source datasets of related traits are available for the target trait. We use each of three traits as target trait and the other two traits as the source traits. We divide the approximately twelve thousand observations into three groups, one for each trait. We perform univariate filtering using only target training data, and the most significant 2000, 3000, 4000 and 5000 genetic variants are used as the predictors. The penalization parameters are selected via a 4-fold cross-validation on target training data. The hyperparameters $(\alpha_k^2, \sigma_k^2, \rho_{kk'})$ are estimated separately within each study using the consistent estimators of Lee et al. (2012) developed for genetic studies. For the target population, estimation is performed using only the training fold.

The held-out test mean-square errors are shown in Table 1. In all the experiments, Trans-Ridge improves the prediction accuracy over the ridge regression using only the target data. The gain in prediction accuracy for HDL is the largest by transferring information from the other two traits. The estimated genetic correlation between HDL and LDL and between HDL and TRI are 0.293 and 0.730 for $p = 4000$, and 0.308 and 0.679 for $p = 5000$. Across all settings, Trans-Ridge dominates target-only ridge. Moreover, Trans-Ridge with the optimal estimation weights sometimes yields better prediction performance than using the optimal prediction weights. We attribute these cases to the robustness of the estimation-weight solution established in Section 6: it is less sensitive to covariate-shift induced by heterogeneity in the feature covariance across subpopulations. In our

PMBB analysis (which includes both Black and White participants), differences in gene-expression covariance across ancestry groups can make the training–test distributions non-identical; under such shift, the optimal estimation weight can outperform.

For comparison, we also applied multi-trait analysis of GWAS summary statistics via MTAG (Turley et al., 2018) and the transfer-learning Lasso regression (Trans-Lasso) of Li et al. (2022). MTAG underperforms TRANS-RIDGE in prediction for all three lipid traits; even when using all 650,281 SNPs, the mean squared errors are 0.909, 0.960, and 0.957 for HDL, LDL, and TRI, respectively. TRANS-LASSO assumes sparsity of genetic effects for each trait (as in standard Lasso) and that cross-trait differences are even sparser. It can be competitive when p is small, but is dominated as p grows. This is commonly observed in PRS applications (Marotta et al., 2021).

7.2 Colorectal cancer prediction using microbiome data

We also demonstrate the advantage of Trans-Ridge for risk prediction using an open-source case-control microbiome study of colorectal cancer (CRC). The predictors are the genera and phyla of the microbiome, along with three demographic factors including age, gender, and BMI (Duvallet et al., 2017). We focus on transfer learning among the three studies, referred to as Zackular, Zeller, and Baxter with sample size of $n = 83, 127, 488$, respectively. These studies include subjects from the USA/Canada, France, and the USA only, and the subjects may have different microbiome profiles due to different diets and lifestyles. Some bacteria are rare among all subjects and are likely not predictive. We remove all bacteria with more than 90% zeros in the training data, retaining roughly 150 microbiome features. The three demographic features are always included. We convert the binary CRC status to -1 and 1 and treat the CRC prediction as a regression problem. The hyperparameters $\sigma_k, \alpha_k, \rho_{kk'}$ used by Trans-Ridge are estimated via Bayesian regression with

Table 2: CRC prediction based on gut microbiome data. The LOOCV AUC is shown. OptEst: Trans-Ridge with estimation-based weight; OptPred: Trans-Ridge with prediction-based weight; Naive: ridge regression fitted with only target data; NAIVEPOOL: ridge regression fitted with the combined data of three studies; nObs: Number of observations

	OptEst	OptPred	Naive	NaivePool	nObs
ZACKULAR	0.845	**0.906**	0.546	0.757	83
ZELLER	0.826	**0.928**	0.727	0.754	127
BAXTER	0.788	**0.927**	0.723	0.708	488

suitable priors (see Supplementary Section **??** for details).

We perform CRC prediction for each study as the target study while the other two studies are treated as source studies. Due to small sample sizes and imbalanced classes (20 % v.s. 80 %), we use a leave-one-out cross-validation (LOOCV) to assess the area under the curve (AUC) metrics of the proposed method and competing methods. For each target data point, we fit a model without it and record its predicted score. At the end, the AUC is computed over all the left-out samples.

The LOOCV AUC is presented in Table 2. Since all studies share the same CRC prediction problem, we also consider fitting a ridge regression on the combined data set. We observe that Trans-Ridge with optimal prediction weight outperforms all other methods across all three studies. This is expected, as it explicitly minimizes the prediction risk rather than the estimations risk. It naturally dominates when model misspecification is limited. Additionally, Trans-Ridge with the optimal estimation weight also outperforms the naive ridge regressions, consistent with our theoretical results. These results again highlight the advantages of employing a transfer learning framework in disease prediction using source data sets.

8. Discussion

We have introduced a flexible transfer learning framework for high dimensional ridge regression, in which the target coefficients are estimated as an optimally weighted blend of source and target ridge

fits. By minimizing either estimation risk or prediction risk, our two proposed weighting schemes adapt to both in-distribution prediction and out-of-distribution robustness. Simulation studies and real data applications (e.g. lipid trait prediction, microbiome based cancer risk) demonstrate that the prediction risk weight excels when train/test covariates align, while the estimation risk weight yields more stable performance under covariate shift.

From a theoretical standpoint, we leverage random matrix asymptotics under general covariance to derive explicit expressions for both types of optimal weights in the high-dimensional limit $p/n \to \gamma$. A key technical ingredient is two new deterministic-equivalent results: Lemma **??** and Lemma **??** establish almost-sure limits for $E_{kk'}$ and $P_{kk'}$ respectively. These limits explicitly couple different studies and remain valid with heterogeneous aspect ratios (γ_k), penalties (λ_k), and study-specific Σ_k. In contrast, Sheng and Dobriban (2020) derive risk formulas only in the isotropic case ($\Sigma = I$), which do not readily extend to multiple populations with heterogeneous n_k or λ_k, nor to cross-study resolvent products.

We assume the hyperparameters α_k and $\rho_{kk'}$ are known in our analysis, as consistent estimators are well established in genetic studies. In other settings, because the hyperparameters are directly interpretable, users can straightforwardly tailor Bayesian specification for their applications. Trans-Ridge is also robust to moderate estimation error in these hyperparameters: they enter the optimal weights only through the bounded matrix $\Sigma_\delta = \mathrm{mat}[\rho_{kk'}\alpha_k\alpha_{k'}]$, and the weight map is a smooth, ridge-regularized function of Σ_δ. Consequently, small perturbations to $\{\alpha_k, \rho_{kk'}\}$ induce only small changes in the weights and the resulting risk.

Trans-Ridge has the same leading computational cost as fitting K separate ridge models and then solving a $K \times K$ system. Using naive linear algebra, forming $\hat{\Sigma}_k = X_k^\top X_k/n_k$ costs $O(n_k p^2)$ per study and inverting $(\hat{\Sigma}_k + \lambda_k I_p)$ costs $O(p^3)$; assembling the optimal weight requires solving

at most two linear systems and one eigenproblem of size K, an $O(K^3)$ overhead that is negligible when $K \ll p$. Moreover, because each inversion has a ridge form, one may apply Woodbury and work in the dual $n_k \times n_k$ space with a Cholesky factorization, reducing the per–study cost to $O(n_k^3 + n_k p)$ when $n_k \ll p$.

Trans-Ridge can also be applied using per-study summary statistics together with estimates of the sample covariance matrices $\hat{\Sigma}_k$. In genetic applications, such summary statistics can be used to estimate (α_k, σ_k) and ridge estimate $\hat{\beta}_k$ as well as the cross-study coefficient correlations $\rho_{kk'}$. Moreover, the cross-study couplings involved in (A, R, V) for estimation, and (C, F, D) for prediction, simplify in the high-dimensional regime: they admit deterministic equivalents that depend only on a small number of scalar trace functionals, namely, $m_k := p^{-1} \operatorname{tr}(\hat{\Sigma}_k + \lambda_k I_p)^{-1}$, $m_k' := p^{-1} \operatorname{tr}(\hat{\Sigma}_k + \lambda_k I_p)^{-2}$, and, for prediction, the analogous quantities v_k and v_k'. When the covariance matrices Σ_k are heterogeneous across studies, standard deterministic-equivalence results justify approximating these quantities via plug-in estimates based on $\hat{\Sigma}_k$.

Supplementary Material

Supplementary materials available online include details of additional lemmas and corollaries, the proofs of all the lemmas, corollaries and theorems, and parameter estimation for real data analysis.

Acknowledgments

We would like to thank Dr. Jiaoyang Huang and Dr. Edgar Dobriban for discussions on random matrix theorems in the derivations. H.L.'s research is supported partially by NIH grants GM123056 and GM129781.

References

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263.

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

Duchi, J. C. and H. Namkoong (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics 49*(3), 1378–1406.

Duvallet, C., S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm (2017). Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature communications 8*(1), 1784.

Faquih, T., A. van Hylckama Vlieg, P. Surendran, A. S. Butterworth, R. Li-Gao, R. de Mutsert, F. R. Rosendaa, R. Noordam, D. van Heemst, K. W. van Dijk, and D. O. Mook-Kanamori (2023). Robust metabolomic age prediction based on a wide selection of metabolites. *medRxiv*.

Ge, J., S. Tang, J. Fan, C. Ma, and C. Jin (2023). Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*.

Hachem, W., P. Loubaton, and J. Najim (2007). Deterministic equivalents for certain functionals of large random matrices.

Hu, Y., M. Li, Q. Lu, et al. (2019). A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics 51*(3), 568–576.

Lee, S., J. Yang, M. Goddard, P. Visscher, and N. Wray (2012). Estimation of pleiotropy between

complex diseases using snp-derived genomic relationships and restricted maximum likelihood. *Bioinformatics 28*(19), 2540–2542.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation, and minimax optimality. *Journal of Royal Statistical Society, series B*.

Marotta, F., R. Mozafari, E. Grassi, A. Lussana, E. Mariella, and P. Provero (2021). Prediction of gene expression from regulatory sequence composition enhances transcriptome-wide association studies. *bioRxiv*.

Márquez-Luna, C., P.-R. Loh, S. A. T. . D. S. Consortium, S. T. . D. Consortium, and A. L. Price (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology 41*(8), 811–823.

Mei, S., W. Fei, and S. Zhou (2011). Gene ontology based transfer learning for protein subcellular localization. *BMC bioinformatics 12*, 44.

Pan, W. and Q. Yang (2013). Transfer learning in heterogeneous collaborative filtering domains. *Artificial intelligence 197*, 39–55.

Rothschild, D., S. Leviatan, A. Hanemann, Y. Cohen, O. Weissbrod, and S. E (2022). An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PLoS ONE 17(3)*, e0265756.

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

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference 90*(2), 227–244.

Shin, H.-C., H. R. Roth, M. Gao, et al. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging 35*(5), 1285–1298.

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

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

Turki, T., Z. Wei, and J. T. Wang (2017). Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access 5*, 7381–7393.

Turley, P., R. Walters, O. Maghzian, A. Okbay, J. Lee, M. Fontana, T. Nguyen-Viet, R. Wedow, M. Zacher, N. Furlotte, 23andMe Research Team, S. S. G. A. Consortium, P. Magnusson, S. Oskarsson, M. Johannesson, P. Visscher, D. Laibson, D. Cesarini, B. Neale, and D. Benjamin (2018). Multi-trait analysis of genome-wide association summary statistics using mtag. *Nat Genet. 50*, 229–237.

Wang, S., X. Shi, M. Wu, and S. Ma (2019). Horizontal and vertical integrative analysis methods for mental disorders omics data. *Scientific Reports*, 1–12.

Zhao, B. and H. Zhu (2019). Cross-trait prediction accuracy of high-dimensional ridge-type estimators in genome-wide association studies. *arXiv preprint arXiv:1911.10142*.

Zhao, Z., L. G. Fritsche, J. A. Smith, B. Mukherjee, and S. Lee (2022). The construction of cross-population polygenic risk scores using transfer learning. *The American Journal of Human Genetics 109*(11), 1998–2008.

Zhou, X., H. Im, and S. Lee (2020). Core greml for estimating covariance between random effects in linear mixed models for complex trait analyses. *Nature Communication 11*, 4208.

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

E-mail: Zhang, Hongzhe, Hongzhe.Zhang@Pennmedicine.upenn.edu;

Li, Hongzhe,hongzhe@upenn.edu.

A. Limiting values for Theorems 4 and 2

Limiting values for $V, A,$ and R in Theorem 2

1. $V \xrightarrow{a.s.} \mathcal{V}$, where for $k = 1, \ldots, K-1$,

$$\mathcal{V}_k \;=\; \rho_{kK}\,\sigma_k\sigma_K\,\alpha_k\alpha_K\left\{1 - \lambda_k\, m_{F_{\gamma_k}}(-\lambda_k)\right\},$$

and

$$\mathcal{V}_K \;=\; \sigma_K^2\alpha_K^2\left\{1 - \lambda_K\, m_{F_{\gamma_K}}(-\lambda_K)\right\}.$$

2. $A \xrightarrow{a.s.} \mathcal{A}$, where for $k = 1, \ldots, K$,

$$\mathcal{A}_{kk} \;=\; \sigma_k^2\alpha_k^2\Big\{1 - 2\lambda_k\, m_{F_{\gamma_k}}(-\lambda_k) + \lambda_k^2\, m'_{F_{\gamma_k}}(-\lambda_k)\Big\},$$

and for $k \neq k'$

$$\mathcal{A}_{kk'} \;=\; \rho_{kk'}\,\sigma_k\sigma_{k'}\alpha_k\alpha_{k'}\Big\{1 - \lambda_k\, m_{F_{\gamma_k}}(-\lambda_k) - \lambda_{k'}\, m_{F_{\gamma_{k'}}}(-\lambda_{k'}) + \lambda_k\lambda_{k'}\,\mathcal{E}_{kk'}\Big\}.$$

3. $R \xrightarrow{a.s.} \mathcal{R}$, where for $k = 1, \ldots, K$,

$$\mathcal{R}_{kk} \;=\; \sigma_k^2\,\gamma_k\Big\{m_{F_{\gamma_k}}(-\lambda_k) - \lambda_k\, m'_{F_{\gamma_k}}(-\lambda_k)\Big\}.$$

Limiting values for $D, C,$ and F in Theorem 4

The limiting values are $D, C,$ and F in Theorem 4 and are given as:

1. $D \xrightarrow{a.s.} \mathcal{D}$, where for $k = 1, \ldots, K-1$,

$$\mathcal{D}_k \;=\; \rho_{kK}\,\sigma_k\sigma_K\,\alpha_k\alpha_K\left[1 - \frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)}\right\}\right],$$

and

$$\mathcal{D}_K \;=\; \sigma_K^2\alpha_K^2\left[1 - \frac{\lambda_K}{\gamma_K}\left\{\frac{1}{\lambda_K\, v_{F_{\gamma_K}}(-\lambda_K)}\right\}\right].$$

2. $C \xrightarrow{a.s.} \mathcal{C}$, where for $k = 1, \ldots, K$,

$$\mathcal{C}_{kk} \;=\; \sigma_k^2\alpha_k^2\left[1 - 2\frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)}\right\} + \frac{\lambda_k^2}{\gamma_k}\,\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k\, v'_{F_{\gamma_k}}(-\lambda_k)}{[\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)]^2}\right],$$

and for $k \neq k'$,

$$\mathcal{C}_{kk'} \;=\; \rho_{kk'}\,\alpha_k\sigma_k\,\alpha_{k'}\sigma_{k'}\left[1 - \frac{\lambda_k}{\gamma_k}\left\{\frac{1}{\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)}\right\} - \frac{\lambda_{k'}}{\gamma_{k'}}\left\{\frac{1}{\lambda_{k'}\, v_{F_{\gamma_{k'}}}(-\lambda_{k'})}\right\} + \lambda_k\lambda_{k'}\,\mathcal{P}_{kk'}\right].$$

3. F is diagonal, and $F \xrightarrow{a.s.} \mathcal{F}$, where

$$\mathcal{F}_{kk} \;=\; \sigma_k^2\gamma_k\left[\frac{1}{\gamma_k}\left\{\frac{1}{\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)} - 1\right\} - \lambda_k\frac{1}{\gamma_k}\,\frac{v_{F_{\gamma_k}}(-\lambda_k) - \lambda_k\, v'_{F_{\gamma_k}}(-\lambda_k)}{[\lambda_k\, v_{F_{\gamma_k}}(-\lambda_k)]^2}\right].$$

B. Algorithm to Implent the Trans-Ridge estimators

This section provides a step-by-step recipe that takes multi-study datasets to the final Trans-Ridge coefficients. Algorithm 1 is the top-level driver: it standardizes the inputs, calls Algorithm 2 to estimate all plug-in quantities and the per-study ridge fits, and then invokes Algorithm 3 to plug these estimates into the theoretically optimal weights from Theorem 2 (estimation risk) and Theorem 4

Algorithm 1 Trans-Ridge for study K (unequal n_k, shared Σ)

Require: Data $\{(Y_k, X_k)\}_{k=1}^K$ with target index K; candidate penalties $\{\lambda_k\}_{k=1}^K$.
Ensure: Transfer learning estimators $\widehat{\beta}_K^{\text{OptEst}}$ and $\widehat{\beta}_K^{\text{OptPred}}$.

1: Within each study, standardize Y_k and columns of X_k to mean 0, variance 1 (Assumption 1).
2: Call Alg. 2 to estimate study-level quantities and ridge fits $\{\widehat{\beta}_k\}_{k=1}^K$.
3: Call Alg. 3 to compute optimal weights $\widehat{W}_E^*$, $\widehat{W}_P^*$.
4: Assemble the transfer learning estimators $\widehat{\beta}_K^{\text{OptEst}} = \sum_{k=1}^K (\widehat{W}_E^*)_k \widehat{\beta}_k$ and $\widehat{\beta}_K^{\text{OptPred}} = \sum_{k=1}^K (\widehat{W}_P^*)_k \widehat{\beta}_k$.

(prediction risk), yielding $\widehat{\beta}_K^{\text{OptEst}}$ and $\widehat{\beta}_K^{\text{OptPred}}$. Unless noted otherwise, we allow unequal sample sizes and assume a shared covariance; practical tuning and variants for equal n_k or heterogeneous Σ_k are summarized in the subsequent remarks.

Remark 3 (Tuning penalties λ_k)**.** To practically select the tuning parameter λ_k for each study k, we reccomend forming a logarithmic grid $\{c\ \gamma_k/\alpha_k^2 : c \in \mathcal{C}\}$ where γ_k/α_k^2 is the (estimated) optimal penalty for the single-study ridge on (Y_k, X_k) (Dobriban and Wager, 2018), and $\mathcal{C}$ is a small set such as $\{1/4, 1/2, 1, 2, 4\}$. Use K-fold CV on the target population to pick $\{\lambda_k\}_{k=1}^K$ for the desired objective (prediction or estimation), then recompute the final Trans-Ridge with full target population.

Remark 4 (Equal n_k and heterogeneous covariances)**.** When n_k are equal, several trace terms simplify (lemma **??** and lemma **??**). If populations have different covariances Σ_k, use the plug-in cross terms $\widehat{P}_{kk'} = \frac{1}{p}\operatorname{tr}\big[\widehat{\Sigma}_K(\widehat{\Sigma}_k + \lambda_k I_p)^{-1}(\widehat{\Sigma}_{k'} + \lambda_{k'} I_p)^{-1}\big]$ and $\widehat{d}_{kK} = \frac{1}{p}\operatorname{tr}\big[(\widehat{\Sigma}_k + \lambda_k I_p)^{-1}\widehat{\Sigma}_K\big]$ as in the heterogeneous-Σ discussion (see the remark 2).

Algorithm 2 Estimation part: study-level parameters, spectra, and ridge fits

Require: $\{(Y_k, X_k)\}_{k=1}^K$, candidate $\{\lambda_k\}$.
Ensure: $\{\widehat{\alpha}_k, \widehat{\sigma}_k\}_{k=1}^K$, $\{\widehat{\rho}_{kk'}\}_{k\neq k'}$, $\{\gamma_k, \widehat{\Sigma}_k, \widehat{m}_k, \widehat{m}'_k, \widehat{v}_k, \widehat{v}'_k\}_{k=1}^K$, $\{\widehat{E}_{kk'}, \widehat{P}_{kk'}\}_{k\neq k'}$, and $\{\widehat{\beta}_k\}_{k=1}^K$.

Signal and Correlation strength

for $k = 1, \ldots, K$ **do**
  Set $n_k = \text{rows}(X_k)$, $p = \text{cols}(X_k)$, $\gamma_k = p/n_k$, and $\widehat{\Sigma}_k = X_k^\top X_k/n_k$.
  Estimate $(\widehat{\alpha}_k, \widehat{\sigma}_k)$ with standard REML/GRM approaches (e.g., Lee et al. (2012)).
for $k \neq k'$ **do**
  Estimate cross-study coefficient correlations $\widehat{\rho}_{kk'}$ using standard estimators (e.g., Lee et al. (2012)).

Spectral terms

for $k = 1, \ldots, K$ **do**
  For a given $\lambda_k > 0$, compute

$$\widehat{m}_k(-\lambda_k) = \frac{1}{p}\,\text{tr}\big[(\widehat{\Sigma}_k + \lambda_k I_p)^{-1}\big], \quad \widehat{m}'_k(-\lambda_k) = \frac{1}{p}\,\text{tr}\big[(\widehat{\Sigma}_k + \lambda_k I_p)^{-2}\big].$$

  Using $\gamma_k = m : n$ relation, set $z = -\lambda_k$ and

$$\widehat{v}_k(-\lambda_k) = \gamma_k\left(\widehat{m}_k(-\lambda_k) - \frac{1}{\lambda_k}\right) + \frac{1}{\lambda_k}, \quad \widehat{v}'_k(-\lambda_k) = \gamma_k\left(\widehat{m}'_k(-\lambda_k) - \frac{1}{\lambda_k^2}\right) + \frac{1}{\lambda_k^2}.$$

Cross-study trace terms by lemma ?? and lemma ??

for $k \neq k'$ **do**

$$\widehat{E}_{kk'} = \frac{1}{\lambda_k\lambda_{k'}}\left\{\lambda_k\widehat{m}_k(-\lambda_k) + \lambda_{k'}\widehat{m}_{k'}(-\lambda_{k'}) + \frac{\lambda_k\widehat{m}_k(-\lambda_k)\,\widehat{m}_{k'}(-\lambda_{k'})}{\widehat{m}_k(-\lambda_k) - \widehat{m}_{k'}(-\lambda_{k'})} - \frac{\lambda_{k'}\widehat{m}_k(-\lambda_k)\,\widehat{m}_{k'}(-\lambda_{k'})}{\widehat{m}_k(-\lambda_k) - \widehat{m}_{k'}(-\lambda_{k'})}\right\}.$$

$$\widehat{P}_{kk'} = \frac{\lambda_k\widehat{m}_k(-\lambda_k) - \lambda_{k'}\widehat{m}_{k'}(-\lambda_{k'})}{\lambda_k\lambda_{k'}\big(\widehat{m}_{k'}(-\lambda_{k'}) - \widehat{m}_k(-\lambda_k)\big)}.$$

Study-wise Ridge estimators

for $k = 1, \ldots, K$ **do**
  $\widehat{\beta}_k = (\widehat{\Sigma}_k + \lambda_k I_p)^{-1} X_k^\top Y_k/n_k$.

Algorithm 3 Assemble part: compute optimal weights and the Trans-Ridge estimators

Require: Outputs of Alg. 2 and penalties $\{\lambda_k\}$.
Ensure: W_E^*, W_P^*.

Optimal estimation weight (Theorem 2):

1: Build $\widehat{V} \in \mathbb{R}^K$ with

$$\widehat{V}_k = \begin{cases} \widehat{\rho}_{kK}\widehat{\sigma}_k\widehat{\sigma}_K\,\widehat{\alpha}_k\widehat{\alpha}_K\big\{1-\lambda_k\,\widehat{m}_k(-\lambda_k)\big\}, & k \neq K, \\ \widehat{\sigma}_K^2\widehat{\alpha}_K^2\big\{1-\lambda_K\,\widehat{m}_K(-\lambda_K)\big\}, & k = K. \end{cases}$$

2: Build $\widehat{A} \in \mathbb{R}^{K\times K}$ with

$$\widehat{A}_{kk} = \widehat{\sigma}_k^2\widehat{\alpha}_k^2\big\{1-2\lambda_k\widehat{m}_k(-\lambda_k)+\lambda_k^2\widehat{m}_k'(-\lambda_k)\big\},$$

$$\widehat{A}_{kk'} = \widehat{\rho}_{kk'}\widehat{\sigma}_k\widehat{\sigma}_{k'}\widehat{\alpha}_k\widehat{\alpha}_{k'}\Big\{1-\lambda_k\widehat{m}_k(-\lambda_k)-\lambda_{k'}\widehat{m}_{k'}(-\lambda_{k'})+\lambda_k\lambda_{k'}\,\widehat{E}_{kk'}\Big\}\ (k\neq k').$$

3: Let $\widehat{R} = \mathrm{diag}\big(\widehat{R}_{11},\ldots,\widehat{R}_{KK}\big)$ with

$$\widehat{R}_{kk} = \widehat{\sigma}_k^2\,\gamma_k\{\widehat{m}_k(-\lambda_k)-\lambda_k\,\widehat{m}_k'(-\lambda_k)\}.$$

4: Compute $\widehat{W}_E^* = (\widehat{A}+\widehat{R})^{-1}\widehat{V}$

Optimal Prediction Weight (Theorem 4):

1: Build $\widehat{D} \in \mathbb{R}^K$ with

$$\widehat{D}_k = \begin{cases} \widehat{\rho}_{kK}\widehat{\sigma}_k\widehat{\sigma}_K\,\widehat{\alpha}_k\widehat{\alpha}_K\left[1-\dfrac{1}{\gamma_k\,\widehat{v}_k(-\lambda_k)}\right], & k \neq K, \\ \widehat{\sigma}_K^2\widehat{\alpha}_K^2\left[1-\dfrac{1}{\gamma_K\,\widehat{v}_K(-\lambda_K)}\right], & k = K, \end{cases}$$

2: Build $\widehat{C} \in \mathbb{R}^{K\times K}$ with

$$\widehat{C}_{kk} = \widehat{\sigma}_k^2\widehat{\alpha}_k^2\left[1-\frac{2}{\gamma_k\,\widehat{v}_k(-\lambda_k)}+\frac{\widehat{v}_k(-\lambda_k)-\lambda_k\,\widehat{v}_k'(-\lambda_k)}{\gamma_k\,\widehat{v}_k^2(-\lambda_k)}\right],$$

$$\widehat{C}_{kk'} = \widehat{\rho}_{kk'}\widehat{\sigma}_k\widehat{\sigma}_{k'}\widehat{\alpha}_k\widehat{\alpha}_{k'}\left[1-\frac{1}{\gamma_k\,\widehat{v}_k(-\lambda_k)}-\frac{1}{\gamma_{k'}\,\widehat{v}_{k'}(-\lambda_{k'})}+\lambda_k\lambda_{k'}\,\widehat{P}_{kk'}\right]\ (k\neq k').$$

3: Let $\widehat{F} = \mathrm{diag}\big(\widehat{F}_{11},\ldots,\widehat{F}_{KK}\big)$ with

$$\widehat{F}_{kk} = \widehat{\sigma}_k^2\gamma_k\left[\frac{1}{\gamma_k}\left(\frac{1}{\lambda_k\widehat{v}_k(-\lambda_k)}-1\right)-\lambda_k\cdot\frac{\widehat{v}_k(-\lambda_k)-\lambda_k\widehat{v}_k'(-\lambda_k)}{\gamma_k\,[\lambda_k\widehat{v}_k(-\lambda_k)]^2}\right].$$

4: Compute $\widehat{W}_P^* = (\widehat{C}+\widehat{F})^{-1}\widehat{D}$