Transfer Learning with Random Coefficients Ridge Regression and Its Applications in Genomics

Hongzhe Li

University of Pennsylvania

June 18, 2026

Thanks

Joint work with Hongzhe Zhang and Arnab Auddy

Zhang H and Li H (2026): Statistica Sinica
Zhang H, Auddy A and Li H (2026): JMLR

Ridge Regression in Genomics and Microbiome Studies

Sparse vs dense models

- Sparse models (e.g. lasso) are effective in identifying trait-associated variants/taxa with large effects.
  - Requires sparsity and sufficiently large effect sizes
  - Transfer Lasso (Li, Cai and Li (2022, JRSSB); Li, Zhang, Cai and Li (2023 JASA)).

- Polygenic models or random-effects ridge regression models are more appropriate for estimating genetic heritability or overall microbiome effects.

- Random-effects ridge regression models provide a way to capture aggregate effects of SNPs or bacterial taxa on outcomes.

Linear random coefficients model

Genetics/Microbiome risk model - each variant/taxon has a small, independent random effect on the outcome, x_i is centered and scaled.

The random coefficients regression framework

$$\underbrace{y}_{\text{Trait, Eg. BMI}} = \beta_1 \underbrace{x_1}_{\text{SNP 1}} + \beta_2 \underbrace{x_2}_{\text{SNP 2}} + \cdots + \beta_p \underbrace{x_p}_{\text{SNP p}} + \epsilon$$

$Var(\epsilon) = \sigma^2$

Random regression coefficient (RRC) assumption:
The coefficients $\beta = (\beta_i, i = 1, \cdots, p)$ are random with

$$E(\beta) = 0 \quad Var[\sqrt{p}\beta] = \alpha^2 \sigma^2 I_P$$

$$\alpha^2 = E[||\beta||_2^2]/\sigma^2 \Rightarrow \text{signal to noise ratio}$$

$$\text{Heritability/Variance explained } h = \alpha^2/(1+\alpha^2)$$

High dimensional asymptotics of ridge prediction

- Ridge regression estimate with tuning parameter λ

$$\hat{\beta}_\lambda = (X^T X + \lambda n I_p)^{-1} X^T y$$

- Dobriban and Wager (2018 AOS) provides results on prediction risk of ridge regression under $p, n \to \infty$ and $p/n \to \gamma > 0$.

- Goal: - Develop transfer learning methods for random-effects ridge regression and characterize their high-dimensional behavior.

Why transfer learning for genetics/microbiome-based prediction?

Goal - improve genetics/microbiome-based prediction for a target population by utilizing data from multiple source/auxiliary populations.

Setup - Leverage available information across related traits or different study populations.

- Cross different traits - e.g., cardio-metabolic traits, psychiatric disorders.
- Cross different studies - study conducted independently for the same trait.

Transfer learning - model setup and assumptions

Target model: $\beta_K = (\beta_{Kj}, j = 1, \cdots, p)$, $Var(\epsilon_K) = \sigma_K^2$

$$\underbrace{y}_{\text{Target Trait}} = \beta_{K1} \underbrace{x_1}_{\text{SNP 1}} + \beta_{K2} \underbrace{x_2}_{\text{SNP 2}} + \cdots + \beta_{Kp} \underbrace{x_p}_{\text{SNP p}} + \epsilon_K$$

Source models, $1, \cdots, K-1$,

$$\underbrace{y}_{\text{Source Trait } k} = \beta_{k1} \underbrace{x_1}_{\text{SNP 1}} + \beta_{k2} \underbrace{x_2}_{\text{SNP 2}} + \cdots + \beta_{kp} \underbrace{x_p}_{\text{SNP p}} + \epsilon_k$$

$\beta_k = (\beta_{kj}, j = 1, \cdots, p)$, $Var(\epsilon_k) = \sigma_k^2$.

RRC assumptions: $E(\beta_k) = 0$, $Var[\sqrt{p}\beta_k] = \alpha_k^2 \sigma_k^2 I_P$.

$Cov(\sqrt{p}\beta_k, \sqrt{p}\beta_l) = \rho_{kl} \alpha_k \alpha_l \sigma_k \sigma_l I_p$ for $k, l = 1, \cdots, K, k \neq l$

ρ_{kl} : genetic/microbiome correlation between outcomes k and l

$X_k \sim Z_k \Sigma^{1/2}$, with Z i.i.d entries with $E[Z_{ij}] = 0, Var[Z_{ij}] = 1$

Transfer learning via ridge regression (TransRidge)

Proposed estimator for β_K in target model:

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

where $\hat{\beta}_k = (X_k^T X + n_k \lambda_k I_p)^{-1} X_k^T Y_k$ is the ridge estimator with data from source k.

Dobriban, E., & Sheng, Y. (2020) for distributed learning.

How to choose the weights W_k?

Optimal Estimation Weight

Estimate $W_k, k = 1, \cdots, K$ by minimizing the estimation risk for β_K

$$\tilde{\beta}_K = \arg\min M_K(W) := E[||\sum_{k=1}^{K} W_k \hat{\beta}_k - \beta_K||^2 | X]$$

Define $\hat{\Sigma}_k = n_k^{-1} X_k^T X_k$ and

$$\begin{aligned} Q_k = (\hat{\Sigma}_k + \lambda_k I_p)^{-1} \hat{\Sigma}_k \quad &\Rightarrow \quad B = [Q_1 \beta_1, \cdots, Q_K \beta_K] \\ R = diag\{R_k\}, R_k \quad &= \quad n_k^{-1} \sigma_k^2 \, \mathrm{tr}[(\hat{\Sigma}_k + \lambda_k I_p)^{-2} \hat{\Sigma}_k]. \end{aligned}$$

$$\Downarrow$$

The optimal estimation weights and the estimation risk are given as

$$W_E^* = (B^T B + R)^{-1} B^T \beta_K$$

$$M_K(W_E^*) = \beta_K^T [I_p - B(B^T B + R)^{-1} B^T] \beta_K$$

Depends on unknown $\beta_1, \cdots, \beta_K$.

$W_E^* \to_{a.s.} \mathcal{W}^*$ and $M_K(W_E^*) \to_{a.s.} \mathcal{M}_K(\mathcal{W}^*)$.

Asymptotic analysis of W_E^* and $M_K(W_E^*)$ under RRC

Asymptotic regime: $n_k, p \to \infty$ while $p/n_k \to \gamma_k > 0$.

For matrix A independent of β_k, β_l

$$\beta_k^T A \beta_l - \rho_{kl}\alpha_k\alpha_l\sigma_k\sigma_l \operatorname{tr}(A)/p \to_{a.s.} 0$$

Need limiting value of

$$\operatorname{tr} Q_k/p, \quad \operatorname{tr} Q_k Q_l/p, \quad \operatorname{tr} S_k S_l/p, \quad \operatorname{tr}[(\hat{\Sigma}_k + \lambda_k I_p)^{-2}\hat{\Sigma}_i]/p$$

where

$$Q_k = (\hat{\Sigma}_k + \lambda_k I_p)^{-1}\hat{\Sigma}_k, \quad S_k = (\hat{\Sigma}_k + \lambda_k I_p)^{-1}$$

Random matrix theorem and spectral distribution.

Asymptotic analysis of W_E^* and $M(W_E^*)$ under RRC

$V := B^T\beta_K$, $A := B^T B$ and R can be consistently estimated through Stieltjes transform m_{F_γ}, m'_{F_γ} of the limit of the empirical spectral distribution of $\hat{\Sigma}_k$ using Marchenko–Pastur theorem.

$$V_k \overset{a.s.}{\to} \mathcal{V}_k = \begin{cases} \rho_{kK}\sigma_k\sigma_K\alpha_k\alpha_K\{1 - \lambda_k m_{F_{\gamma_k}}(-\lambda_k)\}, & k \neq K \\ \sigma_K^2\alpha_K^2\{1 - \lambda_K m_{F_{\gamma_K}}(-\lambda_K)\} \end{cases}$$

$$A_{kk'} \overset{a.s.}{\to} \mathcal{A}_{kk'} = \begin{cases} \rho_{kk'}\sigma_k\sigma_{k'}\alpha_k\alpha_{k'}\{1 - \lambda_k m_{F_{\gamma_k}}(-\lambda_k) - \lambda_{k'} m_{F_{\gamma_{k'}}}(-\lambda_{k'}) + \lambda_k\lambda_{k'}\mathcal{E}_{kk'}\} & k \neq k' \\ \sigma_k^2\alpha_k^2\{1 - 2\lambda_k m_{F_{\gamma_k}}(-\lambda_k) + \lambda_k^2 m'_{F_{\gamma_k}}(-\lambda_k) & k = k' \end{cases}$$

$$R_{kk} \overset{a.s.}{\to} \mathcal{R}_{kk} = \sigma_k^2\gamma_k\{m_{F_{\gamma_k}}(-\lambda_k) - \lambda_k m'_{F_{\gamma_k}}(-\lambda_k)\}$$

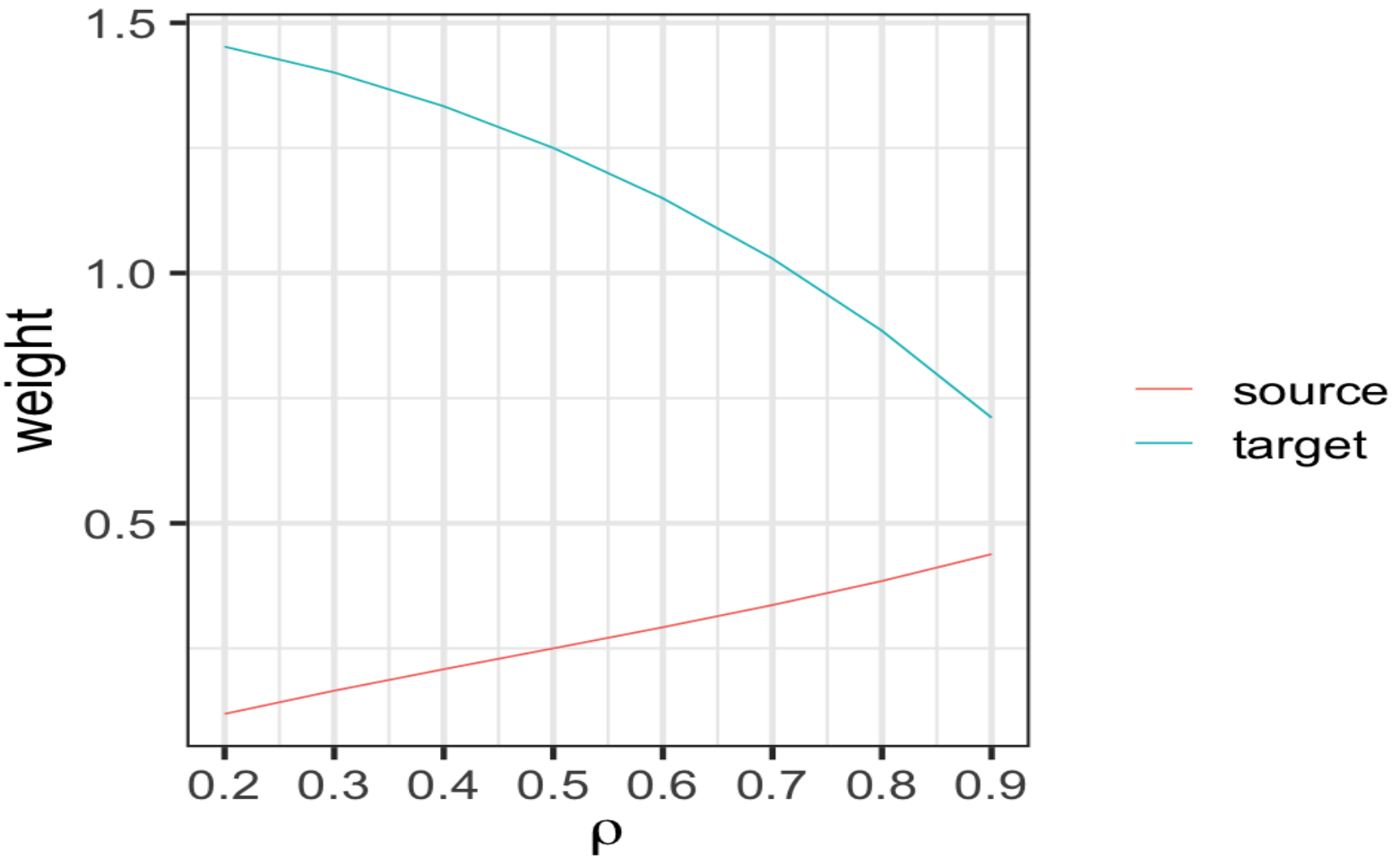
$$\Downarrow$$

$$W_E^* \to_{a.s.} \mathcal{W}_E^* = (\mathcal{A} + \mathcal{R})^{-1}\mathcal{V}$$

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

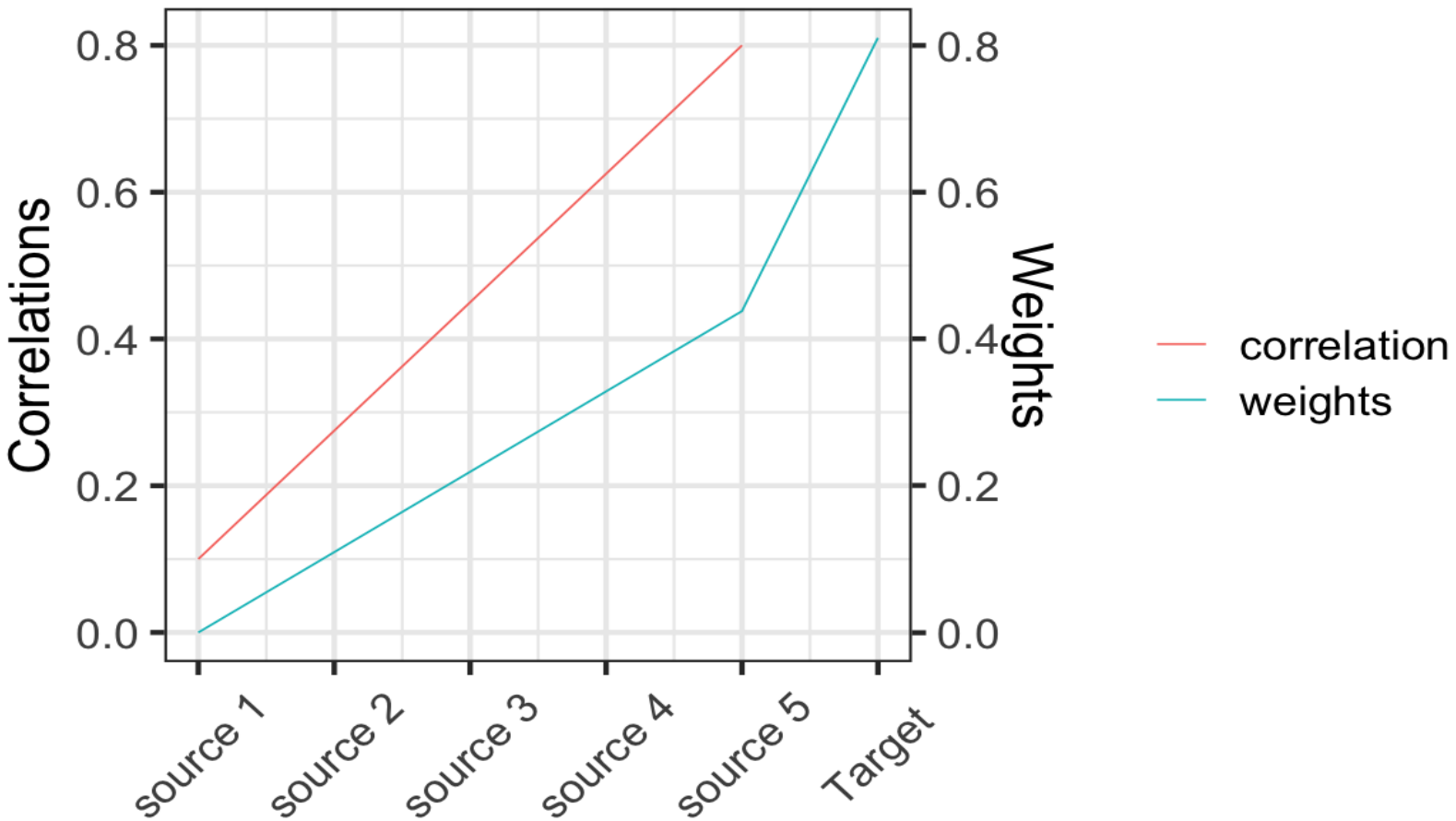
Asymptotic Weights ($\rho_{ij} = \rho$)

$K = 6$, asymptotic weights, $n_k = 100, p = 100$.



Asymptotic Weights -adaptive to different ρ_{ij}

$K = 6$, asymptotic weights, $n_k = 100, p = 100$.



Estimation Risk: Accuracy of Asymptotics

Limiting estimation risk:

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

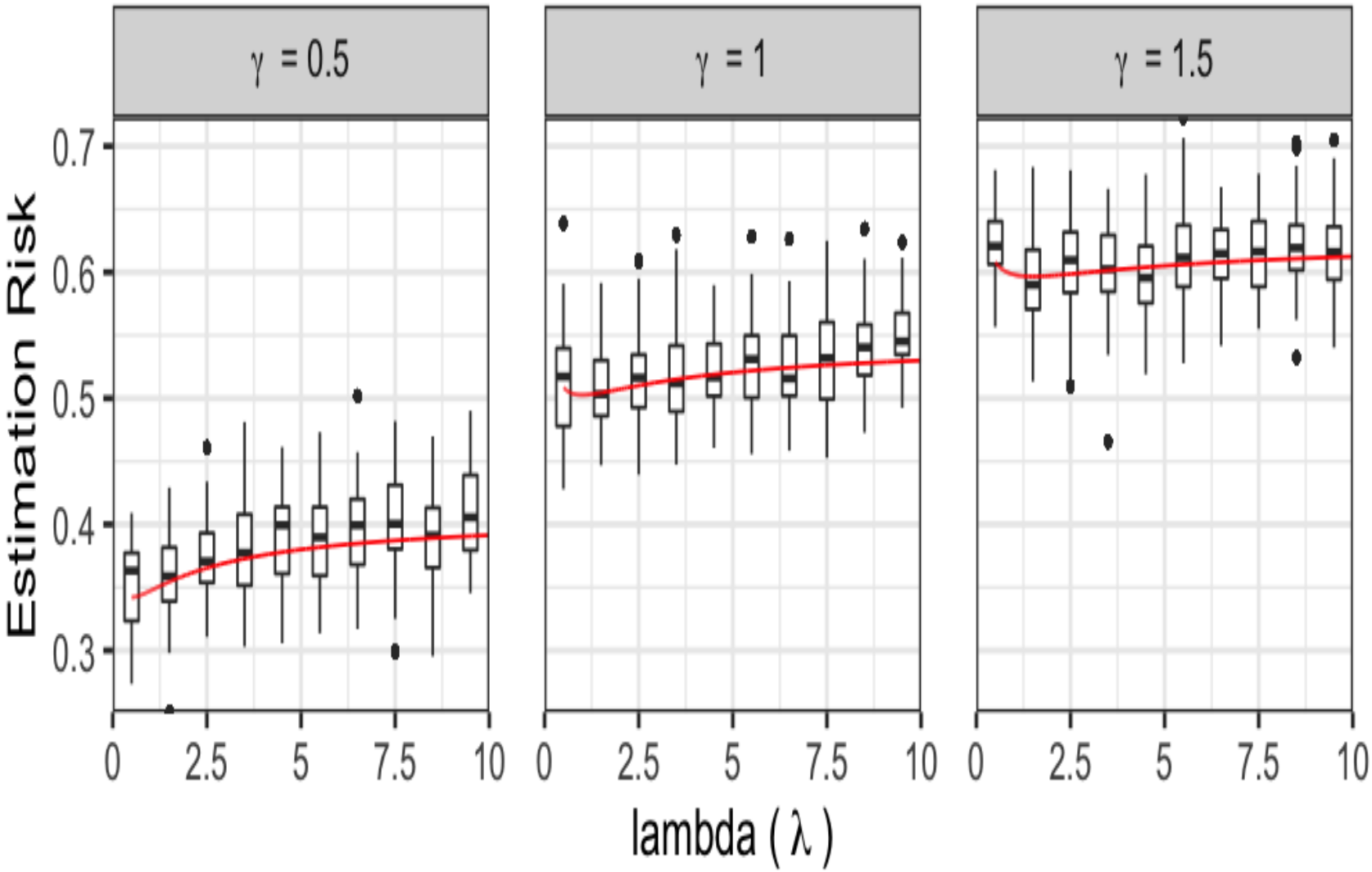
The limiting value $\mathcal{A}$ is different under three set ups:

- equal sample sizes: $\gamma_1 = \cdots = \gamma_K = \gamma,\ \lambda_1 = \cdots = \lambda_K = \lambda$
- isotropic design: $\Sigma = I_p$
- The general case

Estimation Risk - Accuracy of Asymptotics

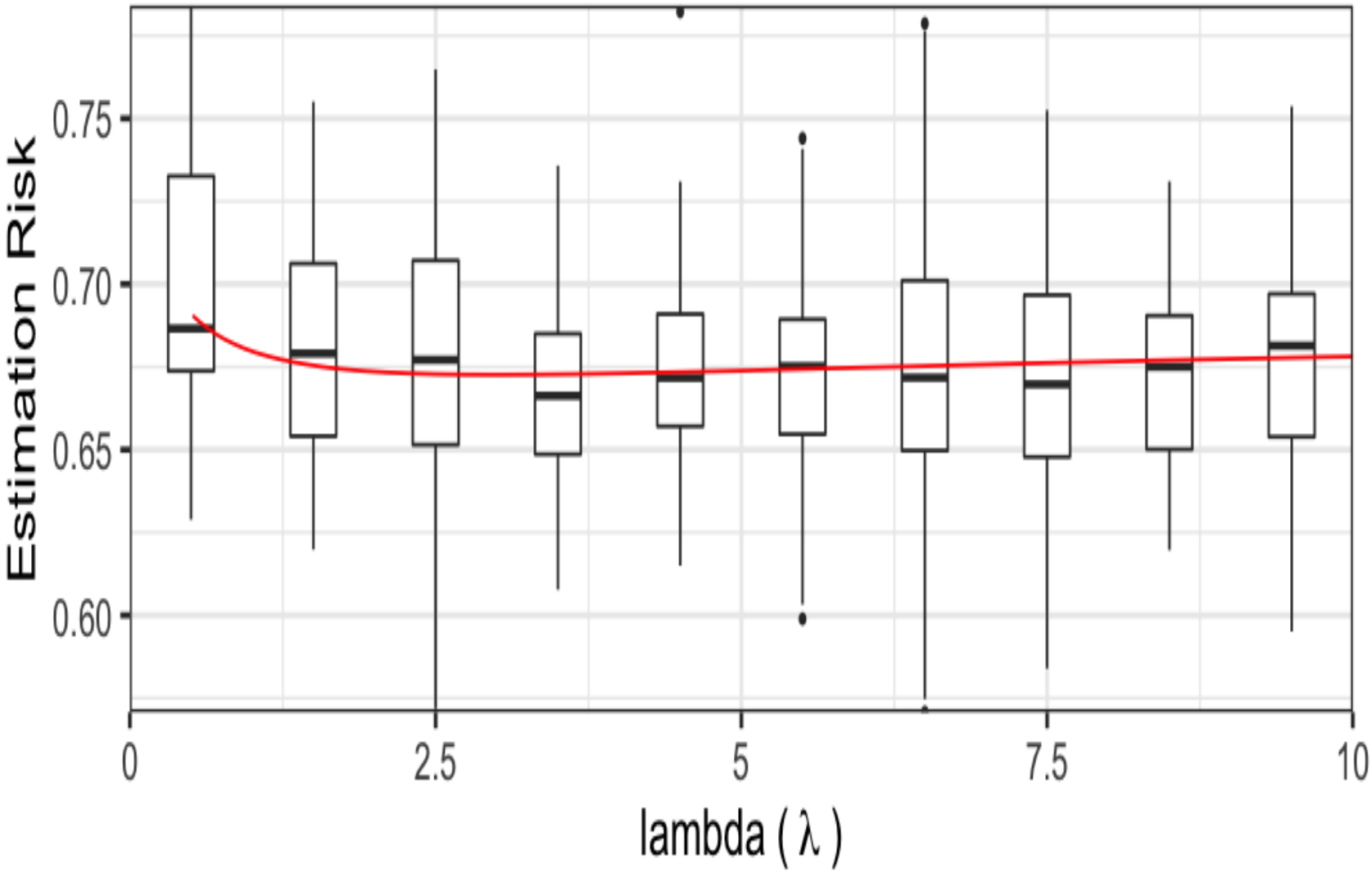
Equal sample sizes $n_k = 500, \rho = 0.5, p = 500 * \gamma$.

$\gamma_1 = \cdots = \gamma_K = \gamma, \ \lambda_1 = \cdots = \lambda_K = \lambda$



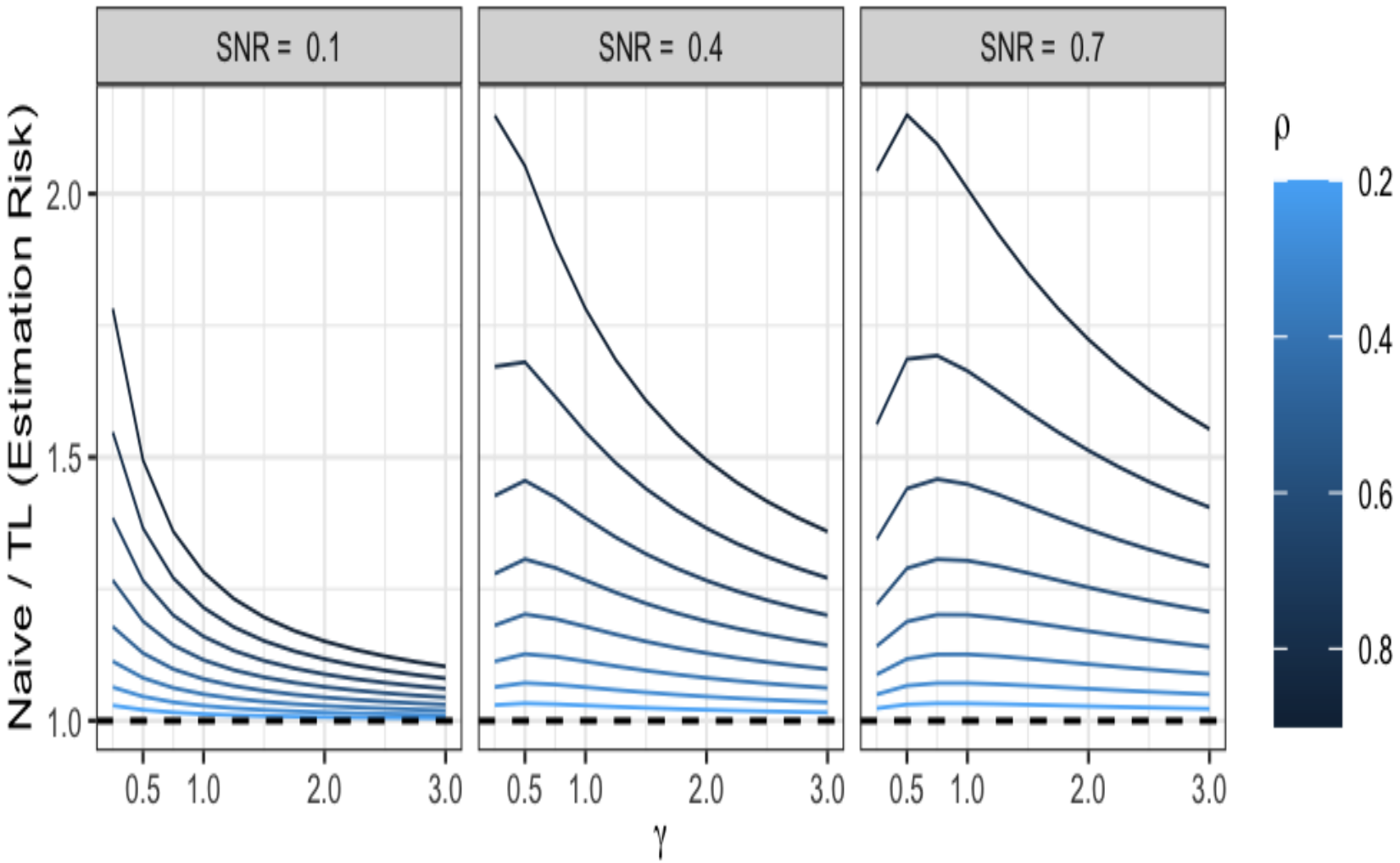
Estimation Risk - Accuracy of Asymptotics

Isotropic design $\Sigma = I_p$, $p = 750$, $n_K = 750$, and the source sample sizes range from 700 to 250.



Relative improvement of estimation risk over target-only ridge

Variance explained $h = 0.091, 0.28, 0.41$.



Weights based on prediction risk

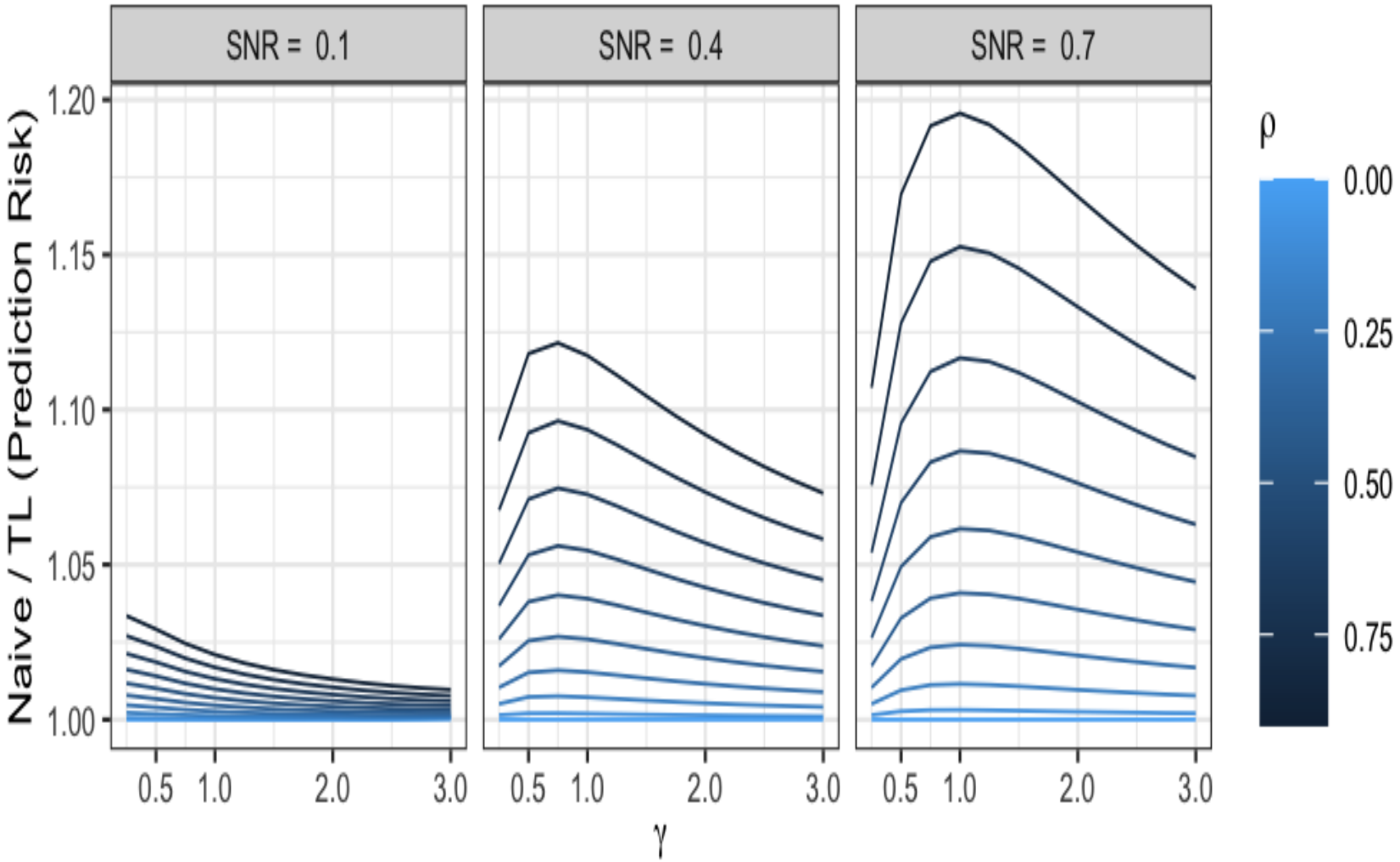
Find $W_k, k = 1, \cdots, K$ that minimizes the prediction risk

$$
\begin{aligned}
r(W) =& E_{\epsilon_1, \cdots, \epsilon_K, x_0}[(y_0 - \sum_{k=1}^{K} W_k \hat{\beta}_k^T x_0)^2 | X] \\
=& \sigma_K^2 + \underbrace{\text{tr}[\Sigma Cov_\epsilon(\sum_{k=1}^{K} W_k \hat{\beta}_k^T - \beta_K | X)]}_{\mathcal{V}_K(\beta)} \\
&+ \underbrace{E_\epsilon(\sum_{k=1}^{K} W_k \hat{\beta}_k^T - \beta_K | X)^T \Sigma E_\epsilon(\sum_{k=1}^{K} W_k \hat{\beta}_k^T - \beta_K | X)}_{\mathcal{B}_K(\beta)}
\end{aligned}
$$

Closed-form weights and their limiting values are derived (Zhang and Li, 2026).

Relative improvement in prediction risk over target-only ridge

variance explained $h = 0.091, 0.28, 0.41$.



Estimation vs prediction based weights

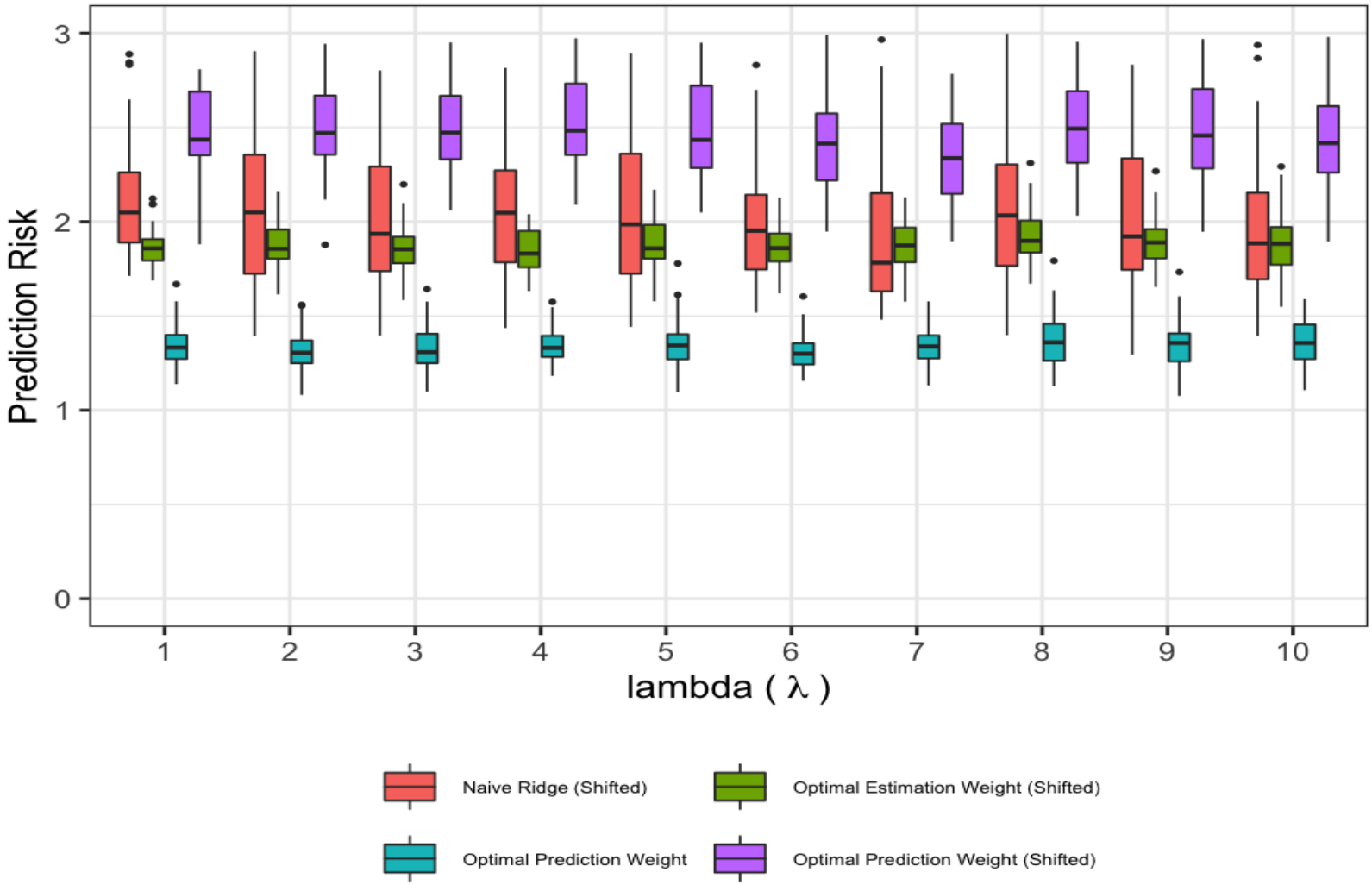
Estimation weights are distributionally robust.

When the testing data x_{K0} has a different distribution from the training data, weights from minimizing the estimation risk is robust to distributions specified by $\mathcal{P} := \{P : E(x) = 0, E_P(||x||_2) \leq c\}$,

$$\min_{W} \max_{x_{K0} \in \mathcal{P}} E_{\epsilon, x_{K0}}(||(\sum_{k=1}^{K} W_k \hat{\beta}_k - \beta_K)^T x_{K0}||^2).$$

Lead to weights determined by the estimation risk.

Comparison of estimation and prediction based weights



Polygenic risk score of lipid traits

Penn Medicine BioBank (PMBB) data.

- Three phenotypes: high density lipoprotein cholesterol (HDL), low density lipoprotein (LDL) cholesterol, triglycerides.
- 11,616 individuals with the lipid data, divide into 3 groups ($n_k = 3872$) so no overlapping samples.
- $n_t = 1000$ as testing data set.
- The top 2000 - 5000 most significant single-nucleotide polymorphisms (SNPs) were selected as predictors via univariate filtering

Results (MSE) - cross-trait transfer

Each trait is a target, and the other two traits are source traits.
MTAG (Turley et al., 2018); Trans-Lasso (Li, Cai and Li, 2022)

	OptEst	OptPred	Target	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	9.953
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

Microbiome and Colorectal Cancer - cross-study transfer

Samples sizes of Zackular (US & Canada), Zeller (France), and Baxter (USA) studies are 83, 127, and 488, respectively.

146 genera and phyla of the microbiomes ($> 10\%$ sample) and three covariates, age, sex, BMI ($p = 149$) remain in the analysis.

Table: Leave-one-out classification error

Target	n_k	Target-only	Sample pooling	TL Ridge(p)	TL Ridge (e)
Zackular	83	48.19	33.73	**28.92**	31.33
Zeller	127	29.92	37.01	**24.41**	25.20
Baxter	488	25.04	30.33	**22.54**	23.57

Regularized Linear Discriminant Analysis: Transfer Learning

Transfer Learning Set up — We have $K-1$ auxilliary studies, for $k \in [K-1]$ and target study K.

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

$$\bar{\mu}_k = 0.5(\mu_{1,k} + \mu_{-1,k}) \quad \delta_k = 0.5(\mu_{1,k} - \mu_{-1,k})$$

RC assumptions: $E(\delta_k) = 0$, $Var[\sqrt{p}\delta_k] = \alpha_k^2 I_p$.

$$Cov(\sqrt{p}\delta_k, \sqrt{p}\delta_{k'}) = \rho_{kk'}\alpha_k\alpha_{k'} I_p \text{ for } k, k' = 1, \cdots, K, k \neq k'$$

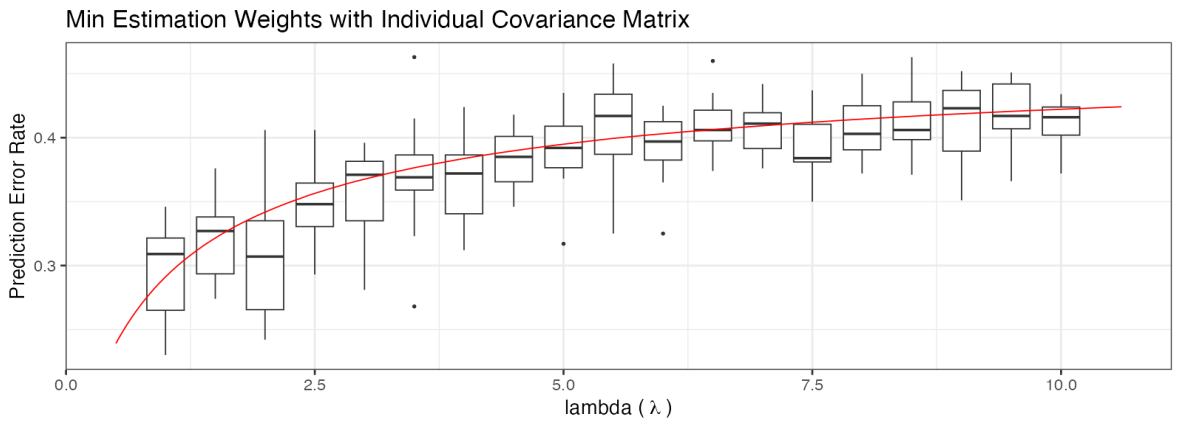
Transfer Learning Estimator — We use a weighted sum of individual discriminant directions (TL-RDA), $\hat{d}_k = (\hat{\Sigma}_k + \lambda_k)^{-1}\hat{\delta}_k$,

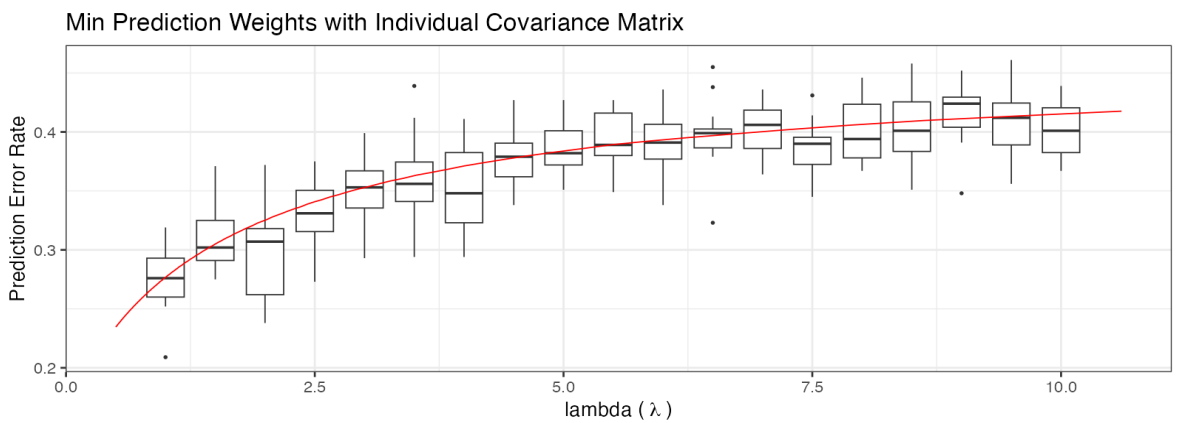
$$\hat{d}(W) = \sum_{k=1}^{K} W_k \hat{d}_k$$

$$\hat{y}_{TL}(x_0) = sign[\hat{d}(W)^T(x_0 - \frac{\hat{\mu}_{-1,K} + \hat{\mu}_{1,K}}{2})]$$

Optimal Weight

Derived consistent estimators for both W_E, W_P as well as the parameters $\alpha_k^2, \rho_{kk'}$, as $n_k, p \to \infty, p/n_k \to \gamma_k$ (Zhang, Auddy and Li, 2026).





Optimal Weight for Classification

In addition to optimizing the score $\hat{d}(\mathrm{w})^\top x_0$, the prediction weights w^P and w_P^P also minimize the misclassification error.

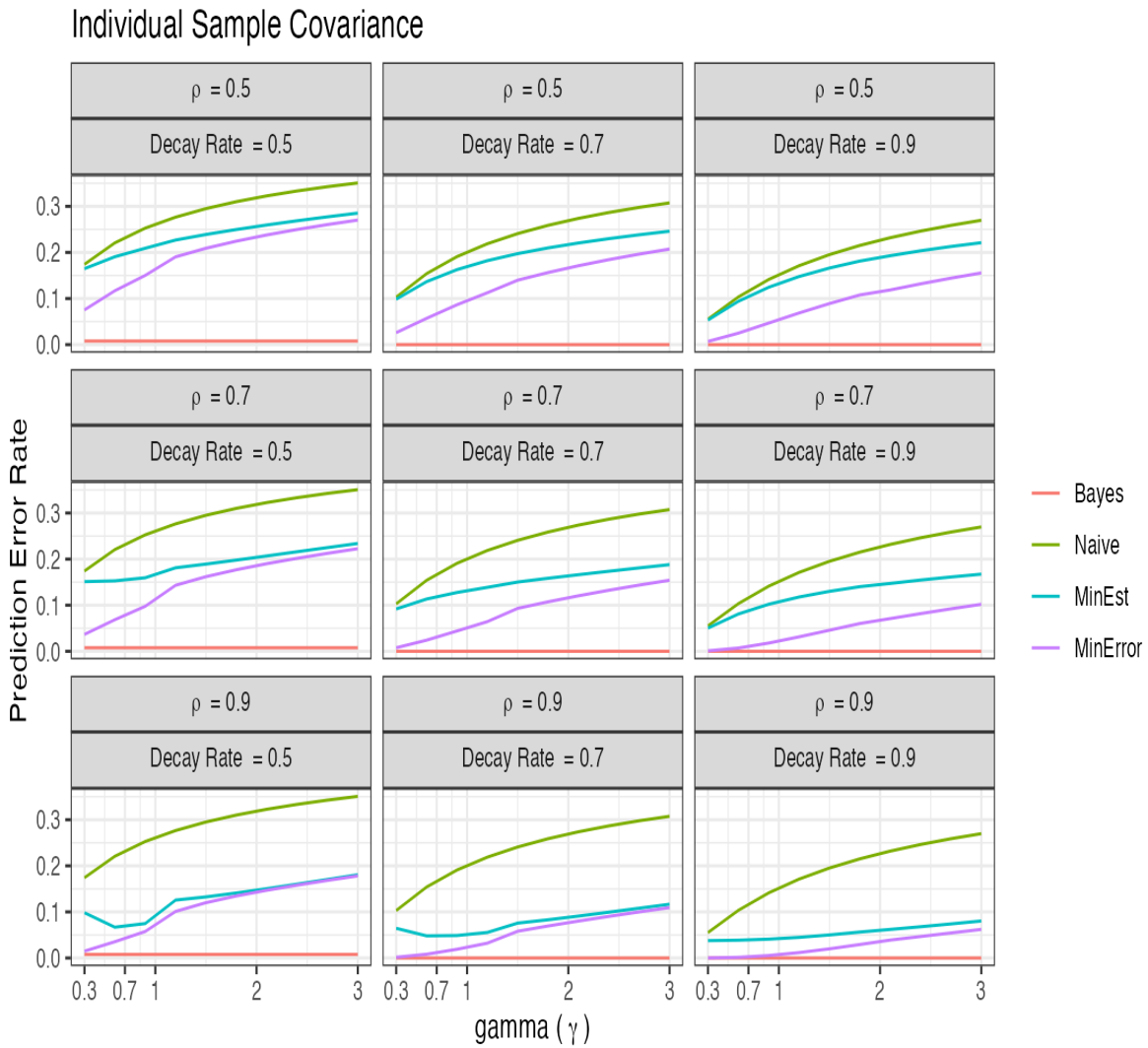
Proposition

The optimal prediction weight minimizes the testing data classification error:

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

w_P^P: pooled covariance matrix is used (more later)

Optimal Prediction Weight is Optimal



Proved $\mathcal{W}^E, \mathcal{W}^P$ always exist and they are unique.

Pooled Sample Covariance

Recall we assumed the population covariance Σ is the same across all populations. A natural extention is to use the pooled sample covariance matrix (TLP-RDA).

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

$$\hat{d}_k^P = (\hat{\Sigma}_P + \lambda_k I)^{-1}\hat{\delta}_k$$

$$\hat{d}^P(W) = \sum_{k=1}^{K} W_k \hat{d}_k^P$$

$$\hat{y}_W^P(x_{K0}) = \text{sign}[\hat{d}^P(W)^\top x_{K0} + \hat{b}_K].$$

All previous results hold for TL-RDA also hold for TLP-RDA. We define the corresponding limiting optimal weights as $\mathcal{W}_P^E, \mathcal{W}_P^P$.

Weight Selection

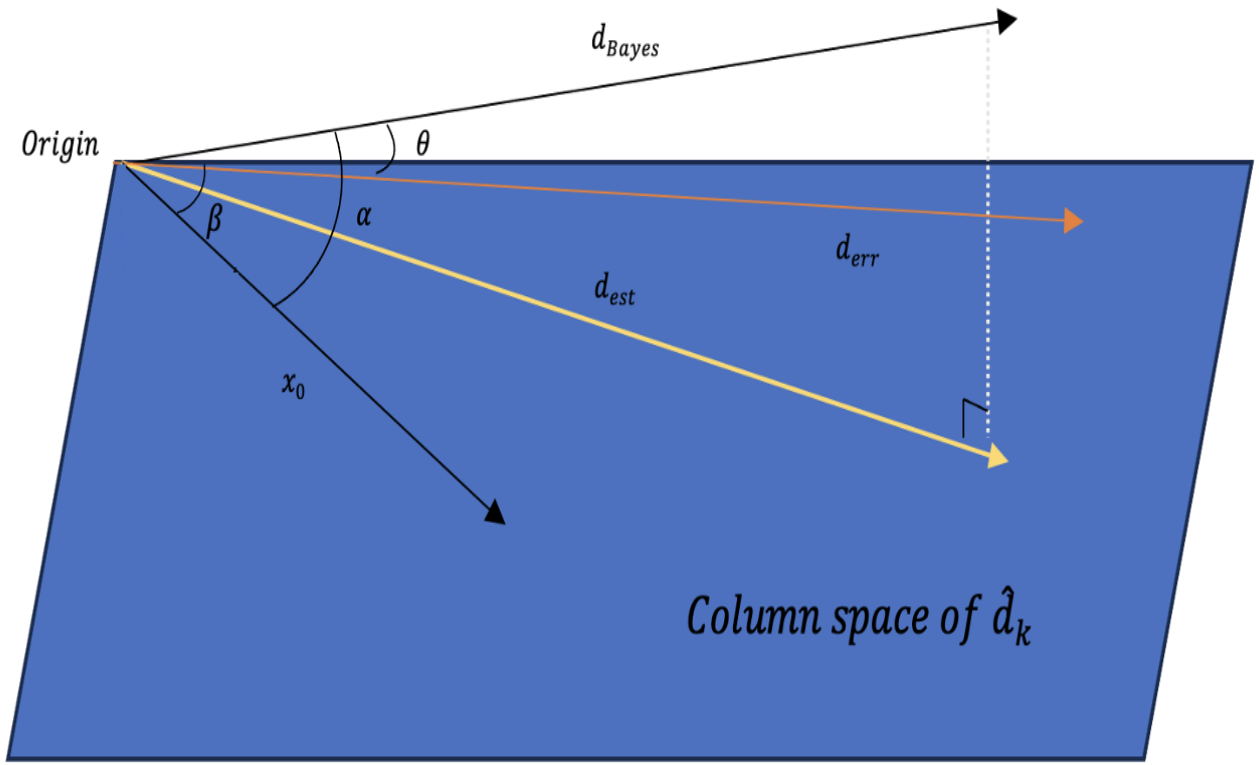
4 types of optimal weights $\Rightarrow$ 4 types of TL estimator.

- Pooled v.s. Individual sample covariance matrix
- Optimal Prediction v.s. Optimal Estimation weights

Geometric Interpretation of the Weights

$$\hat{d}(W) = \sum_{k=1}^{K} W_k \hat{d}_k$$

$$d_{est} := \hat{d}(W_E), \quad d_{err} := \hat{d}(W_p)$$



Pooled v.s. Individual sample covariance matrix

Assume $\Sigma = I_p$, $\gamma_k = \gamma$, $\lambda_k = \lambda = r\left(\gamma - \frac{1}{r+1}\right)$ for some fixed $r > (1-\gamma)_+/\gamma$. For the pooled covariance matrix, we choose $\lambda' = r'\left(\gamma/K - \frac{1}{r'+1}\right)$ for some $r' > (K-\gamma)_+/\gamma$.

1. When $\rho = 1$, $Err(W^P) \geq Err_P(W_P^P)$, if and only if

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

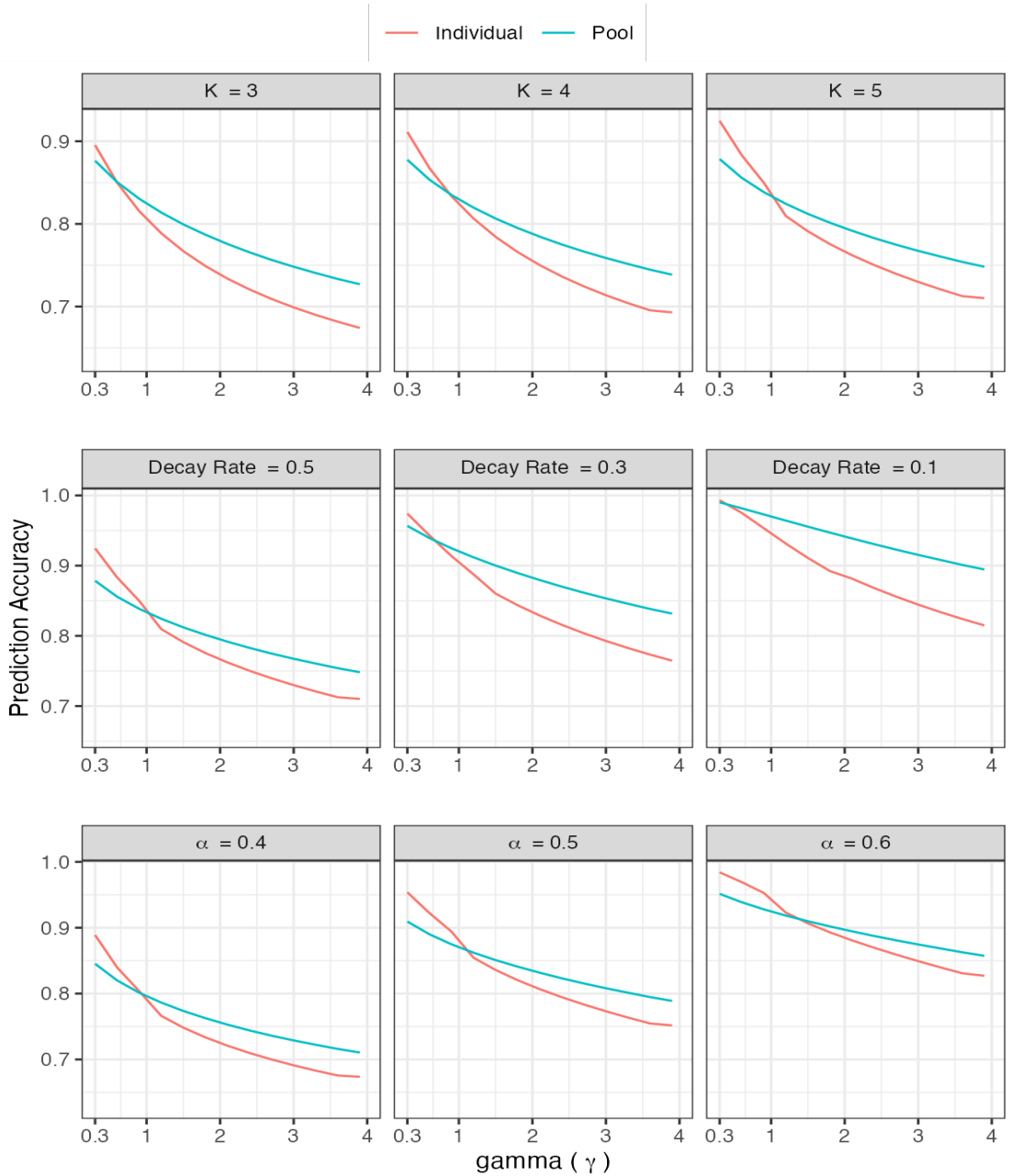
2. When $\rho = 0$, $Err(W^P) \geq Err_P(W_P^P)$, if and only if

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

3. $0 < \rho < 1$, a bit long but the idea is the same.

Pooled covariance is better when dimension ($\gamma := p/n$) is large.

Pooled v.s. Individual sample covariance matrix



Estimation weight v.s. Prediction weight

Proposition (Robustness of Estimation Weight)

For the class of covariate distributions given by

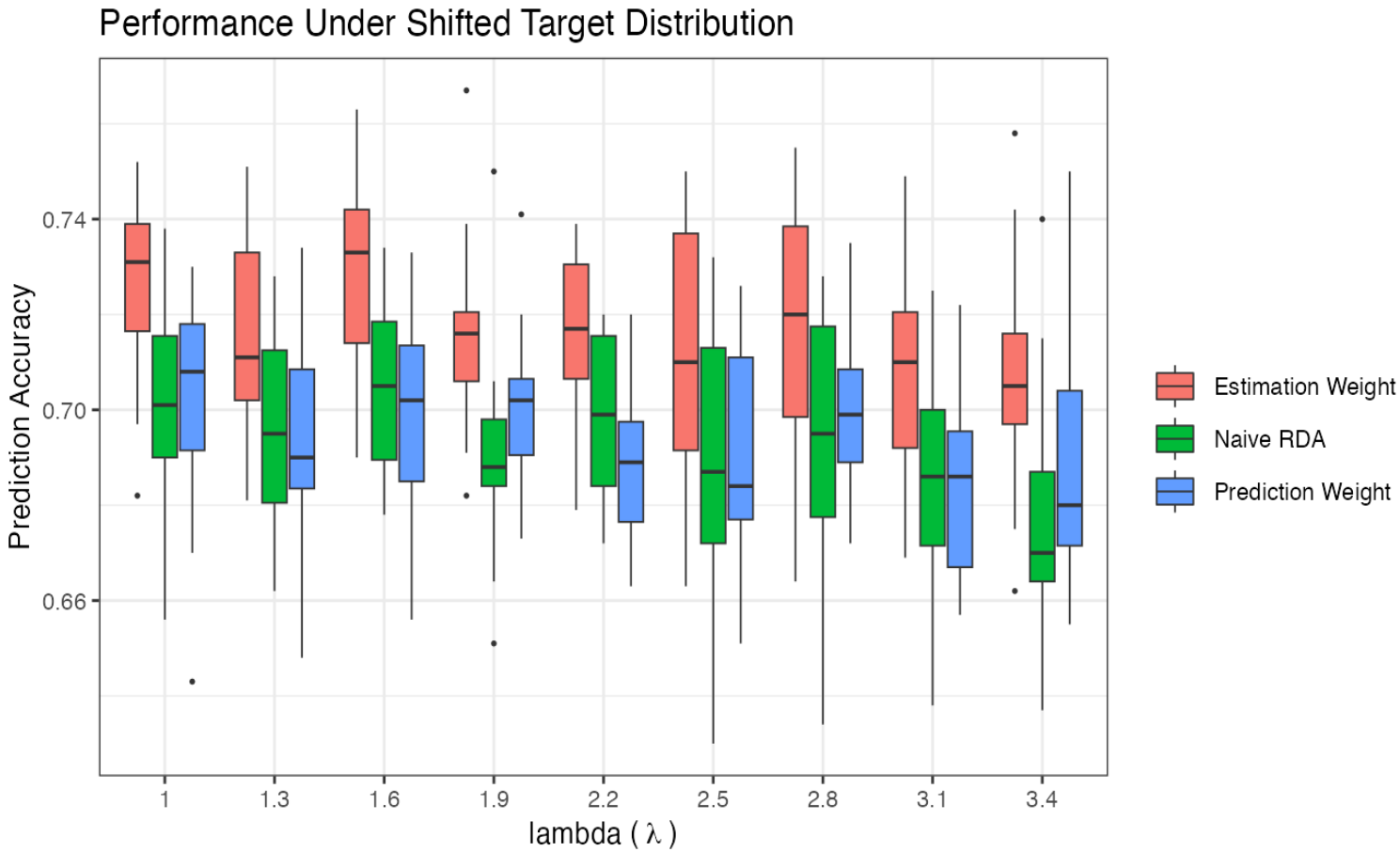
$$\mathcal{P} := \{P : x \sim P, \mathbb{E}_P(\|x\|_2) \leq c\}$$

we have:

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

Optimal Prediction v.s. Optimal estimation weights

The optimal estimation weights beat the optimal prediction weights when the covariance matrix is different in the test data.



Heterogeneous Covariance: Per-Study Σ_k — Both Methods

Relax the common-Σ assumption: each study has its own covariance. This extension applies to both methods.

$$(X_k)_i \mid y_k \sim N(\mu_{y,k}, \Sigma_k), \qquad \text{Bayes direction} = \Sigma_K^{-1}\delta_K.$$

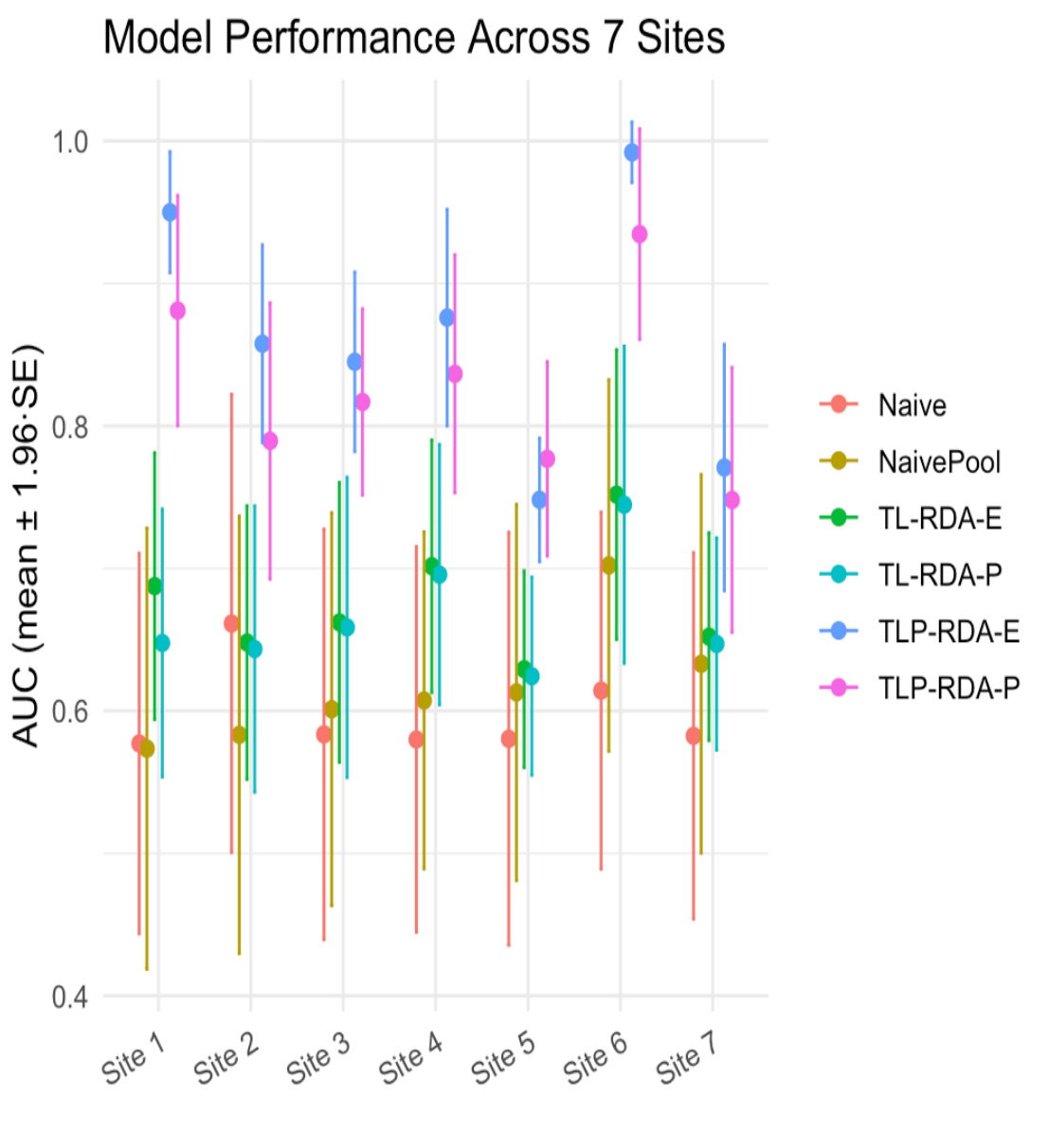
Optimal weights are derived in full for both methods, with no structural assumption on the Σ_k. The cross terms now carry the target covariance:

- **Regression:** $\operatorname{tr}\left[\Sigma_K(\hat{\Sigma}_k + \lambda_k I)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I)^{-1}\right]/p.$
- **Classification:** $\operatorname{tr}\left[(\hat{\Sigma}_k + \lambda_k I)^{-1}(\hat{\Sigma}_{k'} + \lambda_{k'} I)^{-1}\Sigma_K\right]/p.$

Both remain estimable: cross-resolvent traces are observed directly, and target-covariance terms are recovered by plugging in $\hat{\Sigma}_K$ with a deterministic-equivalent correction term.

No shared-eigenbasis or other structural assumption is needed.

MESA proteomics data, site-specific 10 year CVD risk

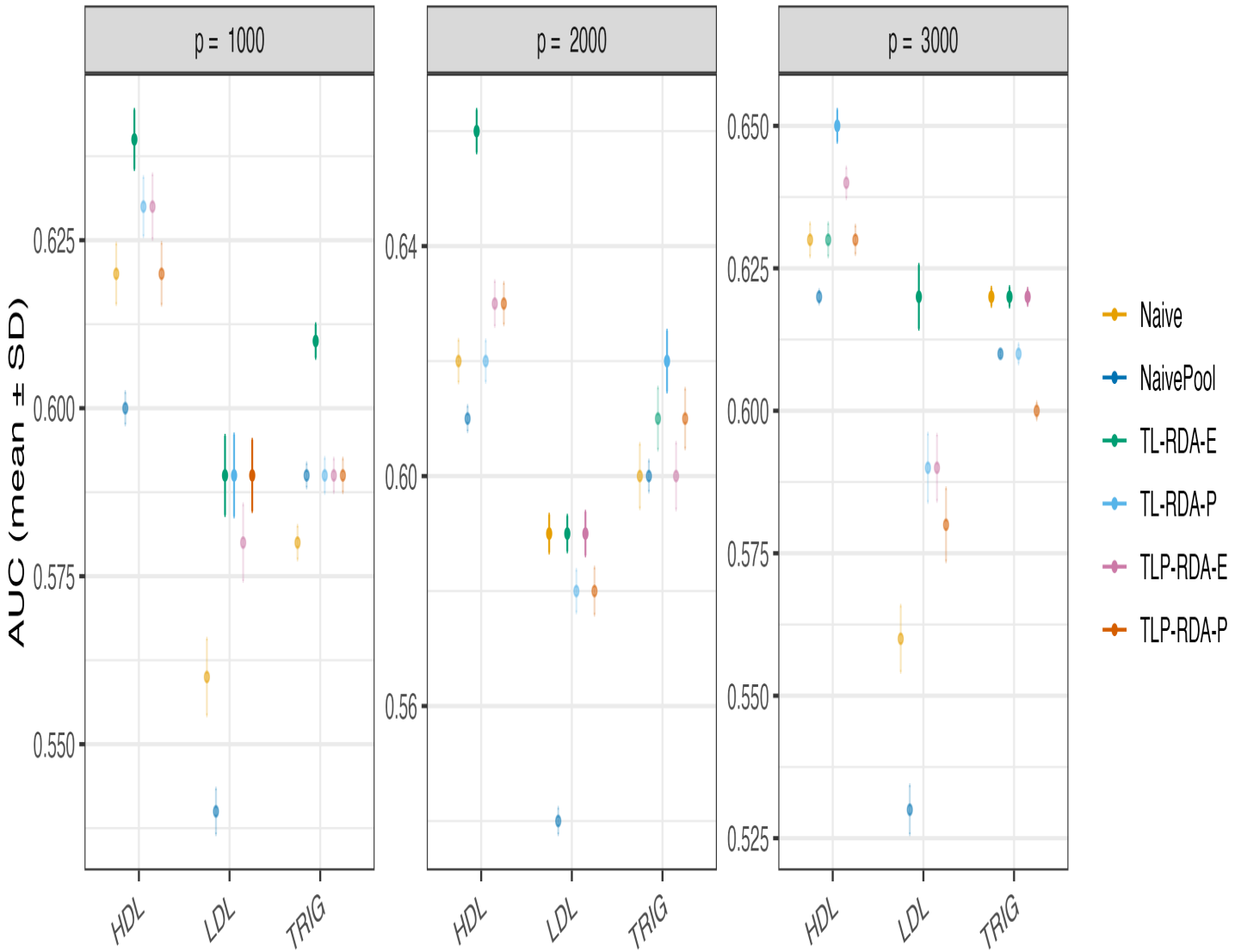


Final_Table_with_Total_Column

Site	City	ASCVD-free	ASCVD	Total
1	Baltimore, Maryland	230	42	272
2	Cleveland, Ohio	223	57	280
3	Ann Arbor and Detroit, Michigan	259	74	333
4	Chicago, Illinois	266	60	326
5	New Orleans, Louisiana	294	115	409
6	Philadelphia, Pennsylvania	125	55	180
7	Oakland, California	326	56	382

Lipid trait classification using genetic variants

$n_k = 3872$, $n_t = 1000$, no overlapping samples, cross-trait analysis



Conclusions

- Develop methods for RRC ridge regression and LDA based transfer learning (Trans-Ridge, Trans-LDA).
- Methods are adaptive to the informativeness of the auxiliary data via estimating ρ_{kl} and choosing the weights in RRC models.
- Methods outperform naive or simple sample-pooling methods, and also Trans-Lasso in PRS prediction when sample size is moderate.