Merging versus Ensembling in Multi-Study Machine Learning:

Theoretical Insight from Random Effects

Zoe Guan[1,2], Giovanni Parmigiani[1,2], and Prasad Patil[1,2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2]Department of Data Sciences, Dana-Farber Cancer Institute

Abstract

A critical decision point when training predictors using multiple studies is whether these studies should be combined or treated separately. We compare two multi-study learning approaches in the presence of potential heterogeneity in predictor-outcome relationships across datasets. We consider 1) merging all of the datasets and training a single learner, and 2) cross-study learning, which involves training a separate learner on each dataset and combining the resulting predictions. In a linear regression setting, we show analytically and confirm via simulation that merging yields lower prediction error than cross-study learning when the predictor-outcome relationships are relatively homogeneous across studies. However, as heterogeneity increases, there exists a transition point beyond which cross-study learning outperforms merging. We provide analytic expressions for the transition point in various scenarios and study asymptotic properties.

I. Introduction

Prediction and classification models trained on a single study often perform considerably worse in external validation than in cross-validation [1, 2]. Their generalizability is compromised by overfitting, but also by various sources of study heterogeneity, including differences in study design, data collection and measurement methods, unmeasured confounders, and study-specific sample characteristics [3]. Using multiple training studies can potentially address these challenges and lead to more replicable prediction models. In many settings, such as precision medicine, multi-study learning is motivated by systematic data sharing and data curation initiatives. For example, the establishment of gene expression databases such as Gene Expression Omnibus [4] and ArrayExpress [5] and neuro-imaging databases such as OpenNeuro [6] has facilitated access to sets of studies that provide comparable measurements of the same outcome and predictors (even if the original measurements are not be comparable, they can often be made comparable through preprocessing and normalization procedures [7, 8]). For problems where such a set of studies is available, it is important to systematically integrate information across datasets when building prediction and classification models.

One approach is to merge all of the datasets and treat the observations as if they are all from the same study (for example, see [9, 10]). The resulting increase in sample size can lead to improved training and better performance when the datasets are relatively homogeneous. Also, the merged

dataset is often representative of a broader reference population than any of the individual datasets. Xu et al. [9] showed that a prognostic test for breast cancer metastases developed from merged data performed better than the prognostic tests developed using individual studies. Zhou et al. [11] proposed hypothesis tests for determining when it is beneficial to pool data across multiple sites for linear regression, compared to using data from a single site.

Another approach is to combine results from separately trained models. Meta-analysis and ensembling both fall under this approach. Meta-analysis combines summary measures from multiple studies to increase statistical power (for example, see [12, 13]). A common combination strategy is to take a weighted average of the study-specific summary measures. In fixed effects meta-analysis, the weights are based on the assumption that there is a single true parameter underlying all of the studies, while in random effects meta-analysis, the weights are derived from a model where the true parameter varies across studies according to a probability distribution. When learners are indexed by a finite number of common parameters, meta-analysis applied to these parameters can be used for multi-study learning, with useful results [12]. Various studies have compared meta-analysis to merging. For effect size estimation, Bravata and Olkin [14] showed that merging heterogeneous datasets can lead to spurious results while meta-analysis protects against such problematic effects. Taminau et al. [15] and Kosch and Jung [16] found that merging had higher sensitivity than meta-analysis in gene expression analysis, while Lagani et al. [17] found that the two approaches performed comparably in reconstruction of gene interaction networks. Ensemble learning methods [18], which combine predictions from multiple models, can also be used to leverage information in a multi-study setting. By combining predictions, ensembling leads to lower variance and higher accuracy, and is applicable to more general classes of learners than meta-analysis. Patil and Parmigiani [19] proposed cross-study learning, defined as weighted average ensembles of prediction models trained on different studies, as an alternative to merging. They showed empirically that when the datasets are heterogeneous, cross-study learning can lead to improved generalizability and replicability compared to merging and meta-analysis.

In this paper, we provide theoretical guidelines for determining whether it is more beneficial to merge or to ensemble. We consider both low-dimensional and high-dimensional linear regression settings by studying merged and cross-study learners (CSLs) based on ordinary least squares (LS) and ridge regression. We hypothesize a mixed effects model for heterogeneity and show that merging has lower prediction error than cross-study learning when heterogeneity is low, but as heterogeneity increases, there exists a transition point beyond which cross-study learning outperforms merging. We characterize this transition point analytically and study it via simulations. We also compare merging and cross-study learning in practice, using microbiome data.

II. Problem Definition

We will use the following matrix notation: $\boldsymbol{I_N}$ is the $N \times N$ identity matrix, $\mathbf{0}_{N \times M}$ is an $N \times M$ matrix of 0's, $\mathbf{0}_N$ is a vector of 0's of length N, $tr(\boldsymbol{A})$ is the trace of matrix $\boldsymbol{A}$, $diag(\boldsymbol{u})$ is a diagonal matrix with $\boldsymbol{u}$ along its diagonal, and $(\boldsymbol{A})_{ij}$ is the entry in row i and column j of matrix $\boldsymbol{A}$. Other notation introduced throughout the paper is summarized in Table 1 of Appendix C.

Suppose we have K comparable studies that measure the same outcome and the same p features, and the datasets have been harmonized so that measurements across studies are on the same scale. For study k, let n_k denote the number of observations, $\boldsymbol{Y_k} \in \mathbb{R}^{n_k}$ the outcome vector, and $\boldsymbol{X_k} \in \mathbb{R}^{n_k \times p}$ the design matrix, where the first column of $\boldsymbol{X_k}$ is a vector of 1's if there is an

intercept. Assume the data are generated from the linear mixed effects model

$$\boldsymbol{Y_k} = \boldsymbol{X_k}\boldsymbol{\beta} + \boldsymbol{Z_k}\boldsymbol{\gamma_k} + \boldsymbol{\epsilon_k} \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of fixed effects, $\boldsymbol{Z_k} \in \mathbb{R}^{n_k \times q}$ is the design matrix for the random effects obtained by subsetting $\boldsymbol{X_k}$, $\boldsymbol{\gamma_k} \in \mathbb{R}^q$ is the vector of random effects with $E[\boldsymbol{\gamma_k}] = \boldsymbol{0_{n_k}}$ and $Cov(\boldsymbol{\beta_k}) = \boldsymbol{G} = diag(\sigma_1^2, \ldots, \sigma_q^2)$ where $\sigma_i^2 \geq 0$, $\boldsymbol{\epsilon_k} \in \mathbb{R}^{n_k}$ is the vector of residual errors with $E[\boldsymbol{\epsilon_k}] = \boldsymbol{0_{n_k}}$ and $Cov(\boldsymbol{\epsilon_k}) = \sigma_\epsilon^2 \boldsymbol{I_{n_k}}$, and $Cov(\boldsymbol{\gamma_k}, \boldsymbol{\epsilon_k}) = \boldsymbol{0_{n_k \times n_k}}$. For $j = 1, \ldots, q$, if $\sigma_j^2 > 0$, then the effect of the corresponding predictor differs across studies, and if $\sigma_j^2 = 0$, then the predictor has the same effect in each study.

The relationship between the predictors and the outcome in a given study can be seen as a perturbation of the population-level effect vector $\boldsymbol{\beta}$. The degree of heterogeneity in predictor-outcome relationships across studies can be summarized by the sum of the variances of the random effects divided by the number of fixed effects: $\overline{\sigma^2} = tr(\boldsymbol{G})/p$. We are interested in comparing the performance of two multi-study learning approaches as $\overline{\sigma^2}$ varies: 1) merging all of the studies and fitting a single linear regression model, and 2) fitting a linear regression model on each study and forming a CSL by taking a weighted average of the predictions.

Learners For low-dimensional settings where $p < n_k$ for all k, we consider merged and cross-study learners based on LS. The LS estimator of $\boldsymbol{\beta}$ based on the merged data is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{2}$$

where $\boldsymbol{Y} = (\boldsymbol{Y_1^T}, \ldots, \boldsymbol{Y_K^T})^T \in \mathbb{R}^{\sum_{k=1}^K n_k}$ and $\boldsymbol{X} = \left[\boldsymbol{X_1^T} | \ldots | \boldsymbol{X_K^T}\right]^T \in \mathbb{R}^{\sum_{k=1}^K n_k \times p}$. The LS estimator based on study k is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}} = (\boldsymbol{X_k^T}\boldsymbol{X_k})^{-1}\boldsymbol{X_k^T}\boldsymbol{Y_k} \tag{3}$$

and the LS cross-study estimator is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}} = \sum_{k=1}^K w_k \hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}} \tag{4}$$

where $\sum_{k=1}^K w_k = 1$.

For high-dimensional settings where $p > n_k$ for some k, we consider merged and cross-study learners based on ridge regression. When $n_k < p$, $\boldsymbol{X_k^T}\boldsymbol{X_k}$ is not invertible, so $\boldsymbol{\beta}$ is not estimable using LS. Ridge regression overcomes the non-invertibility problem [20], which can also arise in low-dimensional settings with highly correlated predictors, by penalizing the ℓ_2-norm of the coefficient vector. The predictors are typically standardized prior to fitting the model since ridge regression shrinks all predictors proportionally (for example, Horel and Kennard's seminal paper [20] assumes the predictors have mean 0 and variance 1). The coefficient estimates based on the standardized data are then transformed back to the original scale. Note that ridge regression is not scale invariant [21], so without loss of generality, we assume that the predictors are scaled but not centered prior to applying ridge regression. We first provide the form of the ridge regression estimators in the case where an intercept is included. Let the scaled versions of $\boldsymbol{X_k}$ and $\boldsymbol{X}$ be denoted by $\tilde{\boldsymbol{X}}_{\boldsymbol{k}} = \boldsymbol{X_k}\boldsymbol{S_k}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{S}$, where $\boldsymbol{S_k}, \boldsymbol{S} \in \mathbb{R}^{p \times p}$ are positive definite scaling matrices.

If scaling is not necessary or desirable (for example, if the predictors are measured in the same units), then set $\boldsymbol{S_k} = \boldsymbol{S} = \boldsymbol{I_p}$. Otherwise, let $\boldsymbol{S_k}$ be diagonal with $(\boldsymbol{S_k})_{11} = 1$ and $(\boldsymbol{S_k})_{jj}$ equal to the inverse standard deviation of column j of $\boldsymbol{X_k}$ for $j > 1$ and let $\boldsymbol{S}$ be diagonal with $(\boldsymbol{S})_{11} = 1$ and $(\boldsymbol{S})_{jj}$ equal to the inverse standard deviation of column j of $\boldsymbol{X}$ for $j > 1$. The merged ridge regression estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}} = \boldsymbol{S}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \lambda \boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y} \tag{5}$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I_p^-}\boldsymbol{S}^{-2})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{6}$$

where $\lambda \geq 0$ is the regularization parameter and $\boldsymbol{I_p^-}$ is obtained from $\boldsymbol{I_p}$ by setting $(\boldsymbol{I_p})_{11} = 0$, so that the intercept is not regularized [21]. The estimator of $\boldsymbol{\beta}$ from study k is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}} = \boldsymbol{S_k}(\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k + \lambda_k \boldsymbol{I_p^-})^{-1}\tilde{\tilde{\boldsymbol{X}}}_k^{\ T}\boldsymbol{Y_k} \tag{7}$$

$$= (\boldsymbol{X}_k^T\boldsymbol{X}_k + \lambda_k \boldsymbol{I_p^-}\boldsymbol{S}_k^{-2})^{-1}\boldsymbol{X}_k^T\boldsymbol{Y_k} \tag{8}$$

and the CSL estimator is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,C}} = \sum_{k=1}^{K} w_k \hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}} \tag{9}$$

If there is no intercept, then we set $(\boldsymbol{S_k})_{11}$ and $(\boldsymbol{S})_{11}$ to be the inverse standard deviations of the first columns of $\boldsymbol{X_k}$ and $\boldsymbol{X}$ respectively and replace $\boldsymbol{I_p^-}$ with $\boldsymbol{I_p}$ in the expressions above.

For simplicity, we assume the weights w_k and the regularization parameters λ and λ_k are predetermined. Note that for linear regression, averaging predictions across study-specific learners is equivalent to averaging the estimated coefficient vectors across study-specific learners and then computing predictions. Thus, cross-study learning based on linear regression is similar to meta-analysis of effect sizes. In particular, when $p = 1$, calculating $\hat{\beta}_{LS,C}$ is equivalent to performing a standard univariate meta-analysis. When $p > 1$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}$ weights each predictor equally in a given study while meta-analytic approaches, which involve either performing separate univariate meta-analyses for each predictor or performing a multivariate meta-analysis (for example, see [22, 23]), do not impose this constraint.

Performance Comparison Given a test set with design matrix $\boldsymbol{X_0}$ and outcome vector $\boldsymbol{Y_0}$, the goal is to identify conditions under which cross-study learning has lower mean squared prediction error (MSPE) than merging, i.e.

$$E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{\cdot,C}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{\cdot,M}}\|_2^2]$$

where the expectations are taken with respect to $\boldsymbol{Y_0}$ and $\|(u_1, \ldots, u_m)^T\|_2 = \sqrt{\sum_{i=1}^m u_i^2}$ is the ℓ_2 norm.

III. Results

We consider two cases for the structure of $\boldsymbol{G}$: equal variances and unequal variances. Let $\sigma^2_{(1)}, \ldots, \sigma^2_{(r)}$ be the distinct values on the diagonal of $\boldsymbol{G}$ and let m_j be the number of random effects with variance $\sigma^2_{(j)}$. In the equal variances case where $r = 1$ and $\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, q$, we

provide a necessary and sufficient condition for the CSL to outperform the merged learner. In the unequal variances case, we provide sufficient conditions under which the CSL outperforms the merged learner and vice versa. These conditions allow us to characterize a transition point in terms of $\overline{\sigma^2}$ between a regime that favors merging and a regime that favors cross-study learning.

In order to present the results more concisely, let $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{R_k} = \boldsymbol{X_k^T X_k}$, $\boldsymbol{M_k} = \boldsymbol{R_k} + \lambda_k \boldsymbol{I_p^-} \boldsymbol{S_k^{-2}}$, and $\boldsymbol{M} = \boldsymbol{R} + \lambda \boldsymbol{I_p^-} \boldsymbol{S^{-2}}$. Also, let $\boldsymbol{\Gamma_{(j)}} \in \mathbb{R}^{q \times m_j}$ be a matrix where $(\boldsymbol{\Gamma_{(j)}})_{il} = 1$ if random effect i is the lth random effect with variance $\sigma^2_{(j)}$ and $(\boldsymbol{\Gamma_{(j)}})_{il} = 0$ otherwise, so that $\boldsymbol{G\Gamma_{(j)}}$ subsets $\boldsymbol{G}$ to the columns corresponding to $\sigma^2_{(j)}$.

Proofs of the results are provided in Appendix A.

i. Least Squares

For the LS results, assume we are in a low-dimensional setting where $n_k > p$ for all k.

i.1 Equal Variances

Definition 1. *Define*

$$\tau_{LS} = \sigma^2_\epsilon \times \frac{q}{p} \times \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1}R_0}) - tr(\boldsymbol{R^{-1}R_0})}{tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{Z_0^T Z_0})} \tag{10}$$

Theorem 1.
Suppose the random effects have equal variances ($\sigma_j^2 = \sigma^2$ for $j = 1, \ldots, q$) and

$$tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{Z_0^T Z_0}) > 0 \tag{11}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2]$ if and only if $\overline{\sigma^2} \geq \tau_{LS}$.

By Theorem 1, for any fixed weighting scheme that does not depend on $\boldsymbol{G}$, satisfies Equation 11, and leads to in $\tau_{LS} > 0$, τ_{LS} represents a transition point from a regime where merging outperforms cross-study learning to a regime where cross-study learning outperforms merging. When equal weights are used and $\boldsymbol{R_k}$ is not identical for all k, it follows from Jensen's operator inequality [24] that Equation 11 holds and the numerator of τ_{LS} is positive, so the transition point always exists.

Corollary 1.1.
Suppose the random effects have equal variances and there exist positive definite matrices $\boldsymbol{A_1}, \boldsymbol{A_2}, \boldsymbol{A_3} \in \mathbb{R}^{p \times p}$ such that as $K \to \infty$,

1. $\frac{1}{K} \sum_{k=1}^K \boldsymbol{R_k} \to \boldsymbol{A_1}$
2. $\frac{1}{K} \sum_{k=1}^K \boldsymbol{R_k^{-1}} \to \boldsymbol{A_2}$
3. $\frac{1}{K} \sum_{k=1}^K \boldsymbol{X_k^T Z_k Z_k^T X_k} \to \boldsymbol{A_3}$
4. $tr(\boldsymbol{A_1^{-1} A_3 A_1^{-1} R_0}) - tr(\boldsymbol{Z_0^T Z_0}) > 0$

where $\rightarrow$ *denotes almost sure convergence. If we set* $w_k = \frac{1}{K}$*, then*

$$\tau_{LS} \rightarrow \sigma_\epsilon^2 \times \frac{q}{p} \times \frac{tr(\boldsymbol{A_2 R_0}) - tr(\boldsymbol{A_1^{-1} R_0})}{tr(\boldsymbol{A_1^{-1} A_3 A_1^{-1} R_0}) - tr(\boldsymbol{Z_0^T Z_0})} \tag{12}$$

For example, suppose all study sizes are equal to n, the predictors are independent and identically distributed within and across studies, and $E[\boldsymbol{R_k}]$, $E[\boldsymbol{R_k^{-1}}]$, $E[\boldsymbol{X_k^T Z_k Z_k^T X_k}] \in \mathbb{R}^{p \times p}$ are positive definite. Then Corollary 1.1 applies with $\boldsymbol{A_1} = E[\boldsymbol{R_k}]$, $\boldsymbol{A_2} = E[\boldsymbol{R_k^{-1}}]$, and $\boldsymbol{A_3} = E[\boldsymbol{X_k^T Z_k Z_k^T X_k}]$. In the special case where $p = q = 1$ and the predictor follows $N(0, v)$, the limit becomes

$$\frac{\sigma_\epsilon^2}{(n-2)v} \tag{13}$$

and the asymptotic transition point is controlled simply by the variance of the residuals, the variance of the predictor, and the study sample size.

i.2 Unequal Variances

Definition 2. *Define*

$$\tau_{LS,1} = \frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1} R_0}) - tr(\boldsymbol{R^{-1} R_0})}{\max\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{14}$$

$$\tau_{LS,2} = \frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1} R_0}) - tr(\boldsymbol{R^{-1} R_0})}{\min\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{15}$$

Theorem 2.

(a) Suppose

$$\max_{j=1,\ldots,r} tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}) > 0 \tag{16}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{LS,C}}\|_2^2] \geq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{LS,M}}\|_2^2]$ *when* $\overline{\sigma^2} \leq \tau_{LS,1}$.

(b) Suppose

$$\min_{j=1,\ldots,r} tr(\boldsymbol{R^{-1}} \sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1} R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}) > 0 \tag{17}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{LS,C}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0 \hat{\beta}_{LS,M}}\|_2^2]$ *when* $\overline{\sigma^2} \geq \tau_{LS,2}$.

Lemma 2.1.

Suppose there exist positive definite matrices $\boldsymbol{A_1}, \boldsymbol{A_2}, \boldsymbol{A_{(j)}} \in \mathbb{R}^{p \times p}$ *such that as* $K \rightarrow \infty$*,*

1. $\frac{1}{K} \sum_{k=1}^K \boldsymbol{R_k} \rightarrow \boldsymbol{A_1}$

2. $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{R_k^{-1}} \to \boldsymbol{A_2}$
3. $\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k} \to \boldsymbol{A_{(j)}}$ *for* $j = 1, \ldots, r$

and we set $w_k = \frac{1}{K}$.

(a) If

$$\max_{j=1,\ldots,r} (tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}})) > 0 \tag{18}$$

then

$$\tau_{LS,1} \to \frac{\sigma_\epsilon^2}{p} \times \frac{tr(\boldsymbol{A_2 R_0}) - tr(\boldsymbol{A_1^{-1} R_0})}{\max_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{19}$$

(b) If

$$\min_{j=1,\ldots,r} (tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}})) > 0 \tag{20}$$

then

$$\tau_{LS,2} \to \frac{\sigma_\epsilon^2}{p} \times \frac{tr(\boldsymbol{A_2 R_0}) - tr(\boldsymbol{A_1^{-1} R_0})}{\min_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{A_1^{-1} A_{(j)} A_1^{-1} R_0}) - tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))} \tag{21}$$

In the unequal variances scenario, we can establish a transition interval such that the merged learner outperforms the CSL when $\overline{\sigma^2}$ is smaller than the lower bound of the interval and the CSL outperforms the merged learner when $\overline{\sigma^2}$ is greater than the upper bound of the interval. Note that the conditions and results for the equal variances scenario are special cases of the conditions and results for the unequal variances scenario. When $r = 1$, we have $m_1 = q$, $\boldsymbol{\Gamma_{(1)}} = \boldsymbol{I_q}$, and $\tau_{LS,1} = \tau_{LS,2} = \tau_{LS}$.

i.3 Optimal Weights

It can be shown (see Appendix A) that the optimal weights for the CSL are given by

$$w_k = \frac{(tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 tr(\boldsymbol{R_k^{-1} R_0}))^{-1}}{\sum_{k=1}^{K}(tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 tr(\boldsymbol{R_k^{-1} R_0}))^{-1}} \tag{22}$$

where the weight for study k is proportional to the inverse MSPE of the LS learner trained on that study.

In the equal variances setting, $tr(\boldsymbol{G Z_0^T Z_0}) = \sigma^2 tr(\boldsymbol{Z_0^T Z_0})$. We saw previously that when the weights satisfy Equation 11 and do not depend on σ^2, τ_{LS} characterizes the value of $\overline{\sigma^2}$ beyond which cross-study learning outperforms merging. The optimal weights depend on σ^2, so τ_{LS} depends on σ^2 under the optimal weighting scheme. Thus, it is difficult to obtain a closed-form expression for the transition point, though numerical methods can be used to solve for the value of $\overline{\sigma^2}$ such that $\overline{\sigma^2} = \tau_{LS}$. In Appendix A, we provide a closed-form approximation of the transition point. Note that the transition point under any fixed weighting scheme provides an upper bound for the transition point under the optimal weighting scheme. We also remark that in the special case where $p = 1$, the optimally weighted CSL has the same variance as the estimator from the true mixed effects model. If σ^2 and σ_ϵ^2 are known, then the CSL will always be at least as efficient as the merged learner when $p = 1$, but in practice, σ^2 and σ_ϵ^2 need to be estimated.

ii. Ridge Regression

Below, we present results for ridge regression that are applicable to both low- and high-dimensional settings.

ii.1 Equal Variances

Definition 3. *Define*

$$\boldsymbol{b}_{R,C} = Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,C}) = -\sum_k w_k \lambda_k \boldsymbol{X_0} \boldsymbol{M}_k^{-1} \boldsymbol{I}_p^- \boldsymbol{S}_k^{-2} \boldsymbol{\beta} \tag{23}$$

$$\boldsymbol{b}_{R,M} = Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,M}) = -\lambda \boldsymbol{X_0} \boldsymbol{M}^{-1} \boldsymbol{I}_p^- \boldsymbol{S}^{-2} \boldsymbol{\beta} \tag{24}$$

$$\tau_R = \frac{q}{p} \times \frac{\sigma_\epsilon^2 \left(\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{R}_k \boldsymbol{M}_k^{-1} \mathbf{R_0}) - tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{R}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) \right) + \boldsymbol{b}_{R,C}^T \boldsymbol{b}_{R,C} - \boldsymbol{b}_{R,M}^T \boldsymbol{b}_{R,M}}{tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}_k^{-1} \mathbf{R_0})} \tag{25}$$

Theorem 3.
Suppose the random effects have equal variances and

$$tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}_k^{-1} \boldsymbol{R_0}) > 0 \tag{26}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,C}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,M}\|_2^2]$ *if and only if* $\overline{\sigma^2} \geq \tau_R$.

ii.2 Unequal Variances

Definition 4. *Define*

$$\tau_{R,1} = \frac{\sigma_\epsilon^2 \left(\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{R}_k \boldsymbol{M}_k^{-1} \mathbf{R_0}) - tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{R}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) \right) + \boldsymbol{b}_{R,C}^T \boldsymbol{b}_{R,C} - \boldsymbol{b}_{R,M}^T \boldsymbol{b}_{R,M}}{p \max\limits_{j=1,\ldots,r} \frac{1}{m_j} (tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}_k^{-1} \boldsymbol{R_0}))} \tag{27}$$

$$\tau_{R,2} = \frac{\sigma_\epsilon^2 \left(\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{R}_k \boldsymbol{M}_k^{-1} \mathbf{R_0}) - tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{R}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) \right) + \boldsymbol{b}_{R,C}^T \boldsymbol{b}_{R,C} - \boldsymbol{b}_{R,M}^T \boldsymbol{b}_{R,M}}{p \min\limits_{j=1,\ldots,r} \frac{1}{m_j} (tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}_k^{-1} \boldsymbol{R_0}))} \tag{28}$$

Theorem 4.

(a) Suppose

$$\max_{j=1,\ldots,r} tr(\boldsymbol{M}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}^{-1} \boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1} \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{\Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{M}_k^{-1} \boldsymbol{R_0}) > 0 \tag{29}$$

Then $E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,C}\|_2^2] \geq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{R,M}\|_2^2]$ *when* $\overline{\sigma^2} \leq \tau_{R,1}$.

(b) Suppose

$$\min_{j=1,\ldots,r} tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) - \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\boldsymbol{R}_0) > 0 \tag{30}$$

Then $E[\|\boldsymbol{Y}_0 - \boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,C}\|_2^2] \leq E[\|\boldsymbol{Y}_0 - \boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,M}\|_2^2]$ *when* $\overline{\sigma^2} \geq \tau_{R,2}$.

Again, the conditions and results for the equal variances scenario are special cases of the conditions and results for the unequal variances scenario.

iii. Interpretation

The covariance matrices of linear regression coefficient estimators can be written as a sum of two components, one driven by between-study variability and one driven by within-study variability. For example, for LS we have

$$Cov(\hat{\boldsymbol{\beta}}_{LS,C}) = \sum_{k=1}^K w_k^2\boldsymbol{\Gamma}\boldsymbol{G}\boldsymbol{\Gamma}^T + \sigma_\epsilon^2\sum_{k=1}^K w_k^2(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times q}$ is the matrix such that $\boldsymbol{Z}_k = \boldsymbol{X}_k\boldsymbol{\Gamma}$, and

$$Cov(\hat{\boldsymbol{\beta}}_{LS,M}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_{k=1}^K[\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k](\boldsymbol{X}^T\boldsymbol{X})^{-1} + \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Since the merged learner ignores between-study heterogeneity, the trace of its first component is generally larger than that of the CSL. However, since the merged learner is trained on a larger sample, the trace of its second component is generally smaller than that of the CSL. The merged and cross-study learners based on LS are unbiased, so the transition point depends on the trade-off between these two components. When $p = 1$, Expression 13 shows that having a higher-variance predictor favors cross-study learning over merging, since increasing the variance of the predictor amplifies the impact of the random effect.

Unlike LS estimators, ridge regression estimators are biased as a result of regularization. The transition point for ridge regression depends on the regularization parameters used on the merged and individual datasets. It also depends on the true coefficient vector $\boldsymbol{\beta}$ through the squared bias terms in the MSPEs of the merged and cross-study learners, so an estimate of $\boldsymbol{\beta}$ is needed to compute the expressions in Theorems 3 and 4. These expressions can vary considerably for different choices of regularization parameters and different values of $\boldsymbol{\beta}$. We did not provide the asymptotic results for ridge regression as $K \to \infty$ (with the study sizes held constant) because this scenario is not entirely fair to the CSL. For $p > n_k$ and sufficiently large K, the merged learner will be in the low-dimensional setting while the CSL will remain in the high-dimensional setting. As $K \to \infty$, the bias term approaches 0 for the merged learner (assuming $\lambda/K \to 0$) but not for the CSL, which suggests that when K is sufficiently large, merging will always yield lower MSPE than cross-study learning. Also, due to the squared bias term in the MSPE, it is not straightforward to find optimal CSL weights for ridge regression.

In general, the transition points for LS and ridge regression depend on the design matrix of the test set. However, the test design matrix drops out when it is a scalar multiple of an orthogonal matrix. For example, this occurs when $p = 1$.

IV. Simulations

We conducted simulations to verify the theoretical results for LS and ridge regression and to compare them to the empirical transition points for three methods for which we could not find a closed-form solution: LASSO, single hidden layer neural network, and random forest. We also made performance comparisons with a linear mixed effects model, univariate random effects meta-analyses, and multivariate random effects meta-analysis. We used the R packages `glmnet, nnet, randomForest, nlme, metafor`, and `mvmeta` for ridge regression/LASSO, neural networks, random forests, linear mixed effects models, univariate meta-analyses, and multivariate meta-analyses respectively.

We considered four simulation scenarios corresponding to the settings in Theorems 1, 2, 3, and 4. We used 4 training studies and 4 test studies of size 40 for all scenarios. We set $\sigma_\epsilon^2 = 1$. For the low-dimensional settings, we set $p = 10$, $q = 5$, and generated 5 $\boldsymbol{\beta}$'s from $N(0, 1)$ and 5 from $N(0, 0.01^2)$, with 5 of the $\boldsymbol{\beta}$'s having random slopes. For the high-dimensional settings, we set $p = 100$, $q = 10$, and generated 30 $\boldsymbol{\beta}$'s from $N(0, 1)$ and 70 from $N(0, 0.01^2)$, with 10 of the $\boldsymbol{\beta}$'s having random slopes. For each simulation scenario, we fixed the predictor values in the training and test sets and the model hyperparameters. Predictor values were sampled from datasets in the `curatedOvarianData` R package [25]. Model hyperparameters were tuned once using 5-fold cross-validation with outcomes generated under $\sigma^2 = 0$. For various values of $\overline{\sigma^2}$, including 0 and the theoretical transition point, we generated random slopes, residual errors, and outcomes for each training and test study according to Model (1), then trained and tested the following approaches: linear mixed effects model, random effects meta-analysis of univariate LS estimates, random effects multivariate meta-analysis, and merged learners and CSLs based on LS, ridge regression, LASSO, neural networks, and random forests. For ridge regression and LASSO, the predictors were standardized prior to model fitting. For linear mixed effects, we fit the true model, using restricted maximum likelihood to estimate $\boldsymbol{G}$. For meta-analysis of univariate LS estimates, we used the DerSimonian and Laird method. For multivariate meta-analysis, we used restricted maximum likelihood to estimate $\boldsymbol{G}$, constraining the covariance matrix to be diagonal. LS, linear mixed effects, and meta-analysis were only applied in the low-dimensional setting. We performed 1000 replicates for each value of $\overline{\sigma^2}$ and estimated the MSPE of each estimator by averaging the squared error across replicates.

As seen in Figures 1 and 2, the empirical transition points for LS and ridge regression agree with the theoretical results from Theorems 1 and 3 (similar figures for Theorems 2 and 4 are provided in Appendix B). The methods all have similar empirical transition points except for random forest, which performed considerably worse than all of the other approaches (see Figure 8 in Appendix). The poor performance of random forest could be because the data were generated from a linear model. The univariate meta-analysis approach also performed poorly 8, which is unsurprising because the generating model is a multivariate model. The performance of the other models relative to the data generating model is summarized in Figure 3 for three values of $\overline{\sigma^2}$. When $\overline{\sigma^2} = 0$, the merged regression learners and multivariate meta-analysis perform as well as or slightly better than the mixed effects model and outperform the CSLs. The merged neural network learner does slightly worse than the regression learners. At the LS transition point, all models perform similarly. Beyond the transition point, the models continue to perform similarly (when heterogeneity is high, all models perform poorly), with the CSLs slightly outperforming the merged learners and multivariate meta-analysis performing as well the mixed effects model. For each of the three values of $\overline{\sigma^2}$, LASSO performed best, even slightly outperforming the mixed

effects model and multivariate meta-analysis. This is likely because several of the true β's were close to 0.

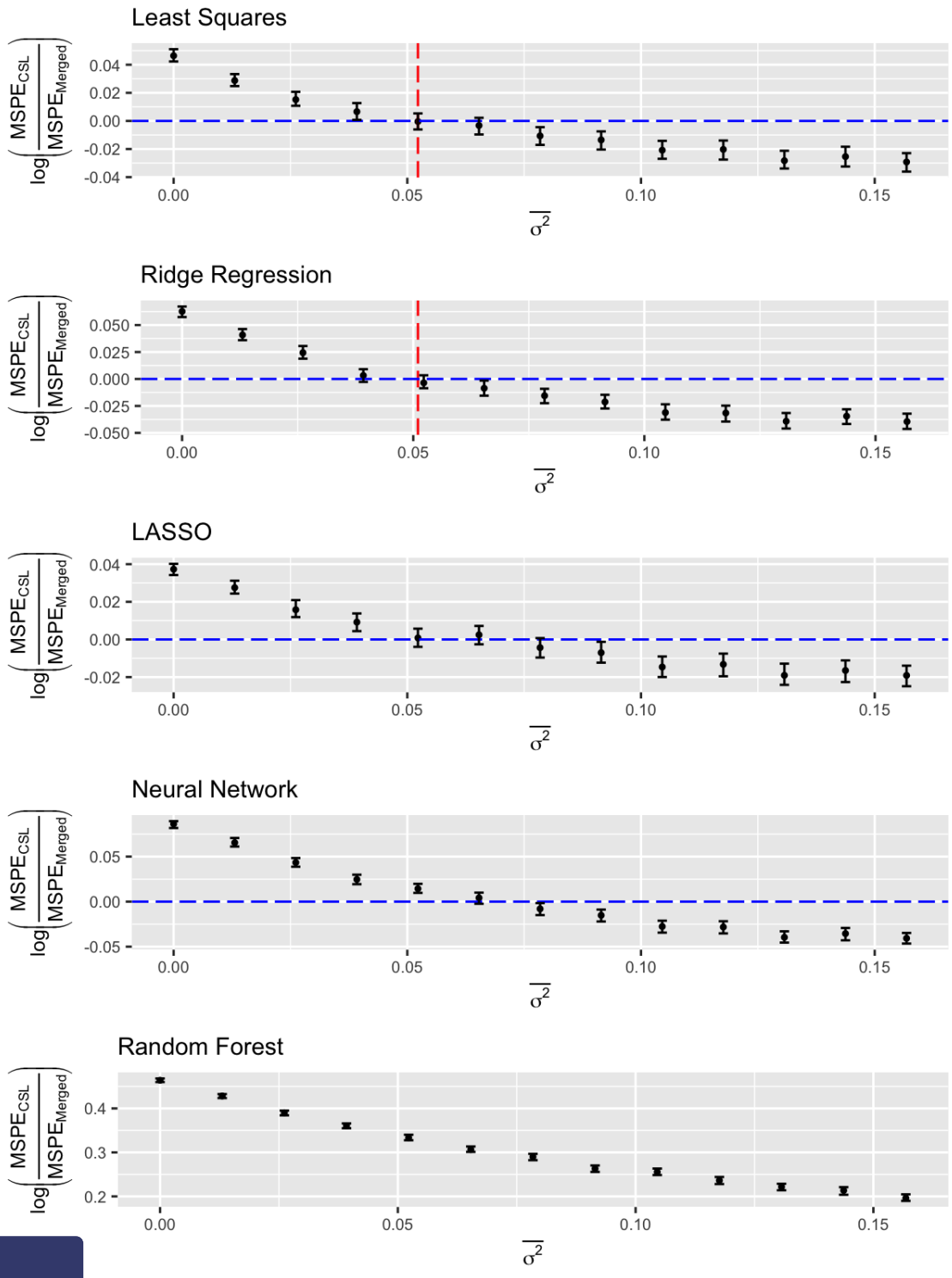

Figure 1: 4 training studies of size 40 with 10 predictors. The red lines correspond to the theoretical transition points from Theorems 1 and 3.

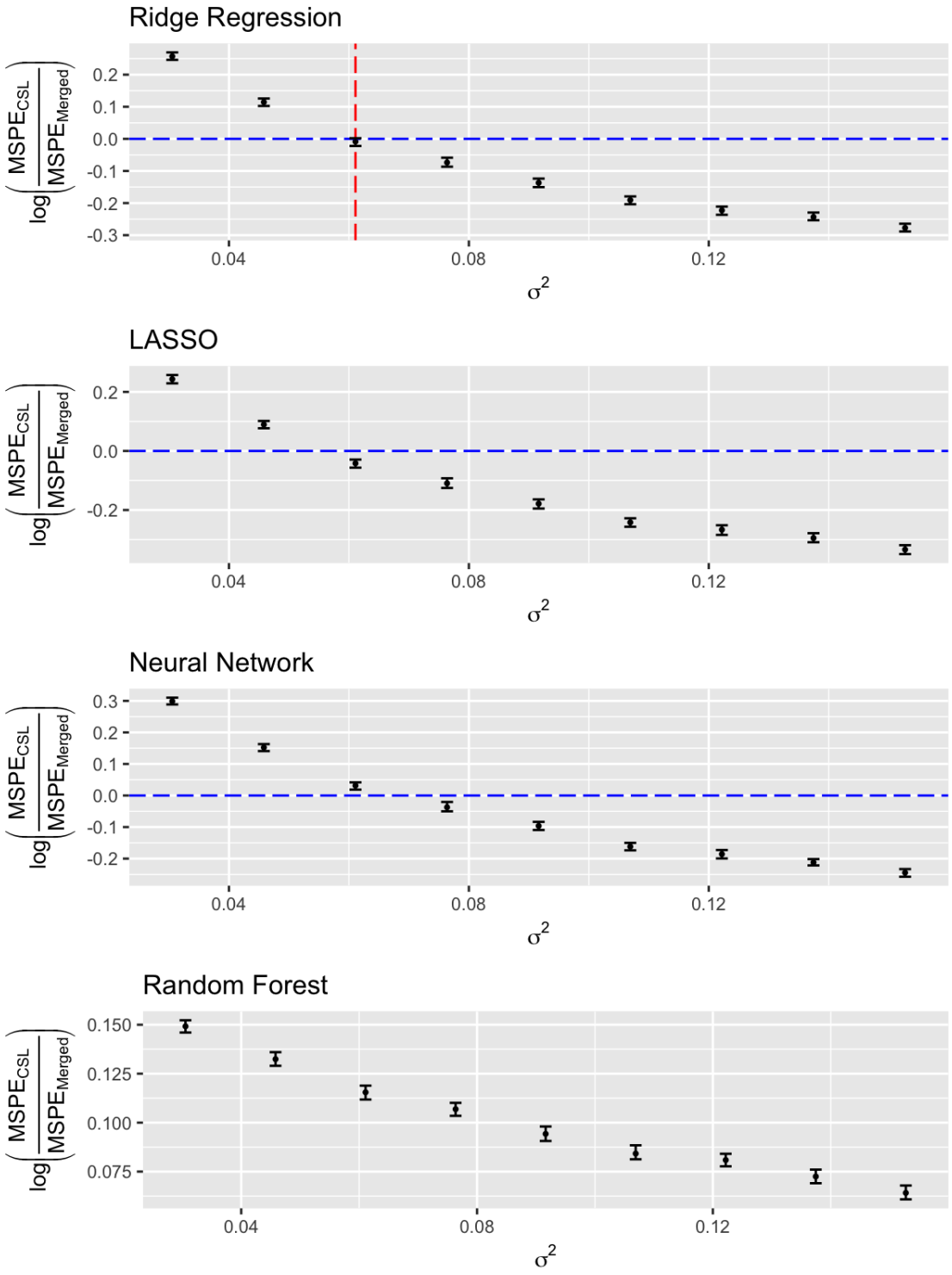

Figure 2: 4 training studies of size 40 with 100 predictors. The red line corresponds to the theoretical transition point from Theorem 3.

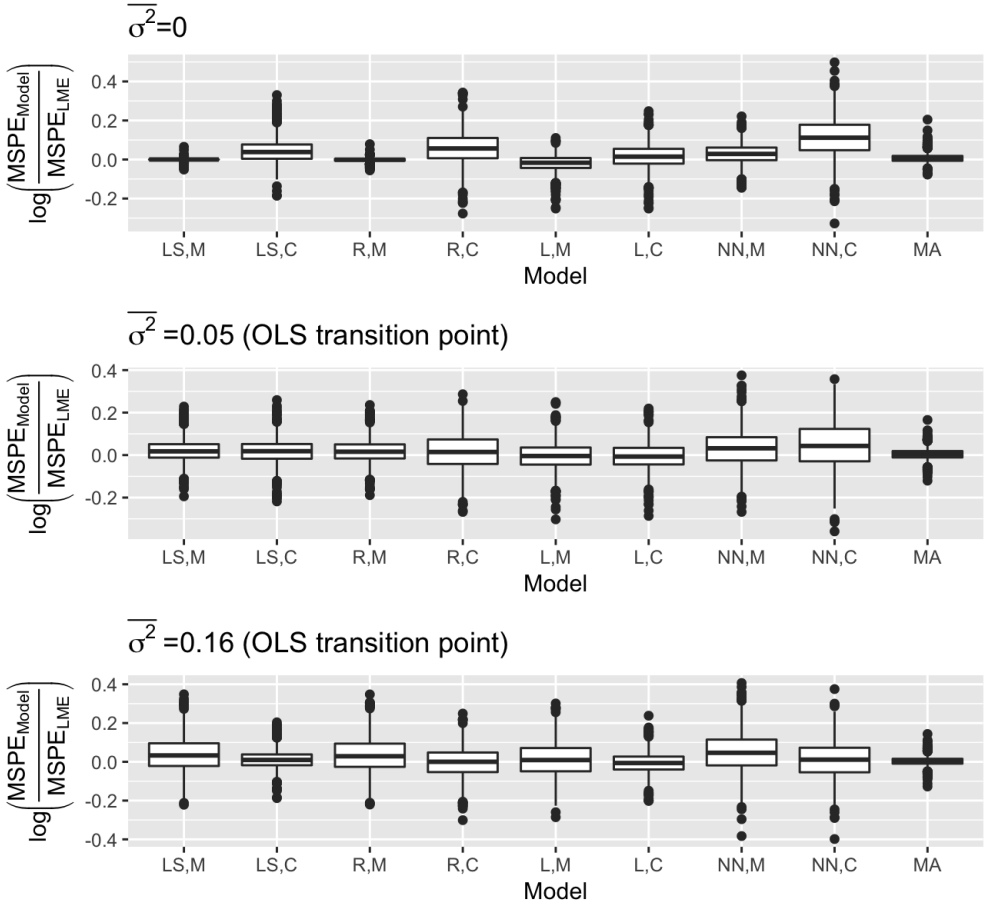

Figure 3: Performance comparisons for three values of $\overline{\sigma^2}$ (the random forest and univariate meta-analysis results are omitted to avoid stretching the y-axis but are provided in Appendix B). LS,M: merged LS learner; LS,C: CSL based on LS; R,M: merged ridge regression learner; R,C: CSL based on ridge regression; L,M: merged LASSO learner; L,C: CSL based on LASSO; NN,M: merged neural network; NN,C: CSL based on neural networks; MA: multivariate meta-analysis.

V. Metagenomics Application

To illustrate in a practical example, we compared the performance of merging and cross-study learning used datasets from the `curatedMetagenomicData` R package [26], which contains a collection of curated, uniformly processed human microbiome data. We focused on three gut microbiome studies that measured cholesterol as well as gene marker abundance in stool, restricting to samples from female patients: 1) Qin et al.'s 2012 study of Chinese type 2 diabetes patients and non-diabetic controls ($n_1 = 151$ samples from independent female patients) [27], 2) Karlsson et al.'s 2013 study of middle-aged European women with normal, impaired or diabetic glucose control ($n_2 = 145$ samples from independent female patients) [28], and 3) Heintz-Buschart et al's 2016 study of patients with a family history of type 1 diabetes ($n_0 = 32$ samples from 13 female patients) [29]. We used merging and cross-study learning to train linear regression models to predict cholesterol, calculated the theoretical transition interval, and evaluated the performance of the two approaches.

We considered two scenarios: 1) training on different subsets of the same study and testing on a held out subset, and 2) training on different studies and testing on an independent study. In the first scenario, we randomly split the Qin et al. 2012 samples into five datasets of approximately equal size, using four for training and the remaining one for testing. We used age and the top five marker abundances most correlated with the outcome in the training set as the predictors. In second scenario, we used the Qin et al. 2012 and Karlsson et al. 2013 datasets for training and the Heintz-Buschart et al. 2016 dataset for testing. We used age and the top twenty marker abundances most correlated with the outcome in the training set as the predictors. In each scenario, we fit merged and CSL versions of LS and ridge regression. We estimated $\boldsymbol{G}$ by fitting a linear mixed effects model using residual maximum likelihood, allowing each predictor to have a random effect. For the CSLs, we used the optimal weights given by Equation 22, plugging in the estimate of $\boldsymbol{G}$. We calculated the theoretical transition bounds from Theorems 2 and 4 and compared them to the estimate of $\overline{\sigma^2}$. We evaluated the performance of the models empirically by calculating the prediction error on the test set.

In the first scenario $\overline{\sigma^2}$ was estimated to be 0.0047^2, $\tau_{LS,1}$ was 0.1018^2, and $\tau_{R,1}$ was 0.0604^2, suggesting that merging was expected outperform cross-study learning. In the test set, the merged versions of LS and ridge regression both had lower prediction error than the respective CSL versions (Figure 4). In the second scenario, $\overline{\sigma^2}$ was estimated to be 18.32^2, $\tau_{LS,2}$ was 15.25^2, and $\tau_{R,2}$ was 4.42^2. In the test set, the CSL versions of LS and ridge regression both had lower prediction error than the respective merged versions (Figure 5).

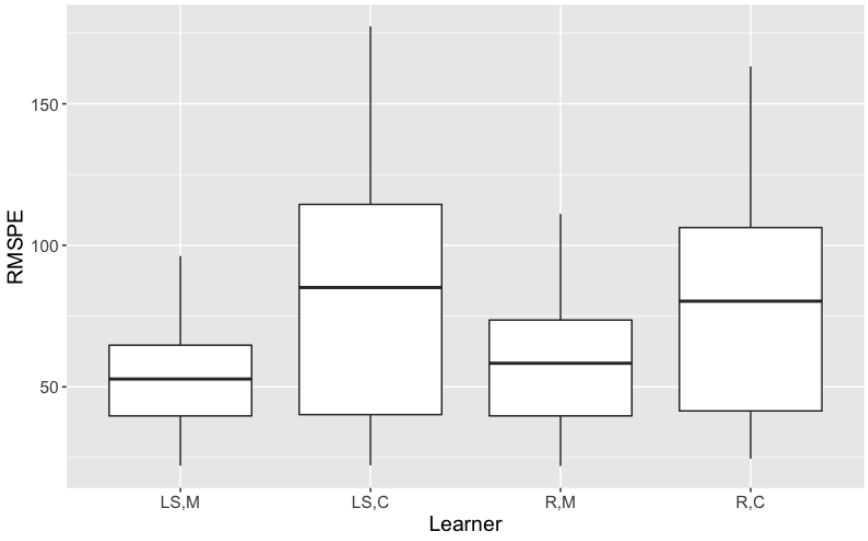

Figure 4: Root mean square prediction error (RMSPE) for the first scenario with bootstrap confidence intervals. LS,M: merged LS learner; LS,C: CSL based on LS; R,M: merged ridge regression learner; R,C: CSL based on ridge regression.

VI. Discussion

The availability of large and increasingly heterogeneous collections of data for training classifiers is challenging traditional approaches for training and validating prediction and classification algorithms. At the same time, it is also opening opportunities for new and more general paradigms.

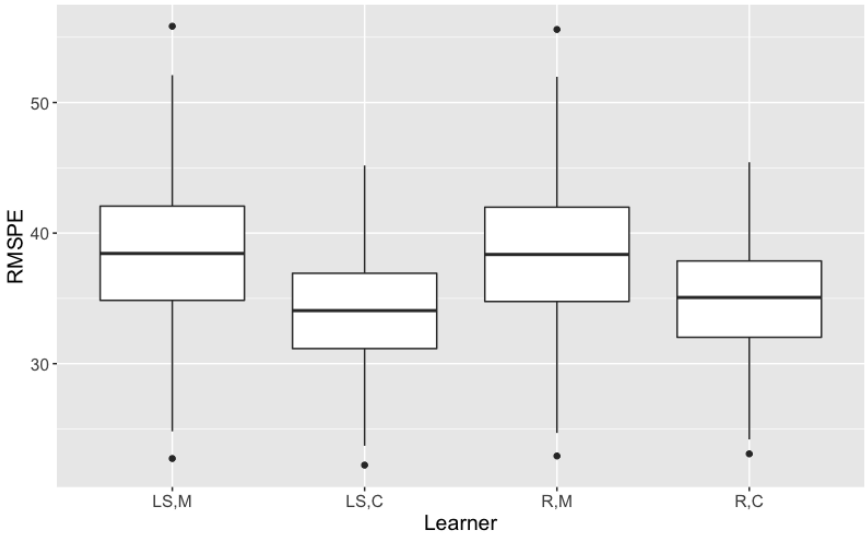

Figure 5: Root mean square prediction error (RMSPE) for the second scenario with bootstrap confidence intervals. LS,M: merged LS learner; LS,C: CSL based on LS; R,M: merged ridge regression learner; R,C: CSL based on ridge regression.

One of these is cross-study machine learning via CSLs, motivated by variation in the relation between predictors and outcomes across collections of similar studies. A natural benchmark for these methods is to combine all training studies, to exploit the power of larger training sample sizes. In previous work [19], *merged learners* perform better than CSLs in low-heterogeneity settings. As heterogeneity increases, however, our earlier simulations indicated a "transition point" in the heterogeneity scale beyond which acknowledging cross-study heterogeneity becomes preferable, and the CSLs outperform the merged learners.

In this paper, we approached this problem analytically for the first time by characterizing cross-study heterogeneity using a linear mixed effects model. We derived closed-form transition points for standard and ridge-regularized linear regression models. We confirmed the analytic results in simulation and demonstrated that when the data are generated by a linear model, the LS and ridge regression solutions can serve as proxies for the transition point under other learning strategies (LASSO, neural network) for which closed-form derivation is difficult. Finally, we estimated the transition point in cases of low and high cross-study heterogeneity in microbiome data and showed how it can be used as a guide for deciding when and when not to merge studies together in the course of learning a prediction rule.

We focused on deriving analytic results for LS and ridge regression because of the opportunity to pursue closed-form solutions. Other widely used methods such as LASSO, neural networks, and random forests are not as easily amenable to a closed-form solution, so we used simulations to study the performance of merging versus ensembling for these methods. In our simulation settings, the merged learners based on LS, ridge regression, LASSO, and neural networks had comparable accuracy, as did the corresponding CSLs. The methods all had similar empirical transition points, perhaps as a consequence of their similar performance. An exception is random forest, which did not reach a transition point within the specified heterogeneity levels, and overall performed less well in general, as is expected in data generated by linear models. The analytic results for LS/ridge regression could potentially serve as an approximation for other

methods that perform comparably, though it is important to consider how the reliability of such an approximation could be affected by the nature of the data and choice of model hyperparameters.

In practice, the analytic transition point and transition interval expressions could be used to help guide decisions about whether to merge data from multiple studies when there is potential heterogeneity in predictor-outcome relationships across the study populations. $\overline{\sigma^2}$ can be estimated from the training data and compared to the theoretical transition points or bounds for LS and/or ridge regression. Various methods can be used to estimate $\overline{\sigma^2}$, including maximum likelihood and method of moments-based approaches used in meta-analysis (for example, see [30]), with the caveat that estimates will be imprecise when the number of studies is small.

Under Model (1), fitting a correctly specified mixed effects model will generally be more efficient than both the merged and cross-study versions of LS. However, more flexible machine learning algorithms can potentially yield better prediction accuracy than the true model. For example, in the the low-dimensional simulations, the mixed effects model was outperformed by either the merged learner or CSL based on LASSO for most levels of heterogeneity. Moreover, fitting a mixed effects model can be computationally difficult when the number of predictors is large and standard mixed effects models are not appropriate for high-dimensional data, though there are methods for penalized mixed effects models [31, 32, 33, 34].

A limitation of our derivations is that they treat the following quantities as known: the subset of predictors with random effects, the CSL weights, and the regularization parameters for ridge regression. In practice, these are usually selected using statistical procedures that introduce additional variability. Furthermore, we obtained closed-form transition point expressions for cases where the CSL weighting scheme does not depend on the variances of the random effects. Such weighting schemes are generally be sub-optimal (for example, the optimal weights for LS given by Equation 22 depend on $\boldsymbol{G}$), so the closed-form results are based on a conservative estimate of the maximal performance of cross-study learning. Another limitation is the assumption that the random effects are uncorrelated, which is often not true in practice.

In summary, although this work is predicated upon the assumption that cross-study heterogeneity manifests through random effects and assumes that weights and regularization parameters are known, we believe it provides a theoretical rationale for multi-study machine learning, and a strong foundation for developing practical rules and guidelines to implement it.

VII. Reproducibility

Code to reproduce the simulations and data application is available at
`https://github.com/zoeguan/transition_point`

VIII. Acknowledgements

Work supported by NIH grants 4P30CA006516-51 (Parmigiani) and 2T32CA009337-36 (Patil), NSERC PGS-D Scholarship (Guan) and NSF grant DMS-1810829 (Patil and Parmigiani). We thank Lorenzo Trippa and Boyu Ren for useful discussions.

References

[1] P. J. Castaldi, I. J. Dahabreh, and J. P. Ioannidis. An empirical assessment of validation practices for molecular classifiers. *Briefings in bioinformatics*, 12(3):189–202, 2011.

[2] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, Jun 2014. PMID: 24931973.

[3] Y. Zhang, C. Bernau, G. Parmigiani, and L. Waldron. The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics*, page kxy044, 2018.

[4] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[5] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research*, 39(suppl_1):D1002–D1004, 2010.

[6] K. Gorgolewski, O. Esteban, G. Schaefer, B. Wandell, and R. Poldrack. OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. *Organization for Human Brain Mapping. Vancouver, Canada*, page 1677, 2017.

[7] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, 14(4):469–490, 2012.

[8] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105–114, 2004.

[9] L. Xu, A. C. Tan, R. L. Winslow, and D. Geman. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC bioinformatics*, 9(1):125, 2008.

[10] Y. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, C.-J. Tsai, and S. Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC bioinformatics*, 5(1):81, 2004.

[11] H. H. Zhou, Y. Zhang, V. K. Ithapu, S. C. Johnson, G. Wahba, and V. Singh. When can Multi-Site Datasets be Pooled for Regression? Hypothesis Tests, l_2-consistency and Neuroscience Applications. *arXiv preprint arXiv:1709.00640*, 2017.

[12] M. Riester, J. M. Taylor, A. Feifer, T. Koppie, J. E. Rosenberg, R. J. Downey, B. H. Bochner, and F. Michor. Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clinical Cancer Research*, 18(5):1323–1333, March 2012.

[13] G. C. Tseng, D. Ghosh, and E. Feingold. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799, 2012.

[14] D. M. Bravata and I. Olkin. Simple pooling versus combining in meta-analysis. *Evaluation & the health professions*, 24(2):218–230, 2001.

[15] J. Taminau, C. Lazar, S. Meganck, and A. Nowé. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN bioinformatics*, 2014, 2014.

[16] R. Kosch and K. Jung. Conducting gene set tests in meta-analyses of transcriptome expression data. *Research synthesis methods*, 2018.

[17] V. Lagani, A. D. Karozou, D. Gomez-Cabrero, G. Silberberg, and I. Tsamardinos. A comparative evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions. *BMC bioinformatics*, 17(5):S194, 2016.

[18] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[19] P. Patil and G. Parmigiani. Training replicable predictors in multiple studies. *Proceedings of the National Academy of Sciences*, 115(11):2578–2583, 2018.

[20] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[21] P. J. Brown. Centering and scaling in ridge regression. *Technometrics*, 19(1):35–36, 1977.

[22] D. Jackson, I. R. White, and S. G. Thompson. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in medicine*, 29(12):1282–1297, 2010.

[23] D. Jackson, R. Riley, and I. R. White. Multivariate meta-analysis: potential and promise. *Statistics in medicine*, 30(20):2481–2498, 2011.

[24] F. Hansen and G. K. Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564, 2003.

[25] B. E. Ganzfried, M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, X. V. Wang, M. Ahmadifar, M. J. Birrer, G. Parmigiani, C. Huttenhower, and L. Waldron. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database (Oxford)*, 2013:bat013, 2013. PMID: 23550061.

[26] E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, et al. Accessible, curated metagenomic data through ExperimentHub. *Nature methods*, 14(11):1023, 2017.

[27] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55, 2012.

[28] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99, 2013.

[29] A. Heintz-Buschart, P. May, C. C. Laczny, L. A. Lebrun, C. Bellora, A. Krishna, L. Wampach, J. G. Schneider, A. Hogan, C. de Beaufort, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology*, 2(1):16180, 2017.

[30] D. Jackson, M. Law, J. K. Barrett, R. Turner, J. P. Higgins, G. Salanti, and I. R. White. Extending DerSimonian and Laird's methodology to perform network meta-analyses with random inconsistency effects. *Statistics in medicine*, 35(6):819–839, 2016.

[31] H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

[32] J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503, 2011.

[33] J. Schelldorfer, P. Bühlmann, and S. V. DE GEER. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

[34] Y. Fan and R. Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043, 2012.

IX. Appendix A: Calculations

i. Covariance Matrices and MSPEs

$$\begin{aligned} Cov(\boldsymbol{Y_k}) &= Cov(\boldsymbol{X_k\beta} + \boldsymbol{Z_k\gamma_k} + \boldsymbol{\epsilon_k}) \\ &= \boldsymbol{Z_k G Z_k^T} + \sigma_\epsilon^2 \boldsymbol{I}_{n_k} \end{aligned}$$

i.1 LS

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}}) &= Cov((\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_k^T Y_k}) \\ &= \boldsymbol{\Gamma G \Gamma}^T + \sigma_\epsilon^2(\boldsymbol{X_k^T X_k})^{-1} \end{aligned}$$

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}) &= Cov(\sum_{k=1}^{K} w_k \hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}}) \\ &= \sum_{k=1}^{K} w_k^2 \boldsymbol{\Gamma G \Gamma}^T + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 (\boldsymbol{X_k^T X_k})^{-1} \end{aligned}$$

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}) &= Cov((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1} \sum_{k=1}^{K} [\boldsymbol{X_k^T Z_k G Z_k^T X_k}](\boldsymbol{X}^T\boldsymbol{X})^{-1} + \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

$$\begin{aligned} E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}})) \\ &= \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma G \Gamma}^T \boldsymbol{X_0^T X_0}) + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_0^T X_0}) \\ &= \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{G Z_0^T Z_0}) + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_0^T X_0}) \\ &= \sum_{j=1}^{r} \sigma_{(j)}^2 tr(\boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z_0^T Z_0 \Gamma}_{(j)}) \sum_{k=1}^{K} w_k^2 + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr((\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_0^T X_0}) \end{aligned}$$

$$\begin{aligned} E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2] &= tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}})) \\ &= tr((\boldsymbol{X}^T\boldsymbol{X})^{-1} \sum_{k=1}^{K} [\boldsymbol{X_k^T Z_k G Z_k^T X_k}](\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) + \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) \\ &= tr(\boldsymbol{G} \sum_{k=1}^{K} \boldsymbol{Z_k^T X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_k^T Z_k}) + \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) \\ &= \sum_{j=1}^{r} \sigma_{(j)}^2 tr(\sum_{k=1}^{K} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z_k^T X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_k^T Z_k \Gamma}_{(j)}) + \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) \\ &= \sum_{j=1}^{r} \sigma_{(j)}^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1} \sum_{k=1}^{K} \boldsymbol{X_k^T Z_k \Gamma}_{(j)} \boldsymbol{\Gamma}_{(j)}^T \boldsymbol{Z_k^T X_k}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) + \sigma_\epsilon^2 tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_0^T X_0}) \end{aligned}$$

i.2 Ridge Regression

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{R,k} &= \boldsymbol{S}_k(\tilde{\boldsymbol{X}}_k^T\tilde{\boldsymbol{X}}_k + \lambda_k\boldsymbol{I}_p^-)^{-1}\tilde{\boldsymbol{X}}_k^T\boldsymbol{Y}_k \\ &= \boldsymbol{S}_k(\boldsymbol{S}_k\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{S}_k + \lambda_k\boldsymbol{I}_p^-)^{-1}\boldsymbol{S}_k\boldsymbol{X}_k^T\boldsymbol{Y}_k \\ &= (\boldsymbol{X}_k^T\boldsymbol{X}_k + \lambda_k\boldsymbol{I}_p^-\boldsymbol{S}_k^{-2})^{-1}\boldsymbol{X}_k^T\boldsymbol{Y}_k \\ &= \boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Y}_k\end{aligned}$$

Similarly,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{R,M} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p^-\boldsymbol{S}^{-2})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \boldsymbol{M}^{-1}\boldsymbol{X}^T\boldsymbol{Y}\end{aligned}$$

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}_{R,k}) &= Cov(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Y}_k) \\ &= \boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T Cov(\boldsymbol{Y}_k)\boldsymbol{X}_k\boldsymbol{M}_k^{-1} \\ &= \boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} + \sigma_\epsilon^2\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} \\ &= \boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} + \sigma_\epsilon^2\boldsymbol{M}_k^{-1}\boldsymbol{R}_k\boldsymbol{M}_k^{-1}\end{aligned}$$

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}_{R,C}) &= Cov(\sum_k w_k\hat{\boldsymbol{\beta}}_{R,k}) \\ &= \sum_k w_k^2\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} + \sigma_\epsilon^2\sum_k w_k^2\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} \\ &= \sum_k w_k^2\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1} + \sigma_\epsilon^2\sum_k w_k^2\boldsymbol{M}_k^{-1}\boldsymbol{R}_k\boldsymbol{M}_k^{-1}\end{aligned}$$

$$\begin{aligned}Bias(\hat{\boldsymbol{\beta}}_{R,C}) &= \sum_k w_k E(\hat{\boldsymbol{\beta}}_{R,k}) - \boldsymbol{\beta} \\ &= \sum_k w_k E(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Y}_k) - \boldsymbol{\beta} \\ &= \sum_k w_k\boldsymbol{M}_k^{-1}\boldsymbol{R}_k\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= -(\sum_k w_k\lambda_k\boldsymbol{M}_k^{-1}\boldsymbol{I}_p^-\boldsymbol{S}_k^{-2})\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}_{R,M}) &= Cov(\boldsymbol{M}^{-1}\boldsymbol{X}^T\boldsymbol{Y}) \\ &= \boldsymbol{M}^{-1}Cov(\boldsymbol{X}^T\boldsymbol{Y})\boldsymbol{M}^{-1} \\ &= \boldsymbol{M}^{-1}Cov(\sum_k \boldsymbol{X}_k^T\boldsymbol{Y}_k)\boldsymbol{M}^{-1} \\ &= \boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T Cov(\boldsymbol{Y}_k)\boldsymbol{X}_k\boldsymbol{M}^{-1} \\ &= \boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1} + \epsilon^2\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1} \\ &= \boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1} + \sigma_\epsilon^2\boldsymbol{M}^{-1}\boldsymbol{R}\boldsymbol{M}^{-1}\end{aligned}$$

$$
\begin{aligned}
Bias(\hat{\boldsymbol{\beta}}_{R,M}) &= E(\boldsymbol{M}^{-1}\boldsymbol{X}^T\boldsymbol{Y}) - \boldsymbol{\beta} \\
&= \boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T E(\boldsymbol{Y}_k) - \boldsymbol{\beta} \\
&= \boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T \boldsymbol{X}_k\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= \boldsymbol{M}^{-1}\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= -\lambda(\boldsymbol{M}^{-1}\boldsymbol{I}_p^- \boldsymbol{S}^{-2})\boldsymbol{\beta}
\end{aligned}
$$

Let $\boldsymbol{b}_{R,C} = Bias(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,C})$ and $\boldsymbol{b}_{R,M} = Bias(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,M})$.

$$
\begin{aligned}
E[\|\boldsymbol{Y}_0 - \boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,C}\|_2^2] &= tr(Cov(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,C})) + \boldsymbol{b}_{R,C}^T\boldsymbol{b}_{R,C} \\
&= \sum_k w_k^2 tr(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k\boldsymbol{R}_0) + \sigma_\epsilon^2\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\mathbf{R_0}) + \boldsymbol{b}_{R,C}^T\boldsymbol{b}_{R,C} \\
&= \sum_k w_k^2 tr(\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\boldsymbol{R}_0\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k) + \sigma_\epsilon^2\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\mathbf{R_0}) + \boldsymbol{b}_{R,C}^T\boldsymbol{b}_{R,C} \\
&= \sum_{j=1}^r \sigma_{(j)}^2\sum_k w_k^2 tr(\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\boldsymbol{R}_0\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{\Gamma}_{(j)}) + \sigma_\epsilon^2\sum_k w_k^2 tr(\boldsymbol{M}_k^{-1}\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}_k^{-1}\mathbf{R_0}) + \boldsymbol{b}_{R,C}^T\boldsymbol{b}_{R,C}
\end{aligned}
$$

$$
\begin{aligned}
E[\|\boldsymbol{Y}_0 - \boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{R,M}\|_2^2] &= tr(Cov(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}_{M,C})) + \boldsymbol{b}_{R,M}^T\boldsymbol{b}_{R,M} \\
&= tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \boldsymbol{b}_{R,M}^T\boldsymbol{b}_{R,M} \\
&= \sum_k tr(\boldsymbol{G}\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0\boldsymbol{M}^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k) + \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \boldsymbol{b}_{R,M}^T\boldsymbol{b}_{R,M} \\
&= \sum_{j=1}^r \sigma_{(j)}^2\sum_k tr(\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0\boldsymbol{M}^{-1}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{\Gamma}_{(j)}) + \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \boldsymbol{b}_{R,M}^T\boldsymbol{b}_{R,M} \\
&= \sum_{j=1}^r \sigma_{(j)}^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \sigma_\epsilon^2 tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{M}^{-1}\boldsymbol{R}_0) + \boldsymbol{b}_{R,M}^T\boldsymbol{b}_{R,M}
\end{aligned}
$$

ii. Theorems

Theorem 1, Corollary 1.1, and Theorem 3 are special cases of Theorem 2, Corollary 2.1, and Theorem 4 when $r = 1$, so we provide proofs for Theorems 2 and 4 (the corollaries follow directly from the theorems).

ii.1 Theorem 2

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{r} m_j \sigma_{(j)}^2}{p} \leq \frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k}^{-1}\boldsymbol{R_0}) - tr(\boldsymbol{R}^{-1}\boldsymbol{R_0})}{\max_{j=1,\dots,r} \frac{1}{m_j}(tr(\boldsymbol{R}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{R}^{-1}\boldsymbol{R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_0}^T\boldsymbol{Z_0}\boldsymbol{\Gamma}_{(j)}))} = \tau_{LS,1}$$

$$\implies \sum_{j=1}^{r} \sigma_{(j)}^2 (tr(\boldsymbol{R}^{-1}\sum_{k=1}^{K}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{R}^{-1}\boldsymbol{R_0}) - \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_0}^T\boldsymbol{Z_0}\boldsymbol{\Gamma}_{(j)}))$$

$$\leq \sigma_\epsilon^2 (\sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k}^{-1}\boldsymbol{R_0}) - tr(\boldsymbol{R}^{-1}\boldsymbol{R_0}))$$

$$\iff E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}\|_2^2]$$

The sufficiency of $\overline{\sigma^2} \geq \tau_{LS,2}$ for the CSL outperforming the merged learner can be shown similarly.

ii.2 Theorem 4

$$\overline{\sigma^2} = \frac{\sum_{j=1}^{r} m_j \sigma_{(j)}^2}{p}$$

$$\leq \frac{\sigma_\epsilon^2 \left(\sum_k w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}^{-1}\mathbf{R_0}) - tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{R_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})\right) + \boldsymbol{b}_{\boldsymbol{R,C}}^T\boldsymbol{b}_{\boldsymbol{R,C}} - \boldsymbol{b}_{\boldsymbol{R,M}}^T\boldsymbol{b}_{\boldsymbol{R,M}}}{p \max_{j=1,\dots,r} \frac{1}{m_j}(tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1}\boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}\boldsymbol{R_0}\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}))} = \tau_{R,1}$$

$$\implies \sum_{j=1}^{r} \sigma_{(j)}^2 (tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M}^{-1}\boldsymbol{R_0}) - \sum_k w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{X_k}^T\boldsymbol{Z_k}\boldsymbol{\Gamma}_{(j)}\boldsymbol{\Gamma}_{(j)}^T\boldsymbol{Z_k}^T\boldsymbol{X_k}\boldsymbol{M_k}^{-1}\boldsymbol{R_0}))$$

$$\leq \sigma_\epsilon^2 \left(\sum_k w_k^2 tr(\boldsymbol{M_k}^{-1}\boldsymbol{R_k}\boldsymbol{M_k}\mathbf{R_0}) - tr(\boldsymbol{M}^{-1}\sum_k \boldsymbol{R_k}\boldsymbol{M}^{-1}\boldsymbol{R_0})\right) + \boldsymbol{b}_{\boldsymbol{R,C}}^T\boldsymbol{b}_{\boldsymbol{R,C}} - \boldsymbol{b}_{\boldsymbol{R,M}}^T\boldsymbol{b}_{\boldsymbol{R,M}}$$

$$\iff E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}\|_2^2] \leq E[\|\boldsymbol{Y_0} - \boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,C}}\|_2^2]$$

iii. Optimal Weights

We want to minimize

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}})) = \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{G}\boldsymbol{Z_0}^T\boldsymbol{Z_0}) + \sigma_\epsilon^2 \sum_{k=1}^{K} w_k^2 tr(\boldsymbol{R_k}^{-1}\boldsymbol{R_0})$$

subject to the constraint $\sum_k w_k = 1$.

Let $f(w_1, \dots, w_K) = tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}))$ and $g(w_1, \dots, w_K) = \sum_{k=1}^{K} w_k$.

Then we can write the problem as

$$\text{minimize } f(w_1, \dots, w_K) \text{ subject to}$$
$$g(w_1, \dots, w_K) = 1$$

Using Lagrange multipliers, we get the system of equations

$$\frac{\partial}{\partial w_k} f(w_1, \ldots, w_K) = \lambda \frac{\partial}{\partial w_k} g(w_1, \ldots, w_K) \qquad \text{(for } k = 1, \ldots, K\text{)}$$
$$g(w_1, \ldots, w_K) = 1$$

Since

$$\frac{\partial}{\partial w_k} f(w_1, \ldots, w_K) = 2w_k \left(tr(\boldsymbol{G}\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)$$

and

$$\frac{\partial}{\partial w_k} g(w_1, \ldots, w_K) = 1$$

the system of equations leads to

$$w_k = \frac{\lambda}{2 \left(tr(\boldsymbol{G}\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)} \qquad \text{(for } k = 1, \ldots, K\text{)}$$
$$\sum_k w_k = 1$$

It follows that

$$w_k = \frac{\left(tr(\boldsymbol{G}\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)^{-1}}{\sum_k \left(tr(\boldsymbol{G}\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)^{-1}}$$

which is proportional to the inverse MSPE of $\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,k}}$.

Plugging the optimal weights into the MSPE of the CSL, $tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}))$, we get

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}})) = \frac{1}{\sum_k \left(tr(\boldsymbol{G}\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)^{-1}}$$

In the equal variances setting,

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}})) = \frac{1}{\sum_k \left(\sigma^2 tr(\boldsymbol{Z}_0^T \boldsymbol{Z}_0) + \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1} \boldsymbol{R}_0) \right)^{-1}}$$

so $tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}))$ is approximately linear in σ^2 when $tr(\boldsymbol{R}_k^{-1}\boldsymbol{R}_0)$ is similar across studies or when σ^2 is large.

Approximation of the optimal weights transition point Suppose we are in the equal variances setting. Let τ be the transition point under equal weights. Let $\boldsymbol{C} = tr(\boldsymbol{Z}_0^T \boldsymbol{Z}_0)$, $\boldsymbol{D_k} = \sigma_\epsilon^2 tr(\boldsymbol{R}_k^{-1}\boldsymbol{R}_0)$, $\boldsymbol{E} = \sigma_\epsilon^2 tr(\boldsymbol{R}\boldsymbol{R}_0)$, and $\boldsymbol{F} = tr(\boldsymbol{R}^{-1} \sum_k \boldsymbol{X}_k^T \boldsymbol{Z}_k \boldsymbol{Z}_k^T \boldsymbol{X}_k \boldsymbol{R}^{-1} \boldsymbol{R}_0)$. Using a first order Taylor expansion about τ,

$$tr(Cov(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}})) \approx \frac{1}{\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}} + \sigma^2 \frac{C \sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-2}}{\left(\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}\right)^2} - \tau \frac{C \sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-2}}{\left(\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}\right)^2}$$

$$\implies \widehat{tr(Cov}(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}) \leq tr(Cov(\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}) \iff \overline{\sigma^2} \geq \frac{q}{p} \times \frac{\frac{1}{\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}} - \tau \frac{C \sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-2}}{\left(\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}\right)^2} - \boldsymbol{E}}{\boldsymbol{F} - \frac{C \sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-2}}{\left(\sum_k (\tau \boldsymbol{C} + \boldsymbol{D_k})^{-1}\right)^2}}$$

X. Appendix B: Additional Plots

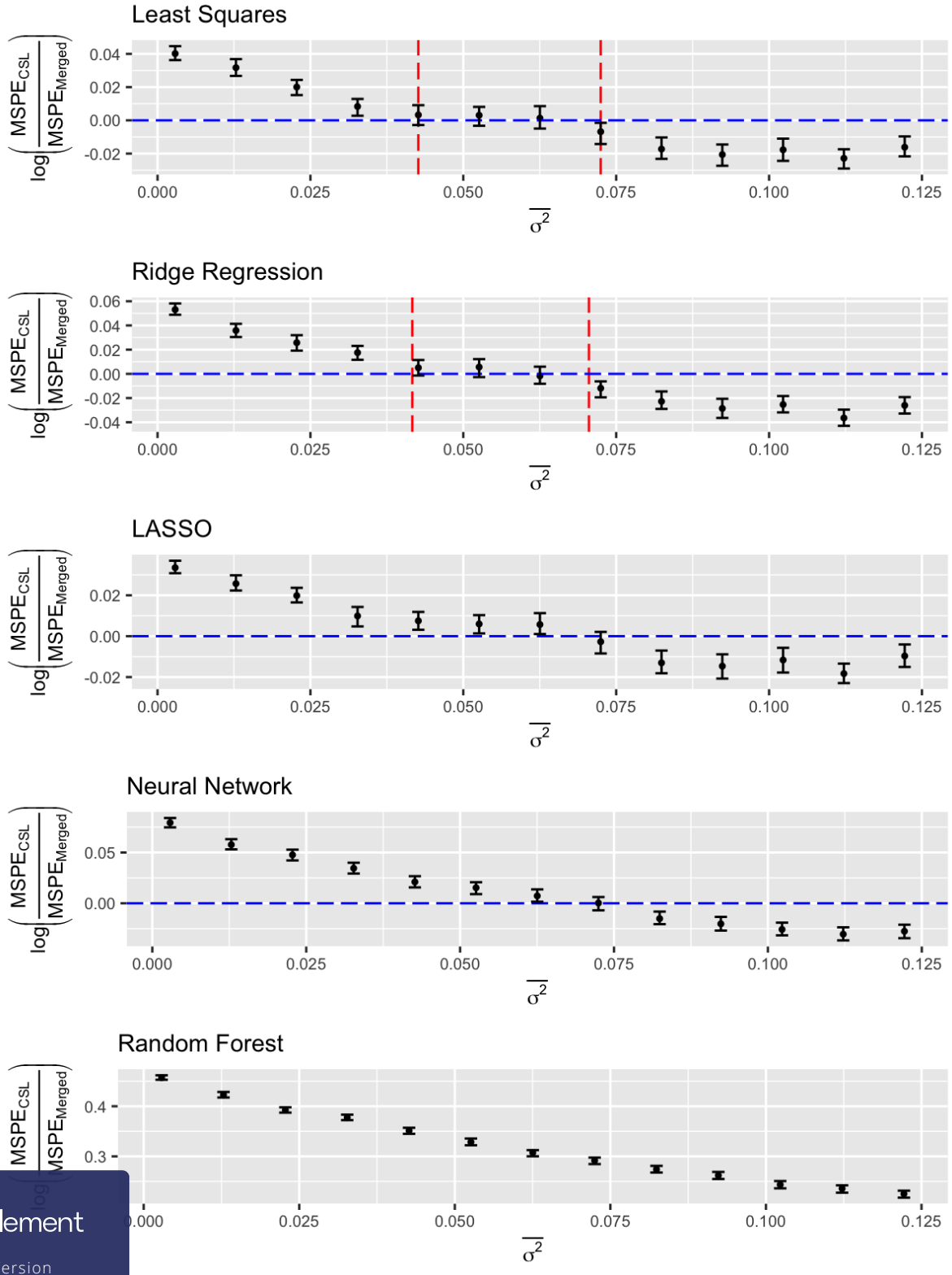

Figure 6: 4 studies of size 40 with 10 predictors. The red lines correspond to the bounds from Theorem 2

.

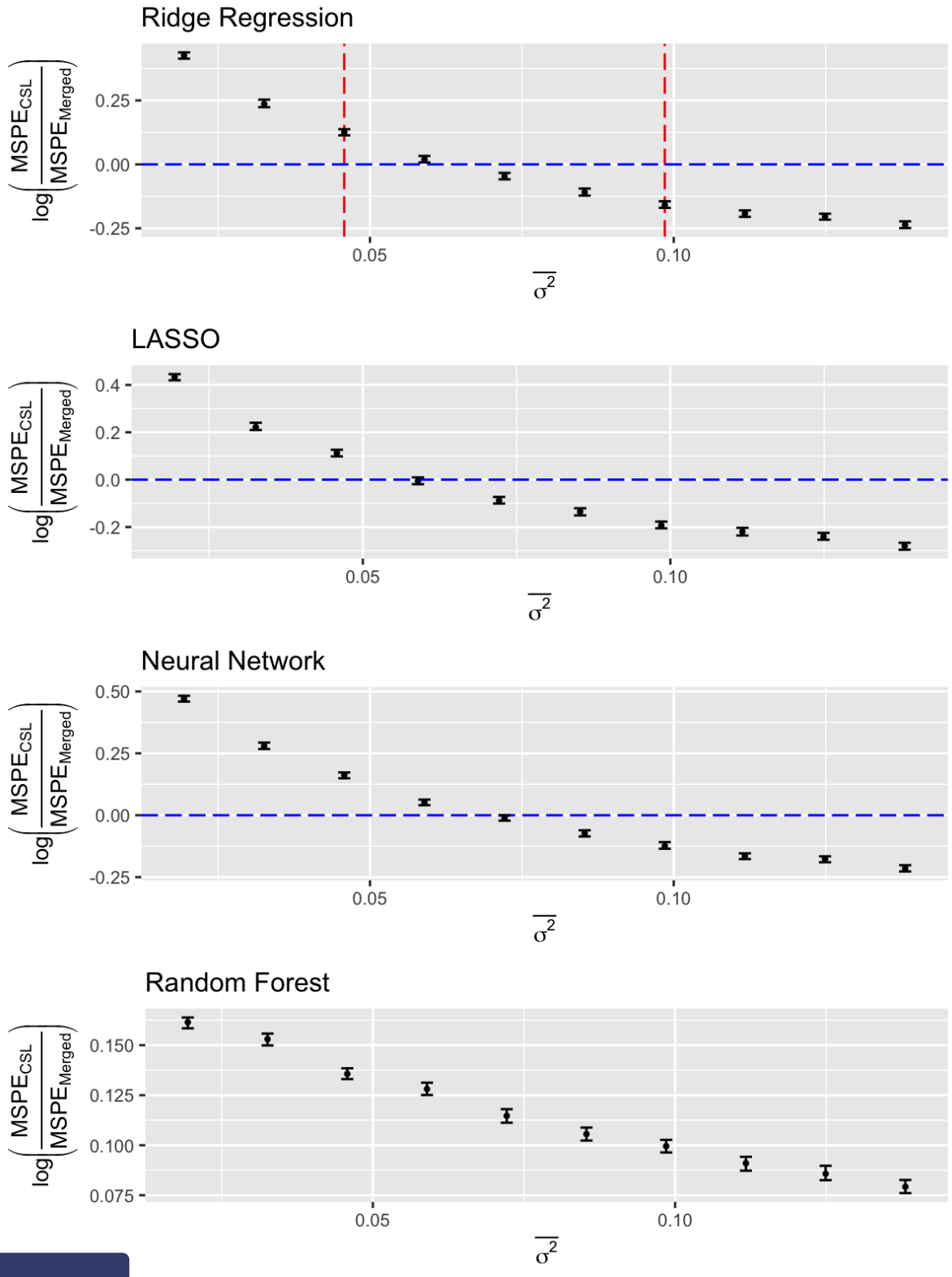

Figure 7. 4 studies of size 40 with 100 predictors. The red lines correspond to the bounds from Theorem 4

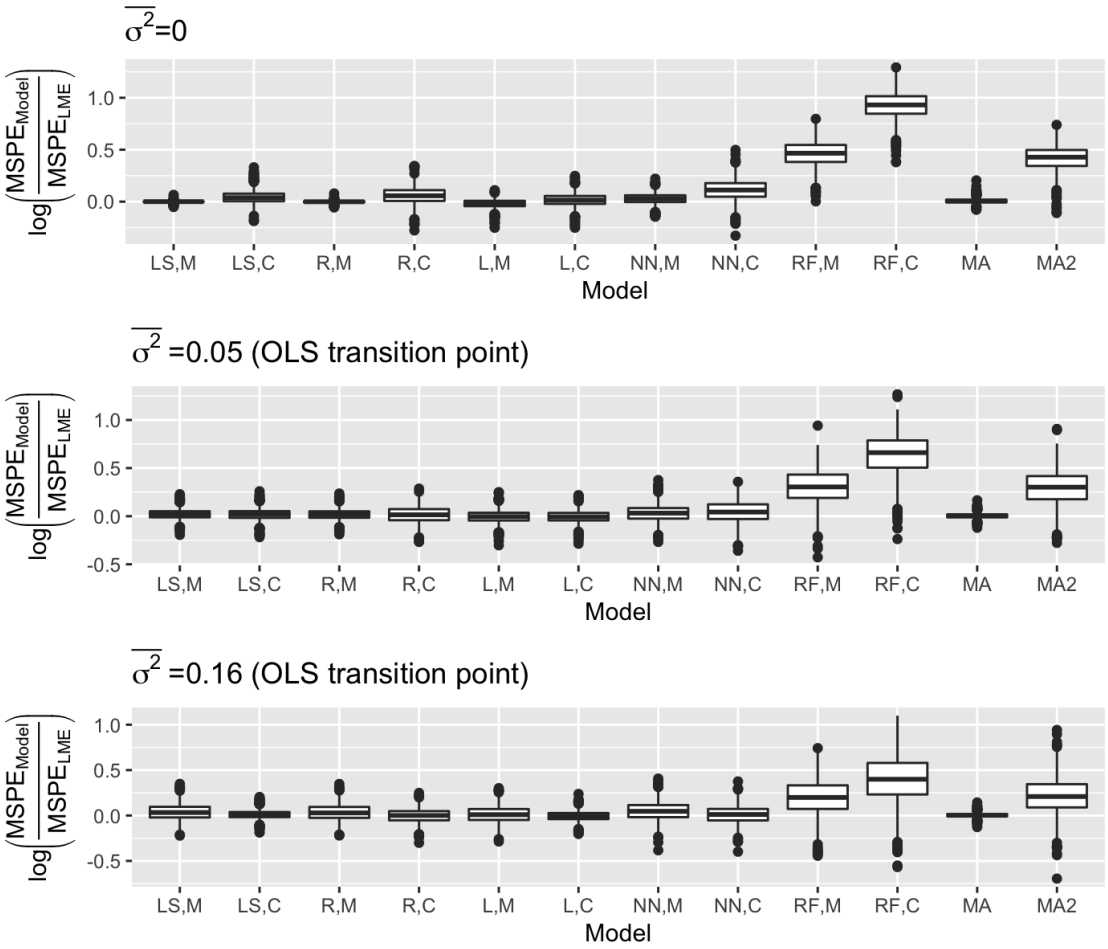

Figure 8: Figure 3 with random forest and univariate meta-analysis results included. LS,M: merged LS learner; LS,C: CSL based on LS; R,M: merged ridge regression learner; R,C: CSL based on ridge regression; L,M: merged LASSO learner; L,C: CSL based on LASSO; NN,M: merged neural network; NN,C: CSL based on neural networks; RF,M: merged random forest; RF,C: CSL based on random forests; MA: multivariate meta-analysis; MA2: multiple univariate meta-analyses.

XI. Appendix C: Notation Table

Variable	Value / Description
K	number of training studies
n_k	sample size in study k ($k = 0$ corresponds to test dataset)
$\boldsymbol{Y_k}$	outcome vector for study k
$\boldsymbol{X_k}$	fixed effects design matrix for study k
$\boldsymbol{Y}$	$(\boldsymbol{Y_1^T}, \ldots, \boldsymbol{Y_K^T})^T$; outcome vector for merged dataset
$\boldsymbol{X}$	$\left[\boldsymbol{X_1^T} \vert \ldots \vert \boldsymbol{X_K^T}\right]^T$; design matrix for merged dataset
p	number of predictors (including the intercept)
q	number of predictors with random effects
r	number of unique variance values for random effects
$\boldsymbol{\Gamma}$	$(\boldsymbol{\Gamma})_{il} = 1$ if predictor i is the lth predictor with a random effect and $(\boldsymbol{\Gamma})_{il} = 0$ otherwise; matrix such that $\boldsymbol{X_k\Gamma}$ subsets $\boldsymbol{X_k}$ to the columns with random effects
$\mathbf{Z_k}$	$\mathbf{X_k}\Gamma$; random effects design matrix for study k
$\boldsymbol{\beta}$	vector of fixed effects
γ_k	vector of random effects
σ_i^2	variance of random effect i
$\boldsymbol{G}$	$diag(\sigma_1^2, \ldots, \sigma_q^2)$; covariance matrix for random effects
$\overline{\sigma^2}$	$tr(\mathbf{G})/p$; heterogeneity summary measure
$\boldsymbol{\epsilon_k}$	vector of residual errors for study k
σ_ϵ^2	variance of residuals
$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,M}}$	$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$; merged LS estimator
w_k	CSL weight for study k
$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS,C}}$	$\sum_k w_k(\boldsymbol{X_k^T X_k})^{-1}\boldsymbol{X_k^T Y_k}$; CSL LS estimator
$\boldsymbol{S_k}$	$\boldsymbol{I_p}$ if scaling is unnecessary, diagonal matrix with $(\boldsymbol{S_k})_{jj}$ = inverse standard deviation of jth column of $\boldsymbol{X_k}$ otherwise; scaling matrix for study k
$\boldsymbol{S}$	$\boldsymbol{I_p}$ if scaling is unnecessary, diagonal matrix with $(\boldsymbol{S})_{jj}$ = inverse standard deviation of jth column of $\boldsymbol{X}$ otherwise; scaling matrix for merged data
$\tilde{\boldsymbol{X}}_{\boldsymbol{k}}$	$\boldsymbol{S_k X_k}$; ridge design matrix for study k
$\tilde{\boldsymbol{X}}$	$\boldsymbol{SX}$; ridge design matrix for merged dataset
$\boldsymbol{I_p^-}$	$\boldsymbol{I_p}$ with $(\boldsymbol{I_p})_{11} = 0$
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}$	$\boldsymbol{S}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \lambda\boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y}$; merged ridge estimator of $\boldsymbol{\beta}$
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}$	$\boldsymbol{S_k}(\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^T\tilde{\boldsymbol{X}}_{\boldsymbol{k}} + \lambda_k\boldsymbol{I_p^-})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{k}}^{\ T}\boldsymbol{Y_k}$; ridge estimator of $\boldsymbol{\beta}$ based on study k
$\hat{\boldsymbol{\beta}}_{\boldsymbol{R,C}}$	$\sum_{k=1}^K w_k\hat{\boldsymbol{\beta}}_{\boldsymbol{R,k}}$; CSL ridge estimator of $\boldsymbol{\beta}$
$\boldsymbol{R_k}$	$\boldsymbol{X_k^T X_k}$
$\boldsymbol{R}$	$\boldsymbol{X}^T\boldsymbol{X}$
$\boldsymbol{M_k}$	$\boldsymbol{R_k} + \lambda_k\boldsymbol{I_p^-}\boldsymbol{S_k^{-2}}$
$\boldsymbol{M}$	$\boldsymbol{R} + \lambda\boldsymbol{I_p^-}\boldsymbol{S^{-2}}$
$\boldsymbol{\Gamma}_{(j)}$	$(\boldsymbol{\Gamma}_{(j)})_{il} = 1$ if random effect i is the lth random effect with variance $\sigma_{(j)}^2$ and $(\boldsymbol{\Gamma}_{(j)})_{il} = 0$ otherwise; matrix such that $\boldsymbol{G\Gamma}_{(j)}$ subsets $\boldsymbol{G}$ to the columns corresponding to $\sigma_{(j)}^2$
τ_{LS}	$\sigma_\epsilon^2 \times \frac{q}{p} \times \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1}R_0}) - tr(\boldsymbol{R^{-1}R_0})}{tr(\boldsymbol{R^{-1}}\sum_{k=1}^K \boldsymbol{X_k^T Z_k Z_k^T X_k R^{-1}R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{Z_0^T Z_0})}$
$\tau_{LS,1}$	$\frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1}R_0}) - tr(\boldsymbol{R^{-1}R_0})}{\max\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{R^{-1}}\sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1}R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))}$
$\tau_{LS,2}$	$\frac{\sigma_\epsilon^2}{p} \frac{\sum_{k=1}^K w_k^2 tr(\boldsymbol{R_k^{-1}R_0}) - tr(\boldsymbol{R^{-1}R_0})}{\min\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{R^{-1}}\sum_{k=1}^K \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k R^{-1}R_0}) - \sum_{k=1}^K w_k^2 tr(\boldsymbol{\Gamma_{(j)}^T Z_0^T Z_0 \Gamma_{(j)}}))}$
$\boldsymbol{b_{R,C}}$	$Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,C}}) = -\sum_k w_k\lambda_k\boldsymbol{X_0 M_k^{-1} I_p^- S_k^{-2}\beta}$
$\boldsymbol{b_{R,M}}$	$Bias(\boldsymbol{X_0}\hat{\boldsymbol{\beta}}_{\boldsymbol{R,M}}) = -\lambda\boldsymbol{X_0 M^{-1} I_p^- S^{-2}\beta}$
τ_R	$\frac{q}{p} \times \frac{\sigma_\epsilon^2\left(\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}R_k M_k^{-1}R_0}) - tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{R_k M^{-1}R_0})\right) + \boldsymbol{b_{R,C}^T b_{R,C}} - \boldsymbol{b_{R,M}^T b_{R,M}}}{tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{X_k^T Z_k Z_k^T X_k M^{-1}R_0}) - \sum_k w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k Z_k^T X_k M_k^{-1}R_0})}$
$\tau_{R,1}$	$\frac{\sigma_\epsilon^2\left(\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}R_k M_k^{-1}R_0}) - tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{R_k M^{-1}R_0})\right) + \boldsymbol{b_{R,C}^T b_{R,C}} - \boldsymbol{b_{R,M}^T b_{R,M}}}{p \max\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M^{-1}R_0}) - \sum_k w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M_k^{-1}R_0}))}$
$\tau_{R,2}$	$\frac{\sigma_\epsilon^2\left(\sum_k w_k^2 tr(\boldsymbol{M_k^{-1}R_k M_k^{-1}R_0}) - tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{R_k M^{-1}R_0})\right) + \boldsymbol{b_{R,C}^T b_{R,C}} - \boldsymbol{b_{R,M}^T b_{R,M}}}{p \min\limits_{j=1,\ldots,r} \frac{1}{m_j}(tr(\boldsymbol{M^{-1}}\sum_k \boldsymbol{X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M^{-1}R_0}) - \sum_k w_k^2 tr(\boldsymbol{M_k^{-1} X_k^T Z_k \Gamma_{(j)} \Gamma_{(j)}^T Z_k^T X_k M_k^{-1}R_0}))}$

Table 1: Notation Table