Cross-study Learning for Generalist and Specialist Predictions

Boyu Ren[1,2], Prasad Patil[3], Francesca Dominici[4], Giovanni Parmigiani[4,5], and Lorenzo Trippa[4,5]

[1]Laboratory for Psychiatric Biostatistics, McLean Hospital

[2]Department of Psychiatry, Harvard Medical School

[3]Department of Biostatistics, Boston University School of Public Health

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[5]Department of Data Sciences, Dana-Farber Cancer Institute

Abstract

Jointly using data from multiple similar sources for the training of prediction models is increasingly becoming an important task in many fields of science. In this paper, we propose a framework for *generalist and specialist* predictions that leverages multiple datasets, with potential heterogenity in the relationships between predictors and outcomes. Our framework uses ensembling with stacking, and includes three major components: 1) training of the ensemble members using one or more datasets, 2) a no-data-reuse technique for stacking weights estimation and 3) task-specific utility functions. We prove that under certain regularity conditions, our framework produces a stacked prediction function with oracle property. We also provide analytically the conditions under which the proposed no-data-reuse technique will increase the prediction accuracy of the stacked prediction function compared to using the full data. We perform a simulation study to numerically verify and illustrate these results and apply our framework to predicting mortality based on a collection of variables including long-term exposure to common air pollutants.

1 Introduction

New advances in technologies, for example biomarker assays in biomedical studies, enable the generation of rich datasets. It is increasingly common for researchers to have access to multiple ($K > 1$) studies, or more generally sets of data, able to answer the same or similar scientific questions (Klein et al., 2014; Kannan et al., 2016; Manzoni et al., 2018). Although datasets from multiple studies may contain the same outcome variable Y and covariates X (for example, patient survival and pre-treatment prognostic profiles in clinical studies), the (X, Y) joint distributions $P_1, \ldots, P_K$ are typically different, due to distinct study populations, study designs and technological artifacts (Simon et al., 2003; Rhodes et al., 2004; Patil et al., 2015; Sinha et al., 2017). In this article, we focus on the task of developing prediction models using multiple datasets, accounting for the heterogeneity across the $(P_k, k = 1, \ldots, K)$ study-specific distributions. We introduce a distinction between two classes of *prediction functions* (PFs) depending on the goal of the prediction problem in the multi-study setting: *generalist* and *specialist* prediction functions.

Generalist predictions are directed to hypothetical future studies $K + 1, K + 2, \ldots$. The training strategy to develop a generalist prediction function depends on relations and similarities between studies. For example, the study-specific geographic areas or assays can be relevant in the development of prediction models. If studies are considered exchangeable, i.e. joint analyses are invariant to permutations of the study indices, then a model which consistently predicts accurately across the available K studies is a good candidate for a generalist use, to predict Y in future studies $k > K$. This class of prediction functions has been studied in the literature (Sutton and Higgins, 2008; Tseng et al., 2012; Pasolli et al., 2016) and several contributions are based on hierarchical models (Warn et al., 2002; Babapulle et al., 2004; Higgins et al., 2009). Similarly, when the exchangeability assumption is inadequate, joint models for multiple studies can incorporate information on relevant relations between studies to construct generalist models(Moreno-Torres et al., 2012). For example, when K studies are collected at different time points $t_1 < t_2 < ... < t_K$, the development of a generalist model can incorporate potential cycles or short-term trends.

Specialist predictions are in contrast directed to predicting future outcomes Y based on covariates X in the context of a specific study k in $\{1, \ldots K\}$ –for example a geographic area– represented by one of the K datasets. Bayesian models can be used to borrow information and leverage $K-1$ datasets in addition to the targeted study k. Typically the degree of heterogeneity of the distributions $(P_1, \ldots, P_K)$ affects the extent of improvement in accuracy that one achieves with multi-study models compared to simpler models developed using only data from study k.

Recently, the use of ensemble methods has been proposed to develop generalist prediction functions based on multi-study data collections (Patil and Parmigiani, 2018; Zhang et al., 2019; Loewinger et al., 2019). In particular, stacking (Wolpert, 1992; Breiman, 1996) is used to combine prediction functions $\{\hat{Y}_k(\cdot), k = 1, \ldots, K\}$, each trained on a single study k, into a single generalist prediction function that targets contexts $k > K$. The weights assigned to each model $\hat{Y}_k$ in stacking are often derived by maximizing a utility function representative of the performance of the resulting prediction function. In this manuscript, our focus will be on collections of exchangeable studies. Nonetheless, the application of stacking does not require this exchangeability assumption, and the optimization of the ensemble weights can be tailored to settings where exchangeabilty is implausible. Importantly, stacking allows investigators to capitalize on multiple machine learning algorithms, such as random forest or neural networks, to train the study-specific functions $\hat{Y}_k$.

We investigate within the stacking framework Patil and Parmigiani (2018) the optimization of the ensemble weights assigned to a collection of single-set prediction functions (SPFs), generated with arbitrary machine learning methods. Each SPF is trained by a single study k or combining multiple studies. The ensemble weights will approximately maximize a utility function U which we estimate using the entire collection of K studies (generalist prediction) or only data in study k (specialist prediction). Notably, stacking as currently implemented in multi-study learning can potentially suffer from over-fitting due to data reuse (DR): the same datasets generate SPFs and contribute (with others) to guiding the optimization of the stacking weights. With the aim of mitigating overfitting we introduce a no data reuse (NDR) procedure that still includes three key components of the staking methodology: the training of SPFs, the estimation of the utility function

U, and the optimal choice of the ensemble weights.

In this manuscript we compare procedures to weight SPFs with and without data reuse. We use the mean squared error (MSE) as our primary metric to evaluate prediction accuracy. Our results prove that, when the number of studies K and the sample sizes n_k become large, both stacking with DR and NDR achieve a performance similar to that of an oracle benchmark. The oracle is defined as the linear combination of the SPFs' limits ($\lim_{n_k} Y_k$) that minimizes the MSE in future studies $k > K$. Our results bound the MSE difference between the oracle ensembles and two stacking procedures, with and without data reuse. We use these asymptotic results to describe similarities between stacking and multi-study Bayesian hierarchical models when the SPFs are linear. Related bounds have been studied for the single-study setting in van der Laan et al. (2006) and in the functional aggregation literature (Juditsky and Nemirovski, 2000; Juditsky et al., 2008). We also illustrate that if the oracle predictions lie within the convex hull of the SPFs limits ($\lim_{n_k} Y_k; k = 1, \ldots, K$), then stacking produces prediction functions that are asymptotically equivalent to the oracle. We finally provide finite sample comparisons of stacking with DR and NDR. We identify a threshold value for the number of datasets K, which depends on the cross-study heterogeneity, below which NDR stacking reduces the MSE.

We apply our NDR and DR stacking procedures to predict mortality in Medicare beneficiaries enrolled before 2002. The datasets contain demographic and health-related information of the beneficiaries at the zipcode-level and measurements of air pollutants. We are interested in predicting the number of deaths per 10,000 person-years. In distinct analyses, we partitioned the database into state-specific datasets ($K = 50$) and in county-specific dataset ($K = 58$). We compare NDR and DR stacking relative performances. The results are aligned with our analytic results. Indeed with hold-out data we verified that in the first analysis, with $K = 10$ state-level datasets (high heterogeneity; the remaining 40 are used as validation datasets), NDR produced generalist predictions with better accuracy than DR. In contrast, with SPFs developed with county-specific datasets (low cross-study hetherogeneity) DR staking predictions are more accurate than with NDR stacking. These comparisons were confirmed by iterated analyses with random sets of $K = 10$ states and $K = 10$ counties.

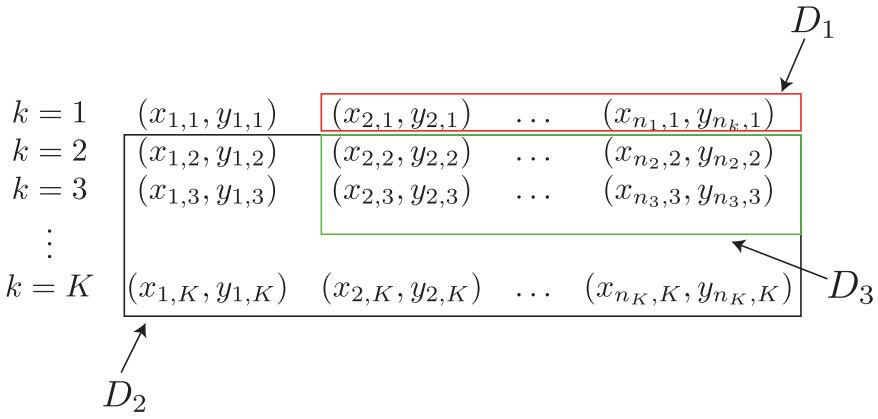


Figure 1: An illustration of the relation between studies and training sets, where $\mathcal{D} = \{D_1, D_2, D_3\}$.

2 Generalist and Specialist Predictions

2.1 Notation

We observe K studies $k = 1, \ldots, K$, with sample sizes n_k. For individual i in study k we have a vector of features $x_{i,k} \in \mathcal{X}$ and the individual outcome $y_{i,k} \in \mathbb{R}$. We use $\mathcal{S} = \{(x_{i,k}, y_{i,k}); i = 1, \ldots, n_k, k = 1, \ldots, K\}$ to indicate the collection of all K datasets. Based on these, we define a library of training sets (LTS), denoted as $\mathcal{D}$, which includes T members $D_1, \ldots, D_T$. Each D_t is a set of (i, k) indices, where $i \in \{1, 2, \ldots, n_k\}$ is the sample index within a study k. The set D_t can include indices with different k values (see for example D_3 in Figure 1). We call a collection $\mathcal{D}$ study-specific if $T = K$ and $D_t = \{(1, t), \ldots, (n_t, t)\}$, with $t = 1, \ldots, K$.

We consider L different learners —a learner is a method of generating a prediction function, such as linear regression, random forest, or a neural network. Training learner ℓ on set $D_t \in \mathcal{D}$ generates a single-study prediction function (SPF) noted as $\hat{Y}_t^\ell : \mathcal{X} \to \mathbb{R}$. The set of all SPFs is $\hat{\mathcal{Y}} = \{\hat{Y}_t^\ell(\cdot); \ell = 1, \ldots, L, t = 1, \ldots, T\}$. Let W be a subset of $\mathbb{R}^{TL}$. With stacking (Wolpert, 1992), we combine the $\hat{\mathcal{Y}}$ components into $\hat{Y}_w : \mathcal{X} \to \mathbb{R}$ via:

$$\hat{Y}_w(\cdot) = \sum_{\ell=1}^{L} \sum_{t=1}^{T} w_{\ell,t} \hat{Y}_t^\ell(\cdot), \tag{1}$$

where $w = (w_{\ell,t}; \ell \leq L, t \leq T)$ is a vector of weights in W.

We want to use $\hat{Y}_w$ for prediction in a target population with unknown joint

(X, Y) distribution π. The performance of $\hat{Y}_w$ is quantified by its expected utility U, quantifying accuracy in the target population:

$$U(w; \pi) = \int_{(x,y)} u(\hat{Y}_w(x), y)\ d\pi(x, y),$$

where $u(\hat{y}, y)$ is a utility function, e.g. $u(\hat{y}, y) = -(\hat{y} - y)^2$. The distribution π is unknown and we estimate $U(w; \pi)$ with

$$\hat{U}(w; \nu) = \sum_{k=1}^{K} \frac{\nu_k}{n_k} \sum_{i=1}^{n_k} u(\hat{Y}_w(x_{i,k}), y_{i,k}). \tag{2}$$

The weights $\nu_k \geq 0$, $\sum_k \nu_k = 1$, are user-specified and are designed to capture the relation between the target π and the set of distributions $P_1, \ldots, P_K$. In this paper, we are interested in generalist prediction, which corresponds to $\nu_k = \frac{1}{K}$, for $k = 1, \ldots, K$, and specialist prediction, which corresponds to $\nu_k = 1$ for study k in $1, \ldots, K$ and 0 for the remaining $K - 1$ studies.

2.2 Generalist prediction

The target distribution π coincides with a future sequence of heterogeneous studies $K + 1, K + 2, \ldots$, and the utility of a generalist prediction function $\hat{Y}_w$ can be represented as

$$U_g(w) = \lim_{I \to \infty} I^{-1} \sum_{i=1}^{I} U(w; P_{K+i}).$$

where the subscript g reminds us that the limit is taken in the generalist case. We will consider scenarios where the above limit is well defined for any $w \in W$ with probability 1. If $P_1, P_2, \ldots$, are exchangeable, i.e. there exists Q such that $P_k | Q \overset{iid}{\sim} Q, k = 1, 2, \ldots$, then $U_g(w)$ can be rewritten as,

$$U_g(w) = \int \left[\int_{(x,y)} u(\hat{Y}_w(x), y) dP(x, y) \right] dQ.$$

Changing the order of integration,

$$U_g(w) = \int_{(x,y)} u(\hat{Y}_w(x), y) dP_0(x, y),$$

where P_0 is the mean of Q, i.e. $P_0(\cdot) = \int P(\cdot) dQ$.

When $\pi = P_0$, we can use $\nu_k = 1/K$ for $k = 1, \ldots, K$ in expression (2) to

approximate $U_g(w)$. Note that in several applications the sequence $P_1, P_2 \ldots$ may not be exchangeable. For example, it can be better modeled by a Markov Chain (Shumway and Stoffer, 2017) i.e. $P_k|P_1, \ldots, P_{k-1} = P_k|P_{k-1}$. Throughout this manuscript we will not need to specify the model Q, but we will assume the exchangeability of the sequence $P_1, P_2, \ldots$.

2.3 Specialist prediction

In this case, the target population distribution π coincides with P_k, for a single $k \in \{1, 2, \ldots, K\}$. The expected utility of a specialist prediction function is

$$U_s(w; k) = U(w; P_k) = \int_{(x,y)} u(\hat{Y}_w(x), y) dP_k(x, y).$$

We can use the empirical distribution of study k to estimate P_k, and the implied specification of ν in (2) is $\nu_i = 1$ for $i = k$ and 0 otherwise.

3 Generalist and specialist stacking

We use stacking for generalist and specialist predictions in multi-study settings. Recall the definition of a stacked prediction function $\hat{Y}_w(\cdot) = \sum_{\ell \leq L, t \leq T} w_{\ell,t} \hat{Y}_t^{\ell}(\cdot)$ based on a set of SPFs $\hat{\mathcal{Y}}$ and weights $w \in W$. We indicate as *oracle* weights

$$w_g = \arg\max_{w \in W} U_g(w)$$
$$w_s^{(k)} = \arg\max_{w \in W} U_s(w; k).$$

Note that $W \subset \mathbb{R}^{TL}$.

Constraints on or penalties applied to select parameters like w can lead to identical results. For example, in several optimization problem constraining an $KL-$dimensional parameter to $W = \{w : \|w\|_2 \leq c\}$ is equivalent to the unconstrained optimization with an L_2 penalty on the parameter. The use of penalties in the estimation of stacking weights has been discussed in Breiman (1996); LeBlanc and Tibshirani (1996). One of the main arguments is that members of the library of SPFs $\hat{\mathcal{Y}}$ tend to be correlated, especially those that are trained on the same set D_t.

3.1 Stacking with data reuse

A direct approach to select w_g and $w_s^{(k)}$ consists in optimizing the $\hat{U}(w;\nu)$ estimates of $U_g(w)$ and $U_s(w;k)$. When the studies are exchangeable $\hat{U}(w;K^{-1}\mathbf{1}_K)$ can be used to select w_g. The estimation of stacking weights attempts to provide values close to the oracle solution w_g. If instead we develop a specialist prediction function for study k, we can optimize $\hat{U}(w;e_k)$ to select $w_s^{(k)}$, where e_k is a $K-$dimensional vector with the k-th component to be one and all others zero. This approach reuses data. Training an SPF $\hat{Y}_t^\ell$ uses part of the data $D_t \subset \mathcal{S}$ that are then reused to compute $\hat{U}(w;\nu)$. Data reuse makes $\hat{U}(w;\nu)$ a biased estimator of $U_g(w)$ and $U_s(w;k)$. In the next paragraph we illustrate a simple example where the bias of $\hat{U}(w;\nu)$ due to data reuse makes the selection of w, denoted as $\hat{w}$, erroneously favors those $\hat{Y}_t^\ell$ generated from studies with large ν_k.

Consider a scenario where $\mathcal{D}$ is study-specific and $K = 2$. Let $u(y, y') = -(y - y')^2$. We only observe $y_{i,k}$ without any covariates and we assume $y_{i,k} \sim N(\mu_k, 1)$ for $k = 1, 2$ where $\mu_k \sim N(0, 1)$. Let $n_1 = n_2 = n$. In this simple example, we generate a library of SPFs with two constant functions $\hat{Y}_1(\cdot) = \bar{y}_1$ and $\hat{Y}_2(\cdot) = \bar{y}_2$, where $\bar{y}_k = n^{-1}\sum_i y_{i,k}$. Under the constraint that $W = \Delta_1$, where Δ_1 is the standard 1-simplex, the weights that optimize $\hat{U}(w;\nu)$ is $\hat{w} = (\hat{w}_1, \hat{w}_2) = (\nu_1, \nu_2)$, while the oracle weights $w_g = (w_{g,1}, w_{g,2})$ that optimize $U_g(w)$ are

$$w_g = \begin{cases} \left(\frac{|\bar{y}_2|}{|\bar{y}_1|+|\bar{y}_2|}, \frac{|\bar{y}_1|}{|\bar{y}_1|+|\bar{y}_2|}\right) & \bar{y}_1 \cdot \bar{y}_2 < 0, \\ (1,0) & |\bar{y}_1| \le |\bar{y}_2|, \quad \bar{y}_1 \cdot \bar{y}_2 \ge 0, \\ (0,1) & |\bar{y}_1| > |\bar{y}_2|, \quad \bar{y}_1 \cdot \bar{y}_2 \ge 0. \end{cases}$$

The oracle weights favor $\hat{Y}_1$, i.e. $w_{g,1} > w_{g,2}$, if $\text{cMSE}(\hat{Y}_1) < \text{cMSE}(\hat{Y}_2)$, where cMSE indicates the conditional MSE of a SPF $\hat{Y}_t^\ell$:

$$\text{cMSE}(\hat{Y}) = \int_{(x,y)} \left(y - \hat{Y}(x)\right)^2 dP_0(x, y).$$

The cMSE measures the actual prediction performance of $\hat{Y}_t^\ell$ across studies given the observed data $(y_{1,1}, \ldots, y_{1,n}, y_{2,1}, \ldots, y_{2,n})$. Note that in our example, $\text{cMSE}(\hat{Y}_k) = |\bar{y}_k|^2 + 2$, $k = 1, 2$. On the other hand, $\hat{w}$ favor $\hat{Y}_k$ whenever $\nu_k > \nu_{k'}$, regardless of the cMSE of each SPF.

To understand the discrepancy described above, we examine the bias of $\hat{U}(w;K^{-1}\mathbf{1}_K)$

to $U_g(w)$, defined as $\mathbb{E}\left(\hat{U}(w; K^{-1}\mathbf{1}_K) - U_g(w)\right)$, where the expectation is taken over all observed data $(y_{1,1}, \ldots, y_{1,n}, y_{2,1}, \ldots, y_{2,n})$. By definition, we have

$$\hat{U}(w; K^{-1}\mathbf{1}_K) - U_g(w) = \underbrace{2(\nu_1 w_2 + \nu_2 w_1)\bar{y}_1\bar{y}_2 + 2(\nu_1 w_1 \bar{y}_1^2 + \nu_2 w_2 \bar{y}_2^2)}_{\text{data-reuse}} + 2 - \nu_1\overline{y_1^2} - \nu_2\overline{y_2^2},$$

where $\overline{y_k^2} = n^{-1}\sum_i y_{i,k}^2$. The first two terms on the right-hand side exist because of data reuse, that is, we evaluate the utility of $w_1\hat{Y}_1 + w_2\hat{Y}_2$ using the training data of $\hat{Y}_1$ and $\hat{Y}_2$.

It follows that

$$\mathbb{E}\left(\hat{U}(w; K^{-1}\mathbf{1}_K) - U_g(w)\right) = \frac{2(n+1)}{n}(\nu_1 w_1 + \nu_2 w_2).$$

We can see data-reuse introduces a non-zero bias to $\hat{U}(w; K^{-1}\mathbf{1}_K)$. This bias term is not always maximized at w_g. In fact, if $\nu_1 \geq \nu_2$, $\nu_1 w_1 + \nu_2 w_2$ is maximized at $w = (1, 0)$. In this case, the bias term would shift $\hat{w}_1$ towards 1, which would make $\hat{w}_1$ larger than $w_{g,1}$ if $w_{g,1} \neq 1$. The strength of this shift increases as ν_1 increases, which explains the reason that $\hat{w}_1$ increases as ν_1 increases.

The effect of the bias term on the discrepancy of $\hat{w}$ to w_g is particularly pronounced when training specialist PFs for study k with $\hat{U}(w; e_k)$. In our example, $\hat{w} = e_k$ regardless of the values of cMSE of $\hat{Y}_k$, which in our setting also captures the prediction accuracy of $\hat{Y}$ on future samples in study k. This result also generalizes to $K > 2$ and to the setting where $L > 1$ with at least one of the single learner using $-u(y, y')$ as its loss function. The specialist PF for study k is then equal to the SPF trained on study k, and we do not borrow any information from other studies, even though they share the same hyper-distribution of mean of the outcome Y with study k.

3.2 Stacking without data reuse

A common approach to limit the effects of data reuse is cross-validation (CV). CV in stacking is implemented by using part of the data for the training of the library of PFs $\hat{\mathcal{Y}}$ and the rest of the data for the estimation of w (see for example Breiman (1996)). How to split the data in multi-study settings is not as obvious as in the

single-study setting due to the multi-level structure of the data. We consider two approaches based on CV. We first introduce their primary characteristics and their precise definitions are deferred to Section 3.2.1 and 3.2.2.

1. **Within-set** (CV_{ws}). For this approach, we assume that sets D_t are mutually exclusive. An M-fold CV_{ws} includes M iterations. At each iteration, we randomly partition each D_t into $D_{t,1}$ and $D_{t,2}$. We use $\{D_{t,1}; t \leq T\}$ to generate the class of SPFs and predict outcomes for samples in $\{D_{t,2}; t \leq T\}$. The final selection of w maximizes a utility estimate that involves all predictions generated across the M iterations.

2. **Cross-set** (CV_{cs}). This approach can handle LTS with overlapped D_t sets and involves a pre-defined number of iterations. At each iteration, we randomly select T' sets $D_t \in \mathcal{D}$ to generate the library of SPFs. We then predict outcomes for samples in the rest of D_t sets using each member of the library. The final selection of w maximizes a utility function that involves predictions generated across all interations.

3.2.1 Within-set CV

We describe the CV_{ws} procedure in the multi-study setting. It can be used to estimate generalist and specialist utilities. An $M-$fold CV_{ws} for stacking includes four steps. Without loss of generality, we assume that $|D_t|$ is divisible by M for $t = 1, \ldots, T$, where $|D_t|$ is the cardinality of D_t.

1. Randomly partition each index set D_t into M equal-sized subsets and denote them as $D_{t,m}, m = 1, \ldots, M$.

2. For every $m = 1, \ldots, M$, train $\hat{Y}^{\ell}_{t,m}$ with $\{(x_{i,k}, y_{i,k}); (i,k) \in D_t \cap D^c_{t,m}\}$ for $\ell = 1, \ldots, L$ and $t = 1, \ldots, T$.

3. For a sample with index (i,k), denote the only index m such that $(i,k) \in D_{t,m}$ by $m(i,k)$. The estimated utility function for w is

$$\hat{U}_{\text{ws}}(w; \nu) = \sum_{k=1}^{K} \frac{\nu_k}{n_k} \sum_{i=1}^{n_k} u\left(\sum_{\ell,t} w_{\ell,t} \hat{Y}^{\ell}_{t,m(i,k)}(x_{i,k}), y_{i,k} \right). \tag{3}$$

and

$$\hat{w}^{\text{ws}} = \arg\max_{w \in W} \hat{U}_{\text{ws}}(w; \nu).$$

4. The CV_{ws} stacked PF is

$$\hat{Y}_w^{\text{ws}} = \sum_{\ell,t} \hat{w}_{\ell,t}^{\text{ws}} \hat{Y}_t^{\ell}.$$

For specialist predictions, CV_{ws} can solve the data-reuse related problem in Section 3.1. In particular, we consider the example in Section 3.1. Assume D_t is study-specific. Denote $\hat{w}_s^{\text{ws}} = \arg\max_w \hat{U}_{\text{ws}}(w; e_1)$ the CV_{ws} selected weights for the specialist PF of study 1 and $\hat{w}_s = \arg\max_w \hat{U}(w; e_1)$. We measure the prediction accuracy of a PF $\hat{Y}_w$ with the expected MSE on study 1 is $\text{MSE}_1(w) = \int_{\mu_1} (\mu_1 - \hat{Y}_w)^2 dP(\mu_1) + 1$, where $P(\mu_1)$ is the distribution of μ_1.

Since there is no analytic expression for $\hat{w}_s^{\text{ws}}$, we use Monte Carlo simulation (1000 replications) to compare $\text{MSE}_1(\hat{w}_s^{\text{ws}})$ and $\text{MSE}_1(\hat{w}_1)$. We set $n = 90$ and $\mu_1 \sim N(0, 0.1)$, where borrowing information from study 2 is beneficial for the estimation of μ_1. When varying M from 3 to 15, we observe that $\mathbb{E}\left[\text{MSE}_1(\hat{w}_1) - \text{MSE}_1(\hat{w}_s^{\text{ws}})\right]$ first increases from 0.0014 to 0.002 then decrease when $M > 8$ to 0.0012 at $M = 15$. Here the expectation is taken over $(y_{1,1}, \ldots, y_{1,n}, y_{2,1}, \ldots, y_{2,n})$. This indicates that CV_{ws}-based approach produces a more accurate PF that stacking with data reuse but the advantage decreases if M is large.

In contrast we illustrate that, if we compare CV_{ws} and stacking with data reuse for generalist predictions, the difference between the resulting estimates of $U(w)$, $\hat{U}(w; \nu) - \hat{U}_{\text{ws}}(w; \nu)$, converges to zero faster than the difference between $\hat{U}(w; \nu)$ and its limit as $n \to \infty$ for any fixed K, rendering $\hat{U}(w; \nu)$ to be asymptotically identical to $\hat{U}(w; \nu)$.

To see this result, we first consider the example in Section 3.1 with fixed $\mu = (\mu_1, \mu_2)^\intercal$ and bounded W. The utility function for stacking with data reuse is $\hat{U}(w; (1/2, 1/2)) = w^\intercal \hat{\Sigma} w - 2\hat{b}^\intercal w + (\overline{y_1^2} + \overline{y_2^2})/2$, where

$$\hat{\Sigma} = \begin{bmatrix} \bar{y}_1^2 & \bar{y}_1 \bar{y}_2 \\ \bar{y}_1 \bar{y}_2 & \bar{y}_2^2 \end{bmatrix}, \quad \hat{b} = \begin{bmatrix} \bar{y}_1 \bar{y} \\ \bar{y}_2 \bar{y} \end{bmatrix},$$

and $\bar{y} = (\bar{y}_1 + \bar{y}_2)/2$. Let $\bar{y}_{k,-m} = (n(M-1)/M)^{-1} \sum_{i \notin D_{k,m}} y_{k,i}$ and $\bar{y}_{k,m} = (n/M)^{-1} \sum_{i \in D_{k,m}} y_{k,i}$. We use a 2-fold $\mathrm{CV}_{\mathrm{ws}}$ to select w for generalist predictions. The associated utility function is $\hat{U}_{\mathrm{ws}}(w; (1/2, 1/2))$ and

$$\hat{U}_{\mathrm{ws}}(w; (1/2, 1/2)) = w^\intercal \hat{\Sigma}_{\mathrm{ws}} w - 2\hat{b}_{\mathrm{ws}}^\intercal w + \frac{\overline{y_1^2} + \overline{y_2^2}}{2},$$

where

$$\hat{\Sigma}_{\mathrm{ws}} = M^{-1} \begin{bmatrix} \sum_{m=1}^{2} \bar{y}_{1,-m}^2 & \sum_{m=1}^{2} \bar{y}_{1,-m}\bar{y}_{2,-m} \\ \sum_{m=1}^{2} \bar{y}_{1,-m}\bar{y}_{2,-m} & \sum_{m=1}^{2} \bar{y}_{2,-m}^2 \end{bmatrix}, \ \hat{b}_{\mathrm{ws}} = (2M)^{-1} \begin{bmatrix} \sum_{m=1}^{2} (\bar{y}_{1,m} + \bar{y}_{2,m})\bar{y}_{1,-m} \\ \sum_{m=1}^{2} (\bar{y}_{1,m} + \bar{y}_{2,m})\bar{y}_{2,-m} \end{bmatrix}.$$

Note by construction, $\bar{y}_k = (M-1)/M \bar{y}_{k,-m} + 1/M \bar{y}_{k,m}$. Therefore,

$$\sum_m \bar{y}_{k,-m} y_{k',-m} = \frac{M^3 - 2M^2 + M}{(M-1)^2} \bar{y}_k \bar{y}_{k'} + \sum_m \left(\bar{y}_{k,m} \bar{y}_{k',m} - \bar{y}_k \bar{y}_{k'} \right),$$

for any $k, k' \in \{1, 2\}$. It is straightforward to show that

$$\mathrm{var}\left(\sum_m \left(\bar{y}_{k,m} \bar{y}_{k',m} - \bar{y}_k \bar{y}_{k'} \right) \right) = \frac{1}{4} \mathrm{var}\left((\bar{y}_{k,1} - \bar{y}_{k,2})(\bar{y}_{k',1} - \bar{y}_{k',2}) \right) = \frac{1 + \mathbb{I}(k = k')}{n^2}.$$

Therefore $\sum_m \bar{y}_{k,-m} y_{k',-m} = 2\bar{y}_k \bar{y}_{k'} + O_p(1/n)$. Similarly, we can prove that $\sum_m (\bar{y}_{k,m} + \bar{y}_{k',m}) \bar{y}_{k,-m} = 2M \bar{y} \bar{y}_k + O_p(1/n)$ for $k, k' \in \{1, 2\}$ and $k \neq k'$. Based on these results, if w is bounded by a finite constant, we have $|\hat{U}_{\mathrm{ws}}(w; (1/2, 1/2)) - \hat{U}(w; (1/2, 1/2))| \leq O_p(1/n)$.

On the other hand, the limit of $\hat{U}(w; (1/2, 1/2))$ as $n \to \infty$ is $w^\intercal \mu \mu^\intercal w - w^\intercal \mu \mu^\intercal \mathbf{1}_2 + \mu^\intercal \mu / 2 + 1$. Since $\bar{y}_k \bar{y}_{k'} = \mu_k \mu_{k'} + O_p(1/\sqrt{n})$ and $\bar{y}_k \bar{y} = \mu_k (\mu_1 + \mu_2)/2 + O_p(1/\sqrt{n})$ by central limit theorem and delta method, $\hat{U}(w; (1/2, 1/2))$ converges to its limit with rate $1/\sqrt{n}$. Hence $|\hat{U}_{\mathrm{ws}}(w; (1/2, 1/2)) - \hat{U}(w; (1/2, 1/2))|$ is ignorable compared to the random fluctuation of $\hat{U}(w; (1/2, 1/2))$ when n is large.

This result on convergence rate also holds when $E(Y|X)$ is linear and $\hat{Y}_t^\ell$ is trained with an ordinary least squares (OLS) regression. Consider K studies,

$$y_{i,k} = \beta_k^\intercal x_{i,k} + \epsilon_{i,k}, \tag{4}$$

where β_k is a study-specific regression coefficient and $\epsilon_{i,k}$ are $N(0, \sigma^2)$ noise terms. The $x_{i,k} \sim N(0, I)$ are *iid* p-dimensional covariate vectors in all K studies. In the

following proposition, we show the respective rates at which $|\hat{U}_{\text{ws}}(w; 1/K\mathbf{1}_K) - \hat{U}(w; 1/K\mathbf{1}_K)|$ and $|\hat{U}(w; 1/K\mathbf{1}_K) - \lim_{n\to\infty} \hat{U}(w; 1/K\mathbf{1}_K)|$ converge to 0.

Proposition 1. *Assume $\mathcal{D}$ is study-specific and $n_k = n$ for $k \in \{1, 2, \ldots, K\}$, where n is divisible by M. Fix $\beta_1, \ldots, \beta_K$ and assume the data are generated with (4). Let $L = 1$ and the single learner be an OLS procedure. If any sub-matrix X'_k formed by $(1-1/M)n$ rows of X_k is invertible for every k, and $u(\hat{y}, y) = -(\hat{y}-y)^2$, then for any $w \in W$, where W is a bounded set in $\mathbb{R}^T$, the following inequality holds*

$$\begin{aligned} \sup_{w\in W} \left|\hat{U}(w; K^{-1}\mathbf{1}_K) - \hat{U}_{ws}(w; K^{-1}\mathbf{1}_K)\right| &\leq O_p(1/n), \\ \sup_{w\in W} \left|\hat{U}(w; K^{-1}\mathbf{1}_K) - \lim_{n\to\infty} \hat{U}(w; K^{-1}\mathbf{1}_K)\right| &\leq O_p(1/\sqrt{n}). \end{aligned} \tag{5}$$

The above proposition indicates that the difference between utility functions in data reuse stacking and CV_{ws} is order of magnitude smaller than the random fluctuation in $\hat{U}(w; K^{-1}\mathbf{1}_K)$, and in turn establishes the asymptotic equivalence of utility functions of stacking with data reuse and CV_{ws}. Since the results in Proposition 1 concerns uniform convergence, the near equivalence of $\hat{U}(w; K^{-1}\mathbf{1}_K)$ and $\hat{U}_{\text{ws}}(w; K^{-1}\mathbf{1}_K)$ can translate into asymptotic equivalence of $\hat{w}$ and $\hat{w}^{\text{ws}}$, provided the limit of $\hat{U}(w; K^{-1}\mathbf{1}_K)$ has a unique maximizer in W.

In Figure 2a, we plot the estimated $\left|\hat{U}(w; K^{-1}\mathbf{1}_K) - \hat{U}_{\text{ws}}(w; K^{-1}\mathbf{1}_K)\right|$ and $|\hat{U}(w; K^{-1}\mathbf{1}_K) - \lim_{n\to\infty} \hat{U}(w; K^{-1}\mathbf{1}_K)|$at $w = K^{-1}\mathbf{1}_K$ as a function of n. We set $K = 20$, $p = 10$ and $\beta_k \sim N(\mathbf{1}_p, I)$. We use a 5-fold CV_{ws}.

3.2.2 Cross-set CV and stacking with no data reuse

In this section, we focus on leave-one-set-out CV_{cs} where T iterations are performed. At each iteration a different D_t is held out. We first introduce this CV scheme when $\mathcal{D}$ is study-specific hence $T = K$:

1. Generate the library of SPFs $\hat{\mathcal{Y}}$ using every set in $\mathcal{D}$. Note that this library remains the same across T iterations.

2. At iteration t, evaluate the utility of w using D_t with SPFs that are not trained on D_t:

$$\hat{U}^{(t)}_{\text{cs}}(w; \nu) = \frac{1}{n_t} \sum_{i=1}^{n_t} u\left(\sum_{\ell,t} \mathbb{I}(t' \neq t) w_{\ell,t'} \hat{Y}^{\ell}_{t'}(x_{i,t}), y_{i,t}\right). \tag{6}$$

3. Combine all $\hat{U}_{\text{cs}}^{(t)}$ across T iterations and evaluated at a scaled w, yielding the utility function $\hat{U}_{\text{cs}}(w;\nu)$ for the selection of w in CV_{cs}:

$$\hat{U}_{\text{cs}}(w;\nu) = \sum_t \nu_t \hat{U}_{\text{cs}}^{(t)}(w;\nu) = \sum_{t=1}^{K} \frac{\nu_t}{n_t} \sum_i u\left(\sum_{\ell,t'} \frac{\mathbb{I}(t' \neq t)}{1-\nu_t} w_{\ell,t'} \hat{Y}_{t'}^{\ell}(x_{i,t}), y_{i,t} \right). \tag{7}$$

The scaling factor $(1-\nu_t)^{-1}$ is used to extrapolate the predicted value given by the full ensemble using the prediction from the partial ensemble. For example, in generalist predictions for exchangeable studies with $\nu_t = K^{-1}$, we expect that $\hat{Y}_k^{\ell}(x) \approx \hat{Y}_{k'}^{\ell}(x)$ for $k' \neq k$ and hence $(\sum_{k,\ell} \hat{Y}_k^{\ell}(x))/(\sum_{k\neq k',\ell} \hat{Y}_k^{\ell}(x)) \approx K/(K-1)$.

4. Let $\hat{w}^{\text{cs}} = \arg\max_{w\in W} \hat{U}_{\text{cs}}(w;\nu)$, CV_{cs} stacked PF is

$$\hat{Y}_w^{\text{cs}} = \sum_{\ell,t} \hat{w}_{\ell,t}^{\text{cs}} \hat{Y}_t^{\ell}.$$

To understand the rationale for (7), we consider applying CV_{cs} for generalist predictions. Note that under exchangeable P_k distributions, $k = 1, \ldots,$ $\hat{U}_{\text{cs}}^{(t)}(w)$ is an unbiased estimator of $U_g(w^{(t)})$, where $w^{(t)}$ is equal to w except for components $w_{\ell,t}$ $\ell = 1, \ldots, L$, which are set to zero:

$$\mathbb{E}\left[\hat{U}_{\text{cs}}^{(t)}(w)\right] = \mathbb{E}\left[U_g(w^{(t)})\right],$$

where the expectation is taken over $\mathcal{S}$. For $\{\nu_t; t \leq K\} \in \Delta_{K-1}$, it follows that

$$\mathbb{E}\left[\sum_{t=1}^{K} \nu_t \hat{U}_{\text{cs}}^{(t)}(w)\right] = \sum_{t=1}^{K} \nu_t \mathbb{E}\left[U_g(w^{(t)})\right].$$

Consider the Taylor expansion of $U_g(w^{(t)})$, $t = 1, \ldots, K$, around $\mathbf{0}$. Since $\sum_t \nu_t = 1$,

$$\sum_t \nu_t U_g(w^{(t)}) = U_g(\mathbf{0}) + \frac{\partial U_g}{\partial w^{\intercal}}(\mathbf{0}) \sum_t \nu_k w^{(t)} + o(\|w\|).$$

By construction $\sum_t \nu_t w^{(t)} = ((1-\nu_t) w_{\ell,t}; \ell = 1, \ldots, L, t = 1, \ldots, K)$. Let S be a $KL \times KL$ diagonal matrix with the diagonal term corresponding to $w_{\ell,t}$ as $1-\nu_t$, we have $\sum_t \nu_t w^{(t)} = Sw$.

Based on the results above, we know that

$$\mathbb{E}\left[\nu_t \sum_{t=1}^{K} \hat{U}_{\text{cs}}^{(t)}(w)\right] = \mathbb{E}U_g(\mathbf{0}) + \mathbb{E}\left[\frac{\partial U_g}{\partial w^\intercal}(\mathbf{0})\right] Sw + o(\|w\|).$$

If w is defined close to $\mathbf{0}$, e.g. $W = \{w : \|w\|_1 \leq 1\}$, the linear term in the above expansion dominates higher order terms of w and we have

$$\mathbb{E}\left(\sum_k \nu_k \hat{U}_{\text{cs}}^{(k)}(w)\right) = \mathbb{E}U_g(Sw) + o(\|w\|).$$

Therefore, a nearly unbiased estimator of $U_g(w)$ is

$$\hat{U}_{\text{cs}}(w;\nu) = \sum_k \nu_k \hat{U}_{\text{cs}}^{(k)}\left(S^{-1}w\right).$$

Expanding the above equation, we get the expression in (7).

In general, $\mathcal{D}$ is not necessarily study-specific and it might contains D_t's that overlap. The utility function for CV_{cs} in the general case can be constructed in the similar manner as when $\mathcal{D}$ is study-specific. In the first place, we modify $\hat{U}_{\text{cs}}^{(t)}(w)$, which estimates the expected utility of the PF combining the library of SPFs with weight $w^{(t)}$, into

$$\hat{U}_{\text{cs}}^{(t)}(w) = \frac{1}{|D_t|} \sum_{(i,k)\in D_t} u\left(\sum_{\ell,t'} \mathbb{I}(s_t \cap s_{t'} = \emptyset) w_{\ell,t'} \hat{Y}_{t'}^{\ell}(x_{i,k}), y_{i,k}\right),$$

where $s_t = \{k : (i,k) \in D_t \text{ for some } i = 1,\ldots,n_k\}$ is the list of studies with at least one sample in D_t. This modified $\hat{U}_{\text{cs}}^{(t)}$ guarantees no-date-reuse even if D_t's are overlapped, since the set of studies that are involved in evaluating the utility of the stacked PF is mutually exclusive to the set of studies that are used in training SPFs in the considered stacked PF.

With the no-data-reuse property, it follows that with exchangeable distributions $P_1, P_2, \ldots$

$$\mathbb{E}\left[\hat{U}_{\text{cs}}^{(t)}(w)\right] = \mathbb{E}\left[U_g(w^{(t)})\right],$$

where $w^{(t)}$ is equal to w except for all elements $w_{\ell,t'}$, such that $s_{t'} \cap s_t \neq \emptyset$ and $\ell = 1,\ldots,L$, which are equal to zero.

In the study-specific $\mathcal{D}$ scenario, each $\hat{U}_{\text{cs}}^{(t)}(w)$ is combined into $\hat{U}_{\text{cs}}(w;\nu)$ ac-

cording a relative importance ν_t. This relative importance can be interpreted as the total probability mass assigned to data in D_t in the empirical distribution of $\mathcal{S}$:

$$\hat{\pi}(x,y) = \sum_k \frac{\nu_k}{n_k} \sum_i \mathbb{I}\left((x,y) = (x_{i,k}, y_{i,k})\right).$$

With this definition, in the general case, the relative importance of D_t is $\gamma_t = \sum_{k \in s_t} \nu_k n_{k,t}/n_k$, where $n_{k,t}$ is the number of samples from study k that are present in D_t. As in the case for study-specific $\mathcal{D}$, we can use these γ_t to combine $\hat{U}^{(t)}_{\text{ws}}(w)$.

With Taylor expansion of $U_g(w^{(t)})$ around $\mathbf{0}$, we get

$$\mathbb{E}\left(\sum_t \gamma_t \hat{U}^{(t)}_{\text{cs}}(w)\right) = \left(\sum_t \gamma_t\right) \mathbb{E}U_g(\mathbf{0}) + \mathbb{E}\left[\frac{\partial U_g}{\partial w^\intercal}(\mathbf{0})\right] \sum_t \gamma_t w^{(t)} + o(\|w\|).$$

Let Γ be a $KL \times KL$ diagonal matrix with the element corresponds to $w_{\ell,t}$ equal to $\sum_{t'} \gamma_{t'} \mathbb{I}(s_{t'} \cap s_t = \emptyset) w_{\ell,t}$. By the definition of $w^{(t)}$, $\sum_t \gamma_t w^{(t)} = \Gamma w$. Therefore we have

$$\left(\sum_t \gamma_t\right)^{-1} \mathbb{E}\left(\sum_t \gamma_t \hat{U}^{(t)}_{cs}(w)\right) = \mathbb{E}U_g(\mathbf{0}) + \mathbb{E}\left[\frac{\partial U_g}{\partial w^\intercal}(\mathbf{0})\right] \frac{\Gamma w}{\sum_t \gamma_t} + o(\|w\|).$$

If the linear term dominates higher order terms in the above expansion, we have

$$\left(\sum_t \gamma_t\right)^{-1} \mathbb{E}\left(\sum_t \gamma_t \hat{U}^{(t)}_{cs}(w)\right) = \mathbb{E}U_g\left(\frac{\Gamma w}{\sum_t \gamma_t}\right) + o(\|w\|),$$

and again, an approximated unbiased estimator of $U_g(w)$ is

$$\left(\sum_t \gamma_t\right)^{-1} \left(\sum_t \gamma_t \hat{U}^{(t)}_{cs}\left(\sum_t \gamma_t \Gamma^{-1} w\right)\right).$$

Expand the above expression, we get the estimated utility function for CV_{cs} for a general $\mathcal{D}$:

$$\hat{U}_{\text{cs}}(w;\nu) = \sum_t \tilde{\gamma}_t |D_t|^{-1} \sum_{(i,k) \in D_t} u\left(\sum_{\ell,t'} \frac{\mathbb{I}\left(s_{t'} \cap s_t = \emptyset\right)}{r_t} w_{\ell,t'} \hat{Y}^{\ell}_{t'}(x_{i,k}), y_{i,k}\right), \quad (8)$$

where $\tilde{\gamma}_t = \gamma_t / (\sum_{t'} \gamma_t)$ and $r_t = \sum_{t'} \tilde{\gamma}_{t'} \mathbb{I}\left(s_{t'} \cap s_t = \emptyset\right)$.

An implicit assumption built in (8) is that none of the r_t's is zero. This is equivalent to the requirement that for each D_t, there exists at least one $D_{t'}$ that

contains a completely different list of studies than D_t. It is not too stringent if we only allow each D_t contains samples from a subset of studies.

We now apply $\hat{U}_{\text{cs}}(w;(1/2,1/2))$ to select w for the example in Section 3.1. Note in this example, $\mathcal{D}$ is study-specific. It follows that

$$\hat{U}_{\text{cs}}(w;(1/2,1/2)) = -\left(2w_1^2\bar{y}_1^2 = 2w_2^2\bar{y}_2^2 - 2\bar{y}_1\bar{y}_2 + \frac{\overline{y_1^2}+\overline{y_2^2}}{2}\right).$$

The maximzier of $\hat{U}_{\text{cs}}(w;(1/2,1/2))$ is $\hat{w}_{\text{cs}} = \big(\bar{y}_2^2/(\bar{y}_1^2+\bar{y}_2^2), \bar{y}_1^2/(\bar{y}_1^2+\bar{y}_2^2)\big)$. Like the oracle weights w_g, $\hat{w}_{\text{cs}}$ depend on $\bar{y}_1$ and $\bar{y}_2$. We can compare the cMSE of PFs specified by $\hat{w}$ and $\hat{w}_{\text{cs}}$:

$$\text{cMSE}(\hat{Y}_w^{\text{cs}}) - \text{cMSE}(\hat{Y}_{\hat{w}}) = \frac{-(\bar{y}_1^2-\bar{y}_2^2)^2(\bar{y}_1+\bar{y}_2)^2}{4(\bar{y}_1^2+\bar{y}_2^2)^2} \leq 0.$$

The equality holds if and only if $\bar{y}_1 = \bar{y}_2$. This comparison shows that CV_{cs} outperforms stacking with data-reuse for selecting generalist PFs.

In light of Proposition 1, which indicates the asymptotic equivalence of CV_{ws} to stacking with DR as $n \to \infty$ with K fixed, we will refer to CV_{cs} as stacking with NDR and will denote $\hat{U}_{\text{cs}}(w;\nu)$ as $\tilde{U}(w;\nu)$ and $\hat{w}_{\text{cs}}$ as $\tilde{w}$ in the remainder of this manuscript. The relative performance of CV_{cs} to stacking with DR, in a general setting, will be discussed in Proposition 3 and a condition under which CV_{cs} outperforms stacking with DR is illustrated.

Remark. CV_{ws} and CV_{cs} have their strengths and limitations. Datasets for selecting w in CV_{cs} are not used to generate $\hat{\mathcal{Y}}$ and are indeed "external". This is not the case for CV_{ws}. For example when $\mathcal{D}$ is study-specific, $\hat{Y}_{t,m}^{\ell}$ trained on $D_{t,-m}$ will be used to predict samples in $D_{t,m}$ from the same study t, which might still lead to optimistic estimation of $U_g(w)$, as observed for CV in model selection (Zhang, 1993). On the other hand, CV_{cs} at each iteration considers linear combinations of sets of SPFs with lower cardinality compared to $\hat{\mathcal{Y}}$, whose cardinality is TL. In addition, CV_{cs} cannot handle specialist predictions for certain types of $\mathcal{D}$. For example, $\hat{U}_{\text{cs}}(w;e_k)$ is not well defined if $\mathcal{D}$ is study-specific.

3.3 Penalization in stacking

Adding a penalty to the utility function $\hat{U}(w;\nu)$ is a common practice for selecting weights w in stacking (Breiman, 1996; LeBlanc and Tibshirani, 1996). Flexible forms of penalties on w can deal with a wide variety of relationships between SPFs in $\hat{\mathcal{Y}}$. For example, group LASSO can be used when SPFs can be organized into related groups. In this section, we leverage this flexibility for specialist predictions when n_k is small.

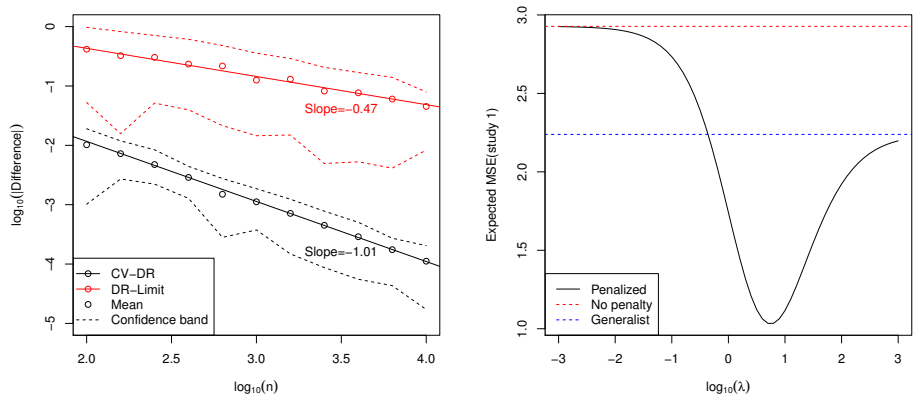
When n_k is small, the estimated prediction accuracy of a PF is highly variable. This disadvantage is further compounded by the fact that under certain conditions, specialist PFs fail to incorporate information from other studies. For instance, stacked PFs for specialist predictions derived from stacking with DR, when OLS regression serves as the single learner, do not put any weights on SPFs derived from studies other than the target study(see Section 3.1). To overcome the above disadvantage arising from small n_k, we introduce a penalized utility function that promotes shrinkage of specialist PFs towards generalist PFs.

The penalized $\hat{U}(w;e_k)$ is defined as follows.

$$\hat{w}_p^{(k)} = \underset{w\in W}{\arg\max}\, \hat{U}(w;e_k) - \lambda\|w-\hat{w}_g\|_2^2, \tag{9}$$

where $\lambda > 0$ is a tuning parameter, $\hat{w}_g = \arg\max_{w\in W} \hat{U}(w;\nu_g)$ and ν_g is a set of study weights used in generalist utility. We use a leave-one-out cross-validation to select the turning parameter λ. For sample i in study k, we generate $\hat{\mathcal{Y}}$ using data with this sample excluded. We then calculate the prediction error of the resulting stacked PF with weights $\hat{w}_p^{(k)}$ on sample i. This procedure is repeated over a set of candidate values for λ. We specify that the candidate values decrease as n_k increases.

In Figure 2b, we illustrate the effect of penalization in training of specialist PFs for a study with small sample size ($n_1 = 10$). As λ increases from 10^{-3}, the expected MSE in study 1, defined as $\int_{(x,y)}(y-\hat{Y}(x))^2 dP_1(x,y)$, of the penalized specialist PFs first decreases, indicating the benefit of shrinking the specialist weights towards the generalist weights. The expected MSE is minimized at $\lambda \approx 8$ and when λ increases beyond it, the expected MSE starts to increase. The details of the distribution assumptions for this example is describe in Section 5.1.



(a) (b)

Figure 2: (**a**) $|U(w; K^{-1}\mathbf{1}_K) - \hat{U}_{\text{ws}}(w; K^{-1}\mathbf{1}_K)|$ and $|U(w; K^{-1}\mathbf{1}_K) - \lim_{n\to\infty} U(w; K^{-1}\mathbf{1}_K)|$ at $w = K^{-1}\mathbf{1}_K$ as a function of n. A 5-fold CV_{ws} is used and we repeat the simulation for 40 times. CV indicates CV_{ws}, DR indicates stacking with data reuse. Both X-axis and Y-axis are $\log_{10}$ transformed. The dashed lines indicate the upper and lower fifth percentile of the differences. The solid lines illustrate the linear regression fitted lines of log-transformed difference on $\log_{10}(n)$. The slopes are labelled beside the two lines. (**b**) Effect of penalization on specialist prediction performance. λ controls the strength of the penalization. Larger λ shrinks the specialist predictions more towards generalist predictions. The two dashed lines indicate the expected MSE when $\lambda = 0$ (No penalty) and $\lambda \to \infty$ (Generalist). Note that the X-axis is $\log_{10}$ transformed.

4 Properties of generalist prediction models

We examine properties of generalist PFs $\hat{Y}_w$ when w are obtained with ($\hat{w}$) and without ($\tilde{w}$) data reuse when $W = \{\|w\|_1 \leq 1\}$. Recall that under the exchangeability assumption, $\hat{w} = \arg\max_{w\in W} \hat{U}(w; K^{-1}\mathbf{1}_K)$ and $\tilde{w} = \arg\max_{w\in W} \tilde{U}(w; K^{-1}\mathbf{1}_K)$. For the remainder of this manuscript, we will assume $\hat{u}(\hat{y}, y) = -(\hat{y} - y)^2$. We work under the assumption that the data generating distribution underlying the multi-study collection is a hierarchical model, and $\mathcal{D}$ will be study-specific. In the last part of this manuscript, we explore and discuss the results derived when this assumption is relaxed.

We present two properties of generalist predictors. First, the expected MSE of the generalist PFs in future $k > K$ studies, as determined by $\hat{w}$ and $\tilde{w}$, converge to the MSE of an oracle PF $Y_{w_g^0}$, and the discrepancy between the MSEs will be bounded by a monotone function of K and $\min_k n_k$. Second, we investigate under which circumstances stacking without data reuse has better MSE compared to stacking with data reuse.

The joint hierarchical model underlying available and future datasets is:

$$\begin{aligned} y_{i,k} &= f_k(x_{i,k}) + \epsilon_{i,k}, \\ f_k &\sim F,\ x_{i,k} \overset{iid}{\sim} F_X, \end{aligned} \tag{10}$$

for $i = 1, \ldots, n_k$ and $k = 1, 2, \ldots$. Here $f_k : \mathbb{R}^p \to \mathbb{R}$, $k \geq 1$, are iid random functions with marginal distribution F. The mean of F is indicated as $f_0 = \int f dF(f)$. Covariate vectors $x_{i,k} \in \mathcal{X}$ have the same distribution F_X with finite second moment across all datasets, and the noise terms $\epsilon_{i,k}$ are independent with mean zero and variance σ^2.

Our propositions 2 and 3 will assume that:

A1. There exists an $M_1 < \infty$ such that for any $k > 0$ and $\ell \leq L$,

$$\sup_{x \in \mathcal{X}} |f_k(x)| \leq M_1,\ a.e. \text{ and } \sup_{x \in \mathcal{X}} |\hat{Y}_k^\ell(x)| \leq M_1,\ a.e.$$

The first a.e. is with respect to the joint distribution of f_k whereas the second a.e. concerns the joint distribution of $\mathcal{S}$.

For example, if $\mathcal{X}$ is a compact set and outcomes Y are bounded, the SPFs trained with a linear regression model with a L_1 constraint on the regression coefficients, i.e. a LASSO regression model, or with tree-based regression models satisfy the assumption.

A2. There exist $M_2 < \infty$, $p_\ell > 0$ and functions $Y_{k,\ell}$ for $k = 1, \ldots, K$, $\ell = 1, \ldots, L$, such that $\sup_{x \in \mathcal{X}} |Y_k^\ell(x)| \leq M_1$ $a.e.$, and,

$$\int_x n_k^{2p_\ell} \left(\hat{Y}_k^\ell(x) - Y_k^\ell(x) \right)^2 dF_X(x) \leq M_2,$$

Here Y_k^ℓ is the limit of $\hat{Y}_k^\ell$ as n_k goes to infinity. For example, if the learner is an OLS model, then $p_\ell < 1/2$.

Let $X_k = (x_{1,k}, \ldots, x_{n_k,k})^\intercal$ and $Y_k = (y_{1,k}, \ldots, y_{n_k,k})$. The predicted outcomes for study k, based on a SPF $\hat{Y}_{k'}^\ell$, is denoted as $\hat{Y}_{k'}^\ell(X_k) = (\hat{Y}_{k'}^\ell(x_{i,k}); i \leq n_k)^\intercal$.

When $u(y, y') = -(y - y')^2$, we have

$$\begin{aligned}\hat{U}(w; K^{-1}\mathbf{1}_K) &= w^\intercal \hat{\Sigma} w - 2\hat{b}^\intercal w + K^{-1}\sum_k n_k^{-1}\sum_i y_{i,k}^2, \\ \tilde{U}(w; K^{-1}\mathbf{1}_K) &= w^\intercal \tilde{\Sigma} w - 2\tilde{b}^\intercal w + K^{-1}\sum_k n_k^{-1}\sum_i y_{i,k}^2,\end{aligned} \tag{11}$$

where $\hat{\Sigma}, \tilde{\Sigma}, \hat{b}$, and $\tilde{b}$ are defined as follows,

$$\hat{\Sigma}_{k,k';\ell,\ell'} = \sum_{i=1}^{K} \frac{\left(\hat{Y}_k^\ell(X_i)\right)^\intercal \hat{Y}_{k'}^{\ell'}(X_i)}{n_i K}, \quad b_{k;\ell} = \sum_{i=1}^{K} \frac{\left(\hat{Y}_k^\ell(X_i)\right)^\intercal Y_i}{n_i K},$$
$$\tilde{\Sigma}_{k,k';\ell,\ell'} = \sum_{i\neq k, i\neq k'} \frac{K\left(\hat{Y}_k^\ell(X_i)\right)^\intercal \hat{Y}_{k'}^{\ell'}(X_i)}{n_i(K-1)^2}, \quad \tilde{b}_{k;\ell} = \sum_{i\neq k} \frac{\left(\hat{Y}_k^\ell(X_i)\right)^\intercal Y_i}{n_i(K-1)}.$$

Note that $\hat{\Sigma}$ and $\tilde{\Sigma}$ are $KL \times KL$ matrices, w, $\hat{b}$ and $\tilde{b}$ are KL-dimensional vectors. $\hat{\Sigma}_{k,k';\ell,\ell'}$ is the element corresponding to $w_{k,\ell}$ and $w_{k',\ell'}$ while $\hat{b}_{k;\ell}$ is the element corresponding to $w_{k,\ell}$.

We define the oracle generalist stacking weights w_g^0 based on the limits of $\hat{Y}_k^\ell$:

$$w_g^0 = \arg\max_{w\in W} \int_{x,y} u(Y_w(x), y) dP_0(x, y),$$

where $Y_w = \sum_{\ell,t} w_{\ell,t} Y_t^\ell$ and P_0 is the average joint distribution of (X, Y) across studies $k \geq 1$. The cross-study MSE associated with a stacking weight w is defined as

$$\psi(w) = \int_{x,y} (y - Y_w(x)) dP_0(x, y) = w^\intercal \Sigma w - 2b^\intercal w + \int_y y^2 dP_0(y),$$

where $\Sigma_{k,k;\ell,\ell'} = \int_x Y_k^\ell(x) Y_{k'}^{\ell'} dF_X(x)$ and $b_{k,\ell} = \int_{x,y} y Y_k^\ell(x) dP_0(x, y)$.

4.1 Generalist models and oracle ensembles

In Proposition 2 we compare $\hat{Y}_{\hat{w}}$ and $\hat{Y}_{\tilde{w}}$ to oracle prediction, using the metrics $\mathbb{E}(\psi(\hat{w}) - \psi(w_g^0))$ and $\mathbb{E}(\psi(\tilde{w}) - \psi(w_g^0))$.

Proposition 2. *Let $L \geq 2$ and $u(x, y) = -(x-y)^2$. Consider K available datasets and future $k > K$ studies from model (10). If (A1) and (A2) hold, then*

$$\mathbb{E}\left(\psi(\hat{w}) - \psi(w_g^0)\right) \leq C_0\sqrt{\log(KL)}K^{-1/2} + C_1(\min_k n_k)^{-\min_\ell p_\ell},$$

and,

$$\mathbb{E}\left(\psi(\tilde{w}) - \psi(w_g^0)\right) \leq C_0'\sqrt{\log(KL)}K^{-1/2} + C_1'(\min_k n_k)^{-\min_\ell p_\ell},$$

where the expectations are taken over the joint distribution of the data $\mathcal{S}$*.* C_0*,* C_0'*,* C_1 *and* C_1' *are constants, independent of* K *and* n_k*.*

The above proposition shows that if we have enough studies and samples in each study, then the estimated generalist PFs $\hat{Y}_{\hat{w}}$ and $\hat{Y}_{\tilde{w}}$ have similar accuracy compared to $Y_{w_g^0}$.

4.2 Generalist predictions with and without data reuse

We compare the prediction accuracy, as indicated by $\psi(\hat{w})$ and $\psi(\tilde{w})$, of generalist PFs trained with and without data reuse. We start from a specific example, followed by a general result on the relative accuracy levels of PFs.

Consider $u(y, y') = -(y - y')^2$ and $L = 1$. Assume that $n_k = n$ and $f_k(x_{i,k}) = \beta_k^\intercal x_{i,k}$, where each component of β_k is an independent $U(\beta_0 - \tau, \beta_0 + \tau)$ random variable for $k = 1, \ldots, K$. Let each component in $x_{i,k} \in R^p$ be a $U(-\sqrt{3}, \sqrt{3})$ random variable and $\epsilon_{i,k}$ be *iid* $U(-\sqrt{3}, \sqrt{3})$ random variables for $i = 1, \ldots, n$, $k = 1, \ldots, K$. Let the learner be an OLS model, therefore $\hat{Y}_k^\ell(x) = \hat{\beta}_k^\intercal x$. Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$.

In this setting, we have $\Sigma_{k,k'} = \beta_k^\intercal \beta_{k'}$, $b_k = \beta_k^\intercal \beta_0$ and

$$\hat{\Sigma}_{k,k'} = (nK)^{-1} \sum_s \hat{\beta}_k^\intercal X_s^\intercal X_s \hat{\beta}_k,\ \hat{b}_k = (nK)^{-1} \sum_s \hat{\beta}_k^\intercal X_s^\intercal Y_s,$$
$$\tilde{\Sigma}_{k,k'} = \frac{K}{n(K-1)^2} \sum_{s \neq k,k'} \beta_k^\intercal X_s^\intercal X_s \hat{\beta}_k,\ \tilde{b}_k = (n(K-1))^{-1} \sum_{s \neq k} \hat{\beta}_k^\intercal X_s^\intercal Y_s.$$

To understand the behavior of $\hat{w}$ and $\tilde{w}$, we first consider the bias of $(\hat{\Sigma}, \hat{b})$ and $(\tilde{\Sigma}, \tilde{b})$ with respect to (Σ, b), as captured by the difference between their expectation over the joint distribution of the observed data $\mathcal{S}$:

$$\mathbb{E}(\hat{\Sigma}(k,k')) - \mathbb{E}(\Sigma(k,k')) = -\frac{p(p+1)}{Kn(n-p-1)}(1 - \delta_{i,j}),$$
$$\mathbb{E}(\tilde{\Sigma}(k,k')) - \mathbb{E}(\Sigma(k,k')) = \begin{cases} (K-1)^{-1}(\beta_0^\intercal \beta_0 + p\tau^2 + \frac{p}{n-p-1}) & i = j, \\ -(K-1)^{-2}\beta_0^\intercal \beta_0 & i \neq j, \end{cases}$$
$$\mathbb{E}(\hat{b}(k)) - b(k)) = \frac{p\tau^2 + p/n}{K},\ \mathbb{E}(\tilde{b}(k)) - b(k)) = 0.$$

The above equalities indicate that stacking without data reuse estimates off-diagonal elements of Σ without bias while zero-out stacking estimates b without bias. However, the equalities don't provide a direct comparisons of the relative performances of stacking with and without data reuse.

The next step is to derive an approximation of $\hat{w}$ and $\tilde{w}$ to compare the stacking procedures based on $\psi(w)$. One approximation considers the optimization of $\hat{U}(w; K^{-1}\mathbf{1}_K)$ and $\tilde{U}(w; K^{-1}\mathbf{1}_K)$ at the limit when $n \to \infty$. In this case, if $W = \mathbb{R}^K$

$$\hat{w} \approx K^{-1}\mathbf{1}_K, \quad \tilde{w} \approx \frac{K-2}{K-1}K^{-1}\mathbf{1}_K + \frac{1}{K-1}\left(\frac{1}{K}S\boldsymbol{\beta}^\intercal\boldsymbol{\beta} - \frac{K-1}{K}I\right)\mathbf{1}_K,$$

where $S = \text{diag}\{\|\beta_1\|_2^{-2}, \ldots, \|\beta_K\|_2^{-2}\}$. Note that each component of $(S\boldsymbol{\beta}\boldsymbol{\beta}^\intercal/K - (K-1)/KI)\mathbf{1}_K$ decreases as τ increases. When $\tau = 0$, each component is equal to K^{-1} whereas when $\tau \to \infty$, the limit is approximately $-(K-1)/K$. We can find $\tilde{w}$ is a shrunk version of $\hat{w}$ towards zero. For study with larger β_k, the strength of shrinkage for w_k tends to be larger. A Monte-Carlo simulation determines that based on the above approximations, $\mathbb{E}(\psi(\hat{w})) > \mathbb{E}(\psi(\tilde{w}))$ when $\tau \gtrsim \sqrt{K}$. This bound is verified by a simulation study (see Figure 3).

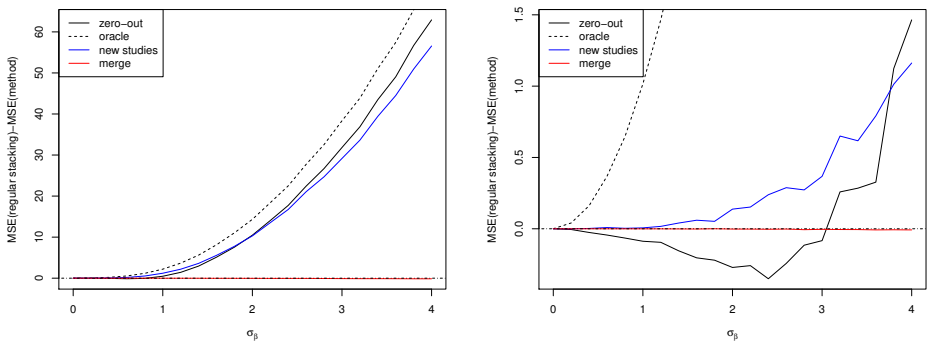


Figure 3: Comparison of stacking PFs constructed with and without data reuse when $K = 2$ (left) and $K = 9$ (right). We set $p = 10$, $\beta_0 = 1_K$, $n = 200$, $\sigma = 1$ and vary τ^2. The difference in $\mathbb{E}\psi_K(w)$ is calculated with 1,000 replications. Oracle is the predictor $\hat{Y}_{w_g}$, "new study" means we train weights with a set of new studies that are not used in constructing $\hat{\mathcal{X}}$ and "merge" means we merge all studies to train a single regression model to serve as the generalist PF.

Motivated by the simulation results, we investigate under what circumstances $\mathbb{E}\psi(\tilde{w})$ is smaller than $\mathbb{E}\psi(\hat{w})$. We present our results in Proposition 3 about the characterization of the relative performance of $\hat{Y}_{\hat{w}}$ and $\hat{Y}_{\tilde{w}}$ in a general setting,

when data are generated from the model (10).

Proposition 3. *Assume the data are generated via* (10) *with* $n_k = n$ *and assumptions A1-A2 hold. Denote*

$$\sigma_f^2 = \int_f \left(\int (f(x) - f_0(x)) dF_X(x) \right)^2 dF(f).$$

There exists $\kappa > 0$ *such that when*

$$8\sqrt{e}(2M_1^2 + M_1\sigma_f)\sqrt{\log((K-1)L)}((K-1)L)^{-1/2} \leq \kappa M_1 \sigma_f \sqrt{\log(KL)}(KL)^{-1/2}, \tag{12}$$

$\mathbb{E}(\psi(\hat{Y}_{\hat{w}})) + C^* n^{-\min_\ell p_\ell} \geq \mathbb{E}(\psi(\hat{Y}_{\tilde{w}}))$, *where* $\mathbb{E}$ *is taken over* $\mathcal{S}$.

σ_f is a metric to measure the heterogeneity across studies since the only difference of one study to the other, based on our model assumption, is $\mathbb{E}(y_{i,k}|x_{i,k}) = f(x_{i,k})$. Note that $(K-1)\log(KL)/(K\log((K-1)L))$ increases as K increases when K is small and starts to decrease to 1 when K gets large. Therefore, if

$$\frac{8\sqrt{e}(2M_1^2 + M_1\sigma_f)}{\kappa M_1 \sigma_f} \leq 1,$$

then $\mathbb{E}\psi(\tilde{w})$ is always smaller than $\mathbb{E}\psi(\hat{w})$ up to a term $C^* n^{-\min_\ell p_\ell}$. If the ratio is larger than 1, only K that is small enough to satisfy (12) can guarantee the superiority of $\tilde{w}$. We also note that $8\sqrt{e}(2M_1^2 + M_1\sigma_f)/(\kappa M_1\sigma_f)$ is a decreasing function in σ_f. This means if σ_f is large, the upper bound for K such that $\mathbb{E}\psi(\tilde{w}) \leq \mathbb{E}\psi(\hat{w})$ will increase.

The proposition provides a rough guideline to select between stacking with and without data reuse. If the number of studies are relatively small, we would prefer stacking without data reuse to stacking with data reuse, as the former outperforms the latter even with low σ_f. On the other hand, when K is large, we might turn to stacking with data reuse more often unless there is strong evidence indicating σ_f is extremely high.

In Figure 6, we examine the relative performance of two stacking approaches across a range of K and cross-study heterogeneity with a simulation study. We can see stacking without data reuse outstrips stacking with data reuse exclusively when τ is above a threshold defined by a function of $\sqrt{K}$. Only when σ

is small with moderate K, stacking with data reuse shows significant advantage over stacking without data reuse.

A more clear-cut recommendation based on Proposition 3 can be challenging since M_1 and κ are unknown and σ_f is also not observed. However, one can adapt a non-parametric model to approximate f_k within each study and estimate these quantities to refine the rough guideline above, which might only be appropriate if n_k is large.

5 Simulation studies

In this section, we first illustrate the effectiveness of the technique in Section 3, proposed for specialist predictions of small studies, through simulated datasets. We then examine the analytical results in Section 4 using numerical examples. We investigate empirically whether the error bound of the estimated stacking predictors in Proposition 2 is tight, and verify that the preferable region of stacking with NDR in comparison to with DR is aligned with our theoretical characterization. We conclude this section with an example illustrating how to extend generalist predictions to non-exchangeable studies.

5.1 Specialist predictions for small studies

We specify the following generative model for the simulated dataset to examine the performance of the specialist predictor derived from the modified utility function (9). In addition, we also consider the performance of the generalist predictor derived based on the utility function $\hat{U}(\hat{Y}_w; 1/K 1_K)$ and the specialist predictor without small-sample based penalization.

$$y_{i,k} = \beta_k^\intercal x_{i,k} + \epsilon_{i,k},$$
$$\beta_k \sim N(1_p, I_p),\ \epsilon_{i,k} \sim N(0, 25),$$

where p is the number of covariates, 1_p is a p-vector of ones and I_p is the $p \times p$ identity matrix. We set $p = 10$ and $K = 5$ with $n_k = 100$ for $k = 2, \ldots, 5$ and n_1 varying from 10 to 50. In Figure 4, we illustrate the RMSEs of the three predictors in consideration when applied to predict new samples in study 1. We set D_t to be all data from study t, $t = 1, \ldots, K$ and ordinary least square (OLS)

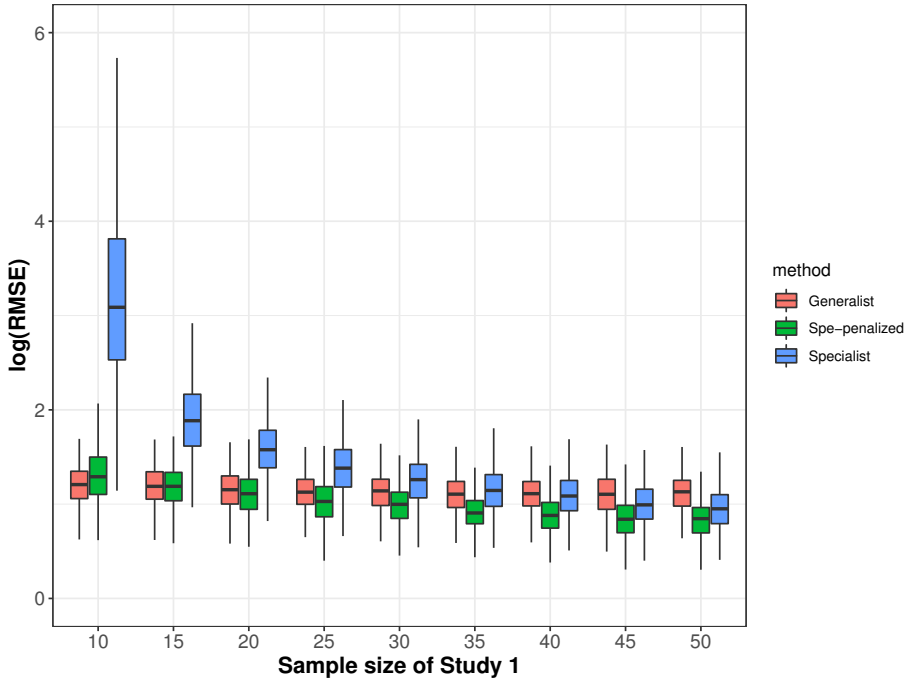


Figure 4: *RMSE of the specialist predictor, the generalist model and the penalized specialist predictor on future samples in study 1.*

regression is the single-set learner. We use negative squared loss as $u(\cdot, \cdot)$ and set $w(n) = 100/n$.

From the results we can see that when n_1 is small (< 40), the generalist predictor outperforms unpenalized specialist predictor. This is as expected since all five studies are similar to each other. The penalized specialist predictor, on the other hand, is not sensitive to n_1 and has the lowest RMSE (except for $n_1 = 10$) among all three predictors.

5.2 Error bound of generalist stacking predictors

We illustrate the difference in the prediction error, $\mathbb{E}\left(\psi(\hat{Y}_{\hat{w}_g})\right) - \mathbb{E}\left(\psi(Y_{w_g^0})\right)$, considered in Proposition 1 with a numeric example and compare the actual difference to the analytic upper bound as n_k and K change. We use a similar generative model specification as in Section 5.1 but specify that each component of β_k follows $U[0, 1]$ and each component of $x_{i,k}$ follows $U[-1, 1]$. In addition, we assume that $\epsilon_{i,k} \sim U[-1, 1]$. The reason to replace the normal distributions with uniform distributions is to satisfy the boundedness assumption for g_k and $\hat{f}_k$. We set $n_k = n$ for all k and calculate with Monte Carlo simulation the difference $\mathbb{E}\left(\psi(\hat{Y}_{\hat{w}_g})\right) - \mathbb{E}\left(\psi(Y_{w_g^0})\right)$ for $n = 100, 200, 400$ as K increases from 5 to 50 with

increment 1 and for $K = 5, 15, 20$ as n increases from 20 to 100 with increment 5. We use the same constraint on w, i.e. $\|w\|_1 \leq 1$ as in the proposition. The results are derived from 1000 simulation replicates and shown in Figure 5.

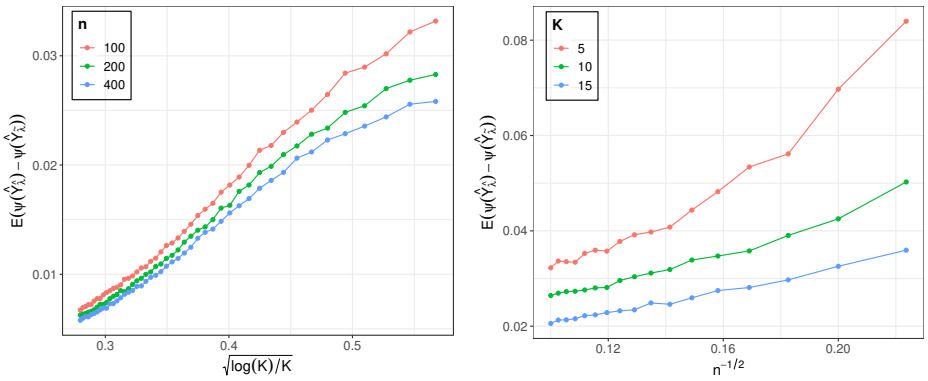


Figure 5: $\mathbb{E}(\psi(\hat{Y}_{\hat{w}_g}) - \psi(Y_{w_g^0}))$ *as a function of* K *and* n*. Note that the values on X-axis in both plots increase as* K *or* n *decreases.*

From the figures we can find except for small K and n, the difference in ψ is approximately a linear function of both $\sqrt{\log K/K}$ or $n^{-1/2}$ when fixing n or K. This indicates that the actual difference in ψ changes at the same order as the upper bound we discovered in Proposition 1. Indeed, under this simulation scenario, the above results implies the upper bound is probably tight and we cannot improve the results about the convergence rate of $\hat{Y}_{\hat{w}_g}$ to $Y_{w_g^0}$.

5.3 Comparison between stacking with and without data reuse

We also perform a simulation analysis to check if the transition bound provided in Proposition 2 correctly delimits the region where stacking without data reuse supersedes stacking with data reuse. We use the same simulation scenario as in Section 5.2 but vary the variance of β_k by changing the range of the corresponding uniform distributions and number of studies K. We then calculate the prediction accuracy on future studies of the generalist predictors derived with stacking with and without data reuse under the constraint that $\|w\|_1 \leq 1$.

5.4 Non-exchangeable studies

To conclude the simulation study section, we present a numeric experiment where we illustrate the flexibility of the stacking approach to incorporate non-exchangeable studies. Specifically, we assume that K studies are collected at time point $t_k = k$

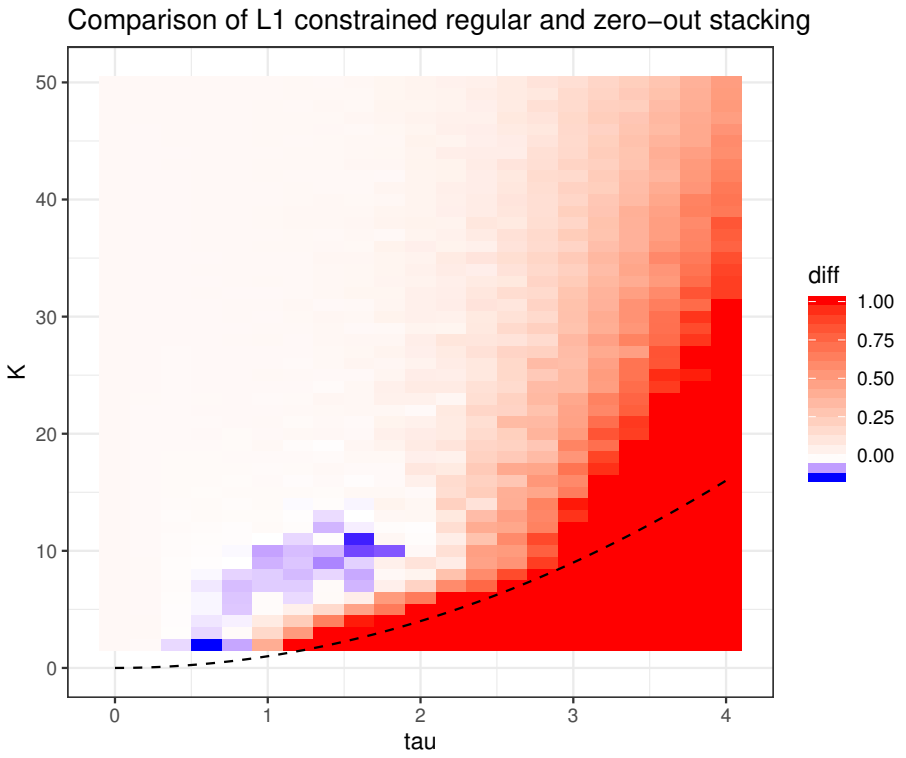


Figure 6: *Difference between standard stacking predictor and zero-out predictor in terms of out-of-study prediction error. Both stacking approaches are constrained with* $\|w\|_1 \leq 1$. *The dash line indicates the upper bound on* K *within which stacking without data reuse has better prediction accuracy.*

and the study-specific regression coefficients β_k follows an AR1 model.

$$\beta_k = \rho\beta_{k-1} + \sqrt{1-\rho^2}\epsilon_k,$$

where ρ is a constant between 0 and 1, which indicates the dependence between studies that are collected at close proximity in time. ϵ_k are independent normal noise with mean zero and covariance matrix I_p. Once we simulate β_k, we use the same generative model for $x_{i,k}$ and $y_{i,k}$ as in Section 5.1.

To account for the non-exchangeability between studies, we set the study-specific weight ν_k based on the distance between study k and the future study, which is assumed to be collected at time $K+1$. Specifically, we assume $\nu_k = 1/(K+1-k)$. The choice here is rather arbitrary but it incorporates the fact that most recent studies will be emphasized when training the generalist predictors for study $K+1$. The performance of this particular choice of stacking weights in the simulated dataset is shown in Figure 7. We consider three different values of ρ, correpsonding to high, medium and no dependence between studies.

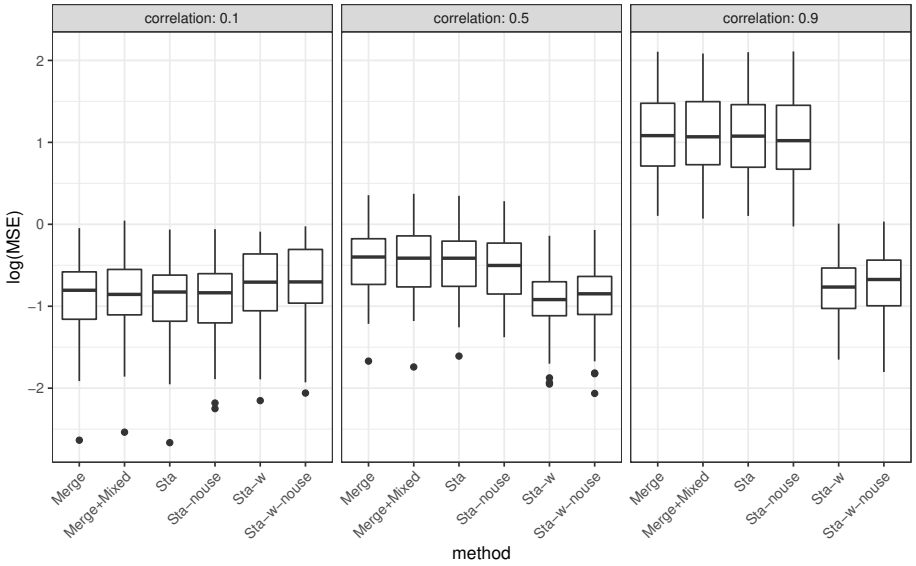


Figure 7: Comparsions of methods when studies are generated using an AR-1 model. ρ indicates the correlation between β_k of two adjacent time points.

6 Application

We apply our generalist predictors on an environmental health dataset containing observed mortality rate (person-year) across 31,414 unique zip codes in the entire U.S. For each ZIP code, the mortality rate is available from 1999 to 2016. The exposure to air pollution agents, such as PM2.5, is calculated for each ZIP code as the average observed levels from 1998 to 1999. In addition to the measurement of air pollution agents and the outcome, we also have access to ZIP code-level demographic covariates. All these covariates are measured before 1999. Demographic covariates consists of temperature, humidity, percentage of ever smokers, black population, median household income, median value of housing, percentage below the poverty level, percentage less than high school education, percentage of owner-occupied housing units, and population density.

We define the generalist prediction task in this dataset as the prediction of ZIP code-level mortality for a state based on data from other states and the specialist prediction task as the prediction for a specific state based on all available data. For generalist predictions, we randomly select 10 states to train an ensemble of state-specific prediction models and use our stacking approaches to combine these models to predict mortality rate for the rest of states. The metric we use to evaluate the performance of the stacked model is the average RMSE across all 40 testing states. We repeat this procedure 20 times. We consider two different

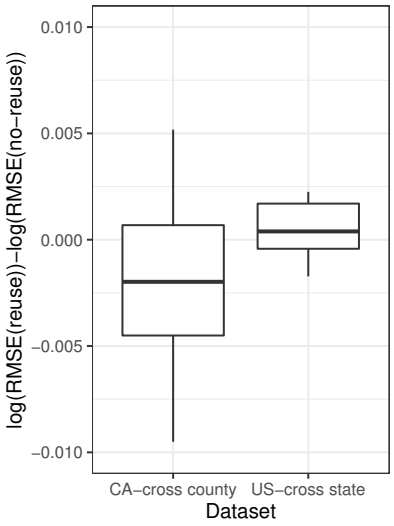


Figure 8: *Comparisons of the performance of stacking-based methods to merging. Stacking-based methods are derived with $L1$ penalty with or without data reuse. Each boxplot illustrates the variability of the prediction accuracy, evaluated with RMSE, of a specific method. The variability for generalist predictions is estimated through 20 replicates of random partitioning of training and testing states. The variability for specialist predictions is estimated through 10-fold cross-validation.*

approaches for generalist problems: stacking with DR and stacking with NDR. The same analysis is then performed for the county-level dataset from California. For this dataset, the number of testing counties are 48. The results are shown in 8.

From the figure we can find that when the dataset contains all states, supposedly with higher between studies heterogeneity, the performance of NDR is slightly better than that of DR, whereas if the dataset contains only county-level studies in California, which has smaller cross-study heterogeneity than the nationwide dataset, the advantage of NDR disappears and DR now has a smaller RMSE than NDR. This result is consistent with what we find in Proposition 3, which indicates for a fixed K, NDR only outperforms DR when the cross-study hetergeneity is large.

Acknowledgements

Research supported by the U.S.A.'s National Science Foundation grant nsf-dms1810829 (BR, LT and GP) and the U.S.A.'s National Institutes of Health grant NCI-5P30CA006516-54 (GP).

References

Babapulle, M. N., L. Joseph, P. Bélisle, J. M. Brophy, and M. J. Eisenberg (2004). A hierarchical bayesian meta-analysis of randomised clinical trials of drug-eluting stents. *The Lancet 364*(9434), 583–591.

Breiman, L. (1996). Stacked regressions. *Machine learning 24*(1), 49–64.

Higgins, J. P., S. G. Thompson, and D. J. Spiegelhalter (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(1), 137–159.

Juditsky, A. and A. Nemirovski (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*, 681–712.

Juditsky, A., P. Rigollet, A. B. Tsybakov, et al. (2008). Learning by mirror averaging. *The Annals of Statistics 36*(5), 2183–2206.

Kannan, L., M. Ramos, A. Re, N. El-Hachem, Z. Safikhani, D. M. Gendoo, S. Davis, D. Gomez-Cabrero, R. Castelo, K. D. Hansen, et al. (2016). Public data and open source tools for multi-assay genomic investigation of disease. *Briefings in bioinformatics 17*(4), 603–615.

Klein, R., K. Ratliff, M. Vianello, R. Adams Jr, S. Bahník, M. Bernstein, K. Bocian, M. Brandt, B. Brooks, C. Brumbaugh, et al. (2014). Data from investigating variation in replicability: A ?many labs? replication project. *Journal of Open Psychology Data 2*(1).

LeBlanc, M. and R. Tibshirani (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association 91*(436), 1641–1650.

Loewinger, G. C., K. K. Kishida, P. Patil, and G. Parmigiani (2019). Covariate-profile similarity weighting and bagging studies with the study strap: Multi-study learning for human neurochemical sensing. *bioRxiv*, 856385.

Manzoni, C., D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis, and R. Ferrari (2018). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics 19*(2), 286–302.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera (2012). A unifying view on dataset shift in classification. *Pattern recognition 45*(1), 521–530.

Pasolli, E., D. T. Truong, F. Malik, L. Waldron, and N. Segata (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS computational biology 12*(7), e1004977.

Patil, P., P.-O. Bachant-Winner, B. Haibe-Kains, and J. T. Leek (2015). Test set bias affects reproducibility of gene signatures. *Bioinformatics 31*(14), 2318–2323.

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

Rhodes, D. R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A. M. Chinnaiyan (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences 101*(25), 9309–9314.

Shumway, R. H. and D. S. Stoffer (2017). *Time series analysis and its applications: with R examples*. Springer.

Simon, R., M. D. Radmacher, K. Dobbin, and L. M. McShane (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute 95*(1), 14–18.

Sinha, R., G. Abu-Ali, E. Vogtmann, A. A. Fodor, B. Ren, A. Amir, E. Schwager, J. Crabtree, S. Ma, C. C. Abnet, et al. (2017). Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (mbqc) project consortium. *Nature biotechnology 35*(11), 1077.

Sutton, A. J. and J. P. Higgins (2008). Recent developments in meta-analysis. *Statistics in medicine 27*(5), 625–650.

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

van der Laan, M. J., S. Dudoit, and A. W. van der Vaart (2006). The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions 24*(3), 373–395.

Warn, D., S. Thompson, and D. Spiegelhalter (2002). Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in medicine 21*(11), 1601–1623.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks 5*(2), 241–259.

Zhang, P. (1993). Model selection via multifold cross validation. *The annals of statistics*, 299–313.

Zhang, Y., W. E. Johnson, and G. Parmigiani (2019). Robustifying genomic classifiers to batch effects via ensemble learning. *bioRxiv*, 703587.

A Proof of Proposition 1

Partition each study evenly into M pieces and denote the covariate matrix for the m-th piece in study k as $X_{k,m}$, the corresponding responses as $Y_{k,m}$. Let $X_{k,-m}$ and $Y_{k,-m}$ denote the entire covariate matrix and outcome vector for study k, excluding $X_{k,m}$ and $Y_{k,m}$. At iteration m, CV_{cs} with study-specific $\mathcal{D}$ fits an OLS model to each study based on $(X_{k,-m}, Y_{k,-m})$. We denote the estimated regression coefficients as

$$\hat{\beta}_{k,m} = (X_{k,-m}^\intercal X_{k,-m})^{-1} X_{k,-m}^\intercal Y_{k,-m}.$$

The utility function for CV_{ws} can be written as

$$\hat{U}_{\text{ws}}(w; K^{-1}\mathbf{1}_K) = (Kn)^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \left\| Y_{k,m} - \sum_{k'=1}^{K} w_{k'} X_{k,m} \hat{\beta}_{k',m} \right\|_2^2 = w^\intercal \Sigma_1 w - 2b_1^\intercal w + \sum_k \|Y_k\|_2^2.$$

The utility function for data reuse stacking is

$$\hat{U}(w; K^{-1}\mathbf{1}_K) = (Kn)^{-1} \sum_{k=1}^{K} \left\| Y_k - \sum_{k'=1}^{K} w_{k'} X_k \hat{\beta}_{k'} \right\|_2^2 = w^\intercal \Sigma_2 w - 2b_2^\intercal w + \sum_k \|Y_k\|_2^2,$$

where $\hat{\beta}_k = (X_k^\intercal X_k)^{-1} X_k^\intercal Y_k$ is the OLS estimate of regression coefficients based on all data from study k.

Σ_1 and Σ_2 are both $K \times K$ matrices and the (i, i')-th element of them are

$$\Sigma_1(i, i') = (Kn)^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \hat{\beta}_{i,m}^\intercal X_{k,m}^\intercal X_{k,m} \hat{\beta}_{i',m}$$
$$\Sigma_2(i, i') = (Kn)^{-1} \sum_{k=1}^{K} \hat{\beta}_i^\intercal X_k^\intercal X_k \hat{\beta}_{i'}.$$

b_1 and b_2 are K-dimensional vectors with the i-th elements

$$b_1(i) = (Kn)^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \hat{\beta}_{i,m}^\intercal X_{k,m}^\intercal Y_{k,m}, \quad b_2(i) = (Kn)^{-1} \sum_{k=1}^{K} \hat{\beta}_i^\intercal X_k^\intercal Y_k.$$

Note that we have the following relationship between $\hat{\beta}_{k,m}$ and $\hat{\beta}_k$:

$$\hat{\beta}_{k,m} = \hat{\beta}_k + (X_k^\intercal X_k)^{-1} X_{k,m}^\intercal (I - P_{k,m})^{-1} (X_{k,m} \hat{\beta}_k - Y_{k,m}), \tag{13}$$

where $P_{k,m} = X_{k,m}(X_k^\intercal X_k)^{-1} X_{k,m}^\intercal$. With the assumptions about the distribution of data and central limit theorem, we have the following characterization:

$$\begin{aligned}
\frac{1}{n} X_k^\intercal X_k &= I_p + O_p(1/\sqrt{n}) \\
\frac{1}{n} X_{m,k}^\intercal X_{m,k} &= \frac{1}{M} I_p + O_p(1\sqrt{n}) \\
\frac{1}{n} X_{m,k}^\intercal Y_{m,k} &= \frac{1}{M} \beta_k + O_p(1\sqrt{n}) \\
\hat{\beta}_k &= \beta_k + O_p(1/\sqrt{n}) \\
\hat{\beta}_{k,m} &= \hat{\beta}_k + O_p(1/\sqrt{n}) \\
P_{k,m} &= O_p(1/n).
\end{aligned} \tag{14}$$

Define $\delta\hat{\beta}_{k,m} = \hat{\beta}_{k,m} - \hat{\beta}_k$. We have

$$\begin{aligned}
n^{-1} \hat{\beta}_{i,m}^\intercal X_{k,m}^\intercal X_{k,m} \hat{\beta}_{i',m} =& \hat{\beta}_i (n^{-1} X_{k,m}^\intercal X_{k,m}) \hat{\beta}_{i'} + \delta\hat{\beta}_{i,m}^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \hat{\beta}_{i'} \\
&+ \hat{\beta}_i^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \delta\hat{\beta}_{i',m} + O_p(1/n).
\end{aligned}$$

Note that by (13) and (14)

$$\begin{aligned}
\hat{\beta}_i^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \delta\hat{\beta}_{i',m} =& \beta_i^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \delta\hat{\beta}_{i',m} + O_p(n^{-1}) \\
=& \beta_i^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) (X_{i'}^\intercal X_{i'})^{-1} X_{i',m}^\intercal (I_{n/M} - P_{i',m})^{-1} (X_{i',m} \hat{\beta}_{i'} - Y_{i',m}) + O_p(n^{-1}) \\
=& \beta_i^\intercal (I + O_p(1/\sqrt{n})) (X_{i'}^\intercal X_{i'})^{-1} X_{i',m}^\intercal (X_{i',m} \hat{\beta}_{i'} - Y_{i',m}) + O_p(1/n).
\end{aligned}$$

Therefore

$$\sum_m \hat{\beta}_i^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \delta \hat{\beta}_{i',m} = \beta_i^\intercal (I + O_p(1/\sqrt{n}))(X_{i'}^\intercal X_{i'})^{-1} \left((X_{i'}^\intercal X_{i'}) \hat{\beta}_{i'} - X_{i'}^\intercal Y_{i'} \right) + O_p(1/n) = O_p(1/n),$$

by the definition of $\hat{\beta}_{i'}$. Similarly, we get $\sum_m \delta \hat{\beta}_{i,m}^\intercal (n^{-1} X_{k,m}^\intercal X_{k,m}) \hat{\beta}_{i'} = O_p(1/n)$. And $|\Sigma_1(i,i') - \Sigma_2(i,i')| = O_p(1/n)$ for every $i, i' \leq K$. The same procedure can be applied to prove $|b_1(i) - b_2(i)| = O_p(1/n)$ by noting that $n^{-1} X_{k,m}^\intercal Y_{k,m} = 1/M \beta_k + O_p(1/\sqrt{n})$. Since w is defined on a bounded set: $\|w\| \leq C$ and K is fixed and finite, we immediately get that for all $w \in W$

$$|w^\intercal (\Sigma_1 - \Sigma_2) w - 2(b_1 - b_2)^\intercal w| \leq C O_p(1/n).$$

B Proof of Proposition 2

Define $\hat{\psi}(w)$ as

$$\hat{\psi}(w) = \hat{U}(w; K^{-1} 1_K) - \frac{\sum_k Y_k^\intercal Y_k}{n_k K} + \int_y y^2 dP_0(y).$$

Similarly, define $\tilde{\psi}(w) = \tilde{U}(w; K^{-1} 1_K) - K^{-1} \sum_k n_k^{-1} Y_k^\intercal Y_k + \int_y y^2 dP_0(y)$.

We first note the following lemma for the upper bounds of two differences $|\psi(\hat{w}) - \psi(w_g^0)|$ and $|\psi(\hat{w}) - \psi(w_g^0)|$.

Lemma 1. *$|\psi(\hat{w}) - \psi(w_g^0)|$ and $\psi(\tilde{w}) - \psi(w_g^0)$ can be bounded as follows.*

$$|\psi(\hat{w}) - \psi(w_g^0)| \leq 2 \sup_{w \in W} |\psi(w) - \hat{\psi}(w)|,$$
$$|\psi(\tilde{w}) - \psi(w_g^0)| \leq 2 \sup_{w \in W} |\psi(w) - \tilde{\psi}(w)|.$$

Proof. We prove the inequality for $\hat{w}$ and similar steps can be followed to verify the other inequality. Note that

$$\psi(\hat{w}) - \psi(w_g^0) = \psi(\hat{w}) - \hat{\psi}(\hat{w}) + \hat{\psi}(\hat{w}) - \hat{\psi}(w_g^0) + \hat{\psi}(w_g^0) - \psi(w_g^0).$$

By definition $\psi(\hat{w}) - \psi(w_g^0) \geq 0$ and $\hat{\psi}(\hat{w}) - \hat{\psi}(w_g^0) \leq 0$, therefore

$$|\psi(\hat{w}) - \psi(w_g^0)| \leq |\psi(\hat{w}) - \hat{\psi}(\hat{w})| + |\hat{\psi}(w_g^0) - \psi(w_g^0)| \leq 2 \sup_{w \in W} |\psi(w) - \hat{\psi}(w)|.$$

□

When $W = \{w : \|w\|_1 \leq 1\}$, we have

$$\sup_{w \in W} |\psi(w) - \hat{\psi}(w)| \leq \|\text{vec}(\Sigma - \hat{\Sigma})\|_\infty + \|b - \hat{b}\|_\infty,$$

where $\|\cdot\|_\infty$ is the L^∞-norm of a vector and vec$(\cdot)$ is the vectorization of a matrix. With Lemma 1, it follows

$$\mathbb{E}[\psi(\hat{w}) - \psi(w_g^0)] \leq 2\mathbb{E}\|\text{vec}(\Sigma - \hat{\Sigma})\|_\infty + 2\mathbb{E}\|b - \hat{b}\|_\infty. \tag{15}$$

The following lemma provides an upper bound for $\mathbb{E}\|\text{vec}(\Sigma - \hat{\Sigma})\|$ and $\mathbb{E}\|\text{vec}(\Sigma - \tilde{\Sigma})\|$.

Lemma 2. *If assumption A1 and A2 hold, we have the following bounds for* $\mathbb{E}\|vec(\Sigma - \hat{\Sigma})\|_\infty$ *and* $\mathbb{E}\|vec(\Sigma - \tilde{\Sigma})\|_\infty$.

$$\mathbb{E}\|vec(\Sigma - \hat{\Sigma})\|_\infty \leq 4\sqrt{2e}M_1^2\sqrt{\log(KL)/K} + 2M_1M_2(\min_k n_k)^{-min_\ell p_\ell},$$
$$\mathbb{E}\|vec(\Sigma - \tilde{\Sigma})\|_\infty \leq 8\sqrt{2e}M_1^2\sqrt{\log(KL)/K} + 4M_1M_2(\min_k n_k)^{-min_\ell p_\ell}.$$

Proof. First note that

$$\hat{A}_{k,\ell;k',\ell'} - A_{k,\ell;k',\ell'} = K^{-1}\sum_{i=1}^{K} n_s^{-1} \sum_{i=1}^{n_s} \left(\hat{Y}_k^\ell(x_{s,i})\hat{Y}_{k'}^{\ell'}(x_{s,i}) - \int_x Y_k^\ell(x) Y_{k'}^{\ell'}(x) dF_X(x) \right).$$

Denote $\int_x Y_k^\ell(x) Y_{k'}^{\ell'} k'^{\ell'}(x) dF_X(x)$ as $\langle Y_k^\ell, Y_{k'}^{\ell'} \rangle$, we have

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_\infty \leq & K^{-1}\sum_s n_s^{-1} \sum_i \left\| \left(\hat{Y}_k^\ell(x_{s,i}) \left(\hat{Y}_{k'}^{\ell'}(x_{s,i}) - Y_{k'}^{\ell'}(x_{s,i}) \right) ; k, k' \leq K, l, l' \leq L \right) \right\|_\infty \\
& + K^{-1}\sum_s n_s^{-1} \sum_i \left\| \left(Y_{k'}^{\ell'}(x_{s,i}) \left(\hat{Y}_k^\ell(x_{s,i}) - Y_k^\ell(x_{s,i}) \right) ; k, k' \leq K, l, l' \leq L \right) \right\|_\infty \\
& + K^{-1} \left\| \left(\sum_s n_s^{-1} \sum_i \left(Y_k^\ell(x_{s,i}) Y_{k'}^{\ell'}(x_{s,i}) - \langle Y_k^\ell, Y_{k'}^{\ell'} \rangle \right) ; k, k' \leq K, l, l' \leq L \right) \right\|_\infty
\end{aligned} \tag{16}$$

By assumption A1, we have

$$\left|\hat{Y}_k^{\ell}(x_{s,i})\left(\hat{Y}_{k'}^{\ell'}(x_{s,i})-Y_{k'}^{\ell'}(x_{s,i})\right)\right| \leq M_1|\hat{Y}_{k'}^{\ell'}(x_{s,i})-Y_{k'}^{\ell'}(x_{s,i})|,\ a.e.$$

Combined with assumption $A2$, we have

$$\mathbb{E}\left\|\left(\hat{Y}_k^{\ell}(x_{s,i})\left(\hat{Y}_{k'}^{\ell'}(x_{s,i})-Y_{k'}^{\ell'}(x_{s,i})\right);k,k'\leq K,l,l'\leq L\right)\right\|_{\infty} \leq M_1M_2(\min_k n_k)^{-\min_\ell p_\ell}.$$

The same upper bound holds the second term to the right-hand side of (16).

Define vector $\alpha_{s,i}=\left(Y_k^{\ell}(x_{s,i})Y_{k'}^{\ell'}(x_{s,i})-\langle Y_k^{\ell},Y_{k'}^{\ell'}\rangle;k,k'\leq K,l,l'\leq L\right)$. Invoke Lemma 2.1 in Juditsky and Nemirovski (2000), we have

$$\begin{aligned}W\left(\sum_{s=1}^{K}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)\leq & W\left(\sum_{s=1}^{K-1}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)+(n_K^{-1}\sum_{i=1}^{n_K}\alpha_{K,i})^{\intercal}\nabla W\left(\sum_{s=1}^{K-1}n_s^{-1}\sum_i\alpha_{s,i}\right)\\ &+c^*(M)\|n_K^{-1}\sum_i\alpha_{K,i}\|_{\infty}^2,\end{aligned}$$

where $M=K^2L^2$, $c^*(M)=4e\log M$, $W(z)=1/2\|z\|_q^2:\mathbb{R}^M\to\mathbb{R}$ and $q=2\log M$. It follows that

$$\mathbb{E}\left[W\left(\sum_{s=1}^{K}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)\right]\leq\mathbb{E}\left[W\left(\sum_{s=1}^{K-1}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)\right]+c^*(M)\mathbb{E}\|n_K^{-1}\sum_i\alpha_{K,i}\|_{\infty}^2, \tag{17}$$

since $\alpha_{k,i}$ and $\alpha_{k',i}$ are independent when $k\neq k'$ and $\mathbb{E}(\alpha_{k,i})=0$. The inequality in (17) implies a recursive relationship and repeatedly apply it for K times we get

$$\mathbb{E}\left[W\left(\sum_{s=1}^{K}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)\right]\leq c^*(M)\sum_{s=1}^{K}n_s^{-1}\mathbb{E}\|\sum_i\alpha_{s,i}\|_{\infty}^2.$$

By assumptions A1 and A2 again, we have $\left|Y_k^{\ell}(x_{s,i})Y_{k'}^{\ell'}(x_{s,i})-\langle Y_k^{\ell},Y_{k'}^{\ell'}\rangle\right|\leq 2M_1^2,\ a.e.$ Therefore,

$$\mathbb{E}\left[W\left(\sum_{s=1}^{K}n_s^{-1}\sum_{i=1}^{n_s}\alpha_{s,i}\right)\right]\leq c^*(M)4KM_1^4=32e\log(KL)KM_1^4.$$

Since $W(z)\geq 1/2\|z\|_{\infty}^2$, it follows

$$K^{-1}\mathbb{E}\|\sum_s n_s^{-1}\sum_i\alpha_{s,i}\|_{\infty}\leq K^{-1}\sqrt{32e\log(KL)KM_1^4}=4\sqrt{2e}M_1^2\sqrt{\log(KL)/K}.$$

The above steps also apply for the bound on $\mathbb{E}\|\tilde{\Sigma}-\Sigma\|_\infty$ by noting that

$$\frac{K\left\|\sum_{s\neq k,k'} n_s^{-1}\sum_i\left(\hat{Y}_k^\ell(x_{s,i})\hat{Y}_{k'}^{\ell'}(x_{s,i})-\langle Y_k^\ell,Y_{k'}^{\ell'}\rangle\right)\right\|_\infty}{(K-1)^2}\leq 2\frac{\left\|\sum_{s\neq k,k'} n_s^{-1}\sum_i\left(\hat{Y}_k^\ell(x_{s,i})\hat{Y}_{k'}^{\ell'}(x_{s,i})-\langle Y_k^\ell,Y_{k'}^{\ell'}\rangle\right)\right\|_\infty}{K-1-\mathbb{I}(k\neq k')}.$$

□

We then prove similar bounds for $\mathbb{E}\|b-\hat{b}\|_\infty$ and $\mathbb{E}\|b-\tilde{b}\|_\infty$.

Lemma 3. *If assumption A1 and A2 hold, we have the following bounds for* $\mathbb{E}\|b-\hat{b}\|_\infty$ *and* $\mathbb{E}\|b-\tilde{b}\|_\infty$.

$$\mathbb{E}\|b-\hat{b}\|_\infty\leq(M_1+\sigma)M_2(\min_k n_k)^{-\min_\ell p_\ell}+8\sqrt{e}(2M_1^2+M_1\sigma)\sqrt{\log(KL)/K},$$
$$\mathbb{E}\|b-\tilde{b}\|_\infty\leq(M_1+\sigma)M_2(\min_k n_k)^{-\min_\ell p_\ell}+8\sqrt{e}(2M_1^2+M_1\sigma)\sqrt{\log((K-1)L)/(K-1)}.$$

Proof. Note that

$$\begin{aligned}\|\hat{b}_{k,\ell}-b_{k,l}\|_\infty=&K^{-1}\left\|\sum_s n_s^{-1}\sum_i\hat{Y}_k^\ell(x_{s,i})y_{s,i}-Y_k^\ell(x_{s,i})y_{s,i}\right\|_\infty\\&+K^{-1}\left\|\sum_s n_s^{-1}\sum_i Y_k^\ell(x_{s,i})y_{s,i}-Y_k^\ell(x_{s,i})f_0(x_{s,i})\right\|_\infty\\&+K^{-1}\left\|\sum_s n_s^{-1}\sum_i Y_k^\ell(x_{s,i})f_0(x_{s,i})-\langle Y_k^\ell,f_0\rangle\right\|_\infty.\end{aligned}$$

Follow the same step as in the proof of Lemma 2 with assumption A1 and A2, as well as Lemma 2.1 in Juditsky and Nemirovski (2000), we have

$$\begin{aligned}\mathbb{E}\|\hat{b}-b\|_\infty\leq&(M_1+\sigma)M_2(\min_k n_k)^{-\min_\ell p_\ell}+(2M_1^2+M_1\sigma)\left(1+\sqrt{4e\log(KL)(K-1)}\right)K^{-1}\\&+4M_1^2K^{-1/2}\sqrt{e\log(KL)}\\\leq&(M_1+\sigma)M_2(\min_k n_k)^{-\min_\ell p_\ell}+8\sqrt{e}(2M_1^2+M_1\sigma)\sqrt{\log(KL)}K^{-1/2}\end{aligned}$$

The proof is completed by noting that

$$\begin{aligned}&\frac{K}{K-1}\left(K^{-1}\sum_{s\neq k}n_s^{-1}\sum_i\hat{Y}_k^\ell(x_{s,i})y_{s,i}-Y_k^\ell(x_{s,i})y_{s,i}\right)\\=&\frac{1}{K-1}\left(\sum_{s\neq k}n_s^{-1}\sum_i\hat{Y}_k^\ell(x_{s,i})y_{s,i}-Y_k^\ell(x_{s,i})y_{s,i}\right).\end{aligned}$$

□

Combining the results in Lemma 2 and Lemma 3 we get the results in Proposition 2.

C Proof of Proposition 3

Let $\lim_{n\to\infty} \hat{U}(w; K^{-1}\mathbf{1}_K) = \hat{U}_0(w)$ and $\lim_{n\to\infty} \tilde{U}(w; K^{-1}\mathbf{1}_K) = \tilde{U}(w)$. The quadratic and linear coefficients for $\hat{U}_0$ and $\tilde{U}_0$ are

$$\hat{\Sigma}_0 = \lim_{n\to\infty} \hat{\Sigma} = \left[\langle Y_k^\ell, Y_{k'}^{\ell'}\rangle; k, k' \le K, \ell, \ell' \le L\right],$$
$$\tilde{\Sigma}_0 = \lim_{n\to\infty} \tilde{\Sigma} = \frac{K}{(K-1)^2}\left[(K-1-\mathbb{I}(k\neq k'))\langle Y_k^\ell, Y_{k'}^{\ell'}\rangle; k, k' \le K, \ell, \ell' \le L\right],$$
$$\hat{b}_0 = \lim_{n\to\infty} \hat{b} = (\langle Y_k^\ell, \bar{Y}^\ell\rangle; k \le K, \ell \le L),$$
$$\hat{b}_0 = \lim_{n\to\infty} \hat{b} = (\langle Y_k^\ell, \bar{Y}_{-k}^\ell\rangle; k \le K, \ell \le L),$$

where $\bar{Y}^\ell = K^{-1}\sum_k Y_k^\ell$ and $\bar{Y}_{-k}^\ell = (K-1)^{-1}\sum_{k'\neq k} Y_{k'}^\ell$. With assumption A2 along with the proof for Lemma 2, we know that

$$\mathbb{E}\left|\hat{U}(\hat{w}; K^{-1}\mathbf{1}_K) - \hat{U}_0(\hat{w})\right| \le Cn^{-\min_\ell p_\ell}.$$

Since both $\hat{U}$ and $\tilde{U}$ are smooth with respect to w, with Taylor expansion and assumptions A1 and A2, we have

$$|\psi(\hat{w}) - \psi(\hat{w}_0)| \le C^* n^{-\min_\ell p_\ell},$$

where $\hat{w}_0 = \arg\max_{w\in W} \hat{U}_0(w)$. The same bound applies for $|\psi(\tilde{w}) - \psi(\tilde{w}_0)|$. Therefore, we can focus on study the difference $\psi(\hat{w}_0) - \psi(\tilde{w}_0)$.

Using results from Lemma 3, we can find an upper bound for $\psi(\tilde{w}_0) - \psi(w_g^0)$:

$$\mathbb{E}\left(\psi(\tilde{w}_0) - \psi(w_g^0)\right) \le 8\sqrt{e}(2M_1^2 + M_1\sigma_f)\sqrt{\log((K-1)L)}(K-1)^{-1/2},$$

here $\sigma_f^2 = \int_f \left(\int_x f(x)dF_X(x) - \int_x f_0(x)dF_X(x)\right)^2 dF(f)$.

We now find a lower bound for $\mathbb{E}(\psi(\hat{w}_0) - \psi(w_g^0))$. Invoking Theorem 3.1 in Juditsky and Nemirovski (2000) by noting that $\hat{U}_0$ is induced with a stacking problem with KL samples observed and the noise associated with the observation

is the deviation of f_k to f_0. With an appropriately chosen large constant κ, the theorem indicates that

$$\mathbb{E}(\psi(\hat{w}_0)) - \mathbb{E}(\psi(w_g^0)) \geq \kappa M_1 \sigma_f \sqrt{\log(KL)}(KL)^{-1/2}.$$

Therefore, if

$$8\sqrt{e}(2M_1^2 + M_1\sigma_f)\sqrt{\log((K-1)L)}((K-1)L)^{-1/2} \leq \kappa M_1 \sigma_f \sqrt{\log(KL)}(KL)^{-1/2},$$

we can find $\mathbb{E}\left(\psi(\hat{w}_0) - \psi(\tilde{w}_0)\right) + C^* n^{-\min_\ell p_\ell} \geq 0$.