Biostatistics (2012), **0**, 0, *pp.* 1–20
doi:10.1093/biostatistics/BoonstraBiostatisticsSubmissionV2

Incorporating Auxiliary Information for Improved Prediction in High Dimensional Datasets: An Ensemble of Shrinkage Approaches

PHILIP S. BOONSTRA*, JEREMY M.G. TAYLOR, BHRAMAR MUKHERJEE

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

philb@umich.edu

Summary

With advancement in genomic technologies, it is common that two high-dimensional datasets are available, both measuring the same underlying biological phenomenon with different techniques. We consider predicting a continuous outcome Y using $\boldsymbol{X}$, a set of p markers which is the best available measure of the underlying biological process. This same biological process may also be measured by $\boldsymbol{W}$, coming from prior technology but correlated with $\boldsymbol{X}$. On a moderately sized sample we have $(Y, \boldsymbol{X}, \boldsymbol{W})$, and on a larger sample we have $(Y, \boldsymbol{W})$. We utilize the data on $\boldsymbol{W}$ to boost prediction of Y by $\boldsymbol{X}$. When p is large and the subsample containing $\boldsymbol{X}$ is small, this is a $p > n$ situation. When p is small, this is akin to the classical measurement error problem; however, ours is not the typical goal of calibrating $\boldsymbol{W}$ for use in future studies. We propose to shrink the regression coefficients $\boldsymbol{\beta}$ of Y on $\boldsymbol{X}$ toward different targets that use information derived from $\boldsymbol{W}$ in the larger dataset. We compare these proposals with the classical ridge regression of Y on $\boldsymbol{X}$, which does

*To whom correspondence should be addressed.

not use $\boldsymbol{W}$. We also unify all of these methods as *targeted* ridge estimators. Finally, we propose a hybrid estimator which is a linear combination of multiple estimators of $\boldsymbol{\beta}$. With optimal choice of weights, the hybrid estimator balances efficiency and robustness in a data-adaptive way to theoretically yield smaller prediction error than any of its constituents. The methods are evaluated via simulation studies. We also apply them to a gene-expression dataset. mRNA expression measured via quantitative real-time polymerase chain reaction (qRT-PCR) is used to predict survival time in lung cancer patients, with auxiliary information from microarray technology available on a larger sample.

Key words: Cross-validation, Generalized Ridge, Mean Squared Prediction Error, Measurement Error

1. Introduction

As sequencing and array technologies change, multiple platforms can measure the same biological quantity of interest. Often, investigators have measurements using an older technology on a large sample and those from a newer technology on a subset of this sample. We are interested in predicting an outcome using the newer measurements, which is a statistical problem of fitting a prediction model for $Y|\boldsymbol{X}$, where Y is the outcome and $\boldsymbol{X}$ is the p-dimensional vector of biomarkers. One such model is a linear regression:

$$Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1) \tag{1.1}$$

On n_A subjects, we have Y, $\boldsymbol{X}$ and $\boldsymbol{W}$, where $\boldsymbol{W}$, also of length p, measures the same biomarkers as does $\boldsymbol{X}$ but with a prior technology. A model for $\boldsymbol{W}|\boldsymbol{X}$ consistent with this motivating context is

$$\boldsymbol{W} = \psi \mathbf{1}_p + \nu \boldsymbol{X} + \tau \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}_p(0, \boldsymbol{I}_p). \tag{1.2}$$

$\boldsymbol{I}_p$ is the identity matrix and ψ, ν and τ are scalars. For notational simplicity, we develop methods under the assumption $\beta_0 = \psi = 0$. Both quantities are estimated in our analyses.

The quantity n_{A} is of modest size, such that $p > n_{\mathrm{A}}$. Additionally, n_{B} observations of Y and $\boldsymbol{W}$ are available. Assume $p < n_{\mathrm{B}}$. Denote subsamples A and B (each assumed to be from the same population) by $(\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}})$ and $(\boldsymbol{y}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}})$, respectively. Using this notation, $\boldsymbol{x}_{\mathrm{B}}$, the set of $\boldsymbol{X}$'s from subsample B, is missing data. Figure 1 gives a schematic representation. $\boldsymbol{x}_{\mathrm{A}}$ is also standardized, ie if x_{ij} is from the ith row and jth column, $\sum_{i=1}^{n_{\mathrm{A}}} x_{ij} = 0$ and $\sum_{i=1}^{n_{\mathrm{A}}} x_{ij}^2 = n_{\mathrm{A}}$, $j = 1, \ldots, p$.

The goal is a prediction model for $Y_{\mathrm{new}}|\boldsymbol{X}_{\mathrm{new}}$ for a new subject: $\boldsymbol{X}_{\mathrm{new}}^{\top}\hat{\boldsymbol{\beta}}$. Predictive performance of $\hat{\boldsymbol{\beta}}$ is measured by mean squared prediction error (MSPE), defined as

$$\begin{aligned}\mathrm{MSPE}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(Y_{\mathrm{new}} - \boldsymbol{X}_{\mathrm{new}}^{\top}\hat{\boldsymbol{\beta}})^2] &= \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\top}\boldsymbol{X}_{\mathrm{new}}\boldsymbol{X}_{\mathrm{new}}^{\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})] \\ &= \sigma^2 + \mathrm{Tr}\big[(\mathrm{Bias}\,\hat{\boldsymbol{\beta}}\,\mathrm{Bias}\,\hat{\boldsymbol{\beta}}^{\top} + \mathrm{Var}\,\hat{\boldsymbol{\beta}})\mathrm{E}[\boldsymbol{X}_{\mathrm{new}}\boldsymbol{X}_{\mathrm{new}}^{\top}]\big], \qquad (1.3)\end{aligned}$$

where Tr indicates the trace operator, and the expectation is over $Y_{\mathrm{new}}, \boldsymbol{X}_{\mathrm{new}}, \boldsymbol{y}_{\mathrm{A}}, \boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{B}}$. We consider two questions: (i) How can the auxiliary information in subsample B be used in the prediction of $Y|\boldsymbol{X}$? (ii) When does using such information lead to improved MSPE?

A simple approach, which ignores subsample B, is ordinary least squares of $\boldsymbol{y}_{\mathrm{A}}$ on $\boldsymbol{x}_{\mathrm{A}}$, i.e. $\hat{\boldsymbol{\beta}}_{\textsc{ols}} = \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) = (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$. However, $(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}})^{-1}$ does not exist for $p > n_{\mathrm{A}}$. Even for $p \leqslant n_{\mathrm{A}}$, multicollinearity of the covariates may lead to variance inflation and numerical instability. Ridge regression (RIDG) (Hoerl and Kennard, 1970) can ameliorate these issues by shrinking coefficients toward zero, i.e. $\hat{\boldsymbol{\beta}}_{\textsc{ridg}} = \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta} = (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$. This can be viewed from a Bayesian perspective: given a normal prior on $\boldsymbol{\beta}$ with mean $\boldsymbol{0}_p$ and precision $\sigma^{-2}\lambda\boldsymbol{I}_p$, the RIDG coefficients are the posterior mode for a given λ. Hoerl and Kennard showed that there exists $\lambda > 0$ which decreases mean squared error, $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$, compared to $\lambda = 0$. RIDG penalizes the L_2 norm; other methods exist which constrain the L_d norm for some d (eg Frank and Friedman, 1993). In contrast to variable selection procedures, which might use an L_1 penalty, our goal is using auxiliary information to boost prediction, and so we restrict attention to ridge-type estimators.

Dempster *and others* (1977) evaluate 57 variants of shrinkage estimators and argue for RIDG. Draper and van Nostrand (1979) are critical of RIDG because of difficulties in choosing the parameter λ. However, Craven and Wahba (1979) and Li (1986) demonstrate the asymptotic optimality of the generalized cross-validation (GCV) function in selecting λ. Simulation studies (Gelfand, 1986; Frank and Friedman, 1993) demonstrate good prediction properties of RIDG for many choices of $\boldsymbol{\beta}$. Rao (1975) generalizes RIDG to allow for different levels of shrinkage between each coefficient. Swindel (1976) proposes ridge estimators which take into account prior information, changing the direction of shrinkage. Casella (1980) and Maruyama and Strawderman (2005) propose variants of ridge estimators with minimax properties. Sclove (1968) adapts the shrinkage estimator of James and Stein (1961) (JS) which, for $p > 3$, uniformly beats the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ in terms of MSE. Gruber (1998) offers a unified treatment of different kinds of JS and ridge estimators from frequentist and Bayesian points of view.

By incorporating subsample B, this may be viewed as a problem of combining multiple estimators. George (1986) proposes JS estimators which shrink toward multiple targets. Green and Strawderman (1991) consider a *targeted* JS estimator: an unbiased estimator is shrunk toward a biased but more efficient estimator so as to minimize MSE under certain assumptions. LeBlanc and Tibshirani (1996) propose linear combinations of regression coefficients to improve prediction error. This bias and variance trade-off in combining estimators has been used in recent genetic studies (Chen *and others*, 2009).

For $p < n_{\mathrm{A}}$, the problem closely resembles that of measurement error (ME) in the covariates, $\boldsymbol{W}$ being an error-prone version of $\boldsymbol{X}$. Fuller (2006) and Carroll *and others* (2006) review ME methods for unbiased and efficient inference on $\boldsymbol{\beta}$. In linear regression, using $\boldsymbol{W}$ instead of $\boldsymbol{X}$ gives biased estimates of $\boldsymbol{\beta}$. However, this substitution is typically not problematic for predicting Y_{new} with $\boldsymbol{W}_{\mathrm{new}}^{\top}\hat{\boldsymbol{\beta}}$. Our prediction model of interest being Y given $\boldsymbol{X}$, this bias in $\hat{\boldsymbol{\beta}}$ from using $\boldsymbol{W}$ instead of $\boldsymbol{X}$ *does* bias $\boldsymbol{X}_{\mathrm{new}}^{\top}\hat{\boldsymbol{\beta}}$ away from Y_{new}. Regression calibration, which fills in each missing $\boldsymbol{X}$ with its conditional expectation given $\boldsymbol{W}$, may provide unbiased estimates of $\boldsymbol{\beta}$ and therefore Y_{new}. In contrast, although the substitution of $\boldsymbol{X}$ by $\boldsymbol{W}$ gives biased estimates of $\boldsymbol{\beta}$, it may reduce the *variance* of estimates of $\boldsymbol{\beta}$ relative to regression calibration (Buzas *and others*, 2005) and consequently reduce MSPE. Even for $p < n_{\text{A}}$, then, it is not evident that the regression calibration algorithm is best for making predictions with $\boldsymbol{X}_{\text{new}}^{\top}\hat{\boldsymbol{\beta}}$.

This paper makes several new contributions. We consider an important but non-standard prediction problem which has not yet received a rigorous mathematical treatment. We introduce a class of targeted ridge estimators, borrowing ideas from the shrinkage and regression calibration literature. We also consider combining an ensemble of targeted ridge estimators, as in Green and Strawderman (1991). In contrast to minimizing MSE, we determine the shrinkage weights adaptively so as to minimize MSPE. Interestingly, one is able to combine two or more *biased* estimators of $\boldsymbol{\beta}$ for better prediction than any individual estimator. This result applies to a linear combination of *any* set of estimates of $\boldsymbol{\beta}$. We evaluate all of these estimators via simulation studies and and a data analysis.

The rest of the paper is organized as follows. In Section 2, we unify RIDG and regression calibration methods under a class of targeted ridge estimators. In Section 3, we propose hybrid estimators which combine multiple estimators with data-adaptive weights to achieve superior prediction. Section 4 presents a simulation study. Section 5 applies the methods, in which survival time (Y) in lung cancer patients is predicted with qRT-PCR data ($\boldsymbol{X}$), with microarray data ($\boldsymbol{W}$) from a larger sample aiding in predictions. Section 6 concludes with a discussion. Most analytical details are in the Supplementary Materials.

2. Targeted Shrinkage

For $p > n_{\text{A}}$, ordinary least squares using subsample A is not applicable. In fact, when $\boldsymbol{X}_{\text{new}}$ is not in the column space of $\boldsymbol{x}_{\text{A}}$, *no* unbiased estimate of $\boldsymbol{X}_{\text{new}}^{\top}\boldsymbol{\beta}$ (using only subsample A) exists (Rao, 1945). A biased alternative is ridge regression (Hoerl and Kennard, 1970),

$$\hat{\boldsymbol{\beta}}_{\text{RIDG}} = (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}}. \tag{2.4}$$

RIDG is equivalent to adding λ to each eigenvalue of $\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}}$, thus allowing the matrix inversion. The coefficient estimates are shrunk to zero, more so for larger values of λ. That the ridge estimator is applicable for $p > n_{\text{A}}$ is crucial in our setting. Shrinkage estimators from Sclove (1968) and Casella (1980) make use of *unbiased* estimators of $\boldsymbol{\beta}$ and hence are not directly applicable for $p > n_{\text{A}}$ situations.

For ridge regression, Craven and Wahba (1979) proposed to select λ using the GCV function, choosing the λ which minimizes

$$\frac{\frac{1}{n_{\text{A}}}(\boldsymbol{y}_{\text{A}} - \boldsymbol{H}(\lambda\boldsymbol{I}_p)\boldsymbol{y}_{\text{A}})^{\top}(\boldsymbol{y}_{\text{A}} - \boldsymbol{H}(\lambda\boldsymbol{I}_p)\boldsymbol{y}_{\text{A}})}{(1 - \text{Tr } \boldsymbol{H}(\lambda\boldsymbol{I}_p)/n_{\text{A}})^2}, \qquad \boldsymbol{H}(\boldsymbol{\Theta}) = \boldsymbol{x}_{\text{A}}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{\Theta})^{-1}\boldsymbol{x}_{\text{A}}^{\top}, \tag{2.5}$$

where $\boldsymbol{\Theta}$ is an arbitrary $p \times p$ positive semi-definite (psd) matrix. Rao (1975) suggested that any psd matrix $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ can replace $\boldsymbol{I}_p$ in (2.4). Swindel (1976) proposed to shrink toward a non-null vector $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. From the Bayesian perspective, these replace the prior precision $\sigma^{-2}\lambda\boldsymbol{I}_p$ in RIDG with $\sigma^{-2}\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ and the prior mean $\mathbf{0}_p$ with $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. The posterior mode is

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}) &= \text{argmin}_{\boldsymbol{\beta}} \tfrac{1}{\sigma^2}(\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\boldsymbol{\beta}) + \tfrac{1}{\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\gamma}_{\boldsymbol{\beta}})^{\top}\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\gamma}_{\boldsymbol{\beta}}) && (2.6)\\ &= (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}}). && (2.7)\end{aligned}$$

Gruber (1998, p.241) calls this a generalized ridge estimator. Because "generalized ridge" has been used for several distinct methods in the shrinkage literature, we instead call this a targeted ridge (TR) estimator, referring to shrinkage toward a target $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. The estimator given in (2.7) implies that there are three terms $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ that determine the general class of TR estimators. As we shall see, different estimators we propose either implicitly or explicitly specify the values for $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$. In particular, RIDG is a TR estimator: $\hat{\boldsymbol{\beta}}_{\text{RIDG}} = \hat{\boldsymbol{\beta}}(\mathbf{0}_p, \lambda, \boldsymbol{I}_p)$.

As stated in (1.3), the MSPE of a TR estimator $\hat{\boldsymbol{\beta}}$ is $\sigma^2+\mathrm{Tr}\big[(\mathrm{Bias}\,\hat{\boldsymbol{\beta}}\,\mathrm{Bias}\,\hat{\boldsymbol{\beta}}^\top+\mathrm{Var}\,\hat{\boldsymbol{\beta}})\mathrm{E}[\boldsymbol{X}_{\mathrm{new}}\boldsymbol{X}_{\mathrm{new}}^\top\big]$. Thus we calculate the MSPE of $\hat{\boldsymbol{\beta}}$ from its bias and variance, taking expectations over the response distribution $\boldsymbol{y}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{B}}$:

$$\begin{aligned}\mathrm{Bias}\,\hat{\boldsymbol{\beta}}&=\mathrm{E}\,\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}+\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}\mathrm{E}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}-\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}-\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta})\\&=\lambda(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}(\mathrm{E}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}-\boldsymbol{\beta})\qquad(2.8)\\\mathrm{Var}\,\hat{\boldsymbol{\beta}}&=(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}(\sigma^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\lambda^2\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}\mathrm{Var}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}\,\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}.\qquad(2.9)\end{aligned}$$

These expressions assume that λ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ are fixed with respect to $\boldsymbol{y}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{B}}$ but allow $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ to be data-dependent. A TR estimator may use a true prior, as in RIDG, in which case $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ is fixed.

We now propose several other TR estimators. If $\boldsymbol{x}_{\mathrm{B}}$ were observed, logical selections of $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ would be $(\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}})^{-1}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{y}_{\mathrm{B}}$ and $\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}$, respectively, with $\lambda=1$, giving the estimator $(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}})^{-1}(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{y}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{y}_{\mathrm{B}})$. In the absence of $\boldsymbol{x}_{\mathrm{B}}$, the naïve inclination is to regress $\boldsymbol{y}_{\mathrm{B}}$ on $\boldsymbol{w}_{\mathrm{B}}$ and use $(\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}})^{-1}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{y}_{\mathrm{B}}$ and $\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}}$ as $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$, that is, use $\boldsymbol{w}_{\mathrm{B}}$ itself as an imputation for $\boldsymbol{x}_{\mathrm{B}}$. We first consider approaches which derive a replacement for the missing $\boldsymbol{x}_{\mathrm{B}}$ which may be better than $\boldsymbol{w}_{\mathrm{B}}$. This is obtained by modeling $\boldsymbol{W}|\boldsymbol{X}$ based on the relationship observed in subsample A and thereby inducing data-driven values of $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$. From the ME perspective, this is regression calibration. These TR estimators fix $\lambda=1$ (data-adaptive estimation of λ may be done using, for example, a GCV criterion).

Structural Regression Calibration (SRC): A distribution on $\boldsymbol{X}$ and the ME model for $\boldsymbol{W}|\boldsymbol{X}$ imply a value of $\mathrm{E}[\boldsymbol{X}|\boldsymbol{W}]$. SRC fills in the missing $\boldsymbol{x}_{\mathrm{B}}$ with its conditional expectation given $\boldsymbol{w}_{\mathrm{B}}$. Assuming $\boldsymbol{X}$ is normal, say $\mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}})$, implies $\boldsymbol{X}|\boldsymbol{W}$ is also normal. Let $\boldsymbol{\theta}=\{\nu,\tau,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\}$. From properties of the conditional distribution of $\boldsymbol{X}|\boldsymbol{W}$,

$$\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}(\boldsymbol{\theta})=\mathrm{E}[\boldsymbol{x}_{\mathrm{B}}|\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\theta}]=\boldsymbol{1}_{n_{\mathrm{B}}}\boldsymbol{\mu}_{\boldsymbol{X}}{}^\top(\boldsymbol{I}_p-\boldsymbol{V}(\boldsymbol{\theta}))+\tfrac{1}{\nu}\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V}(\boldsymbol{\theta})=[\boldsymbol{1}_{n_{\mathrm{B}}},\boldsymbol{w}_{\mathrm{B}}]\boldsymbol{M}(\boldsymbol{\theta}),\qquad(2.10)$$

$$\boldsymbol{M}(\boldsymbol{\theta})=\begin{pmatrix}\boldsymbol{\mu}_{\boldsymbol{X}}{}^\top(\boldsymbol{I}_p-\boldsymbol{V}(\boldsymbol{\theta}))\\\frac{1}{\nu}\boldsymbol{V}(\boldsymbol{\theta})\end{pmatrix},\qquad\boldsymbol{V}(\boldsymbol{\theta})=(\boldsymbol{I}_p+\tfrac{\tau^2}{\nu^2}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})^{-1}\qquad(2.11)$$

(we suppress dependence on $\boldsymbol{\theta}$ of $\boldsymbol{x}_{\mathrm{B}}^{\textsc{src}}(\boldsymbol{\theta})$, $\boldsymbol{M}(\boldsymbol{\theta})$, and $\boldsymbol{V}(\boldsymbol{\theta})$ hereafter). This is a precision-weighted average of $\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\mu}_{\boldsymbol{X}}{}^{\top}$ and $(1/\nu)\boldsymbol{w}_{\mathrm{B}}$. Using (2.7), define $\hat{\boldsymbol{\beta}}_{\textsc{src}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\textsc{src}}}, 1, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\textsc{src}}})$, with $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\textsc{src}}} = (\boldsymbol{x}_{\mathrm{B}}^{\textsc{src}\top}\boldsymbol{x}_{\mathrm{B}}^{\textsc{src}})^{-1}(\boldsymbol{x}_{\mathrm{B}}^{\textsc{src}\top}\boldsymbol{y}_{\mathrm{B}})$ and $\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\textsc{src}}} = \boldsymbol{x}_{\mathrm{B}}^{\textsc{src}\top}\boldsymbol{x}_{\mathrm{B}}^{\textsc{src}}$. In the ME literature, SRC is the standard "Regression Calibration" approach. We append "Structural" (Carroll *and others*, 2006, p.25), referring to a distributional assumption about $\boldsymbol{X}$, to distinguish from its "Functional" alternative, which does not assume this, proposed as follows.

Functional Regression Calibration (FRC): Solving (1.2), $\boldsymbol{W} = \nu\boldsymbol{X} + \tau\boldsymbol{\xi}$, for $\boldsymbol{X}$ gives $\boldsymbol{X} = (1/\nu)\boldsymbol{W} - (\tau/\nu)\boldsymbol{\xi}$. Another natural estimate of $\boldsymbol{x}_{\mathrm{B}}$, and consequently a corresponding $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$, is therefore

$$\boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}}(\boldsymbol{\theta}) = (1/\nu)\boldsymbol{w}_{\mathrm{B}}, \qquad \boldsymbol{\gamma}_{\boldsymbol{\beta}_{\textsc{frc}}} = (\boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}\top}\boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}})^{-1}\boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}\top}\boldsymbol{y}_{\mathrm{B}}, \qquad \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\textsc{frc}}} = \boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}\top}\boldsymbol{x}_{\mathrm{B}}^{\textsc{frc}}. \tag{2.12}$$

This gives a TR estimate defined as $\hat{\boldsymbol{\beta}}_{\textsc{frc}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\textsc{frc}}}, 1, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\textsc{frc}}})$. This imputation for $\boldsymbol{x}_{\mathrm{B}}$ is a scaled version of a substitution of $\boldsymbol{w}_{\mathrm{B}}$ for $\boldsymbol{x}_{\mathrm{B}}$, to which FRC is equivalent when $\nu = 1$, ie under the classical ME model.

In the Supplementary Materials (Appendix A), we conduct extensive analyses which suggest that FRC is preferred over SRC (in terms of MSPE) as any of $\boldsymbol{\beta}^{\top}\boldsymbol{\beta}$, σ^2, or τ/ν increase.

The first rows of Table 1 summarize choices of $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ for RIDG, FRC, and SRC. Assuming non-differential measurement error [NDME] , ie $[Y|\boldsymbol{X}, \boldsymbol{W}] = [Y|\boldsymbol{X}]$, and $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbf{0}_p$, Table 1 also gives $\mathrm{E}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\mathrm{Var}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ for FRC and SRC. Because $\mathrm{E}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\textsc{src}}} = \boldsymbol{\beta}$, SRC provides unbiased estimates of $\boldsymbol{\beta}$ (Equation 2.8).

Remark 2.1 One of the reviewers observed that, when $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ are based on historical data, the prior in the second expression of (2.6) is a power prior (Chen and Ibrahim, 2000), with λ controlling the contribution of the historical data to the posterior.

Remark 2.2 These approaches require estimating $\boldsymbol{\theta} = \{\nu, \tau, \boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\}$. One can regress $\{w_{ij}\}$ on $\{x_{ij}\}$ for $i = 1, \ldots, n_{\mathrm{A}}$ and $j = 1, \ldots, p$ to compute MLEs for ν and τ. If it is required that ν and τ be of a more general form than scalar-valued, the estimation procedure can be modified accordingly. The MLE for $\boldsymbol{\mu}_{\boldsymbol{X}}$ is $\hat{\boldsymbol{\mu}}_{\boldsymbol{X}} = n_{\mathrm{A}}^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\mathbf{1}_{n_{\mathrm{A}}}$, which will be $\mathbf{0}_p$ if $\boldsymbol{x}_{\mathrm{A}}$ is standardized. For $p > n_{\mathrm{A}}$, the required inversion of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = n_{\mathrm{A}}^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}}$ is not possible. An alternative is the shrinkage estimator from Schäfer and Strimmer (2005): since $\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}}$ is standardized, it is simply $\hat{\boldsymbol{\Sigma}}^{*}_{\boldsymbol{X}} = (1-\eta)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} + \eta \boldsymbol{I}_p$, for $\eta \in [0,1]$ chosen data-adaptively. We used the `R` package `corpcor` to choose η targeting a minimum MSE for $\hat{\boldsymbol{\Sigma}}^{*}_{\boldsymbol{X}}$.

Remark 2.3 The bias and variance outlined in Table 1 condition on the true value of $\boldsymbol{\theta}$ and are over and above any bias and variance coming from its estimation. In particular, estimating $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ may pose a challenge to SRC in the high-dimensional setting.

Remark 2.4 One other approach which we do not further explore is modifying FRC or SRC to do adaptive, component-wise shrinkage on $\boldsymbol{\beta}$: a TR estimator where $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ is diagonal and λ is estimated. When λ is not fixed, the GCV approach may be used to choose an appropriate value of λ. The form of this modified GCV criterion is given later on in (3.14), in connection with the hybrid estimator.

3. Hybrid Estimators

While a particular TR estimator may do well for a given set of factors (eg p, n_{B}, $\boldsymbol{\beta}$, τ, etc), none is likely to give small prediction error under all settings. However, a hybrid estimator, that is, an adaptively combined set of *multiple* TR estimators, may yield this flexibility. Given m estimators, $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_m$, and a vector $\boldsymbol{\omega} = \{\omega_1, \omega_2, \ldots, \omega_m\}$ such that $\mathbf{1}_m^{\top}\boldsymbol{\omega} = 1$, let $\boldsymbol{b}(\boldsymbol{\omega}) = \sum_{i=1}^{m} \omega_i \hat{\boldsymbol{\beta}}_i = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \cdots, \hat{\boldsymbol{\beta}}_m]\boldsymbol{\omega}$. The vector $\boldsymbol{\omega}$ determines the contribution from each $\hat{\boldsymbol{\beta}}_i$; a sensible choice for $\boldsymbol{\omega}$ in our situation would be the one which minimizes MSPE($\boldsymbol{b}(\boldsymbol{\omega})$). The following Theorem compares the prediction error of the resulting optimal hybrid estimator, $\boldsymbol{b}(\boldsymbol{\omega}^{\mathrm{opt}})$, to that of its constituents; the result uses the following definition of the "mean cross-product prediction error" between $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{\beta}}_j$:

$$\mathrm{MCPE}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_j) = \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_i)^{\top}\boldsymbol{X}_{\mathrm{new}}\boldsymbol{X}_{\mathrm{new}}^{\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_j)]. \tag{3.13}$$

Theorem 3.1 Let $\boldsymbol{b}(\boldsymbol{\omega}) = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \cdots, \hat{\boldsymbol{\beta}}_m]\boldsymbol{\omega}$ be a hybrid estimator. (i) If Var $\left[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_m)\boldsymbol{v}\right]$ has at least one positive eigenvalue for every $\boldsymbol{v} \in \mathbb{R}^m \backslash \mathbf{0}_m$, then there exists a unique vector $\boldsymbol{\omega}^{\text{opt}}$ which minimizes MSPE$(\boldsymbol{b}(\boldsymbol{\omega}))$ subject to $\mathbf{1}_m^\top \boldsymbol{\omega} = 1$. (ii) Further, let MSPE$(\hat{\boldsymbol{\beta}}_j) = \min_\ell$ MSPE$(\hat{\boldsymbol{\beta}}_\ell)$. If MCPE$(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_i) \neq$ MSPE$(\hat{\boldsymbol{\beta}}_j)$ for some $i \neq j$, then MSPE$(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) <$ MSPE$(\hat{\boldsymbol{\beta}}_j)$.

The proof is in the Supplementary Materials (Appendix B). If the assumptions are satisfied, then, using prediction error as the criterion, $\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})$ will perform better than the best of its constituents. This phenomenon has been observed empirically by Breiman (1996) and LeBlanc and Tibshirani (1996). Fumera and Roli (2005) prove a slightly weaker result for ensembles of classifiers.

Now, MSPE$(\boldsymbol{b}(\boldsymbol{\omega})) = \boldsymbol{\omega}^\top \boldsymbol{P} \boldsymbol{\omega}$, where $\boldsymbol{P}$ is the $m \times m$ matrix with the (ij)th element given by $P_{ij} =$ MCPE$(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$, which is just MSPE$(\hat{\boldsymbol{\beta}}_i)$ when $i = j$. The results from Theorem 3.1 apply when $\boldsymbol{P}$ is known. In practice, however, $\boldsymbol{P}$ and therefore $\boldsymbol{\omega}^{\text{opt}}$ must be estimated. Since P_{ij} is equivalently expressed as $\mathrm{E}[(Y_{\text{new}} - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}_i)(Y_{\text{new}} - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}_j)]$, one might use $(1/n_{\text{A}})(\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\hat{\boldsymbol{\beta}}_i)^\top (\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\hat{\boldsymbol{\beta}}_j)$ as an estimate, but this will be biased. A generalization of a result from Mallows (1973) in the Supplementary Materials (Lemma B.3) gives that, on average, this underestimates P_{ij} by the amount $\sigma^2(\psi_i + \psi_j)$, where $\psi_\ell = \text{Tr}\, \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}_{\boldsymbol{\beta},\ell}^{-1})/n_{\text{A}}$. Borrowing Mallows' idea of adjusting by $\hat{\sigma}^2(\psi_i + \psi_j)$ does not work when there is no good choice of $\hat{\sigma}^2$. We propose as an alternative to adapt the GCV approach:

$$\hat{P}_{ij} = \frac{\frac{1}{n_{\text{A}}}(\boldsymbol{y}^*_{\text{A},i} - \boldsymbol{H}(\lambda_i \boldsymbol{\Omega}_{\boldsymbol{\beta},i}^{-1})\boldsymbol{y}^*_{\text{A},i})^\top (\boldsymbol{y}^*_{\text{A},j} - \boldsymbol{H}(\lambda_j \boldsymbol{\Omega}_{\boldsymbol{\beta},j}^{-1})\boldsymbol{y}^*_{\text{A},j})}{(1-\psi_i)(1-\psi_j)}, \tag{3.14}$$

where $\boldsymbol{y}^*_{\text{A},\ell} = \boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\boldsymbol{\gamma}_{\boldsymbol{\beta},\ell}$. Because $\boldsymbol{y}^*_{\text{A},\ell} - \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}_{\boldsymbol{\beta},\ell}^{-1})\boldsymbol{y}^*_{\text{A},\ell} = \boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\hat{\boldsymbol{\beta}}_\ell$, this is a penalized version of its naïve counterpart. Lemma B.4 provides further justification for this approach.

Note the dual use of the GCV function to calculate $\boldsymbol{b}(\boldsymbol{\omega})$. First, for each ℓ, λ_ℓ is chosen (when required) to minimize $\hat{P}_{\ell\ell}$. Then, fixing these choices of λ_ℓ, (3.14) is employed on the $m(m+1)/2$ pairwise combinations of components in $\boldsymbol{b}(\boldsymbol{\omega})$ to estimate $\boldsymbol{P}$. The particular hybrid estimator we evaluate has three components:

$\hat{\boldsymbol{\beta}}_{\text{HYB}} = [\hat{\boldsymbol{\beta}}_{\text{RIDG}}\, \hat{\boldsymbol{\beta}}_{\text{SRC}}\, \hat{\boldsymbol{\beta}}_{\text{FRC}}]\hat{\boldsymbol{\omega}}^{\text{opt}}$. Following LeBlanc and Tibshirani (1996), in addition to the constraint $\mathbf{1}_m^\top \boldsymbol{\omega} = 1$, we enforce a non-negativity constraint on $\boldsymbol{\omega}$, which improves numerical results.

Remark 3.2 The key aspect that makes $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ practical is that the sum $\sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{HYB}})^\top \boldsymbol{X}_{\text{new}} \boldsymbol{X}_{\text{new}}^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{HYB}})]$ is the quantity to minimize. Estimating either of the terms alone is difficult. Green and Strawderman (1991) propose a similar combination of two estimators which minimizes the MSE of $\boldsymbol{b}(\boldsymbol{\omega})$. For their method, estimation of $\boldsymbol{\omega}^{\text{opt}}$ requires an unbiased $\hat{\boldsymbol{\beta}}_1$ and independent estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. In our case, because MSPE (and not MSE) is of interest, we require neither unbiasedness nor independent estimators.

Remark 3.3 Although Theorem 3.1 proves hyb has a smaller MSPE than any of its constituents when using the true optimal weights $\boldsymbol{\omega}^{\text{opt}}$, for a given dataset with estimated optimal weights $\hat{\boldsymbol{\omega}}^{\text{opt}}$, this uniform dominance may not hold. Numerical performance depends on how accurately (3.14) estimates $\boldsymbol{P}$. As will be seen, $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ (with estimated weights) still performs well across a spectrum of scenarios and closely adapts to the best of its constituents.

4. Simulation Study

We next describe a small simulation study. We fixed $n_{\text{A}} = 50$ and used $n_{\text{B}} \in \{400, 150\}$. The diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ were set to unity, and the off-diagonals were $\rho^{|j_1 - j_2|}$, $\rho \in \{0, 0.75\}$. Using these parameters, $\boldsymbol{x}_{\text{A}}$ and $\boldsymbol{x}_{\text{B}}$ were drawn from $\mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_{\boldsymbol{X}})$. We considered both high ($p = 99$) and low ($p = 5$) dimensional models: $\boldsymbol{\beta} \in \{\{\frac{j}{100}\}_{j=-49}^{j=49}, \{\frac{j}{4}\}_{j=-2}^{j=2}\}$. R^2 values were either 0.1 or 0.4. Thus given $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ and R^2, σ was determined by solving $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2) = R^2$. β_0 was set to zero. $\boldsymbol{y}_{\text{A}}|\boldsymbol{x}_{\text{A}}$ and $\boldsymbol{y}_{\text{B}}|\boldsymbol{x}_{\text{B}}$ were drawn for each combination of $\boldsymbol{\beta}$ and σ from (1.1). This yielded 16 unique simulation settings: two choices each for p, n_{B}, ρ, and R^2. To draw the auxiliary data, we set $\psi = 0$ and $\nu = 1$ and repeated each of the 16 settings for $\tau \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 2\}$, drawing $\boldsymbol{w}_{\text{A}}|\boldsymbol{x}_{\text{A}}$ and $\boldsymbol{w}_{\text{B}}|\boldsymbol{x}_{\text{B}}$ from (1.2).

For four methods (RIDG, SRC, FRC, and HYB), we estimated MSPE by averaging the squared prediction error over 1000 new individuals. Figure 2 plots this empirical MSPE (averaged over 1000 replicates) over τ. For reference, σ^2, the smallest achievable MSPE, is also given. Tables S2 and S3 in the Supplementary Materials provide numeric values of MSPE over all settings.

Remark 4.1 In practice, the analyst estimates β_0 in addition to $\boldsymbol{\beta}$. Following the common prescription for ridge regression, we did not shrink β_0 but instead used a flat "prior" in each of the TR methods.

Effect of τ: RIDG is not affected by τ, as it does not use $\boldsymbol{w}_\mathrm{A}$ or $\boldsymbol{w}_\mathrm{B}$. FRC and SRC are equivalent when τ is very small, close to the complete data case. The MSPE of SRC always rises with τ (this increase is sharp when $p = 99$). However, larger values of τ give favorable shrinkage in FRC. When $p = 99$, the τ for which FRC is best is larger than zero; for $p = 5$, the "optimal" τ is quite small, and the MSPE rises sharply with τ. For $p = 99$, HYB usually predicts very well regardless of τ; when $p = 5$, HYB does a better job of improving upon its constituents when τ is large.

Effect of n_B, p, ρ, R^2 As might be expected, larger values of n_B considerably decrease MSPE for SRC, FRC and, consequently, HYB. Notably, HYB sometimes fares poorly compared to FRC (see Remark 3.3) when $p = 99$, $n_\mathrm{B} = 400$, and $\rho = 0.75$. In the other $p = 99$ scenarios, HYB matches or outperforms every other method. SRC fares poorly when $p = 99$. On the other hand, when $p = 5$, HYB is typically not the best method. Here, all the methods are similarly ranked regardless of other parameter settings, with SRC usually having the smallest MSPE, the exception being the case of $\rho = 0.75$, $R^2 = 0.1$ and $n_\mathrm{B} = 150$ case.

Evaluating MSE: We also evaluated each simulation in terms of MSE of $\hat{\boldsymbol{\beta}}$ (Figure S1 in Supplementary Materials). When $\rho = 0$, the results are virtually the same, up to additive constant. This is to be expected: when $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbf{0}_p$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{I}_p$, $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}) = \sigma^2 + \mathrm{MSE}(\hat{\boldsymbol{\beta}})$. When $\rho = 0.75$, this relationship does not hold, and some rankings of the methods change. However, even though it minimizes prediction error, HYB is the best method overall in terms of MSE, particularly for the $p = 99$ cases.

Appendix C in the Supplementary Materials investigates several violations to the modeling assumptions in this study. The most important result of these studies is that HYB is a flexible method. Under a variety of model settings and violations, HYB is able to efficiently adhere to the best-performing of its constituents.

5. Predicting Survival Time from Gene Expression Measurements

We consider whether gene expression measurements offer information for predicting survival time in patients with lung cancer. Expression data may be collected using microarray technology, which assays the mRNA transcripts of thousands of genes. Alternatively, quantitative real-time polymerase chain reaction (qRT-PCR) amplifies gene expression in a targeted region of DNA so as to precisely measure it. Expression is measured as the number of doublings until a threshold is reached. It is both clinically practical to measure on a new tissue specimen (not requiring the specialized laboratory facilities of microarrays) and typically considered a more precise measurement of gene expression than microarrays.

Our dataset comes from Chen *and others* (2011), who selected $p = 91$ high-correlating genes representing a broad spectrum of biological functions upon which to build a predictive model. Expression on the log-scale using Affymetrix (a microarray technology, $\boldsymbol{W}$) was measured on 439 tumor samples, and qRT-PCR measurements ($\boldsymbol{X}$) were collected on 47 of these tumors. The individual correlations between the qRT-PCR and Affymetrix measurements from the 47 tumors are greater than 0.5 across the 91 genes. Clinical covariates (age, gender and stage of cancer [I-III]) are also available. Because qRT-PCR is the clinically applicable measurement for future observations, the goal is a qRT-PCR + clinical covariate model for predicting log-survival time after surgery (Y). An independent cohort of 101 tumors with qRT-PCR measurements and clinical covariates is available for validation.

11 measurements in the qRT-PCR-only data (out of $47 \times 91 = 4277$ total, or 0.26 percent) were missing;

in order to use all observations, these values were imputed using chained equations and thereafter assumed known. Additionally, four tumors (three in the Affymetrix-only sample and one in the validation sample) had event times less than one month after surgery, and these were removed before analysis. Thus $n_{\mathrm{A}} = 47$, $n_{\mathrm{B}} = 389$, and the validation data contains 100 observations.

Because our methodology was developed for continuous outcomes, censoring necessitated some preprocessing of the data. We first imputed each censored log-survival time from a linear model of the clinical covariates, conditional upon the censoring time. This model was fit to the training data but was applied to censored survival times in both the training and validation data. Given completed log-survival times, we re-fit this same model and calculated residuals from both the training and validation data. These residuals were considered as outcomes, and the question is whether any additional variation in the residuals is explained by gene expression.

Figure S4 (Supplementary Materials) presents the 91 LOESS curves comparing measurements from the 47 tumors using Affymetrix ($\boldsymbol{w}_{\mathrm{A}}$) to qRT-PCR ($\boldsymbol{x}_{\mathrm{A}}$) after standardization. Based on this, we used a gene-specific ME model: $w_{ij} = \psi_j + \nu_j x_{ij} + \tau\xi_{ij}$. We modeled ψ_j and ν_j as random effects, distributed as $\mathcal{N}(\mu_\psi, \sigma^2_\psi)$ and $\mathcal{N}(\mu_\nu, \sigma^2_\nu)$, and used predictions $\{\hat{\psi}_j\}$ and $\{\hat{\nu}_j\}$ to calculate $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$. Violation of the constant τ assumption was also present: gene-specific estimates were in the interval $(0.209, 1.146)$ with the middle 45 in $(0.368, 0.689)$. Considering all genes simultaneously, $\hat{\tau} = .628$. Because our simulations indicate robustness to this assumption, this violation was ignored.

We present results for predicting survival time in the validation data using RIDG, SRC, FRC, and HYB. Table 2 presents numerical results for each of the methods, and Figure 3 plots each estimate of $\boldsymbol{\beta}$ as a kernel density. In terms of MSPE, the best method was RIDG, with an MSPE of 0.620, compared to 8.745 for SRC and 0.781 for FRC. For HYB, $\hat{\boldsymbol{\omega}}^{\mathrm{opt}} = \{1, 0, 0\}$, corresponding to RIDG, SRC, and FRC; so $\hat{\boldsymbol{\beta}}_{\mathrm{HYB}} \equiv \hat{\boldsymbol{\beta}}_{\mathrm{RIDG}}$ and HYB matches the best of its constituents. Plugging in $\hat{\boldsymbol{\beta}} = \mathbf{0}_p$ yields an MSPE of 0.590, which none of the methods can improve upon, suggesting a very weak signal in the set of expression measures for predicting survival. The range of $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$ (and $\hat{\boldsymbol{\beta}}_{\text{HYB}}$), excluding the intercept, is $(-0.019, 0.014)$. For $\hat{\boldsymbol{\beta}}_{\text{SRC}}$, it is $(-0.600, 0.516)$ and for $\hat{\boldsymbol{\beta}}_{\text{FRC}}$, it is $(-0.075, 0.062)$.

Finally, we generated 95% prediction intervals for each observation in the validation sample, using a bootstrap algorithm described in Appendix D of the Supplementary Materials. Table 2 gives the proportion of intervals which included the outcome and the average interval ranges. RIDG/HYB have slight under-coverage (0.91), and SRC and FRC have over-coverage (respectively 1.00 and 0.98).

REMARK 5.1 As in the simulation study, we restricted our optimization of $\boldsymbol{\omega}$ to the subspace of non-negative elements, which on average improves numerical results. In the data analysis, removing the constraint yields $\hat{\boldsymbol{\omega}}^{\text{opt}} = \{1.094, -0.100, 0.006\}$ and an MSPE of 0.601. These results are also presented in Table 2 and Figure 3 denoted as HYB$^{\text{unc}}$.

6. Discussion

Augmenting high-dimensional data with external auxiliary information is useful to boost predictive accuracy. We have described how to quantify this auxiliary information using important ideas from the measurement error and shrinkage literature. The regression calibration algorithm (SRC) yields unbiased estimates of future outcomes but with large variance when p is large. A modified algorithm (FRC) makes a bias-variance trade-off and can give a smaller MSPE. We have also proposed a hybrid estimator (HYB) which is a linear combination of multiple estimators. In addition to point estimates, prediction intervals for capturing uncertainty are also typically of interest. A simple bootstrap algorithm yields prediction intervals but will require some modifications to achieve nominal coverage rates.

HYB stands out as the method of choice. Theorem 3.1 demonstrates its theoretical utility, and, practically, the average performance of $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ across all design and data configurations is encouraging. Importantly, its flexibility is most apparent in the large p scenarios. Because we combined TR estimators, a GCV criterion provides a simple estimate of $\boldsymbol{P}$, the prediction error matrix (3.14), which is required to optimize with respect to $\boldsymbol{\omega}$. When taking linear combinations of arbitrary estimates of $\boldsymbol{\beta}$ for which GCV is not conducive, a challenge is how to estimate $\boldsymbol{P}$; the "632 estimator" of Efron (1983) is one candidate.

Of potential concern is that we have applied our methods, developed for continuous endpoints, to a dataset with censored survival time as the endpoint. In much the same way as ridge regression has been applied to logistic and Cox models, the targeted ridge class may also be adapted to other endpoints. While our theoretical and numerical results have focused only on continuous endpoints, we believe that the ideas and intuition developed will generally transfer to these other endpoints. However, the extension is non-trivial and merits in-depth research, not only for deriving estimators but also in determining the right criterion with which to assess prediction.

That this is a missing data problem can be exploited further than the single imputations considered in this paper. Multiple imputation using chained equations can make repeated draws of the missing $\boldsymbol{x}_{\text{B}}$ as was done in Chen *and others* (2011). Or, by writing out the complete likelihood, a data augmentation/Gibbs sampler algorithm can make alternating draws from the posterior distribution of $\boldsymbol{x}_{\text{B}}$, $\boldsymbol{\beta}$ and the rest of the model's parameters. Apart from the computationally demanding aspects of Bayesian methods, because of the size of p and the large fraction of missing data, a fully Bayesian extension is not automatic. In particular, careful thought must be given to the choice of prior on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$, because the sampler is likely not to converge with non-informative priors.

In a likelihood-based approach, the NDME assumption (ie $[Y|\boldsymbol{X}, \boldsymbol{W}] = [Y|\boldsymbol{X}]$) can be relaxed. Violations to this assumption will change the MSPE of the methods we considered, although our simulations have shown robustness for several of the methods, particularly HYB. However, a likelihood-based method, including fully

Bayesian approaches, may be more sensitive to violations of other model assumptions.

The development of TR estimators assumes that $\boldsymbol{x}_{\mathrm{B}}$ is missing completely at random. More thorough development of these methods under other missingness mechanisms would be of interest. Outcome dependent sampling (ODS) (Weaver and Zhou, 2005; Qin and Zhou, 2011) and two-phase sampling (Neyman, 1938) would be important cases to consider, since designs like these are an appealing way to select the subsample on which expensive measures are taken. It is usually noted that ODS can enhance efficiency but will introduce bias if the sampling mechanism is not properly accounted for in the analysis. However, MSPE is a function of bias and efficiency, thus methods and results from the existing ODS literature that focus on obtaining consistent and unbiased estimates do not directly apply to the prediction context. Also, the high-dimensional aspect of the data implies that standard methods for analyzing two-phase likelihoods would not apply. If a TR estimator which is robust to other missingness mechanisms were developed, it could be included as an ingredient to HYB to balance efficiency and robustness in predictions. To conclude, the vast majority of shrinkage, regression calibration and ODS literature has focused on estimation rather than prediction. The use of these techniques to improve prediction merits further research.

Acknowledgments

This work was supported by the National Science Foundation [DMS1007494]; and the National Institutes of Health [CA129102, CA156608].

Supplementary Materials

Supplementary material is available at `http://biostatistics.oxfordjournals.org`

References

Breiman, Leo. (1996). Stacked regressions. *Machine Learning* **24**, 49–64.

Buzas, Jeffrey, Stefanski, Leonard and Tosteson, Tor. (2005). Measurement error. In: Ahrens, Wolfgang and Pigeot, Iris (editors), *Handbook of Epidemiology*. Springer Berlin Heidelberg, pp. 729–765.

Carroll, Raymond J., Ruppert, David, Stefanski, Leonard A. and Crainiceanu, Ciprian M. (2006). *Measurment Error in Nonlinear Models*, Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Casella, George. (1980). Minimax ridge regression estimation. *The Annals of Statistics* **8**(5), 1036–1056.

Chen, Guoan, Kim, Sinae, Taylor, Jeremy M G, Wang, Zhuwen, Lee, Oliver, Ramnath, Nithya, Reddy, Rishindra M, Lin, Jules, Chang, Andrew C, Orringer, Mark B *and others*. (2011). Development and validation of a qRT-PCR–classifier for lung cancer prognosis. *Journal of Thoracic Oncology* **6**(9), 1481–1487.

Chen, Ming-Hui and Ibrahim, Joseph G. (2000). Power prior distributions for regression models. *Statistical Science* **15**(1), 46–60.

Chen, Yi-Hau, Chatterjee, Nilanjan and Carroll, Raymond J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association* **104**(485), 220–233.

Craven, Peter and Wahba, Grace. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Dempster, A. P., Schatzoff, Martin and Wermuth, Nanny. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association* **72**(357), 77–91.

Draper, Norman R. and van Nostrand, R. Craig. (1979). Ridge regression and James-Stein estimation: Review and comments. *Technometrics* **21**(4), 451–466.

Efron, Bradley. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation.

Journal of the American Statistical Association **78**(382), 316–331.

Frank, Ildiko E and Friedman, Jerome H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**(2), 109–135.

Fuller, Wayne A. (2006). *Measurement Error Models*, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Fumera, Giorgio and Roli, Fabio. (2005). A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(6), 942–956.

Gelfand, Alan E. (1986). On the use of ridge and Stein-type estimators in prediction. *Technical Report* 374, Stanford University.

George, Edward I. (1986). Minimax multiple shrinkage estimation. *The Annals of Statistics* **14**(1), 188–205.

Green, Edwin J. and Strawderman, William E. (1991). A James-Stein type estimator for combining unbiased and possibly biased estimators. *Journal of the American Statistical Association* **86**(416), 1001–1006.

Gruber, Marvin H. J. (1998). *Improving efficiency by shrinkage: the James-Stein and ridge regression estimators*. New York: Marcel Dekker, Inc.

Hoerl, Arthur E. and Kennard, Robert W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.

James, W. and Stein, Charles. (1961). Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1. University of California Press. pp. 361–379.

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

Li, Ker-Chau. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* **14**(3), 1101–1112.

Mallows, C.L. (1973). Some comments on CP. *Technometrics* **15**(4), 661–675.

Maruyama, Yuzo and Strawderman, William E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *The Annals of Statistics* **33**(4), 1753–1770.

Neyman, Jerzy. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33**(201), 101–116.

Qin, Guoyou and Zhou, Haibo. (2011). Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. *Biostatistics* **12**(3), 506–520.

Rao, C. Radhakrishna. (1945). Generalisation of Markoff's theorem and tests of linear hypotheses. *Sankhyā: The Indian Journal of Statistics* **7**(1), 9–16.

Rao, C. Radhakrishna. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* **31**(2), 545–554.

Schäfer, Juliane and Strimmer, Korbinian. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**(1), Article 32.

Sclove, Stanley L. (1968). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association* **63**(322), 596–606.

Swindel, Benee F. (1976). Good ridge estimators based on prior information. *Communications in Statistics* **5**(11), 1065–1075.

Weaver, Mark A and Zhou, Haibo. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100**(470), 459–469.

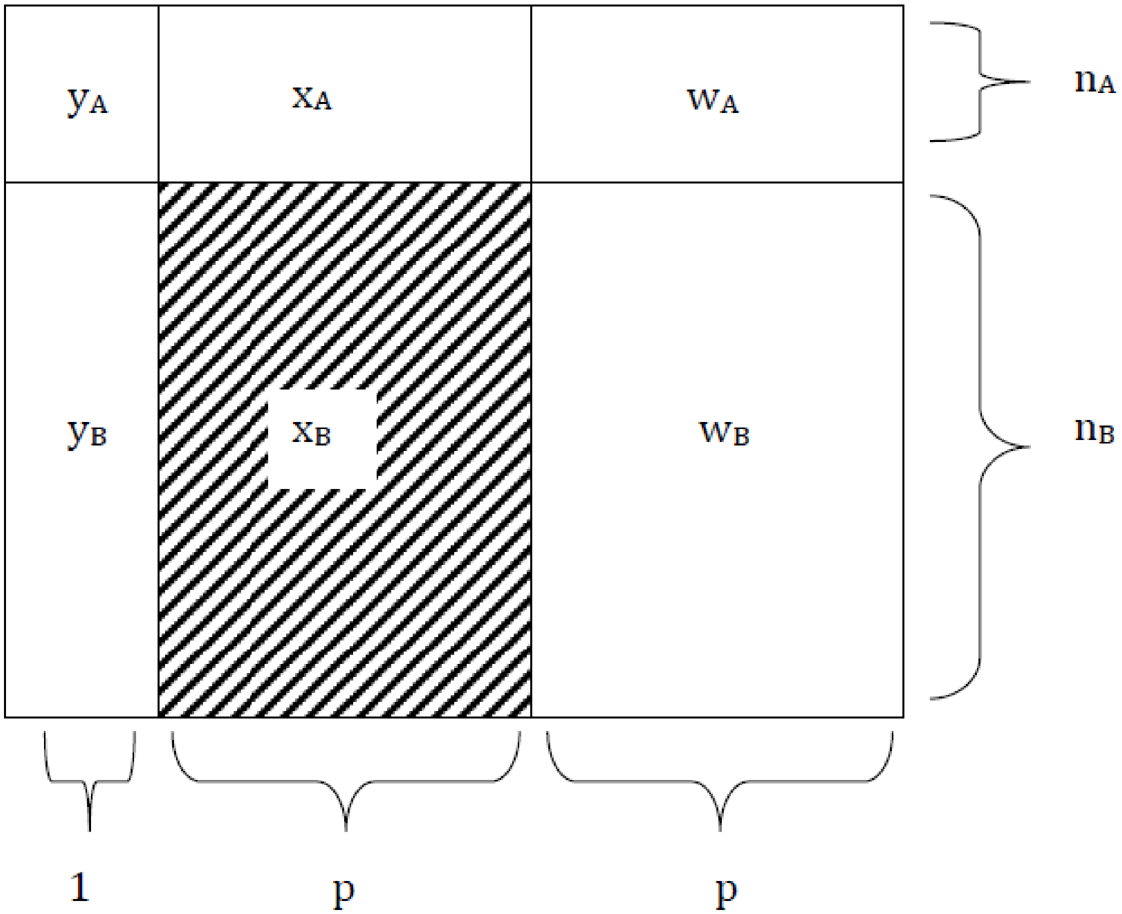

Fig. 1. Schematic representation of the prediction problem: $(\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}})$ is measured on n_{A} subjects and $(\boldsymbol{y}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}})$ is measured on n_{B} subjects. $\boldsymbol{x}_{\mathrm{B}}$ is considered missing. $\boldsymbol{W}$ is a error-prone/noisy version of $\boldsymbol{X}$. The goal is to utilize the data on $\boldsymbol{W}$ to boost prediction of Y by $\boldsymbol{X}$

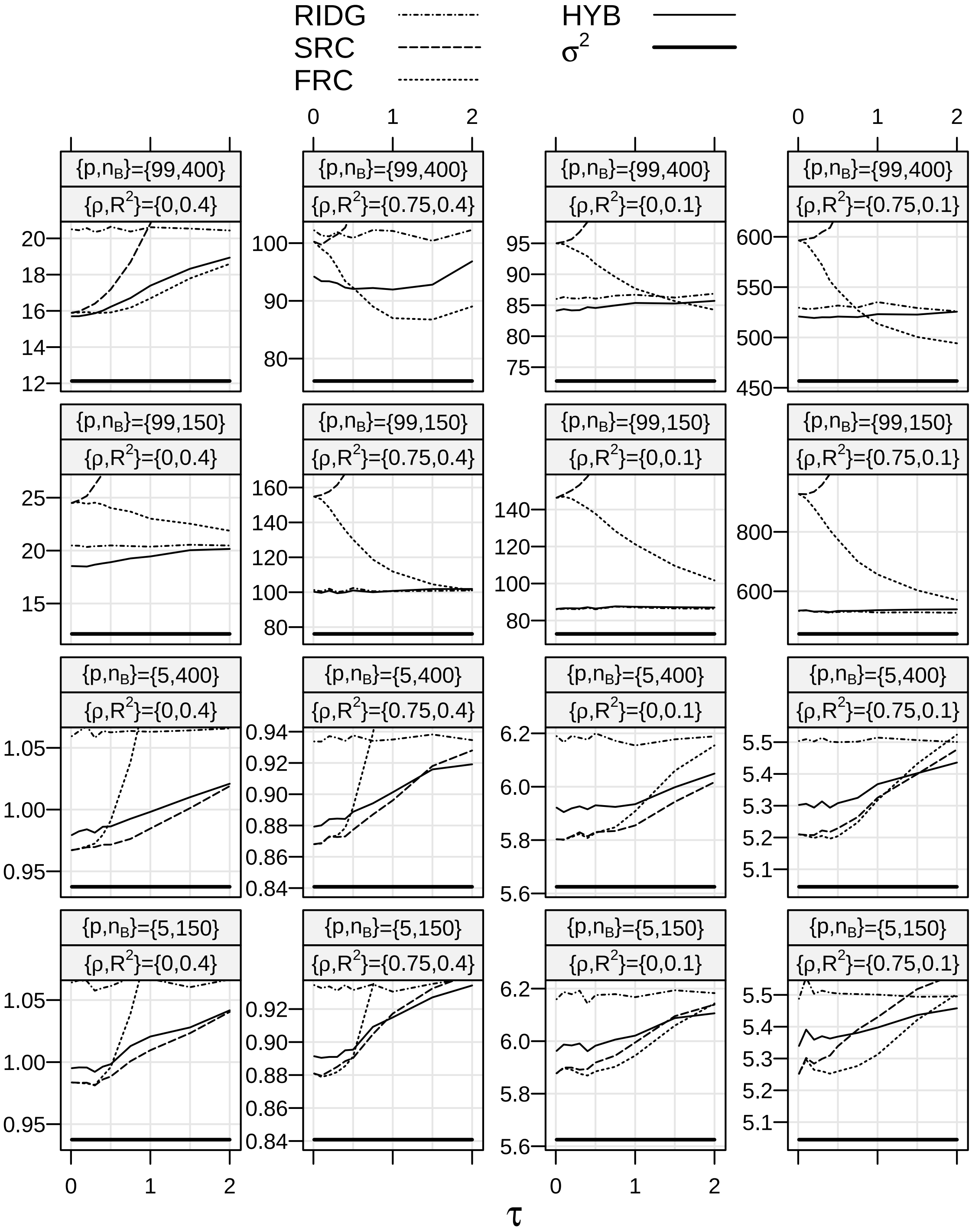

Fig. 2. Empirical MSPE over τ for the simulation study described in Section 4. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2}{\sigma^2}$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}} = 50$, $\beta_0 = \psi = 0$, and $\nu = 1.\sigma^2$, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

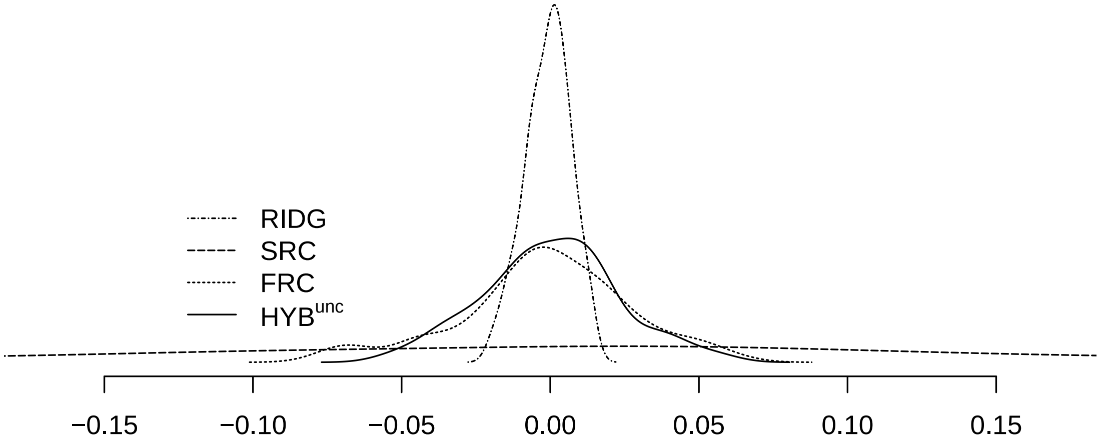

Fig. 3. Kernel density estimate of the 91 elements of $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$, $\hat{\boldsymbol{\beta}}_{\text{SRC}}$, $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ and $\hat{\boldsymbol{\beta}}_{\text{HYB}^{\text{unc}}}$ (the hybrid estimator without the non-negativity constraint – Remark 5.1) from the data analysis. $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ (with the non-negativity constratint) is identically equal to $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$.

Table 1. Key information for several TR estimators, conditioning on the true value of $\boldsymbol{\theta}$. $\kappa = (\tau^2/\nu^2)\boldsymbol{\beta}^\top \boldsymbol{V}\boldsymbol{\beta}$. $\boldsymbol{V} = (\boldsymbol{I}_p + (\tau^2/\nu^2)\boldsymbol{\Sigma}_{\boldsymbol{X}}{}^{-1})$. The '$\lambda = 1$?' column indicates whether λ is fixed at 1 or tuned in a data-adaptive fashion using the general GCV function. The corresponding estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ is given by plugging $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ into (2.7). The expectation and variance of $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$, which are useful for calculating the MSPE of $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$, are over $\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{y}_{\mathrm{B}} | \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{B}}$ under the assumption $[Y|\boldsymbol{X}, \boldsymbol{W}] = [Y|\boldsymbol{X}]$.

Method	$\boldsymbol{\gamma}_{\boldsymbol{\beta}}$	$\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$	$\lambda = 1$?
RIDG	$\boldsymbol{0}_p$	$\boldsymbol{I}_p$	N
FRC	$\nu(\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}})^{-1}\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{y}_{\mathrm{B}}$	$\nu^{-2}\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}}$	Y
SRC	$\nu\boldsymbol{V}^{-1}(\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}})^{-1}\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{y}_{\mathrm{B}}$	$\nu^{-2}\boldsymbol{V}\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}}\boldsymbol{V}$	Y
Method	$\mathrm{E}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}$	$\mathrm{Var}\,\boldsymbol{\gamma}_{\boldsymbol{\beta}}$	
RIDG	–	–	
FRC	$\boldsymbol{V}\boldsymbol{\beta}$	$(\sigma^2 + \kappa)\nu^2(\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}})^{-1}$	
SRC	$\boldsymbol{\beta}$	$(\sigma^2 + \kappa)\nu^2\boldsymbol{V}^{-1}(\boldsymbol{w}_{\mathrm{B}}^\top \boldsymbol{w}_{\mathrm{B}})^{-1}\boldsymbol{V}^{-1}$	

Table 2. Results from the data analysis. $\hat{\text{MSPE}}$ is the empirical MSPE from the validation sample of size 100, $\min(\hat{\boldsymbol{\beta}})$ and $\max(\hat{\boldsymbol{\beta}})$ give the range of the estimate of $\boldsymbol{\beta}$ for each model, Avg. Coverage is the proportion of bootstrap-generated prediction intervals for the validation sample which contained the true outcome, and $\text{Avg}(\hat{Y}_{\text{new}}^{B,97.5} - \hat{Y}_{\text{new}}^{B,2.5})$ gives the average prediction interval length for the validation sample. HYB$^{\text{unc}}$ is the hybrid estimator *without* the non-negativity constraint (Remark 5.1).

	RIDG	SRC	FRC	HYB	HYB$^{\text{unc}}$
$\hat{\text{MSPE}}$	0.620	8.745	0.781	0.620	0.601
$\min(\hat{\boldsymbol{\beta}})$	-0.019	-0.600	-0.075	-0.019	-0.054
$\max(\hat{\boldsymbol{\beta}})$	0.014	0.516	0.062	0.014	0.058
Avg. Coverage	0.91	1.00	0.98	0.91	0.99
$\text{Avg}(\hat{Y}_{\text{new}}^{B,97.5} - \hat{Y}_{\text{new}}^{B,2.5})$	3.372	33.785	4.023	3.372	4.674