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

Philip S. Boonstra[1], Jeremy M.G. Taylor[1], Bhramar Mukherjee[1]
[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109
Email: philb@umich.edu

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 $\boldsymbol{X}$ towards different targets that use information derived from $\boldsymbol{W}$ in the larger dataset, comparing these with the classical ridge regression of Y on $\boldsymbol{X}$, which does 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 $\boldsymbol{\beta}$ and 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 of 91 genes is measured by quantitative real-time polymerase chain reaction (qRT-PCR) and microarray technology on 47 lung cancer patients with microarray measurements available on an additional 392 patients. The goal is to predict uncensored survival time using qRT-PCR. Thus a model which fits well can help identify those patients with the highest risk of death. The methods are evaluated on an independent sample of 101 patients.

KEY WORDS: Cross-validation, Mean Squared Prediction Error, Measurement Error, Ridge Regression, Generalized Ridge

1 Introduction

As sequencing and array technologies change, multiple platforms can measure biologically identical quantities of interest. Often, investigators have a large sample containing measurements from an older technology, with measures from the newer technology available on a subset of this sample. We are interested in predicting an outcome using the newer measures, which is a statistical problem of building 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:

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ (to save on notation, no intercept is included). On a small number of subjects (n_A), we have Y, $\boldsymbol{X}$ and $\boldsymbol{W}$, where $\boldsymbol{W}$ is also of dimension p, representing the same biomarkers as $\boldsymbol{X}$ measured with a prior technology. A model for $\boldsymbol{W}|\boldsymbol{X}$ which is consistent with this motivating context is

$$\boldsymbol{W} = \nu\boldsymbol{X} + \tau\boldsymbol{\xi}, \tag{2}$$

where $\boldsymbol{\xi}$ is multivariate standard normal noise and ν and τ are scalars.

The quantity n_A is of modest size, such that $p > n_\mathrm{A}$. Also available is a larger set of n_B observations of Y and $\boldsymbol{W}$. We assume $p < n_\mathrm{B}$. Denote subsamples A and B by $(\boldsymbol{y}_\mathrm{A}, \boldsymbol{x}_\mathrm{A}, \boldsymbol{w}_\mathrm{A})$ and $(\boldsymbol{y}_\mathrm{B}, \boldsymbol{w}_\mathrm{B})$, respectively (see Figure 1 for a schematic representation). We assume that each sample comes from the same population. Assume further that $\boldsymbol{x}_\mathrm{A}$ is standardized, ie if x_{ij} is the element from the ith row and jth column of $\boldsymbol{x}_\mathrm{A}$, $\sum_{i=1}^{n_\mathrm{A}} x_{ij} = 0$ and $\sum_{i=1}^{n_\mathrm{A}} x_{ij}^2 = n_\mathrm{A}$, $j = 1, \ldots, p$, and that $\boldsymbol{w}_\mathrm{A}$ and $\boldsymbol{w}_\mathrm{B}$ are individually centered, $\sum_{i=1}^{n_\mathrm{A}} w_{ij} = \sum_{i=n_\mathrm{A}+1}^{n_\mathrm{A}+n_\mathrm{B}} w_{ij} = 0$, $j = 1, \ldots, p$. Let $\boldsymbol{\xi}_\mathrm{A}$ and $\boldsymbol{\xi}_\mathrm{B}$ be the unobserved $n_\mathrm{A} \times p$ and $n_\mathrm{B} \times p$ error matrices for subsamples A and B.

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

$$\begin{aligned}\mathrm{MSPE}(\widehat{\boldsymbol{\beta}}) := \mathrm{E}[(Y_\mathrm{new} - \boldsymbol{X}_\mathrm{new}^\top \widehat{\boldsymbol{\beta}})^2] &= \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{X}_\mathrm{new} \boldsymbol{X}_\mathrm{new}^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})] \\ &= \sigma^2 + \mathrm{Tr}\big[(\mathrm{Bias}\,\widehat{\boldsymbol{\beta}}\,\mathrm{Bias}\,\widehat{\boldsymbol{\beta}}^\top + \mathrm{Var}\,\widehat{\boldsymbol{\beta}})\mathrm{E}[\boldsymbol{X}_\mathrm{new}\boldsymbol{X}_\mathrm{new}^\top]\big], \end{aligned} \tag{3}$$

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 (OLS) of $\boldsymbol{y}_{\mathrm{A}}$ on $\boldsymbol{x}_{\mathrm{A}}$, i.e. $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} := \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) = (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$. However, the inversion of $\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}}$ is not possible for $p > n_{\mathrm{A}}$. Even for $p \leq n_{\mathrm{A}}$, 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 the coefficients towards zero, i.e. $\widehat{\boldsymbol{\beta}}_{\mathrm{RIDG}} := \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta} = (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$. This can be viewed from a Bayesian perspective: given a normal prior on $\boldsymbol{\beta}$ with mean $\boldsymbol{0}_p$ and precision $\sigma^{-2}\lambda\boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is the identity matrix, 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, $\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) := \mathrm{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$, compared to $\lambda = 0$.

Dempster et al. (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 simultaneously shrink towards multiple targets. Green and Strawderman (1991) consider a *targeted* JS estimator: an unbiased estimator is shrunk towards 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 et al., 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 et al. (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 $\widehat{Y}(\boldsymbol{W}_{\text{new}})$. Our prediction model of interest being Y given $\boldsymbol{X}$, this bias in $\widehat{\boldsymbol{\beta}}$ from using $\boldsymbol{W}$ instead of $\boldsymbol{X}$ *does* bias $\widehat{Y}(\boldsymbol{X}_{\text{new}})$ 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}. However, the substitution of $\boldsymbol{X}$ by $\boldsymbol{W}$ may reduce the *variance* of estimates of $\boldsymbol{\beta}$ relative to regression calibration (Buzas et al., 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 $\widehat{Y}(\boldsymbol{X}_{\text{new}})$.

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, and evaluate these estimators via analytical and simulation studies. We also propose a cross-validation strategy to select the tuning parameter λ for this class of estimators. Finally, we 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 improved prediction.

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 consider some adaptive shrinkage estimators. We add a data-adaptive/tuning parameter selection feature (3.1) and we propose hybrid estimators which combine multiple estimators with data-adaptive weights (3.2). 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 the predictions. Section 6 concludes with a discussion. Some analytical details are in the Appendix.

2 Targeted Shrinkage

For $p > n_{\text{A}}$, OLS 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),

$$\widehat{\boldsymbol{\beta}}_{\text{RIDG}} = (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}}. \tag{4}$$

RIDG is equivalent to adding λ to each eigenvalue of $\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}}$, 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 $p > n_{\mathrm{A}}$ is crucial in our setting. Shrinkage estimators from Sclove (1968), Strawderman (1978) and Casella (1980) are built upon an *unbiased* estimator of $\boldsymbol{\beta}$ and hence are not directly applicable for $p > n_{\mathrm{A}}$ situations.

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

$$\frac{\frac{1}{n_{\mathrm{A}}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{H}(\lambda\boldsymbol{I}_p)\boldsymbol{y}_{\mathrm{A}})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{H}(\lambda\boldsymbol{I}_p)\boldsymbol{y}_{\mathrm{A}})}{(1 - \mathrm{Tr}\ \boldsymbol{H}(\lambda\boldsymbol{I}_p)/n_{\mathrm{A}})^2}, \qquad \boldsymbol{H}(\boldsymbol{\Theta}) = \boldsymbol{x}_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Theta})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}. \tag{5}$$

Rao (1975) suggested that any positive semi-definite matrix $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ can replace $\boldsymbol{I}_p$ in (4). Swindel (1976) proposed to shrink towards 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 $\boldsymbol{0}_p$ with $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. The posterior mode is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}) &= \mathrm{argmin}_{\boldsymbol{\beta}}\tfrac{1}{\sigma^2}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) + \tfrac{1}{\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\gamma}_{\boldsymbol{\beta}})^{\top}\lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\gamma}_{\boldsymbol{\beta}}) \\ &= (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})^{-1}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}} + \lambda\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}}).\end{aligned} \tag{6}$$

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 towards a target $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. The estimator given in (6) is a main focus of this paper and 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: $\widehat{\boldsymbol{\beta}}_{\mathrm{RIDG}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{0}_p, \lambda, \boldsymbol{I}_p)$.

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. The TR methods we present below fix $\lambda = 1$ (in 3.1, we consider data-adaptive estimation of λ).

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 $\mathcal{N}_p(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X})$, implies $\boldsymbol{X}|\boldsymbol{W}$ is normally distributed. Let $\boldsymbol{\theta} = \{\nu, \tau, \boldsymbol{\mu_X}, \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\}$. From properties of the conditional distribution of $\boldsymbol{X}|\boldsymbol{W}$,

$$\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}}(\boldsymbol{\theta}) := \mathrm{E}[\boldsymbol{x}_\mathrm{B}|\boldsymbol{w}_\mathrm{B}, \boldsymbol{\theta}] = \tfrac{\tau^2}{\nu^2}\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu_X}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{V}(\boldsymbol{\theta}) + \tfrac{1}{\nu}\boldsymbol{w}_\mathrm{B}\boldsymbol{V}(\boldsymbol{\theta}) = [\mathbf{1}_{n_\mathrm{B}}, \boldsymbol{w}_\mathrm{B}]\boldsymbol{M}(\boldsymbol{\theta}), \tag{7}$$

where

$$\boldsymbol{M}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\tau^2}{\nu^2}\boldsymbol{\mu_X}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\boldsymbol{V}(\boldsymbol{\theta}) \\ \frac{1}{\nu}\boldsymbol{V}(\boldsymbol{\theta}) \end{pmatrix} \quad \text{and} \quad \boldsymbol{V}(\boldsymbol{\theta}) = (\boldsymbol{I}_p + \tfrac{\tau^2}{\nu^2}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})^{-1} \tag{8}$$

(we suppress the dependence on $\boldsymbol{\theta}$ of $\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}}(\boldsymbol{\theta})$, $\boldsymbol{M}(\boldsymbol{\theta})$, and $\boldsymbol{V}(\boldsymbol{\theta})$ hereafter). This is a precision-weighted average of $\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu_X}^\top$ and $(1/\nu)\boldsymbol{w}_\mathrm{B}$; assume hereafter that $\boldsymbol{\mu_X} = \mathbf{0}_p$. Using (6), define $\widehat{\boldsymbol{\beta}}_{\mathrm{SRC}} := \widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{SRC}}}, 1, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\mathrm{SRC}}})$, where, $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{SRC}}} = (\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}})^{-1}(\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}\top}\boldsymbol{y}_\mathrm{B})$ and $\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\mathrm{SRC}}} = \boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{SRC}}$. In the ME literature, SRC is the method typically meant by "Regression Calibration". We append "Structural" (Carroll et al., 2006, p.25), meaning it makes a distributional assumption about $\boldsymbol{X}$, to distinguish it from its "Functional" counterpart, which does not make this assumption, proposed as follows.

Functional Regression Calibration (FRC): Solving (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

$$\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}}(\boldsymbol{\theta}) := (1/\nu)\boldsymbol{w}_\mathrm{B}, \qquad \boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{FRC}}} = (\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}})^{-1}\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}\top}\boldsymbol{y}_\mathrm{B}, \qquad \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\mathrm{FRC}}} = \boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}}. \tag{9}$$

This gives a TR estimate defined as $\widehat{\boldsymbol{\beta}}_{\mathrm{FRC}} := \widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{FRC}}}, 1, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\mathrm{FRC}}})$. This imputation for $\boldsymbol{x}_\mathrm{B}$ is a scaled version of a direct 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.

The first rows of Table 1 summarize the choices of $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ for the RIDG, FRC, and SRC methods. An approximate relationship between RIDG and FRC in terms of their functional forms leads to the following properties.

(i) $\boldsymbol{x}_\mathrm{B}^{\mathrm{FRC}}$ is a "contaminated" version of $\boldsymbol{x}_\mathrm{B}$, i.e., it can be viewed as the true covariates with added normal noise.

(ii) $\widehat{\boldsymbol{\beta}}_{\text{FRC}}$ is an approximate Ridge-type estimator.

(iii) Contaminating covariates can thus improve prediction error in our setting.

Proof.
(i) This follows immediately, as by definition, $\boldsymbol{x}_{\text{B}}^{\text{FRC}} = (1/\nu)\boldsymbol{w}_{\text{B}} = \boldsymbol{x}_{\text{B}} + (\tau/\nu)\boldsymbol{\xi}_{\text{B}}$.
(ii) From property (i) and the definition of $\boldsymbol{x}_{\text{B}}^{\text{FRC}}$ in (9), we have:

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1} = \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \tfrac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}}, \qquad \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}} = \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} \tag{10}$$

Plugging these values of $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}$ and $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}}$ into (6) gives that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{FRC}} &= (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \tfrac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}})^{-1}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}) \\ &\approx (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + n_{\text{B}}\tfrac{\tau^2}{\nu^2}\boldsymbol{I}_p)^{-1}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}),\end{aligned} \tag{11}$$

where the last approximation replaces each expression involving $\boldsymbol{\xi}_{\text{B}}$ in the previous line with its marginal expectation. Thus (11) characterizes $\widehat{\boldsymbol{\beta}}_{\text{FRC}}$ as an approximate ridge-type estimator based on the complete data, with the shrinkage parameter $n_{\text{B}}\tau^2/\nu^2$.
(iii) It is known that ridge regression can improve prediction error over ordinary least squares for certain choices of the tuning parameter (Gelfand, 1986; Frank and Friedman, 1993). We have shown in (ii) that FRC is an approximate ridge-type method with tuning parameter depending on τ/ν. We have also shown in (i) that $\boldsymbol{x}_{\text{B}}^{\text{FRC}}$ is a contaminated version of the true unobserved $\boldsymbol{x}_{\text{B}}$. It then follows that one can find choices of τ and ν and construct contaminated covariates $\boldsymbol{x}_{\text{B}} + (\tau/\nu)\boldsymbol{\xi}_{\text{B}}$ to yield smaller prediction error than using the actual observed covariates $\boldsymbol{x}_{\text{B}}$. □

REMARK 1: Following a similar expansion for SRC as above, note that $\boldsymbol{x}_{\text{B}}^{\text{SRC}} = (1/\nu)\boldsymbol{w}_{\text{B}}\boldsymbol{V} = \boldsymbol{x}_{\text{B}}\boldsymbol{V} + (\tau/\nu)\boldsymbol{\xi}_{\text{B}}\boldsymbol{V}$. When we expand $\widehat{\boldsymbol{\beta}}_{\text{SRC}}$ as in (11), we obtain

$$\widehat{\boldsymbol{\beta}}_{\text{SRC}} = (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}}\boldsymbol{V} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}}\boldsymbol{V} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}}\boldsymbol{V} + \tfrac{\tau^2}{\nu^2}\boldsymbol{V}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}}\boldsymbol{V})^{-1}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \boldsymbol{V}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}) \tag{12}$$

From (8), as $\tau/\nu \to \infty$, the elements of $\boldsymbol{V}$ go to zero at a rate proportional to τ^2/ν^2. Thus, for large τ/ν, $\widehat{\boldsymbol{\beta}}_{\text{SRC}}$ is "unstable", because it approximates $(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}})^{-1}\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}}$, the OLS estimate of $\boldsymbol{\beta}$, which does not exist when $n_{\text{A}} > p$.

3 Adaptive Shrinkage Using Cross-Validation

3.1 Targeted Ridge With Adaptive Component-Wise Shrinkage

In the observations following (11) we noted that $\widehat{\boldsymbol{\beta}}_{\text{FRC}}$ can make improved predictions, even over the complete data case, for moderate values of τ/ν, but $\widehat{\boldsymbol{\beta}}_{\text{FRC}}$ will be close to $\mathbf{0}_p$ for very large values of τ/ν. However, τ/ν is fixed for a given dataset and can not be treated as a tuning parameter. To make the amount of shrinkage data-adaptive one can choose a $\lambda \neq 1$ in (6) based on the data. We propose a general version of the GCV function for a TR estimator: minimize

$$\frac{\frac{1}{n_{\text{A}}}(\boldsymbol{y}^*_{\text{A}} - \boldsymbol{H}(\lambda\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}})\boldsymbol{y}^*_{\text{A}})^\top(\boldsymbol{y}^*_{\text{A}} - \boldsymbol{H}(\lambda\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}})\boldsymbol{y}^*_{\text{A}})}{(1 - \text{Tr }\boldsymbol{H}(\lambda\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}})/n_{\text{A}})^2} \tag{13}$$

with respect to λ, where $\boldsymbol{y}^*_{\text{A}} = \boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}}\boldsymbol{\gamma}_{\boldsymbol{\beta}}$ and $\boldsymbol{H}(\cdot)$ is given in (5). Craven and Wahba (1979) chose the original GCV, given in (5), to down-weight the squared cross-validated residuals of highly influential observations, using the diagonals of $\boldsymbol{H}(\cdot)$ to measure influence. Some algebra shows that (13) comes from using these same weights, and so is analogous to GCV for RIDG.

The adaptive parameter λ can be included in the entire class of TR estimators. However, it does little for SRC: the instability of $\widehat{\boldsymbol{\beta}}_{\text{SRC}}$ demonstrated by (12) would not be changed by choosing $\lambda \neq 1$. Rather, a better candidate for adaptive shrinkage is a component-wise version of FRC called FRC.CW: $\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC.CW}}} = \text{diag}(\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})$ and $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC.CW}}} = \boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}}$. This approach retains features of both RIDG and FRC; like RIDG, the shrinkage is component-wise and data-adaptive, but, as in FRC, the amount of shrinkage can vary between components, and the target of shrinkage is derived from subsample B. The corresponding information for FRC.CW is also included in Table 1.

3.2 Hybrid Estimators

Combining multiple estimators is another strategy to incorporate the auxiliary information. Given two estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, let $\boldsymbol{b}(\omega) := \omega\widehat{\boldsymbol{\beta}}_1 + (1-\omega)\widehat{\boldsymbol{\beta}}_2$. A sensible choice of ω would be

that which minimizes

$$\mathrm{MSPE}(\boldsymbol{b}(\omega)) = \omega^2\mathrm{MSPE}(\widehat{\boldsymbol{\beta}}_1) + (1-\omega)^2\mathrm{MSPE}(\widehat{\boldsymbol{\beta}}_2) + 2\omega(1-\omega)\mathrm{MCPE}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2), \tag{14}$$

where $\mathrm{MCPE}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2) = \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_1)^\top \boldsymbol{X}_{\mathrm{new}} \boldsymbol{X}_{\mathrm{new}}^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_2)]$ is a "mean cross-product prediction error". Standard calculus then gives

$$\omega^{\mathrm{opt}} = \frac{\mathrm{MSPE}(\widehat{\boldsymbol{\beta}}_2) - \mathrm{MCPE}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)}{\mathrm{MSPE}(\widehat{\boldsymbol{\beta}}_1) + \mathrm{MSPE}(\widehat{\boldsymbol{\beta}}_2) - 2\mathrm{MCPE}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)}. \tag{15}$$

To combine m estimators, $\boldsymbol{\omega}$ is a length-m vector. Then, $\mathrm{MSPE}(\boldsymbol{b}(\boldsymbol{\omega})) = \boldsymbol{\omega}^\top \boldsymbol{P} \boldsymbol{\omega}$, where $\boldsymbol{P}$ is the $m \times m$ matrix with the (k_1, k_2)th element

$$P_{k_1,k_2} := \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{k_1})^\top \boldsymbol{X}_{\mathrm{new}} \boldsymbol{X}_{\mathrm{new}}^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{k_2})] = \mathrm{MCPE}(\widehat{\boldsymbol{\beta}}_{k_1}, \widehat{\boldsymbol{\beta}}_{k_2}). \tag{16}$$

We seek to minimize $\boldsymbol{\omega}^\top \boldsymbol{P} \boldsymbol{\omega}$ subject to $\mathbf{1}_m^\top \boldsymbol{\omega} = 1$. In general, $\boldsymbol{P} \succeq 0$, so that a global minimum always exists; this minimum is unique if $\boldsymbol{P} \succ 0$. However, the solution can not be analytically written for arbitrary $\boldsymbol{P}$.

The following result compares the performance of the hybrid estimator to that of its constituents.

Theorem 3.1. *Let $\boldsymbol{b}(\boldsymbol{\omega}) := [\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \cdots, \widehat{\boldsymbol{\beta}}_m]\boldsymbol{\omega}$ be a hybrid estimator and $\boldsymbol{\omega}^{opt}$ the weight vector which minimizes MSPE$(\boldsymbol{b}(\boldsymbol{\omega}))$ subject to $\mathbf{1}_m^\top \boldsymbol{\omega} = 1$. Then, MSPE$(\boldsymbol{b}(\boldsymbol{\omega}^{opt})) \leq \min\{$MSPE$(\widehat{\boldsymbol{\beta}}_\ell) : \ell = 1, \ldots, m\}$.*

Thus $\boldsymbol{b}(\boldsymbol{\omega}^{\mathrm{opt}})$ performs no worse 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.

In practice, however, $\boldsymbol{P}$ and therefore $\boldsymbol{\omega}^{\mathrm{opt}}$ are not precisely known and must be estimated. We propose to adapt the GCV approach in (13) such that

$$\widehat{P}_{k_1,k_2} = \frac{\frac{1}{n_{\mathrm{A}}}(\boldsymbol{y}^*_{\mathrm{A},k_1} - \boldsymbol{H}(\lambda_{k_1}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_1})\boldsymbol{y}^*_{\mathrm{A},k_1})^\top (\boldsymbol{y}^*_{\mathrm{A},k_2} - \boldsymbol{H}(\lambda_{k_2}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_2})\boldsymbol{y}^*_{\mathrm{A},k_2})}{(1-\psi_{k_1})(1-\psi_{k_2})}, \tag{17}$$

where $\psi_\ell = \mathrm{Tr}\ \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},\ell})/n_{\mathrm{A}}$ and $\boldsymbol{y}^*_{\mathrm{A},\ell} = \boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\gamma}_{\boldsymbol{\beta},\ell}$. Because $\boldsymbol{y}^*_{\mathrm{A},\ell} - \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},\ell})\boldsymbol{y}^*_{\mathrm{A},\ell} = \boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\widehat{\boldsymbol{\beta}}_\ell$, this is a penalized version of its naïve counterpart.

Note the dual use of the GCV function. To calculate $\boldsymbol{b}(\boldsymbol{\omega})$, (13) is used to choose λ for each

$\widehat{\boldsymbol{\beta}}_\ell$. Then, fixing these choices of λ, (17) is employed on the $m(m+1)/2$ pairwise combinations of components in $\boldsymbol{b}(\boldsymbol{\omega})$ to estimate P. The particular hybrid estimator we evaluate has three contributing estimators, of the form $\widehat{\boldsymbol{\beta}}_{\text{HYB}} = [\widehat{\boldsymbol{\beta}}_{\text{RIDG}}\, \widehat{\boldsymbol{\beta}}_{\text{FRC}}\, \widehat{\boldsymbol{\beta}}_{\text{FRC.CW}}]\boldsymbol{\omega}$. 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 was found to greatly improve numerical results.

REMARK 2: The key aspect that makes $\widehat{\boldsymbol{\beta}}_{\text{HYB}}$ practical is that the sum $\sigma^2 + \text{E}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{HYB}})^\top \boldsymbol{X}_{\text{new}} \boldsymbol{X}_{\text{new}}^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\text{HYB}})]$ is the quantity to minimize. Estimating either of the two terms alone is a difficult task. Green and Strawderman (1991) propose a similar JS type estimator, limited to combining two estimators, which minimizes the MSE of $\boldsymbol{b}(\boldsymbol{\omega})$, deriving an equation similar in structure to (15). For their method, calculation of ω^{opt} requires an unbiased $\widehat{\boldsymbol{\beta}}_1$ and independent estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$. In our case, because MSPE (and not MSE) is of interest, we require neither unbiasedness nor independent estimators.

4 Simulation Study

We conducted a simulation study to evaluate each method, fixing $n_{\text{A}} = 50$ and using $n_{\text{B}} \in \{400, 150\}$. 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 dimensional ($p = 99$) and low dimensional ($p = 5$) 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 $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2) = R^2$. $\boldsymbol{y}_{\text{A}} | \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{y}_{\text{B}} | \boldsymbol{x}_{\text{B}}$ were drawn for each combination of $\boldsymbol{\beta}$ and σ from (1). This yielded 16 unique simulation settings, which are listed in Table 2.

To draw the auxiliary data, we set $\nu = 1$ and repeated each of the 16 settings for $\tau \in \{10^{-4}, 0.1, 0.25, 0.5, 1, 2, 3, 4, 5, 6\}$, drawing $\boldsymbol{w}_{\text{A}} | \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{w}_{\text{B}} | \boldsymbol{x}_{\text{B}}$ from (2).

For five methods (RIDG, SRC, FRC, FRC.CW, and HYB), we estimated the relative MSPE (rMSPE): $\text{MSPE}/\text{MSPE}_{\text{TRUE}} - 1$, where $\text{MSPE}_{\text{TRUE}}$ indicates the MSPE from predicting with the true value of $\boldsymbol{\beta}$, by averaging the squared prediction error over 250 new individuals. Figure 2 plots this estimated rMSPE (the median from 1000 replicates) over values of τ.

Effect of τ: RIDG is not affected by τ, as it does not use $\boldsymbol{w}_{\text{A}}$ or $\boldsymbol{w}_{\text{B}}$. FRC and SRC are equivalent when τ is very small, close to the complete data case. The rMSPE of SRC always rises with τ (this increase is sharp when $p = 99$ [Sims 1-8]). However, larger values of τ give favorable shrinkage in FRC. When $p = 99$, the τ for which FRC performs best is larger than zero; for

$p = 5$ (Sims 9-16), the optimal τ is quite small, and the rMSPE rises sharply with τ. FRC.CW does poorly when τ is small and $p = 99$, but HYB performs quite well, outperforming its three constituents over several values of τ in several scenarios (eg Sims 2-5,7).

Effect of R^2 and ρ: For $p = 99$, all methods have the smallest rMSPE when $R^2 = 0.1$ and $\rho = 0.75$ (Sim 7,8). The change in the ranking of each method is an interaction with τ. For example, in Sim 1 ($\rho = 0$), FRC performs best until $\tau \approx 5$, but in Sim 3 ($\rho = 0.75$) it fares poorly at that point. Changing R^2 and ρ had less effect when $p = 5$ (Sims 9-16).

Effect of n_{B}: The favorable properties of FRC and FRC.CW are a function of n_{B}, with $n_{\mathrm{B}} = 400$ (odd numbered Sims) generally having a smaller rMSPE than $n_{\mathrm{B}} = 150$ (even numbered Sims). HYB is less affected. When $p = 5$ (Sims 9-16), differences between values of n_{B} were smaller.

Effect of p: SRC fares poorly when $p = 99$ (Sims 1-8) and FRC fares poorly when $p = 5$ (Sims 9-16). FRC.CW always does well when $p = 5$ but can sometimes fail when $p = 99$; again, HYB adapts well. When $p = 5$, all the methods are similarly ranked regardless of other parameter settings. Including $\widehat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ as an additional ingredient in $\widehat{\boldsymbol{\beta}}_{\mathrm{HYB}}$ seems reasonable when p is small, but there is no reason to do so otherwise.

Violations to Normality of $\boldsymbol{X}$ Assumption and ME Structure: We also considered the situation where $\boldsymbol{X}$ is drawn from a multivariate t distribution with 5 degrees of freedom, scaled to maintain $\mathrm{Var}\,\boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. Simultaneously, we perturbed (2): instead of $\mathrm{Var}[w_{ij}|x_{ij}] = \tau^2$, the underlying true variance was $\mathrm{Var}[w_{ij}|x_{ij}] = \tau^2|x_{ij}|^{1/4}$. These results (not presented) did not appreciably change the ranking of each method.

When $\boldsymbol{\theta}$ is known: The unbiasedness of $\widehat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ was shown in the case that $\boldsymbol{\theta}$ is known; bias or variance in the estimates of the components of $\boldsymbol{\theta}$, particularly $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ because it is of a large dimension, may increase MSPE beyond our analytical derivations. In our simulation study, we estimated $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ using the shrinkage method of Schäfer and Strimmer (2005). However, that SRC does so poorly in the large p setting does not change if the true $\boldsymbol{\theta}$ is used.

Other values of the true $\boldsymbol{\beta}$ which spread the signal evenly over all components or concentrated the signal in a few elements did not appreciably change our results.

5 Predicting Survival Time from Gene Expression Measurements

We consider whether gene expression measurements offer information for predicting uncensored survival time in patients with lung cancer. Expression data are often collected using microarrays: mRNA from a tissue sample is exposed to a slide or chip containing complementary DNA or oligonucleotides which in turn correspond to known protein-encoding genes. The measured intensity at which the mRNA transcripts bind to the chip is a measure of the level of transcription in the sample. While able to measure thousands of genes at once, microarrays require specialized laboratory facilities for processing.

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 and typically considered a more precise measurement of gene expression than microarrays.

Our dataset comes from Chen et al. (2011), who selected 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) was measured on 439 tumor samples, and qRT-PCR measurements 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 survival time after surgery. An independent cohort of 101 tumors with qRT-PCR measurements and clinical covariates is available for validation. 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 the total sizes of subsample A, subsample B and the validation data were respectively 47, 389, and 100.

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 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 3 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 $w_{ij} = \nu_j x_{ij} + \tau\xi_{ij}$. We modeled ν_j as a random effect, distributed as $\mathcal{N}(\mu_\nu, \sigma_\nu^2)$, and used predictions $\{\widehat{\nu}_j\}$ 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, $\widehat{\tau} = .628$. Because our simulations indicate robustness to this assumption, this violation was ignored.

We present results for predicting uncensored survival time in the validation data using RIDG, FRC, FRC.CW, and HYB (SRC gave poor predictive performance). In addition to analyzing the observed data, we "added noise" to replicate the favorable shrinkage properties that larger values of τ can induce. Independent noise drawn from $N(0, \tau^{*2})$, $\tau^* \in \{0.5, 1.0, 1.5\}$, was added to $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$ (the observed data corresponds to $\tau^* = 0$). We repeated the noise-adding process 200 times and saved the median coefficient estimates of $\boldsymbol{\beta}$. The first three rows of Figure 4 present these 91 coefficient estimates as kernel density plots over increasing levels of τ^*. The range of $\widehat{\boldsymbol{\beta}}_{\mathrm{RIDG}}$, excluding the intercept, is $(-0.019, 0.014)$. The range of $\widehat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ goes from $(-0.071, 0.063)$, when no noise is added, to $(-0.008, 0.010)$, when $\tau^* = 1.5$. Correspondingly, $\widehat{\boldsymbol{\beta}}_{\mathrm{FRC.CW}}$ is $(-0.456, 0.181)$ to $(-0.026, 0.035)$, and $\widehat{\boldsymbol{\beta}}_{\mathrm{HYB}}$ is $(-0.203, 0.077)$ to $(-0.023, 0.026)$.

The bottom row of Figure 4 gives the median MSPE, using the coefficient estimates to predict in the validation data. For RIDG, the MSPE is 0.62. Adding noise improved the MSPE for the other methods; the best was FRC.CW, with an MSPE going from 1.54 to 0.56, but FRC and HYB were very close. The weak signal in the data and the incremental improvements over RIDG point to a situation similar to Sim 7 of Figure 2. Plugging in $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$ yields an MSPE of 0.59, which RIDG does not beat and the candidate methods only slightly improve upon.

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 in this large p situation and can give a smaller MSPE. We have also proposed several adaptive methods which adjust the amount and direction of shrinkage. Considering all methods, HYB stands out as the method of choice. It makes efficient predictions under a range of plausible scenarios. Its performance is particularly notable under the common large n_{B}/small n_{A} scenario, ie when covariates from many observations are observed only with error.

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.

That this is essentially 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 $\boldsymbol{x}_{\mathrm{B}}$ as was done in Chen et al. (2011). Or, by writing out the complete likelihood, a data augmentation/Gibbs sampler algorithm can make alternating draws from the posterior distribution of $\boldsymbol{x}_{\mathrm{B}}$ and $\boldsymbol{\beta}$. However, because of the large fraction of missing data here, further research is needed into whether these methods would offer significant improvement without informative priors, ie shrinkage on $\boldsymbol{\beta}$.

The development of TR estimates 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) would be a particularly important case to consider, since designs such as these are an appealing way to undertake a study without needing to test all samples. It is usually noted that ODS can enhance efficiency but will introduce bias if the sampling mechanism is not properly accounted for. An important area of future research is thus how this additional bias and variance trade-off from ODS affects MSPE.

7 Appendix

The following lemma is used in the proof of THEOREM 3.1.

Lemma 7.1. *Suppose we combine two estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ ($m = 2$), and let ω^{opt} be given by (15). Then MSPE($\boldsymbol{b}(\omega^{opt})$) $\leq \min\{MSPE(\widehat{\boldsymbol{\beta}}_1), MSPE(\widehat{\boldsymbol{\beta}}_2)\}$; the inequality is strict*

if $MCPE(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2) \neq \min\{MSPE(\widehat{\boldsymbol{\beta}}_1), MSPE(\widehat{\boldsymbol{\beta}}_2)\}$.

Proof. Using the P_{k_1,k_2} notation from (16), upon substituting ω^{opt} from (15) into the $\text{MSPE}(\boldsymbol{b}(\omega^{\text{opt}}))$ expression from (14), it is seen that $\text{MSPE}(\boldsymbol{b}(\omega^{\text{opt}})) = \left(\frac{P_{22}-P_{12}}{P_{11}+P_{22}-2P_{12}}\right)^2 P_{11} + \left(\frac{P_{11}-P_{12}}{P_{11}+P_{22}-2P_{12}}\right)^2 P_{22} + 2\frac{(P_{11}-P_{12})(P_{22}-P_{12})}{(P_{11}+P_{22}-2P_{12})^2} P_{12} = \frac{P_{11}P_{22}-P_{12}^2}{P_{11}+P_{22}-2P_{12}}$ This implies that

$$\text{MSPE}(\boldsymbol{b}(\omega^{\text{opt}})) \leq \text{MSPE}(\widehat{\boldsymbol{\beta}}_1) \Leftrightarrow \frac{P_{11}P_{22}-P_{12}^2}{P_{11}+P_{22}-2P_{12}} \leq P_{11} \Leftrightarrow 0 \leq (P_{11}-P_{12})^2 \tag{18}$$

$\text{MSPE}(\boldsymbol{b}(\omega^{\text{opt}})) \leq \text{MSPE}(\widehat{\boldsymbol{\beta}}_2)$ by symmetry. (18) implies that the inequality in the lemma's statement is strict, ie $\text{MSPE}(\boldsymbol{b}(\omega^{\text{opt}})) < \min\{\text{MSPE}(\widehat{\boldsymbol{\beta}}_1), \text{MSPE}(\widehat{\boldsymbol{\beta}}_2)\}$, if and only if $P_{12} \neq \min\{P_{11}, P_{22}\}$ ($P_{12} \neq \max\{P_{11}, P_{22}\}$ by definition of a positive semi-definite matrix). □

Proof. (Theorem 3.1) First show that a given combination of m $\widehat{\boldsymbol{\beta}}$'s (subject to a sum-to-one constraint) is equivalent to a combination of two other $\widehat{\boldsymbol{\beta}}$'s (similarly constrained), and that if the length-m weight vector for the first combination is optimal, so is the length-2 vector for the second combination. Let $\boldsymbol{\omega}^{\text{opt}} = \{\omega_1^{\text{opt}}, \omega_2^{\text{opt}}, \ldots, \omega_m^{\text{opt}}\}^\top$. We have

$$\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}}) = [\widehat{\boldsymbol{\beta}}_1\, \widehat{\boldsymbol{\beta}}_2\, \cdots\, \widehat{\boldsymbol{\beta}}_m]\boldsymbol{\omega}^{\text{opt}} = \sum_{\ell=1}^{m-1} \omega_\ell^{\text{opt}} \widehat{\boldsymbol{\beta}}_\ell + \left(1 - \sum_{\ell=1}^{m-1} \omega_\ell^{\text{opt}}\right) \widehat{\boldsymbol{\beta}}_m = \widetilde{\phi}^{\text{opt}} \widehat{\boldsymbol{\beta}}^* + (1-\widetilde{\phi}^{\text{opt}}) \widehat{\boldsymbol{\beta}}_m,$$

where $\widetilde{\phi}^{\text{opt}} = \sum_{\ell=1}^{m-1} \omega_\ell^{\text{opt}}$ and $\widehat{\boldsymbol{\beta}}^* = (1/\widetilde{\phi}^{\text{opt}}) \sum_{\ell=1}^{m-1} \omega_\ell^{\text{opt}} \widehat{\boldsymbol{\beta}}_\ell$. Define $\boldsymbol{b}^*(\phi) := \phi \widehat{\boldsymbol{\beta}}^* + (1-\phi)\widehat{\boldsymbol{\beta}}_m$. Let ϕ^{opt} be the scalar which minimizes $\text{MSPE}(\boldsymbol{b}^*(\phi))$. We show $\text{MSPE}(\boldsymbol{b}^*(\phi^{\text{opt}})) = \text{MSPE}(\boldsymbol{b}^*(\widetilde{\phi}^{\text{opt}}))$. Observe that $\boldsymbol{b}^*(\phi^{\text{opt}}) = \boldsymbol{b}(\widetilde{\boldsymbol{\omega}}^{\text{opt}})$, where $\widetilde{\boldsymbol{\omega}}^{\text{opt}} = \{\widetilde{\omega}_1^{\text{opt}}, \widetilde{\omega}_2^{\text{opt}}, \ldots, \widetilde{\omega}_m^{\text{opt}}\}^\top$ and $\widetilde{\omega}_\ell^{\text{opt}} := (\phi^{\text{opt}}/\widetilde{\phi}^{\text{opt}})\omega_\ell^{\text{opt}}$ for $\ell = 1, \ldots, m-1$ and $\widetilde{\omega}_m^{\text{opt}} := 1 - \phi^{\text{opt}}$. Then,

$$\begin{aligned}
\text{MSPE}(\boldsymbol{b}^*(\phi^{\text{opt}})) \leq \text{MSPE}(\boldsymbol{b}^*(\widetilde{\phi}^{\text{opt}})) &= \text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \\
&\leq \text{MSPE}(\boldsymbol{b}(\widetilde{\boldsymbol{\omega}}^{\text{opt}})) \quad \text{(by assumption)} \\
&= \text{MSPE}(\boldsymbol{b}^*(\phi^{\text{opt}})).
\end{aligned}$$

Thus $\text{MSPE}(\boldsymbol{b}^*(\phi^{\text{opt}})) = \text{MSPE}(\boldsymbol{b}^*(\widetilde{\phi}^{\text{opt}})) = \text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}}))$. Combining this with Lemma 7.1 gives $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leq \text{MSPE}(\widehat{\boldsymbol{\beta}}_m)$. Because the ordering of $\widehat{\boldsymbol{\beta}}$'s is arbitrary, we can repeat the above steps to show that $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leq \text{MSPE}(\widehat{\boldsymbol{\beta}}_\ell)$ for $\ell = 1, \ldots, m-1$. □

References

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

Buzas, J., Stefanski, L., and Tosteson, T. (2005). Measurement error. In Ahrens, W. and Pigeot, I., editors, *Handbook of Epidemiology*, pages 729–765. Springer Berlin Heidelberg.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurment Error in Nonlinear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

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

Chen, G. et al. (2011). Development and validation of a qRT-PCR–classifier for lung cancer prognosis. *Journal of Thoracic Oncology* **6,** 1481–1487.

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

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

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

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

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

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

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

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

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

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

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage: the James-Stein and ridge*

regression estimators. Marcel Dekker, Inc., New York.

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

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

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

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

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

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

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

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

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

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

Strawderman, W. E. (1978). Minimax adaptive generalized ridge regression estimators. *Journal of the American Statistical Association* **73,** 623–627.

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

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

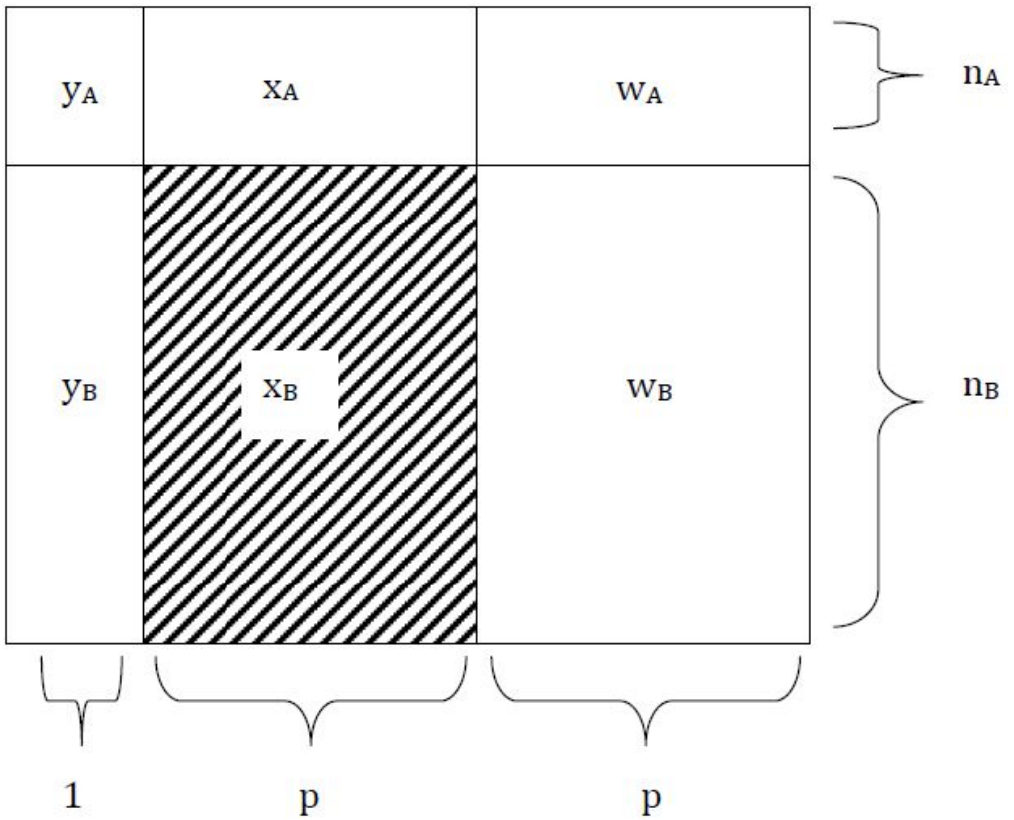

Figure 1: Schematic representation of the prediction problem: $(\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}})$ is measured on n_{A} subjects and $(\boldsymbol{y}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}})$ is measured on n_{B} subjects. $\boldsymbol{x}_{\mathrm{B}}$ 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 $\boldsymbol{X}$

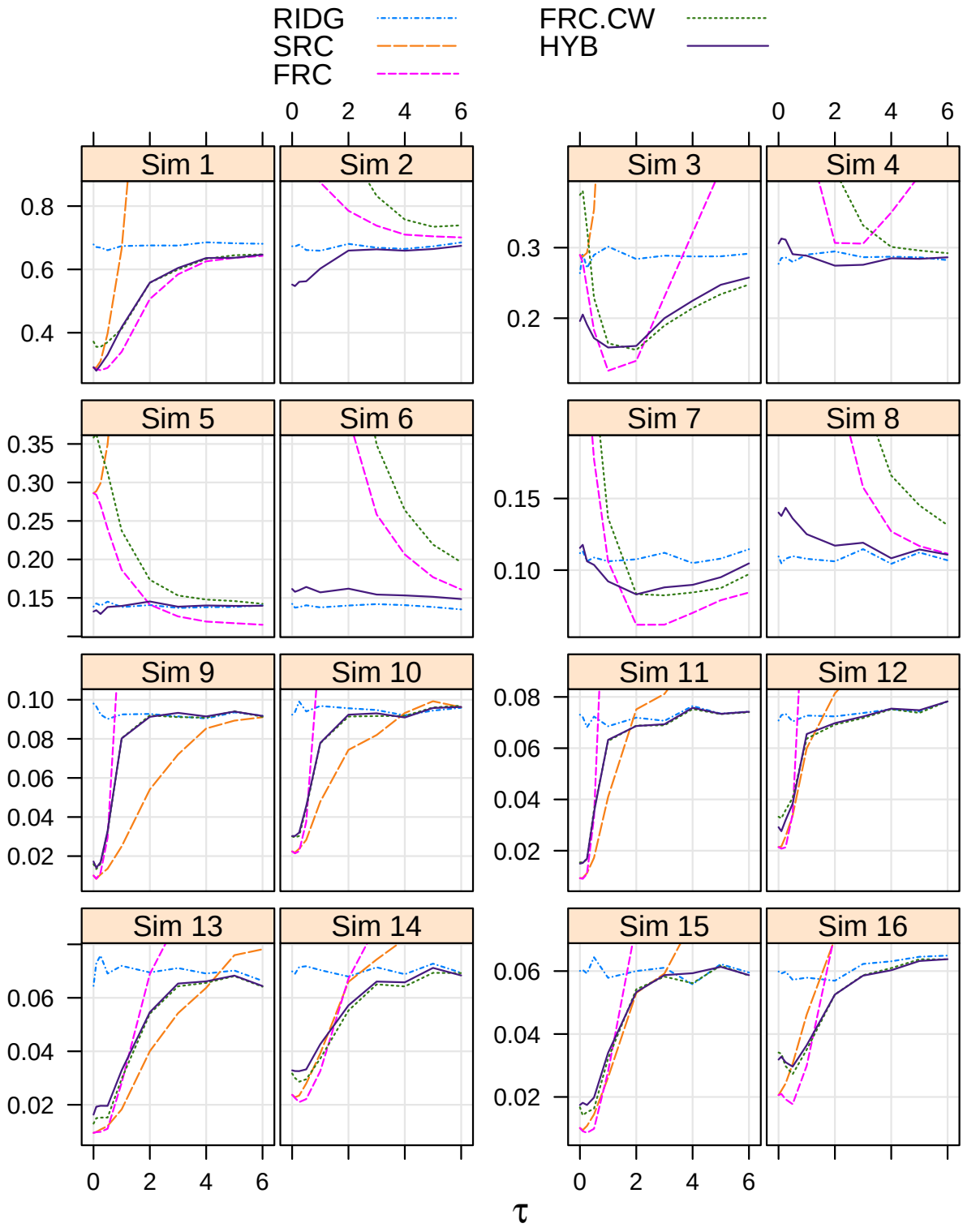

Figure 2: rMSPE for the simulation settings in Table 2.

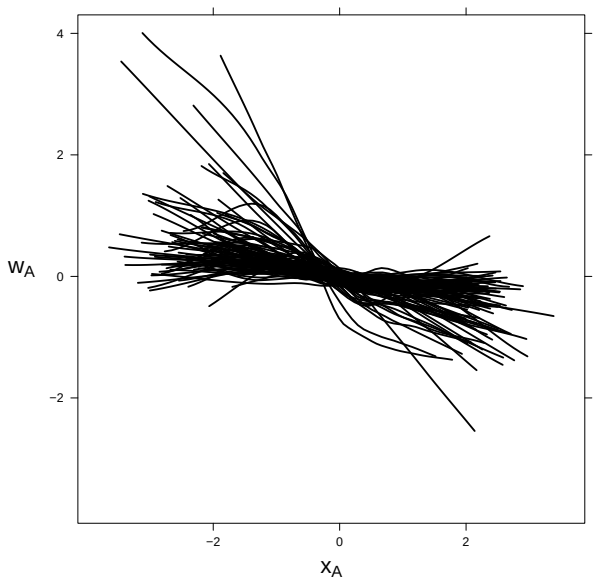

Figure 3: LOESS curves of Affymetrix ($\boldsymbol{w}_{\mathrm{A}}$) by qRT-PCR ($\boldsymbol{x}_{\mathrm{A}}$) measurements for 91 genes from the Chen et al. (2011) data

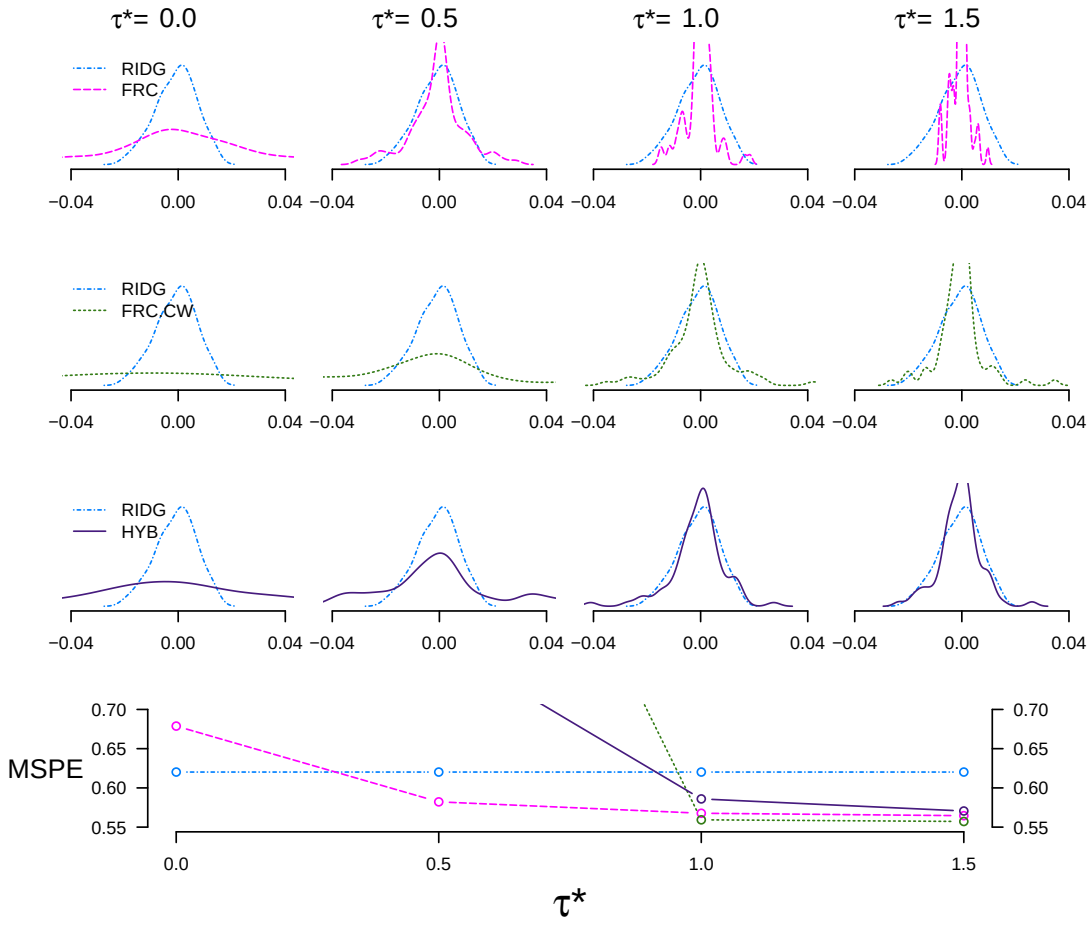

Figure 4: Kernel density estimate of 91 coefficient estimates (top rows) and mean squared prediction error (MSPE - bottom row) from the NSCLC validation data for FRC, FRC.CW and HYB, with differing levels of noise (normal with mean 0 and standard deviation τ^*) added to $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$. Corresponding results from RIDG, which is invariant to τ^*, are given in each panel for reference.

Table 1: Key information for several TR estimators, conditioning on the true value of $\boldsymbol{\theta}$. $\kappa = (\tau^2/\nu^2)\boldsymbol{\beta}^\top \boldsymbol{V}\boldsymbol{\beta}$. $\boldsymbol{V} = (\boldsymbol{I}_p + (\tau^2/\nu^2)\boldsymbol{\Sigma_X}^{-1})$. The '$\lambda = 1$?' column indicates whether λ is fixed at 1 or tuned in a data-adaptive fashion using the general GCV function (13). The corresponding estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma_\beta}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ is given by plugging $(\boldsymbol{\gamma_\beta}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ into (6). The expectation and variance of $\boldsymbol{\gamma_\beta}$, which are useful for calculating the MSPE of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\gamma_\beta}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$, 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	$\boldsymbol{\gamma_\beta}$	$\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$	$\lambda = 1$?
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
FRC.CW	$\nu(\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{w}_\mathrm{B})^{-1}\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{y}_\mathrm{B}$	$\nu^{-2}\mathrm{diag}(\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{w}_\mathrm{B})$	N
Method	$\mathrm{E}\,\boldsymbol{\gamma_\beta}$	$\mathrm{Var}\,\boldsymbol{\gamma_\beta}$	
RIDG	–	–	
FRC	$\boldsymbol{V}\boldsymbol{\beta}$	$(\sigma^2 + \kappa)\nu^2(\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{w}_\mathrm{B})^{-1}$	
SRC	$\boldsymbol{\beta}$	$(\sigma^2 + \kappa)\nu^2 \boldsymbol{V}^{-1}(\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{w}_\mathrm{B})^{-1}\boldsymbol{V}^{-1}$	
FRC.CW	$\boldsymbol{V}\boldsymbol{\beta}$	$(\sigma^2 + \kappa)\nu^2(\boldsymbol{w}_\mathrm{B}^\top \boldsymbol{w}_\mathrm{B})^{-1}$	

Table 2: A legend of simulation settings to match the results in Figure 2.

		ρ	0.0	0.0	0.75	0.75
$\boldsymbol{\beta}$	R^2	n_B	400	150	400	150
$\{\frac{j}{100}\}_{j=-49}^{j=49}$	0.4		Sim 1	Sim 2	Sim 3	Sim 4
$\{\frac{j}{100}\}_{j=-49}^{j=49}$	0.1		Sim 5	Sim 6	Sim 7	Sim 8
$\{\frac{j}{4}\}_{j=-2}^{j=2}$	0.4		Sim 9	Sim 10	Sim 11	Sim 12
$\{\frac{j}{4}\}_{j=-2}^{j=2}$	0.1		Sim 13	Sim 14	Sim 15	Sim 16