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, USA
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 $X$, 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 $W$, coming from a prior technology but correlated with $X$. On a moderately sized sample, we have $(Y, X, W)$, and on a larger sample we have $(Y, W)$. We utilize the data on $W$ to boost the prediction of $Y$ by $X$. When $p$ is large and the subsample containing $X$ 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 $W$ for use in future studies. We propose to shrink the regression coefficients $\beta$ of $Y$ on $X$ toward different targets that use information derived from $W$ in the larger dataset. We compare these proposals with the classical ridge regression of $Y$ on $X$, which does not use $W$. 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 $\beta$. With an optimal choice of weights, the hybrid estimator balances efficiency and robustness in a data-adaptive way to theoretically yield a smaller prediction error than any of its constituents. The methods, including a fully Bayesian alternative, are evaluated via simulation studies. We also apply them to a gene-expression dataset. mRNA expression measured via quantitative real-time polymerase chain reaction is used to predict survival time in lung cancer patients, with auxiliary information from microarray technology available on a larger sample.

Keywords: 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 $Y|X$, where $Y$ is the
Fig. 1. The schematic representation of the prediction problem: \((y_A, x_A, w_A)\) constitutes subsample A, of size \(n_A\), and \((y_B, w_B)\) constitutes subsample B, of size \(n_B\). The covariates represented by \(x_B\) are considered missing. The quantity \(W\) is an error-prone noisy version of \(X\). The goal is to utilize the data on \(W\) to boost the prediction of \(Y\) by \(X\).

outcome and \(X\) is the \(p\)-dimensional vector of biomarkers. One such model is a linear regression:

\[
y = \beta_0 + X^\top \beta + \sigma \varepsilon, \quad \varepsilon \sim N(0, 1).
\]

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

\[
W = \psi 1_p + vX + \tau \xi, \quad \xi \sim N_p(0, I_p).
\]

Here \(I_p\) is the identity matrix and \(\psi, v, \) and \(\tau\) 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 \(p > n_A\). Additionally, \(n_B\) observations of \(Y\) and \(W\) are available. Assume \(p < n_B\). Denote subsamples A and B (each assumed to be from the same population) by \((y_A, x_A, w_A)\) and \((y_B, w_B)\), respectively. Using this notation, \(x_B\), the set of \(X\)'s from subsample B, is missing. Figure 1 gives a schematic representation. \(x_A\) is also standardized, i.e. if \(x_{ij}\) is from the \(i\)th row and \(j\)th column, \(\sum_{i=1}^{n_A} x_{ij} = 0\) and \(\sum_{i=1}^{n_A} x_{ij}^2 = n_A, j = 1, \ldots, p\).

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

\[
\text{MSPE}(\hat{\beta}) = E[(Y_{\text{new}} - X_{\text{new}}^\top \hat{\beta})^2] = \sigma^2 + E[(\beta - \hat{\beta})^\top X_{\text{new}} X_{\text{new}}^\top (\beta - \hat{\beta})]
\]

\[
= \sigma^2 + \text{Tr}([\text{Bias} \hat{\beta} \text{ Bias} \hat{\beta}^\top + \text{Var} \hat{\beta}] E[X_{\text{new}} X_{\text{new}}^\top]),
\]

where \(\text{Tr}\) indicates the trace operator, and the expectation is over \(Y_{\text{new}}, X_{\text{new}}, y_A, x_A, w_A, w_B\). We consider two questions: (i) How can the auxiliary information in subsample B be used in the prediction of \(Y|X\)? (ii) When does using such information lead to an improved MSPE?

A simple approach, which ignores subsample B, is ordinary least squares of \(y_A\) on \(x_A\), i.e. \(\hat{\beta}_{\text{OLS}} = \arg\min_\beta (y_A - x_A \beta)^\top (y_A - x_A \beta) = (x_A^\top x_A)^{-1} x_A^\top y_A\). However, \((x_A^\top x_A)^{-1}\) does not exist for \(p > n_A\). Even for \(p \leq n_A\), multicollinearity of the covariates may lead to variance inflation and numerical instability. Ridge regression (RIDGE) (Hoerl and Kennard, 1970) can ameliorate these issues by shrinking coefficients.
Improving prediction with shrinkage

261
toward zero, i.e. \( \hat{\beta}_{\text{RIDGE}} = \arg\min_{\beta} (y_A - x_A \beta)^\top (y_A - x_A \beta) + \lambda \beta^\top \beta = (x_A^\top x_A + \lambda I_p)^{-1} x_A^\top y_A. \) This can be viewed from a Bayesian perspective: given a normal prior on \( \beta \) with mean \( 0_p \) and precision \( \sigma^{-2} \lambda I_p \), the RIDGE coefficients are the posterior mode for a given \( \lambda \). Hoerl and Kennard showed that there exists \( \lambda > 0 \) which decreases the mean squared error of \( \hat{\beta} \), \( \text{MSE}(\hat{\beta}) = \mathbb{E}[(\hat{\beta} - \beta)^\top (\hat{\beta} - \beta)] \), compared with \( \lambda = 0 \). RIDGE penalizes the \( L_2 \) norm; other methods exist which constrain the \( L_d \) norm for some \( d \) (e.g. 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 RIDGE. Draper and van Nostrand (1979) are critical of RIDGE because of difficulties in choosing the parameter \( \lambda \). However, Craven and Wahba (1979) and Li (1986) demonstrate the asymptotic optimality of the generalized cross-validation (GCV) function in selecting \( \lambda \). Simulation studies (Gelfand, 1986; Frank and Friedman, 1993) demonstrate good prediction properties of RIDGE for many choices of \( \beta \). Rao (1975) generalizes RIDGE to allow for different levels of shrinkage between each coefficient. Swindel (1976) proposes ridge estimators that 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. Selove (1968) adapts the shrinkage estimator of James and Stein (1961) (JS) which, for \( p > 3 \), uniformly beats the maximum likelihood estimate (MLE) of \( \beta \) 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 that 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 \( p < n_A \), the problem closely resembles that of measurement error (ME) in the covariates, \( W \) being an error-prone version of \( X \). Fuller (1987) and Carroll and others (2006) review ME methods for unbiased and efficient inference on \( \beta \). In linear regression, using \( W \) instead of \( X \) gives biased estimates of \( \beta \). However, this substitution is typically not problematic for predicting \( Y_{\text{new}} \) with \( W_{\text{new}}^\top \hat{\beta} \). Our prediction model of interest being \( Y \) given \( X \), this bias in \( \hat{\beta} \) from using \( W \) instead of \( X \) does bias \( X_{\text{new}}^\top \hat{\beta} \) away from \( Y_{\text{new}} \). Regression calibration, which fills in each missing \( X \) with its conditional expectation given \( W \), may provide unbiased estimates of \( \beta \) and therefore \( Y_{\text{new}} \). In contrast, although the substitution of \( X \) by \( W \) gives biased estimates of \( \beta \), it may reduce the variance of estimates of \( \beta \) relative to regression calibration (Buzas and others, 2005) and consequently reduce MSPE. Even for \( p < n_A \), then, it is not evident that the regression calibration algorithm is best for making predictions with \( X_{\text{new}}^\top \hat{\beta} \).

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 (TR) estimators, borrowing ideas from the shrinkage and regression calibration literature. We also consider combining an ensemble of TR 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 \( \beta \) for better prediction than any individual estimator. This result applies to a linear combination of any set of estimates of \( \beta \). We evaluate all of these estimators via simulation studies and a data analysis.

The rest of the paper is organized as follows. In Section 2, we unify RIDGE and regression calibration methods under a class of TR estimators. In Section 3, we propose hybrid estimators that achieve superior prediction by data-adaptively combining multiple estimators. Section 4 gives a fully Bayesian alternative and Section 5 presents a simulation study. Section 6 applies the methods: survival time (\( Y \)) in lung cancer
patients is predicted with quantitative real-time polymerase chain reaction (qRT-PCR) data ($X$), aided by microarray data ($W$) from a larger sample. Section 7 concludes with a discussion. Analytical details are in supplementary material available at Biostatistics online.

2. Targeted shrinkage

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

$$
\hat{\beta}_{\text{RIDG}} = (x_A^\top x_A + \lambda I_p)^{-1} x_A^\top y_A.
$$

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

For ridge regression, Craven and Wahba (1979) proposed to select $\lambda$ using the GCV function, choosing the $\lambda$ that minimizes

$$
(1/n_A)(y_A - H(\lambda I_p)y_A)^\top (y_A - H(\lambda I_p)y_A),
$$

where $H(\lambda I_p) = x_A(x_A^\top x_A + \lambda I_p)^{-1} x_A^\top$.

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 TR estimator, referring to shrinkage toward a target $\gamma_\beta$. The estimator in (2.4) gives the three terms $\{\gamma_\beta, \lambda, \Omega_\beta^{-1}\}$ which determine the general class of TR estimators. As we shall see, different estimators, we propose either implicitly or explicitly specify values for $\{\gamma_\beta, \lambda, \Omega_\beta^{-1}\}$. In particular, RIDG is a TR estimator: $\hat{\beta}^\text{RIDG} = \hat{\beta}(0_p, \lambda, I_p)$.

As stated in (1.3), MSPE($\hat{\beta}$) = $\sigma^2 + \text{Tr}[\text{Bias} \hat{\beta} \hat{\beta}^\top + \text{Var} \hat{\beta}]E[X_{\text{new}}^\top X_{\text{new}}^\top]$. Thus, we calculate the MSPE of a TR estimator from its bias and variance, taking expectations over the response distribution $y_A, y_B | x_A, w_A, w_B$:

$$
\text{Bias} \hat{\beta} = E \hat{\beta} - \beta = (x_A^\top x_A + \lambda \Omega_\beta^{-1})^{-1} (x_A^\top x_A \beta + \lambda \Omega_\beta^{-1} \gamma_\beta - x_A^\top x_A \beta - \lambda \Omega_\beta^{-1} \beta),
$$

$$
\text{Var} \hat{\beta} = (x_A^\top x_A + \lambda \Omega_\beta^{-1})^{-1} (\sigma^2 x_A^\top x_A + \lambda^2 \Omega_\beta^{-1} \gamma_\beta \Omega_\beta^{-1} \gamma_\beta) (x_A^\top x_A + \lambda \Omega_\beta^{-1})^{-1}.
$$

These expressions assume that $\lambda$ and $\Omega_\beta^{-1}$ are fixed with respect to $y_A, y_B | x_A, w_A, w_B$ but allow $\gamma_\beta$ to be data-dependent. A TR estimator may use a true prior, as in RIDG, in which case $\gamma_\beta$ is fixed.
We now propose several other TR estimators. If $x_B$ were observed, logical selections of $\gamma_\beta$ and $\Omega_\beta^{-1}$ would be $(x_B^T x_B)^{-1} x_B^T y_B$ and $x_B^T x_B$, respectively, with $\lambda = 1$, giving the estimator $(x_A^T x_A + x_B^T x_B)^{-1} (x_A^T y_A + x_B^T y_B)$. In the absence of $x_B$, the naïve inclination is to regress $y_B$ on $w_B$ and use $(w_B^T w_B)^{-1} w_B^T y_B$ and $w_B^T w_B$ as $\gamma_\beta$ and $\Omega_\beta^{-1}$, that is, use $w_B$ itself as an imputation for $x_B$. We first consider approaches that derive a replacement for the missing $x_B$ which may be better than $w_B$. This is obtained by modeling $W|X$ based on the relationship observed in subsample $A$ and thereby inducing data-driven values of $\gamma_\beta$ and $\Omega_\beta^{-1}$. From the ME perspective, this is regression calibration. These TR estimators fix $\lambda = 1$ (data-adaptive estimation of $\lambda$ may be done using, e.g. a GCV criterion).

**Structural regression calibration (src):** A distribution on $X$ and the ME model for $W|X$ imply a value of $E[X|W]$. src fills in the missing $x_B$ with its conditional expectation given $w_B$. Assuming that $X$ is normal, say $N_p(\mu_X, \Sigma_X)$, implies that $X|W$ is also normal. Let $\theta = \{v, \tau, \mu_X, \Sigma_X^{-1}\}$. From properties of the conditional distribution of $X|W$,

$$x_B^{SRC}(\theta) = E[x_B|w_B, \theta] = 1_{n_B} \mu_X^T (I_p - V(\theta)) + \frac{1}{v} w_B V(\theta) = [1_{n_B}, w_B] M(\theta),$$

$$M(\theta) = \left(\frac{\mu_X^T (I_p - V(\theta))}{\frac{\tau^2}{\nu} \Sigma_X^{-1}}\right)^{-1}. \quad (2.7)$$

We suppress the dependence on $\theta$ of $x_B^{SRC}(\theta)$, $M(\theta)$, and $V(\theta)$ hereafter. This is a precision-weighted average of $1_{n_B} \mu_X^T$ and $(1/v) w_B$. Using (2.4), define $\hat{\beta}_{SRC} = \hat{\beta}(\gamma_{\beta,ac}^{-1}, 1, \Omega_{\beta,ac}^{-1})$, with $\gamma_{\beta,ac} = (x_B^{SRC^T} x_B^{SRC})^{-1} (x_B^{SRC^T} y_B)$ and $\Omega_{\beta,ac}^{-1} = x_B^{SRC^T} x_B^{SRC}$. 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 $X$, to distinguish from its “Functional” alternative, which does not assume this, proposed as follows.

**Functional regression calibration (frc):** Solving (1.2), $W = v X + \tau \xi$, for $X$ gives $X = (1/v) W - (\tau/v) \xi$. Another natural estimate of $x_B$, and consequently a corresponding $\gamma_\beta$ and $\Omega_\beta^{-1}$, is therefore

$$x_B^{FRC}(\theta) = (1/v) w_B, \quad \gamma_\beta^{FRC} = (x_B^{FRC^T} x_B^{FRC})^{-1} x_B^{FRC^T} y_B, \quad \Omega_\beta^{-1}^{FRC} = x_B^{FRC^T} x_B^{FRC}. \quad (2.9)$$

This gives a TR estimate defined as $\hat{\beta}_{FRC} = \hat{\beta}(\gamma_{\beta,ac}^{-1}, 1, \Omega_{\beta,ac}^{-1})$. This imputation for $x_B$ is a scaled version of a substitution of $w_B$ for $x_B$, to which frc is equivalent when $v = 1$, i.e. under the classical ME model. In supplementary material available at Biostatistics online (Appendix A), we conduct extensive analyses which suggest that frc is preferred over src in terms of MSPE as any of $\beta^T \beta$, $\sigma^2$, or $\tau/v$ increase.

The first rows of Table 1 summarize choices of $(\gamma_\beta, \lambda, \Omega_\beta^{-1})$ for ridg, src, and frc. Assuming that the ME is non-differential, i.e. $E[X|W] = E[X|X]$, and $\mu_X = 0_p$, Table 1 also gives $E\gamma_\beta$ and $\text{Var} \gamma_\beta$ for frc and src. Because $E\gamma_{\beta,ac} = \beta$, from (2.5), src provides unbiased estimates of $\beta$.

**Remark 2.1** One of the reviewers observed that, when $\gamma_\beta$ and $\Omega_\beta^{-1}$ are based on historical data, the prior in the second expression of (2.3) is a power prior (Chen and Ibrahim, 2000), with $\lambda$ controlling the contribution of the historical data to the posterior.

**Remark 2.2** These approaches require estimating $\theta = \{v, \tau, \mu_X, \Sigma_X^{-1}\}$. One can regress $\{w_{ij}\}$ on $\{x_{ij}\}$ for $i = 1, \ldots, n_A$ and $j = 1, \ldots, p$, to compute MLEs for $v$ and $\tau$. If it is required that $v$ and $\tau$ be of a more general form than scalar-valued, the estimation procedure can be modified accordingly. The MLE for $\mu_X$ is $\hat{\mu}_X = n_A^{-1} x_A^T 1_{n_A}$, which will be $0_p$ if $x_A$ is standardized. For $p > n_A$, the required inversion of $\hat{\Sigma}_X = n_A^{-1} x_A^T x_A$ is not possible. An alternative is the shrinkage estimator from Schäfer and Strimmer (2005):
Table 1. Key information for several TR estimators, conditioning on the true value of $\theta$

<table>
<thead>
<tr>
<th>Method</th>
<th>$\gamma_\beta$</th>
<th>$\Omega_\beta^{-1}$</th>
<th>$\lambda = 1?$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIDG</td>
<td>$\theta_p$</td>
<td>$I_p$</td>
<td>N</td>
</tr>
<tr>
<td>FRC</td>
<td>$\nu(w_B^T w_B)^{-\frac{1}{2}}w_B V F_B$</td>
<td>$\nu^{-2}w_B^T w_B V$</td>
<td>Y</td>
</tr>
<tr>
<td>SRC</td>
<td>$\nu V^{-1}(w_B^T w_B)^{-\frac{1}{2}}w_B F_B$</td>
<td>$\nu^{-2}w_B^T w_B V$</td>
<td>Y</td>
</tr>
</tbody>
</table>

$k = (\tau^2/\nu^2)\beta^T V \beta$, $V = (I_p + (\tau^2/\nu^2)\Sigma^{-1}_X)$. The “$\lambda = 1?$” column indicates whether $\lambda$ is fixed at 1 or tuned in a data-adaptive fashion using the general GCV function. The corresponding estimator $\hat{\beta}(\gamma_\beta, \lambda, \Omega_\beta^{-1})$ is given by plugging $(\gamma_\beta, \lambda, \Omega_\beta^{-1})$ into (2.4). The expectation and variance of $\gamma_\beta$, which are useful for calculating the MSPE of $\hat{\beta}(\gamma_\beta, \lambda, \Omega_\beta^{-1})$, are over $y_A, y_B|x_A, w_A, w_B$ under the assumption $\{Y|X, W\} = \{Y|X\}$.

since $x_A^T x_A$ is standardized, it is simply $\hat{\Sigma}_X^* = (1 - \eta)\hat{\Sigma}_X + \eta I_p$, for $\eta \in [0, 1]$ chosen data-adaptively. We used the R package corpcor to choose $\eta$ targeting a minimum MSE for $\hat{\Sigma}_X^*$.

Remark 2.3 The bias and variance outlined in Table 1 condition on the true value of $\theta$ and are over and above any bias and variance coming from its estimation. In particular, estimating $\Sigma_X$ 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 $\beta$: a TR estimator where $\Omega_\beta^{-1}$ is diagonal and $\lambda$ is estimated. When $\lambda$ is not fixed, the GCV approach may be used to choose an appropriate value of $\lambda$. The form of this modified GCV criterion is given later on in (3.2), in connection with the hybrid estimator.

3. Hybrid estimators

While a particular TR estimator may do well for a given set of factors, e.g. $p, n_B, \beta, \tau$, none is likely to give a 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, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m$, and a vector $\omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$ such that $1_m^T \omega = 1$, let $b(\omega) = \sum_{i=1}^m \omega_i \hat{\beta}_i = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m] \omega$. The vector $\omega$ determines the contribution from each $\hat{\beta}_i$; a sensible choice for $\omega$ in our situation would be the one that minimizes MSPE$(b(\omega))$. The following theorem compares the prediction error of the resulting optimal hybrid estimator, $b(\omega^{opt})$, to that of its constituents. The result uses the following definition of the “mean cross-product prediction error” between $\hat{\beta}_i$ and $\hat{\beta}_j$:

$$\text{MCPE}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 + E[(\beta - \hat{\beta}_i)^T X_{\text{new}} X_{\text{new}}^T (\beta - \hat{\beta}_j)].$$

Theorem 3.1 Let $b(\omega) = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m] \omega$ be a hybrid estimator. (i) If $\text{Var}((\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m) \nu]$ has at least one positive eigenvalue for every $\nu \in \mathbb{R}^m \setminus 0_m$, then there exists a unique vector $\omega^{opt}$ which minimizes $\text{MSPE}(b(\omega))$ subject to $1_m^T \omega = 1$. (ii) Further, let $\text{MSPE}(\hat{\beta}_j) = \min_i \text{MSPE}(\hat{\beta}_i)$. If $\text{MCPE}(\hat{\beta}_j, \hat{\beta}_i) \neq \text{MSPE}(\hat{\beta}_j)$ for some $i \neq j$, then $\text{MSPE}(b(\omega^{opt})) < \text{MSPE}(\hat{\beta}_j)$. 
The proof is in supplementary material available at *Biostatistics* online (Appendix B). If the assumptions are satisfied, then, using prediction error as the criterion, \( b(\omega^{\text{opt}}) \) 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\((b(\omega)) = \omega^T P \omega\), where \( P \) is the \( m \times m \) matrix with the \((i,j)\)th element given by \( P_{ij} = \text{MCPE}(\hat{\beta}_i, \hat{\beta}_j) \), which is just MSPE\((\hat{\beta}_i) \) when \( i = j \). The results from Theorem 3.1 apply when \( P \) is known. In practice, however, \( P \) and therefore \( \omega^{\text{opt}} \) must be estimated. Since \( P_{ij} \) is equivalently expressed as \( \text{E}[((Y_{\text{new}} - X_{\text{new}}^T \hat{\beta}_i)(Y_{\text{new}} - X_{\text{new}}^T \hat{\beta}_j)] \), one might use \((1/n_A)(y_A - x_A \hat{\beta}_i)^T(y_A - x_A \hat{\beta}_j)\) as an estimate, but this will be biased. A generalization of a result from Mallows (1973) in supplementary material available at *Biostatistics* online (Lemma B.3) gives that, on average, this underestimates \( P_{ij} \) by the amount \( \sigma^2(\psi_i + \psi_j) \), where \( \psi_i = \text{Tr} H(\lambda_i, \Omega_{\beta_i}^{-1})/n_A \). 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 adapting the GCV approach:

\[
\hat{P}_{ij} = \frac{(1/n_A)(y_{A,i}^e - H(\lambda_i, \Omega_{\beta_i}^{-1})y_{A,i}^e)^T(y_{A,j}^e - H(\lambda_j, \Omega_{\beta_j}^{-1})y_{A,j}^e)}{(1 - \psi_i)(1 - \psi_j)}, \tag{3.2}
\]

where \( y_{A,i}^e = y_A - x_A \gamma_{\beta_i} \). Because \( y_{A,i}^e - H(\lambda_i, \Omega_{\beta_i}^{-1})y_{A,i}^e = y_A - x_A \hat{\beta}_i \), this is a penalized version of its naive counterpart. Lemma B.4 (see supplementary material available at *Biostatistics* online) provides further justification for this approach.

Note the dual use of the GCV function to calculate \( b(\omega) \). First, for each \( \ell, \lambda_{\ell} \) is chosen, when required, to minimize \( \hat{P}_{\ell \ell} \). Then, fixing these choices of \( \lambda_{\ell} \), (3.2) is employed on the \( m(m + 1)/2 \) pairwise combinations of components in \( b(\omega) \) to estimate \( P \). The particular hybrid estimator we evaluate has three components: \( \hat{\beta}_{\text{HYB}} = [\hat{\beta}_{\text{RIDG}}, \hat{\beta}_{\text{SRC}}, \hat{\beta}_{\text{FRC}}]^T \omega^{\text{opt}} \). Following LeBlanc and Tibshirani (1996), in addition to the constraint \( \mu^{\dagger}_\omega = 1 \), we enforce a non-negativity constraint on \( \omega \), which improves numerical results.

**Remark 3.2** The key aspect that makes \( \hat{\beta}_{\text{HYB}} \) practical is that the sum \( \sigma^2 + \text{E}((\beta - \hat{\beta}_{\text{HYB}})^T X_{\text{new}} X_{\text{new}}^T (\beta - \hat{\beta}_{\text{HYB}})) \) 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 \( b(\omega) \). For their method, the estimation of \( \omega^{\text{opt}} \) requires an unbiased \( \hat{\beta}_1 \) and independent estimators \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \). In our case, because MSPE, 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 \( \omega^{\text{opt}} \), for a given dataset with estimated optimal weights \( \hat{\omega}^{\text{opt}} \), this uniform dominance may not hold. Numerical performance depends on how accurately (3.2) estimates \( P \). As will be seen, \( \hat{\beta}_{\text{HYB}} \) with estimated weights still performs well across a spectrum of scenarios and closely adapts to the best of its constituents.

### 4. Bayesian ridge

As a comparison to our proposed TR methods, we consider a fully Bayesian ridge regression which iteratively samples \( x_B \) and all model parameters from their full conditional distributions. We now briefly describe the construction of this approach; most details are in supplementary material available at *Biostatistics* online (Appendix C). We assume the models given in (1.1) and (1.2), along with \( X \sim N_p(\mu_X, \Sigma_X) \). Let \( \phi = \{\beta, \beta_0, \sigma^2, \psi, \nu, \tau^2, \mu_X, \Sigma_X^{-1}\} \) denote the set of all parameters. The likelihood of a complete observation may be factorized as

\[
p(Y, X, W|\phi) = p(Y|X, \phi) p(W|X, \phi) p(X|\phi) \quad (X \text{ observed}). \tag{4.1}
\]
Given the observed data, \( \{y_A, x_A, w_A, r_B, w_B\} \), and \( \phi \), each row of \( x_B \) is independently multivariate Normal; the data augmentation step samples \( x_B \) from this distribution to complete the likelihood. With respect to prior specification, we assume that \( \beta \) is \( N_p(0_p, (\sigma^2/\lambda)I_p) \) (making this a Bayesian ridge) and \( \Sigma^{-1}_X \) is Wishart with \( 3p \) degrees of freedom and scale \( (2p - 1)^{-1}[\text{diag}(\text{Var}[x_A])]^{-1} \), where diag(\text{Var}[x_A]) is the diagonal part of the empirical covariance of \( x_A \). From our numerical studies, this Wishart prior on \( \Sigma^{-1}_X \) ensures the convergence of the algorithm in spite of the large fraction of missing data, represented by \( x_B \). For the remaining components, Jeffrey’s priors were used, that is, \( \beta_0, \ln(\sigma^2), \psi, \nu, \ln(\tau^2), \mu_X \), and \( \ln(\lambda) \), each have flat priors. In summary,

\[
\pi(\phi, \lambda) \propto (\sigma^2 \tau^2 \lambda)^{-1} \left( \frac{\lambda}{\sigma^2} \right)^{p/2} |\Sigma^{-1}_X|^{(2p-1)/2} \exp \left\{ -\frac{1}{2} \frac{\lambda}{\sigma^2} \beta^\top \beta - \frac{2p - 1}{2} \text{Tr}(\text{diag}(\text{Var}[x_A])\Sigma^{-1}_X) \right\}.
\]

(4.2)

The Gibbs steps (Appendix C of supplementary material available at Biostatistics online) sample \( \phi \) from its full conditional distribution derived from the product of the complete data likelihood and the prior. After a burn-in of 2000 iterations, we stored 1000 posterior draws of \( \phi \). For the sake of comparison to the other methods, which yield a point estimate of \( \beta \), we used the posterior mean of \( \beta \), denoted as \( \hat{\beta}_\text{HIERBETAS} \), for predicting future observations.

5. Simulation study

We next describe a small simulation study. We fixed \( n_A = 50 \) and used \( n_B \in \{400, 150\} \). The diagonal elements of \( \Sigma_X \) were set to unity, and the off-diagonals were \( \rho^{\mid i - j\mid}, \rho \in \{0, 0.75\} \). Using these parameters, \( x_A \) and \( x_B \) were drawn from \( N_p(0_p, \Sigma_X) \). We considered both high- \( (p = 99) \) and low \( (p = 5) \)-dimensional models: \( \beta \in \{\{j/100\}_{j=49}, \{j/4\}_{j=2}\} \). The coefficient of determination, \( R^2 \), was either 0.1 or 0.4. Thus, given \( \beta, \Sigma_X \) and \( R^2, \sigma \) was determined by solving \( \beta^\top \Sigma_X \beta / (\beta^\top \Sigma_X \beta + \sigma^2) = R^2 \). The intercept \( \beta_0 \) was set to zero; \( y_A|x_A \) and \( y_B|x_B \) were drawn for each combination of \( \beta \) and \( \sigma \) from (1.1). This yielded 16 unique simulation settings: two choices each for \( p, n_B, \rho, \) and \( R^2 \). To draw the auxiliary data, we set \( \psi = 0 \) and \( \nu = 1 \) and repeated each of the 16 settings using \( \tau \in (0, 2) \), drawing \( w_A|x_A \) and \( w_B|x_B \) from (1.2).

For \( \text{RIDG}, \text{SRC}, \text{FRC}, \text{HYB}, \) and \( \text{HIERBETAS} \), we estimated MSPE by averaging the squared prediction error over 1000 new replicates over \( \tau \). Figure 2 plots this empirical MSPE averaged over 1000 replicates over \( \tau \). For reference, \( \sigma^2 \), the smallest achievable MSPE, is also given. Tables S2 and S3 in supplementary material available at Biostatistics online provide numeric values of the empirical MSPE over all settings. Note that, in practice, the analyst estimates \( \beta_0 \) in addition to \( \beta \). Following the common prescription for ridge regression, we did not shrink \( \beta_0 \) but instead used a flat prior in each of the TR methods and \( \text{HIERBETAS} \).

**Effect of \( \tau \):** \( \text{RIDG} \) is not affected by \( \tau \), as it does not use \( w_A \) or \( w_B \). \( \text{FRC} \) and \( \text{SRC} \) are equivalent when \( \tau \) is very small, close to the complete data case. The MSPE of \( \text{SRC} \) always rises with \( \tau \); this increase is sharp when \( p = 99 \). However, larger values of \( \tau \) give favorable shrinkage in \( \text{FRC} \). When \( p = 99 \), the \( \tau \) for which \( \text{FRC} \) is best is larger than zero; for \( p = 5 \), the “optimal” \( \tau \) is quite small, and the MSPE rises sharply with \( \tau \). For \( p = 99 \), the MSPE of \( \text{HYB} \) and \( \text{HIERBETAS} \) is mostly invariant to \( \tau \), except for when \( \rho = 0 \) and \( R^2 = 0.4 \). When \( p = 5 \), \( \text{HYB} \) does a better job of improving upon its constituents when \( \tau \) is large.

**Effect of \( n_B, p, \rho, R^2 \):** As is expected, larger values of \( n_B \) decrease the MSPE for all methods except \( \text{RIDG} \). Notably, \( \text{HYB} \) sometimes fares poorly compared with \( \text{FRC} \) (see Remark 3.3) when \( p = 99 \), \( n_B = 400 \), and \( \rho = 0.75 \), but otherwise outperforms its constituents in the \( p = 99 \) cases. \( \text{SRC} \) fares poorly when \( p = 99 \). Overall, \( \text{HIERBETAS} \) is the best-performing method in terms of MSPE. The relative difference between all methods is small when \( p = 5 \).
Fig. 2. Empirical MSPE over $\tau$ for the simulation study described in Section 5. Here $p$ stands for the number of covariates, $n_B$ is the size of subsample B, $\rho$ is the first-order auto-regressive correlation coefficient for pairwise combinations of $X$, and $R^2 = \beta^\top \Sigma_X \beta / (\beta^\top \Sigma_X \beta + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_A = 50$, $\beta_0 = \psi = 0$, and $v = 1$. $\sigma^2$, plotted in black, is the smallest possible MSPE for any estimate of $\beta$.

Appendix D in supplementary material available at *Biostatistics* online additionally looks at MSE and investigates several violations to the modeling assumptions in this study. While *HIERBETAS* continues to predict very well, another important result is that *HYB*, which is less expensive computationally, is still flexible: under a variety of model settings and violations, *HYB* is able to efficiently adhere to the best-performing of its constituents.
6. 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, 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, $W$) was measured on 439 tumor samples, and qRT-PCR measurements ($X$) 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.

Eleven measurements in the qRT-PCR-only data, out of $47 \times 91 = 4277$ total, or 0.26%, 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 1 month after surgery, and these were removed before analysis. Thus, $n_A = 47$, $n_B = 389$, and the validation data contain 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 refit 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 (see supplementary material available at Biostatistics online) presents the 91 LOESS curves comparing measurements from the 47 tumors using Affymetrix ($w_A$) to qRT-PCR ($x_A$) after standardization. Based on this, we used a gene-specific ME model: $w_{ij} = \psi_j + v_j x_{ij} + \tau \xi_{ij}$. We modeled $\psi_j$ and $v_j$ as random effects, distributed as $N(\mu_{\psi}, \sigma_{\psi}^2)$ and $N(\mu_{v}, \sigma_v^2)$, and used predictions $\{\hat{\psi}_j\}$ and $\{\hat{v}_j\}$ to calculate $x_{\text{BRC}}^A$ and $x_{\text{BRC}}^A$. For hierbetas, we relaxed the ME model identically. Violation of the constant $\tau$ 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, $\tau = 0.628$. 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, hyb, and hierbetas. Table 2 presents numerical results for each of the methods, and Figure 3 plots each estimate of $\beta$ as a kernel density. In terms of MSPE, the best method was hierbetas, with an MSPE of 0.559, compared to 0.620 (ridg), 0.781 (frc), and 8.745 (src). For hyb, $\omega^{\text{opt}} = \{1, 0, 0\}$, corresponding to ridg, src, and frc; so $\hat{\beta}_{\text{hyb}} = \hat{\beta}_{\text{ridg}}$ and hyb matches the best of its constituents. Plugging in $\hat{\beta} = 0_p$ yields an MSPE of 0.590, which only hierbetas beats, suggesting a very weak signal in the set of expression measures for predicting survival. In our simulation study of low-$R^2$ situations, we observed a similar ranking of methods. The range of $\hat{\beta}_{\text{ridg}}$ and $\hat{\beta}_{\text{hyb}}$, excluding the intercept, is $(-0.019, 0.014)$. For $\hat{\beta}_{\text{src}}$, it is $(-0.588, 0.515)$, for $\hat{\beta}_{\text{frc}}$, it is $(-0.075, 0.062)$, and for $\hat{\beta}_{\text{hierbetas}}$, it is $(-0.027, 0.023)$. From Figure 3, the kernel density estimates of ridg and hierbetas are similar, despite yielding different MSPEs. That
Table 2. Results from the data analysis

<table>
<thead>
<tr>
<th></th>
<th>RIDG</th>
<th>SRC</th>
<th>FRC</th>
<th>HYB</th>
<th>HYBunc</th>
<th>HIERBETAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>M$\text{SPE}$</td>
<td>0.620</td>
<td>8.745</td>
<td>0.781</td>
<td>0.620</td>
<td>0.601</td>
<td>0.559</td>
</tr>
<tr>
<td>min($\hat{\beta}$)</td>
<td>$-0.019$</td>
<td>$-0.588$</td>
<td>$-0.075$</td>
<td>$-0.019$</td>
<td>$-0.053$</td>
<td>$-0.027$</td>
</tr>
<tr>
<td>max($\hat{\beta}$)</td>
<td>0.014</td>
<td>0.515</td>
<td>0.062</td>
<td>0.014</td>
<td>0.057</td>
<td>0.023</td>
</tr>
<tr>
<td>Avg. coverage</td>
<td>0.91</td>
<td>1.00</td>
<td>0.98</td>
<td>0.91</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>Avg($\hat{\gamma}<em>{B.97.5} - \hat{\gamma}</em>{B.2.5}$)</td>
<td>3.372</td>
<td>33.785</td>
<td>4.023</td>
<td>3.372</td>
<td>4.674</td>
<td>3.310</td>
</tr>
</tbody>
</table>

M$\text{SPE}$ is the empirical MSPE from the validation sample of size 100; min($\hat{\beta}$) and max($\hat{\beta}$) give the range of the estimate of $\beta$ for each model, Avg. coverage is the proportion of bootstrap-generated prediction intervals for the validation sample which contained the true outcome, and Avg($\hat{\gamma}_{B.97.5} - \hat{\gamma}_{B.2.5}$) gives the average prediction interval length for the validation sample. HYBunc is the hybrid estimator without the non-negativity constraint (Remark 6.1).

Fig. 3. Kernel density estimate of the 91 elements of $\hat{\beta}_{\text{RIDG}}, \hat{\beta}_{\text{SRC}}, \hat{\beta}_{\text{FRC}}, \hat{\beta}_{\text{HIERBETAS}}$, and $\hat{\beta}_{\text{HYBunc}}$, the hybrid estimator without the non-negativity constraint (Remark 6.1), from the data analysis. $\hat{\beta}_{\text{HYB}}$, with the non-negativity constraint, is identically equal to $\hat{\beta}_{\text{RIDG}}$.

the MSPEs differ despite similar overall shrinkage may be expected given that HIERBETAS is itself a ridge regression that uses more data than RIDG.

Finally, we generated 95% prediction intervals for each observation in the validation sample, using a bootstrap algorithm described in Appendix E of supplementary material available at Biostatistics online. For HIERBETAS, draws from the posterior distribution of $\beta_0$, $\beta$, and $\sigma^2$ naturally yield prediction intervals for future observations. 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). HIERBETAS is closest to nominal, with 0.94.

Remark 6.1 As in the simulation study, we restricted our optimization of $\omega$ to the subspace of non-negative elements, which on average improves numerical results. In the data analysis, removing the constraint yields $\omega^{\text{opt}} = \{1.094, -0.100, 0.006\}$ and an MSPE of 0.601. These unconstrained results are also presented in Table 2 and Figure 3 denoted as HYBunc.

7. 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 ME.
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.

The Bayesian ridge regression, HIERBETAS, proved to be competitive with HYB and typically had smaller MSPE. Also, prediction intervals using HIERBETAS are automatic with draws from the posterior distribution. For the TR methods and HYB, a simple bootstrap algorithm yields prediction intervals but requires some modifications to achieve nominal coverage rates.

Despite this, there are reasons to recommend HYB. First, HYB is flexible: it is a linear combination of estimators, each of which can make different modeling assumptions. For example, RIDG assumes only the outcome regression model in (1.1), FRC additionally assumes the ME model in (1.2), and SRC assumes these two models plus the marginal model: \( X \sim N_p(\mu_X, \Sigma_X) \). Estimators with different modeling assumptions, beyond what we have proposed in this paper, can also be included in HYB, and, from Theorem 3.1, it will theoretically do better than the best of any of these. HIERBETAS does not benefit from this robust model-averaging property. Practically, the average performance of HYB across all design and data configurations is encouraging, and, importantly, its flexibility is most apparent in the large \( p \) scenarios. Second, and more significantly, HYB is very fast to compute, whereas HIERBETAS requires considerably more computational effort. Finally, because HYB combines TR estimators, a GCV criterion provides a simple estimate of \( P \), the prediction error matrix (3.2), which is required to optimize with respect to \( \omega \). In our current research, we are exploring an improved GCV criterion which avoids the tendency of GCV to overfit in small-sample scenarios. This has potential to further improve estimates of \( P \) and, consequently, prediction for HYB. Alternatively, the “632 estimator” of Efron (1983) is another candidate for estimating \( P \).

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 TR 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.

**Supplementary Material**

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

**Acknowledgments**

The authors would like to thank the associate editor and two reviewers for their insightful comments. Conflict of Interest: None declared.

**Funding**

This work was supported by the National Science Foundation [DMS1007494] and the National Institutes of Health [CA129102, CA156608, ES020811].
REFERENCES


[Received February 8, 2012; revised August 27, 2012; accepted for publication August 28, 2012]