Shrinkage Methods Utilizing Auxiliary Information to Improve High-Dimensional Prediction Models

by

Philip S. Boonstra

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2012

Doctoral Committee:

Associate Professor Bhramar Mukherjee, Co-Chair
Professor Jeremy M. G. Taylor, Co-Chair
Assistant Professor Long Nguyen
Professor Trivellore E. Raghunathan

© Philip S. Boonstra 2012

All Rights Reserved

to Laura

and to Simon and Amos – our hybrid estimators

Psalm 19
[1]The heavens declare the glory of God;
the skies proclaim the work of his hands.
[2]Day after day they pour forth speech;
night after night they reveal knowledge.

ACKNOWLEDGMENTS

This document is a product of the support and encouragement of instructors, colleagues, friends and family. Those listed below are a subset, and I apologize to anyone I have left out.

Thanks to my dissertation committee. Beyond giving of her immense statistical knowledge, Bhramar, my co-chair, has been and is still a dedicated mentor. She often provided needed reminders to trust the wisdom of my advisors. Jeremy, also my co-chair, shared with me of his vast statistical breadth and depth time and again. Being his student has uniformly improved my academic writing. Both Bhramar and Jeremy have supported me financially over the past years, provided for me to attend conferences, and gone out of their way to connect me with their statistical collaborators. It has been a pleasure working with them. Long, in addition to serving on my committee, also taught Probabilistic Graphical Models, my favorite class here at Michigan and the one that most impacted my understanding of statistics. And Raghu generously gave of his abundant wisdom in all things missing data, ensuring that no data in this document were modeled unnecessarily.

Thanks also to other instructors and to my peers. Professor Randy Pruim at Calvin College introduced me to statistics as a part of general revelation and showed me the elegance of `R`. Professor Tim Johnson and I had helpful conversations on programming in `C`. Professor Michael Elliott encouraged me during my graduate school experience through classes, office hours, chance meetings, and a perpetual willingness to write me letters of recommendation. Professor Jack Kalbfleisch supported me through his Training Grant and patiently answered my inarticulate statistical questions as I prepared for the qualifying exams. Matthew Flickinger and I had five years' worth of edifying lunchtime conversations. And office colleagues Jared Foster and Oliver Lee offered empathy on the challenges of graduate school.

Finally, thanks to friends and family. The Ann Arbor Christian Reformed Church has been our family here from our first week in town. My frank conversations with Ross Weener were always appreciated, and I enjoyed relaxing evenings playing games of Power Grid, Dominion, and Pandemic with Justin Schiller and Ross. My parents, Rich and Trena, afforded me educational opportunities and made many little sacrifices I see only now, being a parent myself. Our sons Simon and Amos shared with me in countless hours of fruitful diversion playing toys and reading books.

And of course my fantastic and lovely wife, Laura, has supported me through school in every sense of the word and deserves much of the credit for this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

Introduction

In this dissertation we consider making predictions in a high-dimensional dataset for which only surrogate covariates are measured in a large fraction of the data. There is a sparse body of literature for analyzing high-dimensional data, meaning the number of observations is small relative to the number of covariates, with concurrent missing information. Our goal is to contribute novel statistical methodology toward this problem, specifically to improve predictions.

More formally, we predict a continuous outcome Y using high-dimensional covariates $\boldsymbol{X}$. In all observations, a surrogate covariate, $\boldsymbol{W}$, is available, and only in a small number of observations is $\boldsymbol{X}$ observed. We seek to integrate these auxiliary observations, those for which $\boldsymbol{X}$ is missing, into the analysis of the smaller, complete dataset. Using notation that will be introduced in Chapter 2, a schematic representation of the data structure is given in Figure 1. Of primary interest are parameter estimates from the model

$$Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim N\{0,1\}.$$

Estimates of β_0 and $\boldsymbol{\beta}$ will be used in the prediction of future observations, which we assume will contain actual covariate information, meaning $\boldsymbol{X}$ is observed.

The attributes of the problem, ie many observations are missing $\boldsymbol{X}$ with only a surrogate $\boldsymbol{W}$ available, suggest plausible approaches to its analysis. First, this is a problem of measurement error in the covariates, and standard techniques like regression calibra-

tion (Fuller, 1987; Carroll et al., 2006) may replace the missing $\boldsymbol{X}$'s with their conditional expectation. Alternatively, taking a missing data approach, the missing $\boldsymbol{X}$'s can be multiply imputed (Rubin, 2004). Complete data analysis techniques can then be employed, and parameter estimates are obtained by averaging over results from each completed dataset, as described in Little and Rubin (2002). Both the measurement error and missing data approaches will produce estimates of $\boldsymbol{\beta}$, and therefore predictions for future observations of $\boldsymbol{Y}$, that are approximately unbiased.

As will be discussed in more detail in later chapters, a complication from these approaches is that estimates of $\boldsymbol{\beta}$, although approximately unbiased, may have large variance, which will adversely affect predictions. This is due primarily to the high dimension of the problem and further compounded by the large fraction of missing information. Our approaches to this problem make extensive use of shrinkage methods, by which we mean techniques for reducing the variance of parameter estimates at a cost of introducing bias. This trade-off may improve predictions for future observations, the quality of which is a function of both bias and variance.

In the example that first motivated our work, Chen et al. (2011) analyzed a gene-expression microarray dataset of 439 lung adenocarcinomas, with the goal of using gene expression plus clinical covariates to improve predictions of survival time in lung cancer patients relative to using just clinical covariates. The expressions were measured using Affymetrix oligonucleotide microarray technology ($\boldsymbol{W}$). 91 promising genes were identified and re-assayed using quantitative real-time polymerase chain reaction (qRT-PCR, $\boldsymbol{X}$). qRT-PCR is more precise and clinically applicable than Affymetrix, thus the goal was to obtain a qRT-PCR-based prediction model for future use. However, because of tissue availability, only 47 out of 439 tumors were re-assayed by qRT-PCR, creating a high-dimensional, missing data problem. Specifically, Affymetrix data was available for all observations, but the 91 qRT-PCR measurements corresponding to 392 tumors were

unobserved.

Except where otherwise noted, we will assume the framework of the lung adenocarcinoma data in Chen et al. (2011). However, this study is not an isolated example. The ongoing development of array technologies for assaying genomic information has accordingly resulted in high-dimensional datasets with several measures of the same underlying biological process. This scientific context has made important the statistical issue of integrating and synthesizing information from diverse data sources, all bearing relevance to common biological phenomena. The techniques proposed in this dissertation thus have broad applications in that general area. They help to answer the question of integrating existing data sources, for example coming from prior technology, with state-of-the-art measurements to enhance predictions for future patient outcomes. In the remainder of this chapter, we briefly outline this dissertation's contributions. Extensive reviews of the relevant literature are given within each chapter.

In Chapter 2, we consider targeted ridge (TR) regression, a generalization of ridge regression (Hoerl and Kennard, 1970) first discussed by Gruber (1998) that allows for shrinkage of estimates of $\boldsymbol{\beta}$ toward non-zero values. Using ideas from the measurement error literature, we propose several TR estimators. The shrinkage targets are derived using the auxiliary data, that is, the observations for which only surrogate measurements are available, and the amount of shrinkage may be controlled by a tuning parameter, as in ridge regression. We will interchangeably refer to this tuning parameter as a shrinkage parameter, referring to its role in determining the shrinkage of $\boldsymbol{\beta}$. We propose a hybrid estimator, which is a weighted linear combination of multiple TR estimators. The weights are data-adaptively estimated, so that the hybrid estimator may put more weight on the better-predicting TR estimators, giving it a model-averaging flexibility. Moreover, we give sufficient conditions under which the hybrid estimator has *strictly smaller* mean squared prediction error than any of its ingredients, a phenomenon that has been ob-

served empirically by Breiman (1996) and LeBlanc and Tibshirani (1996), among others.

Chapter 3 takes a fully Bayesian perspective to the same problem. Here, the shrinkage parameters are interpreted as hyperparameters, which define the prior distribution of the model parameters, including but not limited to $\boldsymbol{\beta}$. This link between shrinkage parameters and hyperparameters allows a flexibility in terms of which of the model parameters are shrunk and to what extent. In the estimation algorithm, the missing data are iteratively sampled from their full conditional distribution in conjunction with the model parameters and shrinkage parameters. Thus, all unknown quantities – the missing data, the model parameters, and optionally the shrinkage parameters – are treated uniformly, and uncertainty in the imputations is propagated. As an alternative to iterative sampling of the shrinkage parameters from their full conditional distributions, Empirical Bayes methods, which maximize marginal likelihoods, are also employed to calculate explicit estimates, as in Casella (2001).

Our results from Chapters 2 and 3 suggest that the amount of shrinkage, as determined by the choice of tuning parameter, plays a crucial role in the predictive success of a shrinkage method. So as to explore this in more detail, in Chapter 4, we propose novel approaches toward estimating the tuning parameter of a classical ridge regression, given by λ, particularly when the number of observations is small. We adjust the Generalized Cross Validation (GCV) criterion (Craven and Wahba, 1979), which is prone to overfitting in this small-sample scenario. We also propose a "hyperpenalized" likelihood. This shrinks the tuning parameter λ itself and protects against extreme choices. The hyperpenalized likelihood can be maximized jointly with respect to $\boldsymbol{\beta}$ and λ, in contrast to other likelihood-based approaches for choosing λ, which profile or marginalize over $\boldsymbol{\beta}$ (eg Harville, 1977; Wecker and Ansley, 1983; Wahba, 1985; Lee and Nelder, 1996).

In Chapter 5, the hyperpenalized likelihood is adapted to the missing-data context by embedding it into the penalized Expectation-Maximization (EM) algorithm (Dempster

et al., 1977; Green, 1990), allowing for the adaptive shrinkage of maximum likelihood estimates. We apply this so-called hyperpenalized EM (HEM) algorithm to the motivating gene-expression problem. A dissertation-wide comparison of the predictive performance of the Chapters 2 and 3 methods with the HEM algorithm method is conducted in Chapter 5. For reference, all of the methods from Chapters 2, 3, and 5, that is, those which address the missing data problem, are annotated in Tables D1 and D2 of Appendix D. We conclude the dissertation with a discussion of future work and related open problems.

CHAPTER 2

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

2.1 Introduction

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

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

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

$$\boldsymbol{W} = \psi\mathbf{1}_p + \nu\boldsymbol{X} + \tau\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim N_p\{\mathbf{0}_p, \boldsymbol{I}_p\}. \tag{2}$$

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

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

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

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

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

A simple approach, which ignores subsample B, is ordinary least squares of $\boldsymbol{y}_{\mathrm{A}}$ on $\boldsymbol{x}_{\mathrm{A}}$, i.e. $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) = (\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{y}_{\mathrm{A}}$. However, $(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1}$ does not exist 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 coefficients toward zero, i.e. $\hat{\boldsymbol{\beta}}_{\mathrm{RIDG}} = \mathrm{argmin}_{\boldsymbol{\beta}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} = (\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{y}_{\mathrm{A}}$. This can be viewed from a Bayesian perspective: given a normal prior on $\boldsymbol{\beta}$ with mean $\boldsymbol{0}_p$ and

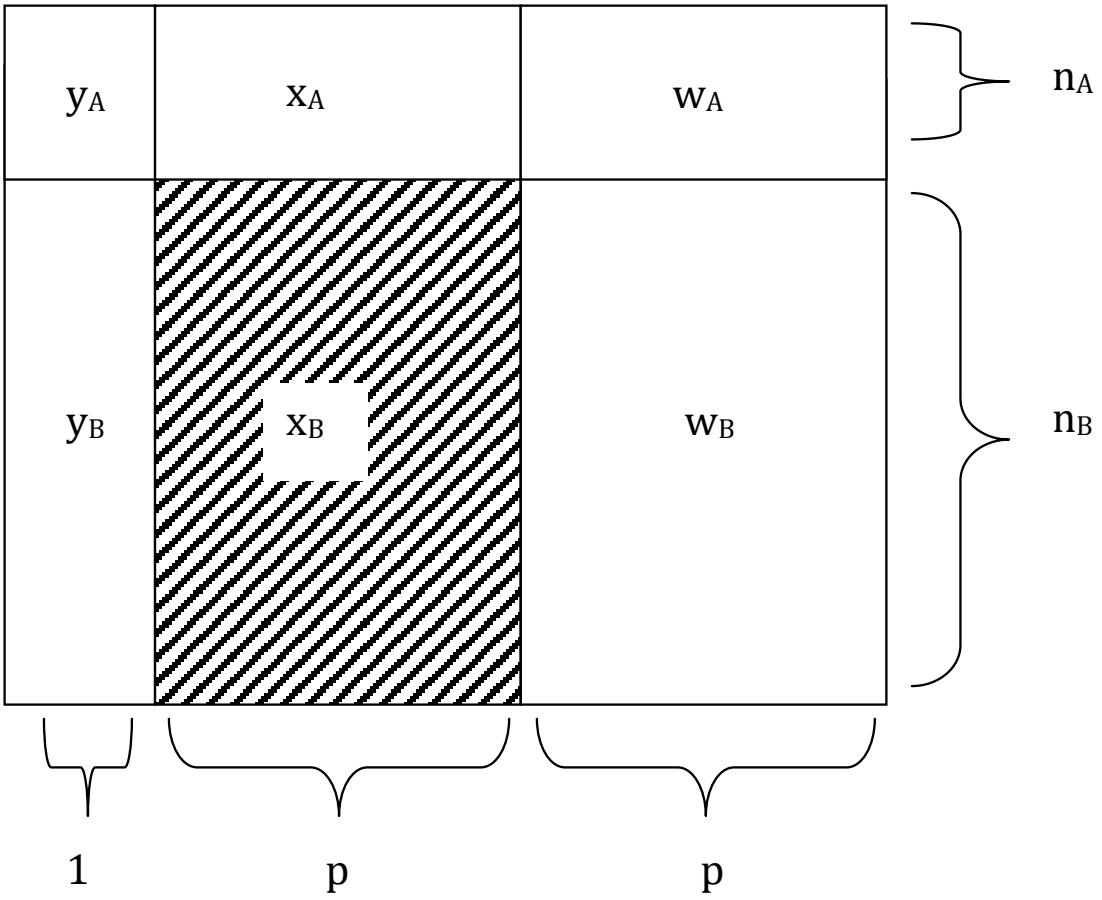

Figure 1: A schematic representation of the prediction problem: $\{\boldsymbol{y}_A, \boldsymbol{x}_A, \boldsymbol{w}_A\}$ constitutes subsample A, of size n_A, and $\{\boldsymbol{y}_B, \boldsymbol{w}_B\}$ constitutes subsample B, of size n_B. $\boldsymbol{x}_B$ is considered missing. $\boldsymbol{W}$ is an error-prone/noisy version of $\boldsymbol{X}$. The goal is to utilize the data on $\boldsymbol{W}$ to boost prediction of Y by $\boldsymbol{X}$.

precision $\sigma^{-2}\lambda \boldsymbol{I}_p$, the RIDG coefficients are the posterior mode for a given λ. Hoerl and Kennard showed that there exists $\lambda > 0$ that decreases mean squared error, $\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$, compared with $\lambda = 0$. RIDG penalizes the ℓ_2 norm; other methods exist that constrain the ℓ_d norm for some d (eg Frank and Friedman, 1993). In contrast to variable selection procedures, which might use an ℓ_1 penalty, our goal is using auxiliary information to boost prediction, and so we restrict attention to ridge-type estimators.

Dempster 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 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. Sclove (1968) adapts the shrinkage estimator of James and Stein (JS, 1961) 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 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 et al., 2009).

For $p < n_A$, the problem closely resembles that of measurement error (ME) in the covariates, $\boldsymbol{W}$ being an error-prone version of $\boldsymbol{X}$. Fuller (1987) 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 $\boldsymbol{W}_{new}^\top \hat{\boldsymbol{\beta}}$. Our prediction model of interest being Y given $\boldsymbol{X}$, this bias in $\hat{\boldsymbol{\beta}}$ from using $\boldsymbol{W}$ instead of $\boldsymbol{X}$ *does* bias $\boldsymbol{X}_{new}^\top \hat{\boldsymbol{\beta}}$ away from Y_{new}. Regression calibration, which fills in each missing $\boldsymbol{X}$ with its conditional expectation given $\boldsymbol{W}$, may provide unbiased estimates of $\boldsymbol{\beta}$ and therefore Y_{new}. In contrast, although the substitution of $\boldsymbol{X}$ by $\boldsymbol{W}$ gives biased estimates of $\boldsymbol{\beta}$, it may reduce the *variance* of estimates of $\boldsymbol{\beta}$ relative to regression calibration (Buzas et al., 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 $\boldsymbol{X}_{new}^\top \hat{\boldsymbol{\beta}}$.

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

The rest of the chapter is organized as follows. In Section 2.2, we unify RIDG and regression calibration methods under a class of targeted ridge estimators. In Section 2.3, we propose hybrid estimators, which combine multiple estimators with data-adaptive weights to achieve superior prediction. Section 2.4 presents a simulation study. Section

2.5 applies the methods, in which survival time, Y, in lung cancer patients is predicted with qRT-PCR data, $\boldsymbol{X}$, with microarray data, $\boldsymbol{W}$, from a larger sample aiding in predictions. Section 2.6 concludes with a discussion. Most analytical details are in Appendix A.

2.2 Targeted Shrinkage

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

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

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

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

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

where $\boldsymbol{\Theta}$ is an arbitrary $p \times p$ positive semi-definite (PSD) matrix. Rao (1975) suggested that any PSD matrix $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$ can replace $\boldsymbol{I}_p$ in (4). Swindel (1976) proposed to shrink toward a non-null vector $\boldsymbol{\gamma}_{\boldsymbol{\beta}}$. From the Bayesian perspective, these replace the prior precision

$\sigma^{-2}\lambda \boldsymbol{I}_p$ in RIDG with $\sigma^{-2}\lambda \boldsymbol{\Omega}_\beta^{-1}$ and the prior mean $\mathbf{0}_p$ with $\boldsymbol{\gamma}_\beta$. The posterior mode is

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1}) = \operatorname{argmin}_\beta \frac{1}{\sigma^2}(\boldsymbol{y}_\mathrm{A} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top(\boldsymbol{y}_\mathrm{A} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta}) + \frac{1}{\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\gamma}_\beta)^\top \lambda \boldsymbol{\Omega}_\beta^{-1}(\boldsymbol{\beta} - \boldsymbol{\gamma}_\beta) \quad (6)$$

$$= (\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1})^{-1}(\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{y}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1}\boldsymbol{\gamma}_\beta). \quad (7)$$

Gruber (1998, p.241) calls this a generalized ridge estimator. Because "generalized ridge" has been used for several distinct methods in the shrinkage literature, we instead call this a targeted ridge (TR) estimator, referring to shrinkage toward a target $\boldsymbol{\gamma}_\beta$. The estimator in (7) gives the three terms $\{\boldsymbol{\gamma}_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1}\}$ that determine the general class of TR estimators. As we shall see, different estimators we propose either implicitly or explicitly specify values for $\{\boldsymbol{\gamma}_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1}\}$. In particular, RIDG is a TR estimator: $\hat{\boldsymbol{\beta}}_\mathrm{RIDG} = \hat{\boldsymbol{\beta}}(\mathbf{0}_p, \lambda, \boldsymbol{I}_p)$.

As stated in (3), $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}) = \sigma^2 + \operatorname{Tr}\big[(\operatorname{Bias}\hat{\boldsymbol{\beta}} \operatorname{Bias}\hat{\boldsymbol{\beta}}^\top + \operatorname{Var}\hat{\boldsymbol{\beta}})\mathrm{E}[\boldsymbol{X}_\mathrm{new}\boldsymbol{X}_\mathrm{new}^\top]\big]$. Thus we calculate the MSPE of a TR estimator from its bias and variance, taking expectations over the response distribution $\boldsymbol{y}_\mathrm{A}, \boldsymbol{y}_\mathrm{B} | \boldsymbol{x}_\mathrm{A}, \boldsymbol{w}_\mathrm{A}, \boldsymbol{w}_\mathrm{B}$:

$$\operatorname{Bias}\hat{\boldsymbol{\beta}} = \mathrm{E}\,\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1})^{-1}(\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} + \lambda \boldsymbol{\Omega}_\beta^{-1}\mathrm{E}\,\boldsymbol{\gamma}_\beta - \boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} - \lambda \boldsymbol{\Omega}_\beta^{-1}\boldsymbol{\beta})$$

$$= \lambda(\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1})^{-1}\boldsymbol{\Omega}_\beta^{-1}(\mathrm{E}\,\boldsymbol{\gamma}_\beta - \boldsymbol{\beta}) \quad (8)$$

$$\operatorname{Var}\hat{\boldsymbol{\beta}} = (\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1})^{-1}(\sigma^2 \boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda^2 \boldsymbol{\Omega}_\beta^{-1}\operatorname{Var}\boldsymbol{\gamma}_\beta\, \boldsymbol{\Omega}_\beta^{-1})(\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \lambda \boldsymbol{\Omega}_\beta^{-1})^{-1}. \quad (9)$$

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

We now propose several other TR estimators. If $\boldsymbol{x}_\mathrm{B}$ were observed, logical selections of $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\Omega}_\beta^{-1}$ 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 $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\Omega}_\beta^{-1}$, that is, use $\boldsymbol{w}_\mathrm{B}$ itself as an imputation for $\boldsymbol{x}_\mathrm{B}$. We first consider approaches that derive a better

replacement for the missing x_B. This is obtained by modeling $W|X$ based on the relationship observed in subsample A and thereby inducing data-driven values of γ_β and Ω_β^{-1}. From the ME perspective, this is regression calibration. These TR estimators fix $\lambda = 1$; data-adaptive estimation of λ may be done using, eg 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 $\theta = \{\nu, \tau, \mu_X, \Sigma_X^{-1}\}$. From properties of the conditional distribution of $X|W$,

$$x_B^{SRC}(\theta) = E[x_B|w_B, \theta] = \mathbf{1}_{n_B}\mu_X^\top(I_p - V(\theta)) + \frac{1}{\nu}w_B V(\theta) = [\mathbf{1}_{n_B}, w_B]M(\theta), \quad (10)$$

$$M(\theta) = \begin{pmatrix} \mu_X^\top(I_p - V(\theta)) \\ \frac{1}{\nu}V(\theta) \end{pmatrix}, \qquad V(\theta) = \left(I_p + \frac{\tau^2}{\nu^2}\Sigma_X^{-1}\right)^{-1}. \quad (11)$$

We suppress the dependence on θ of $x_B^{SRC}(\theta)$, $M(\theta)$, and $V(\theta)$ hereafter. This is a precision-weighted average of $\mathbf{1}_{n_B}\mu_X^\top$ and $(1/\nu)w_B$. Using (7), define $\hat{\beta}_{SRC} = \hat{\beta}(\gamma_{\beta_{SRC}}, 1, \Omega^{-1}_{\beta_{SRC}})$, with $\gamma_{\beta_{SRC}} = (x_B^{SRC\top}x_B^{SRC})^{-1}(x_B^{SRC\top}y_B)$ and $\Omega^{-1}_{\beta_{SRC}} = x_B^{SRC\top}x_B^{SRC}$. In the ME literature, SRC is the standard "Regression Calibration" approach. We append "Structural" (Carroll et al., 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 (2), $W = \nu X + \tau\xi$, for X gives $X = (1/\nu)W - (\tau/\nu)\xi$. Another natural estimate of x_B, and consequently a corresponding γ_β and Ω_β^{-1}, is therefore

$$x_B^{FRC}(\theta) = (1/\nu)w_B, \qquad \gamma_{\beta_{FRC}} = (x_B^{FRC\top}x_B^{FRC})^{-1}x_B^{FRC\top}y_B, \qquad \Omega^{-1}_{\beta_{FRC}} = x_B^{FRC\top}x_B^{FRC}. \quad (12)$$

This gives a TR estimate defined as $\hat{\beta}_{FRC} = \hat{\beta}(\gamma_{\beta_{FRC}}, 1, \Omega^{-1}_{\beta_{FRC}})$. This imputation for x_B is a scaled version of a substitution of w_B for x_B, to which FRC is equivalent when $\nu = 1$,

ie under the classical ME model. In Appendix A.1, we conduct extensive analyses that suggest that FRC is preferred over SRC in terms of MSPE as any of $\boldsymbol{\beta}^\top\boldsymbol{\beta}$, σ^2, or τ/ν increase.

The first rows of Table 1 summarize choices of $(\gamma_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1})$ for RIDG, FRC, and SRC. Assuming non-differential measurement error (NDME) , ie $[Y|\boldsymbol{X}, \boldsymbol{W}] = [Y|\boldsymbol{X}]$, and $\boldsymbol{\mu}_X = \mathbf{0}_p$, Table 1 also gives $\mathrm{E}\,\gamma_\beta$ and $\mathrm{Var}\,\gamma_\beta$ for FRC and SRC. Because $\mathrm{E}\,\gamma_{\beta_{\text{SRC}}} = \boldsymbol{\beta}$, from (8), SRC provides unbiased estimates of $\boldsymbol{\beta}$.

REMARK 1: When γ_β and $\boldsymbol{\Omega}_\beta^{-1}$ are based on historical data, the prior in the second expression of (6) is a power prior (Chen and Ibrahim, 2000), with λ controlling the contribution of the historical data to the posterior.

REMARK 2: These approaches require estimating $\boldsymbol{\theta} = \{\nu, \tau, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}\}$. One can regress $\{w_{ij}\}$ on $\{x_{ij}\}$ for $i = 1, \ldots, n_\text{A}$ and $j = 1, \ldots, p$ to compute MLEs for ν and τ. If it is required that ν and τ be of a more general form than scalar-valued, the estimation procedure can be modified accordingly. The MLE for $\boldsymbol{\mu}_X$ is $\hat{\boldsymbol{\mu}}_X = n_\text{A}^{-1}\boldsymbol{x}_\text{A}^\top\mathbf{1}_{n_\text{A}}$, which will be $\mathbf{0}_p$ when $\boldsymbol{x}_\text{A}$ is standardized. For $p > n_\text{A}$, the required inversion of $\hat{\boldsymbol{\Sigma}}_X = n_\text{A}^{-1}\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A}$ is not possible. An alternative is the shrinkage estimator from Schäfer and Strimmer (2005): since $\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A}$ is standardized, it is simply $\hat{\boldsymbol{\Sigma}}_X^* = (1-\pi)\hat{\boldsymbol{\Sigma}}_X + \pi\boldsymbol{I}_p$, for $\pi \in [0,1]$ chosen data-adaptively. We used the `R` package `corpcor` to choose π targeting a minimum MSE for $\hat{\boldsymbol{\Sigma}}_X^*$.

REMARK 3: The bias and variance outlined in Table 1 condition on the true value of $\boldsymbol{\theta}$ and are over and above any bias and variance coming from its estimation. In particular, estimating $\boldsymbol{\Sigma}_X$ may pose a challenge to SRC in the high-dimensional setting.

REMARK 4: One other approach, which we do not further explore, is modifying FRC or SRC to do adaptive, component-wise shrinkage on $\boldsymbol{\beta}$: a TR estimator where $\boldsymbol{\Omega}_\beta^{-1}$ is diagonal and λ is estimated. When λ is not fixed, the GCV approach may be used to choose an appropriate value of λ. The form of this modified GCV criterion is given later

on in (14), in connection with the hybrid estimator.

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

Table 1: Key information for several TR estimators, conditioning on the true value of $\boldsymbol{\theta}$. $\kappa = (\tau^2/\nu^2)\boldsymbol{\beta}^\top \boldsymbol{V}\boldsymbol{\beta}$. $\boldsymbol{V} = (\boldsymbol{I}_p + (\tau^2/\nu^2)\boldsymbol{\Sigma}_\mathbf{X}^{-1})$. The '$\lambda = 1$?' column indicates whether λ is fixed at 1 or tuned in a data-adaptive fashion using the general GCV function. The corresponding estimator $\hat{\boldsymbol{\beta}}(\gamma_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1})$ is given by plugging $(\gamma_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1})$ into (7). The expectation and variance of γ_β, which are useful for calculating the MSPE of $\hat{\boldsymbol{\beta}}(\gamma_\beta, \lambda, \boldsymbol{\Omega}_\beta^{-1})$, 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 $[Y|\boldsymbol{X}, \boldsymbol{W}] = [Y|\boldsymbol{X}]$.

2.3 Hybrid Estimators

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

$$\mathrm{MCPE}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_j) = \sigma^2 + \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_i)^\top \boldsymbol{X}_\mathrm{new}\boldsymbol{X}_\mathrm{new}^\top(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_j)]. \tag{13}$$

Theorem 2.1. *Let* $\boldsymbol{b}(\boldsymbol{\omega}) = [\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \cdots, \hat{\boldsymbol{\beta}}_m]\boldsymbol{\omega}$ *be a hybrid estimator.*

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

The proof is in Appendix A.2. If the assumptions are satisfied, then, using prediction error as the criterion, $\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})$ will perform better than the best of its constituents. This phenomenon has been observed empirically by Breiman (1996) and LeBlanc and Tibshirani (1996). Fumera and Roli (2005) prove a slightly weaker result for ensembles of classifiers.

Now, $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega})) = \boldsymbol{\omega}^\top \boldsymbol{P} \boldsymbol{\omega}$, where $\boldsymbol{P}$ is the $m \times m$ matrix with the (ij)th element given by $P_{ij} = \text{MCPE}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$, which is just $\text{MSPE}(\hat{\boldsymbol{\beta}}_i)$ when $i = j$. The results from Theorem 2.1 apply when $\boldsymbol{P}$ is known. In practice, however, $\boldsymbol{P}$ and therefore $\boldsymbol{\omega}^{\text{opt}}$ must be estimated. Since P_{ij} is equivalently expressed as $\text{E}[(Y_{\text{new}} - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}_i)(Y_{\text{new}} - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}_j)]$, one might use $(1/n_{\text{A}})(\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}} \hat{\boldsymbol{\beta}}_i)^\top (\boldsymbol{y}_{\text{A}} - \boldsymbol{x}_{\text{A}} \hat{\boldsymbol{\beta}}_j)$ as an estimate, but this will be biased. Lemma A.7 in the Appendices, a generalization of a result from Mallows (1973) gives that, on average, this underestimates P_{ij} by the amount $\sigma^2(\psi_i + \psi_j)$, where $\psi_\ell = \text{Tr}\, \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}_{\boldsymbol{\beta},\ell}^{-1})/n_{\text{A}}$. Borrowing Mallows' idea of adjusting by $\hat{\sigma}^2(\psi_i + \psi_j)$ does not work when there is no good choice of $\hat{\sigma}^2$. We propose as an alternative adapting the GCV approach:

$$\hat{P}_{ij} = \frac{\frac{1}{n_{\text{A}}}(\boldsymbol{y}^*_{\text{A},i} - \boldsymbol{H}(\lambda_i \boldsymbol{\Omega}_{\boldsymbol{\beta},i}^{-1})\boldsymbol{y}^*_{\text{A},i})^\top (\boldsymbol{y}^*_{\text{A},j} - \boldsymbol{H}(\lambda_j \boldsymbol{\Omega}_{\boldsymbol{\beta},j}^{-1})\boldsymbol{y}^*_{\text{A},j})}{(1 - \psi_i)(1 - \psi_j)}, \tag{14}$$

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

Note the dual use of the GCV function to calculate $\boldsymbol{b}(\boldsymbol{\omega})$. First, for each ℓ, λ_ℓ is chosen, when required, to minimize $\hat{P}_{\ell\ell}$. Then, fixing these choices of λ_ℓ, (14) is employed on the

$m(m+1)/2$ pairwise combinations of components in $\boldsymbol{b}(\boldsymbol{\omega})$ to estimate $\boldsymbol{P}$. The particular hybrid estimator we evaluate has three components: $\hat{\boldsymbol{\beta}}_{\text{HYB}} = [\hat{\boldsymbol{\beta}}_{\text{RIDG}}\, \hat{\boldsymbol{\beta}}_{\text{SRC}}\, \hat{\boldsymbol{\beta}}_{\text{FRC}}]\hat{\boldsymbol{\omega}}^{\text{opt}}$. Following LeBlanc and Tibshirani (1996), in addition to the constraint $\mathbf{1}_m^\top \boldsymbol{\omega} = 1$, we enforce a non-negativity constraint on $\boldsymbol{\omega}$, which improves numerical results.

REMARK 5: The key aspect that makes $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ practical is that the sum $\sigma^2 + \text{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{HYB}})^\top \boldsymbol{X}_{\text{new}} \boldsymbol{X}_{\text{new}}^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\text{HYB}})]$ is the quantity to minimize. Estimating either of the terms alone is difficult. Green and Strawderman (1991) propose a similar combination of two estimators that minimizes the MSE of $\boldsymbol{b}(\boldsymbol{\omega})$. For their method, the estimation of $\boldsymbol{\omega}^{\text{opt}}$ requires an unbiased $\hat{\boldsymbol{\beta}}_1$ and independent estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$. In our case, because MSPE, not MSE, is of interest, we require neither unbiasedness nor independent estimators.

REMARK 6: Although THEOREM 2.1 proves HYB has a smaller MSPE than any of its constituents when using the true optimal weights $\boldsymbol{\omega}^{\text{opt}}$, for a given dataset with estimated optimal weights $\hat{\boldsymbol{\omega}}^{\text{opt}}$, this uniform dominance may not hold. Numerical performance depends on how accurately (14) estimates $\boldsymbol{P}$. As will be seen, $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ with estimated weights still performs well across a spectrum of scenarios and closely adapts to the best of its constituents.

2.4 Simulation Study

We next describe a small simulation study. We fixed $n_{\text{A}} = 50$ and used $n_{\text{B}} \in \{400, 150\}$. The diagonal elements of $\boldsymbol{\Sigma}_X$ 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 $N_p\{\mathbf{0}_p, \boldsymbol{\Sigma}_X\}$. We considered both high- ($p = 99$) and low- ($p = 5$) dimensional models: $\boldsymbol{\beta} = \{j/100\}_{j=-49}^{j=49}$ and $\boldsymbol{\beta} = \{j/4\}_{j=-2}^{j=2}$. The coefficient of determination, R^2, was either 0.1 or 0.4. Thus, given $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_X$ and R^2, σ was determined by solving $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2) = R^2$. β_0 was set to zero. $\boldsymbol{y}_{\text{A}} | \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{y}_{\text{B}} | \boldsymbol{x}_{\text{B}}$ were drawn for each combination of $\boldsymbol{\beta}$ and σ from (1). This

yielded 16 unique simulation settings: two choices each for p, n_B, ρ, and R^2. To draw the auxiliary data, we set $\psi = 0$ and $\nu = 1$ and repeated each of the 16 settings for $\tau \in (0, 2)$, drawing $\boldsymbol{w}_A|\boldsymbol{x}_A$ and $\boldsymbol{w}_B|\boldsymbol{x}_B$ from (2).

For four methods, RIDG, SRC, FRC, and HYB, we estimated MSPE by averaging the squared prediction error over 1000 new individuals. Figure 2 plots this empirical MSPE averaged over 1000 replicates over τ. For reference, σ^2, the smallest achievable MSPE, is also given. Tables A2 and A3 in Appendix A provide numeric values of the empirical MSPE over all settings.

REMARK 7: In practice, the analyst estimates β_0 in addition to $\boldsymbol{\beta}$. Following the common prescription for ridge regression, we did not shrink β_0 but instead used a flat "prior" in each of the TR methods.

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

Effect of n_B, p, ρ, R^2 As might be expected, larger values of n_B considerably decrease MSPE for SRC, FRC and, consequently, HYB. Notably, HYB sometimes fares poorly compared with FRC (see Remark 6) when $p = 99$, $n_B = 400$, and $\rho = 0.75$. In the other $p = 99$ scenarios, HYB matches or outperforms every other method. SRC fares poorly when $p = 99$. On the other hand, when $p = 5$, HYB is typically not the best method. Here, all the methods are similarly ranked regardless of other parameter settings, with SRC usually having the smallest MSPE, the exception being the case of $\rho = 0.75$, $R^2 = 0.1$ and $n_B = 150$ case.

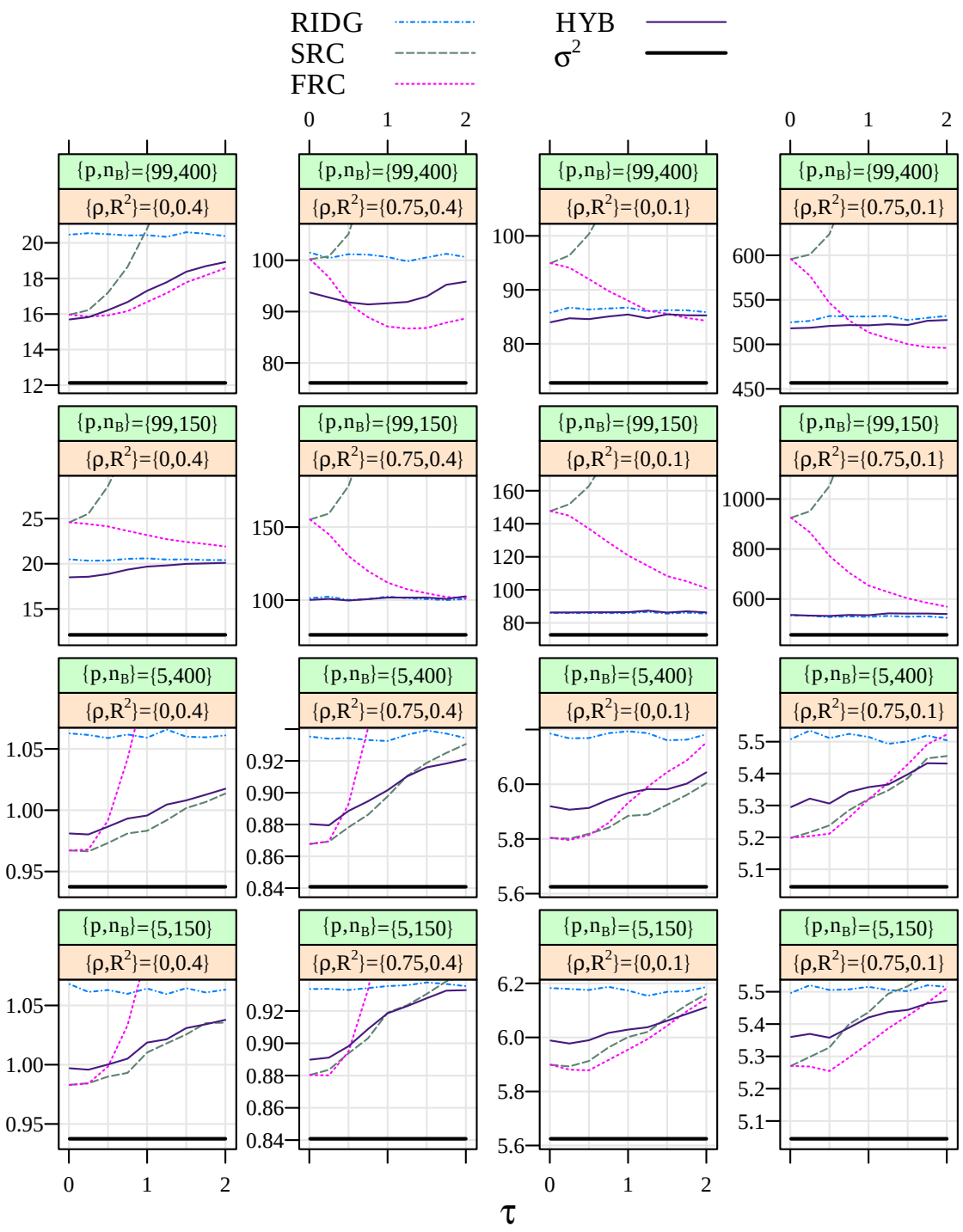

Figure 2: Empirical MSPE over τ for 16 simulation settings described in Section 2.4. p stands for the number of covariates, n_B is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_A = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. σ^2, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

Evaluating MSE: We also evaluated each simulation in terms of MSE of $\hat{\boldsymbol{\beta}}$ (Figure A1 in Appendix A.3). When $\rho = 0$, the results are nearly the same as those for MSPE, up to additive constant. This is to be expected: when $\boldsymbol{\mu_X} = \mathbf{0}_p$ and $\boldsymbol{\Sigma_X} = \boldsymbol{I}_p$, $\text{MSPE}(\hat{\boldsymbol{\beta}}) = \sigma^2 + \text{MSE}(\hat{\boldsymbol{\beta}})$. When $\rho = 0.75$, this relationship does not hold, and some rankings of the methods change. However, even though it minimizes prediction error, HYB is the best method overall in terms of MSE, particularly for the $p = 99$ cases.

Appendix A.3 investigates several violations to the modeling assumptions in this study. The most important result of these studies is that HYB is a flexible method. Under a variety of model settings and violations, HYB is able to efficiently adhere to the best-performing of its constituents.

2.5 Example: Lung Adenocarcinoma Data

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

Our dataset comes from Chen et al. (2011), who selected $p = 91$ high-correlating genes representing a broad spectrum of biological functions upon which to build a predictive model. Expression on the log-scale using Affymetrix (a microarray technology, $\boldsymbol{W}$) was measured on 439 tumor samples, and qRT-PCR measurements ($\boldsymbol{X}$) were collected on 47 of these tumors. The individual correlations between the qRT-PCR and Affymetrix

measurements from the 47 tumors are greater than 0.5 across the 91 genes. Clinical covariates, age, gender and stage of cancer (I-III), are also available. Because qRT-PCR is the clinically applicable measurement for future observations, the goal is a qRT-PCR + clinical covariate model for predicting log-survival time after surgery (Y). An independent cohort of 101 tumors with qRT-PCR measurements and clinical covariates is available for validation.

Eleven measurements in the qRT-PCR-only data, out of $47 \times 91 = 4277$ total, or 0.26 percent, were missing; in order to use all observations, these values were imputed using chained equations and thereafter assumed known. Additionally, four tumors, three in the Affymetrix-only sample and one in the validation sample, had event times less than 1 month after surgery, and these were removed before analysis. Thus $n_{\mathrm{A}} = 47$, $n_{\mathrm{B}} = 389$, 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 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: $w_{ij} = \psi_j + \nu_j x_{ij} + \tau\xi_{ij}$. We modeled ψ_j and ν_j as random effects, distributed as $N\{\mu_\psi, \sigma^2_\psi\}$ and $N\{\mu_\nu, \sigma^2_\nu\}$, and used predictions $\{\hat{\psi}_j\}$ and $\{\hat{\nu}_j\}$ to calculate $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$. Violation of the constant τ assumption was also present: gene-specific estimates were in the interval $(0.209, 1.146)$ with the middle 45 in $(0.368, 0.689)$. Consid-

ering all genes simultaneously, $\hat{\tau} = .628$. Because our simulations indicate robustness to this assumption, this violation was ignored.

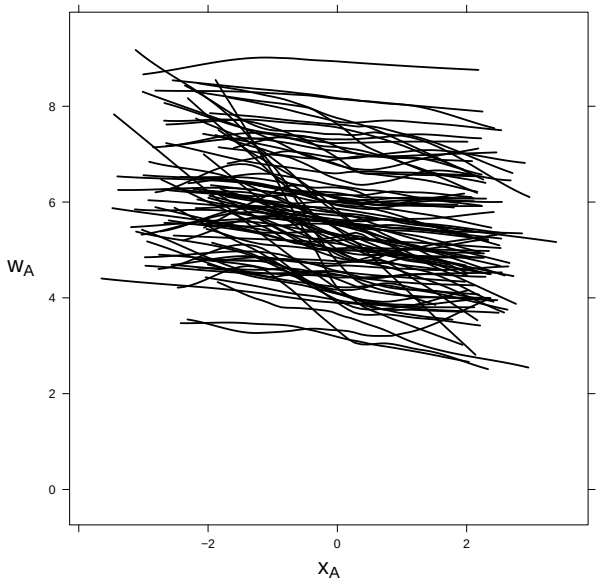

Figure 3: LOESS curves of Affymetrix ($\boldsymbol{w}_{\mathrm{A}}$) by qRT-PCR ($\boldsymbol{x}_{\mathrm{A}}$) measurements for 91 genes from the lung adenocarcinoma data

We present results for predicting survival time in the validation data using RIDG, SRC, FRC, and HYB. Table 2 presents numerical results for each of the methods, and Figure 4 plots each estimate of $\boldsymbol{\beta}$ as a kernel density estimate. In terms of MSPE, the best method was RIDG, with an MSPE of 0.620, compared with 8.745 for SRC and 0.781 for FRC. For HYB, $\hat{\boldsymbol{\omega}}^{\mathrm{opt}} = \{1, 0, 0\}$, corresponding to RIDG, SRC, and FRC; so $\hat{\boldsymbol{\beta}}_{\mathrm{HYB}} \equiv \hat{\boldsymbol{\beta}}_{\mathrm{RIDG}}$ and HYB matches the best of its constituents. Plugging in $\hat{\boldsymbol{\beta}} = \mathbf{0}_p$ yields an MSPE of 0.590, which none of the methods can improve upon, suggesting a very weak signal in the set of expression measures for predicting survival. The range of $\hat{\boldsymbol{\beta}}_{\mathrm{RIDG}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{HYB}}$, excluding the intercept, is $(-0.019, 0.014)$. For $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$, it is $(-0.600, 0.516)$ and for $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$, it is $(-0.075, 0.062)$.

Finally, we generated 95% prediction intervals for each observation in the validation sample, using a bootstrap algorithm described in Appendix A.4. Table 2 gives the proportion of intervals that included the outcome and the average interval ranges. RIDG and

HYB have slight under-coverage (0.91), and SRC and FRC have over-coverage (respectively 1.00 and 0.98).

REMARK 8: As in the simulation study, we restricted our optimization of ω to the subspace of non-negative elements, which on average improves numerical results. In the data analysis, removing the constraint yields $\hat{\omega}^{\text{opt}} = \{1.094, -0.100, 0.006\}$ and an MSPE of 0.601. These results are also presented in Table 2 and Figure 4 denoted as HYB^{unc}.

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

Table 2: Numerical results from analysis of the lung adenocarcinoma data $\hat{\text{MSPE}}$ is the empirical MSPE from the validation sample of size 100, $\min(\hat{\boldsymbol{\beta}})$ and $\max(\hat{\boldsymbol{\beta}})$ give the range of the estimate of $\boldsymbol{\beta}$ for each model, Avg. Coverage is the proportion of bootstrap-generated prediction intervals for the validation sample that contained the true outcome, and $\text{Avg}(\hat{Y}_{\text{new}}^{B,97.5} - \hat{Y}_{\text{new}}^{B,2.5})$ gives the average prediction interval length for the validation sample. HYB^{unc} is the hybrid estimator *without* the non-negativity constraint (Remark 8).

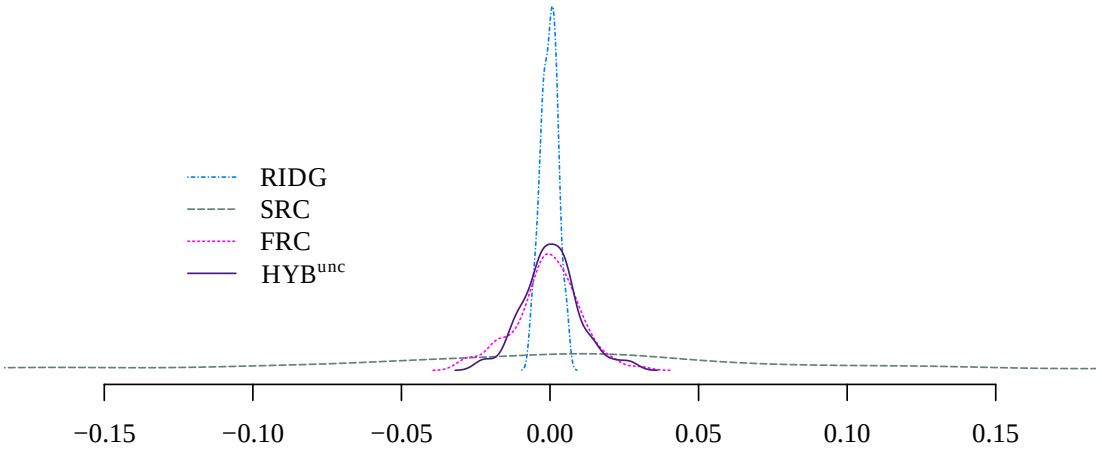

Figure 4: Kernel density estimates of the Chapter 2 methods' 91 coefficient estimates from the lung adenocarcinoma data. HYB^{unc} is the hybrid estimator without the non-negativity constraint (Remark 8). $\hat{\boldsymbol{\beta}}_{\text{HYB}}$, with the non-negativity constraint, is identically equal to $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$.

2.6 Discussion

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

HYB stands out as the method of choice. Theorem 2.1 demonstrates its theoretical utility, and, practically, the average performance of $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ across all design and data configurations is encouraging. Importantly, its flexibility is most apparent in the large p scenarios. Because we combined TR estimators, a GCV criterion provides a simple estimate of $\boldsymbol{P}$, the prediction error matrix (14), which is required to optimize with respect to $\boldsymbol{\omega}$. When taking linear combinations of arbitrary estimates of $\boldsymbol{\beta}$ for which GCV is not conducive, a challenge is how to estimate $\boldsymbol{P}$; the "632 estimator" of Efron (1983) is one candidate.

Our implementation of $\hat{\boldsymbol{\beta}}_{\text{HYB}}$ combined just three estimators: $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$, $\hat{\boldsymbol{\beta}}_{\text{SRC}}$ and $\hat{\boldsymbol{\beta}}_{\text{FRC}}$. Relative to $\hat{\boldsymbol{\beta}}_{\text{RIDG}}$ and $\hat{\boldsymbol{\beta}}_{\text{FRC}}$, $\hat{\boldsymbol{\beta}}_{\text{SRC}}$ predicts unsatisfactorily when $p = 99$. Except for perhaps one panel in Figure 2 (second from the left in the top row), this does not negatively affect the performance of $\hat{\boldsymbol{\beta}}_{\text{HYB}}$, because $\hat{\omega}^{\text{opt}}$ gives little weight to the SRC component. However, this does underscore a practical challenge for the hybrid estimator when combining many $\hat{\boldsymbol{\beta}}$'s, of which a few have much smaller MSPE than the others. Ideally, $\hat{\omega}^{\text{opt}}$ would give little or no weight to the large-MSPE components. We investigated this phenomenon by re-computing the hybrid estimator in the simulation study with

ten ($p = 99$) or three ($p = 5$) additional "SRC-like" TR constituents. We constructed such estimators by sampling, independently of any data, the p components of γ_β from a standard t_5 distribution and the p diagonal components of Ω_β^{-1} from a Gamma distribution with mean n_B and variance $n_B^2/15$. We plugged this randomly sampled prior into (7), using $\lambda = 1$. Compared to the actual hybrid estimator, the increase in prediction error ranged between 2% and 10% in the $p = 99$ cases and 0.5% and 2% in the $p = 5$ cases. This extreme scenario highlights a gap between theoretical optimality and practical implementation but suggests that the performance of the hybrid estimator is relatively unchanged upon the introduction of a large number of ingredients that may not be informative or efficient. It can still discern the better performing constituents, data-adaptively assigning them more weight.

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

That this is a missing data problem can be exploited further than the single imputations considered in this chapter. Multiple imputation using chained equations can make repeated draws of the missing x_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 x_B, β and the rest of the model's parameters. Apart from the computationally demanding aspects of Bayesian methods, because of the size of p and the large fraction of missing data, a fully Bayesian extension

is not automatic. In particular, careful thought must be given to the choice of prior on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_X^{-1}$, because the sampler is likely not to converge with non-informative priors.

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

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

CHAPTER 3

Bayesian Shrinkage Methods for Partially Observed High-Dimensional Data

3.1 Introduction

We consider making predictions of an outcome Y given length-p covariates $\boldsymbol{X}$, using the linear model

$$Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim N\{0,1\}. \tag{15}$$

All observations contain Y and $\boldsymbol{W}$, which is an error-prone length-p surrogate for the true covariate $\boldsymbol{X}$. On a small number of observations of size n_{A}, subsample A, we also observe $\boldsymbol{X}$, which is missing for the remaining subjects, constituting subsample B, of size n_{B}. We assume $n_{\mathrm{A}} < p < n_{\mathrm{B}}$. Complete observations, then, contain an outcome Y, covariates $\boldsymbol{X}$, and surrogates $\boldsymbol{W}$. Subsample A is written as $\{\boldsymbol{y}_{\mathrm{A}}, \boldsymbol{x}_{\mathrm{A}}, \boldsymbol{w}_{\mathrm{A}}\}$ and subsample B as $\{\boldsymbol{y}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}}\}$. The true covariates from subsample B, $\boldsymbol{x}_{\mathrm{B}}$, are unmeasured. The data are schematically presented in Figure 1.

Our goal is a predictive model for $Y|\boldsymbol{X}$ as in (15), but because $\boldsymbol{W}$ is correlated with $\boldsymbol{X}$, subsample B contains information about $\boldsymbol{\beta}$. Moreover, shrinkage of regression coefficients may alleviate problems associated with multicollinearity of covariates. In Chapter, 2, we proposed a class of targeted ridge (TR) estimators of $\boldsymbol{\beta}$, shrinking estimates toward

a target constructed using subsample B, making a bias-variance tradeoff. The amount of shrinkage can be data-adaptive with a tuning parameter, denoted by λ. In our simulation study in Section 2.4, two biased methods, FRC, a regression calibration algorithm, and HYB, a hybrid estimator which is a linear combination of multiple TR estimators with data-adaptive weights, uniformly out-perform structural regression calibration, an unbiased method, in terms of mean-squared prediction error (MSPE):

$$\begin{aligned}\text{MSPE}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) &= \text{E}[(Y_{\text{new}} - \hat{\beta}_0 - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}})^2] \\ &= \sigma^2 + \left(\text{E}[\beta_0 - \hat{\beta}_0 + \boldsymbol{X}_{\text{new}}^\top \boldsymbol{\beta} - \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}]\right)^2 + \text{Var}[\hat{\beta}_0 + \boldsymbol{X}_{\text{new}}^\top \hat{\boldsymbol{\beta}}]. \qquad (16)\end{aligned}$$

However, there are reasons to consider alternative strategies. We show in Appendix A.1 that a TR estimator can be viewed as a missing data technique: make an imputation $\tilde{\boldsymbol{x}}_{\text{B}}$ of the missing $\boldsymbol{x}_{\text{B}}$ and calculate $\hat{\boldsymbol{\beta}}$ treating the data as complete. When the shrinkage is data-adaptive through the tuning parameter λ, there is an intermediate stage: choose λ given $\tilde{\boldsymbol{x}}_{\text{B}}$. Uncertainty in $\tilde{\boldsymbol{x}}_{\text{B}}$ or λ is not propagated in the TR estimators, thus it can be viewed as improper imputation (Little and Rubin, 2002). Moreover, to choose λ, a generalized cross-validation (GCV) criterion was applied to subsample A. Although GCV asymptotically chooses the optimal value of λ (Craven and Wahba, 1979), it can overfit in finite sample sizes, and an approach for estimating λ that also uses information in subsample B is preferred. Finally, constructing prediction intervals corresponding to the point-wise predictions generated by the class of TR estimators requires use of the bootstrap. This resampling process is computationally intensive and provides coverage that may not be nominal.

These reasons, ie characterizing prediction uncertainty and unifying shrinkage, imputation of missing data, and an adaptive choice of λ, motivate a fully Bayesian approach to the same goal of improving predictions using auxiliary data. Consider the generic hierarchical model presented in Figure 5. Known (unknown, respectively) quantities are

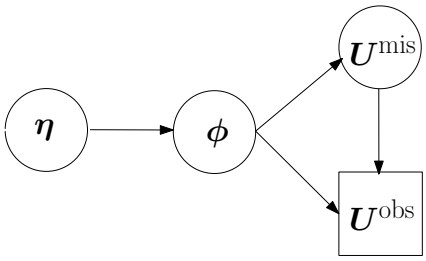

Figure 5: A hierarchical model with missing data $\boldsymbol{U}^{\text{mis}}$ and observed data $\boldsymbol{U}^{\text{obs}}$. The shrinkage penalty parameters $\boldsymbol{\eta}$ are the hyperparameters of $\boldsymbol{\phi}$, the quantity(ies) of primary interest

bounded by square (circular) nodes. Instead of splitting the data into subsamples as in Figure 1, we classify it more broadly into observed ($\boldsymbol{U}^{\text{obs}}$) and missing ($\boldsymbol{U}^{\text{mis}}$) components. Let $\boldsymbol{\phi}$ denote parameters of interest and nuisance parameters in the underlying joint likelihood of $\boldsymbol{U}^{\text{obs}}$ and $\boldsymbol{U}^{\text{mis}}$. Regularization of the high-dimensional $\boldsymbol{\phi}$ is achieved through the shrinkage parameter $\boldsymbol{\eta}$, equivalently interpreted in the figure as the hyperparameters, which index a prior distribution on $\boldsymbol{\phi}$. One can impose another level of hierarchy through a hyperprior distribution on $\boldsymbol{\eta}$. Using $[\cdot]$ and $[\cdot|\cdot]$ to denote marginal and conditional distributions, draws from $[\boldsymbol{U}^{\text{mis}}, \boldsymbol{\phi}, \boldsymbol{\eta}|\boldsymbol{U}^{\text{obs}}]$, the distribution of unknown random quantities conditional on the observed data, constitute proper imputation and incorporate all of the information in the data. Summary values, like posterior means, as well as measurements of uncertainty, like highest posterior density credible intervals and prediction intervals, can easily be calculated based on posterior draws.

Placing the shrinkage parameter $\boldsymbol{\eta}$ in a hierarchical framework allows the flexibility to determine both which components of $\boldsymbol{\phi}$ to shrink and to what extent. As an example of the former, a TR estimator shrink estimates of the regression coefficients $\boldsymbol{\beta}$, tuned by the parameter λ. However, for improved prediction of the outcome Y, it may be beneficial to shrink the parameters generating the missing data x_{B}. In a non-missing-data context, the SCOUT method (Witten and Tibshirani, 2009) shrinks the estimate of $\text{Var}(\boldsymbol{X})$ for better prediction. As for the extent of shrinkage, the hyperparameter-equivalence of the tuning parameters allows for the use of Empirical Bayes algorithms to estimate $\boldsymbol{\eta}$. This has been

used in the Bayesian Lasso (Park and Casella, 2008; Yi and Xu, 2008).

This chapter makes two primary contributions. First, in Section 3.2, we discuss variants of the Gibbs sampler (Geman and Geman, 1984), a key algorithm for fitting hierarchical models with missing data. Here, we keep the context broad, assuming a generic high-dimensional hierarchical model indexed by $\boldsymbol{\phi}$ with missing data $\boldsymbol{U}^{\text{mis}}$ and unspecified hyperparameters $\boldsymbol{\eta}$, as in Figure 5. One variant, Data Augmentation (Tanner and Wong, 1987), is a standard Bayesian approach to missing data, and all unknown quantities have prior distributions. Two others are Empirical Bayes methods: the Monte Carlo Expectation-Maximization algorithm (Wei and Tanner, 1990) and the Empirical Bayes Gibbs sampler (Casella, 2001). Although proposed for seemingly different problems, we argue that the sampling strategies in each are special cases of that in Figure 5: variants of the same general algorithm, which we call EM-within-Gibbs. This framework allows for the chapter's second contribution (Section 3.3): a comparison of several fully Bayesian and Empirical Bayes options and their adaptation in analyses such as ours. Of note in the data are two crucial features: the number of partial observations where $\boldsymbol{X}$ is missing is much larger than the number of complete observations and $\boldsymbol{\phi}$, comprised of β_0, $\boldsymbol{\beta}$, σ^2 plus parameters for modeling the distribution of $\boldsymbol{X}$, is high-dimensional, so that fitting a model with *no* missing data would still be somewhat challenging. Meaningful analysis then requires the regularization, or shrinkage, of $\boldsymbol{\phi}$ via an appropriate specification of the hierarchy and choice of $\boldsymbol{\eta}$. We propose to shrink several different components of $\boldsymbol{\phi}$, making use of the simultaneous interpretation of $\boldsymbol{\eta}$ as a shrinkage penalty and a hyperparameter on $\boldsymbol{\phi}$. We evaluate these methods via a comprehensive simulation study (Section 3.5), also considering robustness of these methods under model misspecification. Finally, we turn to analysis of the Chen et al. dataset, analyzed previously (Section 3.6). Because the likelihood-based approach can offer gains in efficiency but relies more heavily on the underlying modeling assumptions, including the measurement error structure, we include ridge regression (Hoerl and Kennard, 1970) and HYB,

the data-adaptive linear combination of several TR estimators from Chapter 2.

3.2 Gibbs Sampler Variants

In this section, we discuss four existing variants of the Gibbs sampler relevant to our analysis: **DA**, **DA+**, **MCEM**, and **EBGS**. We show that two of these, **MCEM** and **EBGS**, are special cases of a fifth more general variant, which we call **EWiG**. This link between **MCEM** and **EBGS** has not been previously noted. This equivalence leaves three distinct methods: **DA**, **DA+**, and **EWiG**. We define a variant here as characterization of a posterior distribution plus an algorithm for fitting the model. All variants are summarized in Table 3.

Data Augmentation (DA+, DA) (Tanner and Wong, 1987)

$$\textit{Posterior: } [\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}}, \boldsymbol{\eta} | \boldsymbol{U}^{\text{obs}}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{\phi}] \times [\boldsymbol{\phi} | \boldsymbol{\eta}] \times [\boldsymbol{\eta}] \tag{17}$$

Algorithm: at iteration i,

$$\boldsymbol{U}^{\text{mis}(i)} \sim [\boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi}^{(i-1)}] \tag{18}$$

$$\boldsymbol{\phi}^{(i)} \sim [\boldsymbol{\phi} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}(i)}, \boldsymbol{\eta}^{(i-1)}] \tag{19}$$

$$\boldsymbol{\eta}^{(i)} \sim [\boldsymbol{\eta} | \boldsymbol{\phi}^{(i)}]. \tag{20}$$

This is the natural Bayesian treatment of missing data: $\boldsymbol{U}^{\text{mis}}$ and $\boldsymbol{\phi}$ are both unobserved random variables. In **DA+**, which is given above, the hyperparameters $\boldsymbol{\eta}$ are also unknown (Gelfand and Smith, 1990). In **DA**, a value for $\boldsymbol{\eta}$ is chosen. In either case, draws of $\boldsymbol{\phi}$ and $\boldsymbol{U}^{\text{mis}}$ are sequentially made from their conditional posteriors. In **DA+** only, $\boldsymbol{\eta}$ is also sampled from its conditional posterior. Then, in either **DA** or **DA+**, the whole process is iterated. Tanner and Wong prove that iterations will eventually yield a draw from the true posterior distribution of interest, $[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}}, \boldsymbol{\eta} | \boldsymbol{U}^{\text{obs}}]$ for **DA+** or $[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}]$ for **DA**. The full conditional distribution $[\boldsymbol{\phi} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}}, \boldsymbol{\eta}]$ may be difficult to specify. Suppose instead a set of partial conditional distributions is available, $[\boldsymbol{\phi}_J | \boldsymbol{\phi}_{(J)}, \boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}}, \boldsymbol{\eta}]$,

where the set of J's forms a partition of the vector $\boldsymbol{\phi}$. Then under mild conditions, repeated iterative sampling from these partial conditional distributions will also yield draws from the true posterior (Geman and Geman, 1984).

Monte Carlo Expectation-Maximization (MCEM) (Wei and Tanner, 1990)

$$\textit{Posterior: } [\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{\phi}] \times [\boldsymbol{\phi} | \boldsymbol{\eta}] \tag{21}$$

Algorithm: at iteration i,

for $k = 1, \ldots, K$,

$$\boldsymbol{U}^{\text{mis}(i,k)} \sim [\boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi}^{(i-1)}] \tag{22}$$

$$\boldsymbol{\phi}^{(i)} = \text{argmax}_{\boldsymbol{\phi}} \frac{1}{K} \sum_{k=1}^{K} \ln[\boldsymbol{\phi} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}(i,k)}, \boldsymbol{\eta}]. \tag{23}$$

MCEM provides a point estimate of $\boldsymbol{\phi}$ rather than an estimate of the posterior distribution, as with **DA/DA+**. It is a modification of the original EM algorithm (Dempster et al., 1977), replacing an intractable expectation with a Monte Carlo average of multiple imputations. K draws of $\boldsymbol{U}^{\text{mis}}$ are sampled conditional on the current value of $\boldsymbol{\phi}$: $\boldsymbol{\phi}^{(i-1)}$. The expected posterior is updated with a Monte Carlo average and maximized with respect to $\boldsymbol{\phi}$. When $\boldsymbol{\phi}$ has a flat prior, as in the originally proposed **MCEM**, $\{\boldsymbol{\phi}^{(i)}\}$ will converge to the maximum likelihood estimate (MLE) of $\boldsymbol{\phi}$. If an informative prior is specified through a particular choice of $\boldsymbol{\eta}$, the sequence will converge to a *penalized* MLE (Green, 1990).

Empirical Bayes Gibbs Sampling (EBGS) (Casella, 2001)

$$
\begin{aligned}
&\textit{Posterior: } [\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}|\boldsymbol{\phi}] \times [\boldsymbol{\phi}|\boldsymbol{\eta}] \\
&\textit{Algorithm: } \text{at iteration } i, \\
&\qquad \text{for } k = 1, \ldots, K, \\
&\qquad\quad \boldsymbol{\phi}^{(i,k)} \sim [\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}^{(i-1)}] \\
&\qquad \boldsymbol{\eta}^{(i)} = \text{argmax}_{\boldsymbol{\phi}} \frac{1}{K} \sum_{k=1}^{K} \ln[\boldsymbol{\phi}^{(i,k)}|\boldsymbol{\eta}].
\end{aligned}
$$

EBGS allows the data to determine a value for the hyperparameter $\boldsymbol{\eta}$. In the context of Casella, the missing data $\boldsymbol{U}^{\text{mis}}$ are absent. However, $\boldsymbol{\phi}$ is considered missing for purposes of determining $\boldsymbol{\eta}$: choose $\boldsymbol{\eta}$ that maximizes its *marginal* log-likelihood, $\ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\eta}]$. Similar to **MCEM**, an EM-type algorithm can maximize this intractable log-likelihood. K draws of $\boldsymbol{\phi}$ are made from the current estimate of its posterior, and $\boldsymbol{\eta}$ is updated by maximizing a Monte Carlo estimate of $\text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{\eta}]\right]$, where the expectation is over the distribution $[\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}^{(i)}]$. This expected complete-data log-likelihood relates to the desired marginal log-likelihood as follows. First observe that

$$
\begin{aligned}
[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\eta}][\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] &= [\boldsymbol{\phi}|\boldsymbol{\eta}][\boldsymbol{U}^{\text{obs}}|\boldsymbol{\phi}, \boldsymbol{\eta}] \\
&= [\boldsymbol{\phi}|\boldsymbol{\eta}][\boldsymbol{U}^{\text{obs}}|\boldsymbol{\phi}].
\end{aligned}
$$

Let $C = \text{E}\left[\ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\phi}]\right]$, which is constant with respect to $\boldsymbol{\eta}$. Then,

$$
\ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\eta}] = \text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{\eta}]\right] - \text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}]\right] + C.
$$

Because $\text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}]\right] \leq \text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}^{(i)}]\right]$ for any $\boldsymbol{\eta}$, giving this crucial result: maximizing $\text{E}\left[\ln[\boldsymbol{\phi}|\boldsymbol{\eta}]\right]$, or a Monte Carlo approximation thereof, over $\boldsymbol{\eta}$ will increase $\ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\eta}]$ and converge to a local maximum.

EM-within-Gibbs (EWiG)

$$
\begin{aligned}
&\textit{Posterior: } [\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{\phi}] \times [\boldsymbol{\phi} | \boldsymbol{\eta}] \\
&\textit{Algorithm: } \text{at iteration } i, \\
&\qquad \text{for } k = 1, \ldots, K, \\
&\qquad\quad \boldsymbol{U}^{\text{mis}(i,k)} \sim [\boldsymbol{U}^{\text{mis}} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi}^{(i,k-1)}] \\
&\qquad\quad \boldsymbol{\phi}^{(i,k)} \sim [\boldsymbol{\phi} | \boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}(i,k)}, \boldsymbol{\eta}^{(i-1)}] \\
&\qquad \boldsymbol{\eta}^{(i)} = \text{argmax}_{\boldsymbol{\phi}} \frac{1}{K} \sum_{k=1}^{K} \ln[\boldsymbol{\phi}^{(i,k)} | \boldsymbol{\eta}].
\end{aligned}
$$

Importantly, both **MCEM** and **EBGS** allow the lowest level of the hierarchy to be adaptively determined by the data rather than specified a priori. In **MCEM**, this lowest level is $\boldsymbol{\phi}$, and in **EBGS**, it is $\boldsymbol{\eta}$. However, **MCEM** can be expanded in the presence of an unspecified $\boldsymbol{\eta}$ by putting both $\boldsymbol{U}^{\text{mis}}$ and $\boldsymbol{\phi}$ into the imputation step, so $\boldsymbol{\phi}$ is sampled rather than optimized. The maximization step determines $\boldsymbol{\eta}$. This returns to the original goal of **DA+**/**DA**: determining the posterior distribution of $\boldsymbol{\phi}$. Equivalently, we can take the perspective of expanding **EBGS**: add an imputation step for sampling $\boldsymbol{U}^{\text{mis}}$, keeping the maximization step the same. In either case, this yields the same result, which we call EM-within-Gibbs (**EWiG**), given above. Because $\boldsymbol{\eta}$ is unspecified, the hierarchical model here is the same as that given in Figure 5.

In summary, we have asserted that **MCEM** and **EBGS** are special cases of **EWiG**, so we now have three distinct variants, which we apply to our problem in the following section: **DA**, **DA+**, and **EWiG**.

3.3 Specification of the Likelihood and Priors

The discussion so far has been deliberately generic. We now specify a likelihood for our problem of interest, which in turn specifies $\boldsymbol{\phi}$, and apply these Gibbs variants to several

Variant	Posterior	Prior on $\boldsymbol{\eta}$
DA (Tanner and Wong, 1987)	$[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{\phi}] \times [\boldsymbol{\phi} \vert \boldsymbol{\eta}]$	No
DA+ (Gelfand and Smith, 1990)	$[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}}, \boldsymbol{\eta} \vert \boldsymbol{U}^{\text{obs}}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{\phi}] \times [\boldsymbol{\phi} \vert \boldsymbol{\eta}] \times [\boldsymbol{\eta}]$	Yes
MCEM (Wei and Tanner, 1990)	$[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{\phi}] \times [\boldsymbol{\phi} \vert \boldsymbol{\eta}]$	No
EBGS(Casella, 2001)	$[\boldsymbol{\phi} \vert \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}} \vert \boldsymbol{\phi}] \times [\boldsymbol{\phi} \vert \boldsymbol{\eta}]$	No
EWiG	$[\boldsymbol{\phi}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{U}^{\text{obs}}, \boldsymbol{\eta}] \propto [\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} \vert \boldsymbol{\phi}] \times [\boldsymbol{\phi} \vert \boldsymbol{\eta}]$	No

Table 3: A comparison of the general form of the Gibbs sampler variants from Section 3.2 as they were originally proposed. Differences between posteriors depend on the presence of missing data $\boldsymbol{U}^{\text{mis}}$ and whether the hyperparameters $\boldsymbol{\eta}$ are fully specified. Differences in algorithms depend on how the lowest level of the hierarchy, which is unknown, is treated. In particular, **MCEM** differs from **DA** because it returns only an estimate of the posterior mode.

combinations of (i) choices of priors $[\boldsymbol{\phi} \vert \boldsymbol{\eta}]$ and (ii) values of the hyperparameter $\boldsymbol{\eta}$. Translating the quantities in Figure 5 to our problem, we have $\boldsymbol{U}^{\text{obs}} = \{\boldsymbol{y}_{\text{A}}, \boldsymbol{y}_{\text{B}}, \boldsymbol{x}_{\text{A}}, \boldsymbol{w}_{\text{A}}, \boldsymbol{w}_{\text{B}}\}$ and $\boldsymbol{U}^{\text{mis}} = \boldsymbol{x}_{\text{B}}$. A commonly used factorization of the joint likelihood is $[Y, \boldsymbol{X}, \boldsymbol{W}] = [Y \vert \boldsymbol{X}][\boldsymbol{W} \vert \boldsymbol{X}][\boldsymbol{X}]$, which makes a conditional independence assumption $[Y \vert \boldsymbol{X}, \boldsymbol{W}] = [Y \vert \boldsymbol{X}]$. An alternative factorization is $[Y \vert \boldsymbol{X}][\boldsymbol{X} \vert \boldsymbol{W}]$, which we do not consider as it is not consistent with the measurement error structure of $\boldsymbol{W}$ to $\boldsymbol{X}$. We make the following assumptions:

$$Y \vert \boldsymbol{X} \sim N\{\beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2\}, \quad \boldsymbol{W} \vert \boldsymbol{X} \sim N_p\{\psi \mathbf{1}_p + \nu \boldsymbol{X}, \tau^2 \boldsymbol{I}_p\}, \quad \boldsymbol{X} \sim N_p\{\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}. \tag{24}$$

The likelihood has an outcome model relating Y to $\boldsymbol{X}$, a measurement error model relating the error-prone $\boldsymbol{W}$ to $\boldsymbol{X}$, and a multivariate distribution for $\boldsymbol{X}$. Thus,

$$\boldsymbol{\phi} = \{\beta_0, \boldsymbol{\beta}, \sigma, \psi, \nu, \tau, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}.$$

$\boldsymbol{\eta}$ is described below. Of interest is prediction of a new value Y_{new} given $\boldsymbol{X}_{\text{new}}$, eg $\hat{Y}_{\text{new}} = \beta_0^* + \boldsymbol{X}_{\text{new}}^\top \boldsymbol{\beta}^*$, where β_0^* and $\boldsymbol{\beta}^*$ are posterior summaries of β_0 and $\boldsymbol{\beta}$. Uncertainty is quantified using the empirical distribution of $\hat{Y}_{\text{new}}^{(t)} = \beta_0^{(t)} + \boldsymbol{X}_{\text{new}}^\top \boldsymbol{\beta}^{(t)} + \sigma^{2(t)} \varepsilon^{(t)}$, where $\{\beta_0^{(t)}, \boldsymbol{\beta}^{(t)}, \sigma^{2(t)}\}$ is the set of posterior draws and $\varepsilon^{(t)} \overset{\text{iid}}{\sim} N\{0,1\}$. If $\boldsymbol{x}_{\text{B}} = \boldsymbol{U}^{\text{mis}}$ were

observed, the *complete* log-likelihood would be

$$\ell_C = \ln[\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}}|\boldsymbol{\phi}] = \ln[\boldsymbol{y}_{\text{A}}|\boldsymbol{x}_{\text{A}}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_{\text{A}}|\boldsymbol{x}_{\text{A}}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_{\text{A}}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] + \ln[\boldsymbol{y}_{\text{B}}|\boldsymbol{x}_{\text{B}}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_{\text{B}}|\boldsymbol{x}_{\text{B}}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_{\text{B}}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X]. \quad (25)$$

The log-likelihood gives the imputation step:

$$\boldsymbol{x}_{\text{B}}|\boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi} = N_{n_{\text{B}} \times p}\left\{\tilde{\boldsymbol{x}}_{\text{B}}(\boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi}), \boldsymbol{\Gamma}(\boldsymbol{\phi})\right\}, \quad (26)$$

where $\boldsymbol{\Gamma}(\boldsymbol{\phi}) = [\boldsymbol{\beta}\boldsymbol{\beta}^\top/\sigma^2 + (\nu^2/\tau^2)\boldsymbol{I}_p + \boldsymbol{\Sigma}_X^{-1}]^{-1}$ and $\tilde{\boldsymbol{x}}_{\text{B}}(\boldsymbol{U}^{\text{obs}}, \boldsymbol{\phi}) = [(\boldsymbol{y}_{\text{B}} - \beta_0 \mathbf{1}_{n_{\text{B}}})\boldsymbol{\beta}^\top/\sigma^2 + (\nu/\tau^2)(\boldsymbol{w}_{\text{B}} - \psi \mathbf{1}_{n_{\text{B}}}\mathbf{1}_p^\top) + (\mathbf{1}_{n_{\text{B}}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}]\boldsymbol{\Gamma}(\boldsymbol{\phi})$. Note that the mean is an $n_{\text{B}} \times p$ matrix, where each row represents the mean vector corresponding to the $p \times 1$ observation, but the covariance is shared.

REMARK 9: We describe in Appendix A.1 how the performance of SRC and FRC changes with τ^2/ν^2, the noise-to-signal ratio in the ME model. In summary, as τ^2/ν^2 increases, $\hat{\boldsymbol{\beta}}_{\text{SRC}}$ downweights the contribution from the auxiliary data, appropriately so from an imputation perspective, but becomes increasingly instable because it tends to $(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}})^{-1}\boldsymbol{x}_{\text{A}}^\top \boldsymbol{y}_{\text{A}}$ (assuming $\boldsymbol{\mu}_X = \mathbf{0}_p$), which does not exist when $p > n_{\text{A}}$. On the other hand, $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ approaches $\mathbf{0}_p$ as τ^2/ν^2 increases. The imputation step given in (26) resembles that of SRC in that it downweights the contribution from $\boldsymbol{w}_{\text{B}}$ as the noise-to-signal ratio becomes large, but it has two important differences. First, it still makes use of the information about $\boldsymbol{x}_{\text{B}}$ contained in $\boldsymbol{y}_{\text{B}}$, which is invariant to the value of τ^2/ν^2. Second, it is not a mean imputation but rather a random draw from a distribution, so that even if τ^2/ν^2 is very large, the imputation of $\boldsymbol{x}_{\text{B}}$ will be full-rank, an important property for sampling $\boldsymbol{\beta}$. Therefore, even as τ^2/ν^2 increases, the Bayesian methods still efficiently use the available information in subsample B.

The imputation is defined only by the likelihood and is common to all methods we

consider; the differences lie in the choice of prior $[\boldsymbol{\phi}|\boldsymbol{\eta}]$ and the hyperparameter $\boldsymbol{\eta}$. These crucially determine the nature and extent of shrinkage induced on $\boldsymbol{\phi}$. In what follows, we propose several options, which are summarized in Table 4.

FB-FLATBETA As a baseline approach, we apply **DA** to the problem. The choice of prior is

$$[\boldsymbol{\phi}|\boldsymbol{\eta}] \propto (\sigma^2\tau^2)^{-1}|\boldsymbol{\Sigma}_X^{-1}|^{(2p-1)/2} \exp\left\{-\frac{2p-1}{2}\text{Tr}\left(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1}\right)\right\}, \tag{27}$$

where $\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])$ is the diagonal part of the empirical covariance of $\boldsymbol{x}_\text{A}$. This is a Jeffreys prior on each component of $\boldsymbol{\phi}$ except $\boldsymbol{\Sigma}_X^{-1}$ (see Remark 10 below), and $\boldsymbol{\eta}$ is specified. The product of expressions (25) and (27) yields the full conditional distributions of each component of $\boldsymbol{\phi}$. For brevity, we present only the Gibbs steps for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_X^{-1}$. The complete set of full conditional distributions are listed in Appendix B.1.

$$\boldsymbol{\beta} \sim N_p\big\{(\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A} + \boldsymbol{x}_\text{B}^\top\boldsymbol{x}_\text{B})^{-1}(\boldsymbol{x}_\text{A}^\top[\boldsymbol{y}_\text{A} - \beta_0\mathbf{1}_{n_\text{A}}] + \boldsymbol{x}_\text{B}^\top[\boldsymbol{y}_\text{B} - \beta_0\mathbf{1}_{n_\text{B}}]), \sigma^2(\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A} + \boldsymbol{x}_\text{B}^\top\boldsymbol{x}_\text{B})^{-1}\big\}, \tag{28}$$

$$\begin{aligned}\boldsymbol{\Sigma}_X^{-1} \sim W\Big\{&3p + n_\text{A} + n_\text{B}, \\ &\Big((2p-1)\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}]) \\ &\quad + (\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top) + (\boldsymbol{x}_\text{B} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{B} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)\Big)^{-1}\Big\}.\end{aligned} \tag{29}$$

The Wishart distribution with d degrees of freedom, $W\{d, \boldsymbol{S}\}$, is parametrized to have mean $d\boldsymbol{S}$.

REMARK 10: A Jeffreys prior on $\boldsymbol{\Sigma}_X^{-1}$, $\boldsymbol{\Sigma}_X^{-1} \sim W\{0, 0\boldsymbol{I}_p\}$, may result in an improper joint posterior if $n_\text{B} \gg n_\text{A}$ and p is large, ie when the fraction of missing data is very large. From our numerical studies and monitoring of trace plots, even a minimally proper prior on $\boldsymbol{\Sigma}_X^{-1}$, that is, using $p+1$ degrees of freedom, does not ensure a proper posterior. We assume a priori $\boldsymbol{\Sigma}_X^{-1} \sim W\left\{3p, (2p-1)^{-1}[\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])]^{-1}\right\}$, a data-driven choice, the density of which is given in (27). The prior mean of $\boldsymbol{\Sigma}_X^{-1}$ is $3p[\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])]^{-1}/(2p-1)$,

and the prior mean of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is $\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_{\text{A}}])$. Heuristic numeric evidence shows that $3p$ degrees of freedom works well, but we have not demonstrated a theoretical optimality for this choice. Other values that ensure convergence are equally defensible.

We call the Gibbs sampler that uses this mildly informative prior specification FB-FLATBETA. All the other methods we propose will have modified Gibbs steps for two components of $\boldsymbol{\phi}$: $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$. Shrinking $\boldsymbol{\beta}$ is a clear choice: from (24), $\boldsymbol{\beta}$ is closely tied to prediction of $Y|\boldsymbol{X}$. As for $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$, this determines in part the posterior variance of $\boldsymbol{x}_{\text{B}}$ (26); as this variance increases, the a posteriori variance of $\boldsymbol{\beta}$ decreases (28), thereby shrinking draws $\boldsymbol{\beta}$. Other factors in the variance of $\boldsymbol{x}_{\text{B}}$, like τ^2, are additional candidates for shrinkage, but we do not pursue this here.

3.3.1 Adaptive Prior on $\boldsymbol{\beta}$

Since we are interested in regularizing predictions of the outcome Y, a natural candidate for shrinkage via an informative prior is the parameter vector $\boldsymbol{\beta}$, which specifies the conditional mean of $Y|\boldsymbol{X}$. Ridge regression offers favorable predictive capabilities (Frank and Friedman, 1993), and the ℓ_2 penalty on the norm of $\boldsymbol{\beta}$ is conjugate to the normal log-likelihood. For these reasons, we replace the Jeffreys prior on $\boldsymbol{\beta}$ in (27) with

$$[\boldsymbol{\beta}|\sigma^2, \lambda] \propto \left(\frac{\lambda}{\sigma^2}\right)^{p/2} \exp\left\{-\frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}. \tag{30}$$

This Normal prior on $\boldsymbol{\beta}$ is analogous to Bayesian ridge regression. λ is a hyperparameter, ie $\boldsymbol{\eta} = \{\lambda\}$. Conditional upon λ, the Gibbs step for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim N_p\left\{(\boldsymbol{x}_{\text{A}}^\top\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}{}^\top\boldsymbol{x}_{\text{B}} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{x}_{\text{A}}^\top\boldsymbol{y}_{\text{A}} + \boldsymbol{x}_{\text{B}}{}^\top\boldsymbol{y}_{\text{B}}), \sigma^2(\boldsymbol{x}_{\text{A}}^\top\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}{}^\top\boldsymbol{x}_{\text{B}} + \lambda\boldsymbol{I}_p)^{-1}\right\}.$$

Thus, the posterior mean of $\boldsymbol{\beta}$ is shrunk toward zero and with smaller posterior variance. As we have outlined, there are several options for the treatment of λ.

FB-HIBETA-NI Following Gelfand and Smith (1990), we can treat the hyperparameter λ as random (**DA+**) with prior distribution $[\lambda] \propto \lambda^{-1}$. Then, we have the following additional posterior step: $\lambda \sim G\left\{p/2, \boldsymbol{\beta}^\top \boldsymbol{\beta}/(2\sigma^2)\right\}$. This Bayesian ridge regression with posterior sampling of λ is denoted by FB-HIBETA-NI.

EB-HIBETA-NI Alternatively, we may apply **EWiG** to estimate λ. That is, integrate $\log[\boldsymbol{\beta}|\sigma^2, \lambda]$ with respect to the density $[\boldsymbol{\phi}|\boldsymbol{U}^{\text{obs}}, \lambda]$, differentiate with respect to λ, and solve for λ. The resulting **EWiG** update is $\lambda \leftarrow p\left[(1/K)\sum_{k=1}^{K} \boldsymbol{\beta}^{(k)\top}\boldsymbol{\beta}^{(k)}/\sigma^{2(k)}\right]^{-1}$. This is a Monte Carlo estimate of $p\left\{\mathrm{E}\left[(\boldsymbol{\beta}^\top\boldsymbol{\beta})/(\sigma^2)\right]\right\}^{-1}$, the maximum of the marginal likelihood of λ. The update occurs at every Kth iteration of the algorithm using the previous K draws of $\boldsymbol{\beta}$ and σ^2; larger values of K yield a more precise estimate. This Bayesian ridge with an Empirical Bayes update of λ is denoted by EB-HIBETA-NI.

REMARK 11: Although the Bayesian ridge imposes a ℓ_2-type penalty similar to the TR class of estimators, there is an important distinction. TR estimators make use of the surrogate data in subsample B by deriving pre-specified shrinkage targets. Here, the information in the surrogate data has already been used in the imputation step for $\boldsymbol{x}_{\text{B}}$; the Bayesian ridge shrinks $\boldsymbol{\beta}$ toward *zero* to regularize the conditional sampling distribution and therefore predictions.

3.3.2 Adaptive Prior on $\boldsymbol{\Sigma}_X^{-1}$

(EB-HISIGMAX, EB-HIBETASIGMAX) We noted previously that an informative prior on $\boldsymbol{\Sigma}_X^{-1}$ is necessary to ensure a proper joint posterior: $\boldsymbol{\Sigma}_X^{-1} \sim W\left\{3p, (2p-1)^{-1}[\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_{\text{A}}])]^{-1}\right\}$, which has inverse scale matrix $(2p-1)\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_{\text{A}}])$. As we have noted, shrinkage of $\boldsymbol{\Sigma}_X^{-1}$ is closely related to that of $\boldsymbol{\beta}$. This was exploited by Witten and Tibshirani (2009) in the SCOUT procedure, suggesting that prediction can be improved through adaptive

regularization of $\boldsymbol{\Sigma}_X^{-1}$. Leaving the inverse scale matrix unspecified, the prior is

$$[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}] \propto |\boldsymbol{\Lambda}|^{3p/2}|\boldsymbol{\Sigma}_X^{-1}|^{(2p-1)/2}\exp\left\{-(1/2)\text{Tr}\,(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{-1})\right\}. \tag{31}$$

$\boldsymbol{\Lambda}$ is the unknown positive-definite matrix of hyperparameters. The full conditional distribution of $\boldsymbol{\Sigma}_X^{-1}$ changes according to:

$$\boldsymbol{\Sigma}_X^{-1} \sim W\Big\{3p + n_\text{A} + n_\text{B}, \\ \Big(\boldsymbol{\Lambda} + (\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top) + (\boldsymbol{x}_\text{B} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{B} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)\Big)^{-1}\Big\}.$$

$\boldsymbol{\Lambda}$ may be random, or it can be updated with an **EWiG** step. Given the potential difficulty in precisely estimating a fully unconstrained matrix that maximizes the marginal likelihood, we constrain $\boldsymbol{\Lambda}$ to be diagonal. Under this constraint, the **EWiG** update for the ith diagonal of $\boldsymbol{\Lambda}$ is $\Lambda_{ii} \leftarrow 3p\left((1/K)\sum_{k=1}^K \boldsymbol{\Sigma}_X^{-1\,(k)}_{(ii)}\right)^{-1}$. $\boldsymbol{\Sigma}_X^{-1}{}_{(ii)}$ indicates the ith diagonal element of $\boldsymbol{\Sigma}_X^{-1}$. Then, $\boldsymbol{\Lambda} = \text{diag}\{\Lambda_{11}, \ldots, \Lambda_{pp}\}$. This is a Monte Carlo approximation of $3p\text{diag}\{\text{E}[\boldsymbol{\Sigma}_X^{-1}]_{11}, \ldots, \text{E}[\boldsymbol{\Sigma}_X^{-1}]_{pp}\}^{-1}$, the minimizer of $\text{E}\left[\log[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}]\right]$ with respect to $\boldsymbol{\Lambda}$, subject to the diagonal constraint, where $[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}]$ is given in (31). We will refer to this approach as EB-HISIGMAX. Finally, we call the approach that uses both priors in (30) and (31) with **EWiG** updates for λ and $\boldsymbol{\Lambda}$ EB-HIBETASIGMAX. All of these alternatives are summarized in Table 4.

REMARK 12: Adaptively estimating the diagonal inverse scale matrix parameter $\boldsymbol{\Lambda}$ modifies the variance components of $\boldsymbol{X}$. Alternatively, one might apply an **EWiG** update to the degrees of freedom parameter, say d, which modifies the partial correlations of $\boldsymbol{X}$. For example, when $d = p + 1$, the induced prior on each partial correlation is uniform on $[-1, 1]$ (Gelman and Hill, 2006). Larger values of d place more prior mass closer to zero. Allowing the data to specify d is a reasonable alternative; however, we encountered numerical difficulties in implementing this approach. The **EWiG** update

cannot be expressed in closed form and must be estimated numerically. Additionally, the "complete-data log-likelihood" in the M-step is often flat, such that a wide range of values for d will return nearly equivalent log-likelihoods.

Method	$[\boldsymbol{\beta}\|\boldsymbol{\eta}] \propto$	$[\boldsymbol{\Sigma}_X^{-1}\|\boldsymbol{\eta}] \propto \|\boldsymbol{\Sigma}_X^{-1}\|^{(2p-1)/2} \times$	$\boldsymbol{\eta}$	Variant
FB-FLATBETA	1	$\exp\left\{-\frac{2p-1}{2}\text{Tr}\left(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1}\right)\right\}$	$\{\}$	**DA**
FB-HIBETA-NI	$\left(\frac{\lambda}{\sigma^2}\right)^{p/2}\exp\left\{-\frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}$	$\exp\left\{-\frac{2p-1}{2}\text{Tr}\left(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1}\right)\right\}$	$\{\lambda\}$	**DA+**
EB-HIBETA-NI	$\left(\frac{\lambda}{\sigma^2}\right)^{p/2}\exp\left\{-\frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}$	$\exp\left\{-\frac{2p-1}{2}\text{Tr}\left(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1}\right)\right\}$	$\{\lambda\}$	**EWiG**
EB-HISIGMAX	1	$\|\boldsymbol{\Lambda}\|^{3p/2}\exp\left\{-(1/2)\text{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{-1}\right)\right\}$	$\{\boldsymbol{\Lambda}\}$	**EWiG**
EB-HIBETASIGMAX	$\left(\frac{\lambda}{\sigma^2}\right)^{p/2}\exp\left\{-\frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}$	$\|\boldsymbol{\Lambda}\|^{3p/2}\exp\left\{-(1/2)\text{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{-1}\right)\right\}$	$\{\lambda, \boldsymbol{\Lambda}\}$	**EWiG**

Table 4: A summary of all Gibbs samplers and choices of priors considered. $\boldsymbol{\Lambda}$ is constrained to the class of diagonal matrices.

3.4 Estimation Under Predictive Loss

A fitted model may be summarized by measures of uncertainty, eg a posterior predictive interval $(\hat{Y}_\text{new}^{p_L}, \hat{Y}_\text{new}^{p_H})$, as well as point predictions, $\hat{Y}_\text{new} = \beta_0^* + \boldsymbol{X}_\text{new}^\top\boldsymbol{\beta}^*$ using summary values β_0^* and $\boldsymbol{\beta}^*$. These are calculated with draws from the posterior distribution, $\{\boldsymbol{\phi}^{(t)}\}$. Predictive intervals are given by empirical quantiles of $\{Y_\text{new}^{(t)}\}$, where $\hat{Y}_\text{new}^{(t)} = \beta_0^{(t)} + \boldsymbol{X}_\text{new}^\top\boldsymbol{\beta}^{(t)} + \sigma^{2(t)}\varepsilon^{(t)}$ and $\varepsilon^{(t)} \overset{\text{iid}}{\sim} N\{0,1\}$. For point predictions, a summary value of β_0 is given by $\hat{\beta}_0 = (1/T)\sum_t \beta_0^{(t)}$. For $\boldsymbol{\beta}$, we minimize posterior predictive loss. We define the posterior predictive mean by $\boldsymbol{\beta}^\text{ppm} = \text{argmin}_{\boldsymbol{b}}\text{E}_{\boldsymbol{\phi}, X_\text{new}|\boldsymbol{U}_\text{obs}}(\boldsymbol{\beta}-\boldsymbol{b})^\top\boldsymbol{X}_\text{new}\boldsymbol{X}_\text{new}^\top(\boldsymbol{\beta}-\boldsymbol{b})$. This is in contrast to the posterior mean: $\boldsymbol{\beta}^\text{pm} = \text{argmin}_{\boldsymbol{b}}\text{E}_{\beta|\boldsymbol{U}_\text{obs}}(\boldsymbol{\beta}-\boldsymbol{b})^\top(\boldsymbol{\beta}-\boldsymbol{b})$. Estimates of both quantities are given by

$$\hat{\boldsymbol{\beta}}^\text{ppm} = \sum_t\left(\boldsymbol{\Sigma}_X^{(t)} + \boldsymbol{\mu}_X^{(t)}\boldsymbol{\mu}_X^{(t)\top}\right)^{-1}\sum_t\left(\boldsymbol{\Sigma}_X^{(t)} + \boldsymbol{\mu}_X^{(t)}\boldsymbol{\mu}_X^{(t)\top}\right)\boldsymbol{\beta}^{(t)}, \tag{32}$$

$$\hat{\boldsymbol{\beta}}^\text{pm} = (1/T)\sum_t\boldsymbol{\beta}^{(t)}. \tag{33}$$

To summarize, different posterior summaries of $\boldsymbol{\beta}$ come from minimizing different loss functions; we have two estimates of $\boldsymbol{\beta}$ for each method and, as a consequence, two choices of point predictions for Y_{new}. In contrast, we have only one posterior predictive interval, that derived from the empirical quantiles of $\{Y_{\text{new}}^{(t)}\}$.

3.5 Simulation Study

We conducted a simulation study to evaluate these methods. At first, the assumed model of the data satisfied the generating model, as given in (24). We fixed $n_{\text{A}} = 50$ and used $n_{\text{B}} \in \{400, 150\}$. The diagonal elements of $\boldsymbol{\Sigma_X}$ 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 $N\{\mathbf{0}_p, \boldsymbol{\Sigma_X}\}$. We considered both high ($p = 99$) and low ($p = 5$) dimensional models: $\boldsymbol{\beta} = \{j/100\}_{j=-49}^{j=49}$ and $\boldsymbol{\beta} = \{j/4\}_{j=-2}^{j=2}$. R^2 values were either 0.1 or 0.4. Thus given $\boldsymbol{\beta}$, $\boldsymbol{\Sigma_X}$ and R^2, σ^2 was determined by solving $\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} + \sigma^2) = R^2$. β_0 was set to zero. $\boldsymbol{y}_{\text{A}} | \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{y}_{\text{B}} | \boldsymbol{x}_{\text{B}}$ were drawn for each combination of $\boldsymbol{\beta}$ and σ^2. This yielded 16 unique simulation settings: two choices each for p, n_{B}, ρ, and R^2. To draw the auxiliary data, we set $\psi = 0$ and $\nu = 1$ and repeated each of the 16 settings for $\tau \in (0, 2)$, drawing $\boldsymbol{w}_{\text{A}} | \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{w}_{\text{B}} | \boldsymbol{x}_{\text{B}}$ based on the measurement error model in (24).

After a burn-in period of 2500, we stored 1000 posterior draws. We calculated $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ (32) and $\hat{\boldsymbol{\beta}}^{\text{pm}}$ (33). For FB-FLATBETA, FB-HIBETA-NI, EB-HIBETA-NI, EB-HISIGMAX, EB-HIBETASIGMAX, we estimated the MSPE of $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ on 1000 new observations:

$$\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^{\text{ppm}}) = (1/1000) \sum_{j=1}^{1000} (Y_{\text{new},j} - \hat{\beta}_0 - \boldsymbol{X}_{\text{new},j}^\top \hat{\boldsymbol{\beta}}^{\text{ppm}})^2.$$

$\{Y_{\text{new},j}, \boldsymbol{X}_{\text{new},j}\}$ are resampled from the same generating distribution for each simulation. As a comparison, we fit a ridge regression (RIDG) on subsample A only, choosing the tuning parameter with the GCV function, and HYB, the flexible hybrid-type estimator

described in Section 2.3. Figure 6, the primary results of the study, plots $\widehat{\text{MSPE}}$ averaged over 250 simulated datasets over τ. We also estimated MSPE using $\hat{\boldsymbol{\beta}}^{\text{pm}}$. Numerical values are given in Tables B1 and B2. RIDG and HYB yield only one estimate of $\boldsymbol{\beta}$, and are just repeated in the $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ and $\hat{\boldsymbol{\beta}}^{\text{pm}}$ columns. Finally, we computed prediction intervals for the new observations (Section 3.4). Although frequentist in nature, it is still desirable for a Bayesian prediction interval to achieve nominal coverage ; the average coverage rates of $Y_{\text{new},j}$ are given in Table 5.

From Figure 6, when p is large, EB-HIBETASIGMAX uniformly dominates all other methods except EB-HIBETA-NI, which is nearly identical and not plotted. All of the Bayesian methods are similar when p is small; FB-FLATBETA and FB-HIBETA-NI are best by a small margin when $R^2 = 0.1$. EB-HISIGMAX, which corresponds to shrinkage on $\boldsymbol{\Sigma}_X^{-1}$ alone, does a poor job for large p, and FB-FLATBETA predicts only slightly better. RIDG and HYB do not improve upon the best-performing Bayesian methods here. EB-HIBETASIGMAX and EB-HIBETA-NI vary little over the values of τ we evaluated.

Coverage Properties EB-HIBETA-NI and EB-HIBETASIGMAX are the only two methods that maintain close-to-nominal coverage under all scenarios (Table 5). In contrast, larger values of τ considerably decrease the coverage of FB-FLATBETA, FB-HIBETA-NI and EB-HISIGMAX.

Mean Squared Error. The results discussed above and reported in Figure 6 use the estimate of $\boldsymbol{\beta}$ that minimizes predictive loss, that is $\hat{\boldsymbol{\beta}}^{\text{ppm}}$, and are evaluated by MSPE. If instead we use $\widehat{\text{MSE}}(\hat{\boldsymbol{\beta}}^{\text{ppm}})$ or $\widehat{\text{MSE}}(\hat{\boldsymbol{\beta}}^{\text{pm}})$, EB-HIBETA-NI and EB-HIBETASIGMAX remain the preferred methods (results not given).

Violations to Modeling Assumptions As we have noted, these likelihood-based approaches depend on the assumed model approximately matching the true generating model. We inspected robustness through violations to the model assumptions: (i) the distribution of ε is skewed, shifted to maintain a zero mean: $\varepsilon + 1 \sim G\{1, 1\}$, (ii) the measure-

ment error model is misspecified: $\boldsymbol{W}|\boldsymbol{X} \sim N_p\{\psi\mathbf{1}_p + \nu\boldsymbol{X}^2, \tau^2\boldsymbol{I}_p\}$, where we use $\boldsymbol{X}^2$ to denote the element-wise square, or (iii) $\boldsymbol{X}$ comes from a mixture of distributions: $\boldsymbol{X}|Z \sim N_p\{1_{[Z=2]}(3 \times \mathbf{1}_p) - 1_{[Z=3]}(3 \times \mathbf{1}_p), \boldsymbol{\Sigma}_{\boldsymbol{X}}\}$, where $1_{[\cdot]}$ is the indicator function and $Z \overset{iid}{\sim} \text{Unif}\{1, 2, 3\}$.

The results of these modeling violations are given in Tables B3–B8. When ε is skewed (B3,B4), the rankings change little; EB-HIBETA-NI and EB-HIBETASIGMAX are preferred for large p, and all methods are similar for small p. The case is similar for the misspecified measurement error model (B5,B6). Interestingly, when $\boldsymbol{X}$ comes from a mixture of normals (B7,B8), and when p is large, FB-HIBETA-NI often has smaller MSPE than EB-HIBETA-NI and EB-HIBETASIGMAX. Comparing this to Table B1, which has no modeling violations, this is due to a decrease in MSPE for FB-HIBETA-NI rather than increases in MSPE for EB-HIBETA-NI and EB-HIBETASIGMAX. For example, in the MSPE($\hat{\boldsymbol{\beta}}^{\text{PPM}}$) columns of the sixth block down, EB-HIBETA-NI has values 85.0, 85.5, 86.1, 88.4, and 89.3 in Table B1 and values 85.0, 85.5, 87.1, 88.8, and 92.0 in Table B7. However, FB-HIBETA-NI has, respectively, 94.2, 93.1, 93.1, 94.3, and 94.7 (B1) and 84.3, 84.8, 86.6, 87.7, and 90.0 (B7). In summary, EB-HIBETA-NI/EB-HIBETASIGMAX are moderately robust to several strong modeling violations, and the fully Bayesian FB-HIBETA-NI may even have *decreased* MSPE under one of these violations.

3.6 Example: Lung Adenocarcinoma Data

As in Chapter 2, we consider whether gene expression measurements can help predict survival time in lung cancer patients, using data 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 using Affymetrix on the log-scale was measured on 439 tumor samples, and quantitative real-time polyermase chain reaction (qRT-PCR) measurements were collected on a subset of 47 of these. Individual correla-

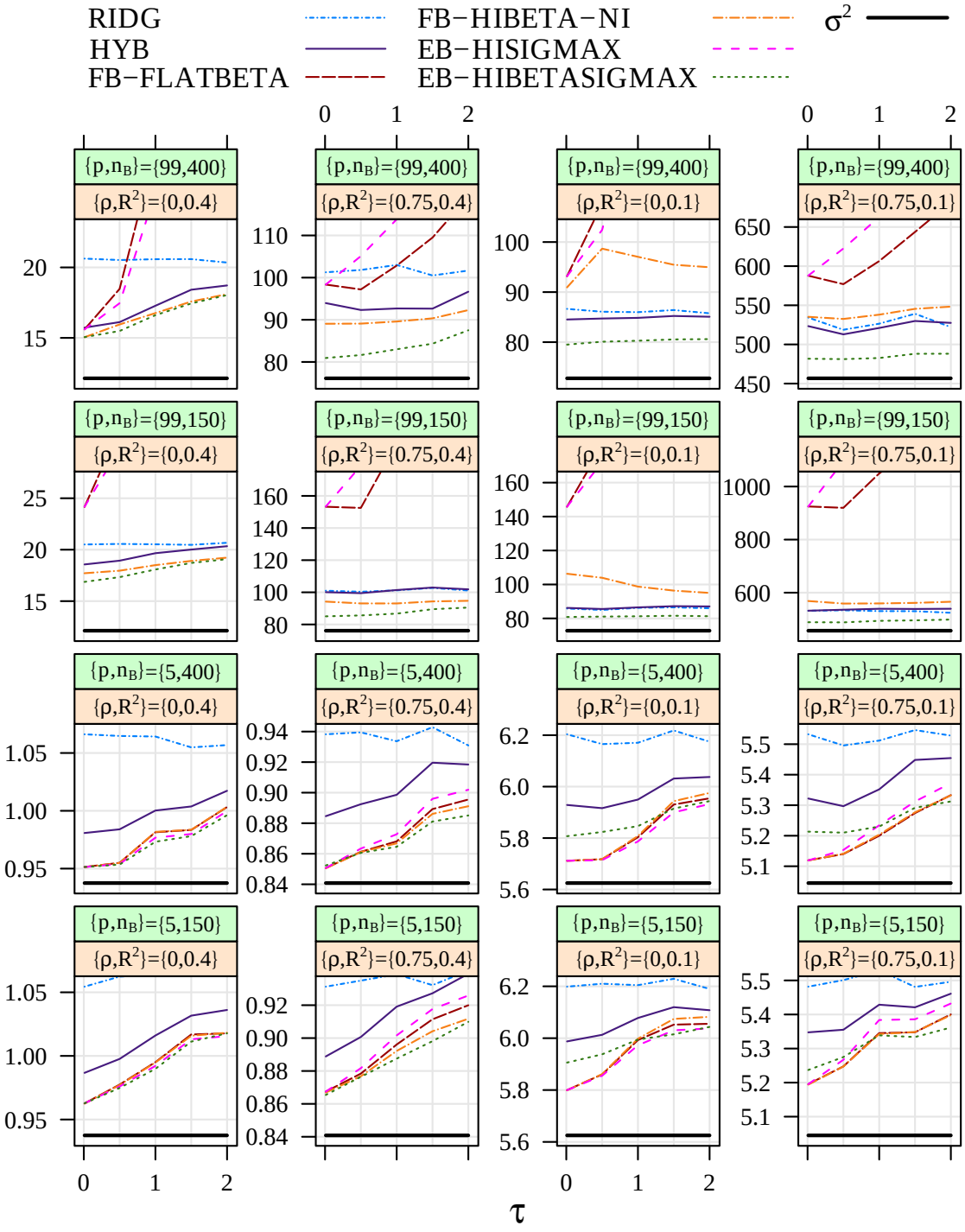

Figure 6: Empirical MSPE over τ for 16 simulation settings described in Section 3.5. For each method, β was estimated with $\hat{\beta}^{\text{ppm}}$ based on 250 independent training datasets, and MSPE was estimated on 250 validation datasets of size 1000. The thick, solid bar (σ^2) corresponds to predictions made using the true generating parameters.

			τ							τ				
$\{\rho, R^2\}$	$\{p, n_B\}$	Method	0.01	0.5	1.0	1.5	2.0	$\{p, n_B\}$	Method	0.01	0.5	1.0	1.5	2.0
0,0.4	99,400	FB-FLATBETA	**949**	**925**	760	518	474	5,400	FB-FLATBETA	**948**	**946**	**939**	**940**	**938**
		FB-HIBETA-NI	**928**	**926**	**933**	**937**	**937**		FB-HIBETA-NI	**947**	**945**	**937**	**936**	**934**
		EB-HIBETA-NI	**949**	**949**	**950**	**951**	**950**		EB-HIBETA-NI	**948**	**947**	**943**	**944**	**944**
		EB-HISIGMAX	**949**	**930**	809	550	474		EB-HISIGMAX	**948**	**947**	**941**	**940**	**938**
		EB-HIBETASIGMAX	**949**	**948**	**948**	**949**	**949**		EB-HIBETASIGMAX	**948**	**948**	**944**	**944**	**943**
0.75,0.4	99,400	FB-FLATBETA	**948**	**922**	882	854	813	5,400	FB-FLATBETA	**948**	**946**	**947**	**943**	**941**
		FB-HIBETA-NI	**930**	910	898	898	899		FB-HIBETA-NI	**946**	**945**	**945**	**942**	**939**
		EB-HIBETA-NI	**948**	**943**	**941**	**939**	**935**		EB-HIBETA-NI	**948**	**947**	**950**	**949**	**947**
		EB-HISIGMAX	**948**	**920**	874	811	721		EB-HISIGMAX	**948**	**945**	**945**	**943**	**941**
		EB-HIBETASIGMAX	**948**	**946**	**943**	**942**	**938**		EB-HIBETASIGMAX	**947**	**946**	**949**	**949**	**948**
0,0.1	99,400	FB-FLATBETA	**949**	**932**	846	592	490	5,400	FB-FLATBETA	**948**	**948**	**945**	**940**	**939**
		FB-HIBETA-NI	919	908	912	913	914		FB-HIBETA-NI	**947**	**947**	**943**	**937**	**937**
		EB-HIBETA-NI	**947**	**949**	**949**	**948**	**949**		EB-HIBETA-NI	**948**	**948**	**947**	**945**	**946**
		EB-HISIGMAX	**949**	**934**	878	628	521		EB-HISIGMAX	**948**	**948**	**945**	**940**	**940**
		EB-HIBETASIGMAX	**948**	**948**	**949**	**948**	**948**		EB-HIBETASIGMAX	**948**	**948**	**947**	**945**	**946**
0.75,0.1	99,400	FB-FLATBETA	**948**	**924**	884	846	800	5,400	FB-FLATBETA	**948**	**947**	**943**	**939**	**936**
		FB-HIBETA-NI	**928**	908	892	886	883		FB-HIBETA-NI	**947**	**945**	**942**	**938**	**934**
		EB-HIBETA-NI	**948**	**947**	**945**	**944**	**943**		EB-HIBETA-NI	**948**	**947**	**947**	**946**	**945**
		EB-HISIGMAX	**948**	**923**	879	812	723		EB-HISIGMAX	**948**	**946**	**943**	**940**	**937**
		EB-HIBETASIGMAX	**948**	**948**	**946**	**945**	**945**		EB-HIBETASIGMAX	**948**	**947**	**947**	**946**	**946**
0,0.4	99,150	FB-FLATBETA	**948**	918	493	442	411	5,150	FB-FLATBETA	**948**	**946**	**942**	**943**	**944**
		FB-HIBETA-NI	919	**931**	**939**	**940**	**938**		FB-HIBETA-NI	**945**	**943**	**938**	**937**	**939**
		EB-HIBETA-NI	**946**	**949**	**949**	**949**	**945**		EB-HIBETA-NI	**948**	**947**	**946**	**947**	**948**
		EB-HISIGMAX	**948**	**922**	468	494	453		EB-HISIGMAX	**948**	**947**	**943**	**942**	**943**
		EB-HIBETASIGMAX	**946**	**947**	**947**	**947**	**944**		EB-HIBETASIGMAX	**948**	**948**	**946**	**946**	**947**
0.75,0.4	99,150	FB-FLATBETA	**947**	880	653	619	548	5,150	FB-FLATBETA	**948**	**946**	**944**	**942**	**941**
		FB-HIBETA-NI	**925**	911	909	911	916		FB-HIBETA-NI	**945**	**943**	**941**	**940**	**938**
		EB-HIBETA-NI	**947**	**942**	**940**	**937**	**939**		EB-HIBETA-NI	**947**	**948**	**948**	**947**	**946**
		EB-HISIGMAX	**947**	854	565	531	461		EB-HISIGMAX	**948**	**946**	**944**	**942**	**942**
		EB-HIBETASIGMAX	**947**	**945**	**943**	**941**	**943**		EB-HIBETASIGMAX	**948**	**947**	**947**	**948**	**947**
0,0.1	99,150	FB-FLATBETA	**949**	**922**	491	455	399	5,150	FB-FLATBETA	**947**	**946**	**943**	**939**	**942**
		FB-HIBETA-NI	899	907	918	**920**	**922**		FB-HIBETA-NI	**944**	**943**	**939**	**934**	**936**
		EB-HIBETA-NI	**946**	**947**	**948**	**947**	**949**		EB-HIBETA-NI	**946**	**946**	**946**	**944**	**946**
		EB-HISIGMAX	**948**	**923**	546	470	443		EB-HISIGMAX	**947**	**946**	**944**	**939**	**941**
		EB-HIBETASIGMAX	**946**	**947**	**947**	**947**	**948**		EB-HIBETASIGMAX	**946**	**946**	**946**	**944**	**947**
0.75,0.1	99,150	FB-FLATBETA	**946**	879	714	616	600	5,150	FB-FLATBETA	**949**	**945**	**942**	**940**	**938**
		FB-HIBETA-NI	**921**	906	897	895	894		FB-HIBETA-NI	**947**	**942**	**938**	**936**	**933**
		EB-HIBETA-NI	**947**	**947**	**945**	**944**	**944**		EB-HIBETA-NI	**949**	**946**	**946**	**947**	**945**
		EB-HISIGMAX	**946**	834	581	533	519		EB-HISIGMAX	**950**	**944**	**942**	**941**	**939**
		EB-HIBETASIGMAX	**947**	**948**	**946**	**946**	**946**		EB-HIBETASIGMAX	**949**	**946**	**946**	**947**	**945**

Table 5: Results from the simulation study described in 3.5: average coverage rates $\times 1000$ corresponding to prediction of new observations, nominally 0.95×1000, for each Bayesian method. Specifically, each value is coverage averaged over 250 simulations, 1000 new observations per simulation. Values that lie in the interval (920, 980) are in **bold**

tions between the qRT-PCR and Affymetrix measurements from the 47 tumors exceed 0.5 across all 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.

11 measurements in the qRT-PCR-only data, out of $47 \times 91 = 4277$ total, or 0.26 percent, were missing; in order to use all observations, these values were imputed using chained equations and thereafter assumed known. Bayesian methods of imputation, like those discussed for imputing $\boldsymbol{x}_B$, are a better approach to handle this missingness but, given the small percentage of missingness, would likely not affect the results. Additionally, four tumors, three in the Affymetrix-only sample and one in the validation sample, had event times less than one month after surgery; these were removed before analysis. Thus $n_A = 47$, $n_B = 389$, and the validation sample is size 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 was fit to the training data but was applied to censored survival times in both the training and validation data. Given completed log-survival times, we re-fit this same model and calculated residuals from both the training and validation data. These residuals were considered as outcomes, and the question is whether any additional variation in the residuals is explained by gene expression. While there are more appropriate ways of dealing with coarsened data and additional covariates in the likelihood-based framework, processing the data this way allowed for a straightforward comparison with the non-likelihood-based methods of RIDG and HYB. Finally, in Chapter 2, we used a gene-specific ME model: $w_{ij} = \psi_j + \nu_j x_{ij} + \tau\tilde{\xi}_{ij}$. To incorporate this modification into

the Bayesian algorithms, we put independent flat priors on ψ_j and ν_j, $j = 1, \ldots, p$. The modified Gibbs steps are included in Appendix B.2

We applied each Bayesian approach, running each chain of the Gibbs sampler for 4000 iterations and storing posterior draws from the subsequent 4000 iterations. Table 6 presents numerical results, and Figure 7 plots $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ from each method as a kernel density estimate.

In terms of MSPE, EB-HIBETA-NI and EB-HIBETASIGMAX were the best performing methods, with MSPEs of 0.554 and 0.555, respectively, using $\hat{\boldsymbol{\beta}}^{\text{ppm}}$. These results are somewhat better than those from RIDG (0.620) and HYB (0.601) as well as FB-FLATBETA (1.595), FB-HIBETA-NI (0.793), and EB-HISIGMAX (1.391). Using $\hat{\boldsymbol{\beta}}^{\text{pm}}$, the estimated posterior mean of $\boldsymbol{\beta}$, the two best methods, EB-HIBETA-NI and EB-HIBETASIGMAX, gave almost identical results, while the remaining Bayesian methods had worse prediction error.

From Figure 7, the induced shrinkage from EB-HIBETA-NI and EB-HIBETASIGMAX is considerable; in both cases, the range of $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is about (-0.008,0.008). Plugging in $\hat{\boldsymbol{\beta}} = 0_p$ yields an MSPE of 0.59, which these two methods improve upon. Considering coverage of the prediction intervals (Table 6), FB-HIBETA-NI (0.94), EB-HIBETA-NI (0.97) and EB-HIBETASIGMAX (0.96) all offer coverage rates that are close to their nominal values, as opposed to both FB-FLATBETA and EB-HISIGMAX, whose coverage rates are less than nominal (both 0.86). We also created prediction intervals for RIDG and HYB using a bootstrap algorithm; these, too, had coverage rates that were not as close to nominal (0.91 and 0.99).

3.7 Discussion

We have considered the problem of shrinking coefficients in a high-dimensional model when a large proportion of the data are missing and predictions for future observations

	RIDG	FB-FLATBETA		EB-HIBETA-NI		EB-HIBETASIGMAX	
		HYB	FB-HIBETA-NI		EB-HISIGMAX		
$\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^{\text{ppm}})$	0.620	0.601	1.595	0.793	0.554	1.391	0.555
$\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^{\text{pm}})$	–	–	1.769	0.893	0.558	1.966	0.559
$\min(\hat{\boldsymbol{\beta}}^{\text{ppm}})$	-0.006	-0.023	-0.120	-0.041	-0.009	-0.083	-0.007
$\max(\hat{\boldsymbol{\beta}}^{\text{ppm}})$	0.006	0.027	0.063	0.064	0.005	0.089	0.007
Avg. Coverage	0.91	0.99	0.86	0.94	0.97	0.86	0.96
$\text{Avg}(\hat{Y}_{\text{new}}^{B,97.5} - \hat{Y}_{\text{new}}^{B,2.5})$	3.37	4.67	4.00	3.31	3.11	3.92	3.10

Table 6: Numerical results from analysis of the lung adenocarcinoma data, with RIDG and HYB included for reference. $\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^{\text{ppm}})$ is the empirical MSPE from the validation sample of size 100, $\min(\hat{\boldsymbol{\beta}}^{\text{ppm}})$ and $\max(\hat{\boldsymbol{\beta}}^{\text{ppm}})$ give the range of the estimate of $\boldsymbol{\beta}$ for each model, Avg. Coverage is the proportion of the prediction intervals, for the validation sample, that contained the true outcome. For RIDG and HYB, these are based on the bootstrap. $\text{Avg}(\hat{Y}_{\text{new}}^{B,97.5} - \hat{Y}_{\text{new}}^{B,2.5})$ gives the average prediction interval length for the validation sample.

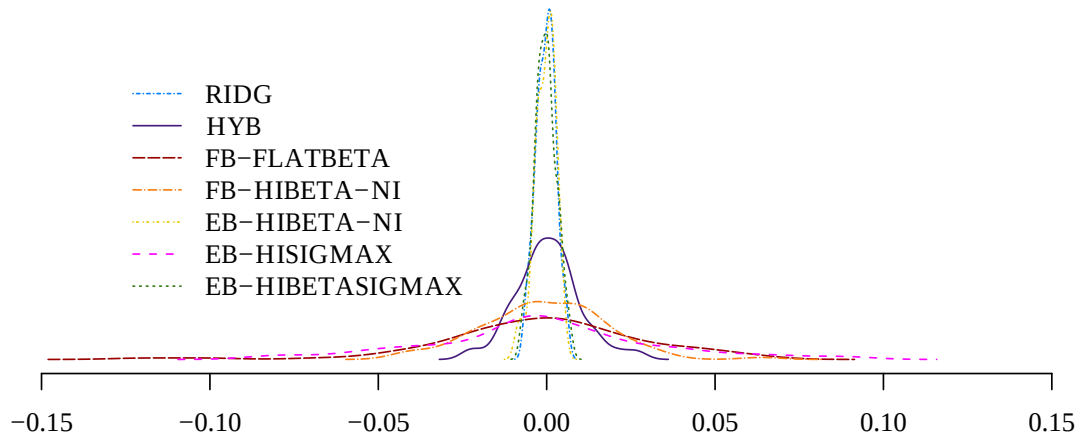

Figure 7: Kernel density estimates of the Chapter 3 methods' 91 coefficient estimates from the lung adenocarcinoma data

are of primary interest. A likelihood-based approach for fitting such a model confers a number of advantages, among these being the inclusion of shrinkage into the likelihood and the proper accounting of uncertainty in predictions coming from the unobserved data. A number of existing Bayesian approaches for the treatment of missing data and/or implementation of shrinkage methods are easily adapted here. We have shown how two such approaches, the Monte Carlo EM (Wei and Tanner, 1990), a Gibbs sampler that multiply imputes missing data, and the Empirical Bayes Gibbs Sampler (Casella, 2001), a Gibbs sampler that adaptively shrinks parameter estimates, actually generalize to the same algorithm, which we call EM-within-Gibbs. Finally, we have proposed several specific choices of prior specification aimed at improving prediction with shrinkage methods.

Two methods, the Bayesian ridge with Empirical Bayes estimation of the shrinkage parameter λ, EB-HIBETA-NI, and Bayesian ridge in conjunction with Empirical Bayes shrinkage of the precision matrix $\boldsymbol{\Sigma}_X^{-1}$, EB-HIBETASIGMAX, stand out as the methods of choice. Our simulation study and data analysis showed them to be best under a number of scenarios using several criteria, including MSPE and prediction coverage, and robust to several modeling violations.

CHAPTER 4

Using Hyperpenalties to Select the Tuning Parameter in Ridge Regression

4.1 Introduction

Suppose we have data, $\{\boldsymbol{y}, \boldsymbol{x}\}$, which are n observations of a continuous outcome Y and p covariates $\boldsymbol{X}$, with the covariate matrix $\boldsymbol{x}$ regarded as fixed. n is small relative to p. We relate Y and $\boldsymbol{X}$ by a linear model, $Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon$, with $\varepsilon \sim N\{0,1\}$. Except where noted, assume $\boldsymbol{y}$ is centered, $\boldsymbol{x}$ is centered and scaled to have unit variance and, consequently, $\hat{\beta}_0 = 0$ (and thus ignored). Up to an additive constant, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}). \tag{34}$$

We consider penalized estimation of $\boldsymbol{\beta}$, with our primary interest being prediction of future observations, rather than variable selection. Thus, we focus on ℓ_2-penalization, ie ridge regression (Hoerl and Kennard, 1970), the significance of which has remained even in the presence of more modern penalization methods (eg Frank and Friedman, 1993; Tibshirani, 1996; Fu, 1998; Zou and Hastie, 2005). Ridge regression may be viewed as a hierarchical linear model, similar to mixed effects modeling. Here the "random effects" are the individual elements of $\boldsymbol{\beta}$. An ℓ_2-penalty implicitly assumes these are jointly and

independently Normal with mean zero and variance σ^2/λ, because the penalty term matches the Normal log-density up to a normalizing constant that does not depend on $\boldsymbol{\beta}$:

$$p_\lambda(\boldsymbol{\beta}, \sigma^2) = -\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + \frac{p}{2}\ln(\lambda) - \frac{p}{2}\ln(\sigma^2). \tag{35}$$

The scalar λ is the ridge parameter and controls the shrinkage of $\boldsymbol{\beta}$ toward zero; larger values yield greater shrinkage. Given λ, the maximum penalized likelihood estimate of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}_\lambda = \text{argmax}_{\boldsymbol{\beta}|\lambda}\left\{\ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2)\right\} = (\boldsymbol{x}^\top\boldsymbol{x} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{x}^\top\boldsymbol{y}. \tag{36}$$

A choice of λ that is too small leads to overfitting of the data, and one that is too large gives underfitting or oversmoothing of the data. When $n - 1 \geq p$, a key result from Hoerl and Kennard (Theorem 4.3, 1970) is that there exists $\lambda > 0$ for which the mean squared error (MSE) of $\boldsymbol{\beta}_\lambda$, $\text{E}[(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)^\top(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)]$, decreases relative to $\lambda = 0$. A strictly positive λ introduces bias in $\boldsymbol{\beta}_\lambda$ but decreases variance, making a bias-variance tradeoff. This result is relevant because prediction error, $\text{E}[(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)^\top\boldsymbol{x}^\top\boldsymbol{x}(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)]$, is closely related to MSE and may correspondingly benefit from such a bias-variance tradeoff. One cannot simply maximize $\ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2)$ jointly with respect to $\boldsymbol{\beta}$, σ^2 and λ, because the expression can be made arbitrarily large by plugging in $\boldsymbol{\beta} = 0$ and letting $\lambda \to \infty$. Typically, λ is selected by optimizing some other other objective function.

Our motivation for this chapter is to investigate selection strategies for λ when n is "small", by which we informally mean $n < p$ or $n \approx p$, the complement being a more standard $n \gg p$ situation. This small-n situation increasingly occurs in modern genomic studies. Our contribution is two-fold. First, we present new ideas for choosing λ in such cases. These include both a small-sample modification to a common existing approach

and novel proposals. Our framework categorizes existing strategies into two classes, based on whether a goodness-of-fit criterion or a likelihood is maximized. Methods in either class are susceptible to over- or underfitting; a third, new class extends the hierarchical perspective of ridge regression, the first level being $\ell(\boldsymbol{\beta}, \sigma^2)$ and the second $p_\lambda(\boldsymbol{\beta}, \sigma^2)$. Following ideas by Takada (1979), who showed that Stein's Positive Part Estimator corresponds to a posterior mode given a certain prior, and, more recently, Strawderman and Wells (2012), who place a hyperprior on the Lasso penalty parameter, we add a third level, defining a "hyperpenalty" on λ. This hyperpenalty induces shrinkage on λ itself and thereby protects against extreme choices of λ. The second contribution follows naturally, namely a comprehensive evaluation of all methods, both existing and novel, in this small-n situation via simulation studies.

The remainder of this chapter is organized as follows. We review current approaches for choosing λ (the first and second classes discussed above) in Sections 4.2 and 4.3 and propose a small-sample modification to one of these methods, generalized cross-validation (GCV, Craven and Wahba, 1979). In Section 4.4, we define a generic hyperpenalty function. Given a hyperpenalty, we define two optimization algorithms, both of which yield a choice of λ: joint optimization (4.4.1) and marginalized optimization (4.4.2). We then present three choices for the form of hyperpenalty (4.4.3). Section 4.5 conducts a comprehensive simulation study. Our results suggest that the existing approaches for choosing λ can be improved upon in many small-n cases. Section 4.6 concludes with a discussion.

4.2 Goodness-of-Fit-Based Methods for the Selection of λ

These methods define an objective function in terms of λ to be minimized. Commonly used is K-fold cross-validation, which partitions observations into K groups, $\kappa(1), \ldots, \kappa(K)$, and calculates $\boldsymbol{\beta}_\lambda$ K times using equation (36), each time leaving out group $\kappa(i)$, to get $\boldsymbol{\beta}_\lambda^{-\kappa(1)}, \boldsymbol{\beta}_\lambda^{-\kappa(2)}$, etc. For $\boldsymbol{\beta}_\lambda^{-\kappa(i)}$, cross-validated residuals are calculated

on the observations in $\kappa(i)$, which did not contribute to estimating $\boldsymbol{\beta}$. The objective function estimates prediction error and is the sum of the squared cross-validated residuals:

$$\lambda_{\text{K-cv}} = \text{argmin}_\lambda \ln \sum_{i=1}^{K} (\boldsymbol{y}_{\kappa(i)} - \boldsymbol{x}_{\kappa(i)} \boldsymbol{\beta}_\lambda^{-\kappa(i)})^\top (\boldsymbol{y}_{\kappa(i)} - \boldsymbol{x}_{\kappa(i)} \boldsymbol{\beta}_\lambda^{-\kappa(i)}). \tag{37}$$

$K = 5$ is a suggested choice; see Hastie et al. (2009) for more details. In the case of $K = n$, some simplification (Golub et al., 1979) gives

$$\lambda_{n\text{-cv}} = \text{argmin}_\lambda \ln \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_\lambda)^2 / (1 - D_{\lambda[ii]} - 1/n)^2, \tag{38}$$

$$\text{with } \boldsymbol{D}_\lambda = \boldsymbol{x}(\boldsymbol{x}^\top \boldsymbol{x} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{x}. \tag{39}$$

$D_{\lambda[ii]}$ is the ith diagonal element of $\boldsymbol{D}_\lambda$ and measures the ith observation's influence in estimating $\boldsymbol{\beta}$. Further discussion of its interpretation is given in Section 4.2.1. From (38), observations for which $D_{\lambda[ii]}$ is large, ie influential observations, have greater weight. GCV multiplies each squared residual in (38) by $(1 - D_{\lambda[ii]} - 1/n)^2 / (1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 1/n)^2$, thereby giving equal weight to all observations. Using the equality $\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}_\lambda = (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)\boldsymbol{y}$, further simplification yields

$$\lambda_{\text{GCV}} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} - 2 \ln(1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 1/n) \right\}. \tag{40}$$

REMARK 13: The input for the functions in (37) and (38) is the non-centered $\boldsymbol{y}$ and unstandardized $\boldsymbol{x}$; for each fold, $\boldsymbol{y}$ is centered and $\boldsymbol{x}$ is standardized according to the model-fitting portions, $\boldsymbol{y}_{-\kappa(i)}$ and $\boldsymbol{x}_{-\kappa(i)}$. $\boldsymbol{D}_\lambda$ is a function of a standardized $\boldsymbol{x}$ and thus depends on the centering and scaling factors, which will change as each fold is sequentially left out. On the other hand, the objection function in (40) contains $\boldsymbol{D}_\lambda$ standardized to all observations. Our notation for $\boldsymbol{D}_\lambda$ is inadequate to reflect this subtlety. Furthermore, re-centering $\boldsymbol{y}$ at each fold implies β_0 is re-estimated; this is reflected by the "$-1/n$" in (38). This term does not appear in the derivations by Golub et al. (1979) nor the

implementation in Chapter 2 given by (5), both of which assume β_0 is known.

Although derived using different principles, other methods reduce to a "model fit + penalty" form similar to (40): Akaike's Information Criterion (AIC, Akaike, 1973) and the Bayesian Information Criterion (BIC, Schwarz, 1978). Respectively, each chooses λ as follows:

$$\lambda_{\text{AIC}} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} + 2(\text{Trace}(\boldsymbol{D}_\lambda) + 2)/n \right\}, \tag{41}$$

$$\lambda_{\text{BIC}} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} + \ln(n)(\text{Trace}(\boldsymbol{D}_\lambda) + 2)/n \right\}. \tag{42}$$

Asymptotically in n, GCV will choose the optimal λ, that is, the one that minimizes the prediction criterion $\text{E}\left[(\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)^\top \boldsymbol{x}^\top \boldsymbol{x} (\boldsymbol{\beta} - \boldsymbol{\beta}_\lambda)\right]$ (Golub et al., 1979; Li, 1986). Further, Golub et al. observe that GCV and AIC asymptotically coincide. BIC asymptotically selects the true underlying model from a set of nested candidate models (Sin and White, 1996; Hastie et al., 2009), so its justification for use in selecting λ, a shrinkage parameter, is weak. In all cases, these methods depend on $n \gg p$. When n is small, extreme overfitting is possible (Wahba and Wang, 1995; Efron, 2001), giving small bias/large variance estimates. A small-sample correction of AIC (AIC_C, Hurvich et al., 1998) and a robust GCV (RGCV_γ, Lukas, 2006) exist:

$$\lambda_{\text{AIC}_C} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} + 2(\text{Trace}(\boldsymbol{D}_\lambda) + 2)/(n - \text{Trace}(\boldsymbol{D}_\lambda) - 3)) \right\}, \tag{43}$$

$$\lambda_{\text{RGCV}_\gamma} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} - 2\ln(1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 1/n) + \ln(\gamma + (1-\gamma)\text{Trace}(\boldsymbol{D}_\lambda^2)/n) \right\}. \tag{44}$$

For AIC_C, the modified penalty is the product of the original penalty, $2(\text{Trace}(\boldsymbol{D}_\lambda) + 2)/n$, and $n/(n - \text{Trace}(\boldsymbol{D}_\lambda) - 3)$. The authors do not consider the possibility of $n - \text{Trace}(\boldsymbol{D}_\lambda) - 3 < 0$, inappropriately giving a negative penalty; in our implementation of AIC_C, we replaced $n - \text{Trace}(\boldsymbol{D}_\lambda) - 3$ with $\max\{\epsilon, n - \text{Trace}(\boldsymbol{D}_\lambda) - 3\}$, with ϵ a small

positive number. As a rule of thumb, Burnham and Anderson (2002) suggest using AIC_C over AIC when $n < 40p$, their threshold for small n, and thus also when $n \approx p$. RGCV_γ adds another penalty to GCV based on a tuning parameter $\gamma \in (0, 1]$, as in (44); we use $\gamma = 0.3$ based on Lukas' recommendation. Small choices of λ are more severely penalized, thereby offering protection against overfitting. To the best of our knowledge, neither AIC_C nor GCV_γ have been extensively studied in the context of ridge regression.

4.2.1 Small-Sample Generalized Cross-Validation

$\text{Trace}(\boldsymbol{D}_\lambda)$, with $\boldsymbol{D}_\lambda$ defined in (39), is the effective number of model parameters (excluding β_0 and σ^2). It is monotonically decreasing with $\lambda > 0$ and in the interval $(0, \min\{n-1, p\})$. The upper bound on $\text{Trace}(\boldsymbol{D}_\lambda)$ is not $\min\{n, p\}$ because the standardization of $\boldsymbol{x}$ reduces its rank by one when $n \leq p$. AIC/BIC do not depend on whether β_0 and σ^2 are properly counted as parameters in the penalty. That is, the counting of these parameters is represented by the "$+2$" expressions in (41) and (42) and thus only additively affects the penalty. For this reason β_0 and σ^2 may be ignored in considering model complexity. On the other hand, from (40), GCV counts β_0 but not σ^2; counting both *will* change the penalty, being on the log-scale. This motivates our proposed small-sample correction to GCV (GCV_C), which properly counts σ^2 as a parameter:

$$\lambda_{\text{GCV}_C} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y} - 2 \ln(\max\{\epsilon, 1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 2/n\}) \right\}, \quad (45)$$

with ϵ a small positive number. As with AIC_C, $1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 2/n$ may be negative, and the objective function is heavily penalized in this case. This is a small-sample correction because the objective functions in (40) and (45) coincide as $n \to \infty$; in particular, the asymptotic optimality of GCV transfers to GCV_C.

The small-sample deficiency of GCV is corrected by GCV_C as follows. If $n - 1 = p$, the

model-fit term in the objective function of (40), $\ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^2 \boldsymbol{y}$, tends to negative infinity as λ decreases. In the extreme case of $\lambda = 0$, each observation is allocated a regression coefficient, and predictions perfectly fit the observations. The penalty term, $-2\ln(1 - \text{Trace}(\boldsymbol{D}_\lambda)/n - 1/n)$, tends to infinity as λ decreases, because $\text{Trace}(\boldsymbol{D}_\lambda)$ approaches $n - 1$. The rates of convergence for the model-fit and penalty terms determine whether GCV chooses a too-small λ. In particular, if the model-fit term approaches negative infinity faster, the objective function is minimized by setting λ as small as possible. Although this phenomenon is most striking in cases for which $n - 1 = p$, as we will see in Section 4.5, the case is similar when $n - 1 < p$. That is, predictions will nearly match observations as λ decreases but remains numerically positive to allow for the matrix inversion in $\boldsymbol{D}_\lambda$, and the penalty term still approaches infinity. The GCV_C penalty, however, is infinite when λ is small enough such that $\text{Trace}(\boldsymbol{D}_\lambda) = n - 2$, rather than $n - 1$, as with GCV. Therefore, the effective number of remaining parameters, beyond σ^2, is allowed to be at most $n - 1$, and perfect fit of the observations to the predictions never occurs because the corresponding penalty is too large.

4.3 Likelihood-Based Methods for the Selection of λ

A second approach treats the ridge penalty in (35) as a log-density. One criterion considers a marginal likelihood, where λ is interpreted as the variance component of a mixed-effects model:

$$\begin{aligned} m(\lambda, \sigma^2) &= \ln \int_{\boldsymbol{\beta}} \exp\{\ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2)\} \, \mathrm{d}\boldsymbol{\beta} \\ &= -\frac{1}{2\sigma^2} \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda) \boldsymbol{y} - \frac{n}{2} \ln(\sigma^2) + \frac{1}{2} \ln \left| \boldsymbol{I}_n - \boldsymbol{D}_\lambda \right|. \end{aligned} \tag{46}$$

From this, $\boldsymbol{y}|\lambda, \sigma^2$ is Multivariate Normal with mean $\boldsymbol{0}_n$ ($\boldsymbol{y}$ is centered) and covariance $\sigma^2(\boldsymbol{I}_n - \boldsymbol{D}_\lambda)^{-1}$. The maximum profile marginal likelihood (MPML) estimate, originally

proposed for smoothing splines (Wecker and Ansley, 1983), profiles $m(\lambda, \sigma^2)$ over σ^2, replacing each instance of σ^2 with $\hat{\sigma}^2_\lambda = \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)\, \boldsymbol{y}/n$, and optimizes the "concentrated" log-likelihood, $m(\lambda, \hat{\sigma}^2_\lambda)$:

$$\lambda_{\text{MPML}} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)\, \boldsymbol{y} - \frac{1}{n} \ln \left| \boldsymbol{I}_n - \boldsymbol{D}_\lambda \right| \right\}. \tag{47}$$

Closely related is the generalized/restricted MPML (GMPML, Harville, 1977; Wahba, 1985), which adjusts the penalty to account for estimation of regression parameters that are not marginalized. Here, only β_0 is not marginalized, so the adjustment is by one degree of freedom (Appendix C.1):

$$\lambda_{\text{GMPML}} = \text{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)\, \boldsymbol{y} - \frac{1}{n-1} \ln \left| \boldsymbol{I}_n - \boldsymbol{D}_\lambda \right| \right\}. \tag{48}$$

In a smoothing-spline comparison of GMPML to GCV, Wahba (1985) found mixed results, with neither method offering uniformly better predictions. For scatterplot smoothers, Efron (2001) notes that GMPML may oversmooth, yielding large bias/small variance estimates.

REMARK 14: Rather than profiling over σ^2, one could jointly optimize $m(\lambda, \sigma^2)$ over λ and σ^2. This can be achieved with iterative gradient ascent steps. We have not found this approach previously used as a selection criterion in ridge regression. Our initial investigation of this (and its restricted likelihood counterpart) gave results similar to MPML and GMPML; consequently, we do not consider them further.

An alternative to the marginal likelihood methods described above is to treat the objective function in (36) as an h-log-likelihood, or "h-loglihood", of the type proposed by Lee and Nelder (1996) for hierarchical generalized linear models. The link between penalized likelihoods (like ridge regression) and the h-loglihood was noted in the paper's ensuing discussion. To estimate σ^2, the dispersion, and λ, the variance component, Lee

and Nelder suggested an iterative profiling approach, yielding the maximum adjusted profile h-loglihood (MAPHL) estimate. In Appendix C.2, we show one iteration proceeds as follows:

$$\sigma^{2(i)} \leftarrow \frac{\left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\right)^\top \left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\right) + \lambda^{(i-1)}\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}}{n-1}, \tag{49}$$

$$\lambda^{(i)} \leftarrow \operatorname{argmin}_\lambda \left\{ \lambda \boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}/\sigma^{2(i)} - \ln\left|\boldsymbol{I}_n - \boldsymbol{D}_\lambda\right| \right\}, \tag{50}$$

$$\boldsymbol{\beta}^{(i)} \leftarrow \boldsymbol{\beta}_{\lambda^{(i)}}, \tag{51}$$

and $\lambda_{\text{MAPHL}} = \lambda^{(\infty)}$.

Finally, Tran (2009) proposed the "Loss-Rank" (LR) method for selecting λ. Its derivation, which we do not give, is likelihood-based, but the criterion resembles that of AIC in (41):

$$\lambda_{\text{LR}} = \operatorname{argmin}_\lambda \left\{ \ln \boldsymbol{y}^\top \left(\boldsymbol{I}_n - \boldsymbol{D}_\lambda\right)^2 \boldsymbol{y} - \frac{2}{n} \ln\left|\boldsymbol{I}_n - \boldsymbol{D}_\lambda\right| \right\}. \tag{52}$$

Tran also suggested a modified penalty term that is dependent on $\boldsymbol{y}$, but this did not give appreciably different results from λ_{LR} in their study.

4.4 Optimization with Hyperpenalties

As we have noted, some existing methods may choose extreme values of λ, particularly when n is small, suggesting a need for a second level of shrinkage, that is, shrinkage of λ itself. We extend the hierarchical framework of (34) and (35) with a "hyperpenalty" on λ, $h(\lambda)$, which gives support for λ over a range of values. The hyperpenalty corresponds to a log-density and therefore must satisfy $\int_0^\infty \exp\{h(\lambda)\}\,\mathrm{d}\lambda < \infty$. Further, $h(\lambda)$ must approach negative ∞ sufficiently fast with λ. We will propose three choices of $h(\lambda)$. When σ^2 is unknown, the hierarchy is completed with a distribution on σ^2; we use the

improper log-density $-\ln(\sigma^2)$ throughout. The "hyperpenalized log-likelihood" is then

$$hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2) = \ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2) + h(\lambda) - \ln(\sigma^2). \tag{53}$$

From the Bayesian perspective, the hyperpenalty is simply a hyperprior on λ, and the hyperpenalized likelihood is the posterior. In contrast to fully Bayesian methods, which characterize the entire posterior, we focus primarily on mode finding, that is, finding a single estimate of $\boldsymbol{\beta}$, σ^2 and λ, which carries with it computational advantages. We consider two options: joint optimization of $\boldsymbol{\beta}$, σ^2 and λ (JO) and optimization of $\boldsymbol{\beta}$ and σ^2 after marginalizing over λ (MO). This hierarchical perspective also relates to our earlier comment in Section 4.1 that one cannot simply optimize $\ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2)$ alone and demonstrates why $h(\lambda)$ must correspond to a log-density. In other words, optimizing (53) using $h(\lambda) = C$, C constant, would yield the same result as optimizing $\ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2)$, namely an infinite hyperpenalized log-likelihood. We first describe JO and MO for a choice of $h(\lambda)$ that satisfies the above conditions but is otherwise general and then propose three specific choices of $h(\lambda)$.

4.4.1 Joint Optimization

JO estimates a joint mode of (53), namely $\{\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma}^2\} \leftarrow \text{argmax}_{\boldsymbol{\beta}, \lambda, \sigma^2}\{hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2)\}$. If this is intractable, the joint maximum may be calculated using conditional maximization steps:

$$\begin{aligned}
\sigma^{2(i)} &\leftarrow \text{argmax}_{\sigma^2}\left\{hp\ell(\boldsymbol{\beta}^{(i-1)}, \sigma^2, \lambda^{(i-1)})\right\} \\
&= \frac{\left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)})^\top(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\right) + \lambda^{(i-1)}\boldsymbol{\beta}^{(i-1)\top}\boldsymbol{\beta}^{(i-1)}}{n + p + 2}, \qquad (54) \\
\lambda^{(i)} &\leftarrow \text{argmax}_\lambda\left\{hp\ell(\boldsymbol{\beta}^{(i-1)}, \sigma^{2(i)}, \lambda)\right\}, \qquad (55) \\
\boldsymbol{\beta}^{(i)} &\leftarrow \text{argmax}_{\boldsymbol{\beta}}\left\{hp\ell(\boldsymbol{\beta}, \sigma^{2(i)}, \lambda^{(i)})\right\} = \boldsymbol{\beta}_{\lambda^{(i)}}, \qquad (56)
\end{aligned}$$

and letting $\{\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma}^2\} = \{\boldsymbol{\beta}^{(\infty)}, \lambda^{(\infty)}, \sigma^{2(\infty)}\}$.

4.4.2 Marginalized Optimization

In MO, λ is marginalized, and the resulting marginal likelihood of $\boldsymbol{\beta}$ and σ^2 is maximized: $\{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\} \leftarrow \operatorname{argmax}_{\boldsymbol{\beta},\sigma^2} \ln \int_\lambda \exp\left\{hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2)\right\} \mathrm{d}\lambda$. An EM algorithm (Dempster et al., 1977) achieves this:

E-step $Q(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\beta}^{(i)}, \sigma^{2(i)}) \leftarrow \mathrm{E}_{\lambda|\boldsymbol{\beta}^{(i)},\sigma^{2(i)}}\left[hp\ell(\boldsymbol{\beta}, \lambda, \sigma^2)\right]$.

M-step $\{\boldsymbol{\beta}^{(i+1)}, \sigma^{2(i+1)}\} \leftarrow \operatorname{argmax}_{\boldsymbol{\beta},\sigma^2} Q(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\beta}^{(i)}, \sigma^{2(i)})$.

Upon convergence, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(\infty)}$. Because $h(\lambda)$ is not a function of $\boldsymbol{\beta}$ and σ^2, calculation of Q, the expected hyperpenalized likelihood, requires only the expectation of $p_\lambda(\boldsymbol{\beta}, \sigma^2)$. From (35), the only expression containing $\boldsymbol{\beta}$ and σ^2 is linear in λ. Therefore, only the expectation of λ is required, and equivalent formulations of the E- and M-steps are written as

$$\sigma^{2(i)} \leftarrow \frac{\left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\right)^\top \left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}^{(i-1)}\right) + \lambda^{(i-1)} \boldsymbol{\beta}^{(i-1)\top} \boldsymbol{\beta}^{(i-1)}}{n + p + 2}, \tag{57}$$

$$\lambda^{(i)} \leftarrow \mathrm{E}_{\lambda|\boldsymbol{\beta}^{(i-1)},\sigma^{2(i)}}[\lambda] = \frac{\int_0^\infty \lambda \exp\{p_\lambda(\boldsymbol{\beta}^{(i-1)}, \sigma^{2(i)}) + h(\lambda)\} \mathrm{d}\lambda}{\int_0^\infty \exp\{p_\lambda(\boldsymbol{\beta}^{(i-1)}, \sigma^{2(i)}) + h(\lambda)\} \mathrm{d}\lambda}, \tag{58}$$

$$\boldsymbol{\beta}^{(i)} \leftarrow \boldsymbol{\beta}_{\lambda^{(i)}}. \tag{59}$$

Thus, the only difference between MO and JO is in the update to λ.

4.4.3 Choice of Hyperpenalty

We now consider choices of the hyperpenalty $h(\lambda)$. Each is parametrized with a shape and rate parameter. The resulting JO and MO updates for λ are given Appendices C.3–C.5.

Gamma (GA) This uses the hyperpenalty $h_{\text{GA}}(\lambda) = (a-1)\ln(\lambda) - \lambda/b$, the log-density of a $G\{a, b\}$ variable. From the Bayesian perspective, this may be natural because it is conjugate to the precision of the Normal distribution, which is one possible interpretation of λ (eg Armagan and Zaretzki, 2010; Tipping, 2001). In contrast, the Gamma distribution may not be amenable to interpreting λ as a *shrinkage* parameter. The density of the random variable $\ln(\lambda)$ induced by $h_{\text{GA}}(\lambda)$ above is negatively skewed (logarithms are the appropriate scale here because of the relative impact of λ). This negative skewness may lead to overfitting (ie small bias/large variance estimates). In small-n situations, it is the variance of $\boldsymbol{\beta}_\lambda$, and not bias, that primarily drives prediction error, meaning that underfitting is the less deleterious error to make. This motivates our consideration of distributions for $\ln(\lambda)$ with no skewness (λ is Log-Normal) or positive skewness (Inv-Gamma).

Log-Normal (LN) This uses $h_{\text{LN}}(\lambda) = -\ln(\lambda) - \ln(b\lambda)^2/(2a)$, which is the log-density of Log-Normal variable with shape a and rate b. If λ is Log-Normal, $\ln(\lambda)$ is Normal, and its density has no skewness.

Inverse-gamma (IG) This has the hyperpenalty $h_{\text{IG}}(\lambda) = -(a+1)\ln(\lambda) - 1/(b\lambda)$, which is the log of an $IG\{a, b\}$ density. In this case, $\ln(\lambda)$ has positive skewness, offering protection against overfitting. Although this is an unusual distributional choice when viewing λ as a precision parameter (the Inv-Gamma distribution is conjugate to the Normal *variance*), the properties of the Inv-Gamma distribution are amenable to the desired behavior of a shrinkage parameter. One drawback to this choice of hyperpenalty is that small values of a may result in an infinite mode (JO) or infinite intergral (MO). This is discussed below.

4.4.4 Specification of the Shape and Rate

Implementing the GA, LN and IG hyperpenalties requires the subjective determination of the shape and rate (a and b, respectively). From the Bayesian perspective, a hyperpenalty quantifies the prior knowledge one is willing to assume about λ, thus it is reasonable to choose a and b by solving for desired moments, $\mathrm{E}\ln(\lambda)$ and $\mathrm{Var}\ln(\lambda)$. Because optimal shrinkage typically increases with p, a simple choice for the mean is $\mathrm{E}\ln(\lambda) = \ln(p)$. $\mathrm{Var}\ln(\lambda)$ needs to allow for a sufficiently flexible hyperpenalty while, in the case of IG, simultaneously ensuring that the optimization has a finite mode (JO) and that the integral is finite (MO). Focusing on IG, in Appendix C.5, we show that $a \geq p/2 + 1$ is sufficient to ensure a finite mode and integral. Now, if $\lambda \sim IG\{a, b\}$, then $\mathrm{Var}\ln(\lambda) = \psi_1(a)$, where $\psi_1(z) = (\mathrm{d}^2/\mathrm{d}z^2)\ln\Gamma(z)$ is the trigamma function, which is decreasing in its argument. Thus, for IG, $a = p/2 + 1$ is a "maximum variance" choice that guarantees a finite value of λ. For the sake of comparison, we use this same choice for all three hyperpenalties. In summary, for GA, LN and IG, we choose a and b to solve $\mathrm{E}\ln(\lambda) = \ln(p)$ and $\mathrm{Var}\ln(\lambda) = \psi_1(p/2 + 1)$.

4.5 Simulation Study

One simulated dataset consists of training and validation data generated from the same model. We consider three choices of $\boldsymbol{\beta}$, all of length $p = 99$:

(i) $\boldsymbol{\beta} = \{1\}_{j=1}^{99}$.

(ii) $\boldsymbol{\beta} = \{j^2/99^2\}_{j=1}^{99}$.

(iii) for $i = 1, \ldots, 99$, $\beta_i | \{\pi_1, \pi_2, \pi_3\} \overset{iid}{\sim} f, f = \pi_1(1/\sqrt{3})t_3 + \pi_2 \mathrm{Exp}\{1\} + \pi_3\delta\{0\}$, $\{\pi_1, \pi_2, \pi_3\} \sim \mathrm{Dir}\{1, 1, 1\}$.

The third choice draws a random $\boldsymbol{\beta}$, ie $\boldsymbol{\beta}$ is randomly drawn for each dataset. The sampling density is a random mixture of scaled t_3, Exponential, and Dirac delta distributions. We use $n \in \{25, 50, 100, 200\}$. The $n \times p$ matrix $\boldsymbol{x}$ is drawn from $N_p\{\mathbf{0}_p, \boldsymbol{\Sigma_X}\}$, where the diagonal elements of $\boldsymbol{\Sigma_X}$ are set to unity, and the (j_1, j_2) off-diagonal element is taken from $\rho_{j_1 j_2} \in \{0.75^{|j_1 - j_2|}, 0.75\}$. For the validation data, a $2000 \times p$ matrix $\boldsymbol{x}_{\text{new}}$ is sampled. $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} + \sigma^2)$ comes from $\{0.1, 0.3, 0.5\}$, which specifies a value for σ^2. Finally, $\boldsymbol{y}|\boldsymbol{x}$ (training data) and $\boldsymbol{y}_{\text{new}}|\boldsymbol{x}_{\text{new}}$ (validation data) are drawn using $\boldsymbol{\beta}$ and σ^2. For each of the 72 combinations of $\boldsymbol{\beta}$, n, $\rho_{j_1 j_2}$, and R^2, 1500 simulated training and validation datasets are sampled.

We tested 16 methods, all of which are listed in Table 7. n-CV is left out because it is approximated by GCV and is computationally expensive, even with the simplification in (38), owing to the need to restandardize $\boldsymbol{x}$ and recalculate $\boldsymbol{D}_\lambda$ with each fold. Also, AIC is replaced with its small-sample correction, AIC_C. Each method differs only in its choice of λ, which determines the estimate of $\boldsymbol{\beta}$ via (36). The criterion by which we evaluate methods on the validation data is relative MSPE, rMSPE(λ), defined as follows:

$$\text{rMSPE}(\lambda) = 1000 \times (\text{MSPE}(\lambda)/\text{MSPE}(\lambda_{\text{opt}}) - 1), \tag{60}$$

$$\text{where MSPE}(\lambda) = (\boldsymbol{y}_{\text{new}} - \boldsymbol{x}_{\text{new}} \boldsymbol{\beta}_\lambda)^\top (\boldsymbol{y}_{\text{new}} - \boldsymbol{x}_{\text{new}} \boldsymbol{\beta}_\lambda) \text{ and } \lambda_{\text{opt}} = \text{argmin}_\lambda \text{MSPE}(\lambda). \tag{61}$$

Thus, rMSPE is a scaled measure of the percentage increase over the best possible MSPE, with zero being the smallest achievable value. We calculated λ_{opt} with an iterative grid search. Tables 8 and 9 give rMSPE averaged over the 1500 simulations for all methods for $\rho_{j_1 j_2} = 0.75^{|j_1 - j_2|}$ and $\rho_{j_1 j_2} = 0.75$, respectively.

AIC_C is the best-performing of the existing goodness-of-fit methods, and, by a large margin, BIC is the worst. 5-CV has small rMSPE only for larger n. GCV, the most commonly used method in the literature, has large average rMSPE in the $n - 1 \leq p$ cases, signifi-

Abbr.	Name	Eqn.	Reference
5-CV	Five-fold Cross Validation	(37)	(Section 7.10, Hastie et al., 2009)
GCV	Generalized Cross Validation	(40)	(Craven and Wahba, 1979)
BIC	Bayesian Information Criterion	(42)	(Schwarz, 1978)
AIC_C	Corrected Akaike's Information Criterion	(43)	(Hurvich et al., 1998)
RGCV_γ	Robust Generalized Cross Validation	(44)	(Lukas, 2006)
GCV_C	Corrected Generalized Cross Validation	(45)	Section 4.2.1
MPML	Maximum Profile Marginal Likelihood	(47)	(Wecker and Ansley, 1983)
GMPML	Generalized Maximum Profile Marginal Likelihood	(48)	(Harville, 1977; Wahba, 1985)
MAPHL	Maximum Adjusted Profile h-Likelihood	(49)-(51)	(Lee and Nelder, 1996)
LR	Loss-Rank	(52)	(Tran, 2009)
GA-JO	Gamma - Joint Optimization	(54)-(56)	Section 4.4
GA-MO	Gamma - Marginal Optimization	(57)-(59)	Section 4.4
LN-JO	Log-Normal - Joint Optimization	(54)-(56)	Section 4.4
LN-MO	Log-Normal - Marginal Optimization	(57)-(59)	Section 4.4
IG-JO	Inv-Gamma - Joint Optimization	(54)-(56)	Section 4.4
IG-MO	Inv-Gamma - Marginal Optimization	(57)-(59)	Section 4.4

Table 7: A list of methods to select the tuning parameter of a ridge regression and their associated references

cantly so when $n - 1 = p$. This is a reflection of plugging in $\lambda = 0$ to achieve an exact fit when predicting the training data (Section 4.2.1). $\text{RGCV}_{0.3}$, its robust counterpart, protects against some of the most drastic overfitting but has larger rMSPE than GCV in other situations. Our proposed small-sample correction, GCV_C, does very well and is on par with AIC_C, exhibiting no extreme overfitting but otherwise closely matching GCV.

Considering the likelihood-based methods, MPML has uniformly larger rMSPE than GMPML, which in turn has uniformly larger rMSPE than MAPHL. These become nearly equal as n increases. LR predicts poorly in some cases, eg when $n = 25$ and occasionally when $R^2 = 0.5$.

Each of our proposed hyperpenalties demonstrates relatively small rMSPE in most situations, with the Inv-Gamma hyperpenalty the overall best-performing method. With few exceptions, its rMSPE is always among the smallest. It is followed closely by the Log-Normal hyperpenalty. As for choice of algorithm, joint optimization (JO) is perhaps slightly preferred to marginalized optimization (MO).

$\boldsymbol{\beta} = \{1\}_{j=1}^{99}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	75	39	*17	*9	98	42	17	*7	97	47	18	*7
GCV	103	62	3255	*8	118	70	1866	*6	105	74	2744	*7
BIC	709	1397	$> 10^4$	53	513	1225	$> 10^4$	93	362	1059	$> 10^4$	89
AIC_C	*20	*17	***11**	***7**	91	45	16	*6	200	72	21	*7
$\text{RGCV}_{0.3}$	*18	*21	1344	22	95	94	950	48	226	176	1421	49
GCV_C	33	31	*19	*8	50	34	25	***6**	*55	34	15	***7**
MPML	577	60	*15	*8	483	92	20	*11	356	183	29	19
GMPML	60	28	*15	*8	87	37	19	*11	93	47	28	19
MAPHL	52	25	*14	*8	66	35	19	*11	67	47	28	19
LR	289	*14	*12	*10	297	33	24	18	267	48	35	25
GA-JO	41	49	54	53	*9	*11	21	28	***29**	***9**	*7	16
GA-MO	40	48	53	52	*9	*10	20	28	*31	*9	*7	16
LN-JO	33	39	43	44	***9**	*8	16	23	*36	*11	*6	14
LN-MO	31	37	41	43	*9	***7**	15	22	*39	*13	***6**	*13
IG-JO	*13	*14	*17	22	18	*8	*8	13	63	22	*6	*10
IG-MO	***10**	***11**	*13	18	25	*11	***7**	*11	78	27	*6	*9
$\boldsymbol{\beta} = \{j^2/99^2\}_{j=1}^{99}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	73	39	*20	*8	96	49	18	*8	96	49	17	*7
GCV	91	60	2656	*7	116	68	2332	*7	107	77	2199	*6
BIC	698	1431	$> 10^4$	53	513	1239	$> 10^4$	95	365	1033	$> 10^4$	91
AIC_C	*19	*17	***13**	***6**	88	47	18	***6**	206	75	22	*7
$\text{RGCV}_{0.3}$	*18	*21	815	23	92	97	976	49	233	182	1001	50
GCV_C	32	29	*18	*7	48	35	18	*7	*55	34	26	***6**
MPML	578	49	*17	*8	480	92	19	*11	361	190	27	18
GMPML	56	29	*16	*8	86	37	18	*11	92	46	26	18
MAPHL	48	26	*15	*8	63	36	18	*11	65	45	26	18
LR	272	*15	*13	*10	287	34	27	18	271	49	37	26
GA-JO	42	50	54	52	*9	*10	19	28	***31**	***9**	*6	15
GA-MO	41	49	53	51	*9	*10	18	27	*32	*9	*6	15
LN-JO	33	39	44	43	*8	*7	14	23	*38	*12	*5	13
LN-MO	31	37	42	42	***8**	***7**	13	21	*41	*13	***5**	*12
IG-JO	*13	*15	*18	22	*17	*9	*7	13	66	23	*6	*10
IG-MO	***10**	***11**	*13	17	24	*12	***6**	*11	81	29	*6	*8
$\beta_i \vert \{\pi_1, \pi_2, \pi_3\} \overset{iid}{\sim} f, f = \pi_1(1/\sqrt{3})t_3 + \pi_2 \text{Exp}\{1\} + \pi_3\delta\{0\}, \{\pi_1, \pi_2, \pi_3\} \sim \text{Dir}\{1,1,1\}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	80	41	*20	*10	90	53	25	*9	103	56	23	*10
GCV	116	66	2691	*9	114	74	1793	*9	112	84	1849	*10
BIC	694	1412	$> 10^4$	41	507	1138	$> 10^4$	128	347	914	$> 10^4$	150
AIC_C	*14	*13	*11	***7**	64	44	22	*9	155	84	33	13
$\text{RGCV}_{0.3}$	*13	*15	880	21	64	78	796	65	168	176	856	81
GCV_C	35	31	*17	*9	44	41	42	*9	61	41	26	*10
MPML	589	56	*15	*8	461	100	19	*8	339	170	19	*10
GMPML	63	27	*15	*8	76	37	19	*8	94	44	19	*9
MAPHL	57	25	*13	*8	57	34	19	*8	64	41	19	*9
LR	301	***12**	***11**	*10	269	33	31	27	254	59	53	46
GA-JO	49	55	54	46	*13	*12	*14	14	***20**	***11**	***7**	*5
GA-MO	48	54	53	46	*12	*11	*13	14	*20	*12	*7	***5**
LN-JO	40	44	44	38	*11	*9	*10	*11	*25	*15	*8	*5
LN-MO	38	42	42	37	***10**	***8**	*9	*10	*26	*16	*8	*5
IG-JO	*17	*17	*18	19	*13	*8	***7**	*7	47	29	*13	*6
IG-MO	***13**	*12	*14	15	*18	*11	*7	***6**	60	36	15	*6

Table 8: Average rMSPE (to the nearest integer) as defined in (60) for 16 methods (summarized in Table 7) over three choices of $\boldsymbol{\beta}$ and twelve combinations of n and R^2, with $\rho_{j_1 j_2} = 0.75^{|j_1 - j_2|}$. Values in **bold** are the column-wise minima, and those with an '*' are less than twice the column-wise minima.

$\boldsymbol{\beta} = \{1\}_{j=1}^{99}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	79	50	19	8	78	45	17	8	71	43	16	*7
GCV	85	58	2745	8	86	59	2556	7	74	55	2322	*6
BIC	362	1007	$> 10^4$	13	345	1009	$> 10^4$	15	332	1001	$> 10^4$	18
AIC_C	*34	18	*8	*4	45	*15	*6	*4	40	*15	*6	*4
$\text{RGCV}_{0.3}$	44	32	1120	*4	64	21	1023	*3	39	*15	1036	*4
GCV_C	40	31	17	7	39	27	24	7	30	25	13	*6
MPML	344	171	13	6	337	189	15	9	331	199	21	13
GMPML	52	28	12	6	57	28	14	9	55	34	19	13
MAPHL	40	27	13	6	40	27	14	9	41	32	20	13
LR	276	26	13	6	281	28	11	*5	282	24	10	*5
GA-JO	*23	*15	16	23	*17	*11	14	22	*14	*10	12	21
GA-MO	*23	*14	16	23	*17	*11	14	22	***14**	*10	12	20
LN-JO	*21	*12	12	18	*16	*9	10	17	*14	*9	9	15
LN-MO	*20	*12	12	17	***16**	***9**	10	16	*15	***9**	8	15
IG-JO	***18**	***8**	***4**	*4	*28	*11	***4**	*4	30	*12	***4**	*5
IG-MO	*24	*13	*5	***3**	38	*15	*5	***3**	37	*14	*4	***4**
$\boldsymbol{\beta} = \{j^2/99^2\}_{j=1}^{99}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	79	48	20	9	83	44	19	8	76	43	16	7
GCV	85	59	2709	8	88	52	2493	7	81	53	2274	*6
BIC	351	1020	$> 10^4$	13	349	1011	$> 10^4$	15	336	989	$> 10^4$	18
AIC_C	*33	17	*8	*4	44	*15	*6	*4	42	*15	*6	***3**
$\text{RGCV}_{0.3}$	43	31	821	*4	64	20	582	*3	41	*15	1151	*4
GCV_C	38	28	17	7	41	26	20	6	34	26	26	*6
MPML	332	157	13	6	341	178	16	9	333	217	20	13
GMPML	49	27	12	6	61	26	15	9	61	35	19	13
MAPHL	39	26	13	6	43	27	15	9	45	33	19	13
LR	268	24	13	6	286	26	11	*6	287	26	10	*5
GA-JO	*22	*14	16	24	*18	*12	14	22	*16	*10	12	20
GA-MO	*22	*14	16	23	*18	*12	14	21	***16**	*10	12	20
LN-JO	*19	*12	12	18	*17	*10	10	16	*17	*8	9	15
LN-MO	*19	*11	12	17	***16**	***9**	10	16	*17	***8**	8	14
IG-JO	***17**	***8**	***4**	*4	*28	*11	***4**	*4	*32	*12	***4**	*5
IG-MO	*22	*12	*5	***3**	37	*15	*5	***3**	39	*14	*4	*4
$\beta_i \vert \{\pi_1, \pi_2, \pi_3\} \overset{iid}{\sim} f, f = \pi_1(1/\sqrt{3})t_3 + \pi_2 \text{Exp}\{1\} + \pi_3\delta\{0\}, \{\pi_1, \pi_2, \pi_3\} \sim \text{Dir}\{1,1,1\}$												
Method/$\{n/R^2\}$	25/0.1	50/0.1	100/0.1	200/0.1	25/0.3	50/0.3	100/0.3	200/0.3	25/0.5	50/0.5	100/0.5	200/0.5
5-CV	79	52	20	8	80	49	20	*9	73	44	*17	*9
GCV	84	68	2558	8	82	54	2503	*8	75	58	2925	*8
BIC	344	984	$> 10^4$	14	322	954	$> 10^4$	25	299	896	$> 10^4$	49
AIC_C	*31	17	*8	*5	47	*17	*9	***6**	47	*20	***12**	***7**
$\text{RGCV}_{0.3}$	39	28	722	*5	66	23	493	*9	51	*23	1142	21
GCV_C	37	33	18	7	42	28	21	*8	*35	25	*22	*8
MPML	326	180	14	7	313	198	*16	*9	296	229	*18	*11
GMPML	48	30	13	6	57	30	*15	*9	57	32	*17	*11
MAPHL	40	29	13	*6	39	27	*15	*9	40	29	*17	*11
LR	257	24	13	7	273	31	*16	*9	255	30	*17	*12
GA-JO	*22	*14	15	20	*18	*12	*12	17	***18**	*12	*14	21
GA-MO	*22	*14	14	19	*18	*12	*12	17	*18	*12	*14	20
LN-JO	*20	*12	11	15	*17	*10	*10	13	*19	***12**	*13	18
LN-MO	*20	*12	11	14	***17**	***10**	*9	13	*19	*12	*13	18
IG-JO	***17**	***8**	***5**	*3	*28	*14	***8**	*7	*36	*19	*16	17
IG-MO	*22	*12	*6	***3**	38	*17	*10	*8	43	*22	*18	18

Table 9: Average rMSPE (to the nearest integer) as defined in (60) for 16 methods (summarized in Table 7) over three choices of $\boldsymbol{\beta}$ and twelve combinations of n and R^2, with $\rho_{j_1 j_2} = 0.75$. Values in **bold** are the column-wise minima, and those with an '*' are less than twice the column-wise minima.

We also inspected the choices of λ for each method. Figure 8 plots histograms of $\ln(\lambda/\lambda_{\text{opt}})$ for the 16 methods from one simulation setting in Table 8 ($\rho_{j_1 j_2} = 0.75^{|j_1 - j_2|}$): $\boldsymbol{\beta} = \{1\}_{j=1}^{99}$, $n = 100$ and $R^2 = 0.1$. When $\ln(\lambda/\lambda_{\text{opt}}) = 0$, the method has selected the optimal λ. From the figure, GCV in particular chooses very extreme values, and GCV_C effectively fixes this. The top-performing methods in this scenario, AIC_C and LR, have bimodal histograms, indicating that a number of simulations chose a very large λ. In contrast, shrinkage from hyperpenalization is evident: IG-MO, which ranks third in rMSPE, closely behind AIC_C and LR, has less spread and is closer to zero. Also noteworthy are the between-hyperpenalty differences, evident from the range of the x-axes. Even though the hyperpenalties are selected so that the first two moments of $\ln(\lambda)$ are equal for each (Section 4.4.4), the Inv-Gamma hyperpenalty allows the data to inform the choice of λ to a greater extent than Gamma.

4.6 Discussion

We have examined strategies for choosing the ridge parameter λ when the sample size n is small relative to p. Our small-sample modification to GCV, called GCV_C, properly counts the number of parameters in the model and uniformly dominates GCV in our simulation study. We also proposed a novel approach using what we call hyperpenalties. These add another level of shrinkage, that of λ itself, by extending the hierarchical model. Relative to existing methods, these can offer superior prediction and protection against choosing extreme values of λ. We also propose two optimization techniques given a choice of hyperpenalty: one that jointly maximizes the hyperpenalized likelihood and one that marginalizes over λ. The heavy positive tail of the Inv-Gamma distribution makes it the most flexible of the hyperpenalties we considered.

One generalizable advantage of our approach is that JO or MO can be embedded within larger optimization routines, eg an EM algorithm. For example, in a missing data prob-

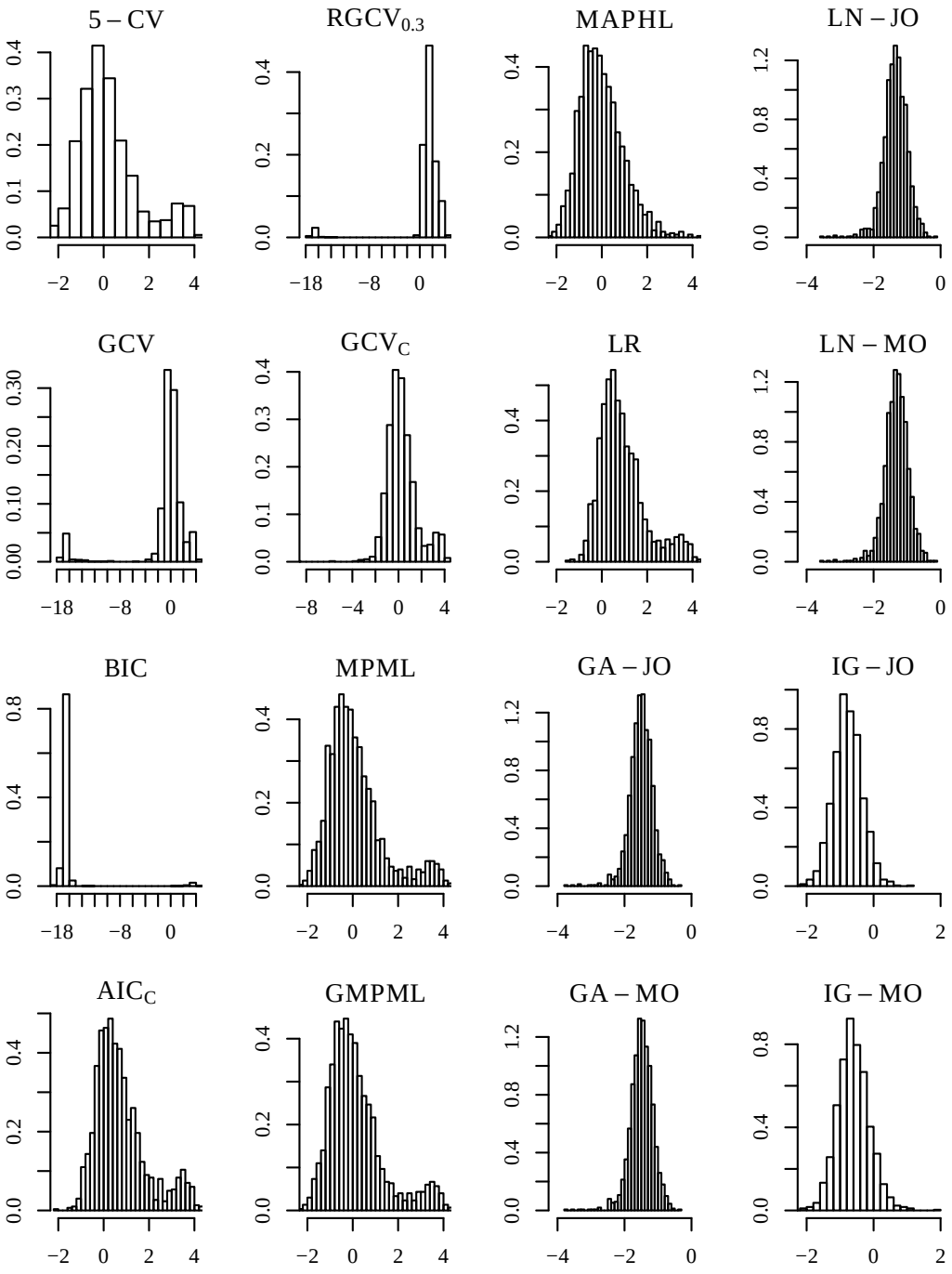

Figure 8: Histograms of $\ln(\lambda/\lambda_{\text{opt}})$ for $\boldsymbol{\beta} = \{1\}_{j=1}^{99}$, $\rho_{j_1 j_2} = 0.75^{|j_1 - j_2|}$, $n = 100$ and $R^2 = 0.1$. $\ln(\lambda/\lambda_{\text{opt}}) = 0$ means that λ was chosen to yield optimal shrinkage.

lem such as in the analysis of the lung adenocarcinoma data, it is not immediately clear how one might do ridge regression concurrently using goodness-of-fit or marginal likelihood approaches. On the other hand, by taking advantage of the conditional independence, specified by the hierarchical framework, between λ and any missing data given the remaining parameters, it is relatively straightforward to embed a JO- or MO-type update for λ within a larger EM algorithm. This is discussed in more detail in Chapter 5.

We have not given a formal definition of "small n", which has implications for recommending a hyperpenalty-based approach such as IG-JO over existing approaches. The asymptotic optimality of the goodness-of-fit methods (5-CV, GCV, AIC_C, and GCV_C) is apparent in our simulation study, because it is when $n = 200$ that we see these methods start to have the smallest rMSPE. So as to make recommendations, we investigated this more thoroughly with a comparison of AIC_C, the best performing existing method, to IG-JO and IG-MO, the best performing proposed methods. In the results summarized in Table 10, p, the length of $\boldsymbol{\beta}$, ranges from 4 to 99; in all cases, $\boldsymbol{\beta}$ is randomly drawn for each simulation from the mixture density described in Section 4.5. In addition to the previous choices of n, we consider $n = 1000$, thereby including the traditional regression situation where n is much larger than p. The criteria for each pairwise comparison are respectively $1000 \times (\text{MSPE}(\lambda_{\text{AIC}_C})/\text{MSPE}(\lambda_{\text{IG-JO}}) - 1)$ and $1000 \times (\text{MSPE}(\lambda_{\text{AIC}_C})/\text{MSPE}(\lambda_{\text{IG-MO}}) - 1)$, so that negative values favor AIC_C. Several insights come from these results. First, when $R^2 = 0.1$, IG-MO does better in a pairwise comparison with AIC_C than IG-JO, as evidenced by larger numbers, but they are about equally good for $R^2 = 0.3$ and $R^2 = 0.5$. Second, typically both IG-JO and IG-MO are equivalent or preferred to AIC_C when $n \approx p$ or $n < p$; in the case of IG-JO, the one exception is the very top left panel: when $\rho_{j_1 j_2} = 0.75^{|j_1 - j_2|}$ and $R^2 = 0.1$. Also of interest is a comparison of the well-known and more frequently used GCV with our proposed methods; we do not present the numerical results, but the interpretation is the same, albeit more dramatic. That is, IG-JO and IG-MO are even more

favored over GCV than AIC_C when $n \approx p$ or $n < p$.

That IG-JO and IG-MO do best when n is small is a reflection of the simple choices of shape and rate parameters in the hyperpenalties. Given proper justification, more complicated choices (eg having the shape and rate depend on both n and p) would make the hyperpenalties even more flexible for large-n situations. However, our studies suggest that a reasonable rule of thumb is that the proposed hyperpenalties, equipped with the suggested shape and rate, are acceptable for a broad range of $n \leq p$ situations.

AIC_C vs. IG-JO		$R^2 = 0.1$					$R^2 = 0.3$					$R^2 = 0.5$				
		$p = 4$	24	49	74	99	$p = 4$	24	49	74	99	$p = 4$	24	49	74	99
$\rho_{j_1 j_2} = 0.75^{\lvert j_1 - j_2 \rvert}$	$n = 25$	-21	-25	-13	-6	-4	27	37	50	54	51	31	66	93	97	103
	50	-7	-21	-11	-6	-3	7	12	21	33	35	8	20	38	46	50
	100	-1	-17	-15	-8	-7	1	0	5	8	16	2	3	10	15	21
	200	-2	-14	-14	-13	-12	0	-3	-2	1	2	1	0	2	4	7
	1000	0	-4	-8	-10	-11	0	-1	-3	-4	-5	0	0	-1	-1	-1
$\rho_{j_1 j_2} = 0.75$	$n = 25$	-11	12	13	17	17	20	25	26	17	16	24	26	17	14	11
	50	-5	5	9	10	9	5	4	7	5	4	6	4	1	4	2
	100	-2	-7	3	4	4	1	-2	0	0	1	3	-6	-6	-6	-5
	200	-3	-14	-6	0	1	1	-8	-5	-2	-2	1	-5	-8	-9	-9
	1000	0	-6	-12	-15	-14	0	-3	-9	-11	-11	0	-2	-5	-8	-9
AIC_C vs. IG-MO		$R^2 = 0.1$					$R^2 = 0.3$					$R^2 = 0.5$				
		$p = 4$	24	49	74	99	$p = 4$	24	49	74	99	$p = 4$	24	49	74	99
$\rho_{j_1 j_2} = 0.75^{\lvert j_1 - j_2 \rvert}$	$n = 25$	1	-7	-3	0	0	33	41	48	49	45	29	59	81	85	90
	50	3	-7	-2	0	1	10	16	22	31	32	8	18	33	39	43
	100	2	-8	-8	-3	-2	2	4	7	9	16	2	4	9	14	18
	200	0	-9	-10	-9	-8	1	-1	-1	2	3	1	0	2	4	6
	1000	0	-3	-7	-9	-10	0	-1	-2	-3	-4	0	0	0	-1	-1
$\rho_{j_1 j_2} = 0.75$	$n = 25$	10	15	11	12	10	29	21	17	9	7	23	17	9	6	4
	50	6	11	8	7	6	9	4	4	1	0	6	-1	-3	0	-1
	100	2	0	5	3	3	2	1	-1	-2	-1	3	-5	-7	-8	-7
	200	-1	-8	-2	1	2	1	-5	-3	-2	-2	1	-4	-7	-10	-10
	1000	0	-5	-10	-13	-12	0	-2	-8	-9	-10	0	-1	-5	-7	-9

Table 10: A pairwise comparison of AIC_C to IG-JO (top table) and AIC_C to IG-MO (bottom): respectively, the average over 1500 simulations of $1000 \times (\text{MSPE}(\lambda_{\text{AIC}_C})/\text{MSPE}(\lambda_{\text{IG-JO}}) - 1)$ and $1000 \times (\text{MSPE}(\lambda_{\text{AIC}_C})/\text{MSPE}(\lambda_{\text{IG-MO}}) - 1)$, rounded to the nearest integer, using the definition of $\text{MSPE}(\lambda)$ in (61). Thus, negative values favor AIC_C. The number p is the dimension of $\boldsymbol{\beta}$, and $\beta_i \sim f$, independently for $i = 1, \ldots, p$, where f is the mixture distribution presented in Section 4.5. $\boldsymbol{\beta}$ is randomly drawn from this mixture distribution with each simulated dataset. n, R^2 and $\rho_{j_1 j_2}$ are also as described in Section 4.5.

CHAPTER 5

A Corrected Hybrid Estimator, Hyperpenalties & Missing Data, and Conclusion

5.1 Introduction

The purpose of this final chapter is threefold. First, we we borrow ideas from Chapters 2–4 to formulate two new approaches to the missing data problem of interest. Specifically, in Section 5.2, we apply the corrected GCV criterion from Chapter 4 to the construction of the hybrid estimator from Chapter 2 to improve its prediction performance in finite samples. In Section 5.3, we propose the hyperpenalized Expectation-Maximization (HEM) algorithm, which extends the penalized EM algorithm (Dempster et al., 1977; Green, 1990) using the hyperpenalized log-likelihood from Chapter 4. This allows for adaptive estimation of shrinkage parameters in a maximum likelihood context with missing data. Our second aim in this chapter is to compare the prediction error corresponding to the best performing methods from Chapters 2 and 3, namely HYB, FB-HIBETA-NI, and EB-HIBETASIGMAX, with these newly proposed ideas. This comparison across methods is presented in Section 5.4. All missing data methods from the dissertation are annotated in Tables D1 and D2 of Appendix D. Third, in Section 5.5 we summarize the contributions of this dissertation and outline several future directions, beyond the scope of this document, that would be worthwhile to undertake.

5.2 Hybrid Estimators Using the Corrected GCV

The corrected hybrid estimator, denoted as HYB$_C$, is a modified version of the hybrid method HYB proposed in Chapter 2. The idea is the same, namely to calculate a linear combination of estimates of $\boldsymbol{\beta}$ from RIDG, SRC and FRC with weights that are data-adaptively estimated using a GCV criterion. We call this "corrected" because it applies the corrected GCV criterion from Section 4.2.1. This correction occurs at two places. First, recall that the regular ridge regression estimate of $\boldsymbol{\beta}$, which only uses subsample A, is a component of HYB. In Chapter 2, the standard GCV criterion was used to select the tuning parameter λ. It is now selected to minimize the following criterion:

$$\frac{\frac{1}{n_{\mathrm{A}}}(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{H}(\lambda \boldsymbol{I}_p)\boldsymbol{y}_{\mathrm{A}})^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{H}(\lambda \boldsymbol{I}_p)\boldsymbol{y}_{\mathrm{A}})}{(1 - \mathrm{Tr}\ \boldsymbol{H}(\lambda \boldsymbol{I}_p)/n_{\mathrm{A}} - 1/n_{\mathrm{A}})^2}, \qquad \boldsymbol{H}(\boldsymbol{\Theta}) = \boldsymbol{x}_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Theta})^{-1}\boldsymbol{x}_{\mathrm{A}}^\top. \tag{62}$$

This differs from the original GCV, given in (5), by the "$-1/n_{\mathrm{A}}$" term in the denominator, accounting for estimation of σ^2. The other components corresponding to SRC and FRC are calculated exactly as before.

Now, given each component, the MSPE of the hybrid estimator is expressed as a quadratic form, $\boldsymbol{\omega}^\top \boldsymbol{P}\boldsymbol{\omega}$, where $\boldsymbol{P}$ is a positive definite prediction error matrix (Section 2.3). Using the corrected GCV, the (i, j)th element of $\boldsymbol{P}$ is now estimated by

$$\hat{P}_{ij} = \frac{\frac{1}{n_{\mathrm{A}}}(\boldsymbol{y}^*_{\mathrm{A},i} - \boldsymbol{H}(\lambda_i \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},i})\boldsymbol{y}^*_{\mathrm{A},i})^\top(\boldsymbol{y}^*_{\mathrm{A},j} - \boldsymbol{H}(\lambda_j \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},j})\boldsymbol{y}^*_{\mathrm{A},j})}{(1 - \psi_i - 1/n_{\mathrm{A}})(1 - \psi_j - 1/n_{\mathrm{A}})}. \tag{63}$$

Comparing this to our original estimate of $\boldsymbol{P}$, as given in (14), the two expressions differ by the "$-1/n_{\mathrm{A}}$" terms in the denominator.

REMARK 15: The corrected GCV given in (62) implicitly assumes β_0 is known, an assumption also made in Chapter 2 for purposes of methodological development. In contrast, in Chapter 4 we assumed β_0 was unknown and to be estimated concurrently with $\boldsymbol{\beta}$.

In applications, β_0 is unknown, and thus the adjustment in (62) and (63) must be by "$-2/n_{\mathrm{A}}$" rather than "$-1/n_{\mathrm{A}}$" to account for this.

5.3 Hyperpenalized EM Algorithm

Here we use notation introduced in Chapter 3. Given observed data $\boldsymbol{U}^{\mathrm{obs}}$ and a set of model parameters $\boldsymbol{\phi}$, the EM algorithm (Dempster et al., 1977) indirectly maximizes a log-likelihood $\ell_O = \ln[\boldsymbol{U}^{\mathrm{obs}}|\boldsymbol{\phi}]$ with respect to $\boldsymbol{\phi}$ that is difficult to maximize directly. It does so by introducing missing data $\boldsymbol{U}^{\mathrm{mis}}$ in such a way that the *complete* log-likelihood $\ell_C = \ln[\boldsymbol{U}^{\mathrm{obs}}, \boldsymbol{U}^{\mathrm{mis}}|\boldsymbol{\phi}]$ is easy to calculate. This is done through successive iterations of the E-step, which calculates $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) = \mathrm{E}[\ell_C|\boldsymbol{\phi}^{(t)}, \boldsymbol{U}^{\mathrm{obs}}]$, where $\boldsymbol{\phi}^{(t)}$ is the maximized value of $\boldsymbol{\phi}$ from the previous iteration, and the M-step, which calculates $\boldsymbol{\phi}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\phi}} Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})$. The penalized EM algorithm (Green, 1990) modifies the M-step with $\boldsymbol{\phi}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\phi}}\{Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) + p_{\boldsymbol{\eta}}(\boldsymbol{\phi})\}$, with $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ a penalty function indexed by a set of given hyperparameters $\boldsymbol{\eta}$. Our proposed *hyper*penalized EM algorithm allows $\boldsymbol{\eta}$ to be unknown, instead giving support to a range of values by way of a hyperpenalty, $h(\boldsymbol{\eta})$. Both $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are sequentially updated, as in the joint optimization (JO) update of Chapter 4. Thus one iteration of the HEM algorithm proceeds as follows:

$$\textbf{E-step} \quad Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) = \mathrm{E}[\ell_C|\boldsymbol{\phi}^{(t)}, \boldsymbol{U}^{\mathrm{obs}}]. \tag{64}$$

$$\textbf{First M-step} \quad \boldsymbol{\phi}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\phi}}\{Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) + p_{\boldsymbol{\eta}^{(t)}}(\boldsymbol{\phi})\}. \tag{65}$$

$$\textbf{Second M-step} \quad \boldsymbol{\eta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\eta}}\{p_{\boldsymbol{\eta}}(\boldsymbol{\phi}^{(t+1)}) + h(\boldsymbol{\eta})\}. \tag{66}$$

As we noted before, the M-steps are generalizations of the JO scheme of Chapter 4. The first M-step updates $\boldsymbol{\phi}$, the model parameters, fixing the current value of $\boldsymbol{\eta}$. The second updates $\boldsymbol{\eta}$, the hyperparameters, fixing $\boldsymbol{\phi}$ at its current value. We consider applications of this algorithm to the prediction problem considered in Chapters 2 and 3. We make

the same modeling assumptions as in Chapters 2 and 3, namely

$$Y|\boldsymbol{X} \sim N\{\beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2\}, \quad \boldsymbol{W}|\boldsymbol{X} \sim N_p\{\psi \mathbf{1}_p + \nu \boldsymbol{X}, \tau^2 \boldsymbol{I}_p\}, \quad \boldsymbol{X} \sim N_p\{\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}.$$

Thus, we have parameters collectively denoted by $\boldsymbol{\phi} = \{\boldsymbol{\beta}, \beta_0, \sigma^2, \psi, \nu, \tau^2, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}$, observed data $\boldsymbol{U}^{\text{obs}} = \{\boldsymbol{y}_\text{A}, \boldsymbol{y}_\text{B}, \boldsymbol{x}_\text{A}, \boldsymbol{w}_\text{A}, \boldsymbol{w}_\text{B}\}$, and missing data $\boldsymbol{U}^{\text{mis}} = \boldsymbol{x}_\text{B}$. The *observed* log-likelihood of the data is

$$\begin{aligned}\ell_O = \ln[\boldsymbol{U}^{\text{obs}}|\boldsymbol{\phi}] = \ln[\boldsymbol{y}_\text{A}|\boldsymbol{x}_\text{A}, \boldsymbol{\beta}, \beta_0, \sigma^2] + \ln[\boldsymbol{w}_\text{A}|\boldsymbol{x}_\text{A}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_\text{A}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] \\ + \ln \int [\boldsymbol{y}_\text{B}|\boldsymbol{x}_\text{B}, \boldsymbol{\beta}, \beta_0, \sigma^2][\boldsymbol{w}_\text{B}|\boldsymbol{x}_\text{B}, \psi, \nu, \tau^2][\boldsymbol{x}_\text{B}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] \, \mathrm{d}\boldsymbol{x}_\text{B}. \qquad (67)\end{aligned}$$

Although this may be calculated analytically, the repeated calculations that a direct maximization would require are computationally difficult. On the other hand, the complete-data log-likelihood, that is, the log-likelihood if $\boldsymbol{x}_\text{B}$ were observed, is equivalent to (25) from Chapter 3 and given by

$$\begin{aligned}\ell_C = \ln[\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}}|\boldsymbol{\phi}] = \ln[\boldsymbol{y}_\text{A}|\boldsymbol{x}_\text{A}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_\text{A}|\boldsymbol{x}_\text{A}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_\text{A}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] \\ + \ln[\boldsymbol{y}_\text{B}|\boldsymbol{x}_\text{B}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_\text{B}|\boldsymbol{x}_\text{B}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_\text{B}|\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X].\end{aligned}$$

This is straightforward to calculate and forms the basis of the E- and M-steps, which we now derive.

5.3.1 E-Step

We adapt the following notation from the Data Augmentation step given in (26), which the E-step closely parallels. Let $\boldsymbol{\Gamma}(\boldsymbol{\phi}) = (\boldsymbol{\beta}\boldsymbol{\beta}^\top/\sigma^2 + (\nu^2/\tau^2)\boldsymbol{I}_p + \boldsymbol{\Sigma}_\text{X}^{-1})^{-1}$, $\boldsymbol{x}_\text{B}^{\text{EM}}(\boldsymbol{\phi}) = ([\boldsymbol{y}_\text{B} - \mathbf{1}_{n_\text{B}}\beta_0]\boldsymbol{\beta}^\top/\sigma^2 + [\nu/\tau^2][\boldsymbol{w}_\text{B} - \psi\mathbf{1}_{n_\text{B}}\mathbf{1}_p^\top] + [\mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_\text{X}^\top]\boldsymbol{\Sigma}_\text{X}^{-1})\boldsymbol{\Gamma}(\boldsymbol{\phi})$, and, given $\boldsymbol{\phi}^{(t)}$ (the current estimates of $\boldsymbol{\phi}$), $\boldsymbol{\Gamma}^{(t)} = \boldsymbol{\Gamma}(\boldsymbol{\phi}^{(t)})$ and $\boldsymbol{x}_\text{B}^{\text{EM}(t)} = \boldsymbol{x}_\text{B}^{\text{EM}}(\boldsymbol{\phi}^{(t)})$. From Appendix D.1, the

expected complete log-likelihood is given by

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= \mathrm{E}[\ell_C|\boldsymbol{\phi}^{(t)}, \boldsymbol{U}^{\text{obs}}] \\
&= -\frac{n_\mathrm{A}+n_\mathrm{B}}{2}\log(\sigma^2\tau^{2p}|\boldsymbol{\Sigma}_{\boldsymbol{X}}|) \\
&\quad - \frac{1}{2\sigma^2}(\boldsymbol{y}_\mathrm{A} - \beta_0\mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top(\boldsymbol{y}_\mathrm{A} - \beta_0\mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta}) \\
&\quad - \frac{1}{2\sigma^2}(\boldsymbol{y}_\mathrm{B} - \beta_0\mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top(\boldsymbol{y}_\mathrm{B} - \beta_0\mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta}) \\
&\quad - \frac{1}{2\tau^2}\mathrm{Tr}\left[(\boldsymbol{w}_\mathrm{A} - \psi\mathbf{1}_{n_\mathrm{A}}\mathbf{1}_p^\top - \nu\boldsymbol{x}_\mathrm{A})^\top(\boldsymbol{w}_\mathrm{A} - \psi\mathbf{1}_{n_\mathrm{A}}\mathbf{1}_p^\top - \nu\boldsymbol{x}_\mathrm{A})\right] \\
&\quad - \frac{1}{2\tau^2}\mathrm{Tr}\left[(\boldsymbol{w}_\mathrm{B} - \psi\mathbf{1}_{n_\mathrm{B}}\mathbf{1}_p^\top - \nu\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)})^\top(\boldsymbol{w}_\mathrm{B} - \psi\mathbf{1}_{n_\mathrm{B}}\mathbf{1}_p^\top - \nu\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)})\right] \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top)\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top)^\top\right] \\
&\quad - \frac{1}{2}\mathrm{Tr}\left[(\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top)\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_{\boldsymbol{X}}^\top)^\top\right] \\
&\quad - \frac{n_\mathrm{B}}{2}\mathrm{Tr}\left(\boldsymbol{\Gamma}^{(t)}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\phi})\right).
\end{aligned} \tag{68}
$$

Calculation of the E-step does not depend on the hyperparameters $\boldsymbol{\eta}$ and therefore remains the same for any choice of $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ and, if required, $h(\boldsymbol{\eta})$.

5.3.2 M-Steps

Recalling the components of the general form of the M-steps in (65) and (66), we consider two choices of penalty $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$; these correspond to the priors on $\boldsymbol{\phi}$ in Chapter 3. The first penalty fixes $\boldsymbol{\eta}$ at a given value and therefore does not require a hyperpenalty $h(\boldsymbol{\eta})$. The second choice of $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ allows $\boldsymbol{\eta}$ to be unspecified and results in data-adaptive shrinkage, as in the analogous Bayesian methods. In this latter case, a hyperpenalty $h(\boldsymbol{\eta})$ must be selected, and we consider three choices, borrowing ideas from Chapter 4.

EM-FLATBETA Maximizing Q given in (68) may result in an infinite likelihood when neither subsamples A nor B contain enough information about $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$. Specifically, this will happen when $p > n_\mathrm{A}$, so the sample covariance of $\boldsymbol{x}_\mathrm{A}$ is singular and, simultaneously, when τ/ν is too large, meaning the surrogate $\boldsymbol{w}_\mathrm{B}$ is too noisy. In these cases, the esti-

mate for $\boldsymbol{\Sigma}_X^{-1}$ approaches singularity. To address this problem of singularity, we mildly penalize the estimate of $\boldsymbol{\Sigma}_X^{-1}$:

$$p_{\boldsymbol{\eta}}(\boldsymbol{\phi}) = \frac{2p-1}{2}\log|\boldsymbol{\Sigma}_X^{-1}| - \frac{2p-1}{2}\text{Tr}\,(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1}). \tag{69}$$

Comparing this to (27), this choice of $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ uses the log-density of the same mildly informative Wishart prior that was placed on $\boldsymbol{\Sigma}_X^{-1}$ in the Bayesian FB-FLATBETA method. The penalty parameter $\boldsymbol{\eta}$ is given, so this is simply a penalized EM algorithm as first outlined by Green (1990). This small amount of shrinkage induced by $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ is necessary to make the algorithm proceed, just as in the Bayesian FB-FLATBETA method of Chapter 3. Following Meng and Rubin (1993), we divide $\boldsymbol{\phi}$ into subvectors and use conditional penalized M-steps to update each component of $\boldsymbol{\phi}$ individually; these are derived in the Appendix. We give only the resulting M-steps for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_X^{-1}$ here:

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A} + \boldsymbol{x}_\text{B}^{\text{EM}(t)\top}\boldsymbol{x}_\text{B}^{\text{EM}(t)} + n_\text{B}\boldsymbol{\Gamma}^{(t)})^{-1}(\boldsymbol{x}_\text{A}^\top[\boldsymbol{y}_\text{A} - \beta_0\mathbf{1}_{n_\text{A}}] + \boldsymbol{x}_\text{B}^{\text{EM}(t)}[\boldsymbol{y}_\text{B} - \beta_0\mathbf{1}_{n_\text{A}}]),\\
\boldsymbol{\Sigma}_X^{-1(t+1)} &= \Big(\frac{(\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{A} - \mathbf{1}_{n_\text{A}}\boldsymbol{\mu}_X^\top) + \boldsymbol{x}_\text{B}^{\text{EM}(t)} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\text{B}^{\text{EM}(t)} - \mathbf{1}_{n_\text{B}}\boldsymbol{\mu}_X^\top)}{n_\text{A}+n_\text{B}+2p-1}\\
&\quad + \frac{n_\text{B}\boldsymbol{\Gamma}^{(t)} + (2p-1)\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])}{n_\text{A}+n_\text{B}+2p-1}\Big)^{-1}.
\end{aligned}$$

Because these are conditional penalized M-steps, any element of $\boldsymbol{\phi}$ on the right-hand side of the above equations is replaced with its value at the previous iteration, eg β_0 is replaced with $\beta_0^{(t)}$ in the $\boldsymbol{\beta}$ update.

EM-HIBETA-GA, EM-HIBETA-LN, EM-HIBETA-IG These are analogs to the Bayesian ridge methods. In all three cases, then, the penalty function is expanded as follows:

$$\begin{aligned}
p_{\boldsymbol{\eta}}(\boldsymbol{\phi}) = &\frac{2p-1}{2}\log|\boldsymbol{\Sigma}_X^{-1}| - \frac{2p-1}{2}\text{Tr}\,(\text{diag}(\hat{\text{Var}}[\boldsymbol{x}_\text{A}])\boldsymbol{\Sigma}_X^{-1})\\
&- \frac{p}{2}\ln(\sigma^2) + \frac{p}{2}\ln(\lambda) - \frac{1}{2\sigma^2}\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}.
\end{aligned} \tag{70}$$

This adds a Normal log-density term to the penalty function of EM-FLATBETA as given in (69). Now, $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ contains a (ridge) parameter λ, that is, $\boldsymbol{\eta} = \{\lambda\}$, for which a value must be chosen. Given λ, the M-step for $\boldsymbol{\beta}$ is modified as follows:

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)} + n_{\mathrm{B}}\boldsymbol{\Gamma}^{(t)} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{x}_{\mathrm{A}}^{\top}[\boldsymbol{y}_{\mathrm{A}} - \beta_0\mathbf{1}_{n_{\mathrm{A}}}] + \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}[\boldsymbol{y}_{\mathrm{B}} - \beta_0\mathbf{1}_{n_{\mathrm{A}}}]).$$

We adaptively choose λ using a hyperpenalized M-step as in (66). In Chapter 4, we considered three choices of hyperpenalty based on the Gamma, Log-Normal, and Inverse-Gamma distributions. Each is indexed by shape and rate parameters (a and b, respectively). We have

$$\begin{aligned}
h_{\mathrm{GA}}(\boldsymbol{\eta}) &= (a-1)\ln(\lambda) - \lambda/b, \\
h_{\mathrm{LN}}(\boldsymbol{\eta}) &= -\ln(\lambda) - \ln(b\lambda)^2/(2a), \\
h_{\mathrm{IG}}(\boldsymbol{\eta}) &= -(a+1)\ln(\lambda) - 1/(b\lambda).
\end{aligned}$$

In each case, we chose a and b to satisfy the moment-matching conditions $\mathrm{E}\ln(\lambda) = \ln(p)$ and $\mathrm{Var}\ln(\lambda) = \psi_1(p/2+1)$, where p is the length of $\boldsymbol{\beta}$, based on the discussion in Section 4.4.4. The hyperpenalized M-steps are respectively as follows (no closed-form exists for the Log-Normal):

$$\begin{aligned}
\textsc{em-hibeta-ga}: &\quad \lambda^{(t+1)} = \frac{p + 2a - 2}{\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/\sigma^{2(t)} + 2b}, \\
\textsc{em-hibeta-ln}: &\quad \lambda^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\eta}}\{p_{\boldsymbol{\eta}}(\boldsymbol{\phi}^{(t)}) + h_{\mathrm{LN}}(\boldsymbol{\eta})\}, \\
\textsc{em-hibeta-ig}: &\quad \lambda^{(t+1)} = \frac{p - 2a - 2 + \sqrt{(p-2a-2)^2 + 8\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/(b\sigma^{2(t)})}}{2\boldsymbol{\beta}^{(t)\top}\boldsymbol{\beta}^{(t)}/\sigma^{2(t)}}.
\end{aligned}$$

In summary, we have a penalized EM algorithm, EM-FLATBETA, and three implementations of the HEM algorithm, EM-HIBETA-GA, EM-HIBETA-LN, and EM-HIBETA-IG, based on different choices of the penalty $p_{\boldsymbol{\eta}}(\boldsymbol{\phi})$ and hyperpenalty $h(\boldsymbol{\eta})$.

5.3.3 Bayesian Ridge with an Informative Hyperprior

FB-HIBETA-NI is the Bayesian ridge that places a Jeffreys hyperprior on λ, equivalently $\lambda \sim G\{0, 0\}$, ie a Gamma distribution with shape and rate equal to zero. In contrast, we define FB-HIBETA-GA to be a Bayesian ridge that places a $G\{a, b\}$ hyperprior on λ, using the same shape and rate as the Gamma hyperpenalty of Chapter 4. Thus, using the language of Section 3.2, it is a **DA+** variant, like FB-HIBETA-NI, and λ is iteratively sampled in conjunction with the other unknown quantities: $\boldsymbol{U}^{\text{obs}}$ and $\boldsymbol{\phi}$. Because the Gamma distribution is conjugate to Normal precision, the posterior step for λ is straightforward and available in closed-form: $\lambda \sim G\{a + p/2, b + \boldsymbol{\beta}^\top \boldsymbol{\beta}/(2\sigma^2)\}$. Apart from this link between FB-HIBETA-GA and FB-HIBETA-NI, we note also that FB-HIBETA-GA explicitly parallels EM-HIBETA-GA: the log-density of the hyperprior in FB-HIBETA-GA is equal to the hyperpenalty in EM-HIBETA-GA. The crucial difference is that the former calculates parameter and hyperparameter estimates by averaging over the posterior and the latter does so by finding the maximum of the posterior, what we also call the hyperpenalized likelihood.

5.4 Comparison Across Methods

In this section, we conduct a dissertation-wide comparison of prediction error from the same simulation study as in Chapter 3 and the lung adenocarcinoma data considered throughout. See Section 3.5 for details on the construction of the simulated datasets and Tables D1 and D2 in Appendix D for concise descriptions and annotations of all the methods. A total of nine methods are evaluated here: HYB (Section 2.3), the corrected HYB (HYB_C, Section 5.2), FB-HIBETA-NI (Section 3.3.1), FB-HIBETA-GA (Section 5.3.3), EB-HIBETASIGMAX (Section 3.3.2), EM-FLATBETA, EM-HIBETA-GA, EM-HIBETA-LN, and EM-HIBETA-IG (Section 5.3.2).

Figure 9 gives empirical MSPEs from the simulation study. For clarity in graphically

presenting results, EM-FLATBETA and EM-HIBETA-LN, which typically perform no better than EM-HIBETA-GA or EM-HIBETA-IG, are not plotted. Based on these results, we make the following observations.

(i) In the $p = 99$ case, the best-predicting method is always EB-HIBETASIGMAX or EM-HIBETA-IG. In the $p = 5$ case, the relative differences between all methods are smaller; HYB and HYB_C predict the worst, and FB-HIBETA-NI, FB-HIBETA-GA, EB-HIBETASIGMAX, EM-HIBETA-GA, and EM-HIBETA-IG are about equally good.

(ii) HYB_C is uniformly better than HYB when $p = 99$, and the two are about equal when $p = 5$.

(iii) FB-HIBETA-GA is uniformly better than FB-HIBETA-NI, indicating that the choice of informative hyperprior on λ improves prediction. Moreover, there are some significant differences between EM-HIBETA-GA and FB-HIBETA-GA. Specifically, EM-HIBETA-GA is almost always preferred to FB-HIBETA-GA when $p = 99$. Thus, maximizing typically yields smaller prediction error than averaging in the high-dimensional case.

(iv) The best performing of the HEM algorithms is EM-HIBETA-IG. As we saw in the simulation study in Chapter 4, the non-conjugate inverse-gamma hyperpenalty predicts considerably better than the conjugate gamma hyperpenalty.

Table 11 gives empirical MSPE from the validation sample of the lung adenocarcinoma data. Here EB-HIBETASIGMAX has the smallest MSPE (0.555), followed by EM-HIBETA-IG (0.597) and HYB (0.601). Consistent with the simulation study results in Figure 9, particularly the small R^2, large n_{B} panels, the best methods are EB-HIBETASIGMAX and EM-HIBETA-IG. In contrast with the simulation study results, HYB_C does not predict better than the original HYB method. The benefit conferred by using HYB_C comes from average performance over many datasets. Differences in average performance between HYB and HYB_C are driven by occasional large differences in prediction error that occur when the

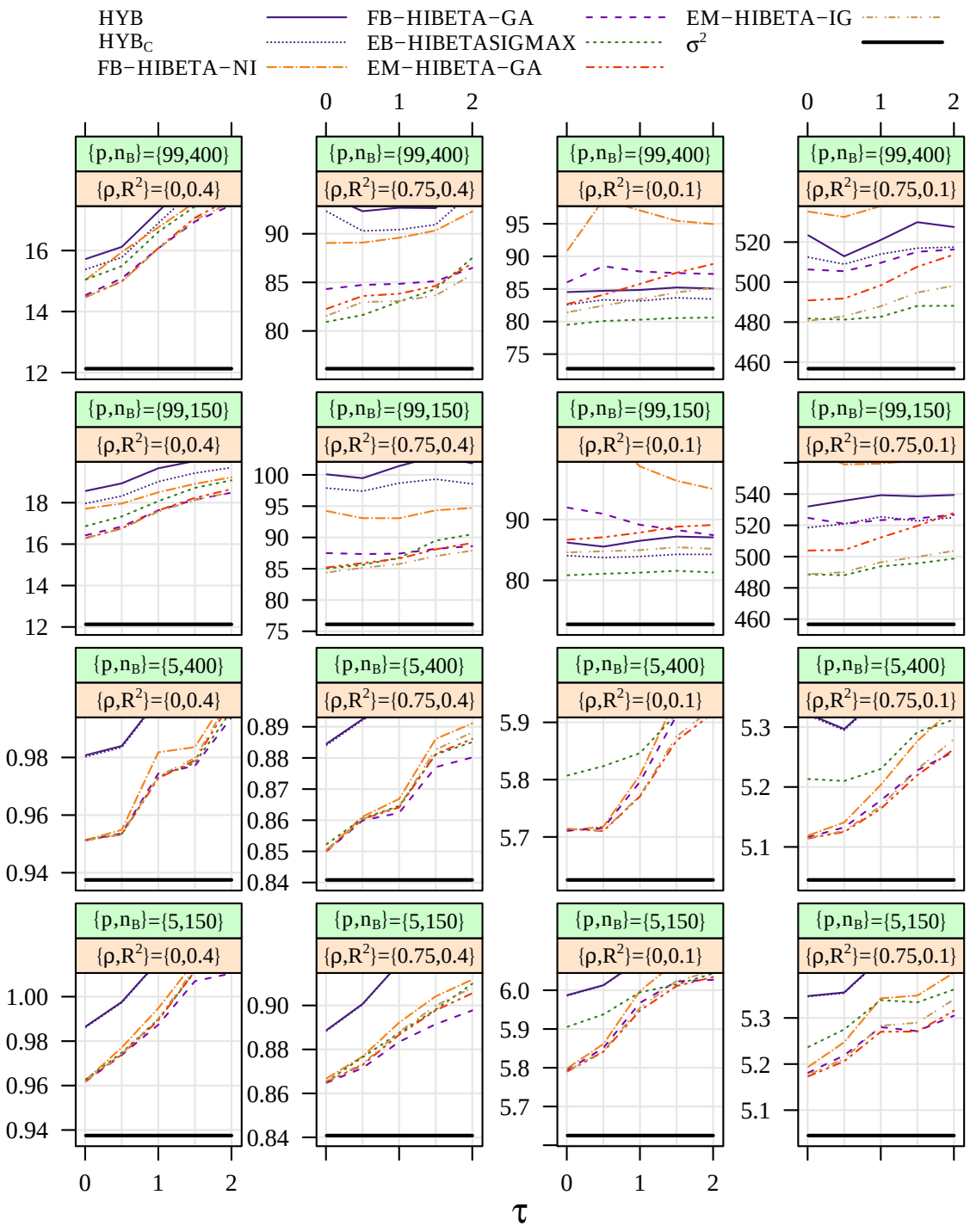

Figure 9: Empirical MSPE of the dissertation's best performing methods over τ for 16 simulation settings. For the Bayesian methods, $\boldsymbol{\beta}$ was estimated with the posterior predictive mean, $\hat{\boldsymbol{\beta}}^{\text{ppm}}$. The thick, solid bar ($\sigma^2$) corresponds to predictions made using the true generating parameters.

	HYB	HYB$_C$	FB-HIBETA-NI	FB-HIBETA-GA	EB-HIBETASIGMAX	EM-FLATBETA	EM-HIBETA-GA	EM-HIBETA-LN	EM-HIBETA-IG
$\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$	0.601	0.617	0.793	0.636	0.555	3.909	0.665	0.642	0.597
$\min(\hat{\boldsymbol{\beta}})$	-0.023	-0.024	-0.041	-0.031	-0.007	-0.156	-0.029	-0.027	-0.023
$\max(\hat{\boldsymbol{\beta}})$	0.027	0.032	0.064	0.026	0.007	0.153	0.034	0.031	0.023
Runtime (sec)	1.4	1.5	404.6	398.6	402.6	71.4	82.6	57.4	70.5

Table 11: Numerical results from analysis of the lung adenocarcinoma data using the dissertation's best-performing methods. $\hat{\text{MSPE}}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ is the empirical MSPE from the validation sample of size 100, $\min(\hat{\boldsymbol{\beta}})$ and $\max(\hat{\boldsymbol{\beta}})$ give the range of the estimate of $\boldsymbol{\beta}$ for each model, and Runtime is the number of seconds required to run each algorithm. The Bayesian methods, FB-HIBETA-NI, FB-HIBETA-GA and EB-HIBETASIGMAX, use $\hat{\boldsymbol{\beta}}$ corresponding to the posterior predictive mean, defined in (32).

standard GCV overfits. We observe no such overfitting in the lung adenocarcinoma data, and HYB and HYB$_C$ are similar.

Finally, Figure 10 graphically summarizes the analysis of the lung adenocarcinoma data with kernel density estimates of the 91 coefficient estimates of $\boldsymbol{\beta}$. The top figure plots estimates from the same seven methods as in Figure 9, and the bottom figure re-plots the three methods with the best MSPE: EB-HIBETASIGMAX, EM-HIBETA-IG and HYB.

As evidenced by these plots, the shrinkage induced by EB-HIBETASIGMAX is considerably greater than that of EM-HIBETA-IG. However, this is not because EB-HIBETASIGMAX shrinks $\boldsymbol{\Sigma}_X^{-1}$ and EM-HIBETA-IG does not shrink $\boldsymbol{\Sigma}_X^{-1}$: from Figure 7 in Chapter 3, the kernel density estimates corresponding to EB-HIBETASIGMAX and EB-HIBETA-NI, which differ only in whether or not $\boldsymbol{\Sigma}_X^{-1}$ is adaptively shrunk, are nearly the same, which is why EB-HIBETA-NI was not presented in this present comparison. Rather, the difference in kernel density estimates between EB-HIBETASIGMAX and EM-HIBETA-IG in Figure 10 is a function of the algorithm used, ie EB-HIBETASIGMAX is Empirical Bayes, using maximum marginal likelihood estimates for the hyperparameters and posterior sampling for the parameters, and EM-HIBETA-IG is a maximum likelihood approach. EM-HIBETA-IG and HYB *do* yield similar amounts of shrinkage, despite the differences in their construction.

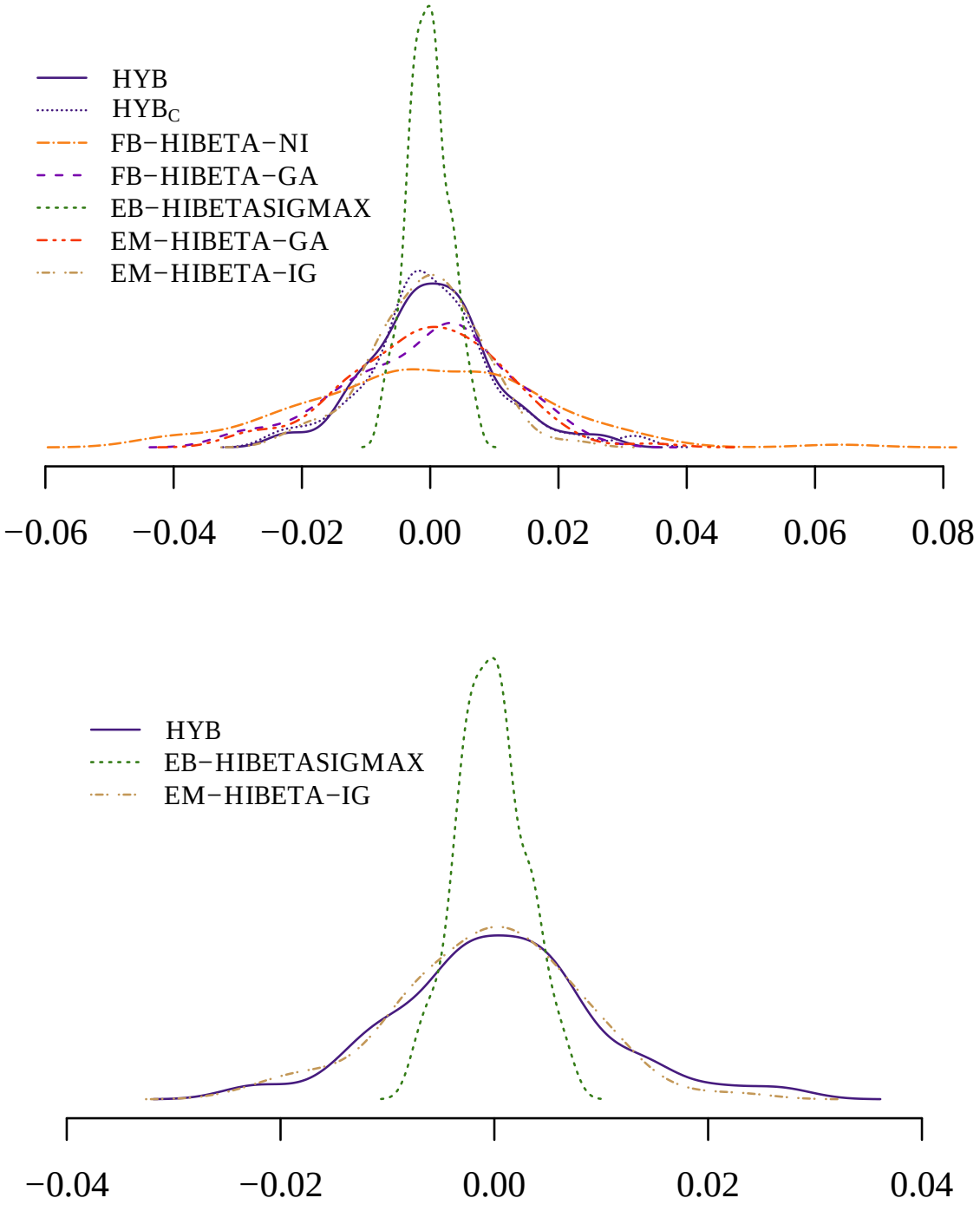

Figure 10: Kernel density estimates of the 91 coefficients from the lung adenocarcinoma data for the dissertation's best performing methods

5.5 Summary and Future Work

We have considered a variety of shrinkage approaches to aid in prediction with high-dimensional data. Represented as a missing data problem, most observations contain only surrogate measurements of the true underlying covariates of interest, and only a few contain measurements of both the true covariates and their surrogates. In this section, we first summarize each chapter and highlight possible future work. We note that many of the ideas between the chapters are interrelated, eg the corrected hybrid estimator from Section 5.2 combines the hybrid estimator from Chapter 2 with the corrected GCV from Chapter 4. In the same way, much of our additional prospective work also seeks to take advantage of interrelated ideas between chapters. Next, we contrast the methods in terms of their limitations and present some general recommendations, taking into account issues like predictive performance, modeling flexibility, and robustness to model misspecification.

5.5.1 Chapter 2

In Chapter 2, we discussed a class of targeted ridge (TR) estimators, which includes as a special case standard ridge regression (RIDG), and a hybrid estimator (HYB), which is a weighted linear combination of multiple TR estimators. HYB possesses several important properties. First, it is flexible, being a linear combination of estimators, each of which can make different modeling assumptions. For example, RIDG assumes only the outcome regression model in (1): $Y|\boldsymbol{X} \sim N\{\beta_0 + \boldsymbol{X}^\top\boldsymbol{\beta}, \sigma^2\}$. FRC additionally assumes the measurement error model in (2): $\boldsymbol{W}|\boldsymbol{X} \sim N_p\{\psi\mathbf{1}_p + \nu\boldsymbol{X}, \tau^2\boldsymbol{I}_p\}$. SRC assumes these two models plus the marginal model: $\boldsymbol{X} \sim N_p\{\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}$. Estimators with different modeling assumptions, beyond what we have proposed in this dissertation, can also be included in HYB. One such estimator could be that which allows for outcome dependent sampling of the observations that comprise subsample A, ie those for which $\boldsymbol{X}$ is measured.

From Theorem 2.1, the hybrid estimator will theoretically predict better than the best of any of its constituents. Practically, the average performance of HYB across many design and data configurations is encouraging, and, importantly, its flexibility is most apparent in the large-p scenarios. Second, HYB is typically very fast to compute, whereas the Bayesian methods of Chapter 3 require considerably more computational effort. Finally, because HYB combines TR estimators, a GCV criterion, which is readily calculated on TR estimators, provides a simple estimate of $\boldsymbol{P}$, the prediction error matrix defined by (13) required to estimate the optimal weight vector $\boldsymbol{\omega}^{\text{opt}}$. In the dissertation-wide comparison of Section 5.4, we used the corrected GCV criterion proposed in Chapter 4 as an alternative, which, from the simulation results, further improved prediction.

5.5.2 Chapter 3

In Chapter 3, we considered Bayesian approaches to this same problem. In this context, shrinkage occurs through specification of priors on the model parameters. The crucial question here is which parameters to shrink, ie choice of prior distributions, and to what extent, ie the choice of hyperparameters. We limited our exploration to shrinkage of the regression coefficients $\boldsymbol{\beta}$ and the marginal precision $\boldsymbol{\Sigma}_X^{-1}$, and it is worthwhile to consider the adaptive shrinkage of other model parameters, like τ^2. The best-predicting methods from Chapter 3 were EB-HIBETA-NI and EB-HIBETASIGMAX, the Bayesian ridge using Empirical Bayes methods to estimate hyperparameters. An alternative course of future work is to apply one of the recommendations from Chapter 4 and equip the fully Bayesian ridge methods, eg FB-HIBETA-NI or FB-HIBETA-GA from Section 5.3.3, with an inverse-gamma hyperprior. The difficulty in this extension is a computational one, because conjugacy between the prior and hyperprior is lost, but there is reason to believe that this approach would be more competitive with the Empirical Bayes approaches, given the performance of EM-HIBETA-IG in Figure 9.

For both Chapters 2 and 3, the adaptive shrinkage of parameters in nonlinear models, such as those with censored or binary outcomes, would also be a valuable extension.

5.5.3 Chapter 4

Motivated by the problem of adaptive estimation of shrinkage parameters, Chapter 4 contributes several novel ideas toward estimating the tuning parameter of a ridge regression. We proposed a modified GCV criterion that corrects the problematic behavior of the standard GCV criterion in ridge regression when $n - 1$, that is, one less than the number of observations, equals p, the dimension of the regression coefficients $\boldsymbol{\beta}$. We described how the standard GCV criterion may induce dramatic overfitting in this case and showed how our corrected GCV effectively fixes this. Improved prediction of the corrected GCV over the standard GCV was observed in many $n \approx p$ situations, and, like the standard GCV criterion, the corrected GCV is asymptotically optimal with n.

The GCV/corrected GCV criteria can be utilized in a ridge regression setup using an entirely different approach. Specifically, rather than adaptively choosing one value of the tuning parameter λ, one could linearly combine multiple estimates of $\boldsymbol{\beta}$ calculated using a supplied grid of values for λ, taking the form of a hybrid estimator from Chapter 2. The role of the corrected GCV would be to estimate the prediction error matrix $\boldsymbol{P}$, which in turn is used to calculate the weight vector $\boldsymbol{\omega}$.

Also in Chapter 4, we proposed the "hyperpenalized" log-likelihood. Added to a penalized log-likelihood is a hyperpenalty, which is equivalent to the log-density of a hyperprior and protects from overfitting by shrinking the shrinkage parameter λ. The hyperpenalized log-likelihood can be maximized jointly with respect to all parameters or maximized after marginalizing over λ. In our simulation studies, the atypical choice of using a hyperpenalty based on the *inverse*-gamma distribution yielded smaller prediction error than the more common gamma distribution. We used simple choices of

the shape and rate of the hyperpenalty function, and there is more work to be done on justifying more complicated choices that would give the hyperpenalty more flexibility. Another extension of this approach is the hyperpenalization of penalty functions used for variable selection, such as the Lasso (Tibshirani, 1996).

5.5.4 Limitations and Recommendations

We broadly classify the potential limitations of our work into three categories, acknowledging that they are not entirely distinct from one another. They are (i) violations to the modeling assumptions, (ii) a more general missing data mechanism, or (iii) alternative likelihood factorizations. Included in the first is misspecification of the error distribution or mean structure of the outcome, measurement error, or marginal models given in (24). Sensitivity to such violations is studied via the simulation studies of Chapters 2 and 3. Results are presented in the Appendices in Figure A4 and Tables B3–B8. In general, all of the methods are fairly robust to these violations. For the Bayesian/Empirical Bayes/hyperpenalized likelihood methods, this robustness highlights the benefit of *adaptive* shrinkage. The hyperprior averages over different models and thus reduces sensitivity to model misspecification.

Focusing on the second category, we have assumed throughout that the missing $\boldsymbol{X}$'s are "missing completely at random" (Little and Rubin, 2002), meaning that the missingness indicator is independent of $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{W}$. In fact, the likelihood-based methods, those described in Chapters 3 and 5, only require independence between the indicator and $\boldsymbol{X}$: from (26) and (64), the imputation steps condition on $\boldsymbol{Y}$ and $\boldsymbol{W}$, so, if instead the indicator depends on $\boldsymbol{Y}$ and/or $\boldsymbol{W}$, they remain equally valid. This contrasts with the TR and HYB methods of Chapter 2, which will be negatively affected, as seen by comparing Figure A3, in which subsample A tends to have larger values of the outcome Y, to Figure 2.

The third category considers alternative likelihood factorizations, of which two naturally

arise. The first allows for $[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{W}] \neq [\boldsymbol{Y}|\boldsymbol{X}]$, an alternative to the more stringent NDME assumption, which is discussed in Section 2.6. The second avoids modeling of the fully observed auxiliary variable $\boldsymbol{W}$: $[Y,\boldsymbol{X}|\boldsymbol{W}] = [Y|\boldsymbol{X}][\boldsymbol{X}|\boldsymbol{W}]$. All methods in this dissertation will break down to some degree if these factorizations more accurately model the true underlying process, as we have assumed a different factorization, namely $[Y,\boldsymbol{X},\boldsymbol{W}] = [Y|\boldsymbol{X}][\boldsymbol{W}|\boldsymbol{X}][\boldsymbol{X}]$. However, we emphasize that our choice of factorization is scientifically motivated by the surrogate and matched relationship between $\boldsymbol{W}$ and $\boldsymbol{X}$. In other words, the ith element of $\boldsymbol{W}$, corresponding to an assay of gene i using microarray technology, is a possibly-scaled, noisier version of the ith element of $\boldsymbol{X}$, the assay of gene i using qRT-PCR, but does not depend on any other element of $\boldsymbol{X}$. Because of this scientific rationale, we do not evaluate in further detail the negative effects of violating this factorization, but this suggests possible refinements of our work.

Integrating all of these individual factors – predictive performance, flexibility, robustness, and other considerations – the Bayesian ridge with an Empirical Bayes update for hyperparameters, EB-HIBETA-NI/EB-HIBETASIGMAX-NI, is a sensible overall choice. Importantly, prediction intervals for quantifying uncertainty are automatic, as outlined in Section 3.4. The obstacles to the Bayesian methods are computational, the most significant being that the algorithms require more advanced programming. In order to efficiently run a Gibbs sampler when p is moderate-to-large, the use of a lower-level language like `C` is required, which mitigates the applicability of our methodology to other problems. Less important but still significant in a simulation study, the Gibbs samplers take longer to run than their non-Bayesian counterparts. Table 11 includes the runtime of each method for the analysis of the lung adenocaricnoma data. The Bayesian methods each took about 7 minutes, compared to about 1 minute for the HEM methods and just 1-2 seconds for the hybrid estimator. Thus, EM-HIBETA-IG maintains most of the advantages of EB-HIBETA-NI/EB-HIBETASIGMAX-NI, the most important being the adaptive shrinkage feature, but with fewer computational challenges. It does not automatically yield auto-

matic prediction intervals, which require variance estimates via the observed information. Segal et al. (1994) propose methodology to calculate the observed information in Green's penalized EM. In principle, this may be extended to the HEM algorithm, yielding variance estimates not only for the parameters but also the hyperparameters. Finally, although HYB never improved upon the more complicated likelihood-based methods, it is computationally straightforward and fast. It is also the most novel of the methods we have considered. As such, there is potential for its improvement, as we have outlined earlier in this section.

In conclusion, we have offered many novel ideas toward this high dimensional prediction problem. Our proposals draw from a broad range of statistical literature, including shrinkage estimation, measurement error, missing data, Bayesian sampling schemes, and penalized regression, and further research can continue the work of this dissertation to connect and build upon each of these.

APPENDICES

Appendix A: Chapter 2 Supplementary Materials

A.1 Analysis of Targeted Ridge Estimators

This section proves some results for TR estimators, first evaluating them as imputations for the missing data, $\boldsymbol{x}_{\text{B}}$, and then evaluating them in terms of MSPE for predicting the outcome Y. Throughout, we condition on the true value of $\boldsymbol{\theta}$ and assume $\boldsymbol{\mu}_{\boldsymbol{X}} = \mathbf{0}_p$.

As demonstrated in their construction, $\hat{\boldsymbol{\beta}}_{\text{SRC}}$ and $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ are equivalent to filling in the missing $\boldsymbol{x}_{\text{B}}$ with $\boldsymbol{x}_{\text{B}}^{\text{SRC}}$ and $\boldsymbol{x}_{\text{B}}^{\text{FRC}}$ and doing OLS on the completed data. Due to Marquardt (1970), RIDG can also be viewed as imputing the missing $\boldsymbol{x}_{\text{B}}$ with $\boldsymbol{x}_{\text{B}}^{\text{RIDG}} = [\sqrt{\lambda}\boldsymbol{I}_p \, \mathbf{0}_p \, \cdots \, \mathbf{0}_p]^\top$, replacing the observed $\boldsymbol{y}_{\text{B}}$ with $\mathbf{0}_{n_{\text{B}}}$, and doing OLS on the completed data. In general, we have the following result for any targeted ridge estimator.

Theorem A.1. *Assuming $n_B > p$, a choice of $(\gamma_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ is equivalent to making imputations $\tilde{\boldsymbol{x}}_B$ and $\tilde{\boldsymbol{y}}_B$ and doing OLS on the* completed *data. That is, $\hat{\boldsymbol{\beta}}(\gamma_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}) = (\boldsymbol{x}_A^\top \boldsymbol{x}_A + \tilde{\boldsymbol{x}}_B^\top \tilde{\boldsymbol{x}}_B)^{-1}(\boldsymbol{x}_A^\top \boldsymbol{y}_A + \tilde{\boldsymbol{x}}_B^\top \tilde{\boldsymbol{y}}_B)$.*

Proof. For any $(\gamma_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ defining a TR estimator in (7), let $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2}$ be such that $\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2\top} = \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1}$. The Cholesky decomposition achieves this but is not the only choice. Then let $\tilde{\boldsymbol{x}}_{\text{B}} = [\sqrt{\lambda}\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2} \, \mathbf{0}_p \, \cdots \, \mathbf{0}_p]^\top$, where $\mathbf{0}_p$ is repeated $n_{\text{B}} - p$ times and $\tilde{\boldsymbol{y}}_{\text{B}} = [\sqrt{\lambda}\gamma_{\boldsymbol{\beta}}^\top \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2} \, 0 \, \cdots \, 0]^\top$, 0 repeated $n_{\text{B}} - p$ times. This gives the desired result. □

Note, although $\boldsymbol{y}_{\text{B}}$ is observed, its value is replaced by $\tilde{\boldsymbol{y}}_{\text{B}}$. Also, choices of $\tilde{\boldsymbol{x}}_{\text{B}}$ and $\tilde{\boldsymbol{y}}_{\text{B}}$ that satisfy the theorem may not be unique. For example, applied to FRC, the algorithm presented in the proof does not yield $\tilde{\boldsymbol{x}}_{\text{B}} = \boldsymbol{x}_{\text{B}}^{\text{FRC}}$ and $\tilde{\boldsymbol{y}}_{\text{B}} = \boldsymbol{y}_{\text{B}}$.

The following result compares $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$ in terms of their expected distance from $\boldsymbol{x}_{\mathrm{B}}$.

Theorem A.2. *The squared Frobenius norm of a matrix $\boldsymbol{S}$ be given by $\|\boldsymbol{S}\|_F^2 = Tr\left[\boldsymbol{S}^\top \boldsymbol{S}\right]$. Then,* $\mathrm{E}_{\boldsymbol{x}_B, \boldsymbol{w}_B}\left[\|\boldsymbol{x}_B^{\mathrm{FRC}} - \boldsymbol{x}_B\|_F^2 - \|\boldsymbol{x}_B^{\mathrm{SRC}} - \boldsymbol{x}_B\|_F^2\right] \geq 0.$

Proof. (Theorem A.2) Using $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}} = (1/\nu)\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V}$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}} = (1/\nu)\boldsymbol{w}_{\mathrm{B}}$,

$$
\begin{aligned}
&\mathrm{E}\,\|\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}} - \boldsymbol{x}_{\mathrm{B}}\|_{\mathrm{F}}^2 \\
&\quad = \mathrm{E}_{\boldsymbol{x}_{\mathrm{B}}}\mathrm{E}_{\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}}}\,\mathrm{Tr}\left[\frac{1}{\nu^2}\boldsymbol{V}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V} - \frac{1}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V} - \frac{1}{\nu}\boldsymbol{V}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right] \\
&\quad = \mathrm{E}_{\boldsymbol{x}_{\mathrm{B}}}\mathrm{Tr}\left[\frac{1}{\nu^2}\boldsymbol{V}(\nu^2\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \tau^2 n_{\mathrm{B}}\boldsymbol{I}_p)\boldsymbol{V} - \frac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} - \frac{\nu}{\nu}\boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right] \\
&\quad = \mathrm{Tr}\left[\frac{1}{\nu^2}\boldsymbol{V}(\nu^2\boldsymbol{\Sigma}_X + \tau^2 n_{\mathrm{B}}\boldsymbol{I}_p)\boldsymbol{V} - \frac{\nu}{\nu}\boldsymbol{\Sigma}_X\boldsymbol{V} - \frac{\nu}{\nu}\boldsymbol{V}\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_X\right] \\
&\quad = \mathrm{Tr}\left[n_{\mathrm{B}}\frac{\tau^2}{\nu^2}\boldsymbol{V}^2 + n_{\mathrm{B}}(\boldsymbol{I}_p - \boldsymbol{V})^2\boldsymbol{\Sigma}_X\right] \qquad (\boldsymbol{V}\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_X\boldsymbol{V}) \\
&\quad = n_{\mathrm{B}}\frac{\tau^2}{\nu^2}\mathrm{Tr}\boldsymbol{V} \qquad (\boldsymbol{\Sigma}_X = \frac{\tau^2}{\nu^2}(\boldsymbol{I}_p - \boldsymbol{V})^{-1}\boldsymbol{V}), \qquad (71)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\,\|\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}} - \boldsymbol{x}_{\mathrm{B}}\|_{\mathrm{F}}^2 &= \mathrm{E}_{\boldsymbol{x}_{\mathrm{B}}}\mathrm{E}_{\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}}}\,\mathrm{Tr}\left[\frac{1}{\nu^2}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}} - \frac{1}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}} - \frac{1}{\nu}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right] \\
&= \mathrm{E}_{\boldsymbol{x}_{\mathrm{B}}}\,\mathrm{Tr}\left[\frac{1}{\nu^2}(\nu^2\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \tau^2 n_{\mathrm{B}}\boldsymbol{I}_p) - \frac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} - \frac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right] \\
&= n_{\mathrm{B}}\frac{\tau^2}{\nu^2}\mathrm{Tr}\boldsymbol{I}_p. \qquad (72)
\end{aligned}
$$

A comparison of expressions (71) and (72), together with the inequality $\mathrm{Tr}\,(\boldsymbol{I}_p - \boldsymbol{V}) \geq 0$ implied by (11) completes the proof. □

Thus, $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}$ is closer on average to $\boldsymbol{x}_{\mathrm{B}}$ than $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$ is to $\boldsymbol{x}_{\mathrm{B}}$, when the assumed model for $\boldsymbol{X}$ is true. This is to be expected given that the assumptions of the SRC algorithm are exactly satisfied; the FRC algorithm does not make explicit use of the model for $\boldsymbol{X}$. However, the regression of the completed data is more relevant in our situation. TR estimators may be evaluated in terms of prediction of the outcome Y, and, from this perspective, this unequivocal preference of SRC over FRC no longer holds.

To show this, we first establish that RIDG and FRC are closely related: $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ is an ap-

proximate ridge-type estimator on the *complete* data, as demonstrated by the following relationship in their functional forms. By definition, $\boldsymbol{x}_{\text{B}}^{\text{FRC}} = (1/\nu)\boldsymbol{w}_{\text{B}} = \boldsymbol{x}_{\text{B}} + (\tau/\nu)\boldsymbol{\xi}_{\text{B}}$, where $\boldsymbol{\xi}_{\text{B}}$ is the unobserved $n_{\text{B}} \times p$ error matrix. From this, and the definition of $\boldsymbol{x}_{\text{B}}^{\text{FRC}}$ in (12), we have:

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1} = \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \frac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}}, \qquad \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}} = \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}. \tag{73}$$

Plugging these values of $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}}$ into (7) gives that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{FRC}} &= \left(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + \frac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{\xi}_{\text{B}}\right)^{-1}\left(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}} + \frac{\tau}{\nu}\boldsymbol{\xi}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}\right) \\ &\approx \left(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{x}_{\text{B}} + n_{\text{B}}\frac{\tau^2}{\nu^2}\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{y}_{\text{A}} + \boldsymbol{x}_{\text{B}}^{\top}\boldsymbol{y}_{\text{B}}\right),\end{aligned} \tag{74}$$

where the last approximation replaces each expression involving $\boldsymbol{\xi}_{\text{B}}$ in the previous line with its marginal expectation. Thus (74) characterizes $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ as an approximate ridge-type estimator based on the complete data, with the shrinkage parameter $n_{\text{B}}\tau^2/\nu^2$. Ridge regression can improve prediction error over OLS for certain choices of the tuning parameter (Gelfand, 1986; Frank and Friedman, 1993). Consequently, $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ may offer improved prediction, even over OLS on the complete data; whether this holds in practice depends crucially on the size of $n_{\text{B}}\tau^2/\nu^2$. As τ^2/ν^2 increases, $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ approaches zero, as seen by the expansion above. Interpreted from the Bayesian perspective, this is because the prior mean, $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}}$, approaches $\mathbf{0}_p$ with τ^2/ν^2, and the prior precision, $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}$, grows without bound with τ^2/ν^2.

Following a similar expansion for SRC as above, note that $\boldsymbol{x}_{\text{B}}^{\text{SRC}} = (1/\nu)\boldsymbol{w}_{\text{B}}\boldsymbol{V} = \boldsymbol{x}_{\text{B}}\boldsymbol{V} +$

$(\tau/\nu)\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V}$ (if $\boldsymbol{\mu}_{\boldsymbol{X}}$ is assumed to be zero). When we expand $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ as in (74), we obtain

$$\hat{\boldsymbol{\beta}}_{\mathrm{SRC}} = \left(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} + \frac{\tau}{\nu}\boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V} + \frac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} + \frac{\tau^2}{\nu^2}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V}\right)^{-1} \times \left(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}} + \frac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}}\right). \tag{75}$$

From (11), as $\tau^2/\nu^2 \to \infty$, the elements of $\boldsymbol{V}$ go to zero at a rate proportional to τ^2/ν^2. Thus, for large τ^2/ν^2, $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ is "unstable", because it approximates $(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}$, the OLS estimate of $\boldsymbol{\beta}$, which does not exist when $p > n_{\mathrm{A}}$. In contrast with the Bayesian interpretation of FRC, in which the prior precision matrix *increases* with τ^2/ν^2, for SRC, the prior precision *decreases* to zero (a flat prior), and using a flat prior when $p > n_{\mathrm{A}}$ yields an improper posterior. From this comparison, we may infer that the MSPE of $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ is unbounded with τ^2/ν^2 (because Var $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ is unbounded), while $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ is not. Next, we more formally compare SRC and FRC in terms of their MSPE.

Theorem A.3. *Let* $\boldsymbol{V}$ *and* $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}$ *be as in* (11) *and* (12)*, respectively. Also, define*

$$\begin{aligned}
\kappa &= (\tau^2/\nu^2)\boldsymbol{\beta}^{\top}\boldsymbol{V}\boldsymbol{\beta}, \\
\boldsymbol{\Delta}_{\sigma}^{\mathrm{SRC}} &= \sigma^2(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}, && (76)\\
\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{SRC}} &= \kappa(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{V}(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}, && (77)\\
\boldsymbol{\Delta}_{\sigma}^{\mathrm{FRC}} &= \sigma^2(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1})^{-1}, && (78)\\
\boldsymbol{\Delta}_{\boldsymbol{\beta}}^{\mathrm{FRC}} &= \kappa(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1})^{-1} \\
&\quad + (\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\beta}\boldsymbol{\beta}^{\top}(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}(\boldsymbol{x}_A^{\top}\boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1})^{-1}. && (79)
\end{aligned}$$

Then using (76)–(79), the MSPE of the SRC *and* FRC *methods can each be expressed as*

$$\sigma^2 + Tr\big[\boldsymbol{\Delta}_{\sigma}\boldsymbol{\Sigma}_X\big] + Tr\big[\boldsymbol{\Delta}_{\boldsymbol{\beta}}\boldsymbol{\Sigma}_X\big].$$

Proof. (THEOREM A.3) The assumption $[Y|\boldsymbol{X},\boldsymbol{W}] = [Y|\boldsymbol{X}]$ gives that $\mathrm{E}[Y|\boldsymbol{W}] = \beta_0 + \mathrm{E}[\boldsymbol{X}|\boldsymbol{W}]\boldsymbol{\beta}$ and $\mathrm{Var}[Y|\boldsymbol{W}] = \sigma^2 + \boldsymbol{\beta}^{\top}\mathrm{Var}[\boldsymbol{X}|\boldsymbol{W}]\boldsymbol{\beta}$. Because $\boldsymbol{X}$ and $\boldsymbol{W}$ are assumed jointly

normal, it is seen that $\mathrm{E}[\boldsymbol{X}|\boldsymbol{W}] = (\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{V}(\boldsymbol{W} - \psi\mathbf{1}_p)/\nu$ and $\mathrm{Var}[\boldsymbol{X}|\boldsymbol{W}] = (\tau^2/\nu^2)\boldsymbol{V}$. Thus, $\mathrm{E}[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{w}_{\mathrm{B}}] = \beta_0\mathbf{1}_{n_{\mathrm{B}}} + [\mathbf{1}_{n_{\mathrm{B}}}, \boldsymbol{w}_{\mathrm{B}}]\boldsymbol{M}\boldsymbol{\beta}$ and $\mathrm{Var}[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{w}_{\mathrm{B}}] = (\sigma^2 + (\tau^2/\nu^2)\boldsymbol{\beta}^\top\boldsymbol{V}\boldsymbol{\beta})\boldsymbol{I}_{n_{\mathrm{B}}}$, where

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{X}}^\top(\boldsymbol{I}_p - \boldsymbol{V}) - (\psi/\nu)\mathbf{1}_p^\top\boldsymbol{V} \\ \frac{1}{\nu}\boldsymbol{V} \end{pmatrix}.$$

These in turn yield the mean and variance of $\gamma_{\beta_{\mathrm{SRC}}}$ and $\gamma_{\beta_{\mathrm{FRC}}}$. Now, assume $\beta_0 = \psi = 0$. With these results and the bias and variance expressions given in (8) and (9), we can write

$$\begin{aligned}
&\mathrm{Bias}\,\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}\,\mathrm{Bias}\,\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}^\top + \mathrm{Var}\,\hat{\boldsymbol{\beta}}_{\mathrm{FRC}} \\
&\quad= (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1} \\
&\qquad\times \left\{\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}(\mathrm{E}\,\gamma_{\beta_{\mathrm{FRC}}} - \boldsymbol{\beta})(\mathrm{E}\,\gamma_{\beta_{\mathrm{FRC}}} - \boldsymbol{\beta})^\top\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1} + \sigma^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\mathrm{Var}\,\gamma_{\beta_{\mathrm{FRC}}}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\right\} \\
&\qquad\times (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1} \\
&\quad= (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1}\left\{\sigma^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + (\sigma^2 + \kappa)\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\right\} \\
&\qquad\times (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1} \\
&\quad= \sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1} \\
&\qquad+ (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1}\left\{\kappa\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\right\}(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1})^{-1}.
\end{aligned}$$

Next, using the identity $\boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1} = \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V}$,

$$\begin{aligned}
&\mathrm{Bias}\,\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}\,\mathrm{Bias}\,\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}^\top + \mathrm{Var}\,\hat{\boldsymbol{\beta}}_{\mathrm{SRC}} \\
&\quad= (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1})^{-1} \\
&\qquad\times \left\{\boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1}(\mathrm{E}\,\gamma_{\beta_{\mathrm{SRC}}} - \boldsymbol{\beta})(\mathrm{E}\,\gamma_{\beta_{\mathrm{SRC}}} - \boldsymbol{\beta})^\top\boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1} + \sigma^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1}\mathrm{Var}\,\gamma_{\beta_{\mathrm{SRC}}}\boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1}\right\} \\
&\qquad\times (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{\Omega}_{\beta_{\mathrm{SRC}}}^{-1})^{-1} \\
&\quad= (\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}\left\{\sigma^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + (\sigma^2 + \kappa)\boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V}\right\}(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1} \\
&\quad= \sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1} + \kappa(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V}(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\beta_{\mathrm{FRC}}}^{-1}\boldsymbol{V})^{-1}.
\end{aligned}$$

□

By taking the difference of the two MSPE expressions for FRC and SRC from THEOREM A.3, the following Corollary characterizes how $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}})$ changes as a function of σ^2 and $\boldsymbol{\beta}$.

Corollary A.4. *$MSPE(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - MSPE(\hat{\boldsymbol{\beta}}_{\text{FRC}}) = \sigma^2 c_1 + \boldsymbol{\beta}^\top (\boldsymbol{C}_2 - \boldsymbol{C}_3)\boldsymbol{\beta}$, where*

$$c_1 = Tr\left[\left\{(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1} - (\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\right\}\boldsymbol{\Sigma}_{\boldsymbol{X}}\right], \tag{80}$$

$$\begin{aligned}\boldsymbol{C}_2 = Tr\Big[&(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V}(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}} \\ &- (\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}}\Big]\left(\frac{\tau^2}{\nu^2}\boldsymbol{V}\right).\end{aligned} \tag{81}$$

$$\boldsymbol{C}_3 = (\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{I}_p - \boldsymbol{V}). \tag{82}$$

When $p = 1$, $c_1, \boldsymbol{C}_2, \boldsymbol{C}_3$ are scalar-valued, and one can show the following:

(i) $c_1 > 0$.

(ii) The sign of $\boldsymbol{C}_2 - \boldsymbol{C}_3$ is equal to that of

$$\frac{\boldsymbol{V}^2(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2}{(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}^2\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2} - \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\frac{(1-\boldsymbol{V})^2}{(\tau^2/\nu^2)\boldsymbol{V}} - 1. \tag{83}$$

(iii) As $\tau^2/\nu^2 \to \infty$,

(a) $c_1^{-1} - \boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} = o(1)$ for $\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} \neq 0$.

(b) $\boldsymbol{C}_2 = o(1)$ for $\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} \neq 0$.

(c) $\boldsymbol{C}_3 - \boldsymbol{\Sigma}_{\boldsymbol{X}} = o(1)$.

Thus fixing all other parameters, (i) indicates that $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}})$ increases with σ^2, making FRC the preferred method for large values of σ^2. From (ii), if $n_{\text{A}} \gg n_{\text{B}}$, (83) is approximated by $\boldsymbol{V}^2 - \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\frac{(1-\boldsymbol{V})^2}{(\tau^2/\nu^2)\boldsymbol{V}} - 1$, because $\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}$ increase linearly

in n_A and n_B, respectively, and therefore $(\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}_{\text{FRC}}}^{-1})^2 \approx (\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{V}^2 \boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}_{\text{FRC}}}^{-1})^2$. Because $0 \leq \boldsymbol{V} \leq 1$, the entire expression is negative in this case, and SRC is preferred to FRC for large values of $\boldsymbol{\beta}^2$. When $n_B > n_A$, there is no clear dominance of SRC over FRC, as the sign of (83) then depends on $\boldsymbol{V}$, which is in turn a function of τ^2/ν^2 and $\boldsymbol{\Sigma}_X^{-1}$.

The effect of an increasing τ^2/ν^2 on $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}})$ gives which method is preferred in the large measurement error case. Replacing c_1, $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ with the limiting values implied by (iii), $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}})$ is approximately $\sigma^2(\boldsymbol{x}_A^\top \boldsymbol{x}_A)^{-1}\boldsymbol{\Sigma}_X - \boldsymbol{\beta}^2\boldsymbol{\Sigma}_X$. The first expression ($\sigma^2(\boldsymbol{x}_A^\top \boldsymbol{x}_A)^{-1}\boldsymbol{\Sigma}_X$) is attributable to $\text{Var}\,\hat{\boldsymbol{\beta}}_{\text{SRC}}$ and the second term ($\boldsymbol{\beta}^2\boldsymbol{\Sigma}_X$) to $\text{Bias}\,\hat{\boldsymbol{\beta}}_{\text{FRC}}$. Thus, when τ^2/ν^2 is large, $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}}) > 0 \Leftrightarrow \sigma^2(\boldsymbol{x}_A^\top \boldsymbol{x}_A)^{-1} > \boldsymbol{\beta}^2$. Moreover, $\boldsymbol{x}_A^\top \boldsymbol{x}_A / n_A$ consistently estimates $\boldsymbol{\Sigma}_X$; some simplification then suggests the approximately equivalent statement $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}}) > 0 \Leftrightarrow (n_A + 1)^{-1} > R^2$, where $R^2 = \boldsymbol{\beta}^2\boldsymbol{\Sigma}_X/(\boldsymbol{\beta}^2\boldsymbol{\Sigma}_X + \sigma^2)$. The dominance of one method over the other thus depends on n_A and the signal in the model.

For $p > 1$, we were not able to prove multivariate versions of the above results; however, extensive simulation studies that evaluate $c_1, \boldsymbol{C}_2, \boldsymbol{C}_3$ (given in Table A1) indicate that the preceding conclusions are still likely to hold in the general p case as long as $p < n_A$. That is, the results above depend crucially on the existence of $(\boldsymbol{x}_A^\top \boldsymbol{x}_A)^{-1}$. When $p > n_A$, as is the case in our motivating example, $p - n_A$ eigenvalues of $\boldsymbol{x}_A^\top \boldsymbol{x}_A + \boldsymbol{V}\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}_{\text{FRC}}}^{-1}\boldsymbol{V}$ (appearing in the expressions for c_1 and $\boldsymbol{C}_2$) may be nearly zero for non-negligible measurement error. Thus the matrix trace, being the sum of reciprocals of the eigenvalues, will be large. This does not affect $\boldsymbol{C}_3$, and so both c_1 and $\text{Tr}(\boldsymbol{C}_2 - \boldsymbol{C}_3)$ tend to be large. Therefore, FRC is favored over SRC as either σ^2 or $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ increase, more so as τ^2/ν^2 increases.

A.2 Analysis of Hybrid Estimators

Lemmas A.5 and A.6 are used in the proof of THEOREM 2.1. We use 'PSD' to describe a positive semi-definite matrix and 'PD' to describe a positive definite matrix.

p	n_B	τ^2/ν^2	c_1	Tr $\boldsymbol{C}_2$	Tr $\boldsymbol{C}_3$	Tr $\boldsymbol{C}_2 - \boldsymbol{C}_3$
1	10	0.01	0.0001	-0.0000	0.0000	-0.0000
1	10	0.25	0.0012	-0.0002	0.0019	-0.0021
1	10	1	0.0038	-0.0011	0.0220	-0.0231
1	10	25	0.0161	-0.0026	0.6187	-0.6213
1	10	100	0.0188	-0.0010	0.8702	-0.8712
1	50	0.01	0.0001	0.0000	0.0000	-0.0000
1	50	0.25	0.0022	0.0000	0.0127	-0.0127
1	50	1	0.0065	-0.0000	0.1105	-0.1105
1	50	25	0.0185	0.0000	0.8557	-0.8557
1	50	100	0.0196	0.0000	0.9612	-0.9612
1	100	0.01	0.0001	0.0000	0.0000	-0.0000
1	100	0.25	0.0019	0.0001	0.0208	-0.0207
1	100	1	0.0059	0.0009	0.1616	-0.1607
1	100	25	0.0181	0.0009	0.8886	-0.8877
1	100	100	0.0195	0.0003	0.9709	-0.9706
1	400	0.01	0.0000	0.0000	0.0001	-0.0001
1	400	0.25	0.0009	0.0001	0.0332	-0.0331
1	400	1	0.0028	0.0010	0.2220	-0.2209
1	400	25	0.0151	0.0034	0.9158	-0.9124
1	400	100	0.0185	0.0014	0.9779	-0.9765
9	10	0.01	0.0006	-0.0000	0.0001	-0.0001
9	10	0.25	0.0130	-0.0088	0.0274	-0.0362
9	10	1	0.0395	-0.0658	0.2747	-0.3406
9	10	25	0.1507	-0.2165	4.5387	-4.7552
9	10	100	0.1821	-0.1458	6.5387	-6.6845
9	50	0.01	0.0010	0.0000	0.0003	-0.0002
9	50	0.25	0.0213	0.0037	0.1219	-0.1183
9	50	1	0.0647	0.0236	1.0202	-0.9966
9	50	25	0.1962	0.0306	7.6113	-7.5807
9	50	100	0.2130	0.0104	8.6141	-8.6037
9	100	0.01	0.0008	0.0000	0.0004	-0.0004
9	100	0.25	0.0184	0.0124	0.1908	-0.1784
9	100	1	0.0564	0.0887	1.4544	-1.3656
9	100	25	0.1905	0.1382	7.9884	-7.8502
9	100	100	0.2098	0.0469	8.7287	-8.6817
9	400	0.01	0.0003	0.0000	0.0007	-0.0007
9	400	0.25	0.0079	0.0110	0.2994	-0.2883
9	400	1	0.0253	0.0850	1.9981	-1.9131
9	400	25	0.1491	0.3707	8.2416	-7.8709
9	400	100	0.1931	0.1794	8.8009	-8.6214
99	100	0.01	0.0258	0.0197	0.0087	0.0110
99	100	0.25	0.5821	8.9737	3.5504	5.4233
99	100	1	1.9040	74.5419	22.2836	52.2583
99	100	25	26.8529	2405.7851	86.6150	2319.1701
99	100	100	100.7154	9693.8778	93.9312	9599.9466
99	400	0.01	0.0049	0.0039	0.0080	-0.0041
99	400	0.25	0.1098	1.7935	3.3435	-1.5500
99	400	1	0.3563	14.4882	21.9265	-7.4383
99	400	25	4.3277	358.4208	90.4184	268.0024
99	400	100	14.9043	1374.9888	96.7380	1278.2507

Table A1: Numerical calculations of c_1, Tr $\boldsymbol{C}_2$, Tr $\boldsymbol{C}_3$, and Tr $\boldsymbol{C}_2 - \boldsymbol{C}_3$ as defined in Equations (80)–(82) in COROLLARY A.4 using the *true* value of $\boldsymbol{\theta} = \{\psi, \nu, \tau, \boldsymbol{\Sigma}_X^{-1}\}$. Each row is averaged over 200 draws of $\boldsymbol{x}_A$, $\boldsymbol{w}_A$ and $\boldsymbol{w}_B$. In all cases, $n_A = 50$, $\psi = 0$, $\nu = 1$, and $\boldsymbol{\Sigma}_X = \boldsymbol{I}_p$.

Lemma A.5. *Given a PSD matrix $\boldsymbol{M}$ with at least one strictly positive eigenvalue and PD matrix $\boldsymbol{N}$, both of the same dimensions, $Tr(\boldsymbol{MN}) > 0$.*

Proof. Suppose the dimension of the matrices is p. Consider the eigendecomposition of $\boldsymbol{M}$, $\boldsymbol{M} = \boldsymbol{Q\Lambda Q}^\top$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_i, \ldots, \lambda_p)$ is the diagonal matrix of eigenvalues of $\boldsymbol{M}$ (in decreasing order) and $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p)$ is the column matrix of corresponding eigenvectors of $\boldsymbol{M}$ (all non-zero). Then,

$$\begin{aligned}
\text{Tr}(\boldsymbol{MN}) &= \text{Tr}(\boldsymbol{Q\Lambda Q}^\top \boldsymbol{N}) \\
&= \text{Tr}(\boldsymbol{\Lambda Q}^\top \boldsymbol{NQ}) \\
&= \sum_{i=1}^{p} \lambda_i (\boldsymbol{q}_i^\top \boldsymbol{N} \boldsymbol{q}_i) \qquad \text{(since } \boldsymbol{\Lambda} \text{ is diagonal)} \\
&\geq \lambda_1 (\boldsymbol{q}_1^\top \boldsymbol{N} \boldsymbol{q}_1) > 0,
\end{aligned}$$

since the largest eigenvalue λ_1 is positive, $\boldsymbol{q}_1$ is non-zero, and $\boldsymbol{N}$ is PD. □

Lemma A.6. *Given estimators $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_m$, define $\boldsymbol{P}$ by* (13) *in the text, ie $P_{ij} = MCPE(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$. If $\text{Var}\left[(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_m)\boldsymbol{v}\right]$ has at least positive eigenvalue for every $\boldsymbol{v} \in \mathbb{R}^m \backslash \boldsymbol{0}_m$, then $\boldsymbol{P}$ is PD.*

Proof. We show $\boldsymbol{v}^\top \boldsymbol{P} \boldsymbol{v} > 0$ for $\boldsymbol{v} \in \mathbb{R}^m \backslash \boldsymbol{0}_m$. Define the following random variable: $\boldsymbol{U}_\ell = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\ell$. Let $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_m)$. Then, $\boldsymbol{P} = \sigma^2 \boldsymbol{1}_m \boldsymbol{1}_m^\top + \text{E}[\boldsymbol{U}^\top \boldsymbol{X}_{\text{new}} \boldsymbol{X}_{\text{new}}^\top \boldsymbol{U}]$. Now, choose $\boldsymbol{v} \in \mathbb{R}^m \backslash \boldsymbol{0}_m$. Then,

$$\begin{aligned}
\boldsymbol{v}^\top \boldsymbol{P} \boldsymbol{v} &= \sigma^2 \boldsymbol{v}^\top \boldsymbol{1}_m \boldsymbol{1}_m^\top \boldsymbol{v} + \boldsymbol{v}^\top \text{E}[\boldsymbol{U}^\top \boldsymbol{X}_{\text{new}} \boldsymbol{X}_{\text{new}}^\top \boldsymbol{U}] \boldsymbol{v} \\
&= \sigma^2 (\boldsymbol{v}^\top \boldsymbol{1}_m)^2 + \text{Var}\left[\boldsymbol{X}_{\text{new}}^\top \boldsymbol{U} \boldsymbol{v}\right] + (\text{E}[\boldsymbol{X}_{\text{new}}^\top \boldsymbol{U} \boldsymbol{v}])^2.
\end{aligned}$$

The first and third expressions are nonnegative. Considering the second expression,

$$\text{Var}\left[\boldsymbol{X}_{\text{new}}^\top \boldsymbol{U} \boldsymbol{v}\right] = \text{Tr}\left(\boldsymbol{\Sigma}_X \text{Var}[\boldsymbol{U} \boldsymbol{v}]\right) + \text{E}[\boldsymbol{X}_{\text{new}}]^\top \text{Var}[\boldsymbol{U} \boldsymbol{v}] \text{E}[\boldsymbol{X}_{\text{new}}] + \text{E}[\boldsymbol{U} \boldsymbol{v}]^\top \boldsymbol{\Sigma}_X \text{E}[\boldsymbol{U} \boldsymbol{v}].$$

The second and third expressions are nonnegative. We show the first is strictly positive:

$$\begin{aligned}\operatorname{Tr}\left(\boldsymbol{\Sigma}_X \operatorname{Var}[\boldsymbol{U}\boldsymbol{v}]\right) &= \operatorname{Tr}\left(\boldsymbol{\Sigma}_X \operatorname{Var}\left[\boldsymbol{\beta}\mathbf{1}_m^\top \boldsymbol{v} - \left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m\right)\boldsymbol{v}\right]\right) \\ &= \operatorname{Tr}\left(\boldsymbol{\Sigma}_X \operatorname{Var}\left[\left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m\right)\boldsymbol{v}\right]\right).\end{aligned}$$

$\boldsymbol{\Sigma}_X$ is PD and, by assumption, $\operatorname{Var}\left[\left(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_m\right)\boldsymbol{v}\right]$ has at least one positive eigenvalue. Applying Lemma A.5, this is strictly positive. □

Proof. (Theorem 2.1)

(i) Being an affine combination, there always exists a feasible solution; existence and uniqueness of $\boldsymbol{\omega}^{\text{opt}}$ follow from $\boldsymbol{P}$ being PD, which in turn comes from Lemma A.6.

(ii) Without loss of generality, suppose $\text{MSPE}(\hat{\boldsymbol{\beta}}_m) = \min_\ell \text{MSPE}(\hat{\boldsymbol{\beta}}_\ell)$. It is always true that $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leq \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$. To see this, define $\boldsymbol{\omega}^{(1)} = \{0, 0, \ldots, 0, 1\}^\top$, and observe that $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{(1)})) = \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$. By definition, $\boldsymbol{\omega}^{\text{opt}}$ will do no worse in terms of MSPE than $\boldsymbol{\omega}^{(1)}$, ie $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leq \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$.

We now demonstrate that a sufficient condition under which this inequality is strict is $\text{MCPE}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\beta}}_i) \neq \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$ for some $i \neq j$. Let $\boldsymbol{\omega}^{\text{opt}} = \{\omega_1^{\text{opt}}, \omega_2^{\text{opt}}, \ldots, \omega_m^{\text{opt}}\}^\top$ and define the $m \times m$ matrix $\boldsymbol{P}$ by (13) in the text, ie $P_{ij} = \text{MCPE}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$. We show that if $\boldsymbol{\omega}^{\text{opt}} = \boldsymbol{\omega}^{(1)}$ (ie, if the best prediction error comes from using only $\hat{\boldsymbol{\beta}}_m$, the estimator with smallest MSPE), then $P_{1m} = P_{2m} = \cdots = P_{mm}$. By contraposition, if $P_{im} \neq P_{mm}$ for some $i \neq m$, then $\boldsymbol{\omega}^{\text{opt}} \neq \boldsymbol{\omega}^{(1)}$, which implies, by the uniqueness of $\boldsymbol{\omega}^{\text{opt}}$, that $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) < \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$ (the required result). For a general $\boldsymbol{\omega}$, $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega})) =$

$\boldsymbol{\omega}^\top \boldsymbol{P}\boldsymbol{\omega}$ will have zero slope at its minimum:

$$\begin{aligned}
&\boldsymbol{\omega}^\top \boldsymbol{P}\boldsymbol{\omega} = \sum_{i=1}^{m-1} P_{ii}\omega_i^2 + 2\sum_{i=2}^{m-1}\omega_i \sum_{j=1}^{i-1} P_{ij}\omega_j + P_{mm}\left(1-\sum_{i=1}^{m-1}\omega_i\right)^2 + 2\left(1-\sum_{i=1}^{m-1}\omega_i\right)\sum_{i=1}^{m-1} P_{im}\omega_i \\
\Rightarrow &\frac{\partial \boldsymbol{\omega}^\top \boldsymbol{P}\boldsymbol{\omega}}{\partial \omega_\ell} = 2P_{\ell\ell}\omega_\ell + 2\sum_{i\neq\ell}^{m-1} P_{\ell i}\omega_i - 2P_{mm}\left(1-\sum_{i=1}^{m-1}\omega_i\right) + 2P_{\ell m}\left(1-\sum_{i=1}^{m-1}\omega_i-\omega_\ell\right) \\
\Rightarrow &\left(\frac{\partial \boldsymbol{\omega}^\top \boldsymbol{P}\boldsymbol{\omega}}{\partial \omega_\ell}\middle| \boldsymbol{\omega}^{\text{opt}} = \boldsymbol{\omega}^{(1)}\right) = -2P_{mm} + 2P_{\ell m} = 0,
\end{aligned}$$

which gives that $P_{1m} = P_{2m} = \ \cdots \ = P_{mm}$. □

Lemma A.7. *Suppose we have two targeted ridge estimators, $\hat{\boldsymbol{\beta}}_{k_1} = \hat{\boldsymbol{\beta}}(\gamma_{\boldsymbol{\beta},k_1}, \lambda_{k_1}, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_1})$ and $\hat{\boldsymbol{\beta}}_{k_2} = \hat{\boldsymbol{\beta}}(\gamma_{\boldsymbol{\beta},k_2}, \lambda_{k_2}, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_2})$, as defined by (7). Let $\psi_\ell = Tr\ \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},\ell})/n_A$. If $\gamma_{\boldsymbol{\beta},k_1}$ and $\gamma_{\boldsymbol{\beta},k_2}$ are not functions of $\boldsymbol{y}_A$, then*

$$\begin{aligned}
&\mathrm{E}\left[(1/n_A)(\boldsymbol{y}_A - \boldsymbol{x}_A\hat{\boldsymbol{\beta}}_{k_1})^\top(\boldsymbol{y}_A - \boldsymbol{x}_A\hat{\boldsymbol{\beta}}_{k_2})\right] \\
&\quad = \sigma^2 + \mathrm{E}\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k_1})^\top \boldsymbol{x}_A^\top \boldsymbol{x}_A(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k_2})/n_A\right] - \sigma^2(\psi_{k_1} + \psi_{k_2}).
\end{aligned} \tag{84}$$

Proof. (Lemma A.7)

$$\begin{aligned}
&(1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_2) \\
&\quad = (1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} + \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} + \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_2) \\
&\quad = (1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) && (85) \\
&\qquad + (1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_1) && (86) \\
&\qquad + (1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_2) && (87) \\
&\qquad + (1/n_{\mathrm{A}})(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta} - \boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_2). && (88)
\end{aligned}$$

Taking expectations, (85) evaluates to σ^2 and (88) to $\mathrm{E}\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)^\top \boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)/n_{\mathrm{A}}\right]$. For

(86),

$$
\begin{aligned}
&(1/n_{\mathrm{A}})\mathrm{E}\left[(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}-\boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_1)\right] \\
&\quad = (1/n_{\mathrm{A}})\mathrm{E}\left[(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}-\boldsymbol{x}_{\mathrm{A}}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}}+\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})^{-1}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}}+\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1}\gamma_{\boldsymbol{\beta},1}))\right] \quad (89) \\
&\quad = -(1/n_{\mathrm{A}})\mathrm{E}(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})\boldsymbol{y}_{\mathrm{A}}) \\
&\quad = -(1/n_{\mathrm{A}})\mathrm{E}(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}(\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \quad (90) \\
&\quad = -(1/n_{\mathrm{A}})\mathrm{E}(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^{\top}\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&\quad = -\sigma^2\mathrm{Tr}\ \boldsymbol{H}(\lambda_1\boldsymbol{\Omega}_{\boldsymbol{\beta},1}^{-1})/n_{\mathrm{A}}.
\end{aligned}
$$

The equality between (89) and (90) assumes that $\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}$ has mean $\mathbf{0}_p$ and is independent of $\gamma_{\boldsymbol{\beta},1}$. The analogous result comes from the expectation of (87). □

The following lemma, a generalization from Golub et al. (1979), provides a condition for the GCV expression being close to the true MSPE expression that it targets.

Lemma A.8. *Let* $R_\ell = \mathrm{E}[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\ell)^{\top}\boldsymbol{x}_A^{\top}\boldsymbol{x}_A(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_\ell)]$*, ie the mean squared error in estimating* $\boldsymbol{x}_A\boldsymbol{\beta}$*. This is a consistent estimate of MSPE*$(\hat{\boldsymbol{\beta}}_\ell)$ *as* n_A *increases, up to the constant* σ^2*. A surrogate for MSPE*$(\hat{\boldsymbol{\beta}}_\ell)$ *is* $\hat{P}_{\ell,\ell}$*, defined in expression* (14)*. The difference in* $\mathrm{E}R_\ell$ *and* $\mathrm{E}\hat{P}_{\ell,\ell}-\sigma^2$ *relative to* $\mathrm{E}R_\ell$ *is*

$$
\frac{\mathrm{E}R_\ell-(\mathrm{E}\hat{P}_{\ell,\ell}-\sigma^2)}{\mathrm{E}R_\ell} = \frac{-2\psi_\ell}{(1-\psi_\ell)^2}+\frac{\psi_\ell^2}{(1-\psi_\ell)^2}\frac{\mathrm{E}R_\ell+\sigma^2}{\mathrm{E}R_\ell}
$$

and so is small when $\psi_\ell = Tr\ \boldsymbol{H}(\lambda_\ell\boldsymbol{\Omega}_{\boldsymbol{\beta},\ell}^{-1})/n_A$ *is small.*

Proof. We have $\hat{P}_{\ell,\ell} = (1-\psi_\ell)^{-2}(1/n_{\mathrm{A}})(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_\ell)^{\top}(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\hat{\boldsymbol{\beta}}_\ell)$. Then,

$$
\begin{aligned}
\frac{\mathrm{E}R_\ell-\mathrm{E}\tilde{R}_\ell+\sigma^2}{\mathrm{E}R_\ell} &= \frac{\mathrm{E}R_\ell+\sigma^2-(1-\psi)^{-2}(\mathrm{E}R_\ell+\sigma^2-2\sigma^2\psi_\ell)}{\mathrm{E}R_\ell} \quad \text{(from Proof of Lemma A.7)} \\
&= \frac{-2\psi_\ell}{(1-\psi_\ell)^2}+\frac{\psi_\ell^2}{(1-\psi_\ell)^2}\frac{\mathrm{E}R_\ell+\sigma^2}{\mathrm{E}R_\ell}.
\end{aligned}
$$

□

A.3 Further Simulation Study Results

Tables A2 and A3 give numerical values of empirical MSPE from Figure 2 in the main text, and Figure A1 gives Empirical Mean Squared Error (MSE) from the same simulation study. Next, we summarize simulation results under various model misspecifications.

α $\boldsymbol{\alpha}$

When $[Y|\boldsymbol{X}, \boldsymbol{W}] \neq [Y|\boldsymbol{X}]$: We repeated each simulation with the alternative generating model $Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta}^* + \boldsymbol{W}^\top \boldsymbol{\alpha} + \sigma^* \varepsilon$. To keep fixed the model of interest, $Y = \beta_0 + \boldsymbol{X}^\top \boldsymbol{\beta} + \sigma\varepsilon$, for a given simulation setting, we set $\boldsymbol{\alpha} = s\boldsymbol{\beta}$, $\boldsymbol{\beta}^* = (1 - s\nu)\boldsymbol{\beta}$ and $\sigma^* = \sigma - s\tau\sqrt{\boldsymbol{\beta}^\top \boldsymbol{\beta}}$ for some $s \in [0, 1]$. Previously, $s = 0$; Figure A1 plots the MSPE when $s = 0.1$. Because σ^* decreases with τ, the MSPE of all methods, including RIDG, also tends to decrease with τ. HYB remains as an attractive choice.

Outcome Dependent Sampling: We repeated each simulation, automatically including the $n_A/2 = 25$ observations in subsample A with the largest values of Y and randomly allocating the remaining observations, as before. MSPE is plotted in Figure A3. As might be expected, since the methods do not account for outcome dependent sampling, the MSPE is typically much larger than in the case of simple random sampling. HYB, being a linear combination of all other methods, increases correspondingly but is still the overall best performing method.

Violations to Normality of $\boldsymbol{X}$ Assumption and ME Structure: We drew $\boldsymbol{X}$ from a multivariate t distribution with 5 degrees of freedom, scaled to maintain $\operatorname{Var} \boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. We simultaneously perturbed (1.2): instead of $\operatorname{Var}[w_{ij}|x_{ij}] = \tau^2$, the underlying true variance was $\operatorname{Var}[w_{ij}|x_{ij}] = \tau^2 |x_{ij}|^{1/4}$. These results are in Figure A4. MSPE actually decreases in this situation, and, again, HYB has MSPE that is smallest or almost so in most scenarios.

When $\boldsymbol{\theta}$ is known: The unbiasedness of $\hat{\boldsymbol{\beta}}_{\text{SRC}}$ 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 $\boldsymbol{\Sigma}_X$ 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 (see Remark 3 in the main text).

We considered other values of the true $\boldsymbol{\beta}$ that spread the signal evenly over all components or concentrated the signal in a few elements. Crucially, consistent with the results in Figure 2, HYB proved to be the most flexible of all methods: small MSPE in each case but not always the smallest.

A.4 Bootstrap Algorithm for Prediction Intervals

Since uncertainty in predictions is typically also of interest to the analyst, we describe a simple method for calculating prediction intervals via the bootstrap. For b in $1, \ldots, B$, repeat the following:

(i) Draw separate bootstrap samples from subsamples A and B, yielding $(\boldsymbol{y}_{\mathrm{A}}^b, \boldsymbol{x}_{\mathrm{A}}^b, \boldsymbol{w}_{\mathrm{A}}^b)$ and $(\boldsymbol{y}_{\mathrm{B}}^b, \boldsymbol{w}_{\mathrm{B}}^b)$. Use these to calculate $\hat{\boldsymbol{\beta}}_{\mathrm{RIDG}}^b$, $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}^b$, etc.

(ii) Let r_{A}^b be the size of the set of remaining observations in subsample A not sampled in step (i). Draw an additional observation from this set, say $(y^{b*}, \boldsymbol{x}^{b*})$, and calculate $e^{b*} = \sqrt{r_{\mathrm{A}}^b / (r_{\mathrm{A}}^b - 1)} \left(y^{b*} - \boldsymbol{x}^{b*} \hat{\boldsymbol{\beta}}^b \right)$, for each of $\hat{\boldsymbol{\beta}}^b = \hat{\boldsymbol{\beta}}_{\mathrm{RIDG}}^b$, $\hat{\boldsymbol{\beta}}^b = \hat{\boldsymbol{\beta}}_{\mathrm{FRC}}^b$, etc (Theorem A.9 gives a rationale for this approach).

(iii) For a new observation with covariate $\boldsymbol{X}_{\mathrm{new}}$, the predicted value of Y_{new} is $\hat{Y}_{\mathrm{new}}^b = \boldsymbol{X}_{\mathrm{new}}^\top \hat{\boldsymbol{\beta}}^b + e^{b*}$.

After B such iterations, the 95% prediction interval for Y_{new} is $(\hat{Y}_{\mathrm{new}}^{B,2.5}, \hat{Y}_{\mathrm{new}}^{B,97.5})$, where $\hat{Y}_{\mathrm{new}}^{B,2.5}$ and $\hat{Y}_{\mathrm{new}}^{B,97.5}$ are the 2.5 and 97.5 percentiles of the B bootstrap predictions.

Theorem A.9. *Suppose V_i is $N\{0, \sigma^2\}$, independently for $i = 1, \ldots, N$, and $U | V_1, \ldots, V_N \sim$ Unif$\{V_1, \ldots, V_N\}$. Then* $\mathrm{E}\left[\mathrm{Var}[U | V_1, \ldots, V_N]\right] = (N-1)\sigma^2/N$.

$\{\rho, R^2\}$	$\{p, n_B\}$	method	$\tau = 0.01$	0.25	0.5	0.75	1	1.25	1.5	1.75	2
0,0.4	99,400	RIDG	20.5	20.5	20.5	20.4	20.4	20.3	20.6	20.5	20.4
		SRC	16.0	16.2	17.2	18.7	20.8	23.4	26.2	29.6	33.4
		FRC	16.0	15.9	**15.9**	**16.2**	**16.7**	**17.2**	**17.8**	**18.2**	**18.6**
		HYB	**15.7**	**15.8**	16.2	16.7	17.3	17.8	18.4	18.7	18.9
0.75,0.4	99,400	RIDG	101.5	100.4	101.2	101.1	100.6	99.8	100.5	101.2	100.6
		SRC	100.2	100.8	105.1	116.9	134.4	161.7	196.1	235.1	280.7
		FRC	100.2	96.7	**91.6**	**88.9**	**87.1**	**86.7**	**86.8**	**87.8**	**88.7**
		HYB	**93.7**	**92.8**	91.8	91.4	91.6	91.9	92.9	95.2	95.8
0,0.1	99,400	RIDG	85.8	86.7	86.4	86.6	86.7	86.0	86.2	86.2	85.9
		SRC	94.9	96.4	100.3	106.8	115.4	125.2	138.2	151.4	167.1
		FRC	94.9	94.1	92.0	89.8	88.0	86.1	85.6	**84.8**	**84.3**
		HYB	**84.0**	**84.7**	**84.6**	**85.1**	**85.4**	**84.7**	**85.5**	85.3	85.3
0.75,0.1	99,400	RIDG	524.8	526.3	531.8	531.3	531.3	531.9	527.3	529.7	531.9
		SRC	595.8	600.8	623.8	678.6	760.8	875.3	1006.9	1160.7	1351.2
		FRC	595.7	577.1	546.6	526.9	**513.4**	**506.6**	**500.3**	**496.8**	**495.9**
		HYB	**518.0**	**518.6**	**520.7**	**521.6**	521.3	522.6	521.7	526.4	527.3
0,0.4	99,150	RIDG	20.5	20.3	20.4	20.5	20.6	20.5	20.5	20.4	20.4
		SRC	24.6	25.5	28.6	33.1	39.1	46.5	55.1	65.8	77.5
		FRC	24.6	24.4	24.1	23.6	23.2	22.7	22.4	22.2	21.9
		HYB	**18.5**	**18.6**	**18.9**	**19.3**	**19.7**	**19.8**	**20.0**	**20.0**	**20.1**
0.75,0.4	99,150	RIDG	101.3	102.3	100.1	**100.6**	102.4	**101.3**	**100.6**	**100.0**	**100.7**
		SRC	155.2	159.2	178.1	215.7	269.6	341.3	428.6	538.5	656.1
		FRC	155.2	145.1	130.2	119.8	111.9	107.3	104.6	102.1	101.3
		HYB	**100.1**	**100.8**	**99.7**	100.6	**101.7**	101.7	101.7	100.8	102.5
0,0.1	99,150	RIDG	**86.1**	**86.0**	**86.0**	**86.0**	**85.9**	**86.7**	**85.6**	**86.2**	**85.6**
		SRC	147.7	151.9	162.8	182.0	207.0	238.7	271.7	315.2	358.0
		FRC	147.7	144.8	137.0	128.7	120.8	114.6	108.3	105.2	101.0
		HYB	86.3	86.3	86.4	86.4	86.5	87.4	86.2	87.0	86.3
0.75,0.1	99,150	RIDG	**534.6**	**533.1**	**528.2**	**530.3**	**528.7**	**532.4**	**529.3**	**530.6**	**524.8**
		SRC	925.1	950.8	1051.6	1249.7	1501.3	1874.5	2271.3	2782.9	3311.8
		FRC	924.9	866.8	772.5	705.7	653.8	627.7	602.8	584.8	569.3
		HYB	536.3	533.6	532.3	536.2	535.1	542.6	541.8	541.9	540.1

Table A2: Numerical values of empirical MSPE for 8 simulation settings described in Section 2.4 when $p = 99$.The smallest MSPE for each τ in each rectangle is in **bold**

$\{\rho, R^2\}$	$\{p, n_B\}$	method	$\tau =$ 0.01	0.25	0.5	0.75	1	1.25	1.5	1.75	2
0,0.4	5,400	RIDG	1.06	1.06	1.06	1.06	1.06	1.07	1.06	1.06	1.06
		SRC	0.97	**0.97**	**0.97**	**0.98**	**0.98**	**0.99**	**1.00**	**1.01**	**1.01**
		FRC	**0.97**	0.97	0.99	1.04	1.11	1.18	1.25	1.30	1.35
		HYB	0.98	0.98	0.99	0.99	1.00	1.00	1.01	1.01	1.02
0.75,0.4	5,400	RIDG	0.94	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.93
		SRC	0.87	**0.87**	**0.88**	**0.89**	**0.90**	0.91	0.92	0.92	0.93
		FRC	**0.87**	0.87	0.89	0.94	1.00	1.08	1.13	1.19	1.22
		HYB	0.88	0.88	0.89	0.89	0.90	**0.91**	**0.92**	**0.92**	**0.92**
0,0.1	5,400	RIDG	6.18	6.17	6.17	6.19	6.19	6.19	6.16	6.16	6.18
		SRC	5.80	5.80	5.82	**5.84**	**5.88**	**5.89**	**5.93**	**5.96**	**6.00**
		FRC	**5.80**	**5.80**	**5.81**	5.86	5.93	5.99	6.04	6.09	6.15
		HYB	5.92	5.91	5.91	5.94	5.97	5.98	5.98	6.00	6.04
0.75,0.1	5,400	RIDG	5.51	5.53	5.51	5.52	5.52	5.49	5.50	5.52	5.50
		SRC	5.20	5.22	5.24	5.29	5.32	**5.35**	**5.39**	5.45	5.46
		FRC	**5.20**	**5.20**	**5.21**	**5.26**	**5.32**	5.37	5.43	5.49	5.52
		HYB	5.29	5.32	5.31	5.34	5.36	5.37	5.40	**5.43**	**5.43**
0,0.4	5,150	RIDG	1.07	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06
		SRC	0.98	0.98	**0.99**	**0.99**	**1.01**	**1.02**	**1.03**	1.04	**1.04**
		FRC	**0.98**	**0.98**	1.00	1.03	1.10	1.16	1.23	1.28	1.32
		HYB	1.00	1.00	1.00	1.01	1.02	1.02	1.03	**1.03**	1.04
0.75,0.4	5,150	RIDG	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.94	0.94
		SRC	0.88	0.88	**0.89**	**0.90**	0.92	0.92	0.93	0.94	0.94
		FRC	**0.88**	**0.88**	0.89	0.93	0.99	1.05	1.11	1.17	1.20
		HYB	0.89	0.89	0.90	0.91	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**
0,0.1	5,150	RIDG	6.18	6.18	6.18	6.19	6.17	6.15	6.17	6.17	6.19
		SRC	5.90	5.89	5.91	5.96	6.00	6.02	6.07	6.12	6.16
		FRC	**5.90**	**5.88**	**5.88**	**5.92**	**5.95**	**5.99**	**6.04**	6.10	6.14
		HYB	5.99	5.98	5.99	6.02	6.03	6.04	6.06	**6.09**	**6.11**
0.75,0.1	5,150	RIDG	5.50	5.52	5.51	5.51	5.52	5.51	5.50	5.52	5.52
		SRC	5.27	5.30	5.33	5.40	5.44	5.49	5.52	5.55	5.58
		FRC	**5.27**	**5.27**	**5.25**	**5.30**	**5.34**	**5.39**	**5.43**	5.47	5.51
		HYB	5.36	5.37	5.36	5.39	5.42	5.44	5.44	**5.46**	**5.47**

Table A3: Numerical values of empirical MSPE for 8 simulation settings described in Section 2.4 when $p = 5$.The smallest MSPE for each τ in each rectangle is in **bold**

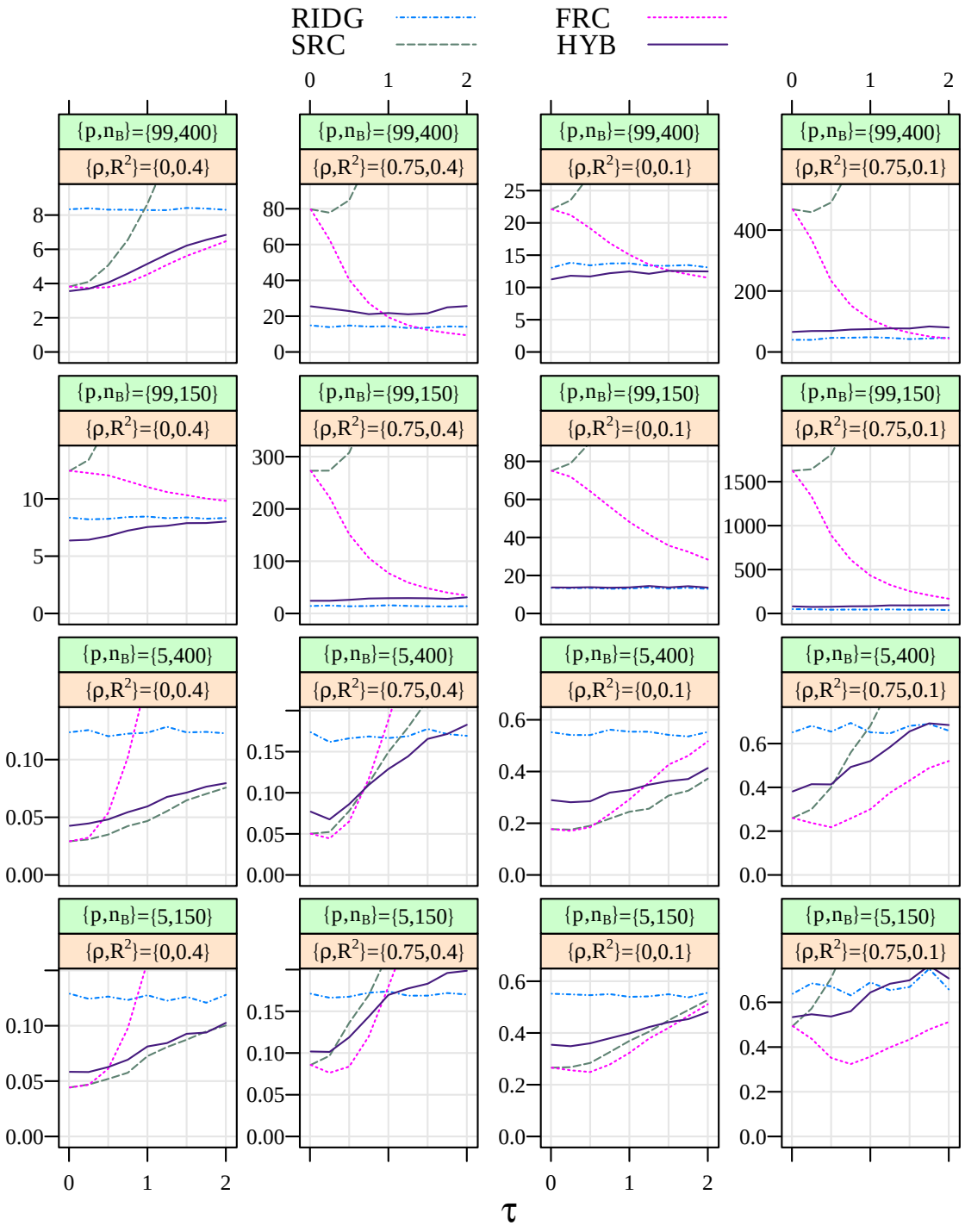

Figure A1: Empirical MSE over τ for 16 simulation settings described in Section 2.4 p stands for the number of covariates, n_B is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_A = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. The smallest possible MSE is zero.

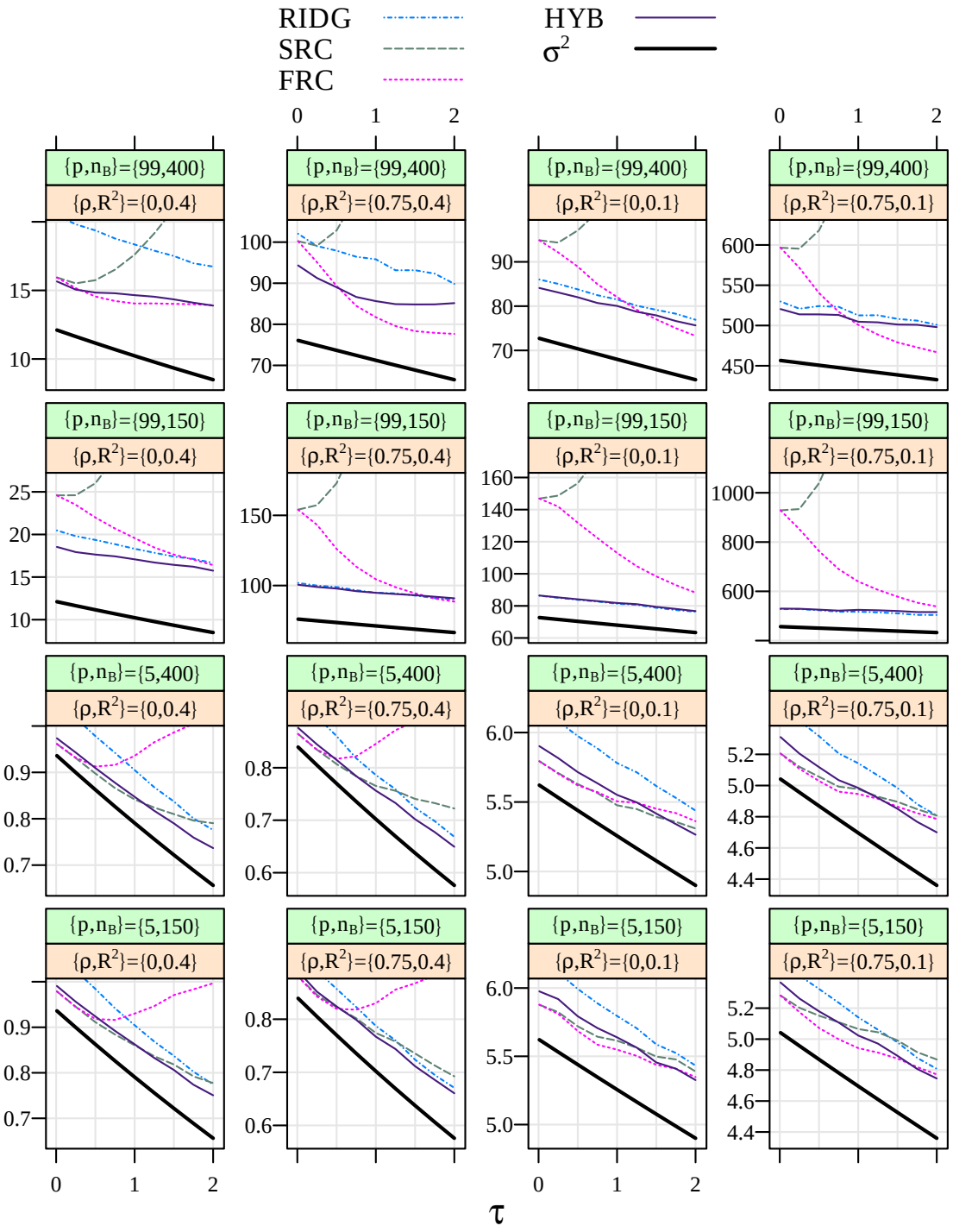

Figure A2: Empirical MSPE over τ for 16 simulation settings described in Section 2.4 when the conditional independence assumption $[Y|\boldsymbol{X},\boldsymbol{W}] = [Y|\boldsymbol{X}]$ is violated. p stands for the number of covariates, n_B is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_A = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. σ^2, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

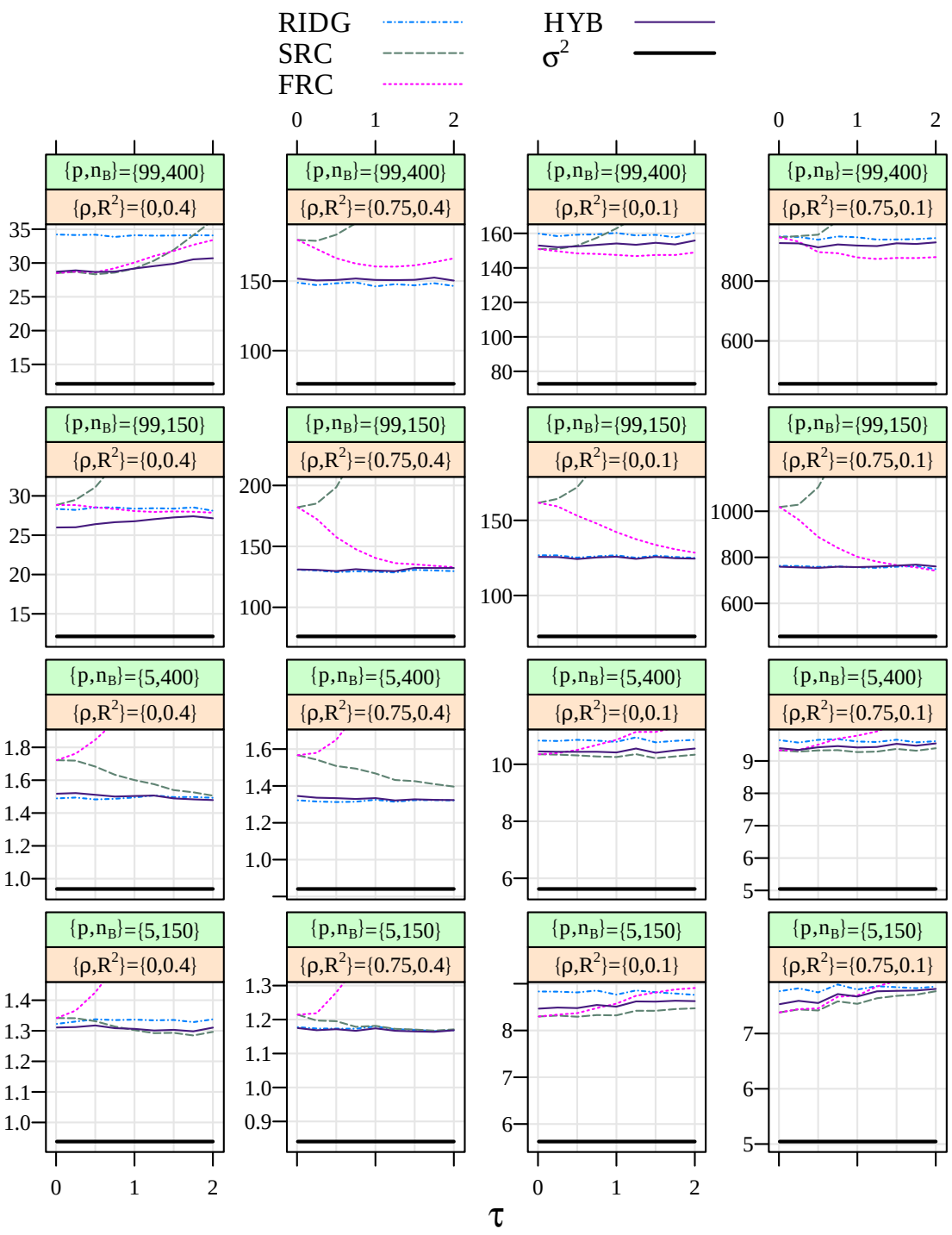

Figure A3: Empirical MSPE over τ for 16 simulation settings described in Section 2.4 under outcome dependent sampling. p stands for the number of covariates, n_B is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_A = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. σ^2, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

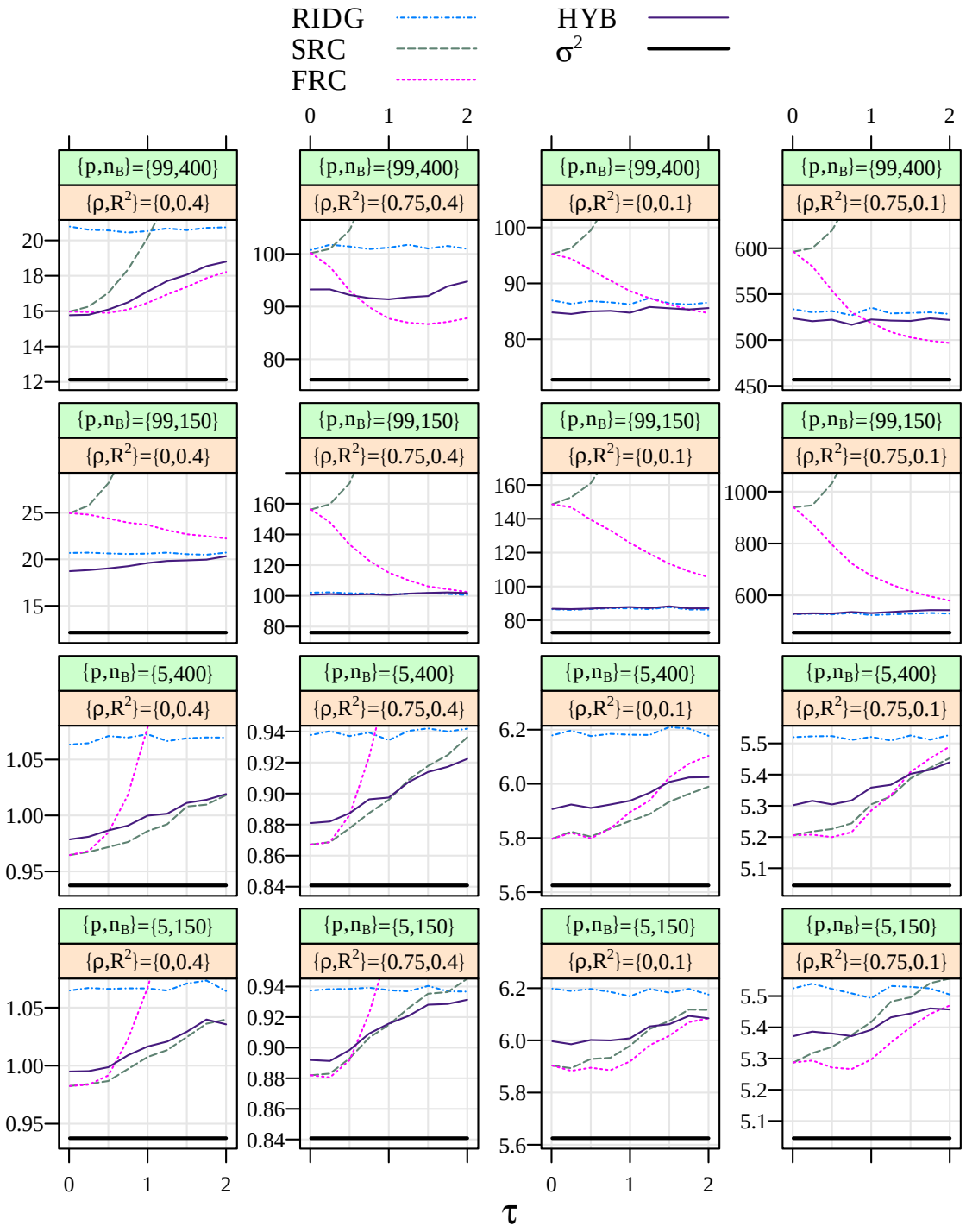

Figure A4: Empirical MSPE over τ for 16 simulation settings described in Section 2.4 under violations to normality of $\boldsymbol{X}$ assumption and ME structure. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2)$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}} = 50$, $\beta_0 = \psi = 0$, and $\nu = 1.\sigma^2$, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

Proof.

$$\mathrm{E}\left[\mathrm{Var}[U|V_1,\ldots,V_N]\right] = \mathrm{E}\left[\frac{1}{N}\sum V_i^2 - \bar{V}^2\right] = \sigma^2 - \sigma^2/N = \frac{N-1}{N}\sigma^2.$$

□

Applying this result to the proposed bootstrap algorithm in the main text, let U be $y^{b*} - \boldsymbol{x}^{b*}\hat{\boldsymbol{\beta}}^b$, a random draw from the r_{A}^b residuals of the observations not sampled in step (i). Ignoring the bias and variance of $\boldsymbol{\beta}^b$, these residuals, corresponding to $V_1,\ldots,V_{r_{\mathrm{A}}^b}$, are approximately $N\{0,\sigma^2\}$. Thus, if $e^{b*} = \sqrt{r_{\mathrm{A}}^b/(r_{\mathrm{A}}^b-1)}\left(y^{b*} - \boldsymbol{x}^{b*}\hat{\boldsymbol{\beta}}^b\right)$, $\mathrm{E}\left[\mathrm{Var}\left[e^{b*}\right]\right]$ is approximately σ^2, which is our justification for using e^{b*} as the prediction error.

Appendix B: Chapter 3 Supplementary Materials

B.1 Enumeration of Gibbs Steps

FB-FLATBETA From (27), the prior is

$$
\begin{aligned}
[\boldsymbol{\phi}] &= [\beta_0, \boldsymbol{\beta}, \sigma^2, \psi, \nu, \tau^2, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}] \\
&\propto (\sigma^2\tau^2)^{-1} |\boldsymbol{\Sigma}_X^{-1}|^{(2p-1)/2} \exp\left\{ -\frac{2p-1}{2} \mathrm{Tr}\left(\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_{\mathrm{A}}]) \boldsymbol{\Sigma}_X^{-1} \right) \right\},
\end{aligned}
$$

where $\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_{\mathrm{A}}])$ is the diagonal part of the empirical covariance of $\boldsymbol{x}_{\mathrm{A}}$. This is a Jeffreys prior on each parameter except $\boldsymbol{\Sigma}_X^{-1}$, and $\boldsymbol{\eta}$ (the hyperparameter) is specified. Using partial conditional distributions implied by the product of expressions (25) and (27): $[\boldsymbol{U}^{\mathrm{obs}}, \boldsymbol{U}^{\mathrm{mis}} | \boldsymbol{\phi}] \times [\boldsymbol{\phi}]$, every Gibbs step is derived as follows:

$$
\begin{aligned}
&\left[\boldsymbol{x}_i | y_i, \boldsymbol{w}_i, \boldsymbol{\beta}, \beta_0, \sigma^2, \psi, \nu, \tau^2, \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1} \right] \\
&\quad \propto \left[y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \beta_0, \sigma^2 \right] \left[\boldsymbol{w}_i | \boldsymbol{x}_i, \psi, \nu, \tau^2 \right] \left[\boldsymbol{x}_i | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1} \right] \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \beta_0 - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 \right\} \exp\left\{ -\frac{1}{2\tau^2} (\boldsymbol{w}_i - \psi \mathbf{1}_p - \nu \boldsymbol{x}_i)^\top (\boldsymbol{w}_i - \psi \mathbf{1}_p - \nu \boldsymbol{x}_i) \right\} \\
&\qquad \times \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_X)^\top \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_X) \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2} \Big(\boldsymbol{x}_i^\top [\boldsymbol{\beta}\boldsymbol{\beta}^\top / \sigma^2] \boldsymbol{x}_i - 2[(y_i - \beta_0)/\sigma^2] \boldsymbol{x}_i^\top \boldsymbol{\beta} \right. \\
&\qquad\qquad \left. + \, [\nu^2/\tau^2] \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2[\nu/\tau^2] \boldsymbol{x}_i^\top [\boldsymbol{w}_i - \psi \mathbf{1}_p] + \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{x}_i - 2 \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X \Big) \right\} \\
&\quad = N_p \left\{ \Gamma \left([(y_i - \beta_0)/\sigma^2] \boldsymbol{\beta} + [\nu/\tau^2] [\boldsymbol{w}_i - \psi \mathbf{1}_p] + \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\mu}_X \right), \Gamma \right\}, \\
&\Gamma = \left(\boldsymbol{\beta}\boldsymbol{\beta}^\top / \sigma^2 + \nu^2/\tau^2 + \boldsymbol{\Sigma}_X^{-1} \right)^{-1}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\boldsymbol{\beta}|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\beta_0,\sigma^2,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{y}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2\right] \\
&\quad\propto \exp\Big\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&\qquad\qquad -\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\Big\} \\
&\quad\propto \exp\Big\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\left[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right]\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\left[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}})+\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}})\right]\right)\Big\} \\
&\quad= N_p\left\{(\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}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}]+\boldsymbol{x}_{\mathrm{B}}^\top[\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}]),\sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}})^{-1}\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\beta_0|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\sigma^2,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{y}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2\right] \\
&\quad\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})\right\} \\
&\qquad\times\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\right\} \\
&\quad\propto \exp\left\{-\frac{1}{2\sigma^2}\left([n_{\mathrm{A}}+n_{\mathrm{B}}]\beta_0^2-2\beta_0(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top\mathbf{1}_{n_{\mathrm{A}}}-2\beta_0(\boldsymbol{y}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top\mathbf{1}_{n_{\mathrm{B}}}\right)\right\} \\
&\quad= N\left\{\frac{(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top\mathbf{1}_{n_{\mathrm{A}}}+(\boldsymbol{y}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top\mathbf{1}_{n_{\mathrm{B}}}}{n_{\mathrm{A}}+n_{\mathrm{B}}},\frac{\sigma^2}{n_{\mathrm{A}}+n_{\mathrm{B}}}\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\sigma^2|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{y}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2\right] \\
&\quad\propto (\sigma^2)^{-n_{\mathrm{A}}/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})\right\} \\
&\qquad\times(\sigma^2)^{-n_{\mathrm{B}}/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\right\}(\sigma^2)^{-1} \\
&\quad= IG\Big\{\frac{1}{2}\left(n_{\mathrm{A}}+n_{\mathrm{B}}\right),\frac{1}{2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&\qquad\qquad\qquad+\frac{1}{2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\Big\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\psi|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{w}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\psi,\nu,\tau^2\right]\left[\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\psi,\nu,\tau^2\right] \\
&\quad\propto \exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})\right\} \\
&\qquad\times\exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})\right\} \\
&\quad\propto \exp\Bigg\{-\frac{1}{2\tau^2}\Big(\psi^2\mathrm{Tr}\,[\mathbf{1}_p\mathbf{1}_{n_{\mathrm{A}}}^\top\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top+\mathbf{1}_p\mathbf{1}_{n_{\mathrm{B}}}^\top\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top] \\
&\qquad\qquad -2\psi\mathrm{Tr}\,[\mathbf{1}_p\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\nu\boldsymbol{x}_{\mathrm{A}})+\mathbf{1}_p\mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\nu\boldsymbol{x}_{\mathrm{B}})]\Big)\Bigg\} \\
&\quad= N\left\{\frac{\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\nu\boldsymbol{x}_{\mathrm{A}})\mathbf{1}_p+\mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\nu\boldsymbol{x}_{\mathrm{B}})\mathbf{1}_p}{(n_{\mathrm{A}}+n_{\mathrm{B}})p},\frac{\tau^2}{(n_{\mathrm{A}}+n_{\mathrm{B}})p}\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\nu|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\psi,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{w}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\psi,\nu,\tau^2\right]\left[\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\psi,\nu,\tau^2\right] \\
&\quad\propto \exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})\right\} \\
&\qquad\times\exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})\right\} \\
&\quad\propto \exp\Bigg\{-\frac{1}{2\tau^2}\Big(\nu^2\mathrm{Tr}\,[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}] \\
&\qquad\qquad -2\nu\mathrm{Tr}\,[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top)+\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top)]\Big)\Bigg\} \\
&\quad= N\left\{\frac{\mathrm{Tr}\,[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top)+\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top)]}{\mathrm{Tr}\,[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}]},\frac{\tau^2}{\mathrm{Tr}\,[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}]}\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\tau^2|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\psi,\nu,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{w}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\psi,\nu,\tau^2\right]\left[\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\psi,\nu,\tau^2\right] \\
&\quad\propto \left(\tau^2\right)^{-(n_{\mathrm{A}}+n_{\mathrm{B}})p/2}\exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})\right\} \\
&\qquad\times\exp\left\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})\right\}(\tau^2)^{-1} \\
&\quad= IG\Bigg\{\frac{1}{2}(n_{\mathrm{A}}+n_{\mathrm{B}})p,\frac{1}{2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}) \\
&\qquad\qquad +\frac{1}{2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})\Bigg\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\boldsymbol{\mu}_X | \boldsymbol{y}_\mathrm{A}, \boldsymbol{x}_\mathrm{A}, \boldsymbol{w}_\mathrm{A}, \boldsymbol{y}_\mathrm{B}, \boldsymbol{x}_\mathrm{B}, \boldsymbol{w}_\mathrm{B}, \boldsymbol{\beta}, \beta_0, \sigma^2, \psi, \nu, \tau^2, \boldsymbol{\Sigma}_X^{-1}\right] \\
&\quad \propto \left[\boldsymbol{x}_\mathrm{A} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}\right] \left[\boldsymbol{x}_\mathrm{B} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}\right] \\
&\quad \propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\,(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top - \frac{1}{2}\mathrm{Tr}\,(\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2}\left([n_\mathrm{A} + n_\mathrm{B}]\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X - 2\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}[\boldsymbol{x}_\mathrm{A}^\top\mathbf{1}_{n_\mathrm{A}} + \boldsymbol{x}_\mathrm{B}^\top\mathbf{1}_{n_\mathrm{B}}]\right)\right\} \\
&\quad = N_p\left\{\frac{\boldsymbol{x}_\mathrm{A}^\top\mathbf{1}_{n_\mathrm{A}} + \boldsymbol{x}_\mathrm{B}^\top\mathbf{1}_{n_\mathrm{B}}}{n_\mathrm{A} + n_\mathrm{B}}, \frac{1}{n_\mathrm{A} + n_\mathrm{B}}\boldsymbol{\Sigma}_X\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\boldsymbol{\Sigma}_X^{-1} | \boldsymbol{y}_\mathrm{A}, \boldsymbol{x}_\mathrm{A}, \boldsymbol{w}_\mathrm{A}, \boldsymbol{y}_\mathrm{B}, \boldsymbol{x}_\mathrm{B}, \boldsymbol{w}_\mathrm{B}, \boldsymbol{\beta}, \beta_0, \sigma^2, \psi, \nu, \tau^2, \boldsymbol{\mu}_X\right] \\
&\quad \propto \left[\boldsymbol{x}_\mathrm{A} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}\right] \left[\boldsymbol{x}_\mathrm{B} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X^{-1}\right] \left[\boldsymbol{\Sigma}_X^{-1}\right] \\
&\quad = \left|\boldsymbol{\Sigma}_X^{-1}\right|^{(n_\mathrm{A}+n_\mathrm{B})/2} \\
&\qquad \times \exp\left\{-\frac{1}{2}\mathrm{Tr}\;\boldsymbol{\Sigma}_X^{-1}\left[(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top) + (\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\right]\right\} \\
&\qquad \times \left|\boldsymbol{\Sigma}_X^{-1}\right|^{(2p-1)/2} \exp\left\{-\frac{2p-1}{2}\mathrm{Tr}\;(\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_\mathrm{A}])\boldsymbol{\Sigma}_X^{-1})\right\} \\
&\quad = W\Big\{3p + n_\mathrm{A} + n_\mathrm{B}, \Big((2p-1)\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_\mathrm{A}]) + (\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top) \\
&\qquad\qquad + (\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{B} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\Big)^{-1}\Big\}.
\end{aligned}
$$

The Inverse-Gamma distribution, $IG\{a, b\}$, is parametrized to have mean $\frac{b}{a-1}$ and the Wishart distribution with d degrees of freedom, $W\{d, \boldsymbol{S}\}$, is parametrized to have mean $d\boldsymbol{S}$.

FB-HIBETA-NI, EB-HIBETA-NI Recall that we replace the Jeffreys prior on $\boldsymbol{\beta}$ in (27) with

$$
[\boldsymbol{\beta} | \sigma^2, \lambda] \propto \left(\frac{\lambda}{\sigma^2}\right)^{p/2} \exp\left\{-\frac{1}{2}\frac{\lambda}{\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}.
$$

The following posterior steps are modified:

$$
\begin{aligned}
&\left[\boldsymbol{\beta}|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\beta_0,\sigma^2,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1},\lambda\right] \\
&\quad\propto \left[\boldsymbol{y}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{\beta}|\sigma^2,\lambda\right] \\
&\quad\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\right\}\exp\left\{-\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\} \\
&\quad\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top\left[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}+\lambda\boldsymbol{I}_p\right]\boldsymbol{\beta}-2\boldsymbol{\beta}^\top\left[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}})+\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}})\right]\right)\right\} \\
&\quad= N_p\Big\{(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}+\lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{x}_{\mathrm{A}}^\top[\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}]+\boldsymbol{x}_{\mathrm{B}}^\top[\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}]), \\
&\qquad\sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}+\lambda\boldsymbol{I}_p)^{-1}\Big\}.
\end{aligned}
$$

$$
\begin{aligned}
&\left[\sigma^2|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1},\lambda\right] \\
&\quad\propto \left[\boldsymbol{y}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{y}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2\right]\left[\boldsymbol{\beta}|\sigma^2,\lambda\right]\left[\sigma^2\right] \\
&\quad\propto (\sigma^2)^{-n_{\mathrm{A}}/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})\right\} \\
&\qquad\times(\sigma^2)^{-n_{\mathrm{B}}/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\right\}(\sigma^2)^{-1} \\
&\qquad\times(\sigma^2)^{-p/2}\exp\left\{-\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\} \\
&\quad= IG\Big\{\frac{1}{2}\left(n_{\mathrm{A}}+n_{\mathrm{B}}+p\right),\frac{1}{2}\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}+\frac{1}{2}(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{A}}-\beta_0\mathbf{1}_{n_{\mathrm{A}}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&\qquad\qquad+\frac{1}{2}(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})^\top(\boldsymbol{y}_{\mathrm{B}}-\beta_0\mathbf{1}_{n_{\mathrm{B}}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\beta})\Big\}.
\end{aligned}
$$

The FB-HIBETA-NI update for λ is given as follows:

$$
\begin{aligned}
&\left[\lambda|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\nu,\tau^2,\boldsymbol{\mu}_{\boldsymbol{X}},\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\right] \\
&\quad\propto \left[\boldsymbol{\beta}|\sigma^2,\lambda\right]\left[\lambda\right] \\
&\quad\propto \lambda^{p/2}\exp\left\{-\frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right\}\lambda^{-1} \\
&\quad= G\left\{\frac{p}{2},\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2\sigma^2}\right\}.
\end{aligned}
$$

To calculate the EB-HIBETA-NI update for λ, observe that $\mathrm{E}_{\boldsymbol{\phi}|\boldsymbol{U}^{\mathrm{obs}},\lambda}\ln[\boldsymbol{\beta}|\sigma^2,\lambda]=(p/2)\ln\lambda-\lambda\mathrm{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma^2\right]/2$. This is maximized with respect to λ when $\lambda=p/\mathrm{E}\left[\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma^2\right]$.

EB-HISIGMAX, EB-HIBETASIGMAX Leaving the inverse scale matrix $\boldsymbol{\Lambda}$ unspecified, the mod-

ified prior on $\boldsymbol{\Sigma}_X^{-1}$ is

$$[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}] \propto |\boldsymbol{\Lambda}|^{3p/2}|\boldsymbol{\Sigma}_X^{-1}|^{(2p-1)/2}\exp\left\{-(1/2)\mathrm{Tr}\,(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{-1})\right\}.$$

$\boldsymbol{\Lambda}$ is the unknown positive-definite matrix of hyperparameters. The Gibbs step for $\boldsymbol{\Sigma}_X^{-1}$ becomes

$$\begin{aligned}
&\left[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{y}_\mathrm{A},\boldsymbol{x}_\mathrm{A},\boldsymbol{w}_\mathrm{A},\boldsymbol{y}_\mathrm{B},\boldsymbol{x}_\mathrm{B},\boldsymbol{w}_\mathrm{B},\boldsymbol{\beta},\beta_0,\sigma^2,\psi,\nu,\tau^2,\boldsymbol{\mu}_X\right] \\
&\quad\propto \left[\boldsymbol{x}_\mathrm{A}|\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X^{-1}\right]\left[\boldsymbol{x}_\mathrm{B}|\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X^{-1}\right]\left[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}\right] \\
&\quad\propto \left|\boldsymbol{\Sigma}_X^{-1}\right|^{(n_\mathrm{A}+n_\mathrm{B})/2}\Big\{-\frac{1}{2}\mathrm{Tr}\,(\boldsymbol{x}_\mathrm{A}-\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x}_\mathrm{A}-\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top\Big\} \\
&\qquad\times\exp\Big\{-\frac{1}{2}\mathrm{Tr}\,(\boldsymbol{x}_\mathrm{B}-\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x}_\mathrm{B}-\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top\Big\} \\
&\qquad\times\left|\boldsymbol{\Sigma}_X^{-1}\right|^{(2p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_X^{-1}\right)\right\} \\
&\quad= W\Big\{3p+n_\mathrm{A}+n_\mathrm{B}, \\
&\qquad\left(\boldsymbol{\Lambda}+(\boldsymbol{x}_\mathrm{A}-\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{A}-\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)+(\boldsymbol{x}_\mathrm{B}-\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{B}-\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\right)^{-1}\Big\}.
\end{aligned}$$

We now derive the Empirical Bayes update for the diagonal inverse-scale matrix $\boldsymbol{\Lambda} = \mathrm{diag}\{\Lambda_{11},\ldots,\Lambda_{pp}\}$. Observe that

$$\mathrm{E}_{\boldsymbol{\phi}|\boldsymbol{U}^\mathrm{obs},\boldsymbol{\Lambda}}\ln\left[\boldsymbol{\Sigma}_X^{-1}|\boldsymbol{\Lambda}\right] \propto p\ln|\boldsymbol{\Lambda}| - \mathrm{Tr}\,(\boldsymbol{\Lambda}\mathrm{E}[\boldsymbol{\Sigma}_X^{-1}]) = 3p\sum_{i=1}^{p}\log\Lambda_{ii} - \sum_{i=1}^{p}\Lambda_{ii}\mathrm{E}[\boldsymbol{\Sigma}_X^{-1}]_{ii}.$$

Thus, each element Λ_{ii} may be optimized individually, yielding the Empirical Bayes update $\Lambda_{ii} \leftarrow 3p\mathrm{E}[\boldsymbol{\Sigma}_X^{-1}]_{ii}^{-1}$.

B.2 Modified Gibbs Steps for Data Analysis

Let $\boldsymbol{\psi} \equiv \{\psi_1,\ldots,\psi_p\}$, $\boldsymbol{\nu} \equiv \mathrm{diag}\{\nu_1,\ldots,\nu_p\}$ (that is, a diagonal matrix with components $\nu_1,\ldots,\nu_p$), and $\{\boldsymbol{e}_j\}$ the set of p-dimensional standard basic vectors. The conditional distributions with individual intercepts and slopes (using flat priors on each ψ_j and ν_j)

are given by

$$
\begin{aligned}
&\Big[\boldsymbol{\psi}|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\boldsymbol{\nu},\tau^2,\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X^{-1}\Big] \\
&\quad\propto \Big[\boldsymbol{w}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\boldsymbol{\psi},\boldsymbol{\nu},\tau^2\Big]\Big[\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\boldsymbol{\psi},\boldsymbol{\nu},\tau^2\Big] \\
&\quad\propto \exp\Big\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\nu})^\top(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\nu})\Big\} \\
&\qquad\times\exp\Big\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\nu})^\top(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\nu})\Big\} \\
&\quad\propto \exp\Big\{-\frac{1}{2\tau^2}\Big(\mathrm{Tr}\,[\boldsymbol{\psi}\mathbf{1}_{n_{\mathrm{A}}}^\top\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top+\boldsymbol{\psi}\mathbf{1}_{n_{\mathrm{B}}}^\top\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top] \\
&\qquad-2\mathrm{Tr}\,[\boldsymbol{\psi}\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\nu})+\boldsymbol{\psi}\mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\nu})]\Big)\Big\} \\
&\quad=\exp\left\{-\frac{1}{2\tau^2}\left((n_{\mathrm{A}}+n_{\mathrm{B}})\sum_{j=1}^{p}\psi_j^2-\sum_{j=1}^{p}\psi_j\boldsymbol{e}_j^\top\Big[(\boldsymbol{w}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\nu})^\top\mathbf{1}_{n_{\mathrm{A}}}+(\boldsymbol{w}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\nu})^\top\mathbf{1}_{n_{\mathrm{B}}}\Big]\right)\right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\Big[\nu|\boldsymbol{y}_{\mathrm{A}},\boldsymbol{x}_{\mathrm{A}},\boldsymbol{w}_{\mathrm{A}},\boldsymbol{y}_{\mathrm{B}},\boldsymbol{x}_{\mathrm{B}},\boldsymbol{w}_{\mathrm{B}},\boldsymbol{\beta},\beta_0,\sigma^2,\psi,\tau^2,\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X^{-1}\Big] \\
&\quad\propto \Big[\boldsymbol{w}_{\mathrm{A}}|\boldsymbol{x}_{\mathrm{A}},\psi,\nu,\tau^2\Big]\Big[\boldsymbol{w}_{\mathrm{B}}|\boldsymbol{x}_{\mathrm{B}},\psi,\nu,\tau^2\Big] \\
&\quad\propto \exp\Big\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{A}}\nu)^\top(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{A}}\nu)\Big\} \\
&\qquad\times\exp\Big\{-\frac{1}{2\tau^2}\mathrm{Tr}\,(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{B}}\nu)^\top(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top-\boldsymbol{x}_{\mathrm{B}}\nu)\Big\} \\
&\quad\propto \exp\Big\{-\frac{1}{2\tau^2}\Big(\mathrm{Tr}\,[\nu^2\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\nu^2\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}]-2\mathrm{Tr}\,[\nu\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top)+\nu\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top)]\Big)\Big\}.
\end{aligned}
$$

From these, the modified Gibbs steps are

$$
\begin{aligned}
\psi_j &\leftarrow N\Big\{\frac{\boldsymbol{e}_j^\top\big[(\boldsymbol{w}_{\mathrm{A}}-\boldsymbol{x}_{\mathrm{A}}\boldsymbol{\nu})^\top\mathbf{1}_{n_{\mathrm{A}}}+(\boldsymbol{w}_{\mathrm{B}}-\boldsymbol{x}_{\mathrm{B}}\boldsymbol{\nu})^\top\mathbf{1}_{n_{\mathrm{B}}}\big]}{(n_{\mathrm{A}}+n_{\mathrm{B}})},\frac{\tau^2}{(n_{\mathrm{A}}+n_{\mathrm{B}})}\Big\}, \\
\nu_j &\leftarrow N\Big\{\frac{\boldsymbol{e}_j^\top\big[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\mathbf{1}_{n_{\mathrm{A}}}\boldsymbol{\psi}^\top)+\boldsymbol{x}_{\mathrm{B}}^\top(\boldsymbol{w}_{\mathrm{B}}-\mathbf{1}_{n_{\mathrm{B}}}\boldsymbol{\psi}^\top)\big]\boldsymbol{e}_j}{\boldsymbol{e}_j^\top\big[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\big]\boldsymbol{e}_j},\frac{\tau^2}{\boldsymbol{e}_j^\top[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}+\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}]\boldsymbol{e}_j}\Big\}.
\end{aligned}
$$

independently for $j=1,\ldots,p$.

$\{\rho, R^2\}$	$\{p, n_B\}$	Method	$\tau = 0.01$	MSPE($\hat{\beta}^{ppm}$) 0.5	1.0	1.50	2.00	0.01	MSPE($\hat{\beta}^{pm}$) 0.5	1.0	1.50	2.00
0,0.4	99,400	RIDG	20.6	20.5	20.6	20.6	20.3	20.6	20.5	20.6	20.6	20.3
		HYB	15.7	16.1	17.3	18.4	18.7	15.7	16.1	17.3	18.4	18.7
		FB-FLATBETA	15.6	18.5	28.4	38.6	41.0	15.6	19.3	32.3	42.1	44.2
		FB-HIBETA-NI	**15.0**	16.0	16.8	17.6	18.1	**15.0**	16.2	17.0	17.8	18.3
		EB-HIBETA-NI	15.0	15.5	16.6	17.5	18.1	15.0	15.6	16.7	17.5	18.1
		EB-HISIGMAX	15.6	17.5	25.3	36.5	40.0	15.6	18.0	28.6	40.0	43.1
		EB-HIBETASIGMAX	15.1	**15.5**	**16.6**	**17.4**	**18.0**	15.1	**15.5**	**16.6**	**17.5**	**18.1**
0.75,0.4	99,400	RIDG	101.3	101.8	103.0	100.5	101.7	101.3	101.8	103.0	100.5	101.7
		HYB	94.0	92.3	92.7	92.6	96.7	94.0	92.3	92.7	92.6	96.7
		FB-FLATBETA	98.4	97.2	102.8	109.5	119.2	98.3	99.2	110.3	124.5	138.8
		FB-HIBETA-NI	89.1	89.1	89.6	90.3	92.3	89.1	89.7	90.9	92.4	94.3
		EB-HIBETA-NI	**80.9**	81.7	**82.4**	**83.1**	**85.8**	**80.9**	81.7	**82.3**	**83.3**	**86.1**
		EB-HISIGMAX	98.3	105.1	113.8	127.8	147.8	98.3	108.9	125.6	148.9	172.2
		EB-HIBETASIGMAX	80.9	**81.6**	83.0	84.3	87.5	80.9	**81.6**	83.0	84.5	87.6
0,0.1	99,400	RIDG	86.6	86.1	86.0	86.4	85.8	86.6	86.1	86.0	86.4	85.8
		HYB	84.5	84.7	84.9	85.2	85.1	84.5	84.7	84.9	85.2	85.1
		FB-FLATBETA	93.2	107.7	142.5	192.7	206.7	93.2	111.8	163.8	214.6	225.5
		FB-HIBETA-NI	90.9	98.7	97.1	95.5	95.0	90.9	100.8	99.6	97.1	96.1
		EB-HIBETA-NI	**79.5**	80.1	80.3	80.6	80.7	**79.5**	80.1	80.3	80.6	80.6
		EB-HISIGMAX	93.2	102.4	129.9	180.9	197.5	93.2	105.3	147.8	202.6	217.1
		EB-HIBETASIGMAX	79.5	**80.1**	**80.3**	**80.5**	**80.6**	79.5	**80.1**	**80.3**	**80.5**	**80.6**
0.75,0.1	99,400	RIDG	534.3	518.7	526.5	539.2	522.1	534.3	518.7	526.5	539.2	522.1
		HYB	523.4	512.9	521.0	529.9	527.5	523.4	512.9	521.0	529.9	527.5
		FB-FLATBETA	587.8	577.0	606.4	643.7	681.6	587.8	588.6	647.9	723.5	785.0
		FB-HIBETA-NI	535.3	532.5	538.0	545.4	548.3	535.2	536.3	546.7	557.7	561.2
		EB-HIBETA-NI	**481.4**	**481.0**	**482.1**	**486.9**	**485.9**	**481.4**	**480.9**	**481.8**	**486.6**	**485.2**
		EB-HISIGMAX	587.9	622.5	662.8	737.6	825.2	587.9	644.0	728.4	851.6	960.6
		EB-HIBETASIGMAX	481.8	481.2	482.6	488.1	488.1	481.8	481.1	482.3	487.6	487.3
0,0.4	99,150	RIDG	20.5	20.6	20.5	20.5	20.7	20.5	20.6	20.5	20.5	20.7
		HYB	18.6	18.9	19.7	20.0	20.3	18.6	18.9	19.7	20.0	20.3
		FB-FLATBETA	24.2	31.7	57.3	59.2	61.2	24.2	34.1	60.9	62.6	64.3
		FB-HIBETA-NI	17.7	18.0	18.5	18.9	19.2	17.7	18.1	18.6	19.0	19.2
		EB-HIBETA-NI	16.9	17.3	18.1	18.7	19.1	16.9	17.4	18.1	18.8	19.1
		EB-HISIGMAX	24.2	29.8	56.6	54.5	56.5	24.2	31.7	59.9	58.2	59.8
		EB-HIBETASIGMAX	**16.9**	**17.3**	**18.1**	**18.7**	**19.1**	**16.9**	**17.3**	**18.1**	**18.7**	**19.1**
0.75,0.4	99,150	RIDG	101.0	100.4	101.2	102.7	101.1	101.0	100.4	101.2	102.7	101.1
		HYB	100.1	99.5	101.4	103.0	101.9	100.1	99.5	101.4	103.0	101.9
		FB-FLATBETA	153.2	152.5	192.8	202.2	219.6	153.2	161.8	212.2	225.4	242.0
		FB-HIBETA-NI	94.2	93.1	93.1	94.3	94.7	94.2	93.4	93.8	95.5	95.9
		EB-HIBETA-NI	**85.0**	**85.5**	**86.1**	**88.4**	**89.3**	**85.0**	**85.4**	**86.1**	**88.6**	**89.7**
		EB-HISIGMAX	153.4	178.8	231.1	241.0	266.2	153.3	193.4	252.9	264.9	287.6
		EB-HIBETASIGMAX	85.1	85.6	86.7	89.5	90.5	85.0	85.5	86.6	89.6	90.8
0,0.1	99,150	RIDG	85.8	85.0	86.2	86.5	86.0	85.8	85.0	86.2	86.5	86.0
		HYB	86.2	85.5	86.5	87.2	87.1	86.2	85.5	86.5	87.2	87.1
		FB-FLATBETA	145.8	181.8	311.3	292.7	301.6	145.8	194.8	332.0	311.1	317.3
		FB-HIBETA-NI	106.3	103.9	98.7	96.4	95.0	106.3	104.9	99.4	96.7	95.3
		EB-HIBETA-NI	**80.8**	81.1	81.3	**81.6**	81.3	**80.8**	81.1	81.3	**81.5**	81.2
		EB-HISIGMAX	145.8	172.1	286.5	285.3	286.2	145.8	182.6	308.2	304.1	303.5
		EB-HIBETASIGMAX	80.8	**81.1**	**81.3**	81.6	**81.3**	80.8	**81.1**	**81.2**	81.5	**81.2**
0.75,0.1	99,150	RIDG	531.8	533.6	530.4	530.1	524.2	531.8	533.6	530.4	530.1	524.2
		HYB	532.1	535.7	539.3	538.5	539.4	532.1	535.7	539.3	538.5	539.4
		FB-FLATBETA	924.8	919.7	1049.4	1143.2	1161.3	924.5	977.5	1169.8	1269.2	1294.6
		FB-HIBETA-NI	568.6	558.9	559.4	561.5	565.8	568.6	561.0	565.1	569.0	574.0
		EB-HIBETA-NI	488.8	**487.8**	**493.0**	**495.3**	**497.1**	488.8	**487.6**	**492.7**	**494.9**	**496.7**
		EB-HISIGMAX	924.6	1095.3	1333.3	1394.0	1390.7	924.3	1185.6	1466.7	1529.8	1527.1
		EB-HIBETASIGMAX	**488.6**	488.1	493.8	495.7	498.7	**488.6**	487.9	493.3	495.2	498.3

Table B1: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 99$. $\hat{\beta}^{pm}$ is the posterior mean and $\hat{\beta}^{ppm}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

			MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_B\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	5,400	RIDG	1.07	1.06	1.06	1.05	1.06	1.07	1.06	1.06	1.05	1.06
		HYB	0.98	0.98	1.00	1.00	1.02	0.98	0.98	1.00	1.00	1.02
		FB-FLATBETA	0.95	0.96	0.98	0.98	1.00	0.95	0.96	0.98	0.99	1.01
		FB-HIBETA-NI	**0.95**	0.95	0.98	0.98	1.00	**0.95**	0.96	0.98	0.99	1.01
		EB-HIBETA-NI	0.95	0.95	0.98	0.98	1.00	0.95	0.95	0.98	0.98	1.00
		EB-HISIGMAX	0.95	0.95	0.98	0.98	1.00	0.95	0.95	0.98	0.98	1.00
		EB-HIBETASIGMAX	0.95	**0.95**	**0.97**	**0.98**	**1.00**	0.95	**0.95**	**0.97**	**0.98**	**1.00**
0.75,0.4	5,400	RIDG	0.94	0.94	0.93	0.94	0.93	0.94	0.94	0.93	0.94	0.93
		HYB	0.88	0.89	0.90	0.92	0.92	0.88	0.89	0.90	0.92	0.92
		FB-FLATBETA	0.85	0.86	0.87	0.89	0.90	0.85	0.86	0.87	0.89	0.90
		FB-HIBETA-NI	**0.85**	0.86	0.87	0.89	0.89	**0.85**	0.86	0.87	0.89	0.89
		EB-HIBETA-NI	0.85	**0.86**	**0.86**	**0.88**	**0.88**	0.85	**0.86**	**0.86**	**0.88**	**0.88**
		EB-HISIGMAX	0.85	0.86	0.87	0.90	0.90	0.85	0.86	0.87	0.90	0.91
		EB-HIBETASIGMAX	0.85	0.86	0.86	0.88	0.89	0.85	0.86	0.87	0.88	0.89
0,0.1	5,400	RIDG	6.20	6.17	6.17	6.22	6.17	6.20	6.17	6.17	6.22	6.17
		HYB	5.93	5.92	5.95	6.03	6.04	5.93	5.92	5.95	6.03	6.04
		FB-FLATBETA	5.71	5.72	5.80	5.93	5.95	5.71	5.72	5.81	5.95	5.98
		FB-HIBETA-NI	**5.71**	5.72	5.81	5.94	5.98	**5.71**	5.72	5.82	5.97	6.00
		EB-HIBETA-NI	5.81	5.83	5.85	5.92	5.95	5.81	5.83	5.85	5.93	5.95
		EB-HISIGMAX	5.71	**5.71**	**5.79**	**5.90**	**5.93**	5.71	**5.71**	**5.79**	5.92	5.95
		EB-HIBETASIGMAX	5.81	5.82	5.85	5.91	5.94	5.81	5.82	5.85	**5.92**	**5.95**
0.75,0.1	5,400	RIDG	5.53	5.50	5.51	5.55	5.53	5.53	5.50	5.51	5.55	5.53
		HYB	5.32	5.30	5.35	5.45	5.45	5.32	5.30	5.35	5.45	5.45
		FB-FLATBETA	5.12	**5.14**	**5.20**	**5.27**	5.33	5.12	**5.14**	**5.21**	5.29	5.35
		FB-HIBETA-NI	5.12	5.14	5.20	5.28	5.33	5.12	5.14	5.21	5.29	5.35
		EB-HIBETA-NI	5.21	5.21	5.23	5.29	**5.31**	5.21	5.21	5.23	**5.29**	**5.31**
		EB-HISIGMAX	**5.12**	5.15	5.23	5.31	5.37	**5.12**	5.16	5.25	5.33	5.40
		EB-HIBETASIGMAX	5.21	5.21	5.23	5.29	5.31	5.21	5.21	5.23	5.29	5.32
0,0.4	5,150	RIDG	1.05	1.06	1.06	1.07	1.06	1.05	1.06	1.06	1.07	1.06
		HYB	0.99	1.00	1.02	1.03	1.04	0.99	1.00	1.02	1.03	1.04
		FB-FLATBETA	0.96	0.98	1.00	1.02	1.02	0.96	0.98	1.00	1.02	1.02
		FB-HIBETA-NI	**0.96**	0.98	0.99	1.02	1.02	**0.96**	0.98	1.00	1.02	1.02
		EB-HIBETA-NI	0.96	0.98	0.99	1.01	1.02	0.96	0.98	0.99	1.01	1.02
		EB-HISIGMAX	0.96	0.98	0.99	1.01	**1.02**	0.96	0.98	0.99	1.02	**1.02**
		EB-HIBETASIGMAX	0.96	**0.97**	**0.99**	**1.01**	1.02	0.96	**0.97**	**0.99**	**1.01**	1.02
0.75,0.4	5,150	RIDG	0.93	0.93	0.94	0.93	0.94	0.93	0.93	0.94	0.93	0.94
		HYB	0.89	0.90	0.92	0.93	0.94	0.89	0.90	0.92	0.93	0.94
		FB-FLATBETA	0.87	0.88	0.90	0.91	0.92	0.87	0.88	0.90	0.91	0.92
		FB-HIBETA-NI	0.87	0.88	0.89	0.90	0.91	0.87	0.88	0.89	0.91	0.91
		EB-HIBETA-NI	**0.87**	**0.88**	**0.89**	**0.90**	**0.91**	**0.87**	**0.88**	**0.89**	**0.90**	**0.91**
		EB-HISIGMAX	0.87	0.88	0.90	0.92	0.93	0.87	0.88	0.90	0.92	0.93
		EB-HIBETASIGMAX	0.87	0.88	0.89	0.90	0.91	0.87	0.88	0.89	0.90	0.91
0,0.1	5,150	RIDG	6.20	6.21	6.20	6.23	6.19	6.20	6.21	6.20	6.23	6.19
		HYB	5.99	6.01	6.08	6.12	6.11	5.99	6.01	6.08	6.12	6.11
		FB-FLATBETA	5.80	5.86	5.99	6.05	6.06	5.80	5.86	6.01	6.07	6.07
		FB-HIBETA-NI	**5.80**	5.86	6.00	6.07	6.08	**5.80**	5.86	6.01	6.10	6.10
		EB-HIBETA-NI	5.90	5.94	6.00	6.02	6.05	5.90	5.94	6.00	6.02	6.05
		EB-HISIGMAX	5.80	**5.86**	**5.97**	6.03	**6.04**	5.80	**5.86**	**5.98**	6.05	6.06
		EB-HIBETASIGMAX	5.91	5.94	5.99	**6.01**	6.04	5.91	5.94	6.00	**6.02**	**6.05**
0.75,0.1	5,150	RIDG	5.48	5.50	5.53	5.48	5.50	5.48	5.50	5.53	5.48	5.50
		HYB	5.35	5.36	5.43	5.42	5.46	5.35	5.36	5.43	5.42	5.46
		FB-FLATBETA	5.20	5.25	5.35	5.35	5.40	5.20	5.25	5.36	5.36	5.42
		FB-HIBETA-NI	**5.19**	**5.25**	5.34	5.35	5.40	**5.19**	**5.25**	5.36	5.36	5.41
		EB-HIBETA-NI	5.24	5.27	**5.34**	**5.33**	**5.36**	5.24	5.27	**5.34**	**5.33**	**5.36**
		EB-HISIGMAX	5.20	5.27	5.38	5.39	5.43	5.20	5.27	5.40	5.40	5.45
		EB-HIBETASIGMAX	5.24	5.28	5.34	5.33	5.36	5.24	5.28	5.34	5.34	5.36

Table B2: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 5$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

				MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_B\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	99,400	RIDG	20.3	20.3	20.6	20.5	20.6	20.3	20.3	20.6	20.5	20.6
		HYB	15.6	16.1	17.6	18.3	19.2	15.6	16.1	17.6	18.3	19.2
		FB-FLATBETA	15.7	18.5	29.2	38.1	40.3	15.7	19.3	33.0	41.5	43.7
		FB-HIBETA-NI	**15.1**	16.0	16.9	17.5	18.3	**15.1**	16.2	17.2	17.7	18.5
		EB-HIBETA-NI	15.3	15.8	16.7	**17.4**	**18.3**	15.3	15.9	16.8	17.4	18.3
		EB-HISIGMAX	15.7	17.5	26.1	36.9	40.3	15.7	18.0	29.6	40.4	43.4
		EB-HIBETASIGMAX	15.3	**15.8**	**16.7**	17.4	18.3	15.3	**15.8**	**16.7**	**17.4**	**18.3**
0.75,0.4	99,400	RIDG	99.6	100.6	100.0	99.3	100.1	99.6	100.6	100.0	99.3	100.1
		HYB	92.9	91.4	91.1	93.8	96.0	92.9	91.4	91.1	93.8	96.0
		FB-FLATBETA	97.7	96.9	102.4	112.2	124.5	97.7	98.9	110.1	126.9	144.0
		FB-HIBETA-NI	88.6	89.0	88.9	91.0	93.1	88.6	89.5	90.2	92.9	95.0
		EB-HIBETA-NI	80.8	81.9	**81.6**	**84.4**	**86.5**	80.8	81.9	**81.5**	**84.5**	**86.7**
		EB-HISIGMAX	97.7	104.5	113.6	131.9	151.9	97.7	108.3	125.9	152.1	175.8
		EB-HIBETASIGMAX	**80.6**	**81.8**	82.2	85.7	88.2	**80.6**	**81.8**	82.1	85.7	88.3
0,0.1	99,400	RIDG	86.3	85.4	86.8	87.1	86.0	86.3	85.4	86.8	87.1	86.0
		HYB	84.3	84.2	84.9	86.0	85.1	84.3	84.2	84.9	86.0	85.1
		FB-FLATBETA	94.4	107.9	140.6	194.4	197.1	94.4	112.1	161.4	216.3	216.8
		FB-HIBETA-NI	92.1	99.0	96.8	95.9	94.6	92.1	101.2	99.3	97.6	95.6
		EB-HIBETA-NI	**79.9**	80.0	80.2	80.9	80.6	79.9	80.0	80.2	80.9	80.5
		EB-HISIGMAX	94.4	102.6	131.1	183.1	193.3	94.4	105.6	148.3	204.9	212.9
		EB-HIBETASIGMAX	79.9	**80.0**	**80.2**	**80.8**	**80.5**	**79.9**	**80.0**	**80.2**	**80.8**	**80.5**
0.75,0.1	99,400	RIDG	527.4	539.0	522.0	527.8	542.9	527.4	539.0	522.0	527.8	542.9
		HYB	519.6	523.9	516.8	522.9	528.7	519.6	523.9	516.8	522.9	528.7
		FB-FLATBETA	589.4	576.6	612.0	645.2	682.0	589.4	588.5	655.7	725.2	783.9
		FB-HIBETA-NI	536.3	531.6	538.3	541.5	550.4	536.2	535.4	547.0	553.6	564.2
		EB-HIBETA-NI	**480.2**	**479.7**	**481.3**	**481.4**	**488.1**	**480.2**	**479.6**	**481.0**	**481.0**	**487.5**
		EB-HISIGMAX	589.3	621.6	675.0	742.8	801.3	589.4	643.2	744.2	856.0	933.9
		EB-HIBETASIGMAX	480.2	479.9	482.2	482.6	489.9	480.2	479.8	481.9	482.0	489.2
0,0.4	99,150	RIDG	20.4	20.2	20.4	20.4	20.5	20.4	20.2	20.4	20.4	20.5
		HYB	18.5	18.7	19.4	19.9	20.2	18.5	18.7	19.4	19.9	20.2
		FB-FLATBETA	24.3	31.4	57.5	57.8	58.8	24.3	33.8	61.0	61.1	62.0
		FB-HIBETA-NI	17.8	18.0	18.4	18.8	19.2	17.8	18.1	18.5	18.9	19.2
		EB-HIBETA-NI	16.8	17.2	**18.0**	18.6	18.9	**16.8**	17.2	**18.0**	18.7	18.9
		EB-HISIGMAX	24.3	29.5	54.6	55.7	58.1	24.3	31.4	57.9	58.9	61.0
		EB-HIBETASIGMAX	**16.8**	**17.2**	18.0	**18.6**	**18.9**	16.8	**17.2**	18.0	**18.6**	**18.9**
0.75,0.4	99,150	RIDG	101.4	100.0	101.3	99.2	98.8	101.4	100.0	101.3	99.2	98.8
		HYB	100.2	99.9	101.2	100.1	102.1	100.2	99.9	101.2	100.1	102.1
		FB-FLATBETA	155.2	153.0	187.0	198.0	229.7	155.1	162.1	207.3	221.8	252.2
		FB-HIBETA-NI	94.8	93.1	93.3	93.6	94.8	94.8	93.4	94.1	94.7	96.1
		EB-HIBETA-NI	85.1	**85.3**	**86.4**	**87.3**	**89.3**	85.1	**85.3**	**86.4**	**87.5**	**89.6**
		EB-HISIGMAX	155.1	178.8	233.8	245.8	269.2	155.1	193.1	256.8	269.2	291.8
		EB-HIBETASIGMAX	**85.1**	85.4	86.9	88.4	90.4	**85.1**	85.4	86.8	88.5	90.6
0,0.1	99,150	RIDG	86.3	87.2	85.7	85.5	86.6	86.3	87.2	85.7	85.5	86.6
		HYB	86.5	87.7	86.7	86.1	87.0	86.5	87.7	86.7	86.1	87.0
		FB-FLATBETA	146.1	185.5	300.0	302.9	295.5	146.1	199.0	322.1	321.2	312.1
		FB-HIBETA-NI	106.2	104.1	98.3	96.1	95.0	106.2	105.1	99.0	96.4	95.3
		EB-HIBETA-NI	**80.8**	81.4	80.9	81.1	81.2	**80.8**	81.4	80.8	81.0	81.1
		EB-HISIGMAX	146.1	174.6	285.5	282.9	282.3	146.1	185.5	307.8	302.2	298.7
		EB-HIBETASIGMAX	80.8	**81.3**	**80.8**	**81.0**	**81.2**	80.8	**81.3**	**80.8**	**80.9**	**81.1**
0.75,0.1	99,150	RIDG	519.8	531.2	527.5	518.1	520.2	519.8	531.2	527.5	518.1	520.2
		HYB	520.6	534.4	534.3	532.9	533.0	520.6	534.4	534.3	532.9	533.0
		FB-FLATBETA	912.4	914.6	1104.0	1117.3	1165.1	912.4	969.5	1214.5	1246.9	1300.9
		FB-HIBETA-NI	566.6	558.8	561.1	559.2	565.2	566.5	560.7	566.8	565.2	574.1
		EB-HIBETA-NI	**487.3**	**490.6**	**495.0**	**493.3**	**498.2**	**487.3**	**490.5**	**494.7**	**492.7**	**497.6**
		EB-HISIGMAX	913.1	1092.6	1335.5	1355.8	1426.1	912.8	1176.9	1461.3	1488.1	1562.7
		EB-HIBETASIGMAX	487.4	491.0	495.8	494.1	499.1	487.4	490.9	495.4	493.4	498.5

Table B3: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 99$ and $\varepsilon + 1 \sim G\{1, 1\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

			MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_B\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	5,400	RIDG	1.05	1.06	1.06	1.06	1.06	1.05	1.06	1.06	1.06	1.06
		HYB	0.98	0.99	0.99	1.01	1.01	0.98	0.99	0.99	1.01	1.01
		FB-FLATBETA	0.95	0.96	0.98	0.99	1.00	0.95	0.96	0.98	0.99	1.00
		FB-HIBETA-NI	**0.95**	0.96	0.98	0.99	1.00	**0.95**	0.96	0.98	0.99	1.00
		EB-HIBETA-NI	0.95	0.95	0.97	0.99	0.99	0.95	0.95	0.97	0.99	0.99
		EB-HISIGMAX	0.95	0.95	0.97	0.99	0.99	0.95	0.96	0.98	0.99	1.00
		EB-HIBETASIGMAX	0.95	**0.95**	**0.97**	**0.99**	**0.99**	0.95	**0.95**	**0.97**	**0.99**	**0.99**
0.75,0.4	5,400	RIDG	0.94	0.93	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.94
		HYB	0.87	0.89	0.91	0.92	0.93	0.87	0.89	0.91	0.92	0.93
		FB-FLATBETA	0.85	0.86	0.88	0.89	0.90	0.85	0.86	0.88	0.90	0.90
		FB-HIBETA-NI	**0.85**	0.86	0.87	0.89	0.90	**0.85**	0.86	0.88	0.89	0.90
		EB-HIBETA-NI	0.85	**0.86**	**0.87**	**0.89**	**0.89**	0.85	**0.86**	**0.87**	**0.89**	**0.89**
		EB-HISIGMAX	0.85	0.86	0.88	0.90	0.91	0.85	0.86	0.88	0.90	0.91
		EB-HIBETASIGMAX	0.85	0.86	0.87	0.89	0.89	0.85	0.86	0.87	0.89	0.89
0,0.1	5,400	RIDG	6.13	6.11	6.17	6.23	6.16	6.13	6.11	6.17	6.23	6.16
		HYB	5.87	5.87	5.96	6.05	6.02	5.87	5.87	5.96	6.05	6.02
		FB-FLATBETA	5.68	5.68	5.81	5.93	5.95	5.68	5.68	5.82	5.95	5.97
		FB-HIBETA-NI	**5.68**	5.68	5.81	5.94	5.97	**5.68**	5.69	5.82	5.97	6.00
		EB-HIBETA-NI	5.77	5.79	5.85	5.94	5.90	5.77	5.79	5.86	5.94	5.91
		EB-HISIGMAX	5.68	**5.68**	**5.78**	**5.90**	5.92	5.68	**5.68**	**5.79**	**5.92**	5.94
		EB-HIBETASIGMAX	5.76	5.78	5.85	5.93	**5.90**	5.76	5.78	5.85	5.93	**5.91**
0.75,0.1	5,400	RIDG	5.49	5.50	5.52	5.47	5.53	5.49	5.50	5.52	5.47	5.53
		HYB	5.28	5.31	5.38	5.37	5.46	5.28	5.31	5.38	5.37	5.46
		FB-FLATBETA	5.07	5.13	**5.19**	5.24	5.33	5.07	5.13	**5.19**	5.26	5.35
		FB-HIBETA-NI	**5.07**	**5.13**	5.19	**5.24**	5.34	**5.07**	**5.13**	5.20	5.26	5.35
		EB-HIBETA-NI	5.16	5.20	5.22	5.25	**5.29**	5.16	5.20	5.22	**5.25**	**5.29**
		EB-HISIGMAX	5.07	5.14	5.22	5.27	5.37	5.07	5.14	5.23	5.29	5.39
		EB-HIBETASIGMAX	5.16	5.20	5.23	5.25	5.30	5.16	5.20	5.23	5.25	5.30
0,0.4	5,150	RIDG	1.07	1.06	1.07	1.07	1.06	1.07	1.06	1.07	1.07	1.06
		HYB	1.00	1.00	1.02	1.03	1.04	1.00	1.00	1.02	1.03	1.04
		FB-FLATBETA	0.97	0.98	1.00	1.02	1.03	0.97	0.98	1.00	1.02	1.03
		FB-HIBETA-NI	**0.97**	0.98	1.00	1.02	1.03	**0.97**	0.98	1.00	1.02	1.03
		EB-HIBETA-NI	0.98	0.97	0.99	1.01	1.03	0.98	0.97	1.00	1.02	1.03
		EB-HISIGMAX	0.97	0.97	1.00	**1.01**	1.03	0.97	0.97	1.00	1.02	1.03
		EB-HIBETASIGMAX	0.98	**0.97**	**0.99**	1.01	**1.03**	0.98	**0.97**	**0.99**	**1.01**	**1.03**
0.75,0.4	5,150	RIDG	0.94	0.93	0.95	0.93	0.93	0.94	0.93	0.95	0.93	0.93
		HYB	0.89	0.90	0.93	0.92	0.93	0.89	0.90	0.93	0.92	0.93
		FB-FLATBETA	0.87	0.88	0.90	0.91	0.92	0.87	0.88	0.91	0.91	0.92
		FB-HIBETA-NI	0.87	0.88	0.90	0.90	0.91	0.87	0.88	0.90	0.90	0.91
		EB-HIBETA-NI	0.87	**0.87**	**0.89**	**0.89**	**0.90**	0.87	**0.87**	**0.90**	**0.89**	**0.90**
		EB-HISIGMAX	0.87	0.88	0.91	0.91	0.92	0.87	0.88	0.91	0.92	0.92
		EB-HIBETASIGMAX	**0.87**	0.87	0.90	0.90	0.90	**0.87**	0.87	0.90	0.90	0.90
0,0.1	5,150	RIDG	6.15	6.15	6.17	6.20	6.15	6.15	6.15	6.17	6.20	6.15
		HYB	5.95	5.96	6.03	6.11	6.10	5.95	5.96	6.03	6.11	6.10
		FB-FLATBETA	5.79	5.81	5.94	6.06	6.13	5.79	5.81	5.95	6.08	6.14
		FB-HIBETA-NI	**5.79**	5.81	5.95	6.08	6.15	**5.79**	5.81	5.97	6.10	6.17
		EB-HIBETA-NI	5.87	5.87	5.97	6.02	**6.03**	5.87	5.87	5.97	6.02	**6.03**
		EB-HISIGMAX	5.79	**5.80**	**5.92**	6.04	6.11	5.79	**5.81**	**5.93**	6.06	6.13
		EB-HIBETASIGMAX	5.87	5.87	5.96	**6.02**	6.03	5.87	5.87	5.96	**6.02**	6.03
0.75,0.1	5,150	RIDG	5.50	5.50	5.51	5.42	5.59	5.50	5.50	5.51	5.42	5.59
		HYB	5.35	5.38	5.42	5.38	5.55	5.35	5.38	5.42	5.38	5.55
		FB-FLATBETA	5.19	**5.27**	5.34	5.34	5.52	5.19	**5.27**	5.35	5.36	5.54
		FB-HIBETA-NI	**5.19**	5.27	5.33	5.34	5.51	**5.18**	5.27	5.35	5.35	5.53
		EB-HIBETA-NI	5.24	5.29	5.33	**5.28**	**5.44**	5.24	5.29	5.33	**5.29**	**5.44**
		EB-HISIGMAX	5.19	5.29	5.37	5.37	5.56	5.19	5.30	5.39	5.39	5.58
		EB-HIBETASIGMAX	5.24	5.29	**5.32**	5.29	5.45	5.24	5.30	**5.32**	5.29	5.45

Table B4: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 5$ and $\varepsilon + 1 \sim G\{1, 1\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

				MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_{\text{B}}\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	99,400	RIDG	20.4	20.7	20.7	20.7	20.6	20.4	20.7	20.7	20.7	20.6
		HYB	16.1	16.3	16.8	17.5	18.1	16.1	16.3	16.8	17.5	18.1
		FB-FLATBETA	17.8	18.9	22.0	28.5	35.8	18.3	19.9	24.5	32.4	39.6
		FB-HIBETA-NI	15.9	16.1	16.4	16.7	17.3	16.1	16.3	16.6	17.0	17.5
		EB-HIBETA-NI	15.6	15.6	16.1	16.7	17.3	15.6	15.6	16.2	16.8	17.4
		EB-HISIGMAX	17.1	17.9	20.4	25.7	33.0	17.4	18.6	22.3	29.1	37.0
		EB-HIBETASIGMAX	**15.6**	**15.6**	**16.1**	**16.7**	**17.3**	**15.6**	**15.6**	**16.1**	**16.7**	**17.3**
0.75,0.4	99,400	RIDG	101.3	104.0	100.5	100.7	102.6	101.3	104.0	100.5	100.7	102.6
		HYB	93.4	93.6	91.9	91.8	95.0	93.4	93.6	91.9	91.8	95.0
		FB-FLATBETA	99.1	98.8	100.0	102.6	107.7	100.8	101.1	104.1	110.0	119.9
		FB-HIBETA-NI	89.1	89.2	89.2	89.4	90.4	89.5	89.7	90.1	90.6	92.1
		EB-HIBETA-NI	82.4	82.8	**81.9**	**82.1**	**83.1**	82.4	82.8	**81.9**	**82.0**	**83.1**
		EB-HISIGMAX	108.4	108.6	110.0	113.9	122.5	111.9	113.3	117.5	125.9	140.6
		EB-HIBETASIGMAX	**81.8**	**82.2**	81.9	82.5	83.9	**81.8**	**82.2**	81.9	82.4	83.9
0,0.1	99,400	RIDG	86.4	87.0	85.4	86.7	87.5	86.4	87.0	85.4	86.7	87.5
		HYB	85.0	85.1	84.0	85.4	86.2	85.0	85.1	84.0	85.4	86.2
		FB-FLATBETA	104.0	109.6	121.4	147.6	179.9	106.6	114.4	133.4	169.3	202.6
		FB-HIBETA-NI	97.7	99.2	98.2	97.9	96.3	99.2	101.4	101.0	100.5	98.3
		EB-HIBETA-NI	80.2	80.1	79.9	80.4	80.6	80.2	80.1	79.9	80.4	80.6
		EB-HISIGMAX	100.0	104.5	114.7	133.5	167.8	101.9	108.0	124.0	152.4	190.4
		EB-HIBETASIGMAX	**80.2**	**80.1**	**79.9**	**80.4**	**80.6**	**80.2**	**80.1**	**79.9**	**80.4**	**80.6**
0.75,0.1	99,400	RIDG	520.9	537.2	534.6	526.4	523.4	520.9	537.2	534.6	526.4	523.4
		HYB	517.8	524.1	526.9	521.1	519.6	517.8	524.1	526.9	521.1	519.6
		FB-FLATBETA	586.6	582.2	594.0	604.7	623.1	596.5	595.6	619.0	647.0	686.3
		FB-HIBETA-NI	529.0	528.7	534.6	533.9	539.6	531.5	532.2	540.7	542.0	550.0
		EB-HIBETA-NI	480.1	480.5	**482.8**	**480.0**	**483.2**	480.1	480.5	482.7	**479.8**	**482.9**
		EB-HISIGMAX	643.5	638.7	652.8	663.9	691.3	664.2	665.0	697.1	731.3	786.0
		EB-HIBETASIGMAX	**479.8**	**480.3**	482.9	480.7	483.8	**479.7**	**480.2**	**482.7**	480.4	483.4
0,0.4	99,150	RIDG	20.3	20.1	20.6	20.4	20.4	20.3	20.1	20.6	20.4	20.4
		HYB	18.8	18.7	19.3	19.6	19.8	18.8	18.7	19.3	19.6	19.8
		FB-FLATBETA	30.0	33.5	53.8	56.1	59.6	31.7	36.3	57.7	59.6	63.0
		FB-HIBETA-NI	18.0	18.0	18.2	18.4	18.8	18.1	18.1	18.3	18.5	18.8
		EB-HIBETA-NI	17.2	17.3	17.8	18.0	18.5	17.2	17.3	17.8	18.1	18.6
		EB-HISIGMAX	28.9	31.6	48.7	56.9	57.8	30.2	33.9	52.5	60.1	61.0
		EB-HIBETASIGMAX	**17.2**	**17.3**	**17.7**	**18.0**	**18.5**	**17.2**	**17.3**	**17.7**	**18.1**	**18.5**
0.75,0.4	99,150	RIDG	102.4	100.1	100.6	102.7	101.1	102.4	100.1	100.6	102.7	101.1
		HYB	101.6	100.3	99.9	101.9	101.8	101.6	100.3	99.9	101.9	101.8
		FB-FLATBETA	164.3	165.7	172.1	183.9	205.1	172.5	176.3	187.5	203.4	226.3
		FB-HIBETA-NI	93.3	93.4	93.0	93.0	93.3	93.4	93.7	93.5	93.8	94.4
		EB-HIBETA-NI	85.1	**85.5**	**85.8**	**86.3**	**87.1**	85.1	**85.5**	**85.7**	**86.3**	**87.3**
		EB-HISIGMAX	198.8	209.2	223.6	235.9	239.3	212.8	225.9	244.5	258.0	261.8
		EB-HIBETASIGMAX	**85.0**	85.6	86.0	86.7	87.9	**85.0**	85.5	85.9	86.7	88.0
0,0.1	99,150	RIDG	86.1	85.3	85.8	86.6	87.3	86.1	85.3	85.8	86.6	87.3
		HYB	86.8	85.8	86.4	87.2	87.9	86.8	85.8	86.4	87.2	87.9
		FB-FLATBETA	180.3	185.6	282.1	304.6	294.3	189.8	200.1	306.1	324.4	312.8
		FB-HIBETA-NI	105.1	103.5	100.6	98.0	96.6	106.0	104.5	101.5	98.7	97.0
		EB-HIBETA-NI	80.9	81.1	**80.9**	81.1	81.3	80.9	81.1	**80.9**	81.1	81.3
		EB-HISIGMAX	172.9	175.8	261.2	294.7	291.0	181.1	188.1	283.0	314.1	307.4
		EB-HIBETASIGMAX	**80.8**	**81.1**	80.9	**81.1**	**81.3**	**80.8**	**81.1**	80.9	**81.0**	**81.2**
0.75,0.1	99,150	RIDG	523.5	525.3	533.3	529.5	521.3	523.5	525.3	533.3	529.5	521.3
		HYB	529.0	529.2	535.9	538.7	532.2	529.0	529.2	535.9	538.7	532.2
		FB-FLATBETA	951.9	966.9	1027.3	1082.1	1116.9	998.9	1028.5	1123.7	1198.5	1250.9
		FB-HIBETA-NI	558.4	558.6	559.3	562.0	561.9	559.5	560.4	562.9	567.6	569.7
		EB-HIBETA-NI	490.2	**489.6**	**491.2**	**492.0**	**494.0**	490.1	**489.5**	**491.0**	**491.6**	**493.7**
		EB-HISIGMAX	1166.4	1217.1	1339.2	1344.3	1367.2	1248.9	1316.0	1457.2	1473.0	1507.7
		EB-HIBETASIGMAX	**490.1**	489.8	491.7	492.7	494.9	**490.0**	489.7	491.4	492.2	494.6

Table B5: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 99$ and $\boldsymbol{W}|\boldsymbol{X} \sim N_p\{\psi\mathbf{1}_p + \nu\boldsymbol{X}^2, \tau^2\boldsymbol{I}_p\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

			MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_{\text{B}}\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	5,400	RIDG	1.06	1.07	1.06	1.06	1.06	1.06	1.07	1.06	1.06	1.06
		HYB	0.99	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00
		FB-FLATBETA	0.96	0.96	0.97	0.97	0.98	0.96	0.96	0.97	0.98	0.99
		FB-HIBETA-NI	0.96	0.96	0.97	0.98	0.98	0.96	0.96	0.97	0.98	0.99
		EB-HIBETA-NI	**0.96**	0.96	0.97	0.97	0.98	0.96	0.96	0.97	0.97	0.98
		EB-HISIGMAX	0.96	**0.96**	**0.97**	0.97	0.98	0.96	**0.96**	**0.97**	0.97	0.98
		EB-HIBETASIGMAX	0.96	0.96	0.97	**0.97**	**0.98**	**0.96**	0.96	0.97	**0.97**	**0.98**
0.75,0.4	5,400	RIDG	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
		HYB	0.89	0.89	0.90	0.90	0.91	0.89	0.89	0.90	0.90	0.91
		FB-FLATBETA	0.87	0.87	0.87	0.87	0.88	0.87	0.87	0.87	0.88	0.88
		FB-HIBETA-NI	0.87	0.87	0.87	0.87	0.88	0.87	0.87	0.87	0.87	0.88
		EB-HIBETA-NI	0.87	0.87	0.87	0.87	0.88	0.87	0.87	0.87	**0.87**	**0.87**
		EB-HISIGMAX	0.87	0.87	0.87	0.88	0.88	0.87	0.87	0.87	0.88	0.89
		EB-HIBETASIGMAX	**0.87**	**0.87**	**0.87**	**0.87**	**0.88**	**0.87**	**0.87**	**0.87**	0.87	0.87
0,0.1	5,400	RIDG	6.17	6.19	6.20	6.18	6.22	6.17	6.19	6.20	6.18	6.22
		HYB	5.92	5.92	5.96	5.96	6.01	5.92	5.92	5.96	5.96	6.01
		FB-FLATBETA	5.73	5.72	5.77	5.81	5.88	5.73	5.72	5.77	5.82	5.89
		FB-HIBETA-NI	5.73	5.72	5.77	5.82	5.89	5.73	5.72	5.78	5.83	5.91
		EB-HIBETA-NI	5.81	5.81	5.85	5.86	5.91	5.81	5.81	5.85	5.86	5.91
		EB-HISIGMAX	**5.73**	**5.72**	**5.76**	**5.79**	**5.85**	**5.73**	**5.72**	**5.76**	**5.80**	**5.87**
		EB-HIBETASIGMAX	5.81	5.81	5.85	5.86	5.90	5.81	5.81	5.85	5.86	5.90
0.75,0.1	5,400	RIDG	5.51	5.51	5.49	5.53	5.52	5.51	5.51	5.49	5.53	5.52
		HYB	5.33	5.34	5.31	5.37	5.40	5.33	5.34	5.31	5.37	5.40
		FB-FLATBETA	5.14	5.17	**5.14**	**5.20**	**5.24**	5.14	5.17	**5.15**	**5.21**	**5.25**
		FB-HIBETA-NI	**5.14**	**5.17**	5.14	5.20	5.24	**5.14**	**5.17**	5.15	5.21	5.25
		EB-HIBETA-NI	5.22	5.22	5.21	5.23	5.26	5.22	5.22	5.21	5.23	5.26
		EB-HISIGMAX	5.15	5.18	5.16	5.23	5.27	5.16	5.19	5.16	5.24	5.29
		EB-HIBETASIGMAX	5.21	5.22	5.21	5.24	5.27	5.21	5.22	5.21	5.24	5.27
0,0.4	5,150	RIDG	1.06	1.06	1.08	1.06	1.07	1.06	1.06	1.08	1.06	1.07
		HYB	1.00	1.00	1.02	1.02	1.03	1.00	1.00	1.02	1.02	1.03
		FB-FLATBETA	0.98	0.98	0.99	1.00	1.01	0.98	0.98	0.99	1.00	1.02
		FB-HIBETA-NI	0.98	0.98	0.99	1.00	1.01	0.98	0.98	0.99	1.00	1.02
		EB-HIBETA-NI	**0.98**	**0.98**	0.99	0.99	1.01	**0.98**	**0.98**	0.99	0.99	1.01
		EB-HISIGMAX	0.98	0.98	**0.99**	0.99	1.01	0.98	0.98	**0.99**	1.00	1.01
		EB-HIBETASIGMAX	0.98	0.98	0.99	**0.99**	**1.01**	0.98	0.98	0.99	**0.99**	**1.01**
0.75,0.4	5,150	RIDG	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.93	0.93	0.94
		HYB	0.91	0.91	0.91	0.92	0.93	0.91	0.91	0.91	0.92	0.93
		FB-FLATBETA	0.89	0.89	0.89	0.90	0.90	0.89	0.89	0.89	0.90	0.91
		FB-HIBETA-NI	0.89	0.88	0.89	0.89	0.90	0.89	0.88	0.89	0.90	0.90
		EB-HIBETA-NI	**0.88**	**0.88**	**0.88**	**0.89**	**0.89**	**0.88**	**0.88**	**0.88**	**0.89**	**0.89**
		EB-HISIGMAX	0.89	0.89	0.90	0.90	0.91	0.89	0.89	0.90	0.90	0.91
		EB-HIBETASIGMAX	0.88	0.88	0.89	0.89	0.89	0.88	0.88	0.89	0.89	0.89
0,0.1	5,150	RIDG	6.18	6.17	6.14	6.18	6.16	6.18	6.17	6.14	6.18	6.16
		HYB	5.98	5.99	5.97	6.04	6.04	5.98	5.99	5.97	6.04	6.04
		FB-FLATBETA	5.84	5.86	5.87	5.94	5.98	5.84	5.86	5.88	5.96	6.00
		FB-HIBETA-NI	5.84	5.86	5.87	5.95	5.99	5.84	5.86	5.88	5.97	6.01
		EB-HIBETA-NI	5.90	5.90	5.91	5.93	5.95	5.90	5.90	5.91	5.93	5.96
		EB-HISIGMAX	**5.83**	**5.85**	**5.86**	**5.92**	5.96	**5.83**	**5.85**	**5.86**	5.93	5.97
		EB-HIBETASIGMAX	5.90	5.91	5.91	5.93	**5.95**	5.90	5.91	5.91	**5.93**	**5.95**
0.75,0.1	5,150	RIDG	5.56	5.52	5.51	5.49	5.51	5.56	5.52	5.51	5.49	5.51
		HYB	5.43	5.39	5.40	5.39	5.43	5.43	5.39	5.40	5.39	5.43
		FB-FLATBETA	5.27	5.27	**5.30**	5.33	5.35	5.27	5.27	**5.31**	5.34	5.37
		FB-HIBETA-NI	**5.26**	**5.27**	5.30	5.32	5.35	**5.27**	**5.27**	5.31	5.34	5.37
		EB-HIBETA-NI	5.31	5.28	5.31	**5.31**	**5.35**	5.31	5.28	5.31	**5.31**	**5.35**
		EB-HISIGMAX	5.29	5.29	5.33	5.36	5.38	5.29	5.29	5.34	5.38	5.40
		EB-HIBETASIGMAX	5.30	5.28	5.31	5.31	5.35	5.30	5.28	5.31	5.31	5.35

Table B6: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 5$ and $\boldsymbol{W}|\boldsymbol{X} \sim N_p\{\psi\mathbf{1}_p + \nu\boldsymbol{X}^2, \tau^2\boldsymbol{I}_p\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

				MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_{\text{B}}\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	99,400	RIDG	21.1	20.8	20.7	20.9	21.0	21.1	20.8	20.7	20.9	21.0
		HYB	15.8	18.1	29.3	38.7	44.7	15.8	18.1	29.3	38.7	44.7
		FB-FLATBETA	15.5	**15.4**	**16.3**	**17.1**	**17.7**	15.5	**15.5**	**16.3**	**17.2**	**17.8**
		FB-HIBETA-NI	**14.9**	15.5	16.5	17.3	17.9	**14.9**	15.5	16.5	17.3	17.8
		EB-HIBETA-NI	15.7	16.6	17.8	18.8	19.5	15.7	16.6	17.8	18.8	19.5
		EB-HISIGMAX	15.5	17.5	22.9	32.5	35.9	15.5	18.0	25.8	36.6	39.8
		EB-HIBETASIGMAX	15.7	16.4	17.1	17.8	18.6	15.7	16.4	17.1	17.8	18.6
0.75,0.4	99,400	RIDG	104.0	102.7	102.6	104.3	101.4	104.0	102.7	102.6	104.3	101.4
		HYB	94.9	94.0	99.6	110.9	121.6	94.9	94.0	99.6	110.9	121.6
		FB-FLATBETA	97.9	88.8	87.1	87.2	86.6	97.9	89.2	87.3	87.2	86.8
		FB-HIBETA-NI	**80.4**	82.8	85.1	86.2	86.2	**80.4**	82.8	84.9	85.8	**85.6**
		EB-HIBETA-NI	81.8	83.6	85.6	87.9	88.1	81.8	83.5	85.5	87.7	87.8
		EB-HISIGMAX	97.9	103.6	110.8	123.7	135.0	97.9	107.2	122.4	144.0	159.9
		EB-HIBETASIGMAX	81.8	**82.0**	**82.4**	**84.5**	**86.1**	81.8	**82.0**	**82.4**	**84.6**	86.3
0,0.1	99,400	RIDG	87.0	88.1	85.5	86.8	87.4	87.0	88.1	85.5	86.8	87.4
		HYB	84.5	88.4	103.0	115.1	129.2	84.5	88.4	103.0	115.1	129.2
		FB-FLATBETA	93.7	88.8	86.3	86.1	85.3	93.7	89.4	87.6	87.7	87.2
		FB-HIBETA-NI	**80.2**	**80.7**	80.7	81.6	81.2	**80.2**	**80.7**	80.8	81.7	81.4
		EB-HIBETA-NI	80.4	80.9	**80.5**	**81.1**	**80.4**	80.4	80.9	**80.5**	**81.1**	**80.5**
		EB-HISIGMAX	93.7	101.3	124.7	159.3	174.6	93.7	103.9	139.4	182.4	197.1
		EB-HIBETASIGMAX	80.4	80.9	80.6	81.1	80.5	80.4	80.9	80.6	81.1	80.5
0.75,0.1	99,400	RIDG	535.6	540.8	552.8	539.8	536.6	535.6	540.8	552.8	539.8	536.6
		HYB	527.6	528.0	577.8	582.6	608.9	527.6	528.0	577.8	582.6	608.9
		FB-FLATBETA	589.9	521.7	503.0	498.0	495.8	589.9	524.7	507.2	503.8	504.1
		FB-HIBETA-NI	**481.0**	**479.6**	**482.6**	**481.6**	**480.4**	**481.0**	**479.6**	**482.8**	**482.0**	**481.0**
		EB-HIBETA-NI	486.4	488.4	490.6	491.5	490.1	486.4	488.4	490.7	491.5	490.3
		EB-HISIGMAX	590.1	615.2	648.5	707.3	756.8	590.0	636.3	713.0	819.9	895.0
		EB-HIBETASIGMAX	486.4	487.6	489.1	490.8	490.9	486.4	487.5	488.9	490.5	490.4
0,0.4	99,150	RIDG	21.1	20.9	21.0	21.0	20.7	21.1	20.9	21.0	21.0	20.7
		HYB	19.0	21.0	27.8	36.8	38.5	19.0	21.0	27.8	36.8	38.5
		FB-FLATBETA	24.6	23.7	22.5	23.0	22.1	24.6	24.0	23.3	24.1	23.3
		FB-HIBETA-NI	**16.7**	**17.2**	**17.8**	**18.5**	**18.8**	**16.7**	**17.2**	**17.8**	**18.4**	**18.7**
		EB-HIBETA-NI	17.4	17.9	18.8	19.6	19.9	17.4	17.9	18.8	19.6	20.0
		EB-HISIGMAX	24.6	29.6	52.9	54.8	54.9	24.6	31.4	56.4	58.2	57.9
		EB-HIBETASIGMAX	17.4	17.8	18.5	19.1	19.3	17.4	17.8	18.5	19.1	19.3
0.75,0.4	99,150	RIDG	102.9	102.7	102.5	103.8	102.0	102.9	102.7	102.5	103.8	102.0
		HYB	101.4	107.0	120.5	136.0	140.2	101.4	107.0	120.5	136.0	140.2
		FB-FLATBETA	153.0	125.0	104.2	99.0	97.9	152.9	126.6	107.4	103.4	102.8
		FB-HIBETA-NI	**84.3**	**84.8**	**86.6**	**87.7**	**90.0**	**84.3**	**84.8**	**86.6**	**87.6**	**89.9**
		EB-HIBETA-NI	85.0	85.5	87.1	88.8	92.0	85.0	85.6	87.1	88.8	92.0
		EB-HISIGMAX	153.0	182.5	218.4	234.3	246.8	153.0	196.8	241.2	258.7	270.5
		EB-HIBETASIGMAX	85.0	85.6	86.8	88.2	90.3	85.0	85.5	86.8	88.4	90.6
0,0.1	99,150	RIDG	86.8	87.5	87.0	86.9	86.2	86.8	87.5	87.0	86.9	86.2
		HYB	86.7	90.5	108.0	114.0	122.8	86.7	90.5	108.0	114.0	122.8
		FB-FLATBETA	146.6	136.1	120.3	109.7	105.4	146.6	137.7	124.6	115.3	112.2
		FB-HIBETA-NI	82.5	82.2	82.1	82.9	82.7	82.5	82.2	82.1	82.8	82.6
		EB-HIBETA-NI	**81.5**	**81.3**	**81.0**	**81.6**	**81.3**	**81.5**	**81.3**	**81.0**	**81.6**	**81.3**
		EB-HISIGMAX	146.6	169.4	277.6	283.1	273.5	146.6	179.3	299.3	302.1	290.6
		EB-HIBETASIGMAX	81.5	81.3	81.2	81.7	81.5	81.5	81.3	81.1	81.7	81.5
0.75,0.1	99,150	RIDG	537.1	546.4	540.0	530.8	535.6	537.1	546.4	540.0	530.8	535.6
		HYB	538.2	558.5	596.7	645.9	646.2	538.2	558.5	596.7	645.9	646.2
		FB-FLATBETA	928.7	731.5	608.1	576.6	563.1	928.4	740.8	627.6	603.5	595.2
		FB-HIBETA-NI	**491.8**	**487.8**	**488.8**	**489.1**	**490.7**	**491.8**	**487.7**	**488.6**	**489.2**	**491.0**
		EB-HIBETA-NI	494.7	493.1	496.7	498.9	499.7	494.7	493.1	496.7	499.0	499.8
		EB-HISIGMAX	928.2	1066.8	1215.5	1314.1	1287.6	928.1	1150.9	1350.6	1468.6	1443.4
		EB-HIBETASIGMAX	494.8	493.3	497.5	500.0	500.6	494.7	493.2	497.3	499.8	500.5

Table B7: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 99$ and $\boldsymbol{X}|Z \sim N_p\{1_{[Z=2]}(3 \times \mathbf{1}_p) - 1_{[Z=3]}(3 \times \mathbf{1}_p), \boldsymbol{\Sigma}_{\boldsymbol{X}}\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

			MSPE($\hat{\boldsymbol{\beta}}^{\text{ppm}}$)					MSPE($\hat{\boldsymbol{\beta}}^{\text{pm}}$)				
$\{\rho, R^2\}$	$\{p, n_{\text{B}}\}$	Method	$\tau = 0.01$	0.5	1.0	1.50	2.00	0.01	0.5	1.0	1.50	2.00
0,0.4	5,400	RIDG	1.06	1.06	1.08	1.08	1.07	1.06	1.06	1.08	1.08	1.07
		HYB	0.98	0.99	1.01	1.02	1.02	0.98	0.99	1.01	1.02	1.02
		FB-FLATBETA	0.95	**0.95**	0.98	1.00	1.00	**0.95**	**0.95**	0.97	0.99	1.00
		FB-HIBETA-NI	0.95	0.96	0.99	1.01	1.02	0.95	0.96	0.99	1.01	1.02
		EB-HIBETA-NI	0.95	0.96	0.98	1.02	1.02	0.95	0.96	0.98	1.01	1.02
		EB-HISIGMAX	**0.95**	0.96	0.97	**0.98**	**0.99**	0.95	0.96	0.97	**0.98**	0.99
		EB-HIBETASIGMAX	0.95	0.95	**0.97**	0.99	0.99	0.95	0.95	**0.97**	0.99	**0.99**
0.75,0.4	5,400	RIDG	0.93	0.93	0.94	0.94	0.94	0.93	0.93	0.94	0.94	0.94
		HYB	0.88	0.88	0.90	0.91	0.92	0.88	0.88	0.90	0.91	0.92
		FB-FLATBETA	0.85	0.86	0.88	0.90	0.91	0.85	0.86	0.88	0.90	0.90
		FB-HIBETA-NI	0.85	0.86	0.90	0.92	0.94	0.85	0.86	0.89	0.92	0.93
		EB-HIBETA-NI	0.85	0.86	0.89	0.92	0.94	0.85	0.86	0.89	0.92	0.93
		EB-HISIGMAX	0.85	0.86	0.88	0.89	0.90	0.85	0.86	0.88	0.89	0.91
		EB-HIBETASIGMAX	**0.85**	**0.86**	**0.87**	**0.88**	**0.89**	**0.85**	**0.86**	**0.87**	**0.88**	**0.89**
0,0.1	5,400	RIDG	6.27	6.28	6.26	6.29	6.26	6.27	6.28	6.26	6.29	6.26
		HYB	5.90	5.92	5.95	6.02	6.04	5.90	5.92	5.95	6.02	6.04
		FB-FLATBETA	5.68	5.70	5.73	5.78	5.82	5.68	5.70	5.73	5.78	5.82
		FB-HIBETA-NI	**5.68**	**5.70**	**5.73**	**5.78**	**5.82**	**5.68**	**5.70**	**5.73**	**5.77**	**5.81**
		EB-HIBETA-NI	5.86	5.87	5.94	6.01	6.05	5.86	5.87	5.94	6.01	6.04
		EB-HISIGMAX	5.68	5.70	5.75	5.82	5.85	5.68	5.70	5.75	5.83	5.86
		EB-HIBETASIGMAX	5.85	5.87	5.93	5.99	6.02	5.85	5.87	5.93	6.00	6.02
0.75,0.1	5,400	RIDG	5.55	5.57	5.57	5.57	5.51	5.55	5.57	5.57	5.57	5.51
		HYB	5.31	5.33	5.36	5.42	5.40	5.31	5.33	5.36	5.42	5.40
		FB-FLATBETA	5.12	5.12	5.15	5.19	5.18	5.12	5.12	5.15	5.18	5.17
		FB-HIBETA-NI	**5.11**	**5.12**	**5.15**	**5.19**	**5.17**	**5.11**	**5.12**	**5.15**	**5.18**	**5.17**
		EB-HIBETA-NI	5.25	5.24	5.31	5.36	5.36	5.25	5.24	5.31	5.35	5.36
		EB-HISIGMAX	5.12	5.15	5.23	5.31	5.30	5.12	5.15	5.25	5.33	5.32
		EB-HIBETASIGMAX	5.25	5.24	5.30	5.32	5.31	5.25	5.24	5.30	5.32	5.31
0,0.4	5,150	RIDG	1.06	1.06	1.08	1.07	1.07	1.06	1.06	1.08	1.07	1.07
		HYB	0.99	1.00	1.02	1.03	1.04	0.99	1.00	1.02	1.03	1.04
		FB-FLATBETA	0.96	**0.97**	0.99	1.01	1.02	0.96	**0.97**	**0.99**	1.00	1.01
		FB-HIBETA-NI	0.97	0.98	1.01	1.03	1.04	0.97	0.98	1.01	1.03	1.04
		EB-HIBETA-NI	0.96	0.98	1.00	1.03	1.05	0.96	0.98	1.00	1.02	1.04
		EB-HISIGMAX	**0.96**	0.97	**0.99**	**1.00**	**1.01**	**0.96**	0.98	0.99	**1.00**	**1.01**
		EB-HIBETASIGMAX	0.96	0.98	0.99	1.01	1.02	0.96	0.98	0.99	1.00	1.02
0.75,0.4	5,150	RIDG	0.94	0.94	0.94	0.93	0.94	0.94	0.94	0.94	0.93	0.94
		HYB	0.89	0.90	0.92	0.92	0.93	0.89	0.90	0.92	0.92	0.93
		FB-FLATBETA	0.87	**0.87**	0.89	0.90	0.91	0.87	**0.87**	0.89	0.89	0.91
		FB-HIBETA-NI	0.87	0.88	0.91	0.93	0.95	0.87	0.88	0.91	0.92	0.94
		EB-HIBETA-NI	0.86	0.87	0.90	0.92	0.93	**0.86**	0.87	0.90	0.91	0.93
		EB-HISIGMAX	0.87	0.88	0.90	0.91	0.92	0.87	0.88	0.90	0.91	0.92
		EB-HIBETASIGMAX	**0.86**	0.87	**0.88**	**0.89**	**0.90**	0.86	0.87	**0.88**	**0.89**	**0.90**
0,0.1	5,150	RIDG	6.30	6.30	6.30	6.25	6.24	6.30	6.30	6.30	6.25	6.24
		HYB	6.00	6.03	6.08	6.07	6.10	6.00	6.03	6.08	6.07	6.10
		FB-FLATBETA	5.79	5.83	5.86	5.90	5.92	5.79	5.83	5.87	5.90	5.93
		FB-HIBETA-NI	**5.79**	**5.82**	**5.85**	**5.88**	**5.91**	**5.79**	**5.82**	**5.85**	**5.88**	**5.91**
		EB-HIBETA-NI	5.97	6.01	6.06	6.12	6.14	5.97	6.01	6.06	6.12	6.14
		EB-HISIGMAX	5.80	5.84	5.91	5.95	5.99	5.80	5.84	5.91	5.97	6.00
		EB-HIBETASIGMAX	5.98	6.01	6.07	6.11	6.12	5.98	6.01	6.07	6.11	6.12
0.75,0.1	5,150	RIDG	5.50	5.55	5.56	5.55	5.55	5.50	5.55	5.56	5.55	5.55
		HYB	5.35	5.37	5.41	5.46	5.51	5.35	5.37	5.41	5.46	5.51
		FB-FLATBETA	5.19	5.20	5.21	5.26	5.29	5.19	5.20	5.22	5.26	5.29
		FB-HIBETA-NI	**5.16**	**5.18**	**5.20**	**5.24**	**5.27**	**5.16**	**5.18**	**5.20**	**5.24**	**5.27**
		EB-HIBETA-NI	5.29	5.34	5.39	5.46	5.48	5.29	5.34	5.39	5.46	5.48
		EB-HISIGMAX	5.19	5.27	5.34	5.39	5.46	5.19	5.27	5.36	5.41	5.48
		EB-HIBETASIGMAX	5.29	5.34	5.37	5.42	5.47	5.29	5.34	5.37	5.42	5.48

Table B8: Numerical values of empirical MSPE for 8 simulation settings described in Section 3.5 when $p = 5$ and $\boldsymbol{X}|Z \sim N_p\{1_{[Z=2]}(3 \times \mathbf{1}_p) - 1_{[Z=3]}(3 \times \mathbf{1}_p), \boldsymbol{\Sigma}_{\boldsymbol{X}}\}$. $\hat{\boldsymbol{\beta}}^{\text{pm}}$ is the posterior mean and $\hat{\boldsymbol{\beta}}^{\text{ppm}}$ is the posterior predictive mean. The smallest MSPEs are in **bold**.

Appendix C: Chapter 4 Supplementary Materials

C.1 Generalized Maximum Profile Marginal Likelihood (GMPML)

In estimating variance components, Harville (1977) suggests to maximize the *restricted* log-likelihood, which offsets the log-likelihood to account for bias introduced from estimating "fixed" effects. Casting ridge regression in the mixed model framework, $\boldsymbol{\beta}$ is treated as random and so does not contribute bias to estimation. However, $\boldsymbol{y}$ is centered and $\boldsymbol{x}$ is standardized, which together implicitly estimate β_0 with $\hat{\beta}_0 = 0$. Thus, there is one unknown parameter hidden in the mean of the distribution $\boldsymbol{y}|\lambda, \sigma^2$, and the restricted marginal log-likelihood, denoted as $m_R(\lambda, \sigma^2)$, is as follows (Section 4.3, Harville, 1977):

$$\begin{aligned} m_R(\lambda, \sigma^2) &= m(\lambda, \sigma^2) - \frac{1}{2} \ln \left| \mathbf{1}_n^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda) \mathbf{1}_n / \sigma^2 \right| \\ &= -\frac{n-1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda) \boldsymbol{y} + \frac{1}{2} \ln \left| \boldsymbol{I}_n - \boldsymbol{D}_\lambda \right| - \frac{1}{2} \ln \mathbf{1}_n^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda) \mathbf{1}_n. \end{aligned}$$

By standardization of $\boldsymbol{x}$, it can be shown that the last term simplifies to a constant: $-(1/2)\ln(n)$. Replacing each instance of σ^2 with the restricted estimate $\hat{\sigma}^2_\lambda = \boldsymbol{y}^\top (\boldsymbol{I}_n - \boldsymbol{D}_\lambda)\boldsymbol{y}/(n-1)$, the optimization in (48) follows.

C.2 Maximum Adjusted Profile h-Likelihood (MAPHL)

The h-loglihood (Lee and Nelder, 1996) is given by

$$\ell_H(\boldsymbol{\beta}, \lambda, \sigma^2) = \ell(\boldsymbol{\beta}, \sigma^2) + p_\lambda(\boldsymbol{\beta}, \sigma^2).$$

When the dispersion and variance components, respectively σ^2 and λ, are unknown, Lee and Nelder propose maximization of the adjusted h-loglihood (Section 4.3, Lee and Nelder, 1996), to simultaneously estimate $\boldsymbol{\beta}$, λ, and σ^2. This, too, is a restricted log-likelihood. In contrast to $m_R(\lambda, \sigma^2)$ above, the h-loglihood must be adjusted for both β_0 and $\boldsymbol{\beta}$, because there is no marginalization. This adjusted h-loglihood is defined as

$$\begin{aligned}
\ell_{HA}(\boldsymbol{\beta}, \lambda, \sigma^2) &= \ell_H(\boldsymbol{\beta}, \lambda, \sigma^2) + \frac{1}{2}\ln(n\sigma^2) + \frac{1}{2}\ln\left|\sigma^2(\boldsymbol{x}^\top\boldsymbol{x} + \lambda)^{-1}\right| \\
&= -\frac{n-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&\quad + \frac{1}{2}\ln\left|\lambda(\boldsymbol{x}^\top\boldsymbol{x} + \lambda)^{-1}\right| + \frac{1}{2}\ln(n) \\
&= -\frac{n-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
&\quad + \frac{1}{2}\ln\left|\boldsymbol{I}_n - \boldsymbol{D}_\lambda\right| + \frac{1}{2}\ln(n).
\end{aligned}$$

Sequentially optimizing $\ell_{HA}(\boldsymbol{\beta}, \lambda, \sigma^2)$ with respect to each of $\boldsymbol{\beta}$, λ and σ^2 yields expressions (49)-(51).

C.3 Gamma Hyperpenalty

We have

$$\begin{aligned}
p_\lambda(\boldsymbol{\beta}, \sigma^2) + h_{\text{GA}}(\lambda) &= -\frac{p}{2}\ln(\sigma^2) + \frac{p}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} + (a-1)\ln(\lambda) - b\lambda \\
&= -\frac{p}{2}\ln(\sigma^2) + \frac{p + 2a - 2}{2}\ln(\lambda) - \lambda\left(\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2\sigma^2} + b\right).
\end{aligned}$$

From the conjugacy of the gamma hyperpenalty, the JO and MO updates immediately follow:

$$\begin{aligned}
&\text{JO} : \text{argmax}_{\lambda|\boldsymbol{\beta},\sigma^2}\{p_\lambda(\boldsymbol{\beta}, \sigma^2) + h_{\text{GA}}(\lambda)\} = \frac{p + 2a - 2}{\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma^2 + 2b}, \\
&\text{MO} : \text{E}_{\lambda|\boldsymbol{\beta},\sigma^2}[\lambda] = \frac{p + 2a}{\boldsymbol{\beta}^\top\boldsymbol{\beta}/\sigma^2 + 2b}.
\end{aligned}$$

C.4 Log-Normal Hyperpenalty

$$\begin{aligned} p_\lambda(\boldsymbol{\beta}, \sigma^2) + h_{\text{LN}}(\lambda) &= -\frac{p}{2}\ln(\sigma^2) + \frac{p}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} - \ln(\lambda) - \frac{1}{2}\frac{\ln(b\lambda)^2}{a} \\ &= -\frac{p}{2}\ln(\sigma^2) + \frac{p-2}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} - \frac{1}{2}\frac{\ln(b\lambda)^2}{a}. \end{aligned}$$

Neither the JO nor MO updates have closed forms, and we use numerical methods to calculate both.

C.5 Inv-Gamma Hyperpenalty

We have

$$\begin{aligned} p_\lambda(\boldsymbol{\beta}, \sigma^2) + h_{\text{IG}}(\lambda) &= -\frac{p}{2}\ln(\sigma^2) + \frac{p}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} - (a+1)\ln(\lambda) - \frac{1}{b\lambda} \\ &= -\frac{p}{2}\ln(\sigma^2) + \frac{p-2a-2}{2}\ln(\lambda) - \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\beta} - \frac{1}{b\lambda}. \end{aligned}$$

The JO update for λ can be computed using the quadratic formula. The MO update is expressed in terms of integrals and must be calculated using numeric methods.

$$\begin{aligned} &\text{JO} : \text{argmax}_{\lambda|\boldsymbol{\beta},\sigma^2}\{p_\lambda(\boldsymbol{\beta}, \sigma^2) + h_{\text{IG}}(\lambda)\} \\ &\quad = \left(p - 2a - 2 + \sqrt{(p-2a-2)^2 + \frac{8\boldsymbol{\beta}^\top\boldsymbol{\beta}}{b\sigma^2}}\right) \Big/ \left(\frac{2\boldsymbol{\beta}^\top\boldsymbol{\beta}}{\sigma^2}\right). \qquad (91) \\ &\text{MO} : \text{E}_{\lambda|\boldsymbol{\beta},\sigma^2}[\lambda] = \frac{\int \exp\{-t\lambda - u/\lambda\}\,\lambda^{v+1}\,\text{d}\lambda}{\int \exp\{-t\lambda - u/\lambda\}\,\lambda^{v}\,\text{d}\lambda}, \quad t = \frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2\sigma^2}, u = 1/b, v = \frac{p-2a-2}{2}. \qquad (92) \end{aligned}$$

As discussed in the main text, caution must be exercised in specifying the choice of a; this is made clear in the expressions immediately above. In particular, from (36), larger values of λ cause $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ to go to zero (recall that $\boldsymbol{\beta}$ and σ^2 are also sequentially updated). If smaller values of $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ in turn increase λ, iterative updates of λ will increase without

bound. For the JO update, this will happen if $p - 2a - 2 > 0$, ie $a < p/2 - 1$: the numerator of (91) will not approach zero as $\boldsymbol{\beta}^\top \boldsymbol{\beta}$ approaches zero, and the denominator will. If $p - 2a - 2 = 0$, then both the numerator and denominator will both go to zero, but the square root will cause the numerator to approach zero more slowly. Thus, a must be such that $p - 2a - 2 < 0$, ie $a > p/2 - 1$.

Similarly, a sufficient condition to ensure that the integrals in the MO update are finite for any value of $\boldsymbol{\beta}$ and σ^2 is if $v + 1 \leq -1$, where $v = (p - 2a - 2)/2$. That is, observe that

$$\int \exp\{-t\lambda - u/\lambda\}\, \lambda^{v+1}\, \mathrm{d}\lambda < \int \lambda^{v+1}\, \mathrm{d}\lambda.$$

Thus, the left-hand side is guaranteed to be finite if $(p - 2a - 2)/2 + 1 \leq -1$, ie $a \geq p/2 + 1$.

Appendix D: Chapter 5 Supplementary Materials

D.1 E-Step

The *complete* log-likelihood is given by

$$
\begin{aligned}
\ell_C &= \ln[\boldsymbol{U}^{\text{obs}}, \boldsymbol{U}^{\text{mis}} | \boldsymbol{\phi}] \\
&= \ln[\boldsymbol{y}_{\text{A}} | \boldsymbol{x}_{\text{A}}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_{\text{A}} | \boldsymbol{x}_{\text{A}}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_{\text{A}} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] \\
&\quad + \ln[\boldsymbol{y}_{\text{B}} | \boldsymbol{x}_{\text{B}}, \beta_0, \boldsymbol{\beta}, \sigma^2] + \ln[\boldsymbol{w}_{\text{B}} | \boldsymbol{x}_{\text{B}}, \psi, \nu, \tau^2] + \ln[\boldsymbol{x}_{\text{B}} | \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X] \\
&= -\frac{n_{\text{A}} + n_{\text{B}}}{2} \log(\sigma^2 \tau^{2p} / |\boldsymbol{\Sigma}_X^{-1}|) \\
&\quad - \frac{1}{2\sigma^2} (\boldsymbol{y}_{\text{A}} - \beta_0 \mathbf{1}_{n_{\text{A}}} - \boldsymbol{x}_{\text{A}} \boldsymbol{\beta})^\top (\boldsymbol{y}_{\text{A}} - \beta_0 \mathbf{1}_{n_{\text{A}}} - \boldsymbol{x}_{\text{A}} \boldsymbol{\beta}) \\
&\quad - \frac{1}{2} \sigma^2 (\boldsymbol{y}_{\text{B}} - \beta_0 \mathbf{1}_{n_{\text{B}}} - \boldsymbol{x}_{\text{B}} \boldsymbol{\beta})^\top (\boldsymbol{y}_{\text{B}} - \beta_0 \mathbf{1}_{n_{\text{B}}} - \boldsymbol{x}_{\text{B}} \boldsymbol{\beta}) \\
&\quad - \frac{1}{2\tau^2} \text{Tr}\ (\boldsymbol{w}_{\text{A}} - \psi \mathbf{1}_{n_{\text{A}}} \mathbf{1}_p^\top - \nu \boldsymbol{x}_{\text{A}})^\top (\boldsymbol{w}_{\text{A}} - \psi \mathbf{1}_{n_{\text{A}}} \mathbf{1}_p^\top - \nu \boldsymbol{x}_{\text{A}}) \\
&\quad - \frac{1}{2\tau^2} \text{Tr}\ (\boldsymbol{w}_{\text{B}} - \psi \mathbf{1}_{n_{\text{B}}} \mathbf{1}_p^\top - \nu \boldsymbol{x}_{\text{B}})^\top (\boldsymbol{w}_{\text{B}} - \psi \mathbf{1}_{n_{\text{B}}} \mathbf{1}_p^\top - \nu \boldsymbol{x}_{\text{B}}) \\
&\quad - \frac{1}{2} \text{Tr}\ \big((\boldsymbol{x}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\mu}_X^\top) \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x}_{\text{A}} - \mathbf{1}_{n_{\text{A}}} \boldsymbol{\mu}_X^\top)^\top + (\boldsymbol{x}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\mu}_X^\top) \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \boldsymbol{\mu}_X^\top)^\top \big).
\end{aligned}
$$

Because every expression containing $\boldsymbol{x}_{\text{B}}$ in ℓ_C is a quadratic form, ie each expression of $\boldsymbol{x}_{\text{B}}$ is either linear or quadratic, $Q(\boldsymbol{\phi} | \boldsymbol{\phi}^{(t)})$ can be calculated as follows. As in the main text, let

$$
\begin{aligned}
\boldsymbol{\Gamma}(\boldsymbol{\phi}) &= (\boldsymbol{\beta}\boldsymbol{\beta}^\top / \sigma^2 + (\nu^2 / \tau^2) \boldsymbol{I}_p + \boldsymbol{\Sigma}_X^{-1})^{-1}, \\
\boldsymbol{x}_{\text{B}}^{\text{EM}}(\boldsymbol{\phi}) &= ([\boldsymbol{y}_{\text{B}} - \mathbf{1}_{n_{\text{B}}} \beta_0] \boldsymbol{\beta}^\top / \sigma^2 + [\nu / \tau^2] [\boldsymbol{w}_{\text{B}} - \psi \mathbf{1}_{n_{\text{B}}} \mathbf{1}_p^\top] + [\mathbf{1}_{n_{\text{B}}} \boldsymbol{\mu}_X^\top] \boldsymbol{\Sigma}_X^{-1}) \boldsymbol{\Gamma}(\boldsymbol{\phi}).
\end{aligned}
$$

Then,

$$\begin{aligned}
\mathrm{E}[\boldsymbol{x}_{\mathrm{B}}|\boldsymbol{\phi}, \boldsymbol{U}^{\mathrm{obs}}] &= \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}}(\boldsymbol{\phi}), \\
\mathrm{E}[\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}|\boldsymbol{\phi}, \boldsymbol{U}^{\mathrm{obs}}] &= \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}}(\boldsymbol{\phi})^{\top}\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}}(\boldsymbol{\phi}) + n_{\mathrm{B}}\boldsymbol{\Gamma}(\boldsymbol{\phi}).
\end{aligned}$$

Setting $\boldsymbol{\Gamma}^{(t)} = \boldsymbol{\Gamma}(\boldsymbol{\phi}^{(t)})$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)} = \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}}(\boldsymbol{\phi}^{(t)})$, $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})$ as given in (68) is easily derived. In words, to calculate $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})$, begin with ℓ_C, replace each instance of $\boldsymbol{x}_{\mathrm{B}}$ with its conditional expectation, and subtract off the term $(n_{\mathrm{B}}/2)\mathrm{Tr}\left(\boldsymbol{\Gamma}^{(t)}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\phi})\right)$. This offset accounts for the uncertainty in the "imputations", $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}$. Its size is based on both the size of subsample B and the estimated covariance of the imputations, and it differentiates the EM algorithm from a basic single imputation approach.

D.2 Hyperpenalized M-Steps

EM-FLATBETA

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= \boldsymbol{\beta}^\top \boldsymbol{x}_\mathrm{A}^\top [\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}}] - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} \\
&\quad + \boldsymbol{\beta}^\top \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} [\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{A}}] - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta} - \frac{n_\mathrm{B}}{2}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_\mathrm{A}^\top [\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}}] - \boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} \\
&\quad + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} [\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{A}}] - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta} - n_\mathrm{B}\boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} \\
\Rightarrow \boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A} + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} + n_\mathrm{B}\boldsymbol{\Gamma}^{(t)})^{-1} (\boldsymbol{x}_\mathrm{A}^\top [\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}}] + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} [\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{A}}]).
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= -\frac{(n_\mathrm{A} + n_\mathrm{B})}{2}\beta_0^2 + \beta_0 (\boldsymbol{y}_\mathrm{A} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{A}} + \beta_0 (\boldsymbol{y}_\mathrm{B} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{B}} \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial \beta_0} &= -\beta_0 (n_\mathrm{A} + n_\mathrm{B}) + (\boldsymbol{y}_\mathrm{A} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{A}} + (\boldsymbol{y}_\mathrm{B} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{B}} \\
\Rightarrow \beta_0^{(t+1)} &= \frac{(\boldsymbol{y}_\mathrm{A} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{A}} + (\boldsymbol{y}_\mathrm{B} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top \mathbf{1}_{n_\mathrm{B}}}{n_\mathrm{A} + n_\mathrm{B}}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= -(n_\mathrm{A} + n_\mathrm{B}) \ln \sigma^2 - \frac{1}{\sigma^2}(\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta}) \\
&\quad - \frac{1}{\sigma^2}(\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta}) \\
&\quad - \frac{n_\mathrm{B}}{\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial \sigma^2} &= -(n_\mathrm{A} + n_\mathrm{B})\frac{1}{\sigma^2} + \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta}) \\
&\quad + \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta}) \\
&\quad + \frac{1}{(\sigma^2)^2}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} \\
\Rightarrow \sigma^{2(t+1)} &= \frac{(\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{A} - \beta_0 \mathbf{1}_{n_\mathrm{A}} - \boldsymbol{x}_\mathrm{A}\boldsymbol{\beta})}{n_\mathrm{A} + n_\mathrm{B}} \\
&\quad + \frac{(\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_\mathrm{B} - \beta_0 \mathbf{1}_{n_\mathrm{B}} - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta})}{n_\mathrm{A} + n_\mathrm{B}} + \frac{n_\mathrm{B}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta}}{n_\mathrm{A} + n_\mathrm{B}}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= -\frac{p(n_{\mathrm{A}}+n_{\mathrm{B}})}{2}\psi^2 + \psi \mathrm{Tr}\left[\mathbf{1}_p\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\nu\boldsymbol{x}_{\mathrm{A}}) + \mathbf{1}_p\mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\nu\boldsymbol{x}_{\mathrm{B}})\right] \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial\psi} &= -\psi p(n_{\mathrm{A}}+n_{\mathrm{B}}) + \mathrm{Tr}\left[\mathbf{1}_p\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\nu\boldsymbol{x}_{\mathrm{A}}) + \mathbf{1}_p\mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\nu\boldsymbol{x}_{\mathrm{B}})\right] \\
\Rightarrow \psi^{(t+1} &= \frac{\mathbf{1}_{n_{\mathrm{A}}}^\top(\boldsymbol{w}_{\mathrm{A}}-\nu\boldsymbol{x}_{\mathrm{A}})\mathbf{1}_p + \mathbf{1}_{n_{\mathrm{B}}}^\top(\boldsymbol{w}_{\mathrm{B}}-\nu\boldsymbol{x}_{\mathrm{B}})\mathbf{1}_p}{(n_{\mathrm{A}}+n_{\mathrm{B}})p}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= \nu\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top)\right] - \frac{\nu^2}{2}\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}\right] + \nu\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top)\right] \\
&\quad - \frac{\nu^2}{2}\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\right] - \frac{n_{\mathrm{B}}\nu^2}{2}\mathrm{Tr}\left[\boldsymbol{\Gamma}^{(t)}\right] \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial\nu} &= \mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top)\right] - \nu\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}}\right] + \mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top)\right] \\
&\quad - \nu\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\right] - n_{\mathrm{B}}\nu\mathrm{Tr}\left[\boldsymbol{\Gamma}^{(t)}\right] \\
\Rightarrow \nu^{(t+1)} &= \frac{\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top) + \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top)\right]}{\mathrm{Tr}\left[\boldsymbol{x}_{\mathrm{A}}^\top\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)\top}\boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)} + n_{\mathrm{B}}\boldsymbol{\Gamma}^{(t)}\right]}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= -p(n_{\mathrm{A}}+n_{\mathrm{B}})\ln\tau^2 - \frac{1}{\tau^2}\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}\right)^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}) \\
&\quad - \frac{1}{\tau^2}\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}}\right)^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}}) - \frac{n_{\mathrm{B}}\nu^2}{\tau^2}\mathrm{Tr}\left(\boldsymbol{\Gamma}^{(t)}\right) \\
\frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial\tau^2} &= -p(n_{\mathrm{A}}+n_{\mathrm{B}})\frac{1}{\tau^2} + \frac{1}{(\tau^2)^2}\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}\right)^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}) \\
&\quad + \frac{1}{(\tau^2)^2}\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}}\right)^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}}) \\
&\quad + \frac{n_{\mathrm{B}}\nu^2}{(\tau^2)^2}\mathrm{Tr}\left(\boldsymbol{\Gamma}^{(t)}\right) \\
\Rightarrow \tau^{2(t+1)} &= \frac{\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}}\right)^\top(\boldsymbol{w}_{\mathrm{A}}-\psi\mathbf{1}_{n_{\mathrm{A}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{A}})}{p(n_{\mathrm{A}}+n_{\mathrm{B}})} \\
&\quad + \frac{\mathrm{Tr}\left(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}}\right)^\top(\boldsymbol{w}_{\mathrm{B}}-\psi\mathbf{1}_{n_{\mathrm{B}}}\mathbf{1}_p^\top-\nu\boldsymbol{x}_{\mathrm{B}})}{p(n_{\mathrm{A}}+n_{\mathrm{B}})} + \frac{n_{\mathrm{B}}\nu^2\mathrm{Tr}\left(\boldsymbol{\Gamma}^{(t)}\right)}{p(n_{\mathrm{A}}+n_{\mathrm{B}})}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) &= \mathrm{Tr}\,\big[\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{A}^\top - \frac{1}{2}\mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X 1_{n_\mathrm{A}}^\top \\
&\quad + \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} - \frac{1}{2}\mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X 1_{n_\mathrm{B}}^\top\big] \\
&= \boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{A}^\top 1_{n_\mathrm{A}} - \frac{n_\mathrm{A}}{2}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X + \boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} 1_{n_\mathrm{B}} - \frac{n_\mathrm{B}}{2}\boldsymbol{\mu}_X^\top\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X \\
\frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial \boldsymbol{\mu}_X} &= \boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{A}^\top 1_{n_\mathrm{A}} - n_\mathrm{A}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_X^{-1}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top} 1_{n_\mathrm{B}} - n_\mathrm{B}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\mu}_X \\
\Rightarrow \boldsymbol{\mu}_X^{(t+1)} &= \frac{\boldsymbol{x}_\mathrm{A}^\top\mathbf{1}_{n_\mathrm{A}} + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top}\mathbf{1}_{n_\mathrm{B}}}{n_\mathrm{A}+n_\mathrm{B}}.
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) + p_{\boldsymbol{\eta}}(\boldsymbol{\phi}) &= (n_\mathrm{A}+n_\mathrm{B})\ln|\boldsymbol{\Sigma}_X^{-1}| - \mathrm{Tr}\,\big[(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}\big] \\
&\quad - \mathrm{Tr}\,\big[(\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)\boldsymbol{\Sigma}_X^{-1}\big] - n_\mathrm{B}\mathrm{Tr}\,\big[\boldsymbol{\Gamma}^{(t)}\boldsymbol{\Sigma}_X^{-1}\big] \\
&\quad - (2p-1)\ln\left|\boldsymbol{\Sigma}_X^{-1}\right| - (2p-1)\mathrm{Tr}\,[\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_\mathrm{A}])\boldsymbol{\Sigma}_X^{-1}] \\
\Rightarrow \boldsymbol{\Sigma}_X^{-1(t+1)} &= \Bigg(\frac{(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{A} - \mathbf{1}_{n_\mathrm{A}}\boldsymbol{\mu}_X^\top) + (\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)^\top(\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} - \mathbf{1}_{n_\mathrm{B}}\boldsymbol{\mu}_X^\top)}{n_\mathrm{A}+n_\mathrm{B}+2p-1} \\
&\quad + \frac{n_\mathrm{B}\boldsymbol{\Gamma}^{(t)} + (2p-1)\mathrm{diag}(\hat{\mathrm{Var}}[\boldsymbol{x}_\mathrm{A}])}{n_\mathrm{A}+n_\mathrm{B}+2p-1}\Bigg)^{-1}.
\end{aligned}
$$

em-hibeta-ga, em-hibeta-ln, em-hibeta-ig Given λ, the M-steps for $\boldsymbol{\beta}$ and σ^2 are modified as follows:

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) + p_{\boldsymbol{\eta}}(\boldsymbol{\phi}) &= \boldsymbol{\beta}^\top\boldsymbol{x}_\mathrm{A}^\top[\boldsymbol{y}_\mathrm{A} - \beta_0\mathbf{1}_{n_\mathrm{A}}] - \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{x}_\mathrm{A}^\top\boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}[\boldsymbol{y}_\mathrm{B} - \beta_0\mathbf{1}_{n_\mathrm{A}}] \\
&\quad - \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta} - \frac{n_\mathrm{B}}{2}\boldsymbol{\beta}^\top\boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} - \frac{1}{2}\lambda\boldsymbol{\beta}^\top\boldsymbol{\beta} \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial\boldsymbol{\beta}} + \frac{\partial p_{\boldsymbol{\eta}}(\boldsymbol{\phi})}{\partial\boldsymbol{\beta}} &= \boldsymbol{x}_\mathrm{A}^\top[\boldsymbol{y}_\mathrm{A} - \beta_0\mathbf{1}_{n_\mathrm{A}}] - \boldsymbol{x}_\mathrm{A}^\top\boldsymbol{x}_\mathrm{A}\boldsymbol{\beta} \\
&\quad + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}[\boldsymbol{y}_\mathrm{B} - \beta_0\mathbf{1}_{n_\mathrm{A}}] - \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}\boldsymbol{\beta} - n_\mathrm{B}\boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} - \lambda\boldsymbol{\beta} \\
\Rightarrow \boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{x}_\mathrm{A}^\top\boldsymbol{x}_\mathrm{A} + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)\top}\boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)} + n_\mathrm{B}\boldsymbol{\Gamma}^{(t)} + \lambda\boldsymbol{I}_p)^{-1} \\
&\quad \times (\boldsymbol{x}_\mathrm{A}^\top[\boldsymbol{y}_\mathrm{A} - \beta_0\mathbf{1}_{n_\mathrm{A}}] + \boldsymbol{x}_\mathrm{B}^{\mathrm{EM}(t)}[\boldsymbol{y}_\mathrm{B} - \beta_0\mathbf{1}_{n_\mathrm{A}}]).
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)}) + p_{\boldsymbol{\eta}}(\boldsymbol{\phi}) = &-(n_{\mathrm{A}} + n_{\mathrm{B}}) \ln \sigma^2 \\
&- \frac{1}{\sigma^2}(\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&- \frac{1}{\sigma^2}(\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta}) \\
&- \frac{n_{\mathrm{B}}}{\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} - p \ln(\sigma^2) - \frac{1}{\sigma^2}\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \\
\Rightarrow \frac{\partial Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(t)})}{\partial \sigma^2} + \frac{\partial p_{\boldsymbol{\eta}}(\boldsymbol{\phi})}{\partial \sigma^2} = &-(n_{\mathrm{A}} + n_{\mathrm{B}}) \frac{1}{\sigma^2} \\
&+ \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta}) \\
&+ \frac{1}{(\sigma^2)^2}(\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta}) \\
&+ \frac{1}{(\sigma^2)^2}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} - p\frac{1}{\sigma^2} + \frac{1}{(\sigma^2)^2}\lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \\
\Rightarrow \sigma^{2(t+1)} = &\frac{(\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{A}} - \beta_0 \mathbf{1}_{n_{\mathrm{A}}} - \boldsymbol{x}_{\mathrm{A}}\boldsymbol{\beta})}{n_{\mathrm{A}} + n_{\mathrm{B}} + p} \\
&+ \frac{(\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta})^\top (\boldsymbol{y}_{\mathrm{B}} - \beta_0 \mathbf{1}_{n_{\mathrm{B}}} - \boldsymbol{x}_{\mathrm{B}}^{\mathrm{EM}(t)}\boldsymbol{\beta})}{n_{\mathrm{A}} + n_{\mathrm{B}} + p} \\
&+ \frac{n_{\mathrm{B}}\boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{(t)}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}}{n_{\mathrm{A}} + n_{\mathrm{B}} + p}.
\end{aligned}
$$

Assumed Models	Type of Algorithm: Ad-hoc	Bayesian	Empirical Bayes	Maximum (hyper-) penalized likelihood
Outcome +ME	FRC (2.2)			
Outcome +ME +Marginal	SRC (2.2)	FB-FLATBETA (3.3)		EM-FLATBETA (5.3.2)
Outcome +penalty on $\boldsymbol{\beta}$				RIDG (2.2)
Outcome +ME +Marginal +penalty on $\boldsymbol{\beta}$		FB-HIBETA-NI (3.3.1) FB-HIBETA-GA (5.3.3)	EB-HIBETA-NI (3.3.1)	EM-HIBETA-GA (5.3.2) EM-HIBETA-LN (5.3.2) EM-HIBETA-IG (5.3.2)
Outcome +ME +Marginal +penalty on $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$			EB-HISIGMAX (3.3.2)	
Outcome +ME +Marginal +penalty on $\boldsymbol{\beta}$ +penalty on $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$			EB-HIBETASIGMAX (3.3.2)	

Table D1: Cross-tabulation of the missing data methods indexed by the assumed models and the algorithm. In parentheses is the section that describes the method. The Outcome model, measurement error (ME) model, and Marginal model for $\boldsymbol{X}$ are all given in (24). The penalty on $\boldsymbol{\beta}$ is given in (30) and the penalty on $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ is given in (31). Two additional methods, HYB (Section 2.3) and HYB_C (Section 5.2), are adaptively weighted linear combinations of RIDG, SRC and FRC and thus average over several model assumptions.

	Name	Description
Chapter 2	RIDG	Ridge regression on subsample A only using GCV to choose λ
	SRC	Structural Regression Calibration.
	FRC	Functional Regression Calibration. Similar to SRC but with no explicit distributional assumption for $\boldsymbol{X}$.
	HYB	Adaptively weighted linear combination of RIDG, SRC and FRC
Chapter 3	FB-FLATBETA	Bayesian data augmentation with flat prior on $\boldsymbol{\beta}$ and mild, fixed shrinkage of $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$.
	FB-HIBETA-NI	Bayesian ridge with Jeffreys hyperprior on λ.
	EB-HIBETA-NI	Bayesian ridge with Empirical Bayes estimation of λ.
	EB-HISIGMAX	Bayesian data augmentation with flat prior on $\boldsymbol{\beta}$ and Empirical Bayes shrinkage on $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$.
	EB-HIBETASIGMAX	Combination of EB-HIBETA-NI and EB-HISIGMAX, ie Empirical Bayes shrinkage on both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$
Chapter 5	HYB_C	HYB with weights estimated using the corrected GCV criterion
	FB-HIBETA-GA	Bayesian ridge with a Gamma hyperprior on λ
	EM-FLATBETA	The penalized maximum likelihood equivalent of FB-FLATBETA
	EM-HIBETA-GA	Maximum hyperpenalized likelihood using the HEM algorithm with a Gamma hyperpenalty
	EM-HIBETA-LN	Maximum hyperpenalized likelihood using the HEM algorithm with a Log-Normal hyperpenalty
	EM-HIBETA-IG	Maximum hyperpenalized likelihood using the HEM algorithm with an Inv-Gamma hyperpenalty

Table D2: Glossary of the missing data methods indexed by chapter.

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* pages 267–281.

Armagan, A. and Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with t priors. *Computational Statistics* **25,** 441–461.

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

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer, New York, 2nd edition.

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, Boca Raton, FL.

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

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics* **2,** 485–500.

Chen, G., Kim, S., Taylor, J. M. G., Wang, Z., Lee, O., Ramnath, N., Reddy, R. M., Lin, J., Chang, A. C., Orringer, M. B., and Beer, D. G. (2011). Development and validation of a qRT-PCR–classifier for lung cancer prognosis. *Journal of Thoracic Oncology* **6,** 1481–1487.

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science* **15,** 46–60.

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., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39,** 1–38.

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.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316–331.

Efron, B. (2001). Selection criteria for scatterplot smoothers. *Annals of Statistics* **29,** 470–504.

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

Fu, W. J. (1998). Penalized regressions: The bridge versus the Lasso. *Journal of Computational and Graphical Statistics* **7,** 397–416.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York.

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, Stanford University.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel hierarchical models*. Cambridge University Press, New York.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the

Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6,** 721–741.

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

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

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.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B* **52,** 443–452.

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

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72,** 320–338.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, 2nd edition.

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

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B* **60,** 271–293.

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.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B* **58,** 619–678.

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.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, NJ, 2nd edition.

Lukas, M. A. (2006). Robust generalized cross-validation for choosing the regularization parameter. *Inverse Problems* **22,** 1883–1902.

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

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **12,** 591–612.

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

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

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

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103,** 681–686.

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.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc., Hoboken, NJ.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

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

Segal, M. R., Bacchetti, P., and Jewell, N. P. (1994). Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **56,** 345–352.

Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* **71,** 207–225.

Strawderman, R. L. and Wells, M. T. (2012). On hierarchical prior specifications and penalized likelihood. In Fourdrinier, D., Éric Marchand, and Rukhin, A. L., editors, *Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman*, volume 8. Institute of Mathematical Statistics.

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

Takada, Y. (1979). Stein's positive part estimator and Bayes estimator. *Annals of the Institute of Statistical Mathematics* **31,** 177–183.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82,** 528–540.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B* **58,** 267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1,** 211–244.

Tran, M. N. (2009). Penalized maximum likelihood for choosing ridge parameter. *Com-*

munications in Statistics **38,** 1610–1624.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* **13,** 1378–1402.

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statistics & Probability Letters* **25,** 105–111.

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.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association* **78,** 81–89.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithms. *Journal of the American Statistical Association* **85,** 699–704.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B* **71,** 615–636.

Yi, N. and Xu, S. (2008). Bayesian Lasso for quantitative trait loci mapping. *Genetics* **179,** 1045–1055.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67,** 301–320.