Biostatistics (2012), **0**, 0, *pp.* 1–23
doi:10.1093/biostatistics/BoonstraBiostatisticsSuppMatV2

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

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

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

philb@umich.edu

APPENDIX

A. Analysis of Targeted Ridge Estimators

This section proves some results for TR estimators, first evaluating them as imputations for the missing data, $\boldsymbol{x}_{\mathrm{B}}$, 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, $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ are equivalent to filling in the missing $\boldsymbol{x}_{\mathrm{B}}$ with $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}}$ and $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$ and doing OLS on the completed data. Due to Marquardt (1970), RIDG can also be viewed as imputing the missing $\boldsymbol{x}_{\mathrm{B}}$ with $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{RIDG}} = [\sqrt{\lambda}\boldsymbol{I}_p \, \mathbf{0}_p \, \cdots \, \mathbf{0}_p]^\top$, replacing the observed $\boldsymbol{y}_{\mathrm{B}}$ with $\mathbf{0}_{n_{\mathrm{B}}}$, and doing OLS on the completed data. In general, we have the following result for any targeted ridge estimator.

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

Proof. For any $(\boldsymbol{\gamma}_{\boldsymbol{\beta}}, \lambda, \boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1})$ defining a TR estimator in (2.6), 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 $\tilde{\boldsymbol{x}}_{\mathrm{B}} = [\sqrt{\lambda}\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2}\,\mathbf{0}_p \cdots \mathbf{0}_p]^{\top}$, where $\mathbf{0}_p$ is repeated $n_{\mathrm{B}} - p$ times and $\tilde{\boldsymbol{y}}_{\mathrm{B}} = [\sqrt{\lambda}\boldsymbol{\gamma}_{\boldsymbol{\beta}}^{\top}\boldsymbol{\Omega}_{\boldsymbol{\beta}}^{-1/2}\,0 \cdots 0]^{\top}$, 0 repeated $n_{\mathrm{B}} - p$ times. This gives the desired result. □

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

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 Let the squared Frobenius norm of a matrix $\boldsymbol{S}$ be given by $\| \boldsymbol{S} \|_{\mathrm{F}}^2 = \mathrm{Tr}\left[\boldsymbol{S}^{\top}\boldsymbol{S}\right]$. Then, $\mathrm{E}_{\boldsymbol{x}_{\mathrm{B}}, \boldsymbol{w}_{\mathrm{B}}}\left[\|\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}} - \boldsymbol{x}_{\mathrm{B}}\|_{\mathrm{F}}^2 - \|\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}} - \boldsymbol{x}_{\mathrm{B}}\|_{\mathrm{F}}^2\right] \geqslant 0$

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[\tfrac{1}{\nu^2}\boldsymbol{V}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V} - \tfrac{1}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V} - \tfrac{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[\tfrac{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} - \tfrac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} - \tfrac{\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[\tfrac{1}{\nu^2}\boldsymbol{V}(\nu^2\boldsymbol{\Sigma}_{\boldsymbol{X}} + \tau^2 n_{\mathrm{B}}\boldsymbol{I}_p)\boldsymbol{V} - \tfrac{\nu}{\nu}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V} - \tfrac{\nu}{\nu}\boldsymbol{V}\boldsymbol{\Sigma}_{\boldsymbol{X}} + \boldsymbol{\Sigma}_{\boldsymbol{X}}\right] \\
&\quad = \mathrm{Tr}\left[n_{\mathrm{B}}\tfrac{\tau^2}{\nu^2}\boldsymbol{V}^2 + n_{\mathrm{B}}(\boldsymbol{I}_p - \boldsymbol{V})^2\boldsymbol{\Sigma}_{\boldsymbol{X}}\right] \qquad (\boldsymbol{V}\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V}) \\
&\quad = n_{\mathrm{B}}\tfrac{\tau^2}{\nu^2}\mathrm{Tr}\boldsymbol{V} \qquad (\boldsymbol{\Sigma}_{\boldsymbol{X}} = \tfrac{\tau^2}{\nu^2}(\boldsymbol{I}_p - \boldsymbol{V})^{-1}\boldsymbol{V}) \qquad \text{(A.1)} \\
&\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[\tfrac{1}{\nu^2}\boldsymbol{w}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}} - \tfrac{1}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{w}_{\mathrm{B}} - \tfrac{1}{\nu}\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[\tfrac{1}{\nu^2}(\nu^2\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \tau^2 n_{\mathrm{B}}\boldsymbol{I}_p) - \tfrac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} - \tfrac{\nu}{\nu}\boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{x}_{\mathrm{B}}^\top\boldsymbol{x}_{\mathrm{B}}\right] \\
&\quad = n_{\mathrm{B}}\tfrac{\tau^2}{\nu^2}\mathrm{Tr}\boldsymbol{I}_p \qquad \text{(A.2)}
\end{aligned}
$$

A comparison of expressions (A.1) and (A.2), together with the inequality $\mathrm{Tr}\,(\boldsymbol{I}_p - \boldsymbol{V}) \geqslant 0$ implied by (2.11) 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 approximate ridge-type estimator on the *complete* data, as demonstrated by the following relationship in their functional forms. By definition, $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}} = (1/\nu)\boldsymbol{w}_{\mathrm{B}} = \boldsymbol{x}_{\mathrm{B}} + (\tau/\nu)\boldsymbol{\xi}_{\mathrm{B}}$, where $\boldsymbol{\xi}_{\mathrm{B}}$ is the unobserved $n_{\mathrm{B}} \times p$ error matrix. From this, and

the definition of $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{FRC}}$ in (2.12), we have:

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1} = \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}} + \tfrac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}}, \qquad \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{FRC}}} = \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}} \tag{A.3}$$

Plugging these values of $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}$ and $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{FRC}}}$ into (2.7) gives that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{FRC}} &= (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}} + \tfrac{\tau^2}{\nu^2}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}})^{-1}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}}) \\ &\approx (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}} + n_{\mathrm{B}}\tfrac{\tau^2}{\nu^2}\boldsymbol{I}_p)^{-1}(\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}} + \boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}}),\end{aligned} \tag{A.4}$$

where the last approximation replaces each expression involving $\boldsymbol{\xi}_{\mathrm{B}}$ in the previous line with its marginal expectation. Thus (A.4) characterizes $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ as an approximate ridge-type estimator based on the complete data, with the shrinkage parameter $n_{\mathrm{B}}\tau^2/\nu^2$. Ridge regression can improve prediction error over OLS for certain choices of the tuning parameter (Gelfand, 1986; Frank and Friedman, 1993). Consequently, $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ may offer improved prediction, even over OLS on the complete data; whether this holds in practice depends crucially on the size of $n_{\mathrm{B}}\tau^2/\nu^2$. As τ/ν increases, $\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$ approaches zero, as seen by the expansion above. Interpreted from the Bayesian perspective, this is because the prior mean, $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\mathrm{FRC}}}$, approaches $\mathbf{0}_p$ with τ/ν, and the prior precision, $\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\mathrm{FRC}}}^{-1}$, grows without bound with τ/ν.

Following a similar expansion for SRC as above, note that $\boldsymbol{x}_{\mathrm{B}}^{\mathrm{SRC}} = (1/\nu)\boldsymbol{w}_{\mathrm{B}}\boldsymbol{V} = \boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} + (\tau/\nu)\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V}$ (if $\boldsymbol{\mu}_{\boldsymbol{X}}$ is assumed to be zero). When we expand $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ as in (A.4), we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{SRC}} &= (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{x}_{\mathrm{B}}\boldsymbol{V} + \tfrac{\tau^2}{\nu^2}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{\xi}_{\mathrm{B}}\boldsymbol{V})^{-1} \\ &\quad \times (\boldsymbol{x}_{\mathrm{A}}^{\top}\boldsymbol{y}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{x}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}} + \tfrac{\tau}{\nu}\boldsymbol{V}\boldsymbol{\xi}_{\mathrm{B}}^{\top}\boldsymbol{y}_{\mathrm{B}})\end{aligned} \tag{A.5}$$

From (2.11), as $\tau/\nu \to \infty$, the elements of $\boldsymbol{V}$ go to zero at a rate proportional to τ^2/ν^2. Thus, for large τ/ν, $\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ 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/ν, 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/ν (because $\mathrm{Var}\hat{\boldsymbol{\beta}}_{\text{SRC}}$ is unbounded), while $\hat{\boldsymbol{\beta}}_{\text{FRC}}$ is not. Next, we more formally compare SRC and FRC in terms of their MSPE.

THEOREM A.3 Let $\boldsymbol{V}$ and $\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}$ be as in (2.11) and (2.12), respectively. Also, define $\kappa = (\tau^2/\nu^2)\boldsymbol{\beta}^\top \boldsymbol{V}\boldsymbol{\beta}$ and

$$\boldsymbol{\Delta}^{\text{SRC}}_{\sigma} = \sigma^2(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}\boldsymbol{V})^{-1} \tag{A.6}$$

$$\boldsymbol{\Delta}^{\text{SRC}}_{\boldsymbol{\beta}} = \kappa(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}\boldsymbol{V})^{-1}\boldsymbol{V}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}\boldsymbol{V}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}\boldsymbol{V})^{-1} \tag{A.7}$$

$$\boldsymbol{\Delta}^{\text{FRC}}_{\sigma} = \sigma^2(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})^{-1} \tag{A.8}$$

$$\begin{aligned}\boldsymbol{\Delta}^{\text{FRC}}_{\boldsymbol{\beta}} = {} & \kappa(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})^{-1}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})^{-1} \\ & + (\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})^{-1}\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta}_{\text{FRC}}})^{-1}.\end{aligned} \tag{A.9}$$

Then using (A.6)–(A.9), the MSPE of the SRC and FRC methods can each be expressed as

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

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

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

These in turn yield the mean and variance of $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{SRC}}}$ and $\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{FRC}}}$. Now, assume $\beta_0 = \psi = 0$. With these results

and equations (2.7) and (2.8), we can write

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

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

$$
\begin{aligned}
&\operatorname{Bias}\hat{\boldsymbol{\beta}}_{\text{SRC}}\operatorname{Bias}\hat{\boldsymbol{\beta}}_{\text{SRC}}^{\top} + \operatorname{Var}\hat{\boldsymbol{\beta}}_{\text{SRC}} \\
&= (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1})^{-1} \\
&\quad \times \left\{\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1}(\operatorname{E}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{SRC}}} - \boldsymbol{\beta})(\operatorname{E}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{SRC}}} - \boldsymbol{\beta})^{\top}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1} + \sigma^2\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1}\operatorname{Var}\boldsymbol{\gamma}_{\boldsymbol{\beta}_{\text{SRC}}}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1}\right\} \\
&\quad \times (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{SRC}}}^{-1})^{-1} \\
&= (\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1}\left\{\sigma^2\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + (\sigma^2+\kappa)\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V}\right\}(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1} \\
&= \sigma^2(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1} + \kappa(\boldsymbol{x}_{\text{A}}^{\top}\boldsymbol{x}_{\text{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}_{\text{A}}^{\top}\boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{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 $\text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{SRC}}) - \text{MSPE}(\hat{\boldsymbol{\beta}}_{\text{FRC}}) = \sigma^2 c_1 + \boldsymbol{\beta}^\top (\boldsymbol{C}_2 - \boldsymbol{C}_3)\boldsymbol{\beta}$, where

$$c_1 = \text{Tr}\left[\left\{(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1} - (\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\right\}\boldsymbol{\Sigma}_{\boldsymbol{X}}\right] \tag{A.10}$$

$$\boldsymbol{C}_2 = \text{Tr}\Big[(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\boldsymbol{V})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}} \tag{A.11}$$

$$- (\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}}\Big]\left(\tfrac{\tau^2}{\nu^2}\boldsymbol{V}\right) \tag{A.12}$$

$$\boldsymbol{C}_3 = (\boldsymbol{I}_p - \boldsymbol{V})\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}(\boldsymbol{I}_p - \boldsymbol{V}) \tag{A.13}$$

. (A.14)

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

$$\frac{\boldsymbol{V}^2(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2}{(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}^2\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2} - \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1}\frac{(1-\boldsymbol{V})^2}{(\tau^2/\nu^2)\boldsymbol{V}} - 1 \tag{A.15}$$

(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 $n_{\text{A}} \gg n_{\text{B}}$, (A.15) 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 $(\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2 \approx (\boldsymbol{x}_{\text{A}}^\top \boldsymbol{x}_{\text{A}} + \boldsymbol{V}^2\boldsymbol{\Omega}_{\boldsymbol{\beta}_{\text{FRC}}}^{-1})^2$. Because $0 \leqslant \boldsymbol{V} \leqslant 1$, the entire expression is negative in this case,

and SRC is preferred to FRC for large values of $\boldsymbol{\beta}^2$. When $n_{\mathrm{B}} > n_{\mathrm{A}}$, there is no clear dominance of SRC over FRC, as the sign of (A.15) then depends on $\boldsymbol{V}$, which is in turn a function of τ^2/ν^2 and $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$.

The effect of an increasing τ^2/ν^2 on $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}) - \mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{FRC}})$ 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), $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}) - \mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{FRC}})$ is approximately $\sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}} - \boldsymbol{\beta}^2\boldsymbol{\Sigma}_{\boldsymbol{X}}$. The first expression $(\sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}})$ is attributable to $\mathrm{Var}\,\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}$ and the second term $(\boldsymbol{\beta}^2\boldsymbol{\Sigma}_{\boldsymbol{X}})$ to $\mathrm{Bias}\,\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}$. Thus, when τ^2/ν^2 is large, $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}) - \mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}) > 0 \Leftrightarrow \sigma^2(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1} > \boldsymbol{\beta}^2$. Moreover, $\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}}/n_{\mathrm{A}}$ consistently estimates $\boldsymbol{\Sigma}_{\boldsymbol{X}}$; some simplification then suggests the approximately equivalent statement $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{SRC}}) - \mathrm{MSPE}(\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}) > 0 \Leftrightarrow (n_{\mathrm{A}}+1)^{-1} > R^2$, where $R^2 = \boldsymbol{\beta}^2\boldsymbol{\Sigma}_{\boldsymbol{X}}/(\sigma^2 + \boldsymbol{\beta}^2\boldsymbol{\Sigma}_{\boldsymbol{X}})$. 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 S1) indicate that the preceding conclusions are still likely to hold in the general p case as long as $p < n_{\mathrm{A}}$. That is, the results above depend crucially on the existence of $(\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}})^{-1}$. When $p > n_{\mathrm{A}}$, as is the case in our motivating example, $p - n_{\mathrm{A}}$ eigenvalues of $\boldsymbol{x}_{\mathrm{A}}^\top \boldsymbol{x}_{\mathrm{A}} + \boldsymbol{V}\boldsymbol{\Omega}_{\hat{\boldsymbol{\beta}}_{\mathrm{FRC}}}^{-1}\boldsymbol{V}$ (appearing in the expressions for c_1 and $\boldsymbol{C}_2$) may be nearly 0 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 $\mathrm{Tr}(\boldsymbol{C}_2 - \boldsymbol{C}_3)$ 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.

B. Analysis of Hybrid Estimators

Lemmas B.1 and B.2 are used in the proof of THEOREM 3.1. 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 S1. 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 (A.10)–(A.13) (each row is averaged over 200 draws of $\boldsymbol{x}_{\mathrm{A}}$, $\boldsymbol{w}_{\mathrm{A}}$ and $\boldsymbol{w}_{\mathrm{B}}$) defined in Corollary 3.4 using the *true* value of $\boldsymbol{\theta} = \{\psi, \nu, \tau, \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\}$. In all cases, $n_{\mathrm{A}} = 50$, $\psi = 0$, $\nu = 1$, and $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{I}_p$.

Lemma B.1 Given a psd matrix $\boldsymbol{M}$ with at least one strictly positive eigenvalue and pd matrix $\boldsymbol{N}$, both of the same dimensions, $\text{Tr}(\boldsymbol{MN}) > 0$.

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,

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

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

Lemma B.2 Given estimators $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_m$, define $\boldsymbol{P}$ by (3.13) in the text, ie $P_{ij} = \text{MCPE}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j)$. 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,

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

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

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

$\boldsymbol{\Sigma}_{\boldsymbol{X}}$ 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 B.1, this is strictly positive. □

Proof. (Theorem 3.1)

(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 B.2.

(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 $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leqslant \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$. 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 $\text{MSPE}(\boldsymbol{b}(\boldsymbol{\omega}^{\text{opt}})) \leqslant \text{MSPE}(\hat{\boldsymbol{\beta}}_m)$.

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 (3.16) 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 optimal value:

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

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

Lemma B.3 Suppose we have two targeted ridge estimators, $\hat{\boldsymbol{\beta}}_{k_1} = \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta},k_1}, \lambda_{k_1}, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_1})$ and $\hat{\boldsymbol{\beta}}_{k_2} = \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}_{\boldsymbol{\beta},k_2}, \lambda_{k_2}, \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},k_2})$, as defined by (2.6). Let $\psi_\ell = \text{Tr } \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},\ell})/n_\text{A}$. If $\boldsymbol{\gamma}_{\boldsymbol{\beta},k_1}$ and $\boldsymbol{\gamma}_{\boldsymbol{\beta},k_2}$ are not functions of $\boldsymbol{y}_\text{A}$, then

$$\text{E}\left[(1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_{k_1})^\top(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_{k_2})\right] = \sigma^2 + \text{E}\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k_1})^\top \tfrac{\boldsymbol{x}_\text{A}^\top \boldsymbol{x}_\text{A}}{n_\text{A}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k_2})\right] - \sigma^2(\psi_{k_1} + \psi_{k_2}). \quad \text{(B.1)}$$

Proof. (Lemma B.3)

$$\begin{aligned}
&(1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_2) \\
&\quad = (1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta} + \boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta} + \boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_2) \\
&\quad = (1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta}) && \text{(B.2)} \\
&\qquad + (1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_1) && \text{(B.3)} \\
&\qquad + (1/n_\text{A})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_2) && \text{(B.4)} \\
&\qquad + (1/n_\text{A})(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_1)^\top(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_2) && \text{(B.5)}
\end{aligned}$$

Taking expectations, (B.2) evaluates to σ^2 and (B.5) to $\text{E}\left[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)^\top \tfrac{\boldsymbol{x}_\text{A}^\top \boldsymbol{x}_\text{A}}{n_\text{A}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\right]$. For (B.3),

$$\begin{aligned}
&(1/n_\text{A})\text{E}\left[(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}\hat{\boldsymbol{\beta}}_1)\right] \\
&\quad = (1/n_\text{A})\text{E}\left[(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{x}_\text{A}\boldsymbol{\beta} - \boldsymbol{x}_\text{A}(\boldsymbol{x}_\text{A}^\top\boldsymbol{x}_\text{A} + \lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})^{-1}(\boldsymbol{x}_\text{A}^\top\boldsymbol{y}_\text{A} + \lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1}\boldsymbol{\gamma}_{\boldsymbol{\beta},1}))\right] && \text{(B.6)} \\
&\quad = -(1/n_\text{A})\text{E}(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})\boldsymbol{y}_\text{A}) \\
&\quad = -(1/n_\text{A})\text{E}(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top(\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})\boldsymbol{y}_\text{A} - \boldsymbol{H}(\lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})\boldsymbol{x}_\text{A}\boldsymbol{\beta}) && \text{(B.7)} \\
&\quad = -(1/n_\text{A})\text{E}(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta})^\top\boldsymbol{H}(\lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})(\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta}) \\
&\quad = -\sigma^2\text{Tr } \boldsymbol{H}(\lambda_1\boldsymbol{\Omega}^{-1}_{\boldsymbol{\beta},1})/n_\text{A}
\end{aligned}$$

The equality between (B.6) and (B.7) assumes that $\boldsymbol{y}_\text{A} - \boldsymbol{x}_\text{A}\boldsymbol{\beta}$ has mean $\boldsymbol{0}_p$ and is independent of $\boldsymbol{\gamma}_{\boldsymbol{\beta},1}$. The analogous result comes from the expectation of (B.4). □

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 B.4 Let $R_\ell = \mathrm{E}[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\ell)^\top \boldsymbol{x}_\mathrm{A}^\top \boldsymbol{x}_\mathrm{A}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\ell)]$, ie the mean squared error in estimating $\boldsymbol{x}_\mathrm{A}\boldsymbol{\beta}$. This is a consistent estimate of $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_\ell)$ as n_A increases, up to the constant σ^2. A surrogate for $\mathrm{MSPE}(\hat{\boldsymbol{\beta}}_\ell)$ is $\hat{P}_{\ell,\ell}$, defined in expression (3.19); this is also equivalent to expression (2.15). 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 $\psi_\ell = \mathrm{Tr}\ \boldsymbol{H}(\lambda_\ell \boldsymbol{\Omega}_{\boldsymbol{\beta},\ell}^{-1})/n_\mathrm{A}$ 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,

$$\begin{aligned}\frac{\mathrm{E}R_\ell - \mathrm{E}\tilde{R}_\ell + \sigma^2}{\mathrm{E}R_\ell} &= \frac{\mathrm{E}R_\ell + \sigma^2 - (1-\psi)^{-2}(\mathrm{E}R_\ell + \sigma^2 - 2\sigma^2\psi_\ell)}{\mathrm{E}R_\ell} \quad \text{(from Proof of LEMMA B.3)} \\ &= \frac{-2\psi_\ell}{(1-\psi_\ell)^2} + \frac{\psi_\ell^2}{(1-\psi_\ell)^2}\frac{\mathrm{E}R_\ell + \sigma^2}{\mathrm{E}R_\ell}\end{aligned}$$

□

C. Further Simulation Study Results

Tables S2 and S3 give numerical values of empirical MSPE from Figure 2 in the main text, and S1 gives Empirical Mean Squared Error (MSE) from the same simulation study. Next, we describe simulation results under various model misspecifications.

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

S2 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 S3. 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 considered the situation where $\boldsymbol{X}$ is drawn from a multivariate t distribution with 5 degrees of freedom, scaled to maintain Var $\boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. Simultaneously, we perturbed (1.2): instead of $\text{Var}[w_{ij}|x_{ij}] = \tau^2$, the underlying true variance was $\text{Var}[w_{ij}|x_{ij}] = \tau^2|x_{ij}|^{1/4}$. These results are in Figure S4. MSPE actually decreases in this situation, and, again, HYB has MSPE that is smallest or almost the smallest in nearly scenario.

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 $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ using the shrinkage method of Schäfer and Strimmer (2005). However, that SRC does so poorly in the large p setting does not change if the true $\boldsymbol{\theta}$ is used (see Remark 2.3 in the manuscript).

We considered other values of the true $\boldsymbol{\beta}$ which 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.

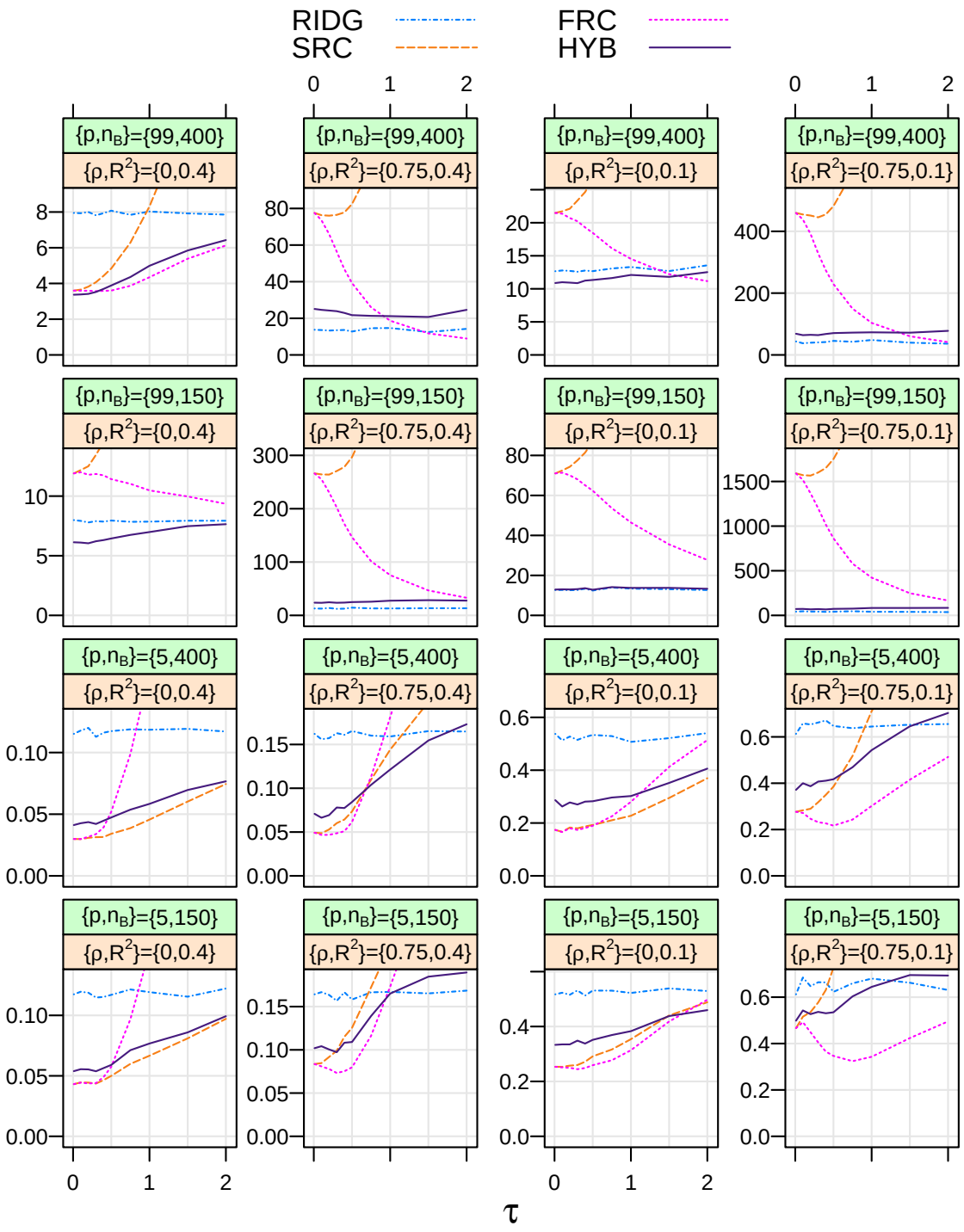

Fig. S1. Empirical Mean Squared Error (MSE) for the simulation study described in Section 4. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2}{\sigma^2}$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}} = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. The smallest possible MSE is zero.

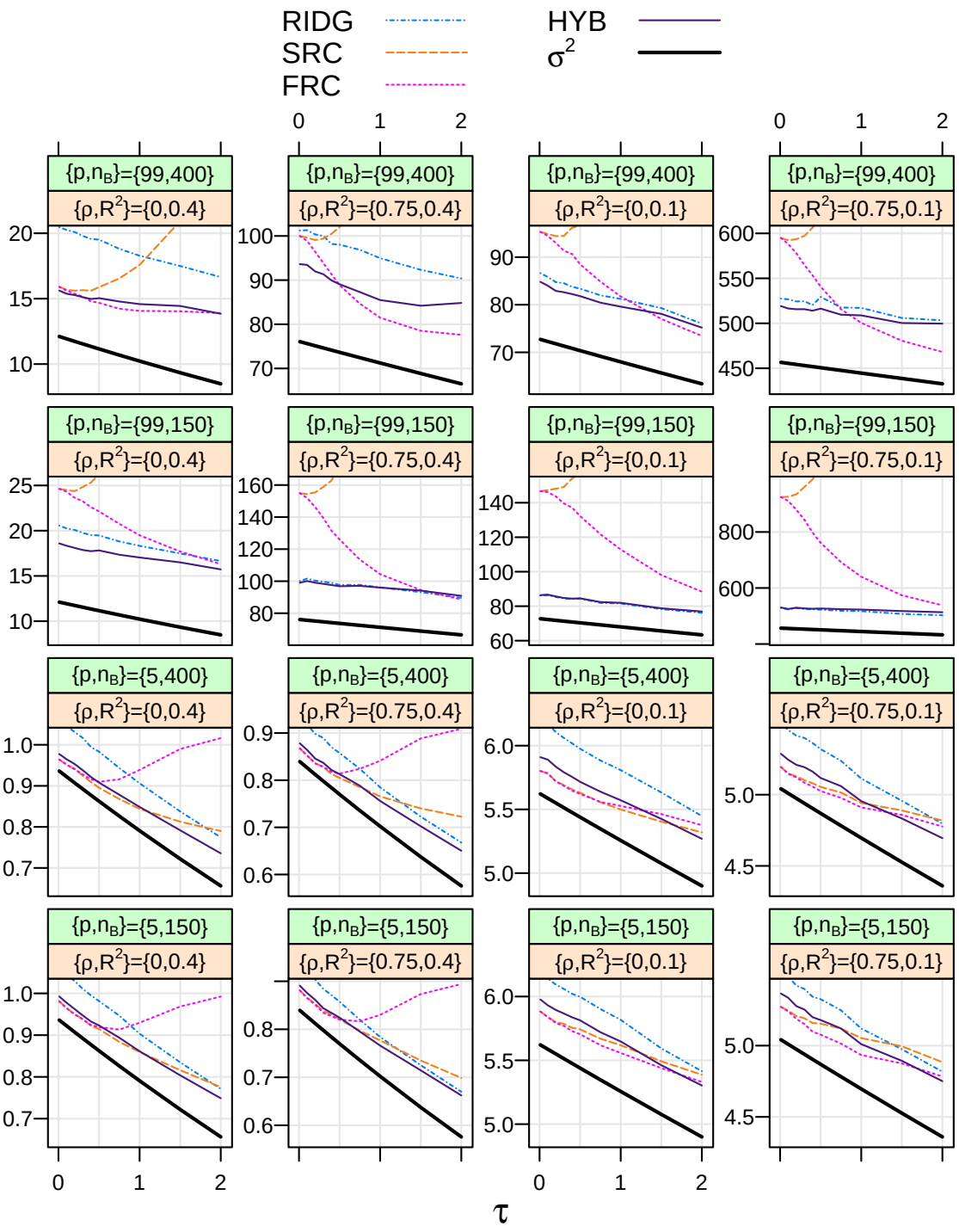

Fig. S2. Empirical MSPE over τ for the simulation study described in Section 4 **when the conditional independence assumption $[Y|\boldsymbol{X},\boldsymbol{W}]=[Y|\boldsymbol{X}]$ is violated**. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2=\frac{\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}+\sigma^2}{\sigma^2}$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}}=50$, $\beta_0=\psi=0$, and $\nu=1$. σ^2, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

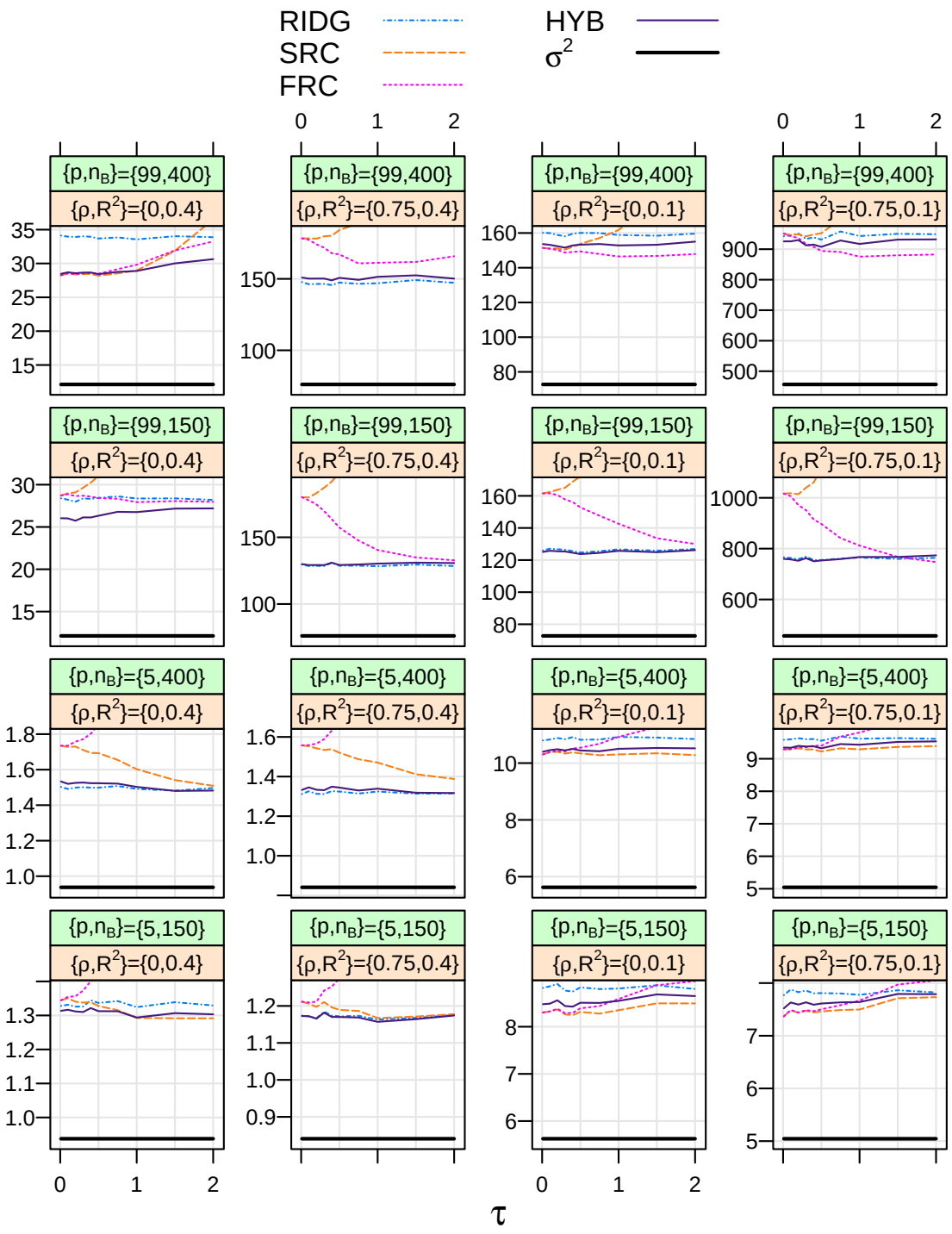

Fig. S3. Empirical MSPE over τ for the simulation study described in Section 4 **under outcome dependent sampling**. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2}{\sigma^2}$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}} = 50$, $\beta_0 = \psi = 0$, and $\nu = 1$. σ^2, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

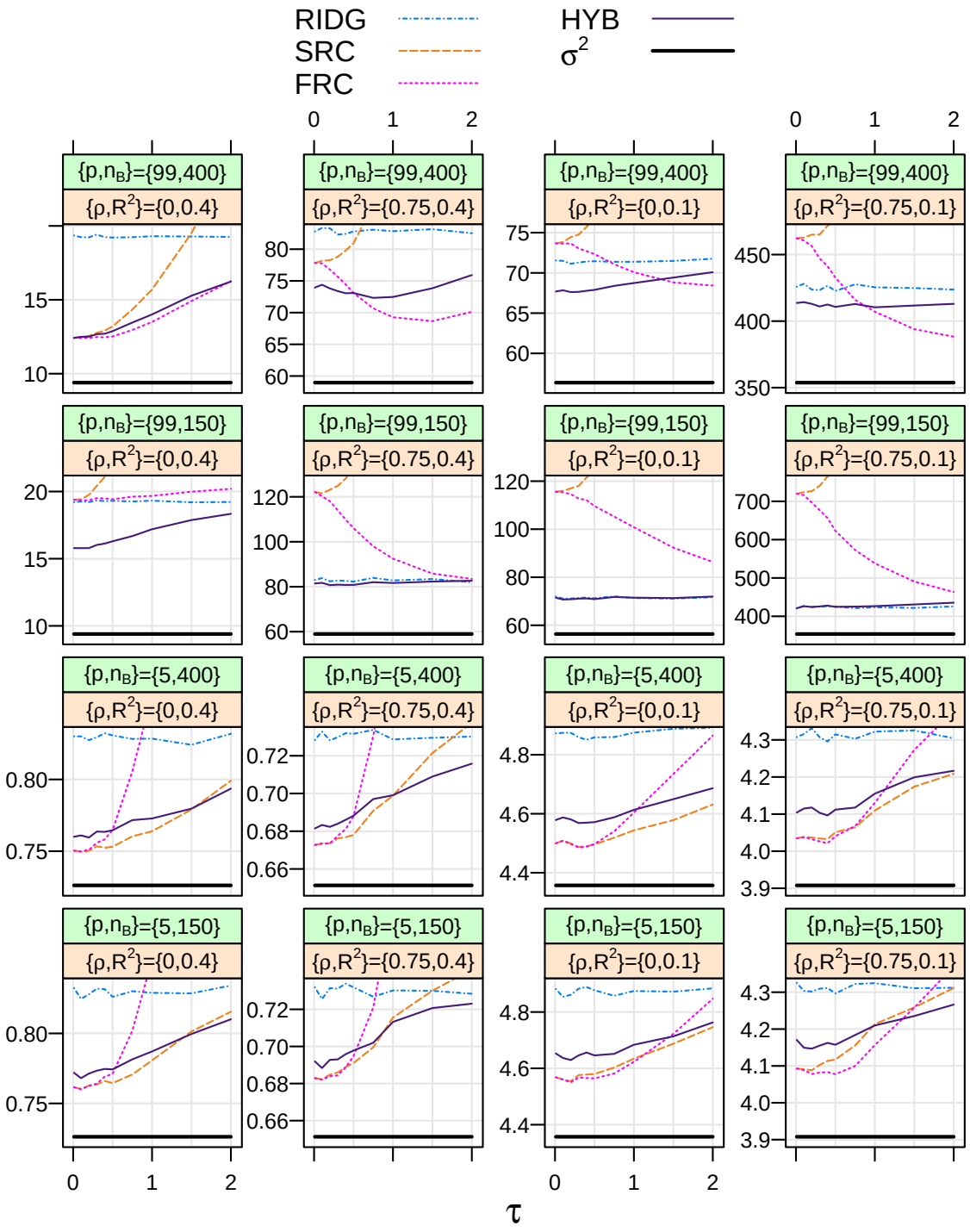

Fig. S4. Empirical MSPE over τ for the simulation study described in Section 4 **under violations to normality of $\boldsymbol{X}$ assumption and ME structure**. p stands for the number of covariates, n_{B} is the size of subsample B, ρ is the first-order auto-regressive correlation coefficient for pairwise combinations of $\boldsymbol{X}$, and $R^2 = \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta} + \sigma^2}{\sigma^2}$. The top strip varies between rows and the bottom strip varies between columns. In all cases, $n_{\mathrm{A}} = 50$, $\beta_0 = \psi = 0$, and $\nu = 1.\sigma^2$, plotted in black, is the smallest possible MSPE for any estimate of $\boldsymbol{\beta}$.

D. 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 $e^{b*} = \sqrt{\frac{r_{\mathrm{A}}^b}{r_{\mathrm{A}}^b - 1}} \left(y^{b*} - \boldsymbol{x}^{b*} \hat{\boldsymbol{\beta}}^b \right)$, 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 D.1 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 D.1 Suppose V_i is $N(0, \sigma^2)$, independently for $i = 1, \ldots, N$, and $U | V_1, \ldots, V_N \sim \mathrm{Unif}\{V_1, \ldots, V_N\}$. Then $\mathrm{E}\left[\mathrm{Var}[U | V_1, \ldots, V_N]\right] = \frac{N-1}{N}\sigma^2$.

Proof. (Theorem D.1)

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

□

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 $e^{b*} =$

$\sqrt{\frac{r_{\mathrm{A}}^b}{r_{\mathrm{A}}^b-1}}\left(y^{b*}-\boldsymbol{x}^{b*}\hat{\boldsymbol{\beta}}^b\right)$, $\mathrm{E}\left[\mathrm{Var}\left[e^{b*}\right]\right]$ is approximately σ^2, which is our justification for using e^{b*} as the prediction error.

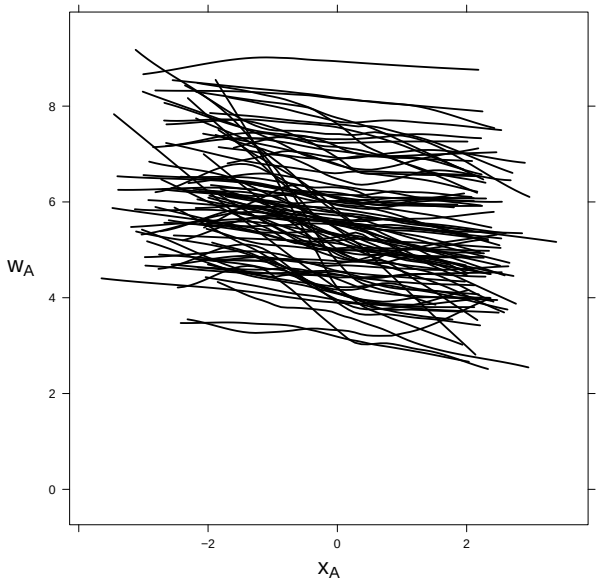

Fig. S5. LOESS curves of Affymetrix ($\boldsymbol{w}_{\mathrm{A}}$) by qRT-PCR ($\boldsymbol{x}_{\mathrm{A}}$) measurements for 91 genes from the Chen et al. (2011) data

References

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(9):1481–1487.

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

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

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612.

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(1):Article 32.

$\{\rho, R^2\}$	$\{p, n_{\mathrm{B}}\}$	method	$\tau = 0.01$	0.1	0.2	0.3	0.4	0.5	0.75	1	1.5	2
{0,0.4}	{99,400}	RIDG	20.5	20.4	20.6	20.4	20.4	20.6	20.4	20.6	20.5	20.4
		SRC	15.9	16.0	16.2	16.4	16.8	17.2	18.7	20.8	26.3	33.4
		FRC	15.9	15.9	15.9	15.9	**15.9**	**15.9**	**16.2**	**16.7**	**17.8**	**18.6**
		HYB	**15.7**	**15.7**	**15.8**	**15.9**	16.0	16.2	16.7	17.4	18.3	18.9
{0.75,0.4}	{99,400}	RIDG	102.2	101.3	101.2	101.9	101.2	100.9	102.3	102.1	100.4	102.3
		SRC	100.2	99.7	100.7	101.5	102.7	106.0	116.8	134.7	195.6	283.5
		FRC	100.2	99.0	98.0	95.8	93.4	92.3	**89.0**	**87.0**	**86.8**	**89.0**
		HYB	**94.2**	**93.4**	**93.4**	**93.1**	**92.3**	**92.1**	92.2	92.0	92.8	96.8
{0,0.1}	{99,400}	RIDG	86.0	86.3	86.1	86.1	86.3	86.1	86.6	86.7	86.2	86.9
		SRC	95.0	95.2	95.7	96.8	98.5	100.1	106.5	115.1	138.2	167.6
		FRC	95.0	94.9	94.2	93.6	92.9	91.7	89.6	87.7	85.7	**84.3**
		HYB	**84.1**	**84.4**	**84.2**	**84.2**	**84.7**	**84.6**	**85.0**	**85.4**	**85.3**	85.7
{0.75,0.1}	{99,400}	RIDG	529.5	528.3	528.6	529.5	530.6	531.7	529.9	535.1	529.2	526.0
		SRC	596.4	597.7	599.1	604.7	608.9	624.2	679.0	757.8	1009.5	1346.0
		FRC	596.3	593.5	583.2	571.9	556.1	546.9	527.0	**513.5**	**500.4**	**494.2**
		HYB	**520.8**	**520.1**	**519.3**	**520.0**	**520.0**	**520.7**	**520.2**	523.1	522.7	525.6
{0,0.4}	{99,150}	RIDG	20.5	20.5	20.3	20.4	20.4	20.5	20.4	20.4	20.6	20.5
		SRC	24.5	24.8	25.2	26.2	27.3	28.5	33.3	39.0	55.9	76.8
		FRC	24.5	24.6	24.4	24.5	24.4	24.0	23.7	23.0	22.5	21.9
		HYB	**18.5**	**18.5**	**18.5**	**18.7**	**18.8**	**18.9**	**19.3**	**19.4**	**20.0**	**20.2**
{0.75,0.4}	{99,150}	RIDG	101.2	100.5	102.0	100.1	100.7	102.4	100.5	**100.6**	**100.8**	**101.0**
		SRC	154.8	155.7	157.7	161.5	167.9	178.0	213.6	270.0	429.2	662.0
		FRC	154.8	153.2	148.4	141.8	135.6	130.2	118.7	111.8	104.5	101.0
		HYB	**100.2**	**99.7**	**100.9**	**99.4**	**99.9**	**101.0**	**100.0**	100.7	101.8	101.9
{0,0.1}	{99,150}	RIDG	**86.0**	**86.3**	**86.2**	**86.1**	**86.9**	**86.1**	**87.5**	**87.1**	**86.5**	**86.4**
		SRC	146.4	148.1	150.4	153.3	157.8	163.5	181.4	207.4	279.2	363.9
		FRC	146.4	147.0	145.9	143.4	140.7	137.7	128.4	121.2	109.5	101.7
		HYB	86.3	86.6	86.6	86.5	87.2	86.5	87.6	87.4	87.2	87.0
{0.75,0.1}	{99,150}	RIDG	**533.7**	**536.1**	**531.1**	**530.9**	**528.0**	**530.7**	**532.2**	**528.6**	**529.4**	**527.8**
		SRC	927.4	926.7	935.1	958.0	994.9	1052.8	1233.0	1533.4	2306.4	3282.8
		FRC	927.2	912.2	879.8	843.7	805.5	774.1	700.3	656.8	603.3	570.8
		HYB	535.3	536.2	531.6	532.7	530.2	533.8	534.2	536.5	538.5	539.3

Table S2. Numerical values of empirical MSPE for the simulation study described in Section 4 for $p = 99$. The smallest MSPE for each τ in each rectangle is in **bold**

$\{\rho, R^2\}$	$\{p, n_{\mathrm{B}}\}$	method	$\tau = 0.01$	0.1	0.2	0.3	0.4	0.5	0.75	1	1.5	2
{0,0.4}	{5,400}	RIDG	1.06	1.06	1.07	1.06	1.06	1.06	1.06	1.06	1.06	1.07
		SRC	0.97	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.98**	**0.98**	**1.00**	**1.02**
		FRC	**0.97**	0.97	0.97	0.97	0.98	0.99	1.04	1.11	1.25	1.35
		HYB	0.98	0.98	0.98	0.98	0.99	0.99	0.99	1.00	1.01	1.02
{0.75,0.4}	{5,400}	RIDG	0.93	0.93	0.94	0.94	0.93	0.94	0.93	0.93	0.94	0.93
		SRC	0.87	0.87	0.87	**0.87**	**0.87**	**0.88**	**0.89**	**0.90**	0.92	0.93
		FRC	**0.87**	**0.87**	**0.87**	0.87	0.88	0.89	0.94	1.00	1.14	1.22
		HYB	0.88	0.88	0.88	0.88	0.88	0.89	0.89	0.90	**0.92**	**0.92**
{0,0.1}	{5,400}	RIDG	6.19	6.17	6.19	6.18	6.18	6.20	6.17	6.15	6.18	6.19
		SRC	5.80	5.80	5.81	5.83	5.81	5.83	**5.83**	**5.85**	**5.94**	**6.02**
		FRC	**5.80**	**5.80**	**5.81**	**5.82**	**5.81**	**5.83**	5.85	5.91	6.06	6.15
		HYB	5.92	5.90	5.92	5.93	5.92	5.93	5.92	5.93	6.00	6.05
{0.75,0.1}	{5,400}	RIDG	5.50	5.51	5.50	5.51	5.50	5.50	5.50	5.51	5.51	5.50
		SRC	5.21	5.21	5.21	5.22	5.22	5.23	5.26	5.33	**5.40**	5.48
		FRC	**5.21**	**5.21**	**5.20**	**5.21**	**5.20**	**5.20**	**5.25**	**5.32**	5.43	5.52
		HYB	5.30	5.31	5.29	5.31	5.29	5.31	5.32	5.37	5.40	**5.44**
{0,0.4}	{5,150}	RIDG	1.06	1.07	1.07	1.06	1.06	1.06	1.07	1.07	1.06	1.07
		SRC	0.98	0.98	0.98	**0.98**	**0.99**	**0.99**	**1.00**	**1.01**	**1.02**	**1.04**
		FRC	**0.98**	**0.98**	**0.98**	0.98	0.99	1.00	1.04	1.10	1.22	1.32
		HYB	1.00	1.00	1.00	0.99	1.00	1.00	1.01	1.02	1.03	1.04
{0.75,0.4}	{5,150}	RIDG	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.94
		SRC	0.88	0.88	0.88	0.89	0.89	**0.89**	**0.90**	0.92	0.93	0.94
		FRC	**0.88**	**0.88**	**0.88**	**0.88**	**0.89**	0.89	0.93	0.99	1.11	1.20
		HYB	0.89	0.89	0.89	0.89	0.89	0.90	0.91	**0.91**	**0.93**	**0.93**
{0,0.1}	{5,150}	RIDG	6.16	6.19	6.18	6.19	6.14	6.18	6.18	6.17	6.19	6.18
		SRC	5.88	5.90	5.90	5.89	5.89	5.92	5.95	5.99	6.10	6.14
		FRC	**5.88**	**5.90**	**5.89**	**5.88**	**5.87**	**5.88**	**5.90**	**5.94**	**6.06**	6.14
		HYB	5.96	5.99	5.98	5.99	5.96	5.98	6.01	6.02	6.09	**6.11**
{0.75,0.1}	{5,150}	RIDG	5.49	5.56	5.50	5.51	5.51	5.50	5.50	5.50	5.49	5.50
		SRC	5.25	5.30	5.28	5.30	5.31	5.34	5.39	5.43	5.52	5.57
		FRC	**5.25**	**5.30**	**5.26**	**5.26**	**5.25**	**5.26**	**5.28**	**5.31**	**5.42**	5.50
		HYB	5.34	5.39	5.36	5.37	5.36	5.37	5.38	5.40	5.44	**5.46**

Table S3. Numerical values of empirical MSPE for the simulation study described in Section 4 for $p = 5$. The smallest MSPE for each τ in each rectangle is in **bold**