

# Post-Estimation Shrinkage in Full and Selected Linear Regression Models in Low-Dimensional Data Revisited

Edwin Kipruto  | Willi Sauerbrei 

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

**Correspondence:** Edwin Kipruto ([edwin.kipruto@uniklinik-freiburg.de](mailto:edwin.kipruto@uniklinik-freiburg.de))

**Received:** 5 December 2023 | **Revised:** 9 July 2024 | **Accepted:** 22 July 2024

**Funding:** This study was supported by Deutsche Forschungsgemeinschaft (Grant SA580/10).

**Keywords:** best subset selection | lasso | post-estimation shrinkage | prediction errors | simulation study

## ABSTRACT

The fit of a regression model to new data is often worse due to overfitting. Analysts use variable selection techniques to develop parsimonious regression models, which may introduce bias into regression estimates. Shrinkage methods have been proposed to mitigate overfitting and reduce bias in estimates. Post-estimation shrinkage is an alternative to penalized methods. This study evaluates effectiveness of post-estimation shrinkage in improving prediction performance of full and selected models. Through a simulation study, results were compared with ordinary least squares (OLS) and ridge in full models, and best subset selection (BSS) and lasso in selected models. We focused on prediction errors and the number of selected variables. Additionally, we proposed a modified version of the parameter-wise shrinkage (PWS) approach named non-negative PWS (NPWS) to address weaknesses of PWS. Results showed that no method was superior in all scenarios. In full models, NPWS outperformed global shrinkage, whereas PWS was inferior to OLS. In low correlation with moderate-to-high signal-to-noise ratio (SNR), NPWS outperformed ridge, but ridge performed best in small sample sizes, high correlation, and low SNR. In selected models, all post-estimation shrinkage performed similarly, with global shrinkage slightly inferior. Lasso outperformed BSS and post-estimation shrinkage in small sample sizes, low SNR, and high correlation but was inferior when the opposite was true. Our study suggests that, with sufficient information, NPWS is more effective than global shrinkage in improving prediction accuracy of models. However, in high correlation, small sample sizes, and low SNR, penalized methods generally outperform post-estimation shrinkage methods.

## 1 | Introduction

Regression modeling is a statistical tool with important practical applications in many fields. The choice of a regression model depends on the aim of the study. For instance, in prediction, a model that includes some noise variables may be acceptable, whereas in descriptive models, a simple model is preferred (Shmueli 2010). For the normal-errors regression models, the Gauss–Markov Theorem states that under certain conditions, the ordinary least squares (OLS) estimator is the best linear unbiased estimator. However, the OLS estimator can exhibit

higher variability in instances with high correlation between covariates, low signal-to-noise ratio (SNR), or small sample size. The latter often results in overfitting, which can lead to poor predictions on new data (Copas 1983; Copas and Long 1991; James et al. 2013; Riley et al. 2021). Several shrinkage methods, which aim to reduce the variance of estimates by introducing a (small) bias, have been proposed to mitigate overfitting. Variable selection using stepwise selection methods can lead to biased regression estimates for selected variables, and shrinkage has also been proposed to reduce this bias (Copas and Long 1991; Miller 2002; Harrell 2015).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Biometrical Journal* published by Wiley-VCH GmbH.

In this study, shrinkage methods are categorized into penalized regression and post-estimation shrinkage methods to avoid confusion. In penalized regression methods, the estimation and shrinkage of regression coefficients are conducted simultaneously. The lasso (Tibshirani 1996) and ridge (Hoerl and Kennard 1970) are two popular penalized methods that will be considered. On the other hand, post-estimation shrinkage is a two-stage procedure. In the first stage, the regression coefficients of a given model (full or selected) are estimated using OLS or standard maximum likelihood estimation method. In the second stage, the shrinkage factors are estimated and used to calculate shrunken regression estimates (van Houwelingen and le Cessie 1990; Sauerbrei 1999). Post-estimation shrinkage methods are a simple alternative to penalized regression methods due to their simplicity and reliance on OLS or standard maximum likelihood estimates (MLE), obviating the need for specialized software. However, these methods cannot be used in high-dimensional data due to unavailability of the OLS or MLE (van Houwelingen and le Cessie 1990; van Houwelingen and Sauerbrei 2013; Dunkler, Sauerbrei, and Heinze 2016).

At least three post-estimation shrinkage approaches have been proposed: global shrinkage (van Houwelingen and le Cessie 1990), parameter-wise shrinkage (PWS; the abbreviation PWSF (Parameterwise Shrinkage Factors) has been used in the literature) (Sauerbrei 1999), and joint shrinkage (JS) (Dunkler, Sauerbrei, and Heinze 2016). Global shrinkage uniformly applies shrinkage to all regression estimates, whereas PWS applies different levels of shrinkage to regression estimates depending on their magnitude. JS is an extension of PWS that is used when variables are structurally grouped, such as dummy variables for a multilevel categorical covariate. A common shrinkage factor is estimated for regression estimates within a group (Dunkler, Sauerbrei, and Heinze 2016). In our study, JS is not applicable as there are no groups of variables, and all covariates are continuous and linearly related to the outcome or have no effect.

In a simulation study, van Houwelingen and Sauerbrei (2013) evaluated the prediction performance of global and PWS in selected models using backward elimination. They reported that PWS gave better predictions than global shrinkage that were comparable to the lasso. However, their evaluation was limited to four scenarios with low correlation and moderate-to-high SNR or  $R^2$  (50% and 71%). In our study, we consider a broader range of simulation scenarios, including high correlation, small sample sizes, and low SNR.

Breiman (1995) proposed the non negative garrote (NNG) and another method that uses a quadratic penalty term on the shrinkage factors, referred to here as quadratic PWS (QPWS). The NNG has been well-studied in the literature (Yuan and Lin 2007; Xiong 2010; Kipruto and Sauerbrei 2022b), but QPWS seems to have been largely overlooked and will be evaluated in our study. The shrinkage behavior of QPWS is similar to PWS, making a comparison between the two approaches necessary.

In this study, we will investigate and compare the shrinkage behavior of post-estimation shrinkage methods in the context of classical linear regression models for low-dimensional data. Our aim is to investigate whether post-estimation shrinkage improves

the predictive accuracy of full and selected models. Furthermore, we propose a modified version of PWS called non-negative PWS (NPWS). The NPWS differs from PWS in that it imposes non-negativity constraints on the shrinkage factors and can be used in a full model, whereas PWS was originally proposed for selected models only. We will investigate whether the non-negativity constraint is useful in improving the prediction performance of PWS.

In full models, we will compare the prediction performance of OLS, post-estimation shrinkage, and ridge (reference model), whereas in selected models, we will compare the prediction performance of best subset selection (BSS), post-estimation shrinkage, and lasso (reference model). All covariates in the training dataset will be standardized to have a mean of zero and a unit variance, and the response variable will be centered to exclude the intercept from the regression model. Additionally, the covariates in the new dataset will also be standardized using the statistics derived from the training data.

Section 2 describes our simulation study following the ADEMP structure, which entails defining aims (A), data-generating mechanisms (D), estimands/targets of analysis (E), and performance (P) measures (Morris, White, and Crowther 2019), whereas Section 3 describes our methods, categorized into three groups: post-estimation shrinkage, penalized regression, and classical variable selection methods. We also provide a detailed explanation of methods for estimating post-estimation shrinkage factors using cross-validation (CV). Section 4 presents our results, categorized into two sets: full and selected models. Finally, Section 5 consists of a discussion and conclusion.

## 2 | Simulation Design

Our simulation study follows the relevant parts of the simulation protocol of Kipruto and Sauerbrei (2022a). However, we decided to separate investigations of post-estimation shrinkage from the broader comparison of variable selection procedures to better understand the former. We also made minor modifications to the original protocol to enable us to investigate situations that were not captured in the original protocol. These changes are discussed in Section 2.1.

We provide a brief description and refer interested readers to the protocol paper for further details. Table A1 shows the summary of the simulation design from the protocol (Kipruto and Sauerbrei 2022a). The corresponding R code is provided at <https://github.com/EdwinKipruto/shrinkage>. Table 1 provides the summary of our simulation study following the ADEMP structure. The aims and target of analysis are clearly stated in Table 1. Thus, we focus our discussion on data generating mechanisms (Sections 2.1), performance measures (Section 2.2), and methods (Section 3) in detail.

### 2.1 | Data Generating Mechanisms

We generated 2000 training datasets per scenario, each consisting of a continuous response variable ( $y$ ) and 15 continuous covariates ( $X$ ). The number of simulation repetitions was set to 2000 to ensure that the Monte Carlo standard error (MCSE) of the model

**TABLE 1** | Simulation design following ADEMP structure.

Aims	<ul style="list-style-type: none"> <li>To compare the shrinkage behavior of post-estimation shrinkage approaches in both full and selected models</li> <li>To evaluate the effectiveness of post-estimation shrinkage approaches in improving prediction accuracy of both full and selected models, and to compare their performance with ridge and lasso as reference models</li> <li>To improve the performance of PWS approach in full models</li> </ul>		
Data generating mechanism (Section 2.1)	<p><b>Training/development dataset</b></p> <ul style="list-style-type: none"> <li><math>X \sim N_p(0, \Sigma)</math> where <math>p = 15</math> and <math>\Sigma \in \mathbb{R}^{p \times p}</math>; <math>\Sigma_{ij} = \rho^{ i-j }</math> for <math>\rho = (0.3, 0.8)</math> indicating low and high correlation, respectively</li> <li><math>Y = X\beta_T + \epsilon</math> where <math>\beta_T \in (\beta_A, \beta_D)</math> and <math>\epsilon \sim N(0, \sigma^2 I_n)</math></li> </ul> <p>True regression coefficients (<math>\beta_T</math>) for 15 covariates  <math>\beta_A</math>: 1.5, 0, 1, 0, 1, 0, 0.5, 0, 0.5, 0, 0.5, 0, -0.5, 0, 0  <math>\beta_D</math>: 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0</p> <p><b>SNR/<math>R^2</math> and sample size (<math>n</math>)</b>  <math>R^2 = \{0.11, 0.20, 0.50, 0.67, 0.80, 0.86\}</math>  <math>n = \{50, 100, 400\}</math></p> <p><b>Number of scenarios (full factorial design) and simulation runs</b>  <math>\beta \times \rho \times R^2 \times n = 2 \times 2 \times 6 \times 3 = 72</math> scenarios  <math>N = 2000</math> simulation repetitions per scenario</p> <p><b>Test dataset</b></p> <ul style="list-style-type: none"> <li>New simulations with the same design as training dataset (<math>n_{\text{test}} = 100,000</math>)</li> </ul> <p><b>Additional analysis for full models</b>  We will use <math>n, \beta_A, \rho, R^2 = (0.11, 0.50, 0.67, 0.80, 0.86)</math> with 30 covariates (15 original covariates and 15 additional noise variables)</p>		
Target of analysis	<ul style="list-style-type: none"> <li>Shrinkage factors for each regression estimate</li> <li>Selection status of each covariate</li> <li>Model prediction errors</li> </ul>		
Methods (Section 3)	Method	Tuning parameters	Initial estimates
	OLS	N/A	N/A
	Global	10-fold CV	N/A
	PWS	10-fold CV	N/A
	NPWS	10-fold CV	N/A
	QPWS	10-fold CV	OLS
	Ridge	10-fold CV	N/A
	Lasso	10-fold CV	N/A
	BSS	10-fold CV	N/A
Performance measures (Section 2.2)	<ul style="list-style-type: none"> <li>Prediction accuracy: RTE and RR</li> </ul>		

Note: Changes made to the original protocol are in bold.

Abbreviations: BSS, best subset selection; CV, cross-validation; NPWS, non-negative parameter-wise shrinkage; OLS, ordinary least square; PWS, parameter-wise shrinkage; QPWS, quadratic parameter-wise shrinkage; RR, relative risk; RTE, relative test error; SNR, signal-to-noise ratio.

error was smaller than 0.005 for better precision (see Kipruto and Sauerbrei (2022a) for details). The  $X$  matrix was generated by sampling from a multivariate normal distribution with a mean vector of zero and covariance matrix  $\Sigma$ , where  $\Sigma_{ij} = \rho^{|i-j|}$  for  $\rho = (0.3, 0.8)$ . Each covariate in  $X$  was standardized to have a mean of zero and unit variance. For  $y$ , we used the formula  $y = X\beta_T + \epsilon$ , where  $\beta_T$  is a vector of true regression coefficients (see Table 1). The error term  $\epsilon_i$  was generated from a normal distribution with a mean of zero and variance of  $\sigma^2$ . The value of  $\sigma^2$  was chosen to achieve the desired SNR (i.e.,  $\sigma^2 = \beta_T^T \Sigma \beta_T / \text{SNR}$ ).

We selected a subset of values for SNR (0.25, 1, and 4) and correlation ( $C2$  and  $C3$ ) from the protocol. Additionally, we added a sample size of  $n = 50$  and SNR values of 0.12, 2, and 6, corresponding to  $R^2$  of about 11%, 67%, and 86%, respectively (as highlighted in Table 1). This enabled us to investigate the performance of methods in small sample sizes and at both low and high SNR levels. To evaluate the performance of each prediction model, we generated an independent test dataset with a sample size of 100,000 using the same design as the training data.

## 2.2 | Performance Measures

We used two performance measures to evaluate the prediction performance of models. These measures are as follows.

### 2.2.1 | Relative Test Error (RTE)

The RTE quantifies the expected test error relative to the Bayes error rate (Hastie, Tibshirani, and Tibshirani 2020). The RTE is simply a standardized form of mean squared prediction error and is defined as follows:

$$\begin{aligned} \text{RTE}(\hat{\beta}) &= \frac{E[(y_0 - X_0^T \hat{\beta})^2]}{\sigma^2} = \frac{(\hat{\beta} - \beta_T)^T \Sigma (\hat{\beta} - \beta_T) + \sigma^2}{\sigma^2} \\ &= \frac{ME + \sigma^2}{\sigma^2}, \end{aligned}$$

where  $X_0$  denotes a random matrix of test covariates,  $y_0$  is a random vector of test response variable,  $\beta_T$  is a vector of true coefficients, and  $\hat{\beta}$  is a vector of estimated coefficients from a regression method such as OLS.  $ME = (\hat{\beta} - \beta_T)^T \Sigma (\hat{\beta} - \beta_T)$  denotes the model error. A null model (model without predictors) has an RTE score of  $(\beta_T^T \Sigma \beta_T + \sigma^2) / \sigma^2 = \text{SNR} + 1$ , whereas a perfect model ( $\hat{\beta} = \beta_T$ ) has a score of 1 as  $ME = 0$ . A good prediction model should have an RTE close to 1.

### 2.2.2 | Relative Risk (RR) in Prediction Model

RR is another metric used to measure the prediction accuracy of models, as employed by Hastie, Tibshirani, and Tibshirani (2020) and Bertsimas, King, and Mazumder (2016). It is defined as follows:

$$\begin{aligned} \text{RR}(\hat{\beta}) &= \frac{E[(X_0^T \hat{\beta} - X_0^T \beta_T)^2]}{E(X_0^T \beta_T)^2} = \frac{(\hat{\beta} - \beta_T)^T \Sigma (\hat{\beta} - \beta_T)}{\beta_T^T \Sigma \beta_T} \\ &= \frac{ME}{\beta_T^T \Sigma \beta_T}. \end{aligned}$$

A null and a perfect model have scores of 1 and 0, respectively. A good prediction model should have an RR close to 0. If a model has an  $\text{RR} > 1$ , its prediction accuracy is worse than a null model.

## 3 | Methods

The methods include post-estimation shrinkage (global, PWS, and NPWS), penalized regression (lasso, ridge, and QPWS), and BSS (with and without shrinkage). Tenfold CV will be used to select the optimal tuning parameters for all procedures, except for OLS, which does not require tuning parameters.

### 3.1 | Post-Estimation Shrinkage Methods

#### 3.1.1 | Global Shrinkage

The global shrinkage method was proposed to improve the prediction accuracy of regression models on new data (van

Houwelingen and le Cessie 1990). This is achieved by uniformly adjusting all regression estimates using a shrinkage factor ( $c$ ), where the estimates are obtained from least-squares or standard maximum likelihood estimation. For ordinary linear models, the adjusted linear predictor is as follows:

$$\hat{y}_i = \hat{\beta}_0^* + c\hat{\beta}_1 x_{i1} + \dots + c\hat{\beta}_p x_{ip}.$$

The intercept  $\hat{\beta}_0^*$  is estimated after obtaining the shrunken estimates. A limitation of global shrinkage is its uniform shrinkage of both small and large coefficients. This can lead to over-shrinkage of large coefficients which hardly need shrinkage, resulting in excessive bias that can be detrimental to prediction. The global shrinkage can be applied to both full and selected models (van Houwelingen and Sauerbrei 2013). The shrinkage factor will be estimated using 10-fold CV (Section 3.1.4).

#### 3.1.2 | Parameter-Wise Shrinkage

PWS approach is an extension of global shrinkage (Sauerbrei 1999), in which each regression estimate from a selected model is shrunk differently. For ordinary linear models, the adjusted linear predictor is as follows:

$$\hat{y}_i = \hat{\beta}_0^{**} + c_1 \hat{\beta}_1 x_{i1} + \dots + c_p \hat{\beta}_p x_{ip},$$

where  $c_j$  is a shrinkage factor for the  $j$ th predictor, and  $\hat{\beta}_0^{**}$  is the intercept estimated after obtaining the shrunken estimates. When very weak effects or noise variables are included in the regression model, the PWS approach often estimates negative shrinkage factors for their regression estimates, indicating that even the signs of the estimates may be wrong. For this reason, PWS is recommended for use after model selection, as most noise variables and variables with very weak effects would have been eliminated, reducing the probability of a “wrong” sign (Sauerbrei 1999). The shrinkage factors ( $\hat{c}_j$ ) will be estimated using 10-fold CV (Section 3.1.4).

#### 3.1.3 | Non-Negative PWS

In regression modeling, especially in uncorrelated settings, shrinkage factors are typically expected to range between zero and one (Breiman 1995). However, negative shrinkage factors can occur when no restriction is imposed on the shrinkage factors, as is the case with the PWS. This study introduces a modified version of the PWS approach, named the NPWS, in which the shrinkage factors are constrained to be non-negative ( $c_j \geq 0$ ). The NPWS estimator preserves the sign of regression estimates and can set estimates of noise variables to zero. The non-negativity constraint enables the application of NPWS to both full and selected models. The shrinkage factors will be estimated using 10-fold CV in conjunction with a constrained least-squares approach (see Section 3.1.4 for more details).

#### 3.1.4 | Estimation of Global and PWS Factors

Leave-one-out CV (LOOCV) was proposed for estimation of global and PWS shrinkage factors (van Houwelingen and le

Cessie 1990; Sauerbrei 1999; Verweij and van Houwelingen 1993). This approach requires fitting a statistical model  $n$  times, where  $n$  is the number of observations, which can be computationally intractable in simulation studies. Here, we used 10-fold CV which involved fitting a statistical model only 10 times, making it more practical for large problem sizes (Kipruto and Sauerbrei 2022b).

Before estimating global or PWS shrinkage factors, it is necessary to specify the outcome  $y$  and covariates  $x_1, \dots, x_d$  for the final model. These covariates may have been selected from a larger set using either subset selection methods or subject matter knowledge (Dunkler, Sauerbrei, and Heinze 2016).

The global shrinkage factors are estimated using  $k$ -fold CV as follows:

1. Randomly divide the set of observations into  $k$  folds of approximately equal size.
2. Hold out the first fold and fit a regression model on the remaining  $k - 1$  folds, resulting in a column vector of regression coefficients  $\hat{\beta}^*$  of dimension  $d \times 1$ . Compute the linear predictor,  $\eta_1^* = x^* \hat{\beta}^*$ , on the observations in the held-out fold, where  $x^*$  denotes the matrix of covariates in the held-out fold.
3. Repeat Step 2 for each fold, holding out a different group of observations each time. This process results in  $k$  sets of linear predictors  $\eta_1^*, \eta_2^*, \dots, \eta_k^*$ .
4. Combine these linear predictors row-wise to form a single matrix  $\eta$  of dimension  $n \times 1$ . Use  $\eta$  as the covariate and  $y$  as the outcome to fit a linear regression model  $E(y|\eta) = c\eta$ , where  $\hat{c}$  is the regression estimate and represents the global shrinkage factor.

On the other hand, the PWS shrinkage factors are estimated by modifying Steps 2 and 4 of the above procedure (Dunkler, Sauerbrei, and Heinze 2016) as follows:

2. Instead of using a single linear predictor  $\eta_1^* = x^* \hat{\beta}^*$ , compute partial linear predictor  $\eta_{1j}^* = x_j^* \hat{\beta}_j^*$  for  $j = 1, \dots, d$ , where  $x_j^*$  denotes a vector of values in the held-out fold for the  $j$ th covariate. This yields a matrix  $\eta_1^* = [\eta_{11}^*, \dots, \eta_{1d}^*]$ .
4. Combine these partial linear predictors row-wise to form a single matrix  $\eta = [\eta_1, \eta_2, \dots, \eta_d]$  of dimension  $n \times d$ . Fit a linear regression model  $E(y|\eta_1, \dots, \eta_d) = c_1 \eta_1 + \dots + c_d \eta_d$ , where  $\eta_j$  represents the column vector of partial predictors for the  $j$ th covariate, and  $c_1, \dots, c_d$  are the regression estimates representing the PWS shrinkage factors.

The final stage (Stage 4) of estimating shrinkage factors involves fitting a regression model with the same response variable ( $y$ ) and (partial) linear predictors ( $\eta$ ) as independent variables. The resulting regression estimates represent the shrinkage factors. Negative shrinkage factors may occur because the regression estimates of partial linear predictors are not constrained to be non-negative, as demonstrated in the application of PWS to full models (Sauerbrei 1999). The NPWS approach follows the same procedure for estimating PWS factors, but with an additional constraint, the regression estimates of partial linear predictors in Stage 4 must be non-negative, which is equivalent to fitting a non-

negative least squares model. This can be achieved using software supporting the estimation of non-negative regression estimates, such as the *nnls* package (Mullen and van Stokkum 2023) or *glmnet* package (with tuning parameter and lower bounds of the coefficients set to zero) (Friedman, Tibshirani, and Hastie 2010) in R.

## 3.2 | Best Subset Selection

The BSS is a traditional approach for variable selection that identifies the best fitting model of each number of variables included. Exhaustive search or leaps and bounds algorithms are often used (Lumley 2020). The best model in OLS models is a model with the smallest residual sum of squares (James et al. 2013). A 10-fold CV was used to choose the best fitting subset of variables. It is well-known that variable selection can result in biased regression estimates that need shrinkage (Miller 2002; Harrell 2015). Without shrinkage, the resulting model may not generalize well to new data (Copas and Long 1991; Riley et al. 2021; Miller 2002). We applied shrinkage to the regression estimates of BSS and compared results without shrinkage. The shrinkage factors were obtained from global, PWS, NPWS, and QPWS approaches.

## 3.3 | Penalized Regression Methods

### 3.3.1 | Ridge and Lasso

Penalized regression methods such as ridge (Hoerl and Kennard 1970) and lasso (Tibshirani 1996) are alternative approaches to post-estimation shrinkage. Ridge and lasso have been extensively studied in the literature, and we will briefly discuss their key concepts. Ridge regression was proposed to address issues of multicollinearity, whereas lasso was proposed to combine variable selection and shrinkage. For ordinary linear regression models, the ridge ( $\alpha = 0$ ) and lasso ( $\alpha = 1$ ) estimators are obtained by minimizing:

$$\frac{1}{2n} \sum_{i=1}^n \left( y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \frac{\lambda}{2} (1 - \alpha) \sum_{j=1}^p \beta_j^2 + \lambda \alpha \sum_{j=1}^p |\beta_j|, \lambda \geq 0,$$

where  $\lambda$  is the tuning parameter that controls the amount of shrinkage that is applied to the regression coefficients. When  $\lambda = 0$ , the penalty term has no effect and the estimates of ridge, lasso, and OLS are identical. When  $\lambda \rightarrow \infty$ , the lasso estimates are zero, whereas ridge estimates approach zero but not exactly zero (James et al. 2013). The tuning parameters  $\lambda$  will be estimated using a 10-fold CV by minimizing mean squared error.

### 3.3.2 | Quadratic Penalty on Shrinkage Factors

Breiman (1995) proposed a method for estimating shrinkage factors that uses a quadratic penalty term on the shrinkage factors. This method can produce negative shrinkage factors for

noise variables in correlated settings, but in uncorrelated settings, all shrinkage factors are non-negative based on the formula provided by Breiman (1995). To avoid negative shrinkage factors, we imposed non-negativity constraints on the shrinkage factors, similar to the NNG (Breiman 1995). As a result, the shrinkage factors for noise variables can be zero, indicating that the variable should be eliminated. We refer to this approach as the QPWS and will compare its performance to PWS, NPWS, and global shrinkage when applied to full and selected models.

Let  $\hat{\beta}_j^{\text{OLS}}$  for  $j = 1, \dots, p$  be the OLS estimate for the  $j$ th variable from the full or selected model. The shrinkage factors  $\hat{c}(\lambda) = (\hat{c}_1, \dots, \hat{c}_p)$  are obtained by minimizing the objective function:

$$\frac{1}{2n} \sum_{i=1}^n \left( y_i - \sum_{j=1}^p c_j \hat{\beta}_j^{\text{OLS}} x_{ij} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^p c_j^2, \quad c_j \geq 0, \quad \lambda \geq 0,$$

where  $\lambda$  is the tuning parameter that controls the amount of shrinkage applied to the OLS initial estimate. When  $\lambda = 0$ , the penalty term has no effect and all shrinkage factors are equal to 1, whereas when  $\lambda \rightarrow \infty$ , the impact of the penalty term is high and all shrinkage factors are equal to zero. The shrunken regression estimate for the  $j$ th variable is given by  $\hat{\beta}_j(\lambda) = \hat{c}_j \hat{\beta}_j^{\text{OLS}}$ . The tuning parameters will be estimated using a 10-fold CV by minimizing mean squared error.

### 3.4 | Notations

We introduce the notations used in Section 4. The full OLS model, followed by global shrinkage, PWS, NPWS, and QPWS approaches, is denoted as Global (F), PWS (F), NPWS (F), and QPWS (F), respectively. Here, ‘‘F’’ enclosed in the brackets denotes the full model. When post-estimation shrinkage is applied after model selection, ‘‘F’’ is replaced by ‘‘S,’’ where S denotes the selected model.

## 4 | Results

This section describes the key findings from our simulation study. A subset of the results is presented, with a focus on scenarios where the true regression coefficients follow beta-type A ( $\beta_A$ ) distribution. Additionally, a summary of all the simulation results is provided in Supporting Information Appendix section. The results are presented in two sets: full models (Section 4.1) and selected models (Section 4.2).

### 4.1 | Full Models

#### 4.1.1 | Comparison of Post-Estimation Shrinkage Factors in Low and High Correlation

Figure 1 compares the shrinkage factors for post-estimation shrinkage approaches in low-correlated settings across different SNR levels. The left and right panels show the average shrinkage factors with one standard error band for regression estimates of signal and noise variables, respectively. By design, global shrinkage factors are identical for all variables in a given dataset, regardless of whether they are signal or noise.

For signal variables (left panel), three key findings are evident. First, shrinkage is higher in low SNR (top left) and decreases as SNR increases across all approaches. Second, in low SNR, global shrinkage tends to shrink large effects ( $x_1$ ) more and weak effects (e.g.,  $x_{13}$ ) less. Third, in low SNR, the PWS on average estimates negative shrinkage factors for weak effects ( $x_7, x_9, x_{11}$ , and  $x_{13}$ ), shifting to positive values as the SNR increases.

For noise variables (right panel), negative shrinkage factors are estimated by PWS approach, both in low and high SNR, a feature that is precluded in NPWS and QPWS approaches due to the non-negativity constraints. The global shrinkage nearly always estimated positive shrinkage factors, but negative shrinkage factors were estimated in a few replications in low SNR (Figure A1). The shrinkage behavior of NPWS and QPWS is similar.

In high correlation (Figure A2), both NPWS and QPWS tend to apply more shrinkage to weak and no effects than global shrinkage, whereas PWS generally estimates negative shrinkage factors for weak effects, except in high SNR.

#### 4.1.2 | Comparison of Prediction Errors

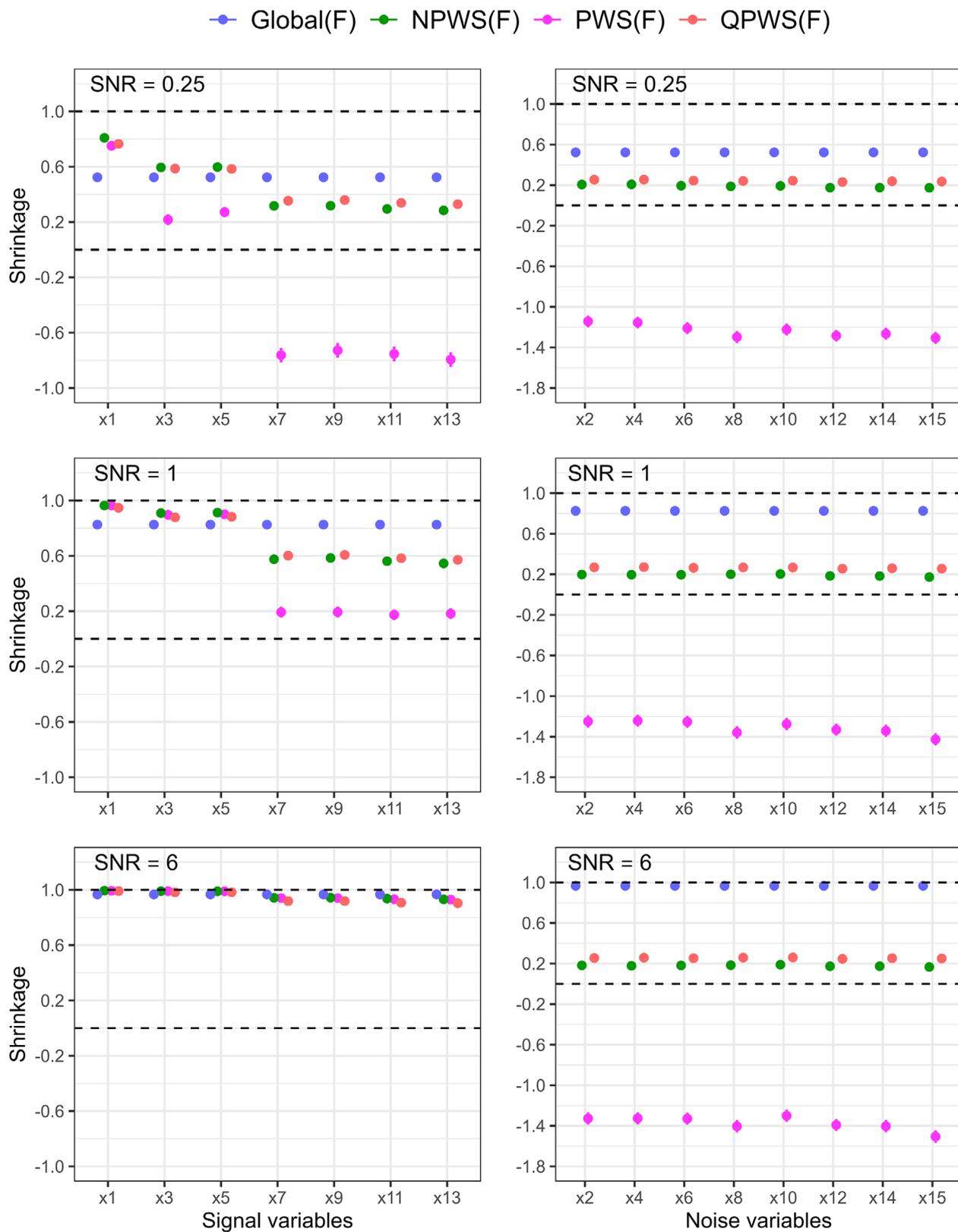
**4.1.2.1 | Effects of SNR on Prediction.** Figure 2 compares the prediction of post-estimation shrinkage methods, OLS, and ridge using RR and RTE metrics, where the latter magnifies the differences. The upper and lower panels are low and high correlation settings, respectively. The average RR and RTE are displayed for each method. The results for NPWS and QPWS were similar, and only NPWS results are reported. Overall, post-estimation shrinkage methods (except PWS) improved the prediction accuracy of OLS models. From RR, we observe that prediction performance improves as SNR increases in all methods, regardless of the amount of correlation.

In low correlation, we observe the following: In low SNR, all shrinkage approaches, except PWS, perform similarly and outperform OLS. In high SNR, NPWS outperforms all other shrinkage approaches, whereas global shrinkage, ridge, and OLS exhibit similar performance (Figures 2 and A3, top right panel). The OLS generally outperformed PWS, likely due to the impact of negative shrinkage factors.

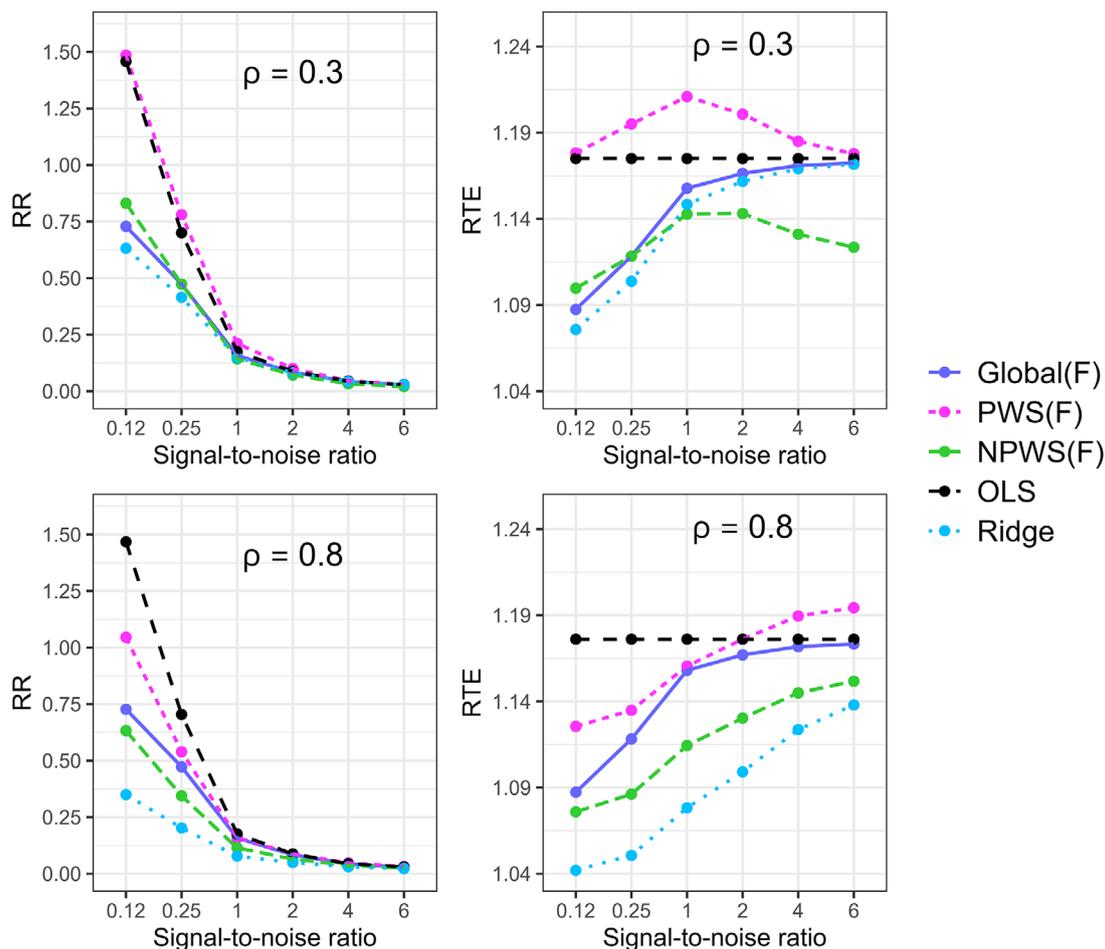
In high correlation, post-estimation shrinkage methods improved prediction accuracy of OLS, particularly in low-to-moderate SNR. Ridge outperformed post-estimation shrinkage methods, whereas NPWS outperformed global shrinkage. As SNR increases, the prediction performance of global shrinkage converges toward that of OLS models, indicating that the shrinkage factors approach one, as illustrated in Figure 1.

**4.1.2.2 | Effects of Sample Size on Prediction.** Figures 3 and A4 shows the relationship between RTE and sample size in low (left panel) and high (right panel) correlation under varying levels of SNR. The upper and lower panels show results for low (0.12) and moderate (1) SNR, respectively. We omitted the results of PWS as it was often inferior to the OLS model.

As sample size increased, we observed that the RTE values of all methods approached nearly perfect accuracy of 1, regardless of



**FIGURE 1** | Full model. Average shrinkage factors with one standard error band (not visible due to small standard errors) for post-estimation shrinkage approaches in low correlation ( $\rho = 0.3$ ) with  $p = 15$ ,  $n = 100$  and beta-type  $A$  ( $\beta_A$ ) for different SNR levels. The average shrinkage factors over 2000 replications for signal (left panel) and noise (right panel) variables are displayed. SNR, signal-to-noise ratio.



**FIGURE 2** | Full model. Average RR and RTE over 2000 replication for methods in low (upper panel) and high correlation (lower panel) for  $n = 100$ ,  $p = 15$  and beta-type A. Prediction is considered good when RR is close to 0 or RTE is close to 1. RTE magnifies small differences in methods. RR, relative risk; RTE, relative test error.

correlation and SNR levels. It is evident that shrinkage is more crucial in small sample sizes ( $n = 50$ ), particularly in low SNR and high correlation (top right panel). However, in large sample sizes ( $n = 400$ ), all methods, including OLS models, perform similarly, suggesting that shrinkage might not be necessary.

#### 4.1.2.3 | Effects of Many Noise Variables on Prediction.

To evaluate the effectiveness of post-estimation shrinkage in complex settings, we developed models with 30 covariates (7 signals and 23 noises), sample size of 50, and different SNRs. The results for NPWS and QPWS were similar, and only NPWS results are reported. From Figure 4, we see that the PWS, which had previously shown poor performance (see Figure 2), now outperforms OLS. This improvement can be attributed to PWS regression estimates being more concentrated around true values than OLS estimates, as seen in the corresponding density plots (not shown).

Among post-estimation shrinkage approaches, global shrinkage tended to perform well in low SNR levels, whereas NPWS performed well in moderate to high SNR ( $\text{SNR} > 1$ ). For a small sample size and low SNR, accurate estimation of PWS factors is challenging. In such cases, global shrinkage approach estimating only one parameter may be more appropriate. Ridge outperformed post-estimation shrinkage methods in high correlation.

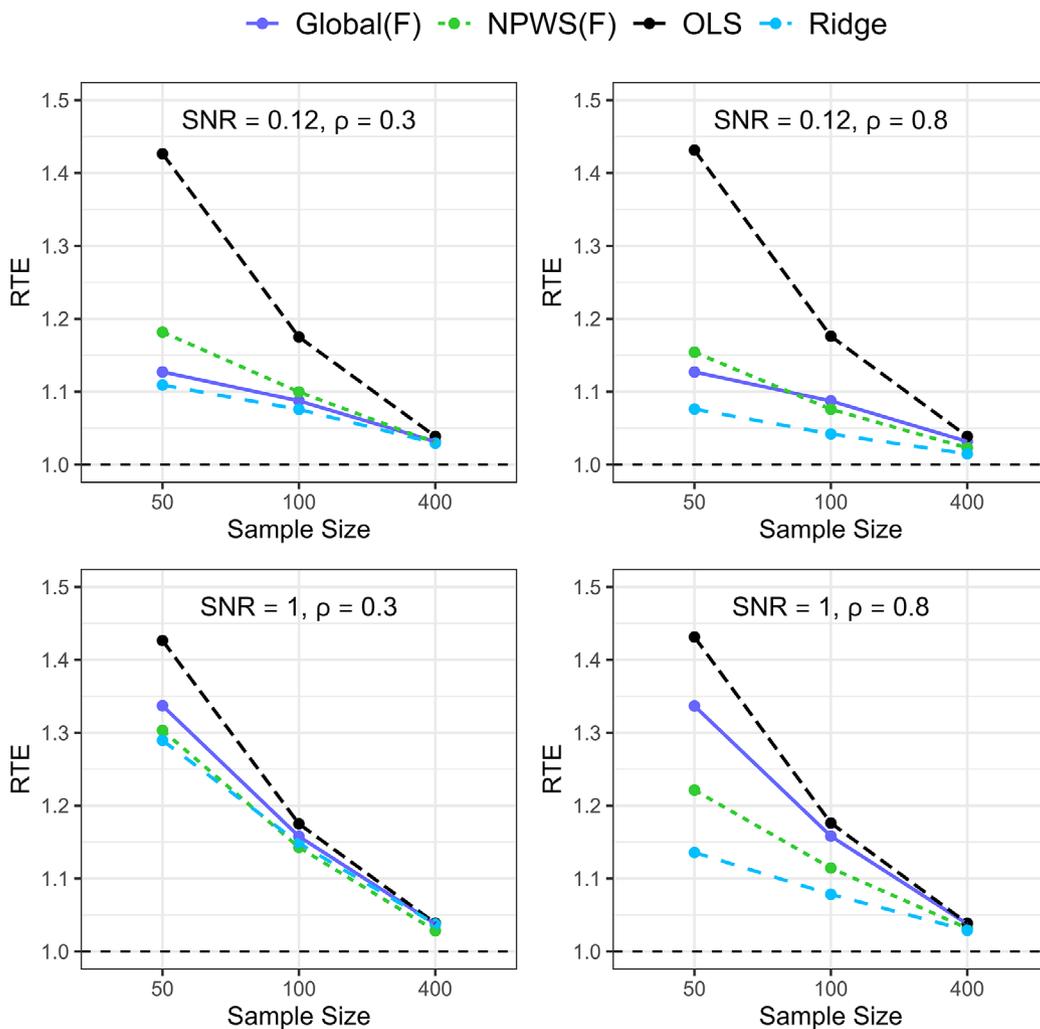
The summary of simulation results in full models is shown in Table A2.

## 4.2 | Selected Models

This section presents the results of post-estimation shrinkage for selected models. We explore the behavior of shrinkage factors of regression estimates of selected noise and signal variables. Additionally, we assess whether post-estimation shrinkage improves the prediction performance of BSS and compare the results with those of lasso.

### 4.2.1 | Behavior of Shrinkage Factors for a Selected Noise Variable in Low-Correlated Settings

We focus on the noise variable  $x_2$ , chosen arbitrarily, which is weakly correlated with other variables. The inclusion frequencies are 18.3%, 17.6%, and 11.5% for sample sizes of 50, 100, and 400, respectively. Figure A5 shows that as the regression estimate of a selected noise variable increases in absolute terms, the amount of shrinkage applied decreases. Conversely, as the estimate approaches zero, the PWS approaches assume that the variable is more likely a noise variable, leading to more shrinkage.



**FIGURE 3** | Full model. RTE as a function of sample size for different methods in low (left panel) and high (right panel) correlation with SNR of 0.12 (upper panel) and 1 (lower panel) for  $p = 15$  covariates with beta-type  $A$  distribution. RTE, relative test error; SNR, signal-to-noise ratio.

The PWS approach may estimate negative shrinkage factors in small (50) and moderate (100) sample sizes, particularly when the regression estimate is close to zero. This is less likely to occur in large sample sizes (400). Generally, global shrinkage applies less shrinkage to estimates close to zero, whereas NPWS and QPWS apply more shrinkage. As the sample size increases from 50 to 400, the regression estimate approaches zero (near the dashed vertical line), and the corresponding shrinkage factor varies depending on the method used.

#### 4.2.2 | Behavior of Shrinkage Factors for a Selected Signal Variable in Low-Correlated Settings

We focus on variable  $x_7$  with the true effects of 0.5. As the effect is relatively weak, the inclusion frequencies were 37.9%, 56.7%, and 98.2% for sample sizes of 50, 100, and 400, respectively. As the sample size increased, the likelihood of selecting the variable also increased, as larger sample sizes provide more power to detect true effects.

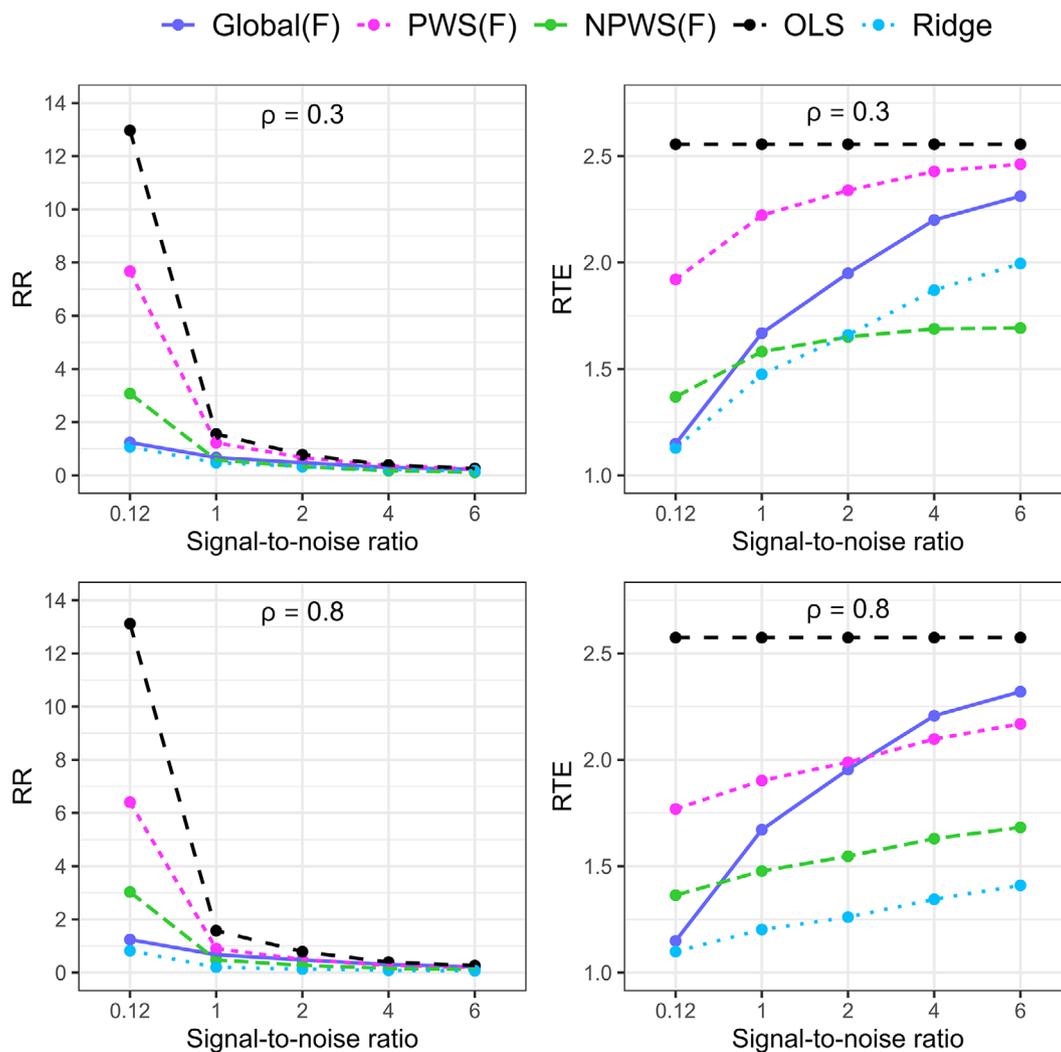
The results show that more shrinkage is applied to estimates that are closer to zero and smaller than true effects in absolute terms.

Conversely, less shrinkage is applied to estimates that are larger than true effects in absolute terms. This trend is more apparent in large sample sizes (Figure 5). The exception to this is global shrinkage, where shrinkage seems to be applied uniformly to all estimates regardless of their deviation from the true value.

Additionally, in small (50) and moderate (100) sample sizes, the PWS can estimate negative shrinkage factors when its estimate is close to zero. This is undesirable because shrinkage factors should not alter the sign of an estimate. In contrast, the NPWS estimates shrinkage factors for estimates close to zero as either zero or very close to zero. It is evident that the choices of shrinkage method and sample size have a significant impact on the estimates obtained from shrinkage methods.

#### 4.2.3 | Usefulness of Shrinkage on Prediction of Subset Selection in Low and High Correlation

Figure 6 compares the prediction performance of BSS, post-estimation shrinkage, and lasso in low (upper panel) and high (lower panel) correlations. The results for NPWS and QPWS were similar, and only NPWS results are reported. Post-estimation



**FIGURE 4** | Full models. Prediction performance of methods in full models with a large number of noise variables under low (upper panel) and high (lower panel) correlation. The analysis involves a sample size of  $n = 50$  and  $p = 30$  covariates (7 signals and 23 noise variables) following a beta-type  $A$  distribution.

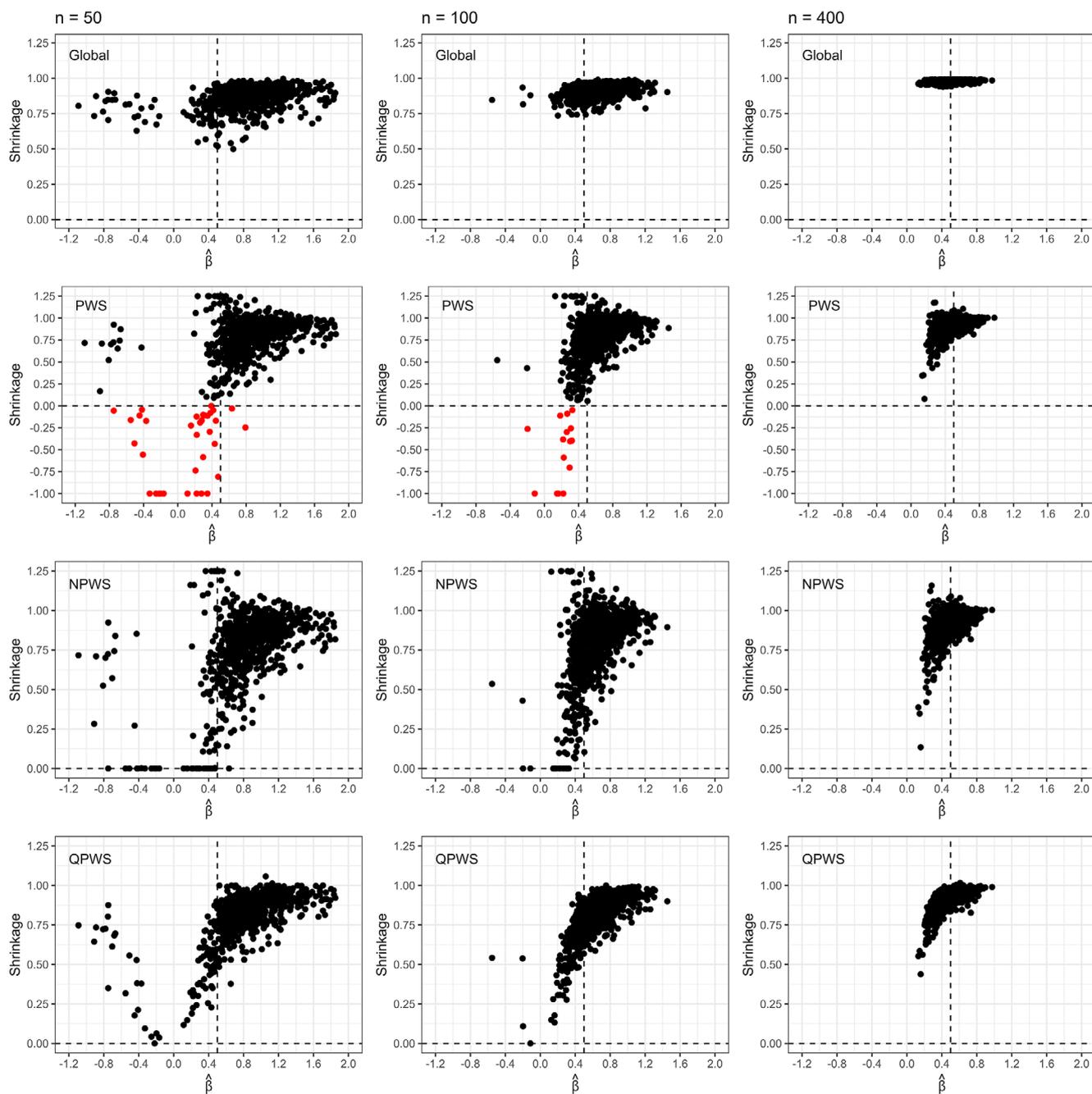
shrinkage improves prediction accuracy of BSS, particularly in low SNR, with negligible impact in high SNR (Figure 6, top left panel). All post-estimation shrinkage approaches performed similarly, with global shrinkage being slightly inferior.

The lasso outperformed BSS and post-estimation shrinkage in small sample sizes (Figure A6), low SNR, and high correlation (Figure 6, and Figures A7 and A8; bottom left panel). However, in moderate (100) and large (400) sample sizes with low correlation and high SNR, BSS and post-estimation shrinkage outperformed the lasso (Figure 6 and Figures A7 and A8; top left panels). The poor performance of the lasso in high SNR was likely due to the selection of many noise variables (Figure 6 and Figures A7, A8, and A9, right panel).

## 5 | Discussion

In our simulation study, we have investigated the prediction performance of post-estimation shrinkage methods in low-dimensional data. Our findings indicate that post-estimation

shrinkage methods (NPWS and global shrinkage) generally outperform OLS in both full and selected models. As expected, when the PWS approach, proposed for use after model selection, was applied to the full model, its prediction performance was inferior to the OLS model due to the impact of negative shrinkage factors. The popular approach of Stein (1956) used for estimating shrinkage factors suffered from the same problem, and a modification was proposed called the “positive part,” where the negative shrinkage factors are set to zero. This modification often outperforms the original Stein approach (Copas 1983). This also explains why NPWS outperformed PWS because when all shrinkage factors are zero, all observations in new data are predicted using the overall mean rather than a linear predictor with incorrect signs for the estimates. Nevertheless, the performance of PWS was comparable to other post-estimation shrinkage methods in selected models, as BSS eliminates some noise variables, resulting in PWS estimating positive shrinkage factors for selected variables in nearly all cases. These results are consistent with the findings of van Houwelingen and Sauerbrei (2013), which showed that PWS improved the prediction performance of backward elimination.

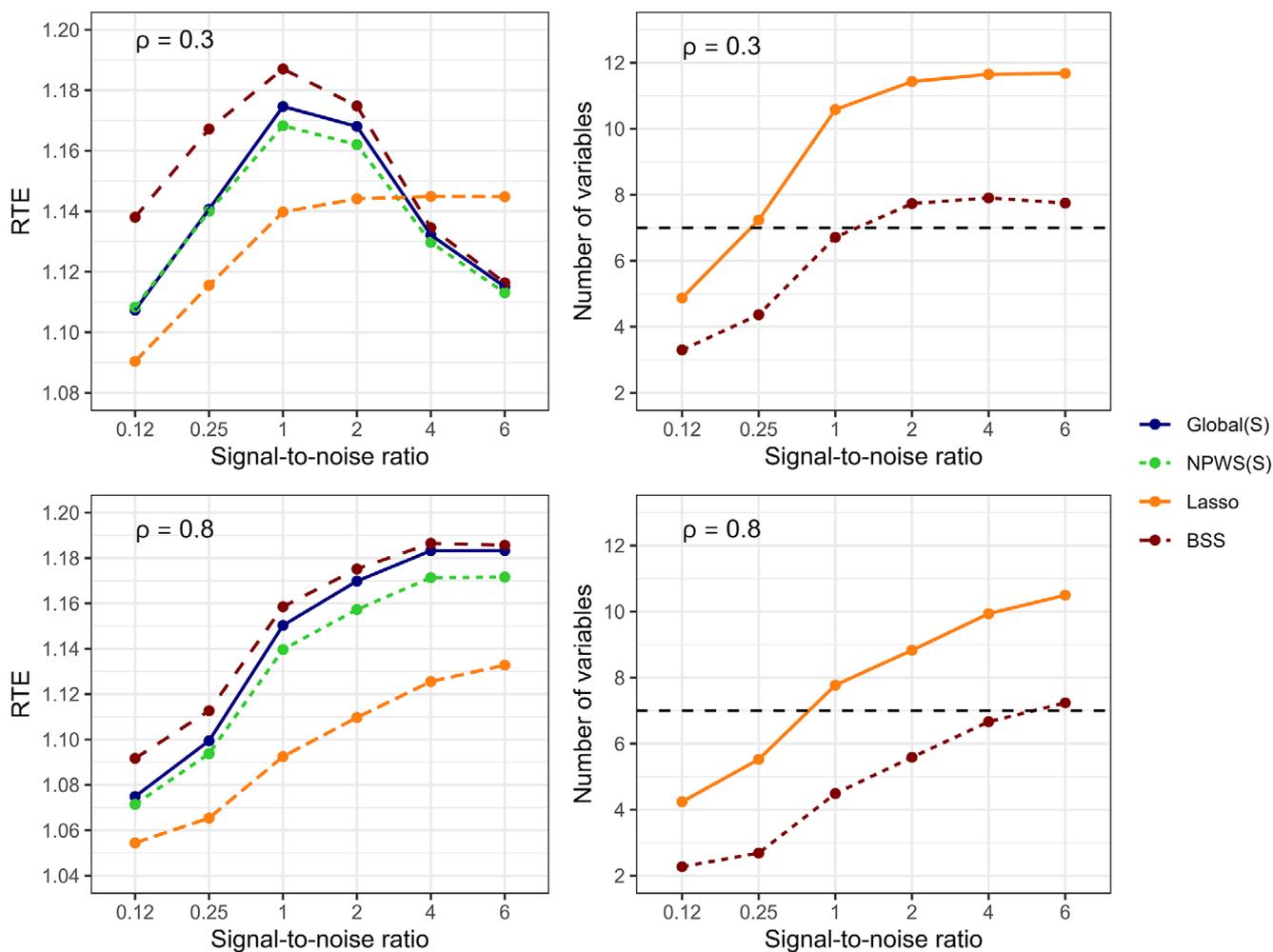


**FIGURE 5** | Selected models. Shrinkage behavior of a signal variable ( $x_7$ ) in selected models for small ( $n = 50$ , left panel), moderate ( $n = 100$ , middle panel), and large ( $n = 400$ , right panel) sample sizes. The scenarios include an SNR of 1, low correlation ( $\rho = 0.3$ ), and  $p = 15$  covariates following beta-type  $A$  distribution. SNR, signal-to-noise ratio.

The prediction performance of NPWS and QPWS was better than global shrinkage in most scenarios, except in small sample sizes with low SNR. Using CV to estimate many shrinkage factors in these scenarios can lead to large uncertainty. This may explain the superior performance of global shrinkage over PWS approaches in such cases, as it involves estimating only one parameter. Moreover, in small sample sizes with low SNR, distinguishing between signal and noise variables is challenging, raising doubts about the necessity of performing any model selection on such situations. However, with sufficient information, PWS approaches tend to outperform global shrinkage as over-shrinking of large effects is avoided.

In scenarios with high correlation, small sample sizes, and low SNR, shrinkage approaches outperformed OLS in prediction accuracy. The uncertainty in OLS estimators is substantial in these situations, and shrinkage methods can effectively reduce variance by introducing small bias, enhancing prediction accuracy in new data.

We evaluated the usefulness of post-estimation shrinkage in selected models and observed that shrinkage is more useful in low SNR and small sample sizes. In these situations, regression estimates of selected variables are known to be highly biased in absolute terms, leading to an increased tendency to overpredict,



**FIGURE 6** | Selected model. The relative test error (left panel) and average number of variables selected (right panel) as functions of SNR, in the low (upper panel) and high (lower panel) correlation setting with  $n = 100$ ,  $p = 15$ , and beta-type A. SNR, signal-to-noise ratio.

and more shrinkage is needed (Copas 1983). Additionally, we compared the prediction performance of post-estimation shrinkage and penalized methods and observed that the performance of the two approaches was comparable in moderate sample sizes with low correlation and moderate to high SNR ( $\text{SNR} > 1$ ). However, in large sample sizes with low correlation and high SNR, the lasso performed poorly compared to both post-estimation shrinkage and BSS. This may be attributed to the selection of a large number of variables and an excessive amount of shrinkage, aligning with findings by Hastie, Tibshirani, and Tibshirani (2020), where BSS outperformed the lasso in high SNR.

In high correlation, low SNR, or small sample sizes, penalized methods generally outperformed post-estimation shrinkage approaches. This could be attributed to the estimation of post-estimation shrinkage factors, which relies on OLS estimates. OLS estimates are highly variable in these situations, which, in turn, can adversely affect the estimation accuracy of shrinkage factors.

## 6 | Conclusion

Our study has demonstrated that post-estimation shrinkage can be an effective tool for improving the prediction performance

of full and selected models, especially in small-to-moderate sample sizes or SNR. However, its usefulness diminishes as the sample size or SNR increases. When the data contain sufficient information, NPWS is more effective than global shrinkage in improving the prediction performance of models. However, in high correlation, small sample size, and very low SNR, penalized methods tend to outperform post-estimation shrinkage methods. Therefore, researchers should carefully consider factors such as the sample size, correlation between covariates, and SNR levels when selecting the appropriate method to achieve the best prediction performance.

### Acknowledgments

The authors gratefully acknowledge Milena Schwotzer (Medical Center-University of Freiburg, Germany) and Jakob Moeller (Medical Center-University of Freiburg, Germany) for invaluable administrative assistance.

Open access funding enabled and organized by Projekt DEAL.

### Conflicts of Interest

The authors declare no conflicts of interest.

## Data Availability Statement

The data and code supporting the findings of this study will be made available at <https://github.com/EdwinKipruto/shrinkage>.

## Open Research Badges



This article has earned an Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available in the [Supporting Information](#) Appendix section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## References

- Bertsimas, D., A. King, and R. Mazumder. 2016. “Best Subset Selection via a Modern Optimization Lens.” *Annals of Statistics* 44, no. 2: 813–852.
- Breiman, L. 1995. “Better Subset Regression Using the Nonnegative Garrote.” *Technometrics* 37, no. 4: 373–384.
- Copas, J. B. 1983. “Regression, Prediction and Shrinkage.” *Journal of the Royal Statistical Society Series B: (Methodological)* 45, no. 3: 311–335.
- Copas, J. B., and T. Long. 1991. “Estimating the Residual Variance in Orthogonal Regression With Variable Selection.” *Journal of the Royal Statistical Society: Series D (The Statistician)* 40, no. 1: 51–59.
- Dunkler, D., W. Sauerbrei, and G. Heinze. 2016. “Global, Parameterwise and Joint Shrinkage Factor Estimation.” *Journal of Statistical Software* 69: 1–19.
- Friedman, J., R. Tibshirani, and T. Hastie. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33, no. 1: 1–22.
- Harrell, F. E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.
- Hastie, T., R. Tibshirani, and R. Tibshirani. 2020. “Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons.” *Statistical Science* 35, no. 4: 579–592.
- Hoerl, A. E., and R. W. Kennard. 1970. “Ridge Regression: Biased Estimation for Non-Orthogonal Problems.” *Technometrics* 12: 55–67.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R* [First edition], 22, 205, and 219. New York: Springer.
- Kipruto, E., and W. Sauerbrei. 2022a. “Comparison of Variable Selection Procedures and Investigation of the Role of Shrinkage in Linear Regression-Protocol of a Simulation Study in Low-Dimensional Data.” *PLoS ONE* 17, no. 10: e0271240.
- Kipruto, E., and W. Sauerbrei. 2022b. *Exhuming Nonnegative Garrote From Oblivion Using Suitable Initial Estimates-Illustration in Low and High-Dimensional Real Data*. <https://doi.org/10.48550/arXiv.2210.15592>.
- Lumley, T. 2020. Leaps: Regression Subset Selection. R package version 3.1. <https://CRAN.R-project.org/package=leaps>.
- Miller, A. J. 2002. *Subset Selection in Regression. Monographs on Statistics and Applied Probability*. London, England: CRC Press.
- Morris, T. P., I. R. White, and M. J. Crowther. 2019. “Using Simulation Studies to Evaluate Statistical Methods.” *Statistics in Medicine* 38, no. 11: 2074–2102.
- Mullen, K. M., and I. H. M. van Stokkum. 2023. nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS). R package version 1.5. <https://CRAN.R-project.org/package=nnls>.
- Riley, R. D., K. I. Snell, G. P. Martin, et al. 2021. “Penalization and Shrinkage Methods Produced Unreliable Clinical Prediction Models Especially When Sample Size Was Small.” *Journal of Clinical Epidemiology* 132: 88–96.
- Sauerbrei, W. 1999. “The Use of Resampling Methods to Simplify Regression Models in Medical Statistics.” *Journal of the Royal Statistical Society Series C: Applied Statistics* 48, no. 3: 313–329.
- Shmueli, G. 2010. “To Explain or to Predict?.” *Statistical Science* 25, no. 3: 289–310.
- Stein, C. 1956. “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal distribution.” In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 197–207. Berkeley: University of California Press.
- Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58, no. 1: 267–288.
- van Houwelingen, H. C., and W. Sauerbrei. 2013. “Cross-Validation, Shrinkage and Variable Selection in Linear Regression Revisited.” *Open Journal of Statistics* 3, no. 2: 79–102.
- van Houwelingen, J. C., and S. le Cessie. 1990. “Predictive Value of Statistical Models.” *Statistics in Medicine* 9, no. 11: 1303–1325.
- Verweij, P. J. M., and H. C. van Houwelingen. 1993. “Cross-Validation in Survival Analysis.” *Statistics in Medicine* 12, no. 24: 2305–2314.
- Xiong, S. 2010. “Some Notes on the Nonnegative Garrote.” *Technometrics* 52, no. 3: 349–361.
- Yuan, M., and Y. Lin. 2007. “On the Non-Negative Garrote Estimator.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69, no. 2: 143–161.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.