Model Averaging with Ridge Regularization*

Alena Skolkova[†]

Abstract

Model averaging is an increasingly popular alternative to model selection. Ridge regression serves a similar purpose as model averaging, i.e. the minimization of mean squared error through shrinkage, though in different ways. In this paper, we propose the ridge-regularized modifications of Mallows model averaging (Hansen, 2007, *Econometrica*, 75) and heteroskedasticity-robust Mallows model averaging (Liu & Okui, 2013, *The Econometrics Journal*, 16) to leverage the capabilities of averaging and ridge regularization simultaneously. Via a simulation study, we examine the finite-sample improvements obtained by replacing least-squares with a ridge regression. Ridge-based model averaging is especially useful when one deals with sets of moderately to highly correlated predictors because the underlying ridge regression accommodates correlated predictors without blowing up estimation variance. A toy theoretical example shows that the relative reduction of mean squared error is increasing with the strength of the correlation. We also demonstrate the superiority of the ridge-regularized modifications via empirical examples focused on wages and economic growth.

Keywords: linear regression; shrinkage; model averaging; ridge regression; Mallows criterion
JEL Classification: C21, C52

[*] I am grateful to Stanislav Anatolyev and Stepan Jurajda for valuable comments and support at all stages of the project. I also thank the participants of the Dissertation Proposal Workshop and the Dissertation Workshop at CERGE-EI and participants of the 6th conference of the Deutsche Arbeitsgemeinschaft Statistik for useful feedback.

[†] CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences, 111 21 Politickych veznu 7, Prague, Czech Republic. Email: alena.skolkova@cerge-ei.cz.

1 Introduction

Model uncertainty is a challenge that is frequently encountered in applied econometrics. The two most common approaches to addressing model uncertainty are model selection and model averaging. While model selection has been the predominant method for decades, the sensitivity of results to the choice of model selection criteria has contributed to the increasing popularity of model averaging techniques.[1] The central question of model averaging is how to assign weights to candidate models optimally. Many different solutions coexist in the literature.[2]

Although model averaging initially developed within the Bayesian paradigm, the literature on frequentist model averaging (FMA) is currently growing rapidly. Within FMA, early contributions were made by Buckland et al. (1997) who suggested that the weight for each model be a function of its value of the Akaike information criterion (hereafter AIC; Akaike, 1970) or the Schwarz-Bayes information criteria (BIC; Schwarz, 1978). Yang (2001) introduced a way to combine candidate models with weights found via sample splitting, thus making weighting schemes more flexible. Hansen (2007, 2008) adopted the Mallows criterion (Mallows, 1973) to model averaging under error homoskedasticity (Mallows model averaging, or MMA), thereby providing a way to find optimal weights without efficiency losses caused by sample splitting. Later, Liu & Okui (2013) introduced a heteroskedastiticty-robust Mallows criterion for model averaging (hereafter HR-MMA).

In this paper, we propose ridge regularized versions of the MMA and HR-MMA estimators that provide better finite-sample prediction performance in terms of the mean squared error (MSE), the ridge model averaging (RMA) estimator and heteroskedasticity-robust ridge model averaging (HR-RMA) estimator, respectively. Ridge regression, introduced by Hoerl and Kennard (1970), is a generalization over OLS regression that aims to reduce the MSE by penalizing large coefficients. A penalization parameter governs the amount of shrinkage (and thus the coefficient biasness) that, in general, makes it possible to trade off a small bias for a significant reduction in variance of estimates, thereby lowering the mean squared error. The gain from ridge regularization tends to be larger in the case of high correlation among predictors.

Building on the idea of least squares averaging by Hansen (2007) we replace ordinary least-squares estimation with a ridge regression, to minimize the consequences of correlation among predictors. Our proposed estimators differ from the MA-Ridge estimator by Zhao et al (2020), which averages across varying regularization parameter values for a single model specification (i.e. across estimators instead of models), and obtains the optimal weights through minimization of the jackknife criterion. Another possible benchmark for our estimator is the jackknife model averaging (JMA)

[1] See also Breiman (1996) where subset selection is shown to be unstable, thus resulting in poor prediction accuracy.

[2] Moral-Benito (2015) and Steel (2017) provide comprehensive reviews of model averaging in economics.

estimator by Hansen and Racine (2012), which is a regularization-free baseline of the MA-Ridge estimator by Zhao et al (2020).

In a Monte Carlo study we compare the finite sample performance of the RMA and HR-RMA estimators with that of the MMA and HR-MMA estimators, as well as several other estimators including weighted BIC (WBIC), Bates-Granger (by Bates and Granger, 1969), and JMA. In general, our simulation design is close to that adopted in Hansen (2007, 2008), while we also examine separately the cases of medium and high correlation among predictors. Although the ridge model averaging estimator does not uniformly MSE-dominate all alternative estimators for all considered specifications, it typically has the best performance over considerable intervals of population R^2.

The reduction in MSE achieved by the RMA can be viewed through the lens of optimal weights. Basically, the set of alternative models includes those with parsimonious specifications (with a relatively small number of regressors and thus a small number of parameters to be estimated), and sophisticated models (with many regressors), as well as variations of these two (moderately parametrized). The optimal weights found via RMA tend to be higher for more sophisticated models, while the weights obtained via different procedures are predominantly distributed between low and moderately parametrized specifications (see Appendix W). This is because the ridge model averaging estimator can use more information from highly parametrized models without inflating the estimation variance, whereas this property is not shared by estimators based on simple least squares estimators.

We also demonstrate how the proposed estimator works in two empirical examples. First, we employ the cross-section earning data used by Hansen and Racine (2012). Second, we use the Barro and Lee (1994) data on cross-country determinants of long-term economic growth. In both examples, there are many possible predictors to be used relative to the sample size. In both examples, ridge-regularized modifications of the MMA and HR-MMA estimators tend to perform better than the baselines, especially in small samples.

This paper proceeds as follows. Section 2 introduces a model averaging estimator in general, and a ridge-regularized model averaging estimator in particular. Section 3 presents a toy example that demonstrates the reduction in MSE achieved via the use of ridge regularization. Section 4 shows the results of a Monte Carlo study that examines the relative performance of several competing estimators in finite samples. Section 5 presents empirical examples. Section 6 concludes.

2 Model Averaging

The setup and notation are taken from Hansen (2007). Consider $\{(y_i, x_i)\}$, $i = 1, ..., n$. Let $\mu_i = \mu(x_i) = \mathbb{E}(y_i | x_i)$ be the conditional mean so that

$$y_i = \mu_i + e_i, \tag{1}$$

where $\mathbb{E}(e_i | x_i) = 0$. For further use of matrix notation define $\mathbf{y} = (y_1, ..., y_n)'$, $\mu = (\mu_1, ..., \mu_n)'$, $\mathbf{e} = (e_1, ..., e_n)'$. The conditional variance $\sigma^2(x_i) = \mathbb{E}(e_i^2 | x_i)$ may depend on x_i.

Consider a set of competitive linear estimators $\{\hat{\mu}^1, ..., \hat{\mu}^M\}$ for the conditional mean μ.[3] Every estimator from this set can be written as $\hat{\mu}^m = \mathbf{P}_m \mathbf{y}$, where operator $\mathbf{P}_m$ does not depend on $\mathbf{y}$. Then the model selection problem is about picking a single estimator from the set $\{\hat{\mu}^1, ..., \hat{\mu}^M\}$. When the selection is guided by the mean-squared error (MSE) criterion, the traditional bias-variance trade-off arises, and thus in principle the model of any complexity may attain a balance.

Compared to model selection, model averaging involves averaging across $\{\hat{\mu}^1, ..., \hat{\mu}^M\}$ to attain further reduction of the MSE. Consider $\mathbf{w} = (w^1, ..., w^M)'$, a vector of non-negative weights such that $\sum_{m=1}^M w^m = 1$. Then for any admissible $\mathbf{w}$, the averaging estimator for μ takes the form

$$\hat{\mu}(\mathbf{w}) \equiv \sum_{m=1}^{M} w^m \hat{\mu}^m = \hat{\mu}\mathbf{w} = \mathbf{P}(\mathbf{w})\mathbf{y}, \tag{2}$$

where $\hat{\mu} = (\hat{\mu}^1, ..., \hat{\mu}^M)$ is the $n \times M$ matrix of first-step estimates, and

$$\mathbf{P}(\mathbf{w}) \equiv \sum_{m=1}^{M} w^m \mathbf{P}_m. \tag{3}$$

For least-squares estimators, $\mathbf{P}_m = \mathbf{P}_m^{LS} \equiv X^m (X^{m\prime} X^m)^{-1} X^{m\prime}$, where x_i^m is the i'th row of X^m, x_i^m is $1 \times k_m$ for $m = 1, 2, ..., M$. In the case of ridge estimators,

$$\mathbf{P}_m^R \equiv X^m (X^{m\prime} X^m + \lambda_m I_{k_m})^{-1} X^{m\prime}$$

for a tuning parameter $\lambda_m \in (0, \infty)$. A particular model corresponds to a choice of predictors x_i^m together with the optimal value of λ_m.

The averaging residual is

[3] The number of competitive estimators M may grow with n but we omit the subscript from M_n for the sake of simpler notation.

$$\hat{e}\left(\mathbf{w}\right)=\mathbf{y}-\hat{\mu}\left(\mathbf{w}\right)=\sum_{m=1}^{M}w^{m}\hat{\mathbf{e}}^{m}=\hat{\mathbf{e}}\mathbf{w},$$

where $\hat{\mathbf{e}}^{m}=\mathbf{y}-\hat{\mu}^{m}$ and $\hat{\mathbf{e}}=\left(\hat{\mathbf{e}}^{1},...,\hat{\mathbf{e}}^{M}\right)$. The Mallows model averaging (MMA) criterion of Hansen (2007) for weight selection is a penalized sum of squared residuals. The weighted average of least-squares residuals is complemented by a penalty term that increases in both error variance and average model complexity conveyed by the trace of the matrix $\mathbf{P}\left(\mathbf{w}\right)$:

$$\begin{aligned}C_{n}\left(\mathbf{w}\right) &= \mathbf{w}'\hat{\mathbf{e}}'\hat{\mathbf{e}}\mathbf{w}+2\hat{\sigma}^{2}\mathrm{tr}\left(\mathbf{P}\left(\mathbf{w}\right)\right)\\ \hat{\mathbf{w}}^{MMA} &= \arg\min_{w\in\mathcal{H}}C_{n}\left(\mathbf{w}\right),\end{aligned}$$

where $\mathcal{H}=\left\{\mathbf{w}\in[0,1]^{M}:\sum w_{m=1}^{M}=1\right\}$, $\hat{\sigma}^{2}$ is a consistent estimate of the error variance.[4]

Define the in-sample mean-squared error

$$L_{n}\left(\mathbf{w}\right)=\left(\mu_{t}-\hat{\mu}\left(\mathbf{w}\right)\right)'\left(\mu_{t}-\hat{\mu}\left(\mathbf{w}\right)\right).$$

Lemma 3 from Hansen (2007) shows unbiasedness (up to a constant) of $C_{n}\left(\mathbf{w}\right)$ for in-sample mean-squared error, $L_{n}\left(\mathbf{w}\right)$, for iid observations. Specifically, he shows that

$$E\left[C_{n}\left(\mathbf{w}\right)\right]=\mathbb{E}\left[L_{n}\left(\mathbf{w}\right)\right]+n\sigma^{2},$$

so that the weights found through minimization of $C_{n}\left(\mathbf{w}\right)$ also minimize $L_{n}\left(\mathbf{w}\right)$, in expectation. In addition, Theorem 1 from Hansen (2007) shows the asymptotic optimality of Mallows' criterion for model selection with independent data if the weights are restricted to a discrete set, in the sense that $L_{n}\left(\hat{\mathbf{w}}\right)/\inf_{\mathbf{w}\in\mathcal{H}_{n}(N)}L_{n}\left(\mathbf{w}\right)\rightarrow_{p}1$, where $\mathcal{H}_{n}\left(N\right)$ restricts the weights w_{m} to the set $\left\{0,\frac{1}{N},\frac{2}{N},...,1\right\}$. Notably, the asymptotic optimality of the Mallows' criterion relies on homoskedasticity of the error term.[5]

To address the case of the heteroskedastic error term, Liu and Okui (2013) introduced a modification of the Mallows' criterion that is heteroskedasticity-robust, the so called HRC_{p} criterion:

$$HRC_{p}\left(\mathbf{w}\right)\equiv\left\|\mathbf{y}-\mathbf{P}\left(\mathbf{w}\right)\mathbf{y}\right\|^{2}+2\mathrm{tr}\left[\Omega\mathbf{P}\left(\mathbf{w}\right)\right],$$

where Ω is an $n\times n$ diagonal matrix with σ_{i}^{2} being the ith diagonal element. The feasible HRC_{p}

[4] Hansen (2007) suggests employing $\hat{\sigma}^{2}$ from the "largest" approximating model.

[5] Wan et al. (2010) provide an alternative proof of the asymptotic optimality that extends the result to a non-discrete weight set.

criterion

$$\widehat{HRC}_p(\mathbf{w}) \equiv \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\sum_{i=1}^{n} \hat{e}_i^2 p_{ii}(\mathbf{w}),$$

where $\hat{e}_i$ is the residual from a preliminary estimation[6] and $p_{ii}(\mathbf{w})$ is the ith diagonal element of $\mathbf{P}(\mathbf{w})$. The weights obtained through minimization of the HRC_p criterion are shown to be asymptotically optimal (see Theorem 2.1 from Liu and Okui, 2013). The same property is shared by its feasible version (under more assumptions, see their Theorem 2.2).[7] For the sake of consistent notation within this paper, the weights obtained via minimization of the $\widehat{HRC}_p$ criterion will be denoted as $\hat{w}^{HR-MMA}$.

Ridge Model Averaging

We define the ridge-regularized MMA estimator (hereafter RMA) as

$$\hat{\mathbf{w}}^{RMA} = \arg\min_{w\in\mathcal{H}} \left[\mathbf{w}'\hat{\mathbf{e}}_R'\hat{\mathbf{e}}_R\mathbf{w} + 2\hat{\sigma}^2\text{tr}\left(P^R(\mathbf{w})\right)\right],$$

where $P^R(\mathbf{w}) = \sum_{m=1}^{M} w^m \mathbf{P}_m^R$ and $\hat{\mathbf{e}}_R = \left(\hat{\mathbf{e}}_R^1, ..., \hat{\mathbf{e}}_R^M\right)$ is a matrix of stacked residuals from ridge regressions for each specification. Thus, ridge regularization affects both terms of the criterion simultaneously. Correspondingly, the heteroskedasticity-robust ridge model averaging (HR-RMA) estimator is defined by

$$\hat{\mathbf{w}}^{HR-RMA} = \arg\min_{w\in\mathcal{H}} \left[\mathbf{w}'\hat{\mathbf{e}}_R'\hat{\mathbf{e}}_R\mathbf{w} + 2\sum_{i=1}^{n} \hat{e}_{iR}^2 p_{ii}^R(\mathbf{w})\right],$$

where $p_{ii}^R(\mathbf{w})$ is the ith diagonal element of $\mathbf{P}^R(\mathbf{w})$. For both the RMA and HR-RMA estimators, $P^R(\mathbf{w})$ is a function of optimal shrinkage values for all models being averaged, i.e. $P^R(\mathbf{w}) = P^R(\mathbf{w}, \lambda^{opt})$. For each separate model m, we estimate λ_m^{opt} via leave-one-out cross-validation that results in asymptotically optimal $\hat{\lambda}_m^{opt}$ (Li, 1987).

Having in mind the results on asymptotic optimality of the Mallows criterion for model averaging by Hansen (2007), and its heteroskedasticity-robust counterpart by Liu and Okui (2013) in the class of linear estimators, we investigate the finite-sample benefits of the proposed regularized modifications from the same class, RMA and HR-RMA, relative to the baselines, MMA and HR-

[6] The authors discuss various possibilities for obtaining $\hat{e}_i$. For instance, in the case of nested models, they recommend using the residuals from the largest model, and this paper follows their recommendation.

[7] Anatolyev (2021) proposes using individual variance estimates that are robust to regressor numerosity.

MMA. For most applications, the right hand side variables tend to be correlated with each other,[8] so the Mallows criterion with underlying ridge regularization of a design matrix is expected to deliver better finite sample properties of the estimates. In the next section we provide a toy example demonstrating the relative performance of the RMA estimator.

3 Theory: Toy Example

In this subsection we consider a toy theoretical example that illustrates the mechanics of the MMA and RMA estimators under homoskedasticity of the error term. First, we derive the MSE for the averaged least-squares and ridge estimates. Then, we derive the optimal shrinkage parameters for two models estimated via the ridge regression, and plug them into the MSE for the averaged ridge estimate. That allows us to find the optimal weights for both estimators.

Let the true unknown model be

$$Y = X_1\beta_1 + X_2\beta_2 + e, \quad \mathbb{E}\left[e|X_1, X_2\right] = 0, \quad \mathbb{E}\left[e^2|X_1, X_2\right] = \sigma^2.$$

Two alternative approximations are $Y = X_1\beta_1 + e_1$ and $Y = X_2\beta_2 + e_2$, i.e. each approximating model includes only a part of the regressors from the true model. The column dimensions of X_1 and X_2 are assumed to be equal, $\operatorname{rank}(X_1) = \operatorname{rank}(X_2) = p$.

Two options are considered: (1) averaging the LS estimates or (2) averaging the ridge estimates for both approximations. Two OLS estimates are given by

$$\hat{\beta}_1^{ols} = \left(X_1'X_1\right)^{-1} X_1'Y \qquad \text{and} \qquad \hat{\beta}_2^{ols} = \left(X_2'X_2\right)^{-1} X_2'Y,$$

and the average least-squares estimate is

$$\tilde{\beta} = w^{ols}\begin{pmatrix} \hat{\beta}_1^{ols} \\ 0 \end{pmatrix} + \left(1 - w^{ols}\right)\begin{pmatrix} 0 \\ \hat{\beta}_2^{ols} \end{pmatrix} = \begin{pmatrix} w^{ols}\hat{\beta}_1^{ols} \\ \left(1 - w^{ols}\right)\hat{\beta}_2^{ols} \end{pmatrix}$$

where w^{ols} is the optimal OLS weight to be determined later.[9] Similarly, two ridge estimates are given by

$$\hat{\beta}_1^{r}\left(\lambda_1\right) = \left(X_1'X_1 + \lambda_1 I_p\right)^{-1} X_1'Y \qquad \text{and} \qquad \hat{\beta}_2^{r}\left(\lambda_2\right) = \left(X_2'X_2 + \lambda_2 I_p\right)^{-1} X_2'Y$$

and the average ridge estimate is

$$\tilde{\beta}\left(\lambda_1, \lambda_2\right) = w^r\begin{pmatrix} \hat{\beta}_1^{r}\left(\lambda_1\right) \\ 0 \end{pmatrix} + \left(1 - w^r\right)\begin{pmatrix} 0 \\ \hat{\beta}_2^{r}\left(\lambda_2\right) \end{pmatrix} = \begin{pmatrix} w^r W_{\lambda_1}\hat{\beta}_1^{ols} \\ \left(1 - w^r\right) W_{\lambda_2}\hat{\beta}_2^{ols} \end{pmatrix}$$

[8] For example, in a high-dimensional dataset, there might be large sample correlations even when the variables are independent, see Fan & Lv (2008).

[9] We assume here that whenever the regressor is missing from the approximating model, the corresponding coefficient is set to 0, as is usually done within the FMA.

where $W_{\lambda_1} = \left(X_1'X_1 + \lambda_1 I\right)^{-1} X_1'X_1$, $W_{\lambda_2} = \left(X_2'X_2 + \lambda_2 I\right)^{-1} X_2'X_2$ and w^r is the optimal ridge weight.

From now on let us assume, for the sake of illustration, that X_1 and X_2 are orthonormal, i.e. $X_1'X_1 = X_2'X_2 = I_p$, and also $X_1'X_2 = \rho I_p$, where ρ mirrors the degree of correlation among the predictors. Then the mean squared error of the average least-squares estimate is

$$\begin{aligned} MSE^{ols}\left(w^{ols}\right) =& p\sigma^2\left[\left(w^{ols}\right)^2 + \left(1-w^{ols}\right)^2\right] + \beta_1^T\left[\left(\left(w^{ols}\right)^2 - 2w^{ols} + 1\right) + \left(1-w^{ols}\right)^2\rho^2\right]\beta_1 \\ &+ \beta_1^T\rho\left[2w^{ols}\left(w^{ols}-1\right) - 2w^{ols}\left(1-w^{ols}\right)\right]\beta_2 \\ &+ \beta_2'\left[\left(w^{ols}\right)^2\rho^2 + \left(\left(1-w^{ols}\right)^2 - 2\left(1-w^{ols}\right) + 1\right)\right]\beta_2, \end{aligned}$$

where p is the common column rank of X_1 and X_2, while the mean squared error of the average ridge estimate is

$$\begin{aligned} MSE^r\left(\lambda_1,\lambda_2,w^r\right) =& p\sigma^2\left[\frac{\left(w^r\right)^2}{\left(1+\lambda_1\right)^2} + \frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}\right] + \\ &+ \beta_1^T\left[\frac{\left(w^r\right)^2 - 2w^r\left(1+\lambda_1\right) + \left(1+\lambda_1\right)^2}{\left(1+\lambda_1\right)^2} + \frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}\rho^2\right]\beta_1 \\ &+ \beta_1^T\rho\left[\frac{2w^r\left(w^r-1-\lambda_1\right)}{\left(1+\lambda_1\right)^2} - \frac{2\left(w^r+\lambda_2\right)\left(1-w^r\right)}{\left(1+\lambda_2\right)^2}\right]\beta_2 \\ &+ \beta_2'\left[\frac{\left(w^r\right)^2}{\left(1+\lambda_1\right)^2}\rho^2 + \left(\frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2} - \frac{2\left(1-w^r\right)}{1+\lambda_2} + 1\right)\right]\beta_2. \end{aligned}$$

Derivations are provided in Appendix E, Part 1. For both $MSE^{ols}\left(w^{ols}\right)$ and $MSE^r\left(\lambda_1,\lambda_2,w^r\right)$ the first term of a sum corresponds to the variance, while the other three terms represent the squared bias.

Before finding the optimal weights for the ridge averaging estimator, the optimal values of λ_1 and λ_2 should be plugged in separately for each ridge regression. Under the assumption that we made

$$\lambda_j^{opt} = \frac{p\sigma^2 + \rho\beta_1'\beta_2}{\beta_j'\beta_j + \rho\beta_1'\beta_2}, \quad j = 1, 2.$$

Derivations are provided in Appendix E, Part 2.

Finally, one can use $MSE^r\left(\lambda_1^{opt},\lambda_2^{opt},w^r\right)$ to find the optimal weights, $0 \leq w^{r,opt} \leq 1$, similar to the optimal weights for the least-squares averaging estimator, $0 \leq w^{ols,opt} \leq 1$. Since the resulting expressions are complicated[10], let us look at the comparative statics.

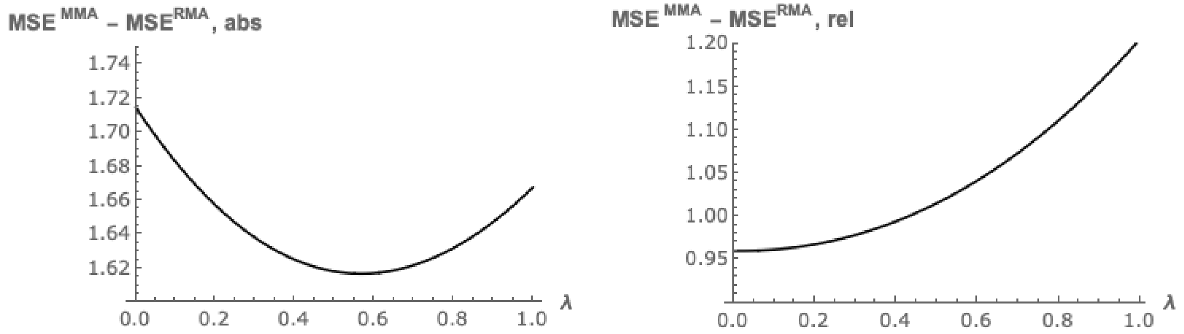


Figure 1: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). Baseline case: $p = 3$, $\sigma^2 = 2$, $\beta_1'\beta_1 = \beta_2'\beta_2 = 1$.

As a baseline case, consider $p = 3$, $\sigma^2 = 2$, $\beta_1'\beta_1 = 1$, $\beta_2'\beta_2 = 1$, $\beta_1'\beta_2 = \sqrt{\beta_1'\beta_1 \cdot \beta_2'\beta_2 - 0.1} = 0.948$. The correlation among the predictors varies between 0 and 1. Figure 1 shows the resulting difference between $MSE^{ols}\left(\widehat{w}^{ols}\right)$ and $MSE^{r}\left(\widehat{\lambda}_1^{opt}, \widehat{\lambda}_2^{opt}, \widehat{w}^{r}\right)$ for $\rho \in [0, 1]$, in absolute terms (left) and relative to $MSE^{r}\left(\widehat{\lambda}_1^{opt}, \widehat{\lambda}_2^{opt}, \widehat{w}^{r}\right)$ (right). Despite the difference itself not being monotonic (in this case, U-shaped), the relative difference is monotonically increasing with the correlation among the predictors. In other words, higher correlation implies larger reduction in the MSE due to ridge regularization, in relative terms.

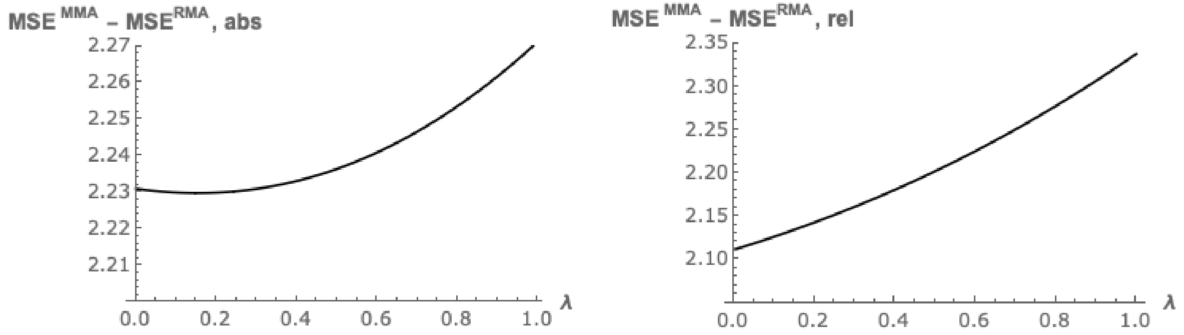


Figure 2: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $\beta_1'\beta_1 = 0.2$

Figures 2, 3, and 4 demonstrate similar outcomes for alternative parameter combinations. In particular, Figure 2 shows the differences in MSE for $\beta_1'\beta_1 = 0.2$, keeping the other parameters the same. In general, the pattern is similar to that for $\beta_1'\beta_1 = \beta_2'\beta_2 = 1$, although the magnitude of $MSE^{ols}\left(\widehat{w}^{ols}\right) - MSE^{r}\left(\widehat{\lambda}_1^{opt}, \widehat{\lambda}_2^{opt}, \widehat{w}^{r}\right)$ is higher in the case of unequal model coefficients. Figure 3 presents the results for the baseline case with the variance of the error term changed to $\sigma^2 = 1$ and $\sigma^2 = 5$, respectively. Overall, the magnitude of the reduction in the MSE is increasing with

[10] Available upon request.

the error variance. Finally, Figure 4 shows the results for the baseline case with the number of predictors changed to $p = 10$. An increase in the number of predictors also leads to a higher magnitude of the reduction in the MSE due to ridge regularization.

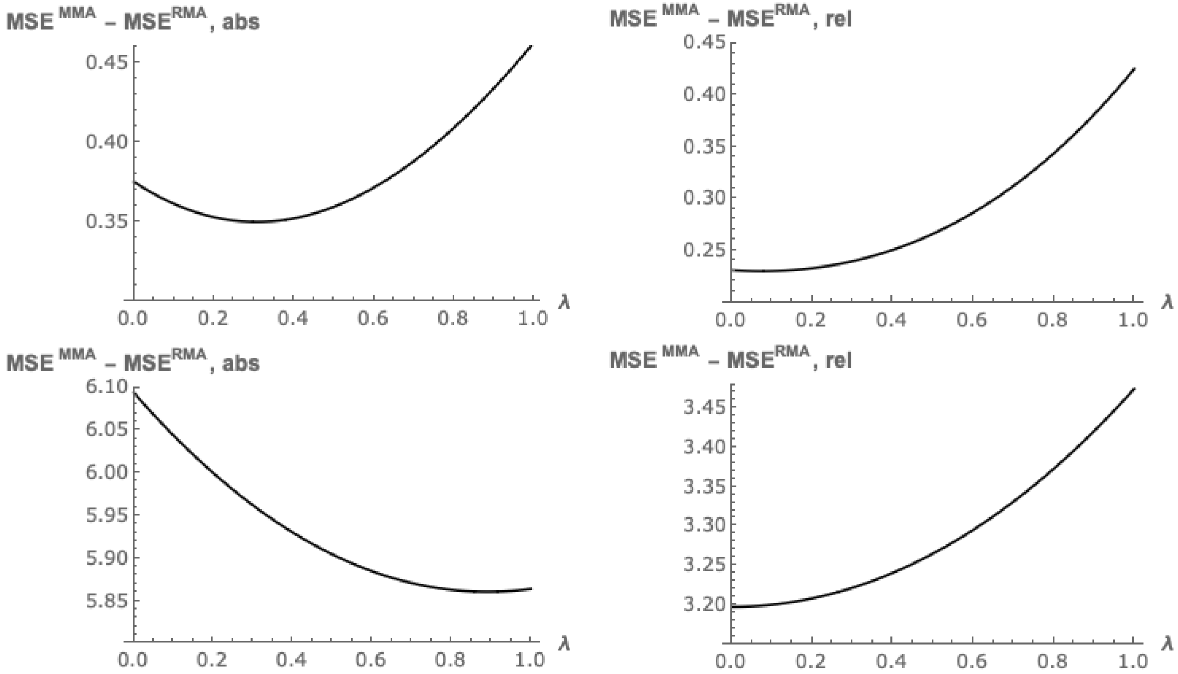


Figure 3: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $\sigma^2 = 1$ (top) and $\sigma^2 = 5$ (bottom)

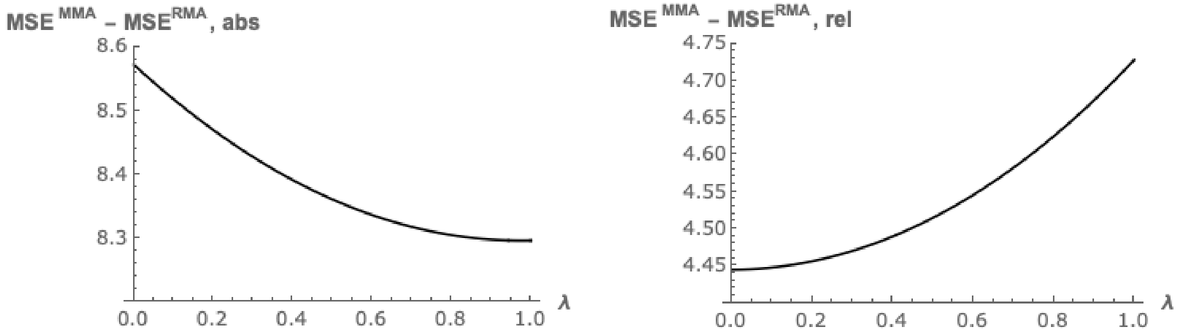


Figure 4: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $p = 10$

In the next section we compare the finite-sample performance of the canonical Mallows model averaging with that also taking advantage of ridge regularization.

4 Finite-Sample Comparison

We now examine the finite-sample performance of the proposed RMA and HR-RMA estimators relative to their closest competitors, the MMA and JMA estimators (Hansen, 2007; Hansen and Racine, 2012), and the HR-MMA estimator (Liu and Okui, 2013), in terms of MSE. Apart from the correlation pattern among predictors, our simulation design combines the features of those from Hansen (2007) and Hansen and Racine (2012). The infinite-order regression model is

$$y_i = \theta_0 + \sum_{k=1}^{\infty} \theta_k x_{ki} + e_i,$$

where x_{ki} are identically distributed $N(0,1)$. All the regressors are equicorrelated with a correlation coefficient 0.5 in case [M](<u>m</u>oderate correlation) and 0.75 in case [H](<u>h</u>igh correlation).[11] The error term e_i is conditionally distributed as $N\left(0, \sigma^2(x_{2i})\right)$, where $\sigma^2(x_{2i}) = x_{2i}^4$. The parameters are set by the rule

$$\begin{aligned} \theta_k &= c\gamma_k \\ \gamma_k &= \frac{k^{\alpha}\beta^{k}}{\sum_{j=1}^{K} j^{2\alpha}\beta^{2j}} \end{aligned}$$

to model various specifications of θ_k. We consider several combinations of α and β. First, for $\alpha = 0.5$, the considered values of β are $[.6, .7, .8, .9]$. Then we fix β at $\beta = 0.7$, and consider $[.25, .5, 1]$ as values for α. The population R^2 varies on a grid from 0.1 to 0.9, so the parameter c is set by the rule $c = \sqrt{R^2/(1-R^2)}$. We examine three sample sizes, $n = 25, 50, 100$ with the maximum model lengths $p = 9, 11, 15$, respectively. In the experiment we also include the weighted BIC criterion (WBIC)[12] and the equal weighting (EW) scheme.[13]

We compare the competing methods based on the mean squared error

$$MSE = \frac{1}{n}(\mu - \widehat{\mu})'(\mu - \widehat{\mu})$$

that is averaged across 5000 simulation draws.

[11] Except for an intercept, x_1.

[12] The least squares model average estimator with the weights $w_m = \exp\left(-\frac{1}{2}\mathrm{BIC}_m\right) / \sum_{j=1}^{M} \exp\left(-\frac{1}{2}\mathrm{BIC}_m\right)$, where $\mathrm{BIC}_m = n \ln \hat{\sigma}_m^2 + \ln(n)\, m$.

[13] The least squares model average estimator with the weights $w_m = 1/M$. EW is uniformly dominated so we do not show it on our graphs for the sake of their better readability.

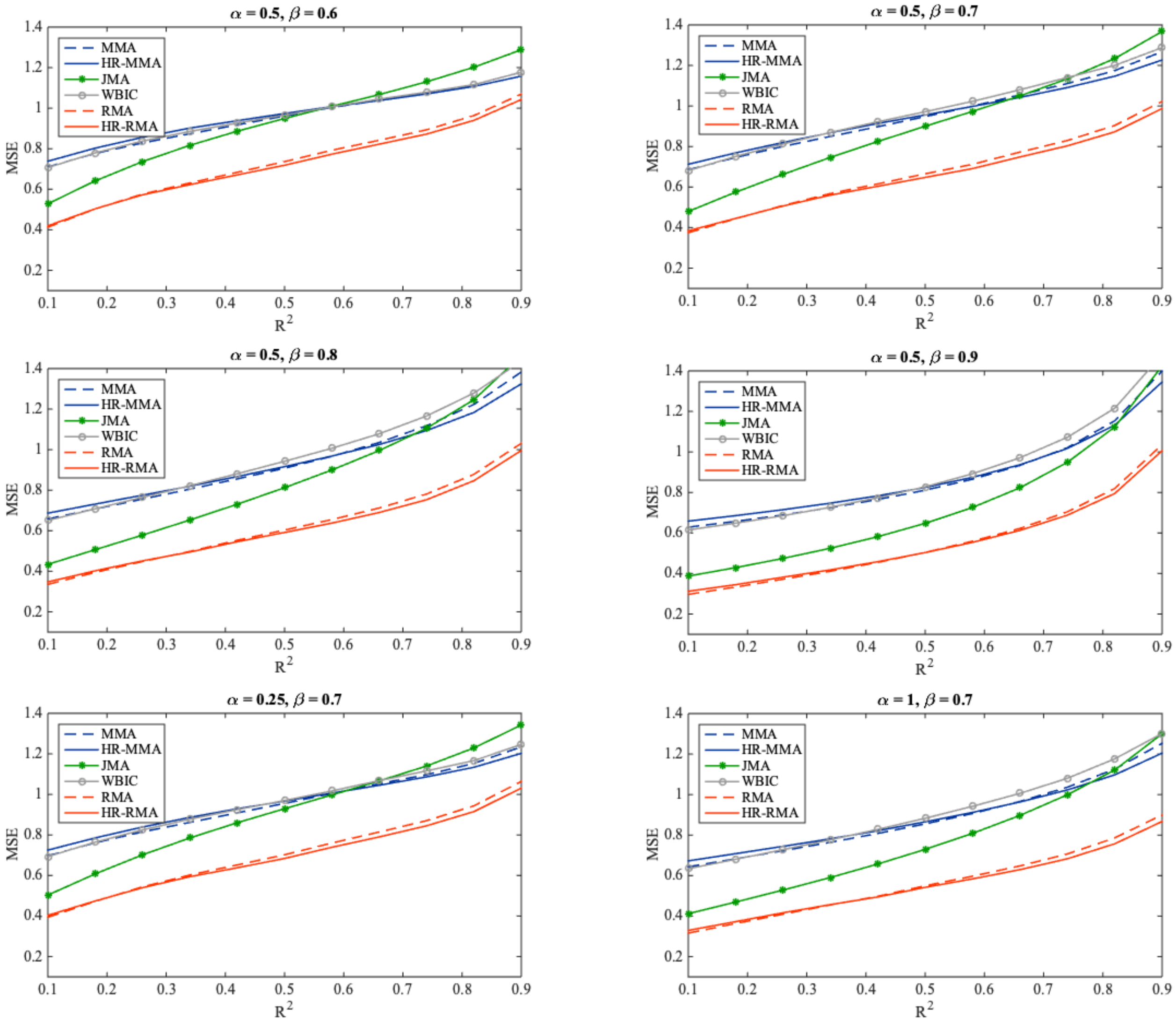


Figure 5: $n = 25$. Case [M] of moderate correlation among predictors.

Figures 5, 6 and 7 present the results for the sample sizes of 25, 50, and 100, respectively, under moderate correlation among the regressors.[14] Each panel of graphs displays average MSE across different values of R^2, varied from 0.1 to 0.9. Overall, the ridge-based model averaging estimators nearly uniformly outperform their alternatives for all sample sizes. In addition, heteroskedasticity robust RMA has a lower MSE than non-robust RMA unless the true R^2 is very low (below about 0.2). The reduction in MSE from using HR-RMA instead of HR-MMA varies between 10% and 53% for $n = 25$, between 6% and 44% for $n = 50$ and between 1% and 44% for $n = 100$. Appendix H presents the results for $n = 100$ in the case [H] of high correlation among the predictors. Although higher correlation does not change the results qualitatively, the improvement achieved by the ridge-based RMA estimators relative to other estimators tends to be more uniformly pronounced under stronger correlation of the regressors.

[14] The shape of coefficients γ_k is shown in Appendix C.

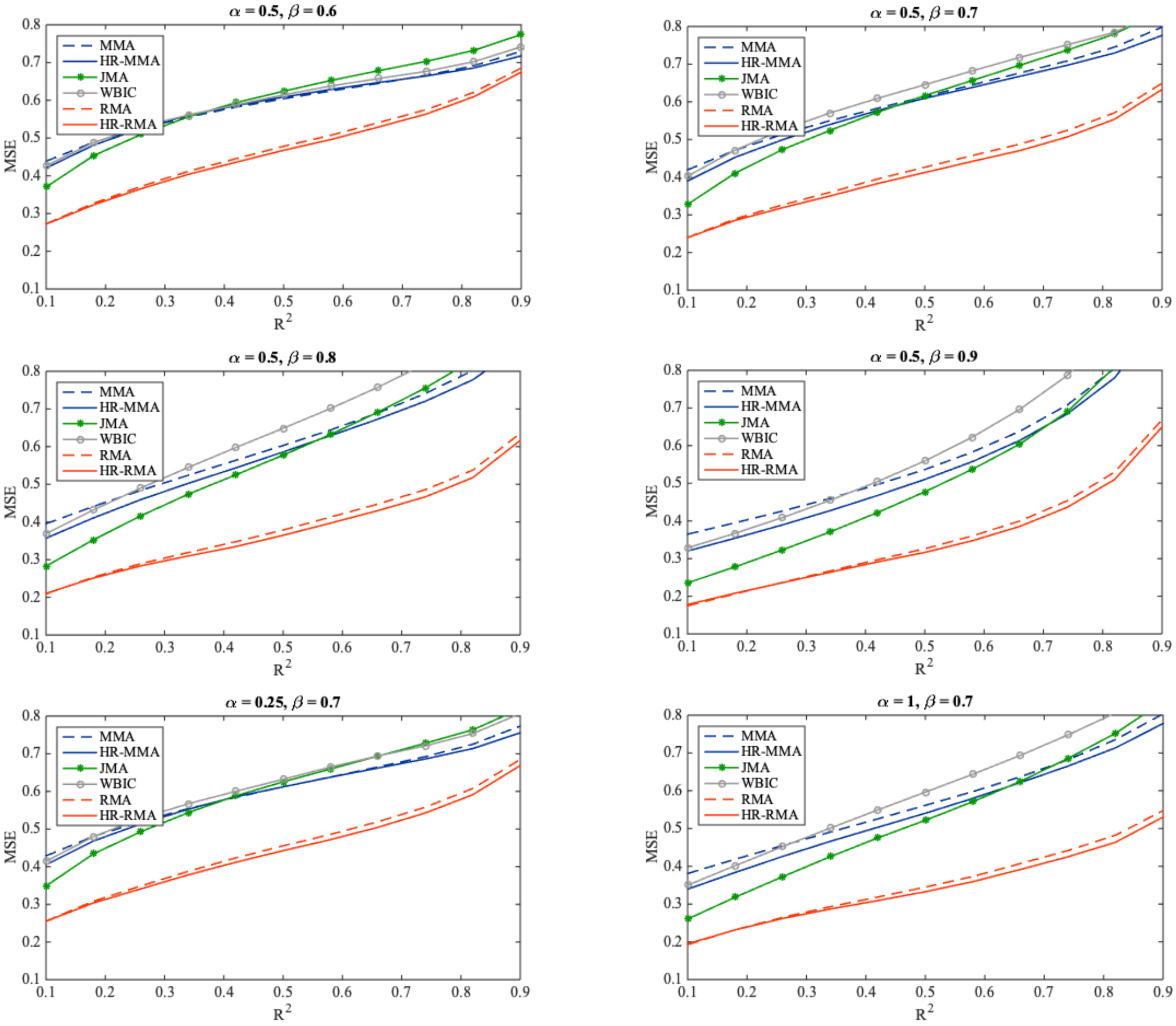


Figure 6: $n = 50$. Case [M] of moderate correlation among predictors

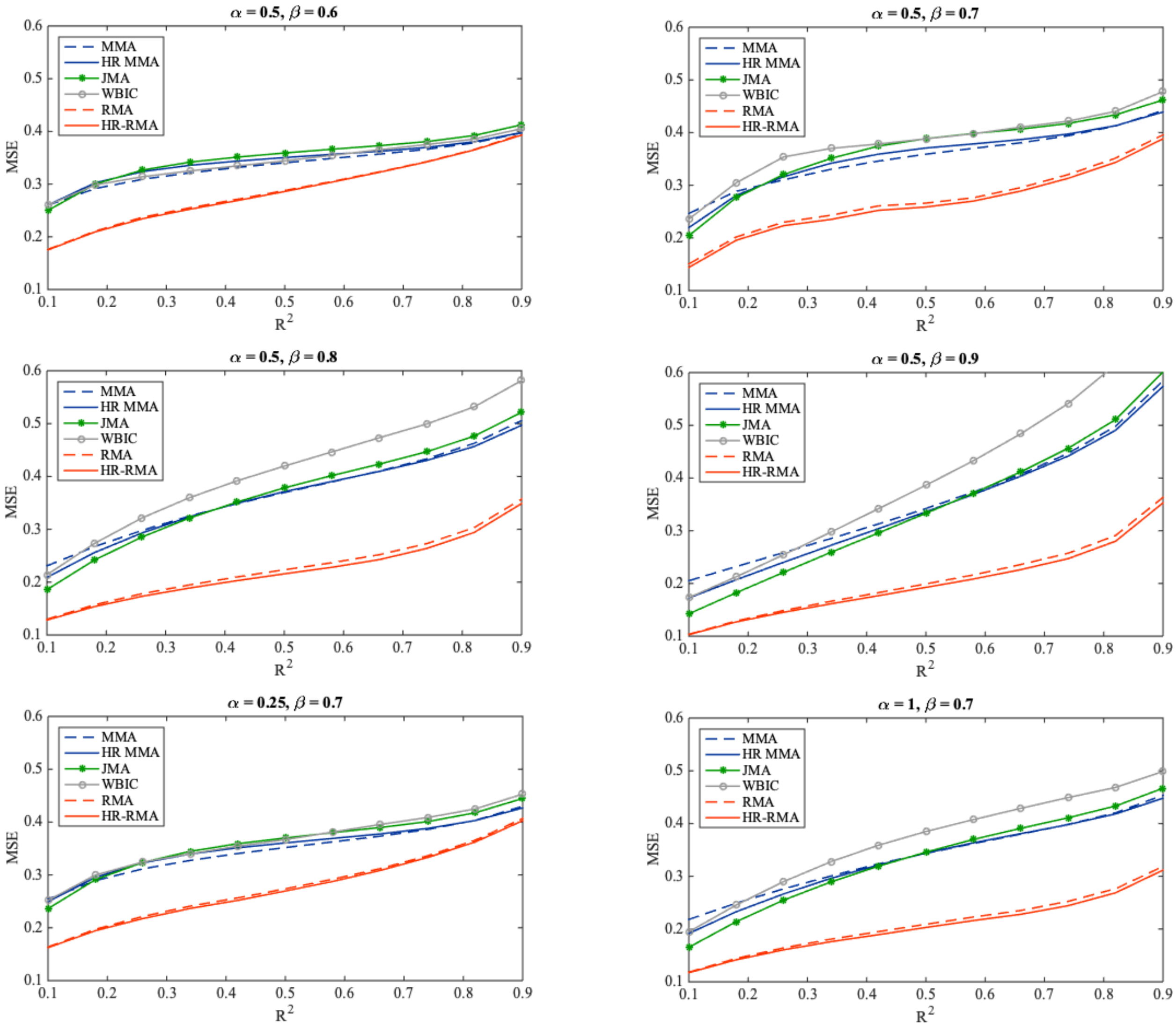


Figure 7: $n = 100$. Case [M] of moderate correlation among predictors

In Appendix W we show the distributions of the optimal weighs over the set of competing models for $n = 100$ with moderately correlated predictors. One can easily see that the weights obtained for the ridge-based estimators tend to favor the larger models, while the optimal weights found via JMA/MMA favor small or moderate model lengths for low and high values of R^2, respectively. The reason is the ability of RMA and HR-RMA to accommodate larger models without inflating the variance, while this property is not shared by estimators based on ordinary least-squares regressions.

In the next section we examine the relative performance of the ridge-based averaging estimators via two real-data examples.

5 Empirical Examples

5.1 Wage Prediction

Similarly to Hansen and Racine (2012), we employ Wooldridge's (2003, pg. 226) 'wage1' cross-sectional dataset, a random sample (526 observations) from the US Current Population Survey for the year 1976.[15] There is uncertainty about the best model for the log of average hourly earnings, so a set of thirty models ranging from the unconditional mean ($k = 1$) through a full model that includes $k = 30$ variables is considered. Explanatory variables include non-dummy variables educ, exper, tenure and dummy variables female, married, nonwhite, numdep, smsa, northcen, south, west, construc, ndurman, trcommpu, trade, services, profserv, profoss, clerocc, servocc, and interaction terms nonwhite×educ, nonwhite×exper, nonwhite×tenure, female×educ, female×exper, female×tenure, married×educ, married×exper, married×tenure.

Then, as in Hansen and Racine (2012), the sample is randomly split into a training portion n_1 and an evaluation portion of size $n_2 = n - n_1$. We compare the same methods mentioned in the previous section: MMA, HR-MMA, JMA, WBIC, RMA and HR-RMA. For each model we compute its average square prediction error (ASPE) using the evaluation set of observations. The procedure is repeated for 100 splits, then the median ASPE over 100 random splits is reported. The size of the training portion is varied, $n_1 = 50, 75, 100, 200, 300, 400, 500$. All numbers in the Table 1 are normalized by the corresponding ASPE of HR-MMA, so the entries lower than 1 indicate superior performance relative to the HR-MMA estimator.

Table 1: Out-of-sample predictive efficiency. Entries less than one indicate superior performance relative to the HR-MMA estimator.

n_1	MMA	JMA	WBIC	RMA	HR-RMA
50	0.7131	0.6935	0.8066	0.6047	0.6272
75	0.9338	0.9012	1.1341	0.8473	0.8731
100	0.9540	0.9389	1.1850	0.9034	0.9214
200	0.9966	0.9952	1.0266	0.9857	0.9903
300	1.0014	1.0018	1.0081	0.9970	0.9929
400	1.0020	1.0044	1.0073	0.9939	0.9946
500	0.9987	1.0052	1.0453	1.0074	1.0072

Table 1 shows that both ridge-based model averaging estimators (RMA and HR-RMA columns) deliver improvement in predictive efficiency comparable to that achieved by the MMA, HR-MMA and JMA methods in finite samples. The benefits of RMA and HR-RMA are especially pronounced for smaller sample sizes, though they tend to persist for larger samples as well. Moreover, for smaller samples (n_1 = 50, 75, 100) random splits result relatively often in the singular design

[15] See http://fmwww.bc.edu/ec-p/data/wooldridge/WAGE1.des for a full description of the data.

matrix, thus increasing the motivation for regularization from a practitioner's perspective. HR-RMA tends to have marginally lower out-of-sample predictive efficiency relative to RMA, thus demonstrating a price to pay for robustness to heteroskedasticity.

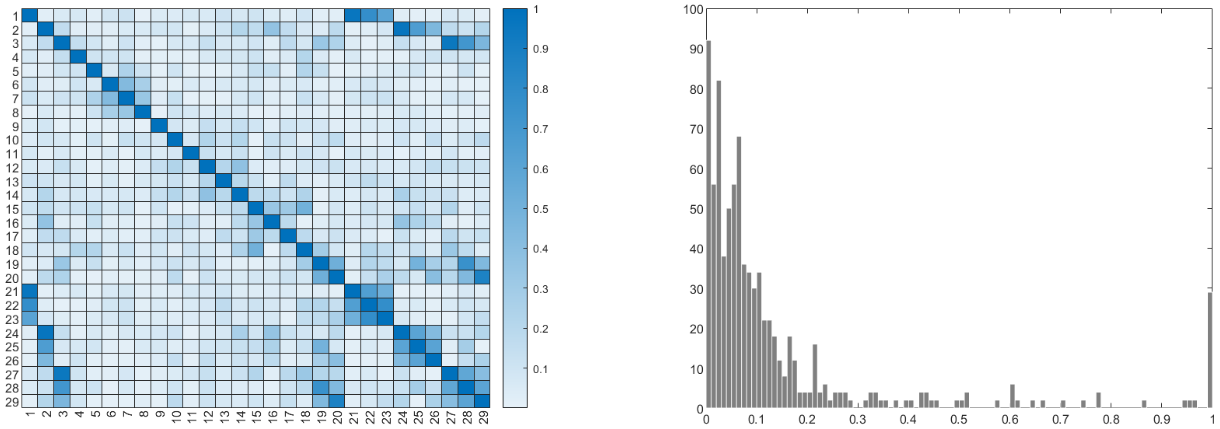


Figure 8: Correlation heatmap and correlation histogram for the wage predictors, in-sample portion of the data $n_1 = 500$. The absolute values of correlations are employed.

This example illustrates the scope of the benefit achieved by the use of ridge-regularized model averaging estimators under relatively low correlations among the predictors. Figure 8 presents a heatmap and histogram for pairwise correlations[16] among the variables for $n_1 = 500$. Notably, the variables are mostly low to moderately correlated, though the correlations are high enough for the ridge regularization to be beneficial. The next subsection presents another example, with moderately to highly correlated predictors, where the relative benefits from using the ridge-based model averaging estimators are even larger.

5.2 Growth Determinants

Next, we work with the dataset collected by Barro and Lee (1994) on cross-country determinants of long-term economic growth. Overall, the dataset includes 60 potential predictors of the average growth rate of GDP between 1960 and 1985 for 90 countries. We use this dataset to predict the growth rate via averaging across different combinations of predictors in the model. The intercept and the logarithm of the initial GDP are always included,[17] and only nested models are considered.

We employ three different schemes for sample-splitting to compare the performance of all estimators:

[16] Absolute values of pairwise correlations are used for the sake of visibility.

[17] Similarly to Belloni et al (2011a) and Giannone at al. (2021) who employ the same dataset for the purpose of prediction.

(Leave-one-out) use all but one country for model estimation to make the predictions for the remaining country, do this for each country,

(Out-of-sample-5) randomly select 85 (out of 90) countries for model estimation to make the predictions for the remaining 5 countries, make 500 draws, then average the results across them,

(Out-of-sample-10) randomly select 80 countries for model estimation to make the predictions for the remaining 10 countries, make 500 draws, then average the results across them.
For each scheme, we compute the average squared prediction error across 1/5/10 countries, respectively. We compare the same methods as before, and all presented statistics are again normalized with respect to the HR-MMA. Table 2 shows that all methods outperform the HR-MMA estimator. Both the RMA and HR-RMA tend to deliver smaller prediction error than the MMA, while the performance of the RMA is similar to that of the JMA. Remarkably, the oldest method, WBIC, does especially well in this example.

Table 2: Average squared prediction error in long-run growth regression (all numbers are normalized over those for HR-MMA)

	MMA	JMA	WBIC	RMA	HR-RMA
Leave-one-out	0.7489	0.4193	0.3324	0.4851	0.7815
Out-of-sample-5	0.6347	0.4422	0.3718	0.4770	0.6109
Out-of-sample-10	0.5861	0.4043	0.3312	0.4369	0.5294

Figure 10 presents the correlation heatmap and histogram, similarly the previous empirical example. Unlike in the previous example, here the predictors are moderately to highly correlated. Correspondingly, in this example we observe bigger improvement attained by the RMA and HR-RMA methods relative to that in the previous example, where the predictors are low to moderately correlated (say, for the sample sizes $n_1 = 75$ and $n_1 = 100$ in the wage prediction example, which are close to the sample sizes employed in the example of the current subsection).

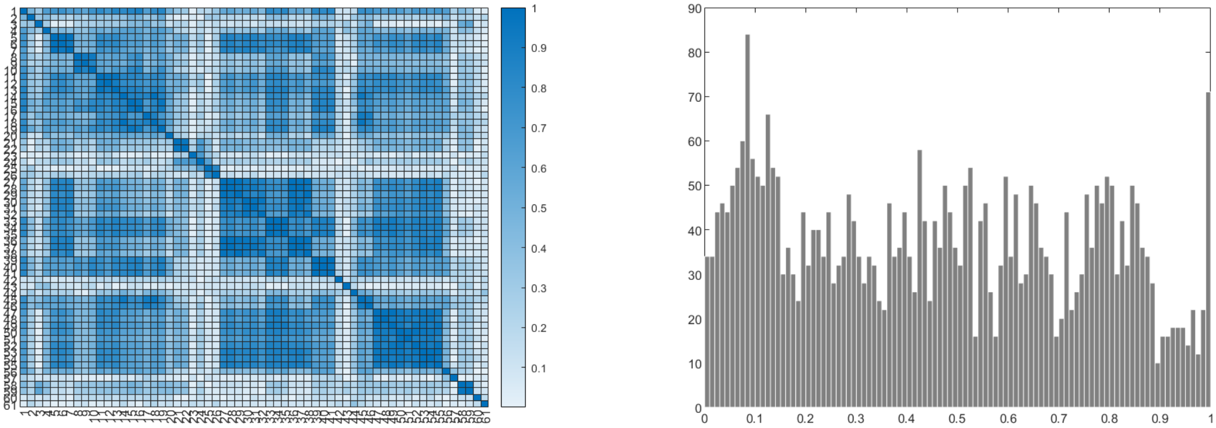


Figure 9: Correlation heatmap and correlation histogram for the growth predictors. The absolute values of correlations are employed.

6 Conclusion

This paper promotes the use of ridge-regularized model averaging estimation. Although the proposed RMA and HR-RMA estimators do not dominate the alternatives uniformly over the parameter space, in most cases they outperform others over the considerable interval of the population R^2. The improvement achieved by ridge regularization may be partially attributed to the changes of weight distribution: the optimal weights found via RMA/HR-RMA tend to be higher for more sophisticated models, while the weights obtained via other procedures are predominantly distributed among low and moderately parametrized specifications.

Two empirical examples also demonstrate the benefits of the ridge-regularized model averaging estimators. Specifically, the RMA tends to deliver better predictions than the MMA, while the HR-RMA outperforms the HR-MMA, especially in small samples. Notably, in both examples the RMA performs better or comparably to the JMA, which may be more computationally intensive. Although in this paper we utilize a rather demanding cross-validation procedure to select the optimal degree of regularization, there are alternative ways to set up the shrinkage parameter (see, for example, Hansen and Kozbur, 2014). While other data-driven approaches may result in the shrinkage parameter deviating from the optimal value, their use may still be beneficial, as shown by Hansen and Kozbur, in particular.

References

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4), 451-468.

Barro, R. J., & Lee, J. W. (1994). Sources of economic growth. In *Carnegie-Rochester conference series on public policy* (Vol. 40, pp. 1-46). North-Holland.

Belloni, A., V. Chernozhukov, and C. Hansen (2011a): "Inference for high-dimensional sparse econometric models," in *Advances in Economics and Econometrics – World Congress of Econometric Society 2010.*

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6), 2350-2383.

Buckland, S., Burnham, K., & Augustin, N. (1997). Model selection: an integral part of inference. *Biometrics*, 603-618.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5), 849-911.

Giannone D., Lenza M., & Primiceri G. (2021). Economic Predictions with Big Data: The Illusion of Sparsity. Econometrica, 89(5), p. 2409-2437.

Goldenshluger, A. (2009). A universal procedure for aggregating estimators. *The Annals of Statistics*, 37(1), 542-568.

Hansen, B. (2007). Least squares model averaging. *Econometrica*, 75(4), 1175-1189.

Hansen, B. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2), 342-350.

Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182(2), 290-308.

Hansen, B., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1), 38-46.

Hjort, N., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464), 879-899.

Hjort, N., & Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101(476), 1449-1464.

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 958-975.

Liao, J., Zou, G., Gao, Y., & Zhang, X. (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics*, 223(1), 190-221.

Liu, Q., & Okui, R. (2013). Heteroscedasticity-robust C_p model averaging. *The Econometrics Journal*, 16(3), 463-472.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4), 661-675.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys*, 29(1), 46-75.

Kapetanios, G., Labhard, V., & Price, S. (2008). Forecasting using Bayesian and information-theoretic model averaging: an application to UK inflation. *Journal of Business & Economic Statistics*, 26(1), 33-41.

Pesaran, M., Schleicher, C., & Zaffaroni, P. (2009). Model averaging in risk management with an application to futures markets. *Journal of Empirical Finance*, 16(2), 280-305.

Steel, M. F. (2017). Model averaging and its use in economics. *arXiv preprint arXiv:1709.08221.*

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2), 277-283.

Wooldridge, J. M. (2003). Introductory Econometrics-A Modern Approach. *Thomson South-Western.*

Xie, T. (2015). Prediction model averaging estimator. *Economics Letters*, 131, 5-8.

Xie, T. (2019). Forecast bitcoin volatility with least squares model averaging. *Econometrics*, 7(3), 40.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454), 574-588.

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how?. *Journal of the American Statistical Association*, 100(472), 1202-1214.

Zhao, S., Liao, J., & Yu, D. (2020). Model averaging estimator in ridge regression and its large sample properties. *Statistical Papers*, 61(4), 1719-1739.

Appendix C

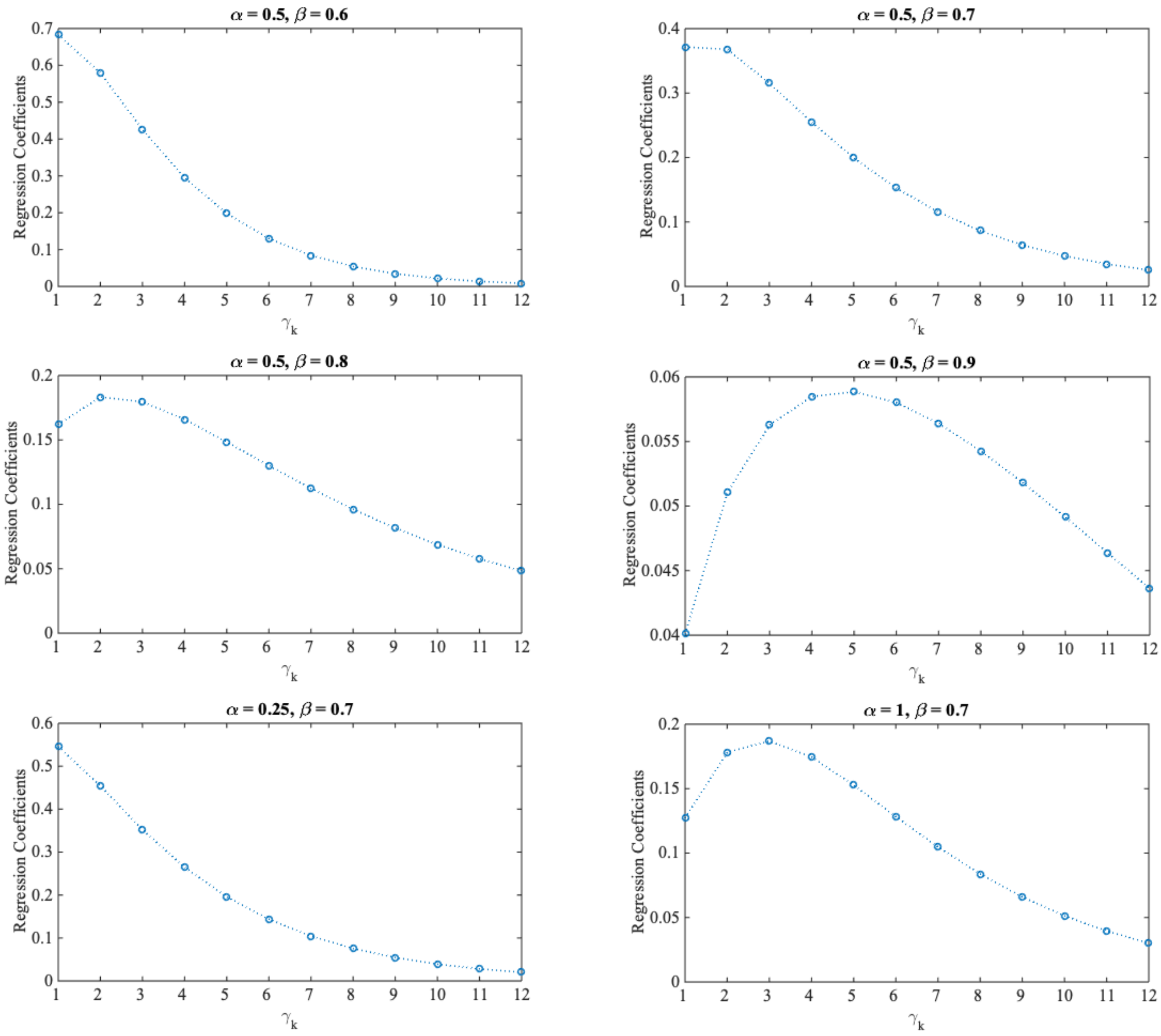


Figure 10: Simulation study: regression coefficients (all graphs are truncated along the horizonal axis)

Appendix E

Part 1

The averaged OLS estimate:

-

$$\begin{aligned}\tilde{\beta} =& w^{ols}\left(\begin{array}{c}\hat{\beta}_1^{ols}\\ 0\end{array}\right)+\left(1-w^{ols}\right)\left(\begin{array}{c}0\\ \hat{\beta}_2^{ols}\end{array}\right)=\left(\begin{array}{c}w^{ols}\hat{\beta}_1^{ols}\\ \left(1-w^{ols}\right)\hat{\beta}_2^{ols}\end{array}\right)\\ =&\left[\begin{array}{cc}w^{ols}&0\\ 0&1-w^{ols}\end{array}\right]\left(\begin{array}{c}\hat{\beta}_1^{ols}\\ \hat{\beta}_2^{ols}\end{array}\right)=W^{ols}\hat{\beta}^{ols}\\ =&\left(\begin{array}{c}w^{ols}\beta_1+w^{ols}\left(X_1'X_1\right)^{-1}X_1'\left(X_2\beta_2+e\right)\\ \left(1-w^{ols}\right)\beta_2+\left(1-w^{ols}\right)\left(X_2'X_2\right)^{-1}X_2'\left(X_1\beta_1+e\right)\end{array}\right)\end{aligned}$$

The bias of the averaged OLS estimate:

$$\begin{aligned}E\left[\tilde{\beta}-\beta\right] =& E\left[w^{ols}\left(\begin{array}{c}\hat{\beta}_1^{ols}\\ 0\end{array}\right)+\left(1-w^{ols}\right)\left(\begin{array}{c}0\\ \hat{\beta}_2^{ols}\end{array}\right)-\left(\begin{array}{c}\beta_1\\ \beta_2\end{array}\right)\right]\\ =& E\left[\left(\begin{array}{c}\beta_1\left(w^{ols}-1\right)+w^{ols}\left(X_1'X_1\right)^{-1}X_1'X_2\beta_2+w^{ols}\left(X_1'X_1\right)^{-1}X_1'e\\ -\beta_2w^{ols}+\left(1-w^{ols}\right)\left(X_2'X_2\right)^{-1}X_2'X_1\beta_1+\left(1-w^{ols}\right)\left(X_2'X_2\right)^{-1}X_2'e\end{array}\right)\right]\\ =& E\left[\begin{array}{c}\beta_1\left(w^{ols}-1\right)+w^{ols}\left(X_1'X_1\right)^{-1}X_1'X_2\beta_2\\ -\beta_2w^{ols}+\left(1-w^{ols}\right)\left(X_2'X_2\right)^{-1}X_2'X_1\beta_1\end{array}\right]\end{aligned}$$

The variance of the averaged OLS estimate:

$$\begin{aligned}Var\left[\tilde{\beta}\right] =& Var\left[\left(\begin{array}{c}w^{ols}\left(X_1'X_1\right)^{-1}X_1'e\\ \left(1-w^{ols}\right)\left(X_2'X_2\right)^{-1}X_2'e\end{array}\right)\right]\\ =&\left[\begin{array}{cc}\left(w^{ols}\right)^2\sigma^2\left(X_1'X_1\right)^{-1} & w^{ols}\left(1-w^{ols}\right)\sigma^2\left(X_1'X_1\right)^{-1}X_1'X_2\left(X_2'X_2\right)^{-1}\\ w^{ols}\left(1-w^{ols}\right)\sigma^2\left(X_2'X_2\right)^{-1}X_2'X_1\left(X_1'X_1\right)^{-1} & \left(1-w^{ols}\right)^2\sigma^2\left(X_2'X_2\right)^{-1}\end{array}\right]\end{aligned}$$

The average of ridge estimates:

$$
\begin{aligned}
\tilde{\beta}\left(\lambda_1,\lambda_2\right) =& w^r W_{\lambda_1}\begin{pmatrix}\hat{\beta}_1^{ols}\\ 0\end{pmatrix} + \left(1-w^r\right)W_{\lambda_2}\begin{pmatrix}0\\ \hat{\beta}_2^{ols}\end{pmatrix} = \begin{pmatrix} w^r W_{\lambda_1}\hat{\beta}_1^{ols}\\ \left(1-w^r\right)W_{\lambda_2}\hat{\beta}_2^{ols}\end{pmatrix}\\
=&\begin{bmatrix} w^r W_{\lambda_1} & 0\\ 0 & \left(1-w^r\right)W_{\lambda_2}\end{bmatrix}\begin{pmatrix}\hat{\beta}_1^{ols}\\ \hat{\beta}_2^{ols}\end{pmatrix} = W^r_{\lambda_1\lambda_2}\hat{\beta}^{ols}\\
=&\begin{pmatrix} w^r W_{\lambda_1}\beta_1 + w^r W_{\lambda_1}\left(X_1'X_1\right)^{-1}X_1'\left(X_2\beta_2+e\right)\\ \left(1-w^r\right)W_{\lambda_2}\beta_2 + \left(1-w^r\right)W_{\lambda_2}\left(X_2'X_2\right)^{-1}X_2'\left(X_1\beta_1+e\right)\end{pmatrix}
\end{aligned}
$$

where $W_{\lambda_1} = \left(X_1'X_1+\lambda I_p\right)^{-1}X_1'X_1$ and $W_{\lambda_2} = \left(X_2'X_2+\lambda I_p\right)^{-1}X_2'X_2$.
The bias of the averaged ridge estimate:

$$
\begin{aligned}
E\left[\tilde{\beta}\left(\lambda_1,\lambda_2\right)-\beta\right] =& E\left[\begin{pmatrix} w^r W_{\lambda_1}\hat{\beta}_1^{ols}\\ \left(1-w^r\right)W_{\lambda_2}\hat{\beta}_2^{ols}\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right]\\
=& E\left[\begin{pmatrix} \left(w^r W_{\lambda_1}-I_p\right)\beta_1 + w^r W_{\lambda_1}\left(X_1'X_1\right)^{-1}X_1'\left(X_2\beta_2+e\right)\\ \left(\left(1-w^r\right)W_{\lambda_2}-I_p\right)\beta_2 + \left(1-w^r\right)W_{\lambda_2}\left(X_2'X_2\right)^{-1}X_2'\left(X_1\beta_1+e\right)\end{pmatrix}\right]\\
=& E\left[\begin{pmatrix} \left(w^r W_{\lambda_1}-I_p\right)\beta_1 + w^r W_{\lambda_1}\left(X_1'X_1\right)^{-1}X_1'X_2\beta_2\\ \left(\left(1-w^r\right)W_{\lambda_2}-I\right)\beta_2 + \left(1-w^r\right)W_{\lambda_2}\left(X_2'X_2\right)^{-1}X_2'X_1\beta_1\end{pmatrix}\right]
\end{aligned}
$$

The variance of the averaged ridge estimate:

$$
\begin{aligned}
Var\left[\tilde{\beta}\left(\lambda_1,\lambda_2\right)\right] =& Var\begin{bmatrix} w^r W_{\lambda_1}\left(X_1'X_1\right)^{-1}X_1'e\\ \left(1-w^r\right)W_{\lambda_2}\left(X_2'X_2\right)^{-1}X_2'e\end{bmatrix}\\
=&\begin{bmatrix} V_{11} & V_{21}'\\ V_{21} & V_{22}\end{bmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
V_{11} =& \left(w^r\right)^2\sigma^2 W_{\lambda_1}\left(X_1'X_1\right)^{-1}W_{\lambda_1}'\\
V_{21} =& w^r\left(1-w^r\right)\sigma^2 W_{\lambda_2}\left(X_2'X_2\right)^{-1}X_2'X_1\left(X_1'X_1\right)^{-1}W_{\lambda_1}'\\
V_{22} =& \left(1-w^r\right)^2\sigma^2 W_{\lambda_2}\left(X_2'X_2\right)^{-1}W_{\lambda_2}'
\end{aligned}
$$

The mean squared error of the averaged ridge estimate $\tilde{\beta}\left(\lambda_1,\lambda_2\right)$:

$$
\mathrm{MSE}\left(\tilde{\beta}\left(\lambda_1,\lambda_2\right)\right) = \mathrm{tr}\left[Var\left[\tilde{\beta}\left(\lambda_1,\lambda_2\right)\right]\right] + \left[E\left(\tilde{\beta}\left(\lambda_1,\lambda_2\right)-\beta\right)\right]^2
$$

The variance component is the trace of the variance matrix:

$$\begin{aligned}\text{tr}\left[Var\left[\tilde{\beta}\left(\lambda_1,\lambda_2\right)\right]\right] =& \text{tr}\left[V_{11}\right]+\text{tr}\left[V_{22}\right]\\ =& \text{tr}\left[\left(w^r\right)^2 W_{\lambda_1}\left(X_1'X_1\right)^{-1}W_{\lambda_1}'\right]+\text{tr}\left[\left(1-w^r\right)^2 W_{\lambda_2}\left(X_2'X_2\right)^{-1}W_{\lambda_2}'\right]\end{aligned}$$

From now on assume $X_1'X_1 = I_p = X_2'X_2$ and $X_1'X_2 = X_2'X_1 = \rho I_p$:

$$\begin{aligned}\text{tr}\left[\left(w^r\right)^2 W_{\lambda_1}\left(X_1'X_1\right)^{-1}W_{\lambda_1}'\right] &= \left(w^r\right)^2 tr\left[W_{\lambda_1}\left(X_1'X_1\right)^{-1}W_{\lambda_1}'\right]\\ &= \left(w^r\right)^2 tr\left[W_{\lambda_1}W_{\lambda_1}'\right] = \left(w^r\right)^2\frac{p}{\left(1+\lambda_1\right)^2}\\ \text{tr}\left[\left(1-w^r\right)^2 W_{\lambda_2}\left(X_2'X_2\right)^{-1}W_{\lambda_2}'\right] &= \left(1-w^r\right)^2\frac{p}{\left(1+\lambda_2\right)^2}\end{aligned}$$

Therefore,

$$\text{tr}\left[Var\left[\tilde{\beta}\left(\lambda_1,\lambda_2\right)\right]\right] = \left(w^r\right)^2\frac{p}{\left(1+\lambda_1\right)^2}+\left(1-w^r\right)^2\frac{p}{\left(1+\lambda_2\right)^2} \tag{4}$$

Now its squared bias:

$$\begin{aligned}\left[E\left(\tilde{\beta}\left(\lambda_1,\lambda_2\right)-\beta\right)\right]^2 =& E\left[\begin{array}{c}\left(w^rW_{\lambda_1}-I\right)\beta_1+w^rW_{\lambda_1}\rho I_p\beta_2\\ \left(\left(1-w^r\right)W_{\lambda_2}-I\right)\beta_2+\left(1-w^r\right)W_{\lambda_2}\rho I_p\beta_1\end{array}\right]^T\times\\ &\times E\left[\begin{array}{c}\left(w^rW_{\lambda_1}-I\right)\beta_1+w^rW_{\lambda_1}\rho I_p\beta_2\\ \left(\left(1-w^r\right)W_{\lambda_2}-I\right)\beta_2+\left(1-w^r\right)W_{\lambda_2}\rho I_p\beta_1\end{array}\right]\end{aligned}$$

$$\begin{aligned}\left[E\left(\tilde{\beta}\left(\lambda_1,\lambda_2\right)-\beta\right)\right]^2 =& \beta_1^T\left[\frac{\left(w^r\right)^2-2w^r\left(1+\lambda_1\right)+\left(1+\lambda_1\right)^2}{\left(1+\lambda_1\right)^2}+\frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}\rho^2\right]\beta_1 && (5)\\ &+\beta_1^T\rho\left[\frac{2w^r\left(w^r-1-\lambda_1\right)}{\left(1+\lambda_1\right)^2}-\frac{2\left(w^r+\lambda_2\right)\left(1-w^r\right)}{\left(1+\lambda_2\right)^2}\right]\beta_2 && (6)\\ &+\beta_2'\left[\frac{\left(w^r\right)^2}{\left(1+\lambda_1\right)^2}\rho^2+\left(\frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}-\frac{2\left(1-w^r\right)}{1+\lambda_2}+1\right)\right]\beta_2 && (7)\end{aligned}$$

So, the desired MSE is (10) + (11) + (12) + (13).

Part 2

For the first model estimated via ridge:

$$
\begin{aligned}
MSE\left[\hat{\beta}_1\left(\lambda_1\right)\right] =& E\left[\left(W_{\lambda_1}\hat{\beta}_1-\beta_1\right)'\left(W_{\lambda_1}\hat{\beta}_1-\beta_1\right)\right]\\
=& E\left[\hat{\beta}_1{}'W'_{\lambda_1}W_{\lambda_1}\hat{\beta}_1\right]-E\left[\beta_1'W_{\lambda_1}\hat{\beta}_1\right]-E\left[\hat{\beta}_1{}'W'_{\lambda_1}\beta_1\right]+E\left[\beta_1^{'}\beta_1\right]\\
=& E\left[\hat{\beta}_1{}'W'_{\lambda_1}W_{\lambda_1}\hat{\beta}_1\right]-E\left[\beta_1'W'_{\lambda_1}W_{\lambda_1}\hat{\beta}_1\right]-E\left[\hat{\beta}_1'W'_{\lambda_1}W_{\lambda_1}\beta_1\right]+E\left[\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1\right]\\
&-E\left[\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1\right]+E\left[\beta_1'W'_{\lambda_1}W_{\lambda_1}\hat{\beta}_1\right]+E\left[\hat{\beta}_1'W'_{\lambda_1}W_{\lambda_1}\beta_1\right]\\
&-E\left[\beta_1'W_{\lambda_1}\hat{\beta}_1\right]-E\left[\hat{\beta}_1'W'_{\lambda_1}\beta_1\right]+E\left[\beta_1^{'}\beta_1\right]\\
=& E\left[\left(\hat{\beta}_1-\beta_1\right)'W'_{\lambda_1}W_{\lambda_1}\left(\hat{\beta}_1-\beta_1\right)\right]\\
&-\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1+\beta_1'W'_{\lambda_1}W_{\lambda_1}E\left[\hat{\beta}_1\right]+E\left[\hat{\beta}_1'\right]W'_{\lambda_1}W_{\lambda_1}\beta_1\\
&-\beta_1'W_{\lambda_1}E\left[\hat{\beta}_1\right]-E\left[\hat{\beta}_1'\right]W'_{\lambda_1}\beta_1+\beta_1^{'}\beta_1
\end{aligned}
$$

$$
\begin{aligned}
MSE\left[\hat{\beta}_1\left(\lambda_1\right)\right] =& E\left[\left(\hat{\beta}_1-\beta_1\right)'W'_{\lambda_1}W_{\lambda_1}\left(\hat{\beta}_1-\beta_1\right)\right]\\
&-\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1+\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1+\beta_1'W'_{\lambda_1}W_{\lambda_1}\beta_1\\
&-\beta_1'W_{\lambda_1}\beta_1-\beta_1'W'_{\lambda_1}\beta_1+\beta_1^{'}\beta_1\\
&+\beta_1'W'_{\lambda_1}W_{\lambda_1}B+B'W'_{\lambda_1}W_{\lambda_1}\beta_1\\
&-\beta_1'W_{\lambda_1}B-B'W'_{\lambda_1}\beta_1\\
=& E\left\{\left(\hat{\beta}-\beta\right)'W'_{\lambda}W_{\lambda}\left(\hat{\beta}-\beta\right)\right\}+\beta'\left(W_\lambda-I_{pp}\right)'\left(W_\lambda-I_{pp}\right)\beta\\
&+\beta_1'W'_{\lambda_1}W_{\lambda_1}B+B'W'_{\lambda_1}W_{\lambda_1}\beta_1-\beta_1'W_{\lambda_1}B-B'W'_{\lambda_1}\beta_1
\end{aligned}
$$

where $B=\left(X_1'X_1\right)^{-1}X_1'X_2\beta_2$.

Under $X_1'X_1=I$,

$$\begin{aligned}
MSE\left[\hat{\beta}_1\left(\lambda_1\right)\right] =& \frac{p\sigma^2}{\left(1+\lambda_1\right)^2}+\frac{\lambda_1^2}{\left(1+\lambda_1\right)^2}\beta_1'\beta_1 \\
&+\beta_1'\left(I+\lambda_1 I\right)^{-1}\left(I+\lambda_1 I\right)^{-1}X_1'X_2\beta_2 \\
&+\beta_2 X_2'X_1\left(I+\lambda_1 I\right)^{-1}\left(I+\lambda_1 I\right)^{-1}\beta_1 \\
&-\beta_1'\left(I+\lambda_1 I\right)^{-1}X_1'X_2\beta_2 \\
&-\beta_2'X_2'X_1\left(I+\lambda_1 I\right)^{-1}\beta_1 \\
=& \frac{p\sigma^2}{\left(1+\lambda_1\right)^2}+\frac{\lambda_1^2}{\left(1+\lambda_1\right)^2}\beta_1'\beta_1 \\
&+\frac{1}{\left(1+\lambda_1\right)^2}\beta_1'X_1'X_2\beta_2+\frac{1}{\left(1+\lambda_!\right)^2}\beta_2X_2'X_1\beta_1 \\
&-\frac{1}{1+\lambda_1}\beta_1'X_1'X_2\beta_2-\frac{1}{1+\lambda_1}\beta_2'X_2'X_1\beta_1
\end{aligned}$$

$$\begin{aligned}
MSE\left[\hat{\beta}_1\left(\lambda_1\right)\right] =& \frac{p\sigma^2}{\left(1+\lambda_1\right)^2}+\frac{\lambda_1^2}{\left(1+\lambda_1\right)^2}\beta_1'\beta_1 \\
&+\frac{2}{\left(1+\lambda_1\right)^2}\beta_1'X_1'X_2\beta_2 \\
&-\frac{2}{1+\lambda_1}\beta_1'X_1'X_2\beta_2
\end{aligned}$$

The derivative w.r.t. λ_1 provides us with the optimal value of shrinkage for the first model:

$$-\tfrac{2p\sigma^2}{\left(1+\lambda_1\right)^3}+\tfrac{2\lambda_1\left(1+\lambda_1\right)^2-2\lambda_1^2\left(1+\lambda_1\right)}{\left(1+\lambda_1\right)^4}\beta_1'\beta_1-\tfrac{4}{\left(1+\lambda_1\right)^3}\beta_1'X_1'X_2\beta_2+\tfrac{2}{\left(1+\lambda_1\right)^2}\beta_1'X_1'X_2\beta_2=0$$

$$\lambda_1^{opt}=\frac{p\sigma^2+\beta_1'X_1'X_2\beta_2}{\beta_1'\beta_1+\beta_1'X_1'X_2\beta_2}.$$

Similarly, for the second model:

$$\lambda_2^{opt}=\frac{p\sigma^2+\beta_1'X_1'X_2\beta_2}{\beta_2'\beta_2+\beta_1'X_1'X_2\beta_2}.$$

Appendix H

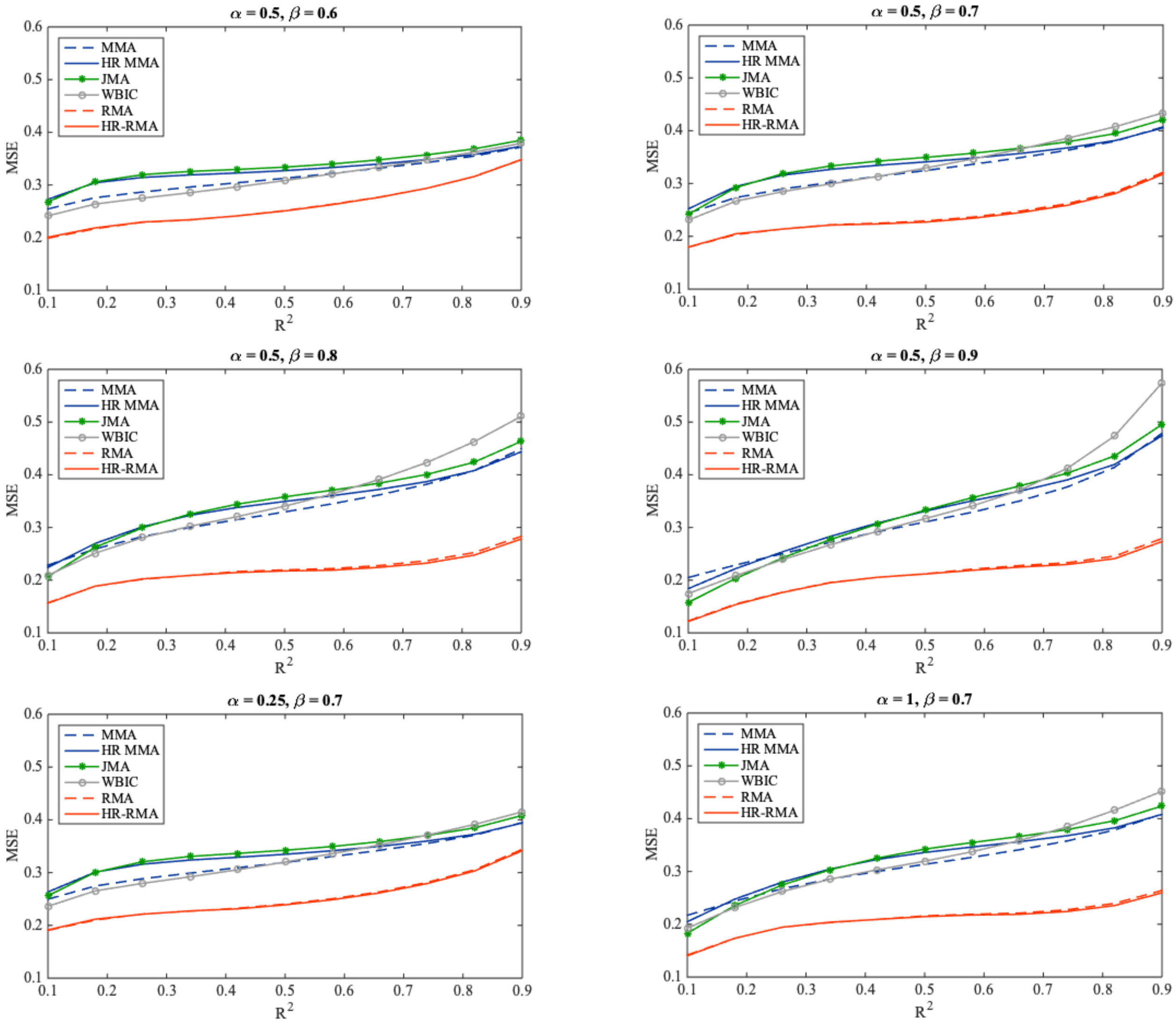


Figure 11: $n = 100$. Case [H] of moderate correlation among predictors

Appendix W

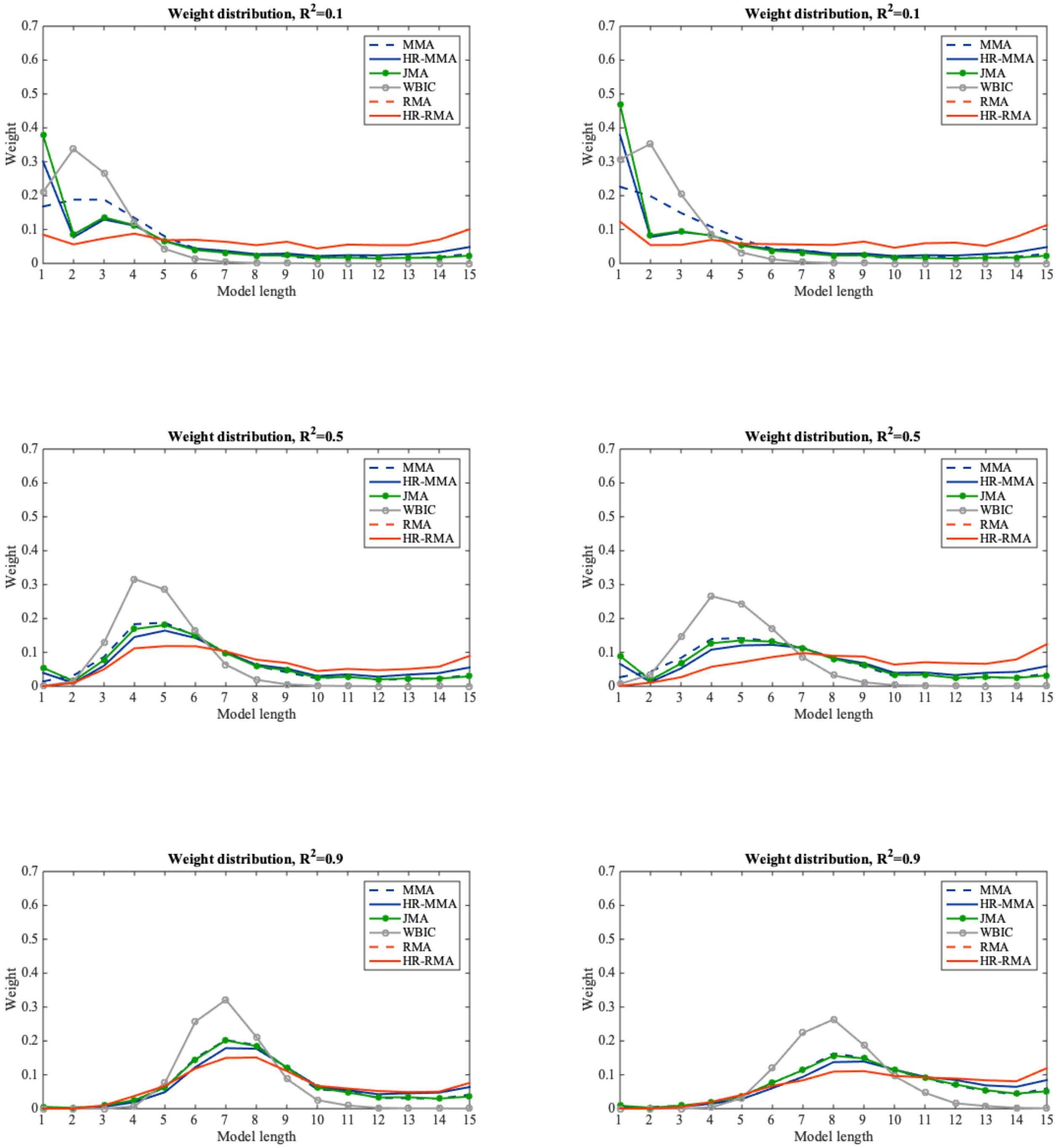


Figure 12: Optimal weights, $\alpha = 0.5$, $\beta = [0.6, 0.7]$ (left to right), $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.

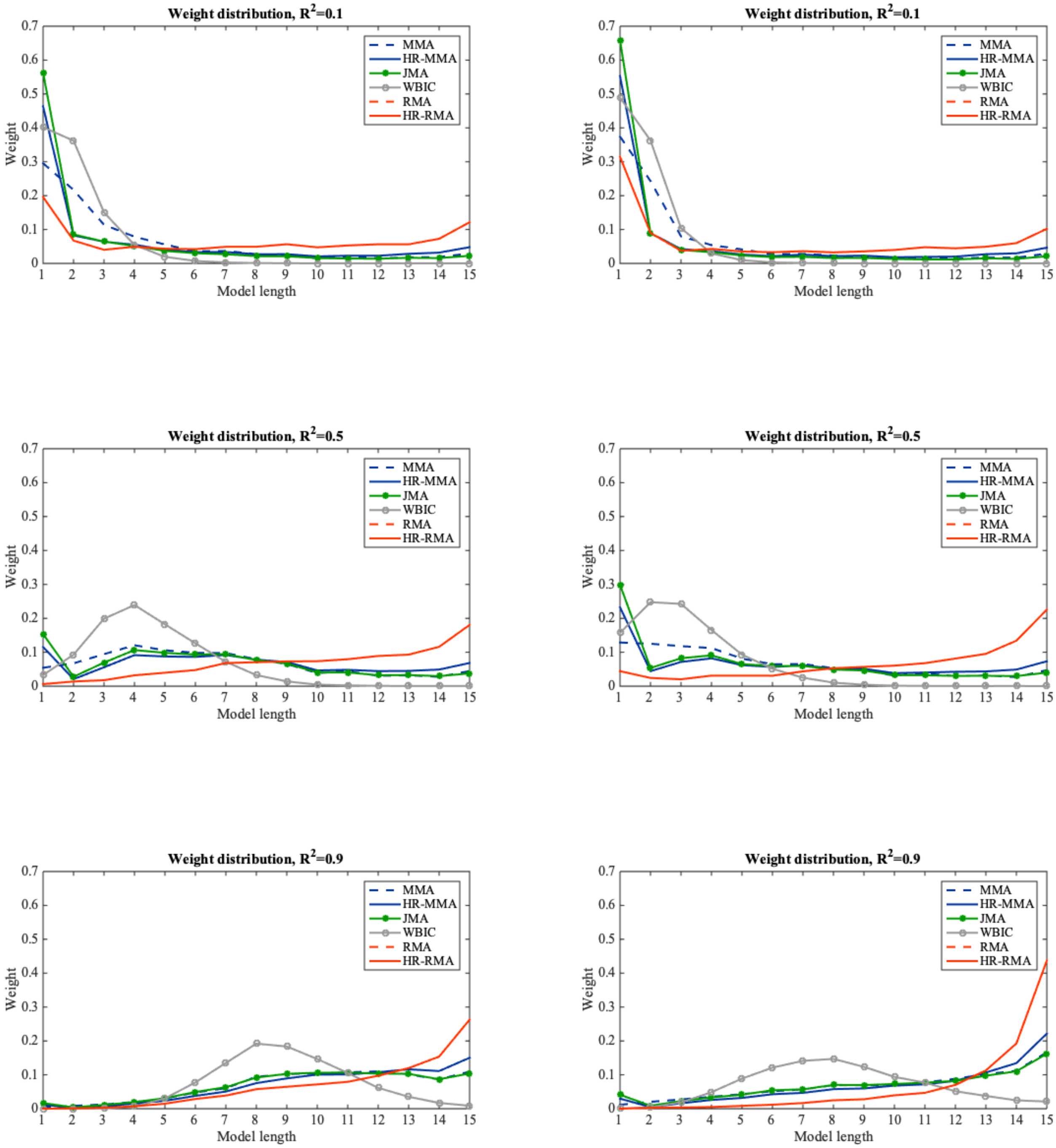
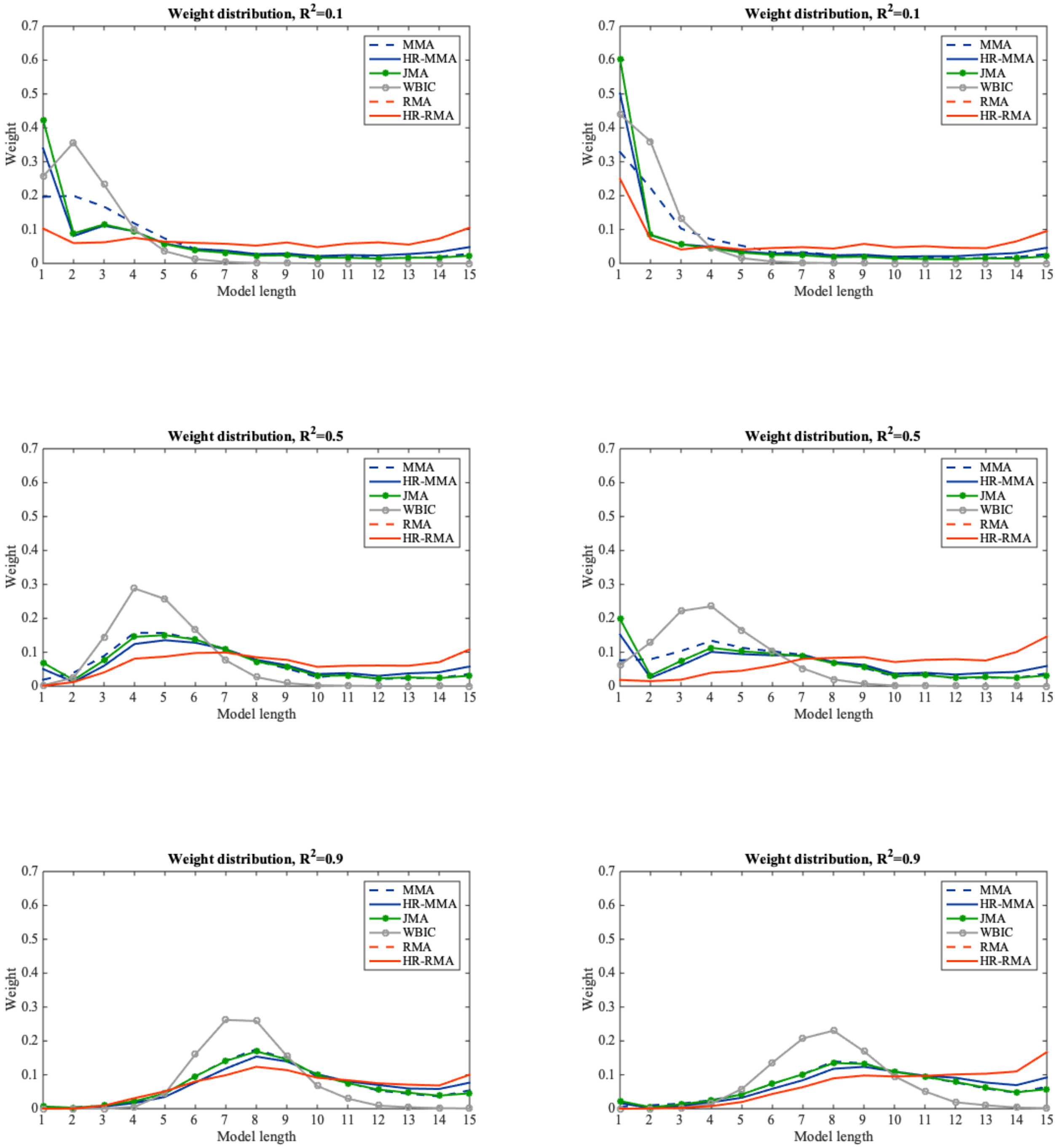


Figure 13: Optimal weights, $\alpha = 0.5$, $\beta = [0.8, 0.9]$ (left to right), $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.


Weight distribution, R^2=0.1

Weight distribution, R^2=0.1

Weight distribution, R^2=0.5

Weight distribution, R^2=0.5

Weight distribution, R^2=0.9

Weight distribution, R^2=0.9

Figure 14: Optimal weights, $\alpha = [0.25, 1]$ (left to right), $\beta = 0.7$, $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.