CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Center for Economic Research and Graduate Education

Dissertation Thesis

2023 **Alena Skolkova**

CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Center for Economic Research and Graduate Education

Alena Skolkova

Essays on Model Uncertainty and Model Averaging

Dissertation Thesis

Prague 2023

Author: **Alena Skolkova**
Supervisor: **Prof. Ing. Štěpán Jurajda, Ph.D.**
Year of the defense: 2023

References

SKOLKOVA, Alena. Essays on Model Uncertainty and Model Averaging. Praha, 2023. 64 pages. Dissertation thesis (PhD.). Charles University, Faculty of Social Sciences, Center for Economic Research and Graduate Education – Economics Institute. Supervisor Prof. Ing. Štěpán Jurajda, Ph.D.

Abstract

In the first chapter of this dissertation I study the properties of a model averaging estimator with ridge regularization. I propose the ridge-regularized modifications of Mallows model averaging (Hansen, 2007, *Econometrica*, 75) and heteroskedasticity-robust Mallows model averaging (Liu and Okui, 2013, *The Econometrics Journal*, 16) to leverage the capabilities of averaging and ridge regularization simultaneously. Via a simulation study, I 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 two-model theoretical example shows that the relative reduction of mean squared error is increasing with the strength of the correlation. I also demonstrate the superiority of the ridge-regularized modifications via empirical examples focused on wages and economic growth.

The second chapter focuses on the use of elastic-net regression for instrumental variable estimation. I investigate the relative performance of the lasso and elastic-net estimators for fitting the first-stage as part of IV estimation. Because elastic-net includes a ridge-type penalty in addition to a lasso-type penalty, it generally improves upon lasso in finite samples when correlations among the instrumental variables are not negligible. I show that IV estimators based on the lasso and elastic-net first-stage estimates can be asymptotically equivalent. Via a Monte Carlo study, I demonstrate the robustness of the sample-split elastic-net IV estimator to deviations from approximate sparsity, and to correlation among instruments that may be high-dimensional. Finally, I provide an empirical example that demonstrates potential improvement in estimation accuracy gained by the use of IV estimators based on elastic-net.

The third chapter, a joint work with S. Anatolyev, contributes to wider use of advanced conventional methods for dealing with instrumental variable regression with many, possibly weak, instruments in Stata. We introduce a STATA command, `mivreg`, that implements consistent estimation and testing in linear IV regressions with many instruments, which may be weak.

Abstrakt

V první kapitole této disertační práce navrhuji a studuji vlastnosti modelového průměrovacího estimátoru s hřebenovou regularizací. Navrhuji ridge-regularizační modifikace Mallowsova průměrování modelu(Hansen, 2007, *Econometrica*, 75) a Mallowsova průměrování modelu robustního vůči heteroskedasticitě(Liu and Okui, 2013, *The Econometrics Journal*, 16), abychom současně využili schopnosti průměrování a ridge regularizace. Prostřednictvím

simulační studie dokumentuji vylepšení na konečném vzorku dat, což je důsledkem nahrazení nejmenších čtverců ridge regresí. Průměrování na základě ridge modelu je zvláště užitečné, když se zabýváme množstvími středně až vysoce korelovaných prediktorů, protože základní ridge regrese se korelované prediktory akomoduje, aniž by došlo k nafouknutí rozptylu odhadů. Jednoduchý teoretický příklad ukazuje, že relativní snížení střední kvadratické chyby roste se silou korelace. Na empirických příkladech, zaměřených na mzdy a ekonomický růst, dále demonstruji přednost ridge regularizovaných modifikací.

Druhá kapitola se zaměřuje na použití elastic net regrese pro instrumentální odhad proměnných. Zkoumám relativní výkon odhadů lassa a elastic net pro predikované hodnoty prvního stupně jako součást odhadu IV. Jelikož elastic net obsahuje kromě penalizace typu lasso penalizaci typu ridge, obecně se oproti lassu v konečných vzorcích zlepšuje, když korelace mezi instrumentálními proměnnými nejsou zanedbatelné. Ukazuji, že IV odhady založené na odhadech lassa a elastic net v prvním stupni mohou být asymptoticky ekvivalentní. Prostřednictvím Monte Carlo studie demonstruji robustnost estimátoru elastic net IV s rozděleným vzorkem dat vůči odchylkám od přibližné řídkosti a vůči korelaci mezi potenciálně mnohorozměrnými instrumenty. Nakonec uvádím empirický příklad, který demonstruje potenciální zlepšení přesnosti odhadu získané použitím IV odhadů založených na elastic net.

Třetí kapitola, společná práce se S. Anatolyevem, přispívá k širšímu využití pokročilých konvenčních metod pro práci s regresí s mnoha, potenciálně slabými, instrumentálními proměnnými ve Statě. Zavádíme příkaz, mivreg, který implementuje konzistentní odhad a testování v lineárních IV regresích s mnoha instrumenty, které mohou být slabé.

Keywords

Econometrics; Model averaging; Ridge regularization; IV estimation; Elastic-net regression

Klíčová slova

Ekonometrie; Modelové průměrování; Hřebenová regularizace; IV odhad; Elastic net regrese

Length of the work: 120,373 characters with spaces, without abstract and appendices

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on
25.10.2023

Alena Skolkova

Acknowledgement

This dissertation thesis would not have existed without the continuous support from my supervisor, Professor Štěpán Jurajda, whose patience does not have finite moments.

Table of Contents

Introduction

Increasing availability of covariate-rich datasets creates new challenges encountered in applied econometrics. While classic model selection methods have been predominant for dealing with model uncertainty for decades, more modern methods with embedded regularization often have favourable asymptotic and finite-sample properties. Each chapter of this thesis presents a setup in which data dimensionality deteriorates the performance of traditional methods, and highlights ways to address the issues.

The first chapter contributes to the literature on model uncertainty and model averaging for prediction problems. When a model for determination of a specific variable is not precisely dictated by theory, one often faces a trade-off between a parsimonious model with few variables and a sophisticated model with potentially high-dimensional sets of predictors. While a parsimonious model delivers estimates with a low variance and large bias, a sophisticated model tends to do exactly the opposite. Therefore, combining models with different numbers of variables generally reduces the mean squared error of the resulting predictions. Many methods for finding the optimal combination exist. A leading method is based on generalization of the Mallows (1973) model selection criterion to the Mallows model averaging criterion by Hansen (Hansen, 2007).

I propose a ridge-regularized Mallows model averaging estimator. The ridge model averaging estimator (RMA) ensures better finite-sample properties via ridge regularization of the design matrices corresponding to the models being averaged. In principle, ridge regression and model averaging serve a similar purpose, minimization of the mean squared error through shrinkage, though in different ways. While model averaging, e.g., as in Hansen (2007), reduces the asymptotic mean squared error, ridge regularization leads to finite-sample improvements. Therefore, combining model averaging with ridge regularization results in an estimator that inherits asymptotic optimality, and, in addition, yields better finite-sample properties due to ridge regularization.

I suggest ridge-based modifications of both Mallows model averaging (Hansen, 2007) and heteroskedasticity-robust Mallow model averaging (Liu and Okui, 2013). A tractable theoretical example with two models demonstrates that the relative reduction of the mean squared error is increasing with the strength of predictor correlatedness. Via a simulation study, I examine the finite-sample improvements obtained by replacing ordinary least-squares with a ridge regression for model averaging prediction. Ridge-based model averaging is shown to be superior when one deals with sets of moderately to highly correlated predictors, because underlying ridge regressions accommodate correlated predictors without blowing up estimation variance. I also show the superiority of the ridge-regularized estimator modifications via empirical examples focused on wages and economic growth.

The second chapter contributes to the literature on estimation of treatment effects in a non-experimental setting with many instrumental variables (IV). While the use of many instruments improves estimation accuracy, dealing with high-dimensional sets of instrumental variables can be complicated, and often requires instrument selection or regularization of the first-stage regression. Currently, lasso is established as the most popular method for

simultaneous variable selection and regularization.

I advocate the use of elastic-net in place of lasso in the first-stage regression. The motivation is twofold. First, elastic-net combines lasso regularization with ridge penalization, and thus it generally improves over lasso in finite samples if correlations among the instrumental variables are significant. Second, by attaining a balance between lasso and ridge penalties, elastic-net accommodates deviations of the first-stage equation from a sparse structure, and thus is a robust alternative to lasso, which relies heavily on the sparsity assumption.

I claim that IV estimators that employ lasso and elastic-net first-stage predictions under sparsity are asymptotically equivalent. Via a Monte Carlo study, I demonstrate the robustness to correlation among the instruments and deviations from sparsity of the sample-split IV estimation based on elastic-net first-stage estimates. The cross-fitted elastic-net IV estimator tends to performs similarly to the sample-split version, though sometimes it results in minor test size distortions. Finally, I provide an empirical example that employs the proposed methods to estimation of returns to schooling. The example demonstrates the cross-fitted elastic-net IV estimator that results in the point estimate without a clear bias towards the OLS estimate, while delivering the smallest standard errors. As expected, the sample-split elastic-net IV estimator appears to be more vulnerable to random splits of the real data. However, similarly to the cross-fitted elastic-net IV estimator, it continues to produce reasonable estimates even when its lasso-based counterpart does not select any variables into the first-stage regression, and thus fails to deliver any estimates.

The third chapter, a joint work with S. Anatolyev, contributes to wider use of advanced conventional methods to deal with instrumental variable regression with many instruments, which may be weak, in Stata. Over recent decades, econometric tools for handling instrumental variable regressions with many instruments have been developed. However, practitioners rarely use appropriate tools because they are not available in popular econometric packages, STATA in particular. We introduce a STATA command, `mivreg`, that implements consistent estimation and testing in linear IV regressions with many (possibly weak) instruments. The command `mivreg` covers both homoskedastic and heteroskedastic environments, estimators that are both non-robust and robust to error non-normality and projection matrix limit, both parameter tests and specification tests, and both with and without correction for the existence of moments. We also run a small simulation experiment using `mivreg` and illustrate how `mivreg` works with real data.

1 Model Averaging with Ridge Regularization

Published as CERGE-EI Working Paper Series No 758.

1.1 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 was 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 1974) 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 (Hansen 2007, Hansen 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 and Okui (2013) introduced a heteroskedastiticty-robust Mallows criterion for model averaging (hereafter HR-MMA).

In this paper, I 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 the heteroskedasticity-robust ridge model averaging (HR-RMA) estimator, respectively. The ridge regression, introduced by Hoerl and Kennard (1970), is a generalization of the 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), I replace ordinary least-squares estimation with a ridge regression to minimize the consequences of correlation among predictors. The 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 the proposed

[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 (2020) provide comprehensive reviews of model averaging in economics.

estimator is the jackknife model averaging (JMA) estimator by Hansen and Racine (2012), which is a regularization-free baseline of the MA-Ridge estimator by Zhao et al. (2020). However, the jackknife model averaging by Hansen and Racine (2012) is based on OLS regressions, and thus is not suited for the cases when the number of predictors approaches or exceeds the sample size.

In a Monte Carlo study I 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. Our simulation design is close to that adopted in Hansen (2007, 2008), while I 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 few regressors), and sophisticated models (with many regressors), as well as moderately parametrized models. 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. 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.

I demonstrate how the proposed estimator works in two empirical examples. I employ the cross-section earning data used by Hansen and Racine (2012) and 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 general model averaging estimator, and a ridge-regularized model averaging estimator. Section 3 presents a two-model 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.

1.2 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 $\left\{\hat{\mu}^1, ..., \hat{\mu}^M\right\}$ 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 $\left\{\hat{\mu}^1, ..., \hat{\mu}^M\right\}$. 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 $\left\{\hat{\mu}^1, ..., \hat{\mu}^M\right\}$ to attain further reduction of the MSE. Consider $\mathbf{w} = \left(w^1, ..., w^M\right)'$, 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 $\hat{\mu} = \left(\hat{\mu}^1, ..., \hat{\mu}^M\right)$ 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

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

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 that is conveyed by the trace of the matrix $\mathbf{P}(\mathbf{w})$:

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

[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.

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

$$L_n(\mathbf{w}) = (\mu_t - \hat{\mu}(\mathbf{w}))'(\mu_t - \hat{\mu}(\mathbf{w})).$$

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

$$E[C_n(\mathbf{w})] = \mathbb{E}[L_n(\mathbf{w})] + n\sigma^2,$$

so that the weights found through minimization of $C_n(\mathbf{w})$ also minimize $L_n(\mathbf{w})$, 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 $L_n(\hat{\mathbf{w}}) / \inf_{\mathbf{w}\in\mathcal{H}_n(N)} L_n(\mathbf{w}) \to_p 1$, where $\mathcal{H}_n(N)$ 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:

$$HRC_p(\mathbf{w}) \equiv \|\mathbf{y} - \mathbf{P}(\mathbf{w})\mathbf{y}\|^2 + 2\text{tr}[\Omega\mathbf{P}(\mathbf{w})],$$

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

I 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

[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.

[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.

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

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, I 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, I investigate the finite-sample benefits of the proposed regularized modifications from the same class, RMA and HR-RMA, relative to the baselines of MMA and HR-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, I provide a toy example demonstrating the relative performance of the RMA estimator.

1.3 Theory: A Two-Model Example

In this subsection I consider a toy theoretical example that illustrates the mechanics of the MMA and RMA estimators under homoskedasticity of the error term. First, I derive the MSE for the averaged least-squares and ridge estimates. Then, I 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

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

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, $\text{rank}(X_1) = \text{rank}(X_2) = p$.

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

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

and the average least-squares estimate is

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

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

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

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

and the average ridge estimate is

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

where $W_{\lambda_1} = (X_1'X_1 + \lambda_1 I)^{-1} X_1'X_1$, $W_{\lambda_2} = (X_2'X_2 + \lambda_2 I)^{-1} X_2'X_2$ 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

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

Derivations are provided in Appendix 1.E, Part 1. For both $MSE^{ols}\left(w^{ols}\right)$ and $MSE^r(\lambda_1, \lambda_2, w^r)$, the first term of the 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

[9] I 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.

I made earlier,

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

Derivations are provided in Appendix 1.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.

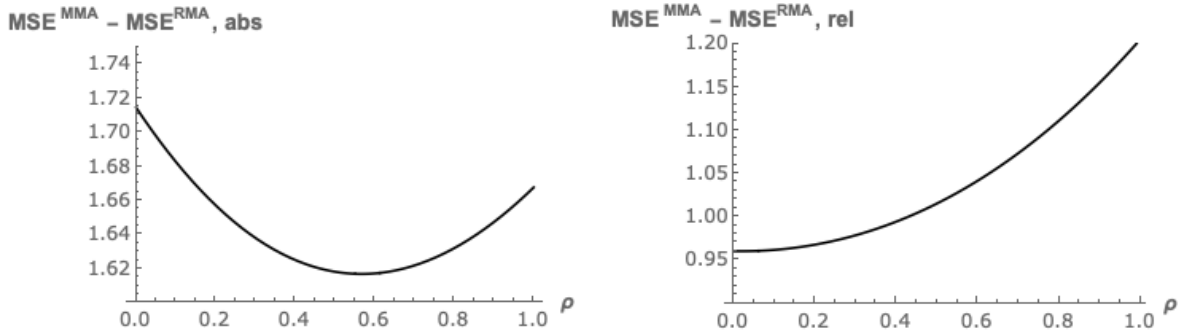

Figure 1: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). Baseline case: $p = 3$, $\sigma^2 = 2$, $\beta_1'\beta_1 = \beta_2'\beta_2 = 1$.

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.

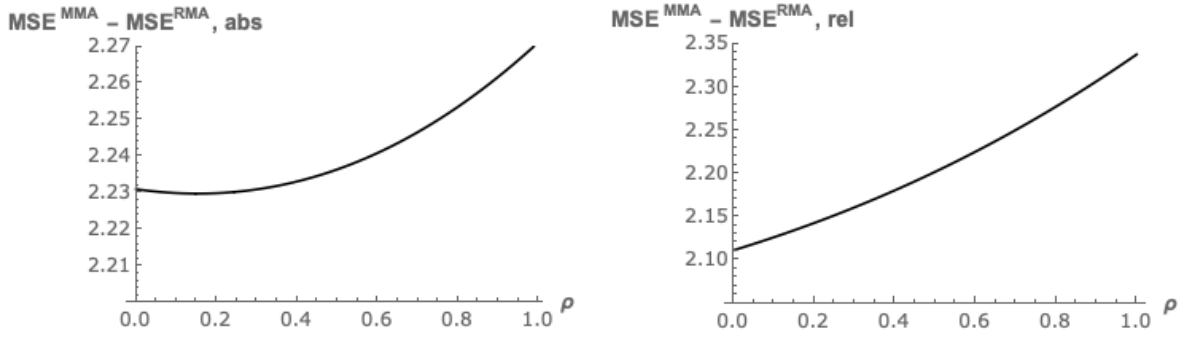

Figure 2: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $\beta_1'\beta_1 = 0.2$

[10] Available upon request.

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 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.

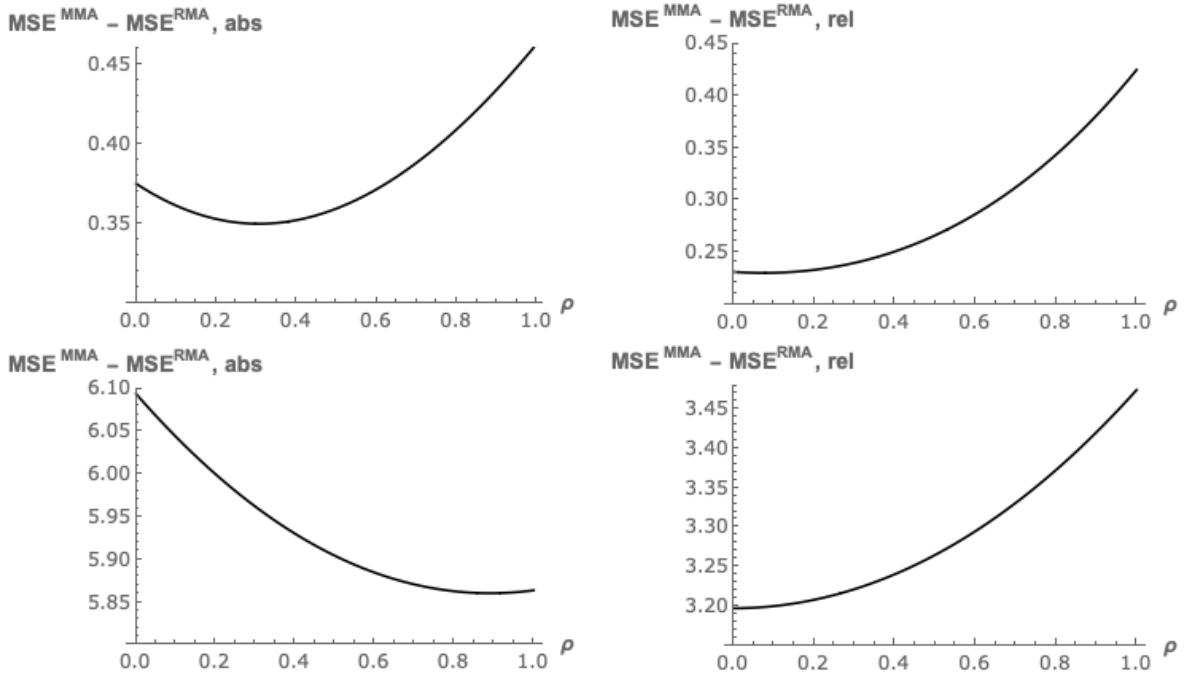

Figure 3: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $\sigma^2 = 1$ (top) and $\sigma^2 = 5$ (bottom)

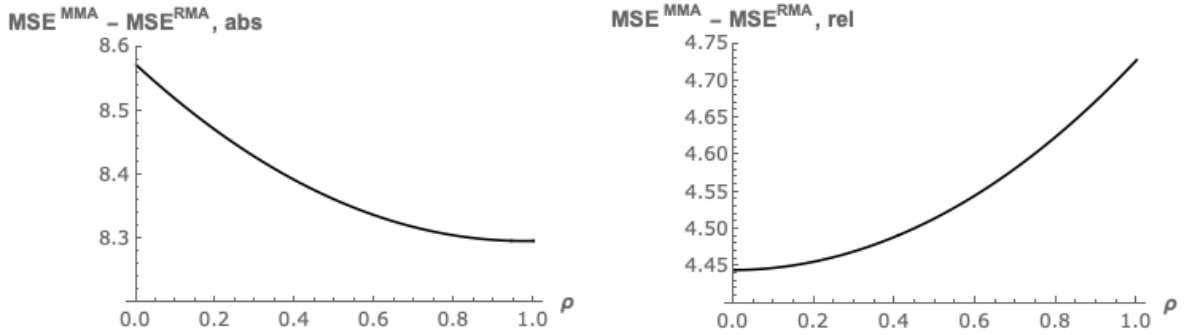

Figure 4: Difference in MSE given the optimal weights: in absolute terms (left) and normalized over the MSE of the RMA estimator (right). $p = 10$

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

1.4 Finite-Sample Comparison

I 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](moderate correlation) and 0.75 in case [H](high correlation).[11] The error term e_i is conditionally distributed as $N(0, \sigma^2(x_{2i}))$, where $\sigma^2(x_{2i}) = x_{2i}^4$. The parameters are set by the rule

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

to model various specifications of θ_k. I consider several combinations of α and β. First, for $\alpha = 0.5$, the considered values of β are $[.6, .7, .8, .9]$. Then I 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 $c = \sqrt{R^2/(1-R^2)}$. I examine three sample sizes, $n = 25, 50, 100$ with the maximum model lengths $p = 9, 11, 15$, respectively. In the experiment I also include the weighted BIC criterion (WBIC)[12] and the equal weighting (EW) scheme.[13]

I 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 $w_m = \exp\left(-\frac{1}{2}\text{BIC}_m\right) / \sum_{j=1}^{M} \exp\left(-\frac{1}{2}\text{BIC}_m\right)$, where $\text{BIC}_m = n \ln \hat{\sigma}_m^2 + \ln(n)\, m$.

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

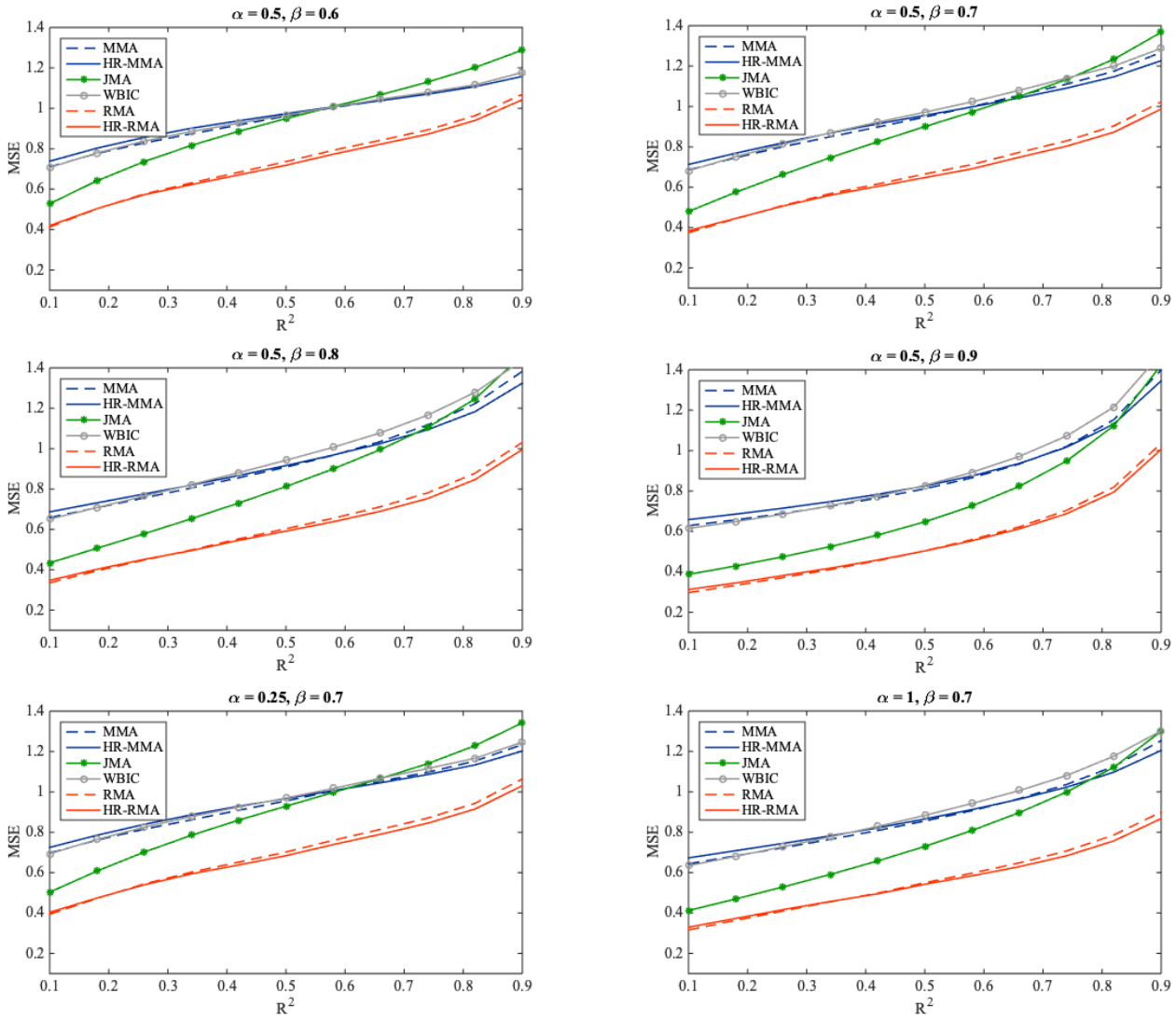

Figure 5: $n = 25$. Case [M] of moderate correlation among predictors.

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 1.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 1.C.

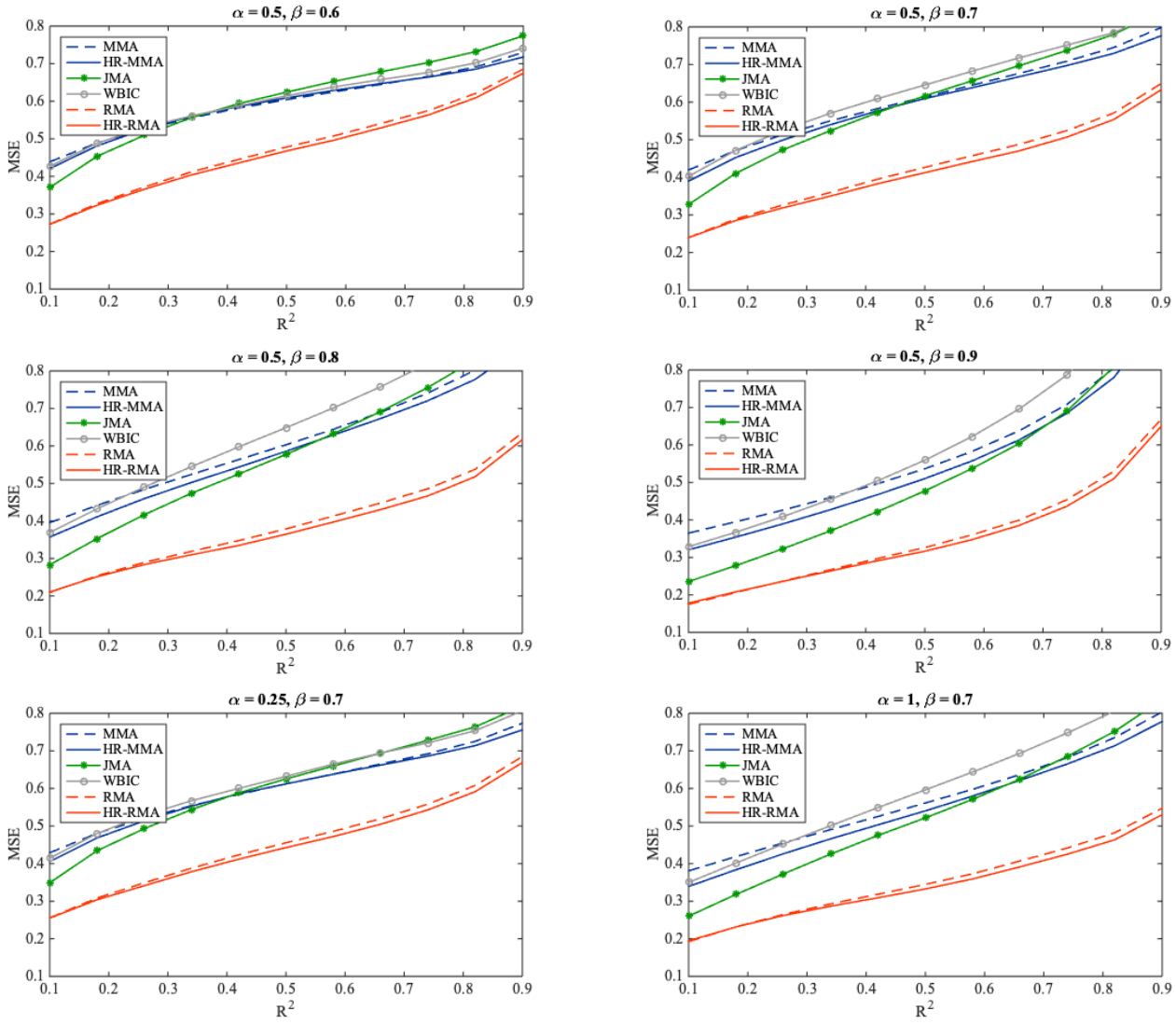

Figure 6: $n = 50$. Case [M] of moderate correlation among predictors

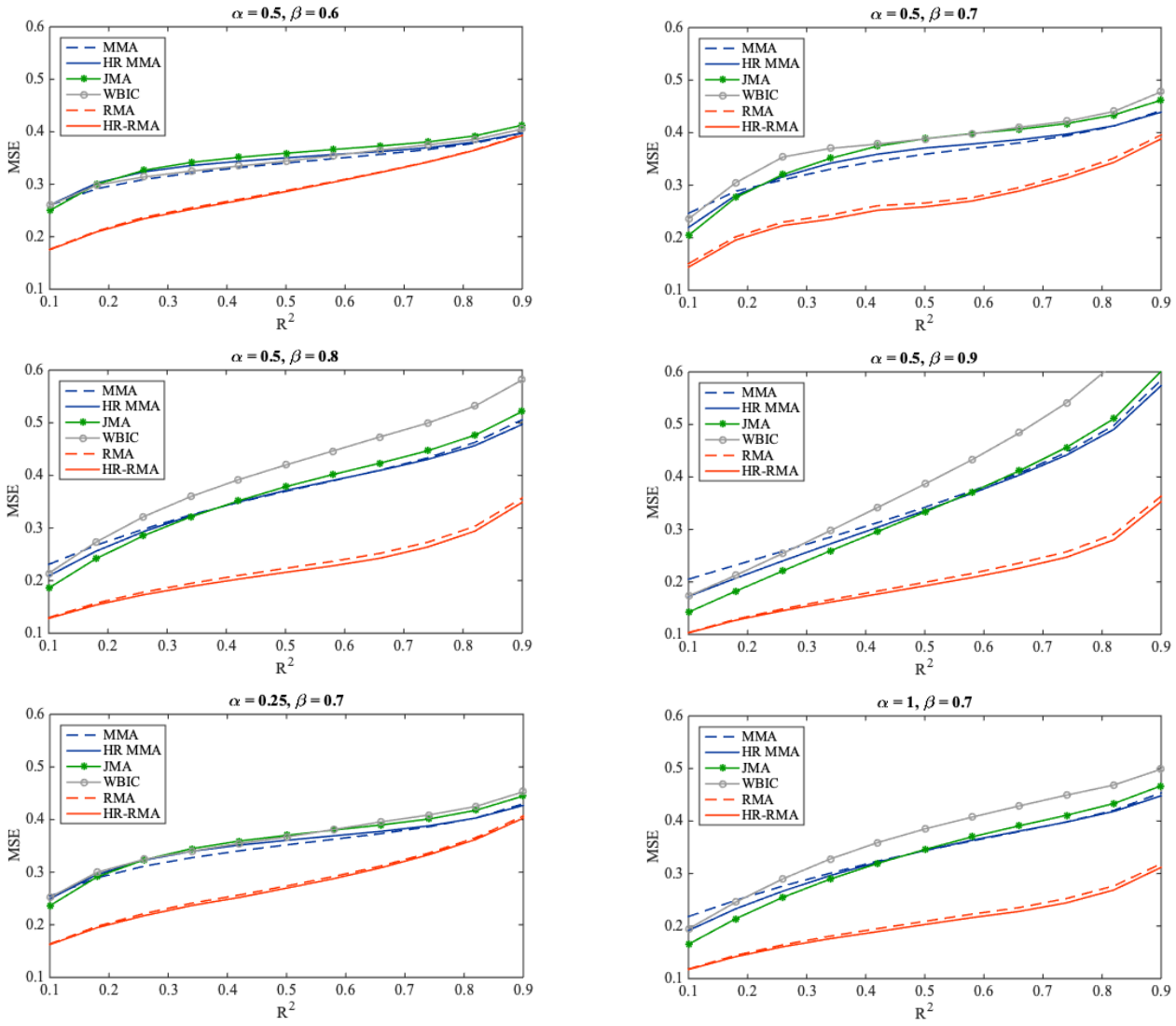

Figure 7: $n = 100$. Case [M] of moderate correlation among predictors

In Appendix 1.W I 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 I examine the relative performance of the ridge-based averaging estimators via two real-data examples.

1.5 Empirical Examples

1.5.1 Wage Prediction

Similarly to Hansen and Racine (2012), I employ Wooldridge's (Wooldridge 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 $n_2 = n - n_1$. I compare the same methods mentioned in the previous section: MMA, HR-MMA, JMA, WBIC, RMA and HR-RMA. For each model I 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 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.

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

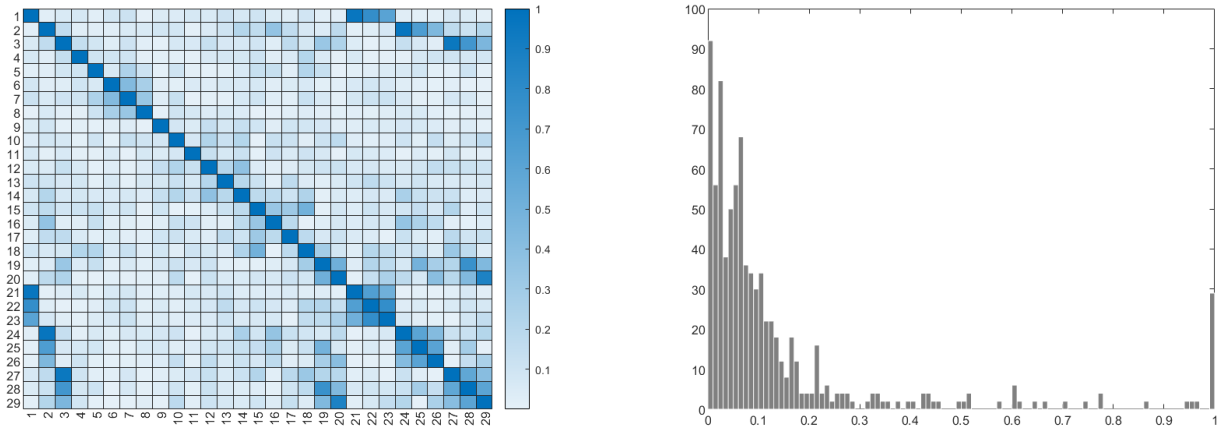

Figure 8: Correlation heatmap and correlation histogram for the wage predictors, in-sample portion of the data $n_1 = 500$. The absolute values of correlations are employed.

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.

1.5.2 Growth Determinants

Next, I 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. I 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.

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

(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,

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

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

(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, I compute the average squared prediction error across 1/5/10 countries, respectively. I 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 9 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 I 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).

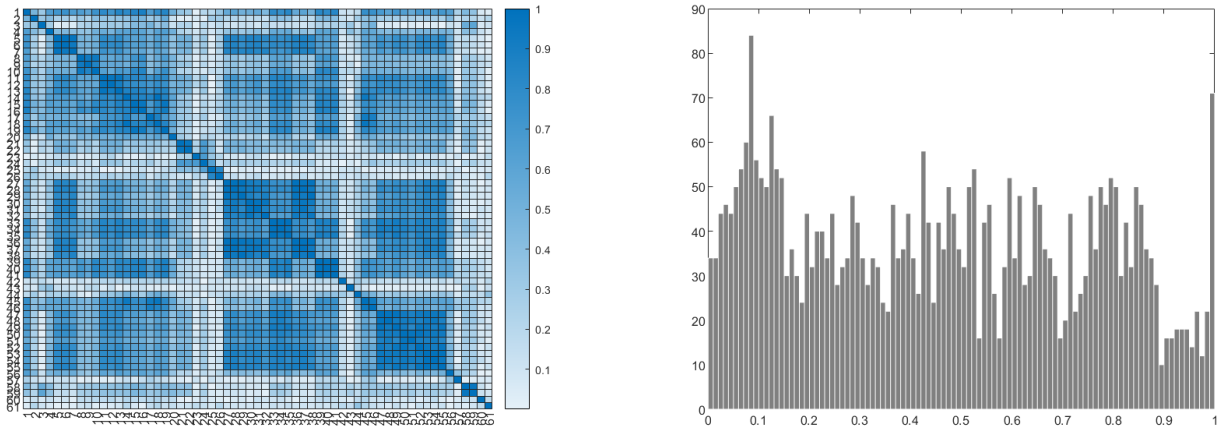

Figure 9: Correlation heatmap and correlation histogram for the growth predictors. The absolute values of correlations are employed.

1.6 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 a considerable interval of the population R^2. The improvement achieved by ridge regularization may be partially attributed to changes of the 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 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 I 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.

Appendix 1.C

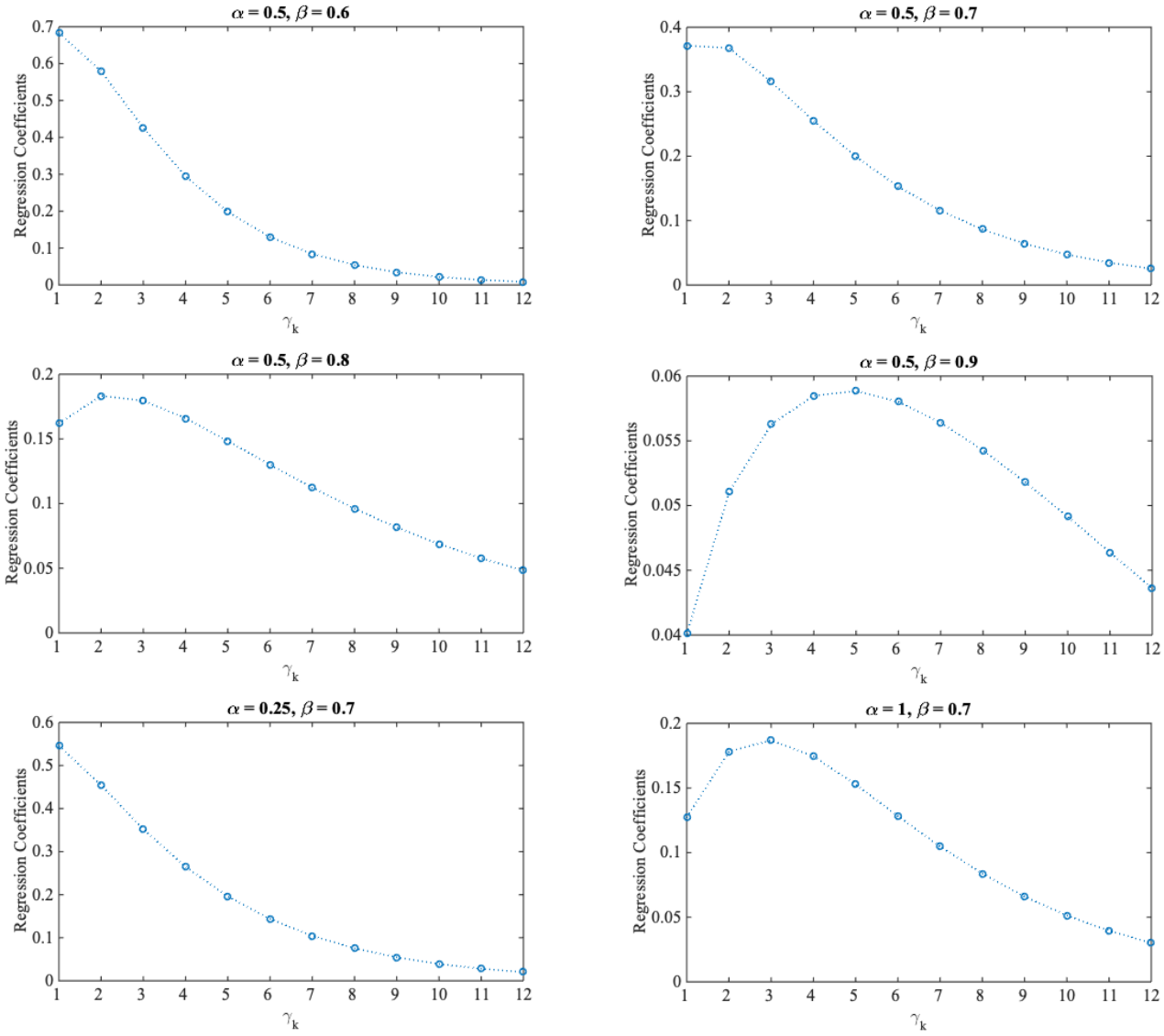

Figure 10: Simulation study: regression coefficients (all graphs are truncated along the horizonal axis)

Appendix 1.E

Part 1

The averaged OLS estimate:

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

The bias of the averaged OLS estimate:

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

The variance of the averaged OLS estimate:

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

The average of ridge estimates:

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

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

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

The variance of the averaged ridge estimate:

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

where

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

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

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

The variance component is the trace of the variance matrix:

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

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

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

Therefore,

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

Now its squared bias:

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

$$\left[E\left(\tilde{\beta}\left(\lambda_1,\lambda_2\right)-\beta\right)\right]^2 = \beta_1^T\left[\frac{\left(w^r\right)^2-2w^r\left(1+\lambda_1\right)+\left(1+\lambda_1\right)^2}{\left(1+\lambda_1\right)^2}+\frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}\rho^2\right]\beta_1 \quad (5)$$

$$+\beta_1^T\rho\left[\frac{2w^r\left(w^r-1-\lambda_1\right)}{\left(1+\lambda_1\right)^2}-\frac{2\left(w^r+\lambda_2\right)\left(1-w^r\right)}{\left(1+\lambda_2\right)^2}\right]\beta_2 \quad (6)$$

$$+\beta_2'\left[\frac{\left(w^r\right)^2}{\left(1+\lambda_1\right)^2}\rho^2+\left(\frac{\left(1-w^r\right)^2}{\left(1+\lambda_2\right)^2}-\frac{2\left(1-w^r\right)}{1+\lambda_2}+1\right)\right]\beta_2 \quad (7)$$

So, the desired MSE is (4) + (5) + (6) + (7).

Part 2

For the first model estimated via ridge:

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

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

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

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

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

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

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

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

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 1.H

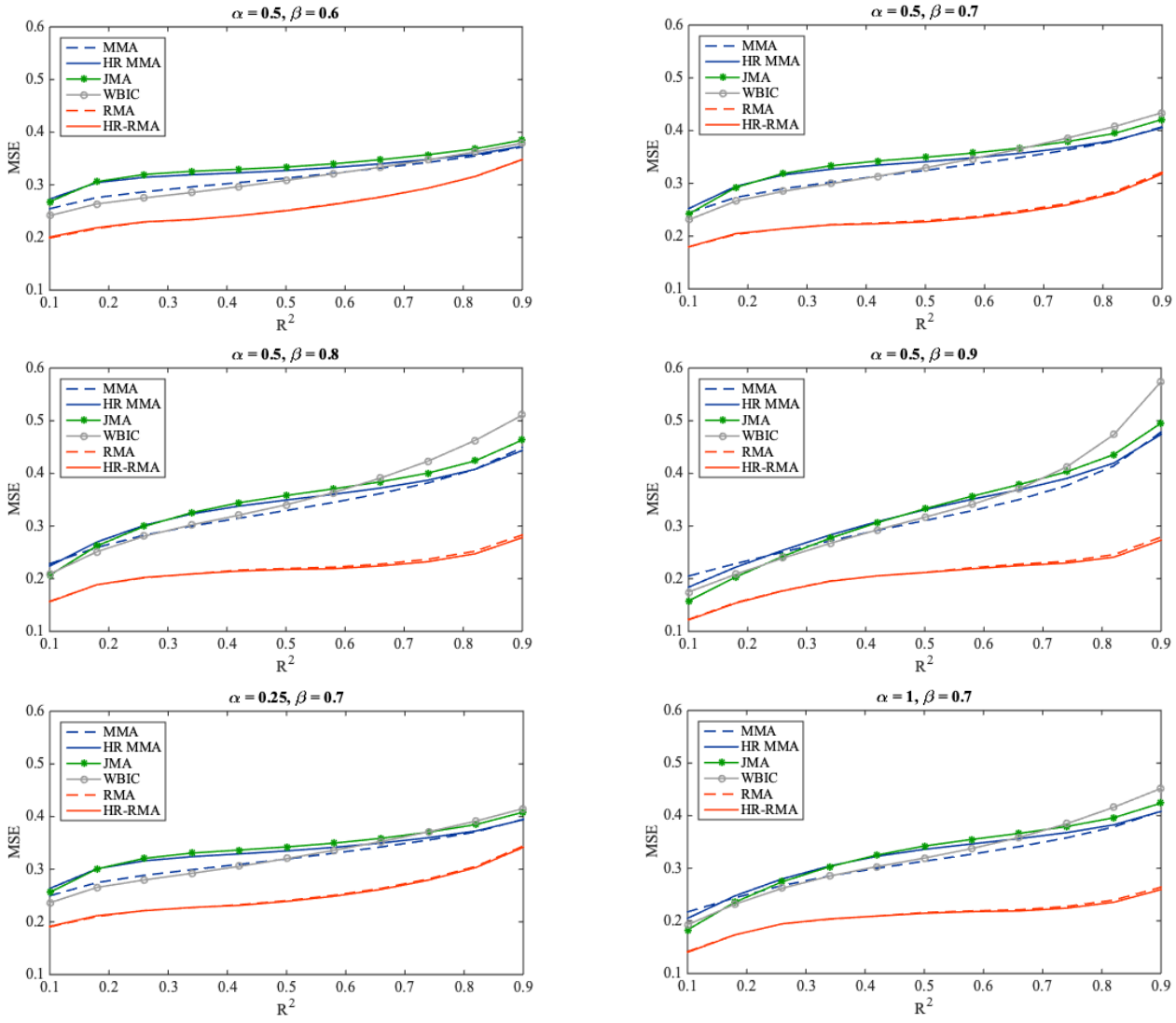

Figure 11: $n = 100$. The case [H] of high correlation among predictors

Appendix 1.W

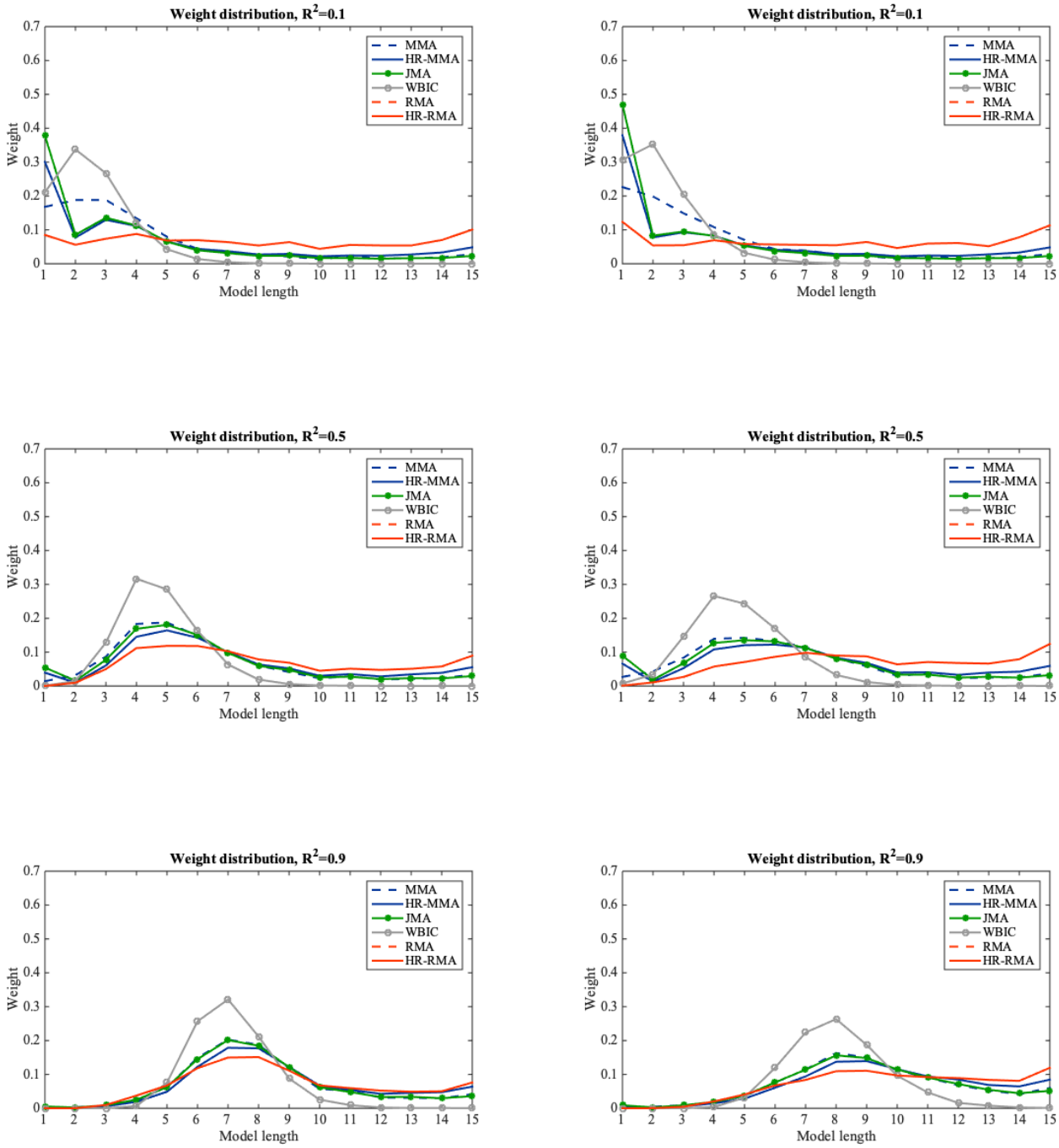

Figure 12: Optimal weights, $\alpha = 0.5$, $\beta = [0.6, 0.7]$ (left to right), $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.

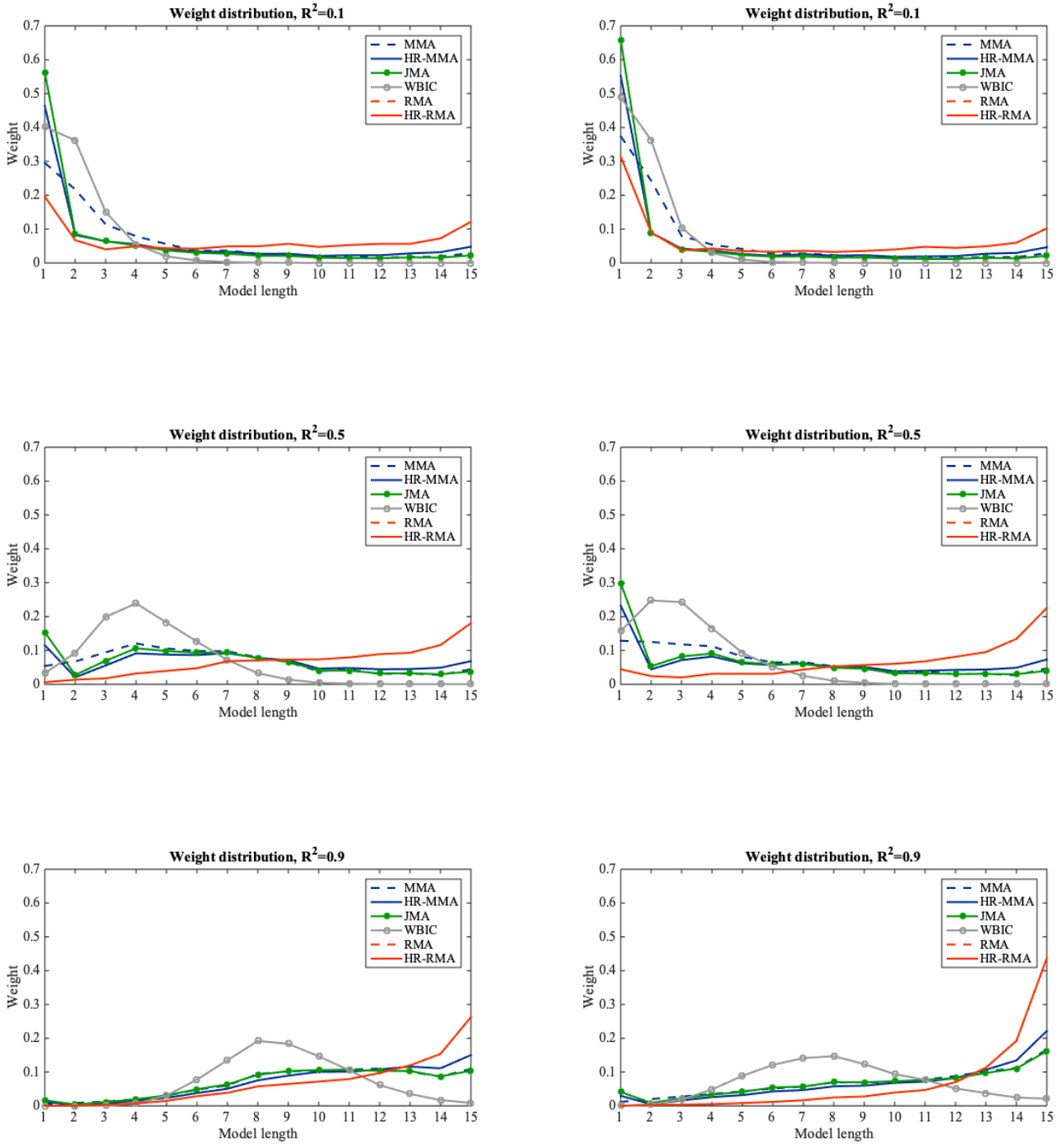

Figure 13: Optimal weights, $\alpha = 0.5$, $\beta = [0.8, 0.9]$ (left to right), $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.

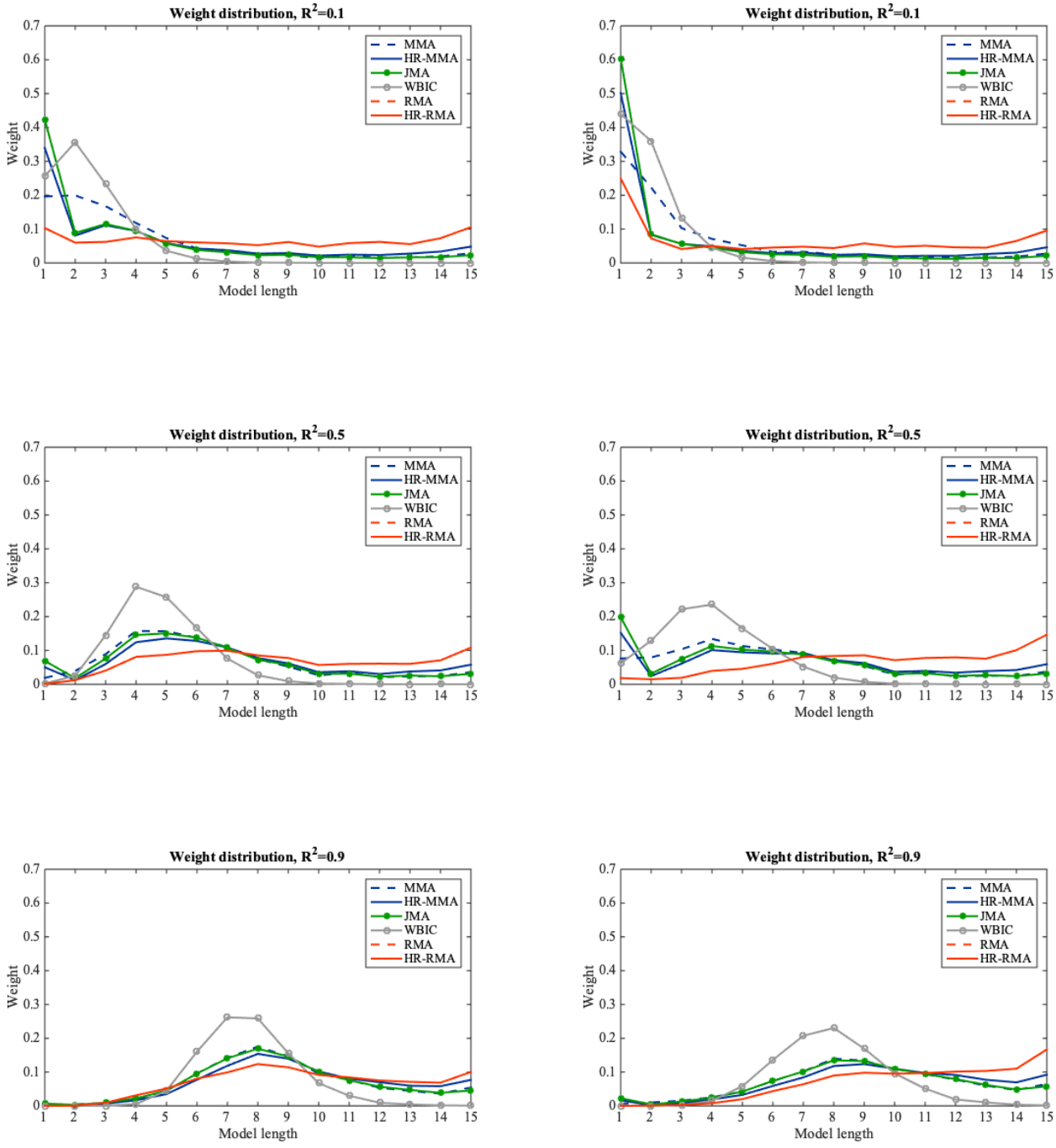

Figure 14: Optimal weights, $\alpha = [0.25, 1]$ (left to right), $\beta = 0.7$, $R^2 = [0.1, 0.5, 0.9]$ (top to bottom). The case [M] of moderate correlation among predictors.

2 Instrumental Variable Estimation with Many Instruments Using Elastic-Net IV

Published as CERGE-EI Working Paper Series No 759.

2.1 Introduction

The instrumental variables (IV) regression is a common tool for identification of treatment effects under regressor endogeneity. From the theoretical perspective, researchers would like to utilize as much exogenous variation in the explanatory variables as possible, as it increases the precision of IV estimates: Newey (1990), Amemiya (1974) and Chamberlain (1987) motivate the use of many instruments for the purpose of nonparametric estimation of optimal instruments. However, conventional GMM-type estimators, such as 2SLS, tend to be substantially biased when the number of instrumental variables is not small relative to the sample size: see Bekker (1994a) and Newey and Smith (2004).

The problem of many instruments may be circumvented in various ways. The use of statistical methods with imbedded regularization is increasingly popular among economists. Regularization techniques allow one to deal with ill-posed inverse problems, and date back to Tikhonov (1943). Such methods include the ridge regression (Hoerl and Kennard 1970), lasso (Tibshirani 1996), the penalized maximum likelihood estimation (Hastie et al. 2009), and boosting (Buhlmann 2006), among others. There are several alternative regularization procedures used as part of IV estimators: ridge and James-Stein type shrinkage applied to the first stage by Hansen and Kozbur (2014) and Spiess (2017), respectively; lasso for estimation of both the first stage and the reduced form by Belloni, Chen, Chernozhukov and Hansen (2012, hereafter BCCH); applications of random forests and deep neural networks by Wager and Athey (2018) and Farrell et al. (2021), respectively. In this list, BCCH stands out due to the extreme popularity of lasso as a regularization technique that is often employed under sparsity. In sparse models, there is a small number of variables[18] that convey most of the impact of all covariates in the response variable. Lasso represents the simplest sparse modeling approach that allows simultaneous variable selection and coefficient estimation.

The key assumption needed for lasso to produce a meaningful solution is the sparsity of the underlying model (see Section 2.1 for the definition of sparsity). The sparsity assumption may be justified in structural economic equations, where few variables participate in determinating an outcome variable. However, the lasso estimator is also promoted as a universal workhorse for pure prediction tasks. Despite the popularity of the sparse modeling framework, the adequacy of the sparsity (or approximate sparsity) assumption is often questionable. For example, Giannone et al. (2021) find evidence against sparsity for a collection of empirical applications from macroeconomics, microeconomics, and finance, where sparsity is routinely assumed without pretesting.

Furthermore, the simplicity of the lasso approach has its costs even under sparsity. For example, Zou and Hastie (2005, hereafter ZH) stress three limitations of classical lasso: (1) if

[18] $s = o(n)$,where n is the sample size.

predictors are highly correlated as a whole, the prediction performance of the ridge regression dominates that of lasso (first observed in Tibshirani 1996), as with highly correlated predictors the lasso solution paths tend to be unstable ; (2) in the $p > n$ case, when the number of variables p exceeds the number of observations n, lasso selects at most n variables; (3) if there are groups of predictors within which pairwise correlations are high, lasso generally selects only one variable from each group. ZH propose an alternative estimator – elastic-net (EN) – that successfully eliminates these shortcomings of lasso.[19] Through a simulation study and empirical examples they show that elastic-net often outperforms lasso in terms of prediction accuracy. In addition, EN essentially combines the properties of lasso and ridge , thus being able to accommodate some DGP's deviations from sparsity.

Of the three above-mentioned conditions under which the performance of lasso may be improved, at least the first two directly relate to IV estimation. Economists often estimate a causal effect based on a dataset at hand with many characteristics available for every unit (possibly $p > n$), where many serve as potential instruments (including the basic instrumental variables, their interactions and transformations). These instruments, however, tend to be moderately or highly correlated, leading to unstable lasso solution paths.[20] Thus, by using lasso to tackle the first-stage prediction problem, one faces exactly the scenario under which the performance of lasso may be improved via an additional ridge-type regularization, therefore justifying the use of the elastic-net technique.

This paper contributes to the literature on IV estimation with many instruments by considering the use of the elastic-net approach for estimating the first-stage regression. While the lasso (and post-lasso) IV estimator by BCCH and the ridge jackknife IV estimator by Hansen and Kozbur (2014) stem from the sparsity and the density of the first-stage relationship, respectively, I propose the elastic-net IV estimator (ENIV), which fits between those two. Similarly to lasso, elastic-net with a properly selected penalty parameter is shown to have oracle properties[21] under sparsity. Consequently, the results of BCCH on consistency and asymptotic normality (under possible non-Gaussianity and heteroskedasticity of the error term) of a generic sparsity-based IV estimator can be applied to the proposed elastic-net IV estimator. At the same time, in the case of no sparsity, elastic-net is by construction capable of acting like a ridge regression. Thus, for elastic-net with data-driven parameters (a penalty level, and a weighting parameter reflecting the degree of DGP sparsity), the proposed estimator should be robust to the unknown degree of sparsity of first-stage relationships.

To address the issue of overfitting (see, for example Chernozhukov et al. 2018), I consider sample-split and cross-fit versions of the basic elastic-net IV estimator (SS-ENIV and CF-ENIV, respectively), and compare them with the lasso-based analogues. I study the relative performance of the proposed estimators via simulations. Specifically, I compare the resulting IV estimates in terms of the median absolute bias, median absolute deviation and rejection rate. The SS-ENIV and CF-ENIV estimators perform well relative to the lasso-based alternatives, regardless of the signal's sparsity.

[19] Elastic-net reduces to lasso in an orthogonal design, where lasso is optimal, see Donoho et al. (1995).

[20] Under high dimensionality of the problem, even when the instrumental variables are independent, there might be large sample correlations, see Fan and Lv (2008).

[21] i.e. to achieve the rate of convergence that is very close to the oracle rate $\sqrt{s/n}$ achievable when the true model is known.

Additionally, I demonstrate the potential gains of the EN-based IV estimation based on the classic empirical investigation from Angrist and Krueger (1991), who look at the causal effect of schooling on earnings. The identification strategy and data from Angrist and Krueger (1991) provide many available instrumental variables for schooling. While employing as many of them as possible potentially leads to higher accuracy of the estimated causal effect, it also leads to biases and inferential problems. Therefore, the use of instrument selection or regularization techniques is justified, thus making the example suitable for testing the performance of the EN-based IV estimators.

The plan of this paper is as follows. In Section 2 I describe an instrumental variables setup and overview the regularization-based methods for estimation of optimal instruments. In Section 3 I present and discuss the results of a simulation study that examines the performance of the proposed estimator relative to its closest competitors. Section 4 provides an empirical example to demonstrate potential improvement in estimation accuracy gained by the use of IV estimators based on elastic-net.

2.2 The Instrumental Variables Model

The problem setup is similar to that from BCCH, simplified to the case of a scalar endogenous variable. The model is $y_i = d_i'\delta_0 + e_i$, where y_i is a scalar outcome, d_i is a k_d-vector of variables, and δ_0 denotes the true value of a vector-valued parameter δ. The first element of d_i is endogenous, while the remaining elements of d_i constitute a vector of exogenous covariates w_i. The disturbance term e_i is such that $E[e_i|z_i] = 0$, where z_i is a k_z-vector of instrumental variables.

As a motivation, suppose the disturbance term is conditionally homoskedastic, $E[e_i^2|z_i] = \sigma^2$. For a k_d-vector of instruments $A(z_i)$, the standard IV estimator of δ_0 takes the form

$$\hat{\delta} = \left(\mathbb{E}_n\left[A(z_i)\, d_i'\right]\right)^{-1} \mathbb{E}_n\left[A(z_i)\, y_i\right],$$

where $\{(z_i, d_i, y_i), i = 1, ..., n\}$ is i.i.d. sample, $\mathbb{E}_n[f] := \mathbb{E}_n[f(z_i)] := \sum_{i=1}^n f_i/n$. Given instruments $A(z_i)$,

$$\sqrt{n}\left(\hat{\delta} - \delta_0\right) \to^d \mathcal{N}\left(0, Q_0^{-1}\Omega_0 Q_0^{-1}\right),$$

where $Q_0 = E[A(z_i)\, d_i']$ and $\Omega_0 = \sigma^2 E\left[A(z_i)\, A(z_i)'\right]$. Employing the optimal instrument $A(z_i) = D(z_i) = E[d_i|z_i]$ achieves the semiparametric efficiency bound for estimating δ_0, with the asymptotic variance $\Lambda^* = \sigma^2\left\{E\left[D(z_i)\, D(z_i)'\right]\right\}^{-1}$ (see Chamberlain 1987).

2.2.1 Regularized Estimation Methods for Optimal Instruments

In practice, the optimal instrument $D(z_i)$ is not known, and many ways to estimate it exist in the literature. Suppose there is a large set of instruments,

$$f_i := (f_{i1}, ..., f_{ip})' := (f_1(z_1), ..., f_p(z_1))'$$

available for estimation of conditional expectation $D(z_i)$, and the number of instruments p is possibly larger than the sample size n. In BCCH, the optimal instrument $D(z_i)$ is assumed

to be approximately sparse, i.e. a function $D(z_i)$ is deemed to be well-approximated by a function of unknown $1 \leq s \ll n$ instruments:

$$\begin{aligned} D(z_i) &= f_i^{'}\beta_0 + a(z_i), \\ \|\beta_0\|_0 \leq s &= o(n), \left[\mathbb{E}_n a(z_i)^2\right]^{1/2} \leq c_s \lesssim_P \sqrt{s/n}. \end{aligned}$$

The identities of s relevant instruments, i.e. $T = \text{support}(\beta_0) = \{j \in \{1, ..., p\} : |\beta_{0j}| > 0\}$, are meant to be a priori unknown. The sparsity assumption requires that at most s instruments approximate the conditional expectation $D(z_i)$ so that the approximation error $a(z_i)$ does not exceed the conjectured size $\sqrt{s/n}$ of the error of the infeasible estimator that "knows" the identity of these s relevant instruments (the "oracle estimator").

Lasso

The first stage regression equation is

$$d_i = D(z_i) + v_i, \quad E[v_i|z_i] = 0.$$

For the sample $\{(z_i, d_i), i = 1, ..., n\}$, consider estimators of the optimal instrument $D(z_i)$ of the form

$$\widehat{D}_i := \widehat{D}(z_i) = f_i'\widehat{\beta},$$

where $\widehat{\beta}$ is the sparse estimator based on regressors f_i and d_i as the dependent variable. The sparse estimator sets all but a small fraction of the coefficient estimates $\hat{\beta}_j$ to 0. Let $Q(\beta)$ denote the least squares criterion function, $\widehat{Q}(\beta) := \mathbb{E}_n\left[(d_i - f_i'\beta)^2\right]$, then the lasso estimator employed in BCCH is defined as a solution to

$$\widehat{\beta}_L \in \arg\min_{\beta \in R^p} \widehat{Q}(\beta) + \lambda^L \left\|\widehat{\Upsilon}\beta\right\|_1,$$

where λ^L is the penalty level and $\widehat{\Upsilon} = \text{diag}(\widehat{\gamma}_1, ..., \widehat{\gamma}_p)$ is a diagonal matrix with data-dependent weights, also called penalty loadings. The basic lasso estimator, with all penalty loadings set to 1, was introduced by Tibshirani (1996) as a technique for simultaneous estimation and variable selection. Basically, lasso shrinks the coefficients toward 0 as λ^L increases, and some coefficient estimates are set to 0 for large enough λ^L.

Lasso has been shown to be variable selection consistent, i.e. to be able to discover the correct model specification, under suitable conditions (see Meinshausen and Buhlmann 2004). Initially, the weighted/adaptive version of lasso (with data-dependent penalty loadings) was proposed in Zou (2006) in response to debates about whether the lasso estimator is an oracle procedure (Fan and Li 2001; Meinshausen and Buhlmann 2004). For the data-dependent and cleverly chosen loadings[22], the adaptive lasso estimator is shown to enjoy oracle properties. Relatively recently, BCCH have proposed novel penalty loadings that

[22]Zou (2006) suggests the weight vector $\hat{w} = 1/\left|\hat{\beta}\right|^{\gamma}$, where $\hat{\beta}$ is a root-n consistent estimator for β, and $\gamma > 0$.

result in sharp convergence rates for the lasso estimator under possible non-Gaussianity and heteroskedasticity.

Having estimated the optimal instrument via lasso, let $\widehat{D}_i$ be a vector of instruments that also includes the vector of exogenous covariates w_i

$$\widehat{D}_i = \left(\widehat{D}\left(z_i\right), w_i'\right)'.$$

Then the resulting lasso-IV estimator

$$\widehat{\delta}^L = \mathbb{E}_n\left[\widehat{D}_i d_i'\right]^{-1} \mathbb{E}_n\left[\widehat{D}_i y_i\right] \tag{8}$$

is shown to achieve the efficiency bound asymptotically, $\sqrt{n}\left(\widehat{\delta^L} - \delta_0\right) =_d N\left(0, \Lambda^*\right) + o_P\left(1\right)$. The IV estimator with the lasso-based optimal instrument is root-n consistent and asymptotically normal (see Theorem 3 of BCCH). Moreover, consistency and asymptotic normality continues to hold for any generic sparsity-based method achieving specific near-oracle performance bounds (see Theorem 4 of BCCH), and I exploit this result in the next section.

Elastic-Net IV Estimator

Although lasso is aimed at high-dimensional problems, its performance may be deteriorated by the correlation among predictors, which often takes place in high-dimensional settings. Zou and Hastie (2005) point out that the lasso solution paths are unstable (i.e. not smooth) when predictors are highly correlated. The relevance of this issue is stressed by Fan and Lv (2008) who show that even with the independent predictors the maximum sample correlation can be large, as long as the dimensionality is high. In addition, ZH notice that for high-dimensional problems with $p \gg n$, lasso is incapable of selecting more than p variables into the model. Consequently, they propose an alternative penalized estimator, elastic-net (EN),

$$\hat{\beta}^{EN} = \arg\min_{\beta}\left\{\sum_{i=1}^{N}\left(d_i - \sum_{j=1}^{p} f_{ij}\beta_j\right)^2 + \lambda^{EN}\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right| + \left(1-\alpha\right)\beta_j^2\right)\right\},$$

which involves an l_2-penalty in addition to lasso's l_1-penalty. The first term of the penalty, $\lambda^{EN}\sum_{j=1}^{p}\alpha\left|\beta_j\right|$ encourages a sparse solution, as does the lasso penalty, while the second term, $\lambda^{EN}\sum_{j=1}^{p}\left(1-\alpha\right)\beta_j^2$, regularizes the covariance matrix, and encourages equality of the coefficients on highly correlated predictors. ZH shows that elastic-net may be interpreted as a stabilized[23] version of lasso (p. 308, Theorem 2), and can therefore improve upon lasso.

In the statistical literature, the performance of the elastic-net estimator is usually analyzed under a restrictive assumption of the Gaussian and homoskedastic error term. For example, when Gaussian and homoskedastic noise is assumed, Jia and Yu (2010) study the model selection properties of the elastic-net estimator in the asymptotic framework where the

[23]Stabilization is achieved via replacement of the sample covariance matrix $\widehat{\Sigma}$ with its shrunken (towards the identity matrix) version.

number of variables p grows with the sample size n. They provide sufficient conditions for elastic-net to be model selection consistent[24], as well as theoretical and simulation examples demonstrating when elastic-net can consistently select the true model, while lasso fails to do so.[25] Further, Ghosh (2011) considers adaptive elastic-net that generalizes elastic-net in the same way that adaptive lasso generalizes lasso, thus expanding the set of conditions under which elastic-net performs consistent variable selection. The adaptive elastic-net estimator uses a more flexible l_1-penalty for consistent variable selection, while the ridge-type penalty term stays unchanged[26] and continues to regularize the solution path:

$$\hat{\beta}_{ada}^{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(d_i - \sum_{j=1}^{p} f_{ij}\beta_j \right)^2 + \lambda_1^{EN} \sum_{j=1}^{p} w_j \left|\beta_j\right| + \lambda_2^{EN} \sum_{j=1}^{p} \beta_j^2 \right\},$$

where the weight estimate $\widehat{w}_j = 1/|\widehat{\beta}_j|^{\gamma}$, $j = 1, ..., p$, for some $\gamma > 0$, with the ordinary least squares estimator $\widehat{\beta}^{OLS}$ being a possible choice of $\widehat{\beta}$. Under suitable conditions, the adaptive elastic-net estimator is shown to have oracle properties (variable selection consistency and asymptotic normality, see Theorem 3.2).

However, the breakthrough results of Theorem 4 in BCCH on root-n consistency and asymptotic normality apply to a wide class of sparsity-based methods that encompasses the elastic-net estimator. Consequently, to get the desired asymptotic properties of the elastic-net estimator under possible non-Gaussianity and heteroskedasticity of the error term, it is enough to establish the near-oracle bounds that are required by BCCH's Theorem 4. I use the result from Zou and Hastie (2006) about transformation of the elastic-net problem into an equivalent lasso problem on augmented data to show that the elastic-net estimator performs closely enough to the oracle under sparsity, in the sense of meeting sufficient conditions of BCCH's Theorem 4.

Proposition 1. For $\left(\lambda_1^{EN}, \lambda_2^{EN}\right)$ such that $\gamma = \lambda_1^{EN}/\sqrt{1+\lambda_2^{EN}} = \lambda_{opt}^{L}$, where λ_{opt}^{L} denotes the optimal penalty for the lasso-estimator, the elastic-net estimator obeys the near-oracle performance bounds:

$$\left\| \widehat{D}_i^{EN} - D_i \right\|_{2,n} \leq_p \sqrt{\frac{s \log\left(n+p\right)}{n+p}}$$
$$\left\| \widehat{\delta}^{EN} - \delta_0 \right\|_1 \leq_p \sqrt{\frac{s^2 \log\left(n+p\right)}{n+p}}$$

Therefore, the elastic-net estimator can perform a variable selection and estimation similarly to the lasso estimator. Once the sufficient conditions of Theorem 4 in BCCH continue to hold, one can rely on the existing results regarding consistency and asymptotic normality of generic sparsity-based IV estimators obtained in BCCH. In other words, the IV estimators based on elastic-net and lasso can be asymptotically equivalent under sparsity and the appropriate choice of the penalty parameters $\left(\lambda_1^{EN}, \lambda_2^{EN}\right)$. At the same time, ridge regularization often

[24] Jia and Yu (2010) also state a specific condition for the inconsistency of the elastic-net estimator.

[25] See also Yuan and Lin (2007) for a similar study for fixed p.

[26] In principle, adaptive weights can also be placed on an l_2 penalty, but it is not necessary to guarantee the oracle properties of the adaptive elastic-net estimator examined in Ghosh (2011).

leads to finite-sample improvement, so the relative finite-sample performance of the IV estimators based on elastic-net (with a ridge-type penalty) and lasso (without a ridge-type penalty) is of interest, and is investigated in Section 3 of this paper.

Sample-Split and Cross-Fit Elastic-Net IV Estimator

In principle, one could employ $\widehat{D}_i = f_i'\widehat{\beta}^{EN}$ for $\widehat{D}_i$ in (2.2.1) to define an IV estimator with a EN-regularized first stage. However, as noted in Hansen and Kozbur (2014), among others, this direct approach would typically introduce a so-called regularization bias (similar to other methods involving regularization).[27] In general, the least shrunk coefficients correspond to the instruments that are most highly correlated with the first stage noise, thus contaminating the exclusion restriction. The use of sample-splitting or jackknifing is a common way of lowering the regularization bias. I employ the sample-splitting technique to preserve the exclusion restriction, thus defining

$$\hat{\beta}_{I_1}^{EN} = \arg\min_{\beta} \left\{ \sum_{i \in I_1} \left(d_i - \beta_0 - \sum_{j=1}^{p} f_{ij}\beta_j \right)^2 + \lambda^{EN} \sum_{j=1}^{p} \left(\alpha \left|\beta_j\right| + (1-\alpha)\,\beta_j^2 \right) \right\},$$

which is the elastic-net estimate from an elastic-net regression of d on f with regularization parameters $\left(\lambda^{EN}, \alpha\right)$ using the random subset of observations I_1 (a half of the sample, in the simplest case). The estimator $\widehat{D}_i$ for the i^{th} unit is then defined as $\widehat{D}_i = f_i'\hat{\beta}_{I_1}^{EN}$. Finally, I define the sample-split ENIV estimator as

$$\widehat{\delta}^{SS-ENIV} = \left(\sum_{i \in I_1^c} f_i'\hat{\beta}_{I_1}^{EN} d_i' \right)^{-1} \sum_{i \in I_1^c} f_i'\hat{\beta}_{I_1}^{EN} y_i,$$

where $I_1^c \cap I_1 = \emptyset$.

By splitting the sample into halves, I break the correlation between $\widehat{D}_i$ and e_i that is not asymptotically negligible. Although the elastic-net regularization causes some loss of signal due to coefficient shrinkage (similar to other regularization methods), a data-driven choice of $\left(\lambda^{EN}, \alpha\right)$ is expected to result in quality signal extraction from a high-dimensional set of instruments, whether sparse or dense. For example, for $\alpha = 0$ and positive λ^{EN}, the elastic-net IV estimator reduces to the ridge IV estimator. I suggest choosing the shrinkage parameter based on the optimization of a first stage cross-validation criterion due to popularity and availability of cross-validation tools in R, Python, Stata, etc.[28] In general, for not very large datasets one can replace a sample-splitting approach with a jackknifing procedure to fit the first stage, thus generalizing the sample-split ENIV estimator to the jackknife ENIV estimator.

[27] See Chernozhukov et al. (2018) for an extended discussion of the regularization bias and de-biased estimation.

[28] The use of cross-validation is yet to be theoretically justified for elastic-net, despite being a widely spread practice. See Chetverikov et al. (2021), which justifies the practice of using cross-validation to choose the penalty parameter for lasso.

Another possible approach is cross-fitting. Cross-fitting estimators are also based on the idea of sample-splitting. First, the sample is partitioned into I_1 and I_2, and only observations from I_1 are used to get $\hat{\beta}_{I_1}^{EN}$, whereas only observations from I_2 are used to produce $\widehat{\delta}_{12} = \left(\sum_{I_2} f_i' \hat{\beta}_{I_1}^{EN} d_i'\right)^{-1} \times \sum_{i \in I_2} f_i' \hat{\beta}_{I_1}^{EN} y_i$. Then the subsamples are swapped so that $\hat{\beta}_{I_2}^{EN}$ and $\widehat{\delta}_{21} = \left(\sum_{I_1} f_i' \hat{\beta}_{I_2}^{EN} d_i'\right)^{-1} \times \sum_{i \in I_1}^{N} f_i' \hat{\beta}_{I_2}^{EN} y_i$ are obtained in an analogous way. Consequently, the cross-fit elastic-net IV estimator is defined as $\widehat{\delta}^{CF-ENIV} = \left(\widehat{\delta}_{12} + \widehat{\delta}_{21}\right)/2$. This way both subsamples (symmetrically) contribute to the resulting estimate, thus increasing its efficiency. I adopt the algorithm by Anatolyev and Mikusheva (2022, Section 3.2) to estimate the variance of $\widehat{\delta}^{CF-ENIV}$ in a way that accounts for the correlation between $\widehat{\delta}_{12}$ and $\widehat{\delta}_{21}$.[29] Finally, sample-split and cross-fit lasso-based IV estimators, which act as benchmarks in the following section, are defined analogously.

2.3 Simulation study

The design of this simulation study closely follows that of Hansen and Kozbur (2014). I demonstrate the performance of the IV estimators employing elastic-net, and compare it with the performance of lasso-based IV estimators, and the ridge jackknife IV estimator (RJIVE) by Hansen and Kozbur (2014). Let the data generating process be

$$\begin{aligned} y_i &= x_i \delta_0 + e_i \\ x_i &= Z_i' \Pi + u_i \end{aligned}$$

with

$$(e_i, u_i) \sim N\left(0, \begin{pmatrix} \sigma_e^2 & \sigma_{eu} \\ \sigma_{eu} & \sigma_u^2 \end{pmatrix}\right),$$

where x_i is the scalar treatment variable, and $\delta_0 = 1$ is the parameter of interest. The sample size $n = 100$, $\sigma_e^2 = 2$, and $\text{corr}(e_i, u_i) = 0.6$. The remaining parameters are varied within the simulation study.

I consider two instrument designs: binary and continuous (Gaussian). Real datasets typically employ very different combinations of both binary and continuous instruments, thus motivating examination of the two extreme cases: (i) all instruments are binary, and (ii) all instruments are continuous. The continuous instrument design considers correlated Gaussian instruments drawn with mean 0 and variance $\text{var}(Z_{ij}) = 0.3$. The correlation between Gaussian instruments is given by $\text{corr}(Z_{ij}, Z_{ik}) = 0.8^{|j-k|}$. The binary design is motivated by the presence of many categorical variables, which often takes place in practice. In this design, all instruments are drawn from $Z_{ij} \in \{0, 1\}$ with $\Pr(Z_{ij} = 1) = 0.8$ such that the pairwise correlations are close to $\text{corr}(Z_{ij}, Z_{ik}) = 0.8^{|j-k|}$.[30] For each design, the number of instruments is set to $K = 95$ or $K = 190$.

[29] Anatolyev and Mikusheva (2022) propose the algorithms for constructing a four-split estimator. I use a version simplified to a case with only two splits.

[30] First, I make draws from the standard normal distribution, and apply Cholecky's decomposition to generate the Gaussian instruments Z_{ij}^0 with correlations $\text{corr}\left(Z_{ij}^0, Z_{ik}^0\right) = 0.8^{|j-k|}$. Then I set $Z_{ij} = \mathbb{I}_{\left\{Z_{ij}^0 > 0.8\right\}}$.

In addition to alternation of the instrument design, I also vary the first-stage coefficients Π to generate dense, sparse, and mixed first-stage signal structures. In the dense scenario, $\Pi = (\iota_{0.4K}, 0_{0.6K})'$, where ι_p is a $1 \times p$ vector of ones, and 0_q is a $1 \times q$ vector of zeros. In the sparse scenario $\Pi = (3\iota_5, 0_{K-5})'$, so only five instruments are relevant. Finally, in the mixed scenario, $\Pi = (3\iota_5, \iota_{0.4K}, 0_{0.6K-5})'$. By varying the noise σ_u^2 in the first-stage regression, I control the strength of the instrument set measured by the concentration parameter $\mu^2 = N\Pi' E[Z_i' Z_i]\Pi/\sigma_u^2$. To model the cases of the weak and strong signal provided by the instruments, I set $\mu^2 = 30$ and $\mu^2 = 150$, respectively.

I consider three IV estimators based on elastic-net: elastic-net IV estimator (ENIV), sample-split elastic-net IV estimator (SS-ENIV), and cross-fit elastic-net IV estimator (CF-ENIV). Their lasso-based counterparts are Lasso-IV, SS-Lasso-IV, and CF-Lasso-IV. I also report the results for RJIVE and the 2SLS estimator. In addition, I present the results for the post-Lasso-IV estimator described in BCCH[31], as well as its sample-split version (SS-post-Lasso-IV). The penalty levels for ENIV, SS-ENIV, and CF-ENIV is chosen through cross-validation.

The reported results are obtained by averaging across 1500 draws for each setting. For each estimator, I present the median bias (Med. Bias), the median absolute deviation (MAD), and the rejection rate for a 5%-level test of $H_0 : \delta_0 = 1$ (RP 5%). For the post-Lasso estimator with lasso sometimes selecting no instruments into the first stage regression, I calculate the median bias and the median absolute deviation conditional on the lasso estimator selecting at least one variable. In such a case, a failure to reject the null is recorded.

Table 2.1 shows the results for $K = 95$. Panels A and B focus on the results for weak instruments ($\mu^2 = 30$), Panels C and D report the results for a stronger signal ($\mu^2 = 150$). For the weak sparse signal, Lasso-IV, post-Lasso-IV, RJIVE, SS-ENIV, and CF-ENIV result in reasonable rejection frequencies, with RJIVE and SS-ENIV being among the most accurate. However, for the dense weak signal, only RJIVE, SS-ENIV and CF-ENIV continue to have approximately the correct size (CF-ENIV tends to over-reject but not as much as the Lasso-based estimators).

For the mixed design, only RJIVE and SS-ENIV deliver accurate test size. Overall, SS-ENIV tends to produce more precise rejection rates when the true non-zero coefficients on the instruments vary in magnitude (the case of a mixed signal), compared to the case of the equal coefficient magnitude[32], which is often examined as part of simulation exercises in the literature (e.g. in Hansen and Kozbur, 2014, among others). In practice, there is often no good reason to expect a signal to be evenly distributed across all instruments that explain a decent share of variance in x_i, the treatment variable. Whereas RJIVE tends to result in rejection frequencies slightly above the nominal test size, the opposite is true for SS-ENIV.

With a strong sparse signal, most Lasso-based estimators produce adequate rejection frequencies, as expected. RJIVE, SS-ENIV and CF-ENIV retain rather accurate test size irrespective of the data structure when the signal is strong. Notably, CF-ENIV performs better with strong signals (sparse, dense, or mixed) than with weak signals. The SS-ENIV estimator proves to be a good alternative to RJIVE when dealing with a strong mixed signal, similarly

[31] BCCH recommend the penalty level to be proportional to $\sqrt{n \log K}$. I employ the same penalty as in Hansen and Kozbur (2014), namely $2.2\sqrt{2n \log(2K)}\sigma_u\sigma_e$.

[32] All first-stage variables are standardized before ridge/lasso/elastic-net estimation is performed.

to the case of a weak mixed signal discussed above.

Table 2.1. Simulation Results many instruments K = 95

	Sparse Signal			Dense Signal			Mixed Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration parameter = 30. Binary Instruments								
Lasso-IV	0.009	0.015	0.091	0.017	0.018	0.237	0.012	0.013	0.201
SS-Lasso-IV	0.003	0.023	0.009	0.004	0.04	0.011	0.000	0.024	0.007
post-Lasso-IV	0.010	0.015	0.111	0.016	0.017	0.253	0.012	0.013	0.249
SS-post-Lasso-IV	0.003	0.023	0.008	0.004	0.038	0.013	0.000	0.024	0.009
CF-Lasso-IV	0.014	0.015	0.000	0.002	0.025	0.000	0.007	0.015	0
RJIVE	-0.001	0.020	0.047	-0.001	0.011	0.055	-0.001	0.010	0.052
ENIV	0.022	0.022	0.405	0.020	0.020	0.448	0.015	0.015	0.466
SS-ENIV	0.000	0.028	0.038	0.001	0.020	0.056	0.000	0.015	0.048
CF-ENIV	0.001	0.022	0.104	0.000	0.015	0.098	-0.001	0.012	0.095
	B. Concentration parameter = 30. Gaussian Instruments								
Lasso-IV	0.005	0.011	0.076	0.011	0.012	0.210	0.007	0.008	0.177
SS-Lasso-IV	0.002	0.016	0.031	0.002	0.030	0.005	0.002	0.015	0.012
post-Lasso-IV	0.007	0.011	0.104	0.011	0.012	0.224	0.008	0.009	0.215
SS-post-Lasso-IV	0.002	0.016	0.029	0.001	0.030	0.005	0.003	0.015	0.011
CF-Lasso-IV	0.004	0.010	0.001	0.004	0.022	0.000	0.006	0.009	0.000
RJIVE	-0.001	0.014	0.051	-0.002	0.010	0.041	0.000	0.008	0.053
ENIV	0.012	0.014	0.284	0.013	0.013	0.421	0.010	0.010	0.459
SS-ENIV	0.001	0.019	0.041	0.001	0.018	0.037	0.002	0.013	0.043
CF-ENIV	0.001	0.014	0.101	0.001	0.014	0.123	0.001	0.010	0.119
	C. Concentration parameter = 150. Binary Instruments								
Lasso-IV	0.005	0.014	0.065	0.012	0.014	0.133	0.008	0.010	0.130
SS-Lasso-IV	0.000	0.022	0.047	0.000	0.022	0.048	-0.001	0.016	0.043
post-Lasso-IV	0.005	0.014	0.068	0.013	0.014	0.155	0.009	0.010	0.144
SS-post-Lasso-IV	-0.001	0.022	0.047	0.001	0.020	0.047	0.000	0.016	0.047
CF-Lasso-IV	-0.001	0.016	0.000	-0.001	0.016	0.000	-0.001	0.013	0.000
RJIVE	-0.002	0.017	0.052	0.000	0.011	0.063	-0.001	0.009	0.063
ENIV	0.012	0.016	0.149	0.016	0.016	0.218	0.012	0.012	0.233
SS-ENIV	0.000	0.022	0.052	0.001	0.016	0.060	-0.001	0.013	0.057
CF-ENIV	-0.001	0.015	0.054	0.000	0.012	0.053	-0.001	0.010	0.071
	D. Concentration parameter = 150. Gaussian Instruments								
Lasso-IV	0.002	0.010	0.064	0.010	0.011	0.175	0.005	0.007	0.113
SS-Lasso-IV	0.000	0.015	0.058	0.000	0.02	0.045	-0.001	0.012	0.048
post-Lasso-IV	0.004	0.010	0.076	0.010	0.011	0.186	0.006	0.007	0.145
SS-post-Lasso-IV	0.000	0.015	0.057	0.001	0.018	0.045	-0.001	0.012	0.048
CF-Lasso-IV	-0.001	0.011	0.006	0.000	0.014	0.000	0.000	0.009	0.002
RJIVE	0.000	0.011	0.057	0.000	0.009	0.065	-0.001	0.006	0.055
ENIV	0.008	0.011	0.132	0.012	0.012	0.251	0.008	0.009	0.225
SS-ENIV	0.000	0.015	0.054	0.001	0.013	0.060	0.000	0.010	0.047
CF-ENIV	-0.001	0.011	0.056	0.000	0.010	0.079	0.000	0.007	0.070

Note: Results are based on 1500 simulation replications. I report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators: the Lasso IV and post-Lasso IV estimators of Belloni et al. (2012, Lasso-IV and post-Lasso-IV), their sample-split versions (SS-Lasso-IV and SS-post-Lasso-IV), the cross-fit Lasso IV estimator, the RJIVE by Hansen and Kozbur (2014, RJIVE), and three estimators proposed in this paper: the elastic-net IV estimator (ENIV), the sample-split elastic-net IV estimator (SS-ENIV) and the cross-fit elastic-net IV estimator (CF-ENIV).

Table 2.2 shows the results for $K = 190$. Panels A and B again focus on the results for weak instruments ($\mu^2 = 30$), Panels C and D report the results for a stronger signal ($\mu^2 = 150$).

For the weak sparse signal, some Lasso-based estimators have reasonable rejection frequencies, although RJIVE and SS-ENIV tend to be superior in terms of bias and rejection rate, irrespective of sparsity. With the weak signal and mixed data structure, RJIVE and SS-ENIV perform similarly, although the sample-split elastic-net IV estimator seems to be more prone to under-rejection. With the strong sparse signal, Lasso-based estimators (Lasso-IV, SS-Lasso-IV, post-Lasso-IV, SS-post-Lasso-IV) most often result in relatively adequate rejection frequencies, the same holds for RJIVE, SS-ENIV, and CF-ENIV. With the strong mixed signal, binary or Gaussian, SS-ENIV tends to produce slightly lower rejection frequencies than RJIVE, including the case of Gaussian instruments when both estimators slightly over-reject.

To sum up the results of the simulation study, the IV estimators based on elastic-net constitute a safe alternative to those based on lasso under an unknown degree of sparsity. In particular, the sample-split elastic-net IV estimator tends to dominate its lasso-based counterpart, the sample-split lasso IV estimator, as well as other lasso-based IV estimators, in terms of bias and test accuracy. In addition, the performance of the sample-split elastic-net IV estimator is comparable to that of the ridge jackknife IV estimator. SS-ENIV tends to result in slightly lower rejection frequencies than RJIVE, thus being superior in the settings when both estimators over-reject. RJIVE shows minor over-rejection in most settings considered with the mixed signal, thereby motivating further investigation of the relative performance of RJIVE and SS-ENIV estimators in various settings with uneven distribution of explanatory power across the instrumental variables. Finally, data generating processes with alternative degrees of sparsity are also worth examining.

Figure 15 presents frequency plots for the penalty ratio from first-stage regressions estimated via elastic-net. The elastic-net penalty ratio is $a/(a+b)$ where a and b come from representing the elastic-net penalty term $\lambda\left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right)$ as $a|\beta_j| + b\beta_j^2$. The penalty ratio is chosen through cross-validation.[33] For the ratio 1.0 the penalty is an l_1-penalty (lasso-type), whereas for the ratio 0.0 it is an l_2-penalty (ridge-type).

[33] I use a Python package, sklearn.linear_model.ElasticNetCV, to fit the first-stage via elastic-net, with a prespecified grid [0.01, 0.03, .05, .07, .1, .2, .5, .8, .9, 0.93, .95, 0.97, .99, 1]. For each value of the penalty ratio, the grid for a parameter α, which is also estimated through cross-validation, consists of 100 values and is defined automatically as part of the ElasticNetCV package.

Table 2.2. Simulation Results many instruments K = 190

	Sparse Signal			Dense Signal			Mixed Signal		
	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%	Med. Bias	MAD	RP 5%
	A. Concentration parameter = 30. Binary Instruments								
Lasso-IV	0.010	0.015	0.103	0.016	0.016	0.329	0.013	0.013	0.290
SS-Lasso-IV	0.001	0.032	0.009	0.007	0.038	0.003	0.004	0.025	0.001
post-Lasso-IV	0.010	0.015	0.120	0.015	0.015	0.359	0.013	0.013	0.333
SS-post-Lasso-IV	0.001	0.032	0.008	0.005	0.039	0.003	0.003	0.026	0.002
CF-Lasso-IV	0.025	0.025	0.000	0.009	0.027	0.000	0.006	0.012	0.000
RJIVE	0.000	0.025	0.042	0.000	0.009	0.043	0.000	0.008	0.048
ENIV	0.026	0.026	0.498	0.018	0.018	0.720	0.016	0.016	0.729
SS-ENIV	0.001	0.034	0.037	0.000	0.015	0.043	0.002	0.013	0.043
CF-ENIV	0.001	0.026	0.132	0.000	0.012	0.108	0.001	0.010	0.124
	B. Concentration parameter = 30. Gaussian Instruments								
Lasso-IV	0.005	0.011	0.074	0.010	0.01	0.275	0.007	0.008	0.231
SS-Lasso-IV	0.000	0.016	0.031	0.010	0.025	0.001	0.002	0.019	0.005
post-Lasso-IV	0.007	0.011	0.124	0.010	0.01	0.315		0.008	0.274
SS-post-Lasso-IV	0.000	0.016	0.030	0.011	0.023	0.001	0.002	0.019	0.005
CF-Lasso-IV	0.005	0.010	0.000	0.015	0.015	0.000	0.007	0.009	0.000
RJIVE	-0.001	0.017	0.052	0.000	0.008	0.050	-0.001	0.007	0.042
ENIV	0.013	0.015	0.348	0.011	0.011	0.688	0.009	0.009	0.641
SS-ENIV	0.001	0.019	0.044	0.002	0.013	0.045	0.001	0.013	0.025
CF-ENIV	0.001	0.015	0.110	0.002	0.011	0.129	0.001	0.010	0.141
	C. Concentration parameter = 150. Binary Instruments								
Lasso-IV	0.005	0.014	0.069	0.011	0.012	0.208	0.010	0.01	0.180
SS-Lasso-IV	-0.001	0.021	0.049	0.000	0.023	0.039	0.001	0.018	0.029
post-Lasso-IV	0.005	0.014	0.074	0.012	0.012	0.235	0.010	0.011	0.217
SS-post-Lasso-IV	0.000	0.021	0.051	0.001	0.020	0.042	0.001	0.016	0.032
CF-Lasso-IV	0.001	0.015	0.001	0.001	0.016	0.002	0.000	0.014	0.001
RJIVE	-0.001	0.018	0.053	0.000	0.008	0.059	0.000	0.007	0.049
ENIV	0.015	0.018	0.201	0.015	0.015	0.400	0.014	0.014	0.430
SS-ENIV	0.001	0.021	0.052	0.000	0.012	0.061	0.000	0.010	0.042
CF-ENIV	0.000	0.015	0.047	0.000	0.009	0.053	0.000	0.007	0.055
	D. Concentration parameter = 150. Gaussian Instruments								
Lasso-IV	0.003	0.010	0.055	0.009	0.010	0.239	0.007	0.008	0.207
SS-Lasso-IV	-0.001	0.016	0.043	0.002	0.018	0.021	0.000	0.014	0.036
post-Lasso-IV	0.004	0.010	0.060	0.010	0.010	0.274	0.008	0.008	0.259
SS-post-Lasso-IV	0.000	0.015	0.043	0.001	0.016	0.024	0.000	0.014	0.033
CF-Lasso-IV	-0.001	0.011	0.002	0.002	0.015	0.001	0.000	0.010	0.002
RJIVE	0.000	0.012	0.045	-0.001	0.006	0.053	0.000	0.006	0.064
ENIV	0.009	0.012	0.146	0.012	0.012	0.440	0.010	0.010	0.444
SS-ENIV	0.000	0.016	0.037	0.000	0.010	0.039	0.000	0.009	0.053
CF-ENIV	0.000	0.011	0.046	0.000	0.007	0.083	0.000	0.006	0.088

Note: Results are based on 1500 simulation replications. I report Median Bias (Med. Bias), Median absolute deviation (MAD) and rejection frequency for a 5% level test (RP 5%) for nine different estimators: the Lasso IV and post-Lasso IV estimators of Belloni et al. (2012, Lasso-IV and post-Lasso-IV), their sample-split versions (SS-Lasso-IV and SS-post-Lasso-IV), the cross-fit Lasso IV estimator, the RJIVE by Hansen and Kozbur (2014, RJIVE), and three estimators proposed in this paper: the elastic-net IV estimator (ENIV), the sample-split elastic-net IV estimator (SS-ENIV) and the cross-fit elastic-net IV estimator (CF-ENIV).

A combination of both l_1 and l_2 penalties is employed when the cross-validation procedure results in a value between 0 and 1. The results for a sparse, dense, and mixed DGP are shown in the first, second and third column of plots, respectively. As before, panels A, B,

C and D correspond to various instrument designs. Only the case with $p = 95$ is presented, since the results for the case with $p = 190$ look very similar.

When fitting the right combination of both l_1 and l_2 penalties to a first-stage relationship, the elastic-net estimator is quite successful in detecting a sparse structure, and thus often sets the penalty ratio to 1 in this case. When dealing with non-sparse first-stage relationships, the distribution of the penalty ratio is more even, with massive point mass on 0 and 1, and also on the intermediate values if the signal is strong ($\mu = 150$). Thus, the elastic-net estimator is performing better in combining l_1 and l_2 penalties when facing a strong signal, whereas it tends to often converge to a corner solution (imposing no ridge-type penalty, or no lasso-type penalty at all) when dealing with a weak signal ($\mu = 30$). In addition, the graphs presented indicate the need for a finer grid to search over for the best penalty ratio (especially around the middle value), for a better fit to the unknown sparsity of the data at hand.

2.4 Empirical Example

In this section, I demonstrate the application of the EN-based IV estimators to the classic example from the many-instrument literature – Angrist and Krueger (1991). The coefficient of interest in this example is the causal effect of schooling on earnings, and the schooling endogeneity is addressed through the use of instrumental variables. The data from Angrist and Krueger (1991) potentially allow one to employ many instruments for identification of the treatment effect, and there is a rich literature on consequences of alternative IV-choice decisions, in terms of both point estimate's and inference quality, driven by the numerosity and weakness of the available instrumental variables (Bound et al. 1995; Angrist et al. 1999; Staiger and Stock 1997; Hansen et al. 2008a).

The simple model under consideration is

$$\begin{aligned} \log(\text{wage}_i) &= \alpha\, \text{Schooling}_i + W_i'\gamma + \varepsilon_i \\ \text{Schooling}_i &= Z_i'\Pi_1 + W_i'\Pi_2 + u_i \end{aligned}$$

where ε_i and u_i satisfy $E[\varepsilon_i | W_i, Z_i] = E[u_i | W_i, Z_i] = 0$, $\log(\text{wage}_i)$ is a log of individual wage, Schooling_i is individual years of completed schooling, W_i is a vector of control variables and Z_i is a vector of instrumental variables that affect the wage only through the education channel. The data come from the 1980 U.S. Census and represent 329,509 men born between 1930 and 1939. The control set consists of 510 variables: a constant, 9 year-of-birth dummies, 50 state-of-birth dummies and 450 state-of-birth $\times$ year-of-birth cross-products. I employ three alternative sets of instruments, varying from three quarter-of-birth dummies to a full set of interactions with state-of-birth and year-of-birth control variables W_i, i.e. a total of 1,527 instrumental variables. By the identification argument of Angrist and Krueger (1991), α, the IV coefficient on Schooling_i, is a causal effect of education on earnings.

I report the results for three instrument sets in Table 3.3. For each set of instrumental variables, I present the estimates from conventional 2SLS, post-Lasso, SS-post-Lasso, ENIV, SS-ENIV, and CF-ENIV. For the estimators involving sample-splitting, I report two estimates (separated by / in Table 3.3) that result from swapping the sample halves used for fitting the

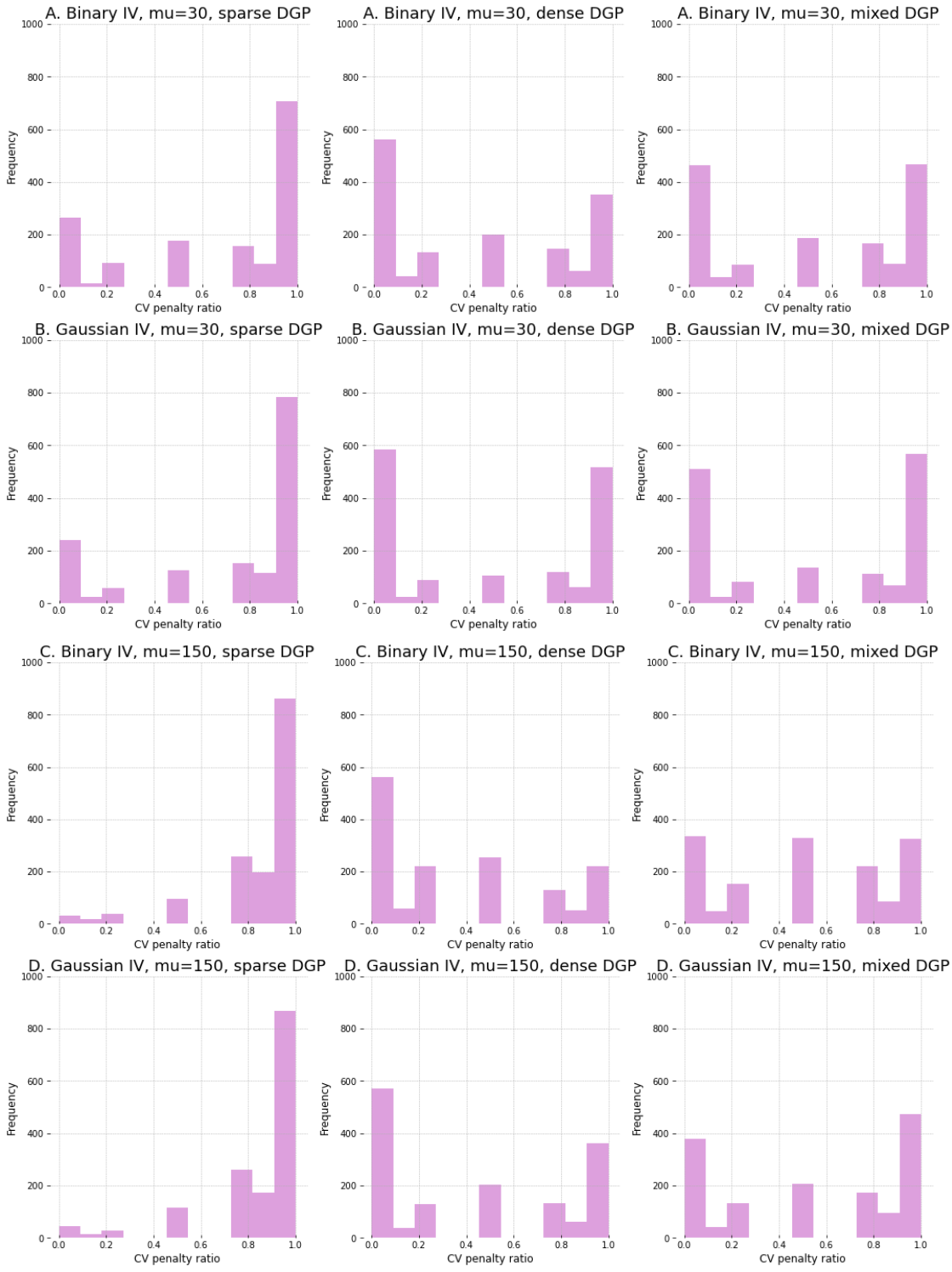

Figure 15: The penalty ratio chosen through cross-validation as part of the first-stage elastic-net regression. Cross-validation is performed on a grid from 0 to 1. Graphs show the frequency of each value being selected. For the penalty ratio 1 the penalty is an l_1-penalty; for the penalty ratio 0 it is an l_2-penalty; for the penalty ratio between 0 and 1 it is a combination of both. The case with $p = 95$ instruments and $n = 100$ observations is presented.

first stage. This way I demonstrate the sensitivity of the point estimates that takes place despite the large sample at hand.

Table 3.3

	2SLS	post-Lasso	SS-post-Lasso	RJIVE	ENIV	SS-ENIV	CF-ENIV
			A. 3 instruments				
Coefficient	0.108	0.111	0.097 / 0.112	0.109	0.108	0.098 / 0.118	0.108
St. error	0.020	0.0205	0.034 / 0.039	0.020	0.020	0.027 / 0.029	0.020
			B. 180 instruments				
Coefficient	0.093	0.112	0.097 / 0.112	0.106	0.093	0.103 / 0.114	0.108
St. error	0.010	0.017	0.034 / 0.039	0.016	0.010	0.026 / 0.027	0.009
			C. 1527 instruments				
Coefficient	0.071	0.086	0.097	0.107	0.074	0.079 / 0.145	0.112
St. error	0.005	0.025	0.039	0.017	0.005	0.061 / 0.064	0.004

Panel A uses the three main quarter-of-birth dummies from Angrist and Krueger (1991). As expected, all estimators considered result in similar point estimates and standard errors. Due to the high strength of each of the small number of instrumental variables being used, the methods involving regularization impose a small regularization penalty, thus leading to nearly identical results as 2SLS.

Panel B employs 180 instruments including the three quarter-of-birth dummies and their cross-products with the 9 year-of-birth dummies and 50 state-of-birth dummies. This set is also used in Angrist and Krueger (1991), with the aim of increasing the efficiency of the estimates. As expected, the 2SLS estimate is biased toward the OLS estimate of 0.0673. The same applies to ENIV that actually employs approximately as many instruments as 2SLS does. Post-Lasso, SS-post-Lasso, SS-ENIV, and CF-ENIV tend to deliver adequate estimates, though the instability of the estimators involving sample splitting is noticeable. The post-Lasso estimator does not have a downward bias, while CF-ENIV results in the smallest estimated standard error.

In Panel C, I show results based on the full set of 1527 instrumental variables. Even stronger bias of the 2SLS estimate towards the OLS estimate is observed. In this case, the SS-post-Lasso estimator tends to select no variables into the first stage regression (therefore, only a single number is provided). The post-Lasso, SS-post-Lasso, ENIV estimators now also result in a substantial downward bias. However, the CF-ENIV still delivers a reasonable point estimate, and also the smallest estimated standard error as well.

2.5 Conclusion

In this paper, I propose elastic-net instrumental variable estimators to deal with high-dimensional sets of instruments. The proposed estimators can be asymptotically equivalent to the lasso-based IV estimators but have better sampling properties if correlations among the instruments are not negligible. In addition, the IV estimators based on elastic-net are robust to deviations of the first-stage regression from sparsity. These features make the elastic-net IV estimators a valuable alternative to the lasso IV estimators for policy evaluation.

Appendix 2

Proof of Proposition 1.

Lemma 1 from Zou and Hastie (2006) shows that the naive elastic-net criterion

$$L\left(\lambda_1^{EN}, \lambda_2^{EN}, \beta\right) = |y - X\beta|^2 + \lambda_1^{EN} |\beta|_1 + \lambda_2^{EN} |\beta|_2$$

can be written as the lasso criterion

$$L\left(\gamma, \beta^*\right) = |y^* - X^*\beta^*|^2 + \gamma |\beta^*|_1,$$

where $\gamma = \lambda_1^{EN}/\sqrt{1+\lambda_2^{EN}}$, $\beta^* = \sqrt{1+\lambda_2^{EN}}\beta$, and an augmented data set (y^*, X^*) is defined by

$$X^*_{(n+p)\times p} = \left(1+\lambda_2^{EN}\right)^{-1/2} \begin{pmatrix} X \\ \sqrt{\lambda_2^{EN}} I \end{pmatrix}, \qquad y^*_{(n+p)} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

Then, for $\widehat{\beta}^* = \arg\min_\beta L\left(\gamma, \beta^*\right)$,

$$\widehat{\beta}^{EN} = \frac{1}{\sqrt{1+\lambda_2^{EN}}} \widehat{\beta}^*.$$

Having the elastic-net problem represented as the lasso problem, we can directly apply the results from Corollary 1 by BCCH on lasso's convergence rates under non-Gaussian and heteroskedastic errors. For a properly chosen γ,

$$\left\|\widehat{D}_i^* - D_i^*\right\|_{2,n} \lesssim_P \sqrt{\frac{s \log\left(p \vee (n+p)\right)}{n+p}} = \sqrt{\frac{s \log\left(n+p\right)}{n+p}}$$

and therefore,

$$\left\|\widehat{D}_i^{EN} - D_i\right\|_{2,n} \lesssim_P \sqrt{\frac{s \log\left(n+p\right)}{n+p}}.$$

Similarly, using the second inequality from Corollary 1,

$$\left\|\widehat{\beta}^* - \beta^*\right\|_1 \lesssim_P \sqrt{\frac{s^2 \log\left(n+p\right)}{n+p}},$$

and it can be written as

$$\left\|\widehat{\beta}^{EN} - \beta\right\|_1 \lesssim_P \frac{1}{\sqrt{1+\lambda_2^{EN}}} \sqrt{\frac{s^2 \log\left(n+p\right)}{n+p}} \leq \sqrt{\frac{s^2 \log\left(n+p\right)}{n+p}},$$

thus giving us a sufficient condition for Theorem 4 by BCCH to hold.

3 Many Instruments: Implementation in STATA

Published as Anatolyev, Stanislav, and Alena Skolkova. 2019. "Many instruments: Implementation in Stata." *The Stata Journal* 19(4), 849-866.

3.1 Introduction

Instrumental variables (IV) estimation and inference have long been a distinctive method in applied microeconometric analysis and have often spurred advances in econometric theory. The IV methods were designed to address endogeneity bias from OLS in estimating a causal/treatment effect in structural models (such as an effect of smoking on health, returns to education, or demand elasticity), see Angrist and Krueger (2001). At the dawn of the 21st century, both theory and practice were extended to accommodate such complications as weak instruments, numerous instruments, and combinations thereof. It was established that the empiricist's workhorse, the two-stage least-squares (2SLS) estimator, fails to deliver consistent estimates and results in invalid inference when such complications arise, and alternative approaches to estimation and inference were proposed. The quick progress in econometric theory did not, however, carry over to empirical practice as fast.

The seminal article by Bekker (1994b) proposed an alternative asymptotic approximation for linear normal homoskedastic IV regressions with many instrumental variables, together with consistent estimation and construction of valid standard errors within the new paradigm of dimension asymptotics. Since then, there has been a significant progress in the theory of estimation and testing in IV regressions with many, possibly weak, instruments. Many new or modified versions of old estimators and tests have been proposed, including, among others, limited information maximum likelihood (LIML), bias-corrected 2SLS, several versions of jackknife IV estimators, and so on. In an important article, Hansen et al. (2008b) proposed extensions of estimation and inference methods based on LIML, when, in particular, the structural and first stage errors are not necessarily normal and when the instruments may be weak as a group. More recently, Hausman et al. (2012) showed that the leading 'homoskedastic' estimators fail to deliver consistency in heteroskedastic models, and proposed their 'heteroskedastic' modifications. Specification testing tools were developed in Anatolyev and Gospodinov (2011) and Lee and Okui (2012) for the homoskedastic case and in Chao et al. (2014) for the heteroskedastic case.

The state-of-the-art theoretical literature has converged to suggesting estimation based on LIML and its Fuller (1977)-type correction that remedies the problem of non-existence of moments. Parameter inference is based on consistent estimation of up to four terms in the asymptotic variance, while specification testing is based on asymptotically normal (or asymptotically equivalent possibly adjusted chi-squared) distribution of the overidentifying test statistic. The literature has shown that all these tools are robust to weakness of the instruments as a group (though weakness of a lesser degree than that would jeopardize identification). We describe these tools in brief in the following sections; see the recent survey Anatolyev (2019) for more technical details as well as the history of theoretical developments and suggestions of empirical strategies.

Despite the theoretical advances, practitioners rarely use appropriate tools because of their non-availability in popular econometric packages, STATA in particular. The present contribution aims at filling this void. We introduce a STATA command, `mivreg`, that implements consistent estimation and testing in linear IV regressions with many, possibly weak, instruments. This command covers both homoskedastic and heteroskedastic environments, estimators that are both non-robust and robust to error non-normality and projection matrix limit, both parameter tests and specification tests. Even though, as noted above, a number of other consistent estimators have been proposed, we build up `mivreg` around the leading LIML estimator and its Fuller (1977) correction as suggested by the state-of-the-art literature.

In Section 2, we set out the model and introduce necessary notation. In Sections 3 and 4, we describe estimation and testing tools pertaining to the homoskedastic and heteroskedastic models, respectively. In Section 5, we present the new command, `mivreg`. In Section 5, we illustrate how `mivreg` works in simulations and compare it with the classical command `ivregress` in Section 6. Finally, in Section 7, we illustrate how `mivreg` works with real data.

3.2 Model

The structural equation is

$$y_i = x_i'\beta_0 + e_i,$$

where β_0 is $k\times 1$ vector of structural coefficients of interest, or in matrix notation, $Y = X\beta_0+e$, where $Y = (y_1, ..., y_n)'$ is $n \times 1$, $X = (x_1, ..., x_n)'$ is $n \times k$, and $e = (e_1, ..., e_n)'$ is $n \times 1$. The first stage equation is

$$x_i = z_i'\Gamma + u_i,$$

where z_i is $\ell \times 1$ vector of instruments and Γ is $\ell \times k$ matrix of first stage coefficients, or in matrix notation, $X = Z\Gamma + U$, where $U = (u_1, ..., u_n)'$ is $n \times k$. We assume that the rank of instrument matrix $Z = (z_1, ..., z_n)'$ equals its column dimension ℓ. The structural and first stage errors follow

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} | z_i \sim \mathcal{D}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \Psi_i' \\ \Psi_i & \Omega_i \end{pmatrix} \right),$$

for some distribution $\mathcal{D}$, normal $\mathcal{N}$ being a possibility. Under conditional homoskedasticity, $\sigma_i^2 = \sigma^2$, $\Psi_i = \Psi$ and $\Omega_i = \Omega$ for all $i = 1, ..., n$.

Introduce the projection matrices associated with the instruments

$$P = Z\left(Z'Z\right)^{-1}Z', \quad M = I_n - P.$$

The $(i, j)^{th}$ element of P is denoted P_{ij}. Let us also denote by D the diagonal matrix with diagonal elements of P on the main diagonal: $D = \text{diag}\left\{P_{ii}\right\}_{i=1}^n$. By $\overline{P_{ii}^2}$ we denote an average of diagonal elements of P squared: $\overline{P_{ii}^2} = n^{-1}\text{tr}\left(D^2\right)$.

3.3 Homoskedastic case

In the conditionally homoskedastic case, correct parameter estimation and inference was developed in Bekker (1994b) and Hansen et al. (2008b). Specification testing was dealt with in Anatolyev and Gospodinov (2011) and Lee and Okui (2012).

3.3.1 Point estimation

Under many instruments, 2SLS estimation is inconsistent. The leading consistent estimator is the limited information maximum likelihood (LIML) estimator

$$\hat{\beta}_{LIML} = \arg\min_{\beta} \frac{(Y - X\beta)' P (Y - X\beta)}{(Y - X\beta)' (Y - X\beta)}.$$

Numerically, instead of the above optimization problem, it can be found via the eigenvalue problem:

$$\hat{\beta}_{LIML} = \bar{H}^{-1} X' \mathring{P} Y,$$

where

$$\bar{H} = X' \mathring{P} X,$$

and $\mathring{P} = P - \bar{\alpha} I_n$, and $\bar{\alpha}$ is the smallest eigenvalue of the matrix $(\mathring{X}'\mathring{X})^{-1}\mathring{X}' P \mathring{X}$, where $\mathring{X} = (Y, X)$.

The LIML estimator has a disadvantage that even its low order moments do not exist. A simple Fuller (1977) adjustment solves the moment problem:

$$\tilde{\alpha} = \frac{\bar{\alpha} - (1 - \bar{\alpha})\varsigma/n}{1 - (1 - \bar{\alpha})\varsigma/n}. \tag{9}$$

This adjustment leads to the FULL estimator, where $\bar{\alpha}$ is replaced by $\tilde{\alpha}$ everywhere. It is usually advised to use the value $\varsigma = 1$ in practice.

Denote the vector of LIML or FULL residuals by $\hat{e}$, then

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n - k}$$

is the residual variance.

3.3.2 Variance estimation

Under error normality and/or asymptotically constant diagonal of P, the asymptotic variance is estimated by

$$\bar{V} = n\bar{H}^{-1}\bar{\Sigma}_0\bar{H}^{-1},$$

where

$$\bar{\Sigma}_0 = \hat{\sigma}^2 \left((1 - \bar{\alpha})^2 \bar{X}' P \bar{X} + \bar{\alpha}^2 \bar{X}' (I_n - P) \bar{X} \right),$$

and

$$\bar{X} = X - \hat{e}\frac{\hat{e}' X}{\hat{e}'\hat{e}}$$

(Bekker (1994b), Hansen et al. (2008b)).

Under error non-normality and asymptotically variable diagonal of P, the asymptotic variance is estimated by

$$\bar{V}_R = n\bar{H}^{-1} \left(\bar{\Sigma}_0 + \bar{\Sigma}_A + \bar{\Sigma}_A' + \bar{\Sigma}_B \right) \bar{H}^{-1},$$

where the subscript R stands for 'robust', and in addition

$$\bar{\Sigma}_A = \left(\sum_{i=1}^{n}\left(P_{ii} - \frac{\ell}{n}\right)(PX)_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\left(M\bar{X}\right)_i\right)'$$

and

$$\bar{\Sigma}_B = \frac{\overline{P_{ii}^2} - (\ell/n)^2}{1 - 2\ell/n + \overline{P_{ii}^2}}\sum_{i=1}^{n}\left(\hat{e}_i^2 - \hat{\sigma}^2\right)\left(M\bar{X}\right)_i\left(M\bar{X}\right)_i'$$

(Hansen et al. (2008b)).

The variance estimates $\bar{V}$ and $\bar{V}_R$ are a basis of parameter inference. For example, the standard error for j^{th} parameter can be computed as $\sqrt{\bar{V}_{jj}/n}$.

3.3.3 Specification testing

Consider the conventional J statistic

$$J = \frac{\hat{e}'P\hat{e}}{\hat{\sigma}^2} = (n-k)\,\bar{\alpha},$$

and the bias-corrected J statistic

$$J_R = J - \frac{\ell}{n}\frac{\hat{e}'\hat{e}}{\hat{\sigma}^2} = (n-k)\left(\bar{\alpha} - \frac{\ell}{n}\right),$$

where the subscript R stands for 'robust'.

Under error normality and/or asymptotically constant diagonal of P, the Anatolyev and Gospodinov (2011) test prescribes rejecting correct model specification at significance level ϕ when the value of J exceeds $q_{\phi^*}^{\chi^2(\ell-k)}$, the $(1-\phi^*)$-quantile of the chi-squared with $\ell - k$ degrees of freedom, where

$$\phi^* = \Phi\left(\sqrt{1 - \frac{\ell}{n}}\cdot\Phi^{-1}(\phi)\right).$$

Under error non-normality and asymptotically variable diagonal of P, the Lee and Okui (2012) test prescribes rejecting correct model specification at significance level ϕ when the value of

$$\frac{J_R}{\sqrt{n\hat{V}^J}}$$

exceeds $q_\phi^{\mathcal{N}(0,1)}$, the $(1-\phi)$-quantile of the standard normal. Here,

$$\hat{V}^J = 2\frac{\ell}{n}\left(1 - \frac{\ell}{n}\right) + \left(\overline{P_{ii}^2} - \left(\frac{\ell}{n}\right)^2\right)\left(\frac{\overline{\hat{e}_i^4}}{\hat{\sigma}^4} - 3\right).$$

3.4 Heteroskedastic case

In the conditionally heteroskedastic case, correct parameter estimation and inference were developed in Hausman et al. (2012). Specification testing was dealt with in Chao et al. (2014).

3.4.1 Point estimation

The HLIM ('heteroskedastic LIML') estimator is

$$\hat{\beta}_{HLIM} = \arg\min_{\beta} \frac{(Y - X\beta)'(P - D)(Y - X\beta)}{(Y - X\beta)'(Y - X\beta)}$$

Numerically, it can be found via the eigenvalue problem:

$$\hat{\beta}_{HLIM} = \bar{H}^{-1} X' \mathring{P} Y,$$

where

$$\bar{H} = X' \mathring{P} X,$$

and $\mathring{P} = P - D - \bar{\alpha} I_n$, and $\bar{\alpha}$ is the smallest eigenvalue of the matrix $(\mathring{X}'\mathring{X})^{-1}\mathring{X}'(P - D)\mathring{X}$, where $\mathring{X} = (Y, X)$. Similarly to FULL, the Fuller (1977) adjustment (9) leads to HFUL ('heteroskedastic FULL') estimation.

Denote the vector of HLIM or HFUL residuals by $\hat{e}$, then

$$\hat{\sigma}^2 = \frac{\hat{e}'\hat{e}}{n - k}$$

is the residual variance.

3.4.2 Asymptotic variance estimation

Hausman et al. (2012) provide a valid and robust variance estimator for the HLIM estimator:

$$\bar{V} = n\bar{H}^{-1}\bar{\Sigma}\bar{H}^{-1},$$

where

$$\bar{\Sigma} = \sum_{i=1}^{n}((P\bar{X})_i(P\bar{X})_i' - P_{ii}\bar{X}_i(P\bar{X})_i' - P_{ii}(P\bar{X})_i\bar{X}_i')\hat{e}_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}^2 \bar{X}_i \bar{X}_j' \hat{e}_i \hat{e}_j, \tag{10}$$

where

$$\bar{X} = X - \hat{e}\frac{\hat{e}'X}{\hat{e}'\hat{e}}.$$

The variance estimate $\bar{V}$ is a basis of parameter inference. For example, the standard error for j^{th} parameter can be computed as $\sqrt{\bar{V}_{jj}/n}$.

3.4.3 Specification testing

Chao et al. (2014) generalize the specification J test for the heteroskedastic case. Their statistic is based on the jackknife modification of J statistic's quadratic form:

$$J = \frac{\hat{e}'(P-D)\hat{e}}{\sqrt{\hat{V}^J}} + \ell,$$

where

$$\hat{V}^J = \frac{1}{\ell}\sum_{i\neq j}\hat{e}_i^2 P_{ij}^2 \hat{e}_j^2 = \frac{1}{\ell}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{e}_i^2 P_{ij}^2 \hat{e}_j^2 - \sum_{i=1}^{n} P_{ii}^2 \hat{e}_j^4\right) \tag{11}$$

is an estimate of the variance of the modified quadratic form.

The test is one-sided, and the decision rule is reject the null of instrument validity if the value of J exceeds $q_\phi^{\chi^2(\ell-k)}$, the $(1-\phi)$-quantile of the $\chi^2(\ell-k)$ distribution.

3.5 Command `mivreg`

3.5.1 Functionality

The command `mivreg` implements estimation, inference on individual parameters and specification testing under many, possibly weak, instruments. The default 'hom' (for 'homoskedastic') option is based on the LIML or FULL estimators, the 'het' (for 'heteroskedastic') option is based on the HLIM or HFUL estimators. Within the 'hom' version, the 'robust' option leads to the Hansen–Hausman–Newey variance estimator and Lee–Okui specification test, while the default non-robust variation computes the Bekker variance estimator and Anatolyev–Gospodinov specification test. The 'hetero' version implements the Hausman–Newey–Woutersen–Chao–Swanson variance estimator and Chao–Hausman–Newey–Swanson–Woutersen specification test. By default, the estimators used are LIML or HLIM; the 'fuller' option makes the Fuller correction with parameter $\varsigma = 1$, and so the FULL or HFUL estimators are used instead.

3.5.2 Syntax

`mivreg` *depvar* [*indepvars*] (*varlist1* = *varlist2*) [*if*] [*in*] [, `hom` `het` `robust` `fuller` `level`(#)]

`by`, `rolling`, `statsby` and `xi` are allowed.

3.5.3 Description

The command `mivreg` performs estimation, inference on individual parameters and specification testing under many possibly weak instruments. The dependent variable *depvar* is modeled as a linear function of *indepvars* and *varlist1*, using *varlist2* (along with *indepvars*) as instruments for *varlist1*.

3.5.4 Options

`hom` uses the LIML (default) or FULL (in combination with `full` option) estimator.

`het` uses the HLIM (default) or HFUL (in combination with `full` option) estimator.

`robust` leads, under `hom` option, to the Hansen–Hausman–Newey variance estimator and the Lee–Okui specification test, while the default non-robust variation computes the Bekker variance estimator and the Anatolyev–Gospodinov specification test; under `het` option, to the Hausman–Newey–Woutersen–Chao–Swanson variance estimator and the Chao–Hausman–Newey–Swanson–Woutersen specification test.

`fuller` makes the Fuller correction with parameter $\varsigma = 1$, which leads to the FULL (in combination with `hom` option) or HFUL (in combination with `het` option) estimator.

`level(#)` sets the confidence level; the default is `level(95)`.

3.5.5 Saved results

`mivreg` saves the following in `e()`:

Scalars			
`e(N)`	number of observations	`e(F1)`	first-stage F statistic
`e(rmse)`	root mean squared error	`e(df_m_F1)`	first-stage model degrees of freedom
`e(F)`	model F statistic	`e(df_r_F1)`	first-stage residual degrees of freedom
`e(df_m)`	model degrees of freedom	`e(r2_1)`	first-stage R^2
`e(df_r)`	residual degrees of freedom	`e(jval)`	model J statistic
`e(r2)`	R^2	`e(jpv)`	J-test p-value
`e(r2_a)`	adjusted R^2		
Macros			
`e(model)`	hom or het	`e(instd)`	instrumented variables
`e(title)`	title in estimation output	`e(insts)`	instruments
`e(depvar)`	name of dependent variable	`e(properties)`	`b V`
Matrices			`b V`
`e(b)`	coeffcient vector	`e(V)`	variance-covariance matrix of the estimators
Functions			
`e(sample)`	marks estimation sample		

3.5.6 Computational notes

First, throughout we avoid storing $n \times n$ matrices like P and I_n in the memory. For example, we compute $\bar{H} = X' (P - \bar{\alpha} I_n) X$ as

$$\bar{H} = X' Z (Z'Z)^{-1} Z' X - \bar{\alpha} X' X.$$

Second, the last term in (10) can be alternatively computed without double summations over n observations (Hausman et al. (2012)):

$$\sum_{p=1}^{\ell} \sum_{r=1}^{\ell} \left(\sum_{i=1}^{n} \tilde{Z}_{ip} \tilde{Z}_{ir} \bar{X}_i \hat{e}_i \right) \left(\sum_{j=1}^{n} Z_{jp} Z_{jr} \bar{X}_j \hat{e}_j \right)',$$

where $\tilde{Z} = Z(Z'Z)^{-1}$. Similarly, the full double summation in (11) can analogously be computed as

$$\sum_{p=1}^{\ell}\sum_{r=1}^{\ell}\left(\sum_{i=1}^{n}\tilde{Z}_{ip}\tilde{Z}_{ir}\hat{e}_i^2\right)\left(\sum_{j=1}^{n}Z_{jp}Z_{jr}\hat{e}_j^2\right).$$

3.6 Simulations

3.6.1 Artificial data

We demonstrate how `mivreg` works with two sets of artificial data. The artificial data are generated from the Monte-Carlo setup in Hausman et al. (2012). The estimated equation is

$$y = \beta_1 + \beta_2 x_2 + e,$$

and the first stage equation is

$$x_2 = \gamma z_1 + u_2,$$

where $z_1 \sim N(0,1)$ and $u_2 \sim N(0,1)$. The instrument vector is

$$z = \left(1, z_1, z_1^2, z_1^3, z_1^4, z_1 d_1, ..., z_1 d_{\ell-5}\right)',$$

where $d_j \in \{0,1\}$ with $\Pr\{d_j = 1\} = \frac{1}{2}$ independent of z_1. The structural disturbance is given by

$$e = 0.30u_2 + \sqrt{\frac{1-0.30^2}{\phi^2 + 0.86^4}}\left(\phi v_1 + 0.86 v_2\right),$$

with $v_1 \sim N(0,1)$ in the homoskedastic case and $v_1 \sim N(0, z_1^2)$ in the heteroskedastic case, and $v_2 \sim N(0, 0.86^2)$, both v_1 and v_2 being independent of u_2. Samples of size $n = 400$ are generated, with $\ell = 30$ instruments, the instrument strength γ is chosen so that the concentration parameter equals $n\gamma^2 = 32$. The parameter ϕ is set at the value 0.8 which in the heteroskedastic case corresponds to $R^2 \approx 0.25$ in the skedastic regression. The true values of β_1 and β_2 are set at 1.

Note that the instrument vector is such that the diagonal of P is asymptotically heterogeneous (see Anatolyev and Yaskov (2017)). In the homoskedastic case, simplifications due to error normality pertaining to variance estimation and specification testing (see subsections 3.2 and 3.3) are applicable.

3.6.2 Simulation results

In this section, we report output statistics resulting in simulations from using `mivreg` and compare it with that when the STATA command `ivregress` was used.[34] The reported results are obtained from 10,000 simulations.

[34]For example, to compute 2SLS-related statistics, `ivregress 2sls y one (x = z*), nocons robust` was used.

Table 3.1. Percentiles of simulated distribution of various estimators.

Estimator	Homoskedastic case					Heteroskedastic case				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
command `ivregress`										
2SLS	0.93	1.06	1.14	1.23	1.35	0.85	1.02	1.14	1.26	1.43
GMM	0.91	1.05	1.14	1.23	1.37	0.85	1.02	1.14	1.26	1.42
LIML	0.47	0.83	1.00	1.16	1.42	−4.08	−0.27	0.49	1.07	4.48
command `mivreg`										
LIML	0.47	0.83	1.00	1.16	1.42	−4.08	−0.27	0.49	1.07	4.48
FULL	0.52	0.84	1.01	1.17	1.41	−1.14	−0.03	0.56	1.09	2.77
HLIM	0.43	0.82	1.00	1.17	1.43	0.15	0.76	1.01	1.22	1.62
HFUL	0.52	0.84	1.01	1.17	1.43	0.30	0.79	1.02	1.22	1.60

Note: The true value of the parameter is unity.

First, we focus on point estimates. Table 3.1 collects percentiles of simulated distributions of 2SLS, LIML and GMM estimators produced by `ivregress`, and LIML, FULL, HLIM and HFUL estimators produced by `mivreg`. Naturally, the LIML rows coincide.

The 2SLS and GMM estimators (whose results are very similar) are always rightward biased, as expected. In the homoskedastic case, all the other estimators deliver unbiased estimation. The LIML estimator is a bit more concentrated towards the center than HLIM, which reflects higher efficiency of the former. The Fuller versions are more concentrated away from the tails, which reflects their resistance to outliers. In the heteroskedastic case, LIML and FULL have severe negative biases, which reflects their inconsistency. Their 'heteroskedastic' versions, HLIM and HFUL, are both median unbiased. While the HLIM estimator is susceptible to outliers, especially in the left tail, its Fuller version, HFUL, exhibits much tighter and more symmetric distribution.

Table 3.2 contains actual rejection rates corresponding to the 5% nominal rate for the two sided t-test of the null $H_0 : \beta_2 = 1$ marked as $t_{\beta_2=1}$, the Wald test of the null $H_0 : \beta_1 = \beta_2 = 1$ marked as $W_{\beta_1=\beta_2=1}$, and the specification test marked as $J_{E[ze]=0}$. The 2SLS and LIML tests produced by `ivregress` come in two forms: non-robust and robust to heteroskedasticity. In the specification tests (which are available only for efficient estimators), the Basmann (1957) variance estimator is used. The test statistics produced by `mivreg` use the following estimators and robustness regimes:[35] non-robust LIML, non-robust FULL, robust LIML, robust FULL, HLIM, and HFUL.

As expected, severe size distortions are exhibited by conventional parameter tests based on

[35]Note again the different use of the term 'robust': the classical tests produced by `ivregress` may be robust to heteroskedasticity; of course, they are not robust to instrument numerosity. The tests produced by `mivreg` may or may not be robust, within natural robustness to many possibly weak instruments, to error non-normality and asymptotically variable diagonal of the projection matrix.

Table 3.2. Actual rejection rates for parameter and specification tests

Estimator	Homoskedastic case			Heteroskedastic case		
	$t_{\beta_2=1}$	$W_{\beta_1=\beta_2=1}$	$J_{E[ze]=0}$	$t_{\beta_2=1}$	$W_{\beta_1=\beta_2=1}$	$J_{E[ze]=0}$
			command `ivregress`			
non-robust 2SLS	22.0%	17.7%	6.2%			
robust 2SLS				14.9%	13.1%	–
GMM	33.9%	31.8%	2.5%	26.8%	24.4%	2.1%
non-robust LIML	12.0%	9.6%	3.0%			
robust LIML				1.6%	1.3%	–
			command `mivreg`			
non-robust LIML	4.1%	4.3%	3.0%	9.4%	4.6%	60.1%
non-robust FULL	4.2%	4.5%	2.4%	9.3%	4.7%	56.8%
robust LIML	4.0%	4.3%	2.1%	9.2%	4.5%	54.2%
robust FULL	4.2%	4.5%	1.7%	9.2%	4.6%	50.9%
HLIM	4.7%	4.9%	2.8%	5.4%	4.9%	3.5%
HFUL	5.0%	5.2%	2.9%	5.7%	5.1%	3.4%

Note: The nominal significance level of all tests is 5%.

2SLS, GMM and LIML.[36] In the homoskedastic case, all the `mivreg` tests exhibit similar behavior, with much smaller distortions, though the 'heteroskedastic' versions seem to be more reliable. In the heteroskedastic case, the latter are the only valid ones theoretically, and do deliver rejection rates close to nominal. The Fuller correction does not significantly affect these rejection rates. The results of specification testing point at huge distortions if one relies on 'homoskedastic' specification tests when in fact the homoskedasticity assumption is violated. One must avoid using them in heteroskedastic environments as one is too much likely to receive a signal of instrument invalidity when in fact the instruments are valid.

3.7 Example with real data

We illustrate the use of `mivreg` using real data from a well-known application to the married female labor supply of Mroz (1987). The number of observations is 428.[37]

The left-side variable is working hours `hours`, the only endogenous right-side variable is log wages `lwage`; there are also 6 exogenous controls: `nwifeinc`, `educ`, `age`, `kidslt6`, `kidsge6`, and the constant `one`. The list of basic instruments includes, in addition to the 6 exogenous controls, 8 exogenous variables: `exper`, `expersq`, `fatheduc`, `motheduc`, `hushrs`, `husage`,

[36]The conventional specification tests do not exhibit too much of distortions in this particular design; however, in general they may well do; see Anatolyev and Gospodinov (2011).

[37]The data can be found at `http://www.stata.com/data/jwooldridge/eacsap/mroz.dta`. We use only the records that correspond to women in labor force.

Table 3.3. Various estimates of wage coefficient for married female labor supply

Options	Estimator	Instruments	Estimate	(Standard error)
		command `reg`		
`robust`	OLS	–	−17.4	(81.4)
		command `ivregress`		
`robust`	2SLS	basic only	1179.1	(185.2)
`robust`	2SLS	extended	536.4	(101.5)
		command `mivreg`		
`hom`	LIML	extended	1120.6	(195.3)
`hom robust fuller`	FULL	extended	1110.0	(197.2)
`het robust fuller`	HFUL	extended	1058.3	(170.5)

`huseduc`, and `mtr`, resulting in 14 instruments in total. The basic instruments are pretty strong as a group: the first-stage F statistic equals 183.5. We also consider an extended set of instruments – the basic instruments plus all their cross-products ('interactions'), the total numerosity amounting to 92. The use of the extended instrument set is meant to possibly enhance estimation efficiency by exploiting information in the instruments more actively. However, while the conventional tools are suitable for the basic set of instruments, the extended instrument set evidently requires handling via many-instrument asymptotics: the ratio of the number of instruments to the sample size is sizable: $\ell/n \approx 0.215$.

Table 3.3 presents various estimates for the slope coefficient of log wages: OLS, heteroskedasticity-robust 2SLS (employing the basic and extended instrument sets), as well as three many-instrument-robust estimators – LIML, FULL and HFUL (employing the extended instrument set) – whose STATA output will appear below.

Evidently, due to unaccounted endogeneity, OLS estimation from applying the `reg` command is inconsistent; the numerical value of the OLS estimate is even negative revealing a big endogeneity bias. The (more than twofold!) difference between the two 2SLS estimates points at invalidity of conventional tools and the `ivregress` command when instruments are many. The LIML, FULL and HFUL point estimates produced by the `mivreg` command are quite in line with the 2SLS estimate that uses only the basic instruments.[38] There is a small difference between 'homoskedastic' LIML and FULL point estimates and the 'heteroskedastic' HLIM point estimate. Though not too big, this difference makes the HFUL estimate more trustworthy.[39] The smaller standard error of HLIM compared to that of 2SLS may be interpreted as a gain in efficiency from using the extended instrument set.

[38]Note also from the STATA outputs that all three corresponding specification tests produce very high p-values and agree on the model validity.

[39]Mroz (1987) reports a similar 2SLS estimate using a short list of instruments (line 2 in his Table IV), but 2SLS estimates also get a lot smaller with longer lists of instruments (lines 3–6 in Table IV). Eventually, Mroz (1987) adopts smaller estimates than ones seeming correct from our experiments.

The STATA outputs produced by the command `mivreg` to deliver the three many-instrument-robust estimators appear next.

Example

The STATA output for LIML estimation with option `hom`:

```
. mivreg hours nwifeinc educ age kidslt6 kidsge6 one (lwage = nwifeinc educ ///
> age kidslt6 kidsge6 one exper expersq fatheduc motheduc hushrs husage ///
> huseduc mtr *X*) if inlf==1 , hom
MIVREG estimation (HOM)
First-stage summary
------------------------------
F(  86,   336)  =    2.07
Prob > F        =  0.0000
R-squared       =  0.8471
------------------------------
                                      Number of obs   =     428
                                      F(   7,   421)  =   95.74
                                      Prob > F        =  0.0000
                                      R-squared       = -0.5157
                                      Adj R-squared   = -0.5373
                                      Root MSE        =  1.1e+03
LIML estimation
Bekker variance estimation
```

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lwage	1120.595	195.3494	5.74	0.000	736.6134	1504.577
nwifeinc	-7.890468	5.261348	-1.50	0.134	-18.23225	2.451317
educ	-133.1851	31.79141	-4.19	0.000	-195.6748	-70.69543
age	-9.954741	7.918058	-1.26	0.209	-25.51859	5.609111
kidslt6	-246.5892	143.8619	-1.71	0.087	-529.3663	36.18793
kidsge6	-65.87681	44.77805	-1.47	0.142	-153.8932	22.13958
one	2345.98	487.9451	4.81	0.000	1386.868	3305.092

```
Instrumented:  lwage
Instruments:   nwifeinc educ age kidslt6 kidsge6 one exper expersq fatheduc
               motheduc hushrs husage huseduc mtr educXnwifeinc ageXnwifeinc
               ageXeduc kidslt6Xnwifeinc kidslt6Xeduc kidslt6Xage
               kidsge6Xnwifeinc kidsge6Xeduc kidsge6Xage kidsge6Xkidslt6
               experXnwifeinc experXeduc experXage experXkidslt6 experXkidsge6
               expersqXnwifeinc expersqXeduc expersqXage expersqXkidslt6
               expersqXkidsge6 expersqXexper fatheducXnwifeinc fatheducXeduc
               fatheducXage fatheducXkidslt6 fatheducXkidsge6 fatheducXexper
               fatheducXexpersq motheducXnwifeinc motheducXeduc motheducXage
               motheducXkidslt6 motheducXkidsge6 motheducXexper
               motheducXexpersq motheducXfatheduc hushrsXnwifeinc hushrsXeduc
               hushrsXage hushrsXkidslt6 hushrsXkidsge6 hushrsXexper
               hushrsXexpersq hushrsXfatheduc hushrsXmotheduc husageXnwifeinc
               husageXeduc husageXage husageXkidslt6 husageXkidsge6
               husageXexper husageXexpersq husageXfatheduc husageXmotheduc
               husageXhushrs huseducXnwifeinc huseducXeduc huseducXage
               huseducXkidslt6 huseducXkidsge6 huseducXexper huseducXexpersq
               huseducXfatheduc huseducXmotheduc huseducXhushrs huseducXhusage
               mtrXnwifeinc mtrXeduc mtrXage mtrXkidslt6 mtrXkidsge6 mtrXexper
               mtrXexpersq mtrXfatheduc mtrXmotheduc mtrXhushrs mtrXhusage
               mtrXhuseduc
------------------------------------------------------------------------------

AG specification test:
J statistic  =  0.1748
Prob > J =  0.8059
------------------------------------------------------------------------------
```

The STATA output for FULL estimation with options `hom robust fuller`:

```
. mivreg hours nwifeinc educ age kidslt6 kidsge6 one (lwage = nwifeinc educ ///
> age kidslt6 kidsge6 one exper expersq fatheduc motheduc hushrs husage ///
> huseduc mtr *X*) if inlf==1 , hom robust fuller

MIVREG estimation (HOM)

First-stage summary
------------------------------
F(  86,   336)  =    2.07
Prob > F        =  0.0000
R-squared       =  0.8471
------------------------------

                                Number of obs   =       428
                                F(   7,   421)  =     96.46
                                Prob > F        =    0.0000
                                R-squared       =   -0.5013
                                Adj R-squared   =   -0.5227
                                Root MSE        =   1.1e+03

FULL estimation
HHN variance estimation

------------------------------------------------------------------------------
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |   1109.999   197.2334     5.63   0.000      722.314    1497.684
    nwifeinc |  -7.856532   5.235509    -1.50   0.134    -18.14753    2.434462
        educ |  -132.0795   31.83561    -4.15   0.000     -194.656   -69.50294
         age |  -9.934026   7.879563    -1.26   0.208    -25.42221    5.554159
     kidslt6 |  -247.4823   143.2961    -1.73   0.085    -529.1472    34.18265
     kidsge6 |   -66.3344   44.59569    -1.49   0.138    -153.9923    21.32355
         one |   2343.827   485.5647     4.83   0.000     1389.394     3298.26
------------------------------------------------------------------------------

Instrumented:  lwage
Instruments:   nwifeinc educ age kidslt6 kidsge6 one exper expersq fatheduc
               motheduc hushrs husage huseduc mtr educXnwifeinc ageXnwifeinc
               ageXeduc kidslt6Xnwifeinc kidslt6Xeduc kidslt6Xage
               kidsge6Xnwifeinc kidsge6Xeduc kidsge6Xage kidsge6Xkidslt6
               experXnwifeinc experXeduc experXage experXkidslt6 experXkidsge6
               expersqXnwifeinc expersqXeduc expersqXage expersqXkidslt6
               expersqXkidsge6 expersqXexper fatheducXnwifeinc fatheducXeduc
               fatheducXage fatheducXkidslt6 fatheducXkidsge6 fatheducXexper
               fatheducXexpersq motheducXnwifeinc motheducXeduc motheducXage
               motheducXkidslt6 motheducXkidsge6 motheducXexper
               motheducXexpersq motheducXfatheduc hushrsXnwifeinc hushrsXeduc
               hushrsXage hushrsXkidslt6 hushrsXkidsge6 hushrsXexper
               hushrsXexpersq hushrsXfatheduc hushrsXmotheduc husageXnwifeinc
               husageXeduc husageXage husageXkidslt6 husageXkidsge6
               husageXexper husageXexpersq husageXfatheduc husageXmotheduc
               husageXhushrs huseducXnwifeinc huseducXeduc huseducXage
               huseducXkidslt6 huseducXkidsge6 huseducXexper huseducXexpersq
               huseducXfatheduc huseducXmotheduc huseducXhushrs huseducXhusage
               mtrXnwifeinc mtrXeduc mtrXage mtrXkidslt6 mtrXkidsge6 mtrXexper
               mtrXexpersq mtrXfatheduc mtrXmotheduc mtrXhushrs mtrXhusage
               mtrXhuseduc
------------------------------------------------------------------------------

LO specification test:
J statistic (bias-corrected) = -0.0382
Prob > J =  0.8752
------------------------------------------------------------------------------
```

The STATA output for HFUL estimation with options `het robust fuller`:

```
. mivreg hours nwifeinc educ age kidslt6 kidsge6 one (lwage = nwifeinc educ ///
> age kidslt6 kidsge6 one exper expersq fatheduc motheduc hushrs husage ///
> huseduc mtr *X*) if inlf==1 , hom robust fuller
MIVREG estimation (HOM)
First-stage summary
------------------------------
F(  86,   336)  =    2.07
Prob > F        =  0.0000
R-squared       =  0.8471
------------------------------
                                Number of obs   =       428
                                F(   7,   421)  =     96.46
                                Prob > F        =    0.0000
                                R-squared       =   -0.5013
                                Adj R-squared   =   -0.5227
                                Root MSE        =   1.1e+03
FULL estimation
HHN variance estimation
```

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lwage	1109.999	197.2334	5.63	0.000	722.314	1497.684
nwifeinc	-7.856532	5.235509	-1.50	0.134	-18.14753	2.434462
educ	-132.0795	31.83561	-4.15	0.000	-194.656	-69.50294
age	-9.934026	7.879563	-1.26	0.208	-25.42221	5.554159
kidslt6	-247.4823	143.2961	-1.73	0.085	-529.1472	34.18265
kidsge6	-66.3344	44.59569	-1.49	0.138	-153.9923	21.32355
one	2343.827	485.5647	4.83	0.000	1389.394	3298.26

```
. mivreg hours nwifeinc educ age kidslt6 kidsge6 one (lwage = nwifeinc educ ///
> age kidslt6 kidsge6 one exper expersq fatheduc motheduc hushrs husage ///
> huseduc mtr *X*) if inlf==1 , het robust fuller
MIVREG estimation (HET)
First-stage summary
------------------------------
F(  86,   336)  =    2.07
Prob > F        =  0.0000
R-squared       =  0.8471
------------------------------
                                Number of obs   =       428
                                F(   7,   421)  =    124.27
                                Prob > F        =    0.0000
                                R-squared       =   -0.4339
                                Adj R-squared   =   -0.4543
                                Root MSE        =   1.1e+03
HFUL estimation
HNWCS variance estimation
```

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lwage	1058.269	170.4895	6.21	0.000	723.1527	1393.386
nwifeinc	-8.041127	4.708921	-1.71	0.088	-17.29705	1.214798
educ	-133.5581	29.08721	-4.59	0.000	-190.7323	-76.38381
age	-10.71399	8.313921	-1.29	0.198	-27.05596	5.627971
kidslt6	-274.0719	166.8757	-1.64	0.101	-602.0853	53.94151
kidsge6	-81.38394	43.17962	-1.88	0.060	-166.2584	3.49055
one	2485.039	466.6137	5.33	0.000	1567.856	3402.222

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. Journal of Econometrics, 2(2):105–110.

Anatolyev, S. (2019). Many instruments and/or regressors: a friendly guide. Journal of Economic Surveys, 33:689–726.

Anatolyev, S. (2021). Mallows criterion for heteroskedastic linear regressions with many regressors. Economics Letters, 203:109864.

Anatolyev, S. and Gospodinov, N. (2011). Specification testing in models with many instruments. Econometric Theory, 27:427–441.

Anatolyev, S. and Mikusheva, A. (2022). Factor models with many assets: strong factors, weak factors, and the two-pass procedure. Journal of Econometrics, 229(1):103–126.

Anatolyev, S. and Yaskov, P. (2017). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. Econometric Theory, 33:717–738.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. Journal of Applied Econometrics, 14(1):57–67.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives, 15:69–85.

Barro, R. J. and Lee, J.-W. (1994). Sources of economic growth. In Carnegie-Rochester Conference Series on Public Policy, volume 40, pages 1–46. Elsevier.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica, 25:77–83.

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

Bekker, P. A. (1994a). Alternative approximations to the distributions of instrumental variable estimators. Econometrica: Journal of the Econometric Society, pages 657–681.

Bekker, P. A. (1994b). Alternative approximations to the distributions of instrumental variable estimators. Econometrica, 62:657–681.

Belloni, A., Chernozhukov, V., and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. arXiv preprint arXiv:1201.0220.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association, 90(430):443–450.

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

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. Biometrics, pages 603–618.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. Annals of Statistics, 34(2):559–583.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics, 34(3):305–334.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R., and Woutersen, T. (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. Journal of Econometrics, 178:15–21.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. The Econometrics Journal, 21(1).

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. The Annals of Statistics, 49(3):1300–1317.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? Journal of the Royal Statistical Society: Series B (Methodological), 57(2):301–337.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360.

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

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. Econometrica, 89(1):181–213.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. Econometrica, 45:939–954.

Ghosh, S. (2011). On the grouped selection and model complexity of the adaptive elastic net. Statistics and Computing, 21:451–462.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. Econometrica, 89(5):2409–2437.

Hansen, B. E. (2007). Least squares model averaging. Econometrica, 75(4):1175–1189.

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

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. Journal of Econometrics, 167(1):38–46.

Hansen, C., Hausman, J. A., and Newey, W. K. (2008a). Estimation with many instrumental variables. Journal of Business & Economic Statistics, 26(4):398–422.

Hansen, C., Hausman, J. A., and Newey, W. K. (2008b). Estimation with many instrumental variables. Journal of Business & Economics Statistics, 26:398–422.

Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. Journal of Econometrics, 182(2):290–308.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. Quantitative Economics, 3:211–255.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when p»n. Statistica Sinica, pages 595–611.

Lee, Y. and Okui, R. (2012). Hahn–hausman test as a specification test. Journal of Econometrics, 167:133–139.

Li, K.-C. (1987). Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics, pages 958–975.

Liu, Q. and Okui, R. (2013). Heteroscedasticity-robust cp model averaging. The Econometrics Journal, 16(3):463–472.

Mallows, C. L. (1973). Some comments on cp. Technometrics, 15(4):661.

Meinshausen, N. and Buhlmann, P. (2004). Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Technical report, ETH Zürich.

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

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica, 55:765–799.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. Econometrica, 58(4):809–37.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. Econometrica, 72(1):219–255.

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

Spiess, J. (2017). Bias reduction in instrumental variable estimation through first-stage shrinkage. arXiv preprint arXiv:1708.06443.

Staiger, D. and Stock, J. (1997). Instrumental variables regression with weak instruments. Econometrica, 65(3):557–586.

Steel, M. F. (2020). Model averaging and its use in economics. Journal of Economic Literature, 58(3):644–719.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Tikhonov, A. (1943). On the stability of inverse problems. In Dokl. Akad. Nauk SSSR, volume 39, pages 195–198.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.

Wan, A. T., Zhang, X., and 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.

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

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):143–161.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.