Dissertation review:
Alena Skolkova: Essays on Model Uncertainty and Model Averaging

Lukáš Lafférs (reviewer, Matej Bel University)

October 2023

The submitted thesis consists of three chapters: two are sole-authored and one was written with one coauthor. The common topic is model averaging and model uncertainty.

Chapter 1 - Model Averaging with Ridge Regularization

The first chapter of the thesis introduces a model averaging estimator based on a ridge regularization. It extends the work of Hansen (2007) that considers Mallows model averaging estimator (MMA) which optimizes a criterion that consists of weighted least squares residuals (which captures the quality of the fit) and of the trace of weighted projection matrices (which controls the complexity of the averaged model). The main idea of the paper is to replace the residuals from the ordinary least squares estimator with the residuals from the ridge regression and also replace the projection matrices with its regularized counterparts. Such extension is referred to as ridge-regularized MMA (RMA). The efficiency gains from such approach are expected in situations when there is a high correlation among regressors. The paper is structured as follows: it first defines the estimator, then provides a two-sample simplified setup where the reduction of a mean squared error is explained, then provides a simulation study documenting its finite-sample performance and concludes with two empirical illustrations.

In the introductory section the value added of the paper and connections to the previous literature are clearly communicated. In the next section, the ridge-regularized estimator, together with its heteroskedasticity robust variant are defined. In order to understand what is the source of improvements of RMA relative to the MMA, the mean squared errors (MSE) of both estimators are derived and compared for a simple case where the averaging is made across two models only. This is a useful and a practical approach. In the expressions for MSE the optimal values for the regularization parameters were plugged in and then it was optimized via the weights. This allows to compare MSE of RMA to the MSE of MMA, at least in an ideal situation where the optimal regularization parameters are known. Given that the analytical formulas are long and complicated, differences in MSEs are shown for different scenarios (e.g. in terms of variance of the noise term or in the number of regressors), where relative difference between MSEs was increasing in the degree of correlation between the two model matrices X_1 and X_2. This comparative statics exercise show an intuitive result that the with higher-dimensions the relative gains tend to increase.

Finite-sample simulations show large improvements in MSE of RMA relative to MMA. Main insight is that it that the ridge regularized variant places much larger weight on models with more parameters (Appendix W.1). While the setup in the theoretical section assumed orthonormal design

(most likely for the sake of simplicity), the simulation section shows, that the gains from using RMA are somewhat larger if regressors in the model matrix X_j are also correlated, but the difference does not appear to be dramatic (Figure 1.7 vs Figure 1.11).

The novel estimators are also studied in terms of prediction performance on empirical data based on two important economic problems - wage equation and growth determinants. While the RMA uniformly dominates all the considered comparison methods in terms of out-of-sample predictive efficiency in the wage equation example, in the growth determinants example it was a weighted BIC (WBIC) method that performed the best.

Here is the list of questions and comments that are mostly in terms of clarification of the presentation.

Comments and suggestions:

- The formulas for the MSEs are interesting on their own as they allow to decompose the improvement (RMA vs MMA) into the bias and variance part. This explicit quantification of the trade-off that leads to an improvement in MSE could be explained in a greater depth.
- It may be interesting for the reader to see the optimal weights in the Figures 1.1 to 1.4 and also the MSEs, not only the differences.
- Consider model 1. The expression for the optimal regularization parameter λ_1^{opt} that is plugged in the MSE formula before the optimization across weights consists of both ρ and β_2. But these quantities cannot be recovered on the basis of information from X_1 alone, that is, in estimation of model 1. It is interesting to see if this bears any relevance for how can the results from the comparative statics (section 1.3) be applicable for empirical implementations.
- Even though the simulation setup has some degree of heteroskedasticity built in, performance of RMA and HR-RMA are very similar. A reader may be interested why this is the case.
- It would be helpful to expand a bit on the discussion of the limitations of RMA and HR-RMA. Especially determine in what situations RMA would be expected to fail or be sub-optimal. For example it is interesting to shed some light on what features of the dataset made the WBIC perform best in the second empirical example.

A few minor comments:

- terms $\hat{e}_i$ and $p_{ii}(w)$ are being referred to, but are not being used in the expression for HRC_p (p.9),
- it may be more precise to refer to the second term (p.8) as to "a complexity of an average model" instead of "average model complexity",
- $\hat{e}_{iR}^2$ is not defined - based on the text, I assume (in the light of footnote 7) that it is the squared residual from the largest model,
- It is unclear what is the source of the statistical uncertainty (p.13), and why there are "hat" symbols used (like $\hat{\lambda}_1$). The way I read this section is that there is no simulation involved here, so there is no source of uncertainty. It may be helpful to clarify.
- in Appendix 1.W I cannot see the RMA (red dashed line) line in the graphs. I assume that it overlaps with the HR-RMA so well that is not visible to the naked eye(?).

Notation/typos:

- P^R operator (p.10) is used without a bold font in contrast to the $\mathbf{P}$ operator (p.8),
- the trace operator is used both with parenthesis and with brackets,
- $\hat{\beta}_1^r(\theta_1)$ is used in the paper, but $\hat{\beta}_2(\theta_1)$ in the appendix,
- the covariance matrix is outside margins of the paper (p.25).

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

The second chapter studies instrumental variable estimation in a situation when many instruments are available. It suggests an improvement over the widely used Lasso-IV estimator (Belloni, Chen, Chernonozhukov and Hansen, 2012), where first and second stages are estimated with lasso (Tibshirani, 1996) and sample splitting (or its variants) are used to control an overfitting bias. The improvement is related to the fact that in many empirical applications, instruments tend to be highly correlated and lasso-path may be unstable. Therefore this paper suggests to use Elastic-net estimation (combination of lasso and ridge regularization, Zou and Hastie, 2005) instead of the lasso so that correlated regressors can be handled. This requires to set additional penalty parameter that controls the relative importance of lasso relative to the ridge penalty.

The simulation study shows a meaningful improvement of the sample-split ENIV estimators relative to the Lasso counterparts. The simulation design varies important features of the data-generating process such as the number and the type of instruments, the degree of sparsity and the strength of the instruments. The paper also includes an empirical example on returns to education where the method is demonstrated and compared to alternatives.

This paper justifies that the proposed sample-split ENIV estimators could be a safer alternative to widely used lasso variants.

Comments and suggestions:

- while the ENIV extension appears attractive given the simulation study, it comes at the price of selecting an additional nuisance parameter that controls the relative importance of the ridge vs lasso penalty. This new degree of freedom involves both a cost in terms of computing time and also risk that the additional parameter would not be chosen well, especially in a real world situations (where we cannot control the data generating process (DGP) as in the simulation study). It would be interesting to be a bit more explicit about the costs and risks, if these are any relevant.
- from a more applied point of view: The choice of the grid for this extra parameter for cross-validation is important - where it should be finer or more coarse. Readers may appreciate if there are any practical data-driven ways guide this choice.
- among the existing methods, ridge jackknife IV estimator performed particularly well. One notable situation when SS-ENIV was able to slightly out-compete RJIVE (Hansen and Kozbur, 2014) was in the situation with mixed instruments strengths - which is empirically relevant scenario. The difference was, however, rather small and this opens up avenues for further investigation.

Minor comments:

- Proof of the Proposition 1 is based on the fact the EN problem can be reformulated as a lasso problem in the sense of Zou and Hastie (2006) as shown in the Appendix. It may improve the clarity of the exposition if all the sufficient conditions of Theorem 4 in Belloni, Chen, Chernonozhukov and Hansen (2012) be stated more formally, at least in the Appendix.

Notation/typos:

- notation appears to be somewhat inconsistent at places: e.g. $\lesssim_P$ (p.39) vs o_P (p.40) vs $\leq_p$ (p.42)
- different metrics are undefined, e.g. $||\cdot||_0$ (p.39), $||\cdot||_1$ (p.41), $||\cdot||_{2,n}$ (p.42),

- "Cholecky" (p.45)
- N appears undefined in the definition of the concentration parameter (p.45)
- "level" (p.45)

Chapter 3 - Many Instruments: Implementation in STATA

The last chapter of the thesis was coauthored with Stanislav Anatolyev and was published in the The Stata Journal in 2019.

This paper introduced a STATA command `mivreg`, that implements the estimators and testing procedures in linear instrumental variable regression models in a situations with many instruments, following the newest literature.

The usefulness of this paper is apparent. It bridges the gap between the theoretical econometric advancements and the empirical practice. The structure of the paper is that it presents the estimators, explains the syntax of `mivreg`, illustrate it on simulations, compares it with a `ivregress` command and the illustrate it with real data.

The paper is polished and I applaud the authors for the careful implementation. I installed the `mivreg` command and tried it on my computer.

I have no comments for this paper.

Overall assesment

The research in this thesis makes some novel, original and distinct contributions in the area of model uncertainty and model averaging. It satisfies both the formal and content requirements for a PhD thesis in economics.

The issues raised in this review are minor relative to the scholarly contributions. These comments do not require any changes of the dissertation. Instead, they could be taken as suggestions for future research or journal submissions.

I am very pleased to recommend the dissertation for a defense.

Lukáš Lafférs