Problem Set # 3

Econ 722

Spring, 2018

Instructions: Answer each of the following. Problems 2 and 3 require numerical calculations. For these questions please append clearly commented source code along with your write-up. All solutions must be submitted electronically on Canvas by 11:59pm on Sunday, April 15th. Late problem sets will not be accepted: it is much better to turn in partial solutions rather than nothing at all. You may discuss these problems with your classmates, but if you work together please list the names of the students with whom you have collaborated at the top of your solutions.

1. Consider a regression model of the form $y_t = \mathbf{x}_t'\beta + \epsilon_t$ where $\mathbf{x}_t$ is $(p \times 1)$ and satisfies $T^{-1}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' = \mathbf{I}_p$ and $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$ where σ^2 is finite. You may treat the regressors as *fixed* rather than random in your derivations.

   (a) What is the MLE for β in this setting? Derive its finite-sample distribution. Is the MLE consistent for β?

   (b) Derive a closed-form expression for the Ridge Regression estimator of β in this setting, expressed in terms of the MLE and the shrinkage parameter λ. Use this result to write out the finite-sample distribution of the Ridge Estimator.

   Important Note: It will be helpful to factor a T from λ in your derivation. When you do this, you can still call the re-scaled shrinkage parameter λ rather than λ/T. Remember: we are free to choose λ so its precise scaling is irrelevant. Writing things this way will make your expression for the Ridge estimator match the (slightly different) example from the lecture notes and should help to avoid confusion in the parts that follow below.

   (c) First, suppose that we choose a fixed positive value for λ. Explain why the corresponding Ridge estimator will *not* be consistent for β as $T \to \infty$ in this case. Next suppose that, instead of fixing λ we decide to allow it to change with sample size.

State sufficient conditions on the sequence λ_T to ensure that the Ridge estimator is consistent.

(d) Derive a closed-form expression for the LASSO estimator in this example, expressed in terms of the MLE and the shrinkage parameter λ.

Important Note: As in the Ridge example above, and for the same reason, if you encounter the term $T\lambda$ you can forget about the T and just call this λ.

(e) For sufficiently large λ, LASSO shrinks some coefficients all the way to zero and hence can be used to carry out variable selection. Derive an exact finite sample expression for the probability that LASSO decides to "exclude" regressor j, that is $P(\widehat{\beta}_j^{Lasso} = 0)$. Explain the intuition with the help of one or more plots.

(f) Now suppose we allow the LASSO shrinkage parameter to depend on sample size. Prove that $\lim_{T\to\infty} \lambda_T = 0$ implies that the probability of LASSO excluding a relevant regressor, i.e. one with a non-zero coefficient, converges to zero.

(g) Now consider the case of an *irrelevant regressor*, i.e. $\beta_j = 0$. What is the probability that LASSO excludes such a regressor? If we allow λ_T to depend on sample size, what condition on the *rate* at which $\lambda_T \to 0$ is required to ensure that LASSO excludes irrelevant regressors with probability approaching one in the limit? What happens if $\lambda_T \to 0$ at a slower rate?

(h) Combining the two preceding parts gives a *consistency* result for LASSO: provided that λ_T converges to zero at a sufficiently fast rate, all relevent regressors are selected and all irrelevant regressors excluded with probability approaching one in the limit. Crucially, however, this result depended on β_j being *fixed*. Suppose instead that we consider a sequence of local parameter values $\beta_{j,T} = \delta/\sqrt{T}$ where δ is a constant. When $\delta \neq 0$, this captures in asymptotic form the idea of "small but nonzero" coefficients. How do the results of the preceding parts change under these asymptotics? Discuss your findings.

2. This question and the one that follows it are based on Stock and Watson (JBES, 2012) "Generalized Shrinkage Methods for Forecasting Using Many Predictors." You can download the paper, a supplemental appendix, data and replication files from Mark Watson's webpage at `https://www.princeton.edu/~mwatson/publi.html`

(a) Read the Stock and Watson (2012) paper. Provide a brief (one or two paragraph) summary of the main findings in the paper.

(b) Familiarize yourself with the data set. The transformed series which Stock and Watson use in their estimation are posted at `http://ditraglia.com/econ722/SW2012data.csv`. Take a look at Section B of the Supplemental Appendix to understand how the raw data (you can find them in the replication zip file on Mark's webpage) are transformed into the data that I have posted. Compare the GDP growth series in the Stock and Watson data set to a GDP growth series that you construct from FRED data.

(c) Compute the first 20 principal components from 108 lower-level disaggregates.[1] How much variation in each of the 108 series is explained by the first or by the first two principal components? Think carefully about how to present this information in an efficient manner.

3. In this question you will construct diffusion index forecasts of real GDP (GDP 251), consumption (GDP252), and real government consumption expenditures (GDP265) following Stock & Watson (2012). The data you will need for this exercise are posted at `http://ditraglia.com/econ722/SW2012data.csv`. For each series and each part of this question you will construct one-step-ahead, pseudo-out-of-sample forecasts based on a rolling window of the most recent 100 observations and use RMSE to compare the different forecasting methods. (The procedure is detailed in section 3.1 of the paper.) Note that you will *not* use cross-validation in this question.

   (a) First try to replicate Stock and Watson's (2012) one-step-ahead RMSE results for the AR(4) model and the OLS model, both relative to the DFM5. These appear in Table S-8 in the online supplemental appendix and are described in section 4.1 of the paper.

   (b) An AR(4) model may be somewhat too complicated to serve as a reasonable benchmark for the DFM5. Augment your results from the preceding part by adding an AR(1) model as well as two model-selection based AR forecasting procedures: one using AIC and another using BIC. How do the results compare? Do they differ across the three series?

   (c) Although Stock & Watson (2012) explore a number of shrinkage estimators in their paper, Ridge and Lasso are not among them. Try using Ridge and Lasso rather

[1] Although the paper gives the total number of disaggregates as 109, there in fact only 108 in the replication dataset. This agrees with the count based on table B.1 of the Supplementary Appendix.

than OLS for the forecasting regression based on the first 50 PCs. Remember: since the design is orthogonal, there is a simple closed form for both Lasso and Ridge. Experiment with a variety of values for the shrinkage parameters. Try to find values that give similar performance to the DFM5. Can you find a level of shrinkage that beats the best AR benchmark forecast? How do your results compare across the three series?