**Final Project in Selected topics in Computational Statistics**

2023-24B / Dr. Amit Moscovich Eiger

Chosen paper: RIDGE REGRESSION: STRUCTURE, CROSS-VALIDATION, AND SKETCHING

Sifan Liu (Stanford University)

Edgar Dobriban (University of Pennsylvania)

Submitted by: Rony Zafrir

* What is the main contribution of the paper?

The paper studies ridge regression thoroughly- points out a few problems that it entails and offers solutions.

The first section, discusses the structure of the estimator obtained by ridge regression. They presented a precise representation of the estimator- one that takes into consideration the big data setting that we're in (where sample size and number of features grows to infinity, yet the ratio converges to a constant).

The second one revolves around CV- since it is the go-to method for choosing the regularization parameter. CV causes bias for estimating the error rate, due to using a smaller number of observations than the total number of observations that we have. Dobriban and Liu proposed a simple-to-calculate bias correction parameter, and I found it surprising and fascinating that it's not dependent on any unknown parameter!

The third and final part they talk about how to actually compute it in practice, in the same big data setting. They showed that using sketching or random projection, in order to approximate the covariance matrix, produces quite accurate results and can in fact reduce computation costs.

* Which parts did you try to implement/reproduce?

I've chosen to focus on the cross-validation section- since we've learned about it this semester in Statistical Computation and I thought that it’s a great opportunity to implement the material, and also because I found their findings regrading this subject to be the most surprising and interesting. I reproduced Figure 2: Cross-validation on the Million Song Dataset. The dataset contains prediction of the release year of a song; based on its audio features. Songs are mostly western, commercial tracks released in 1922 until 2011.

* Did you encounter any difficulties?

Personally, the main difficulty was dealing with a scientific paper for the first time, which resulted in skimming through many papers that I didn’t think were fitting, yet it was resolved after finding one that I found interesting and doable. After reading it for a few times before moving on to recreating a graph, I grew more comfortable with the methods and theory that were presented in the paper.

* How fast was it to run the method? Did you take any steps to make the computation faster?

My program took 25-35 seconds to run. I did my best to write efficient code but didn't take further actions to make it run faster.

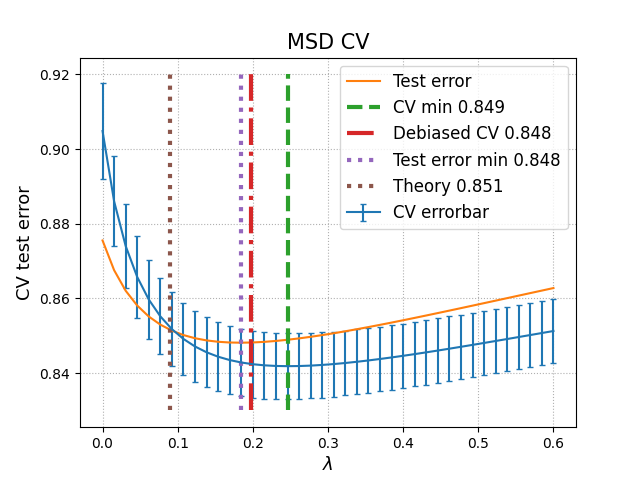
* Were there any numerical issues? If so, did you do anything to handle them? Did you test the numerical accuracy?

I didn’t encounter numerical issues. While coding I kept in mind some of the ideas and tips that were presented in class, like using library functions when possible, and doing the minimal needed number of arithmetic operations. When calculating the CV error for the error-bars; in order to minimize possible error accumulation – I implemented summation from the smallest to largest number (since the first ones are multiplied by the larger error terms), which was discussed in class as well.

* What did you learn from this experience?

I've learned about many concepts that I didn’t encounter during my bachelor's degree, such as EDS, sketching, etc. Moreover, it allowed me to think more deeply on concepts that I've learned previously (CV and ridge regression), as well as implementing cross-validation. In addition, I learned to be aware of numerical and computational aspects when coding which is going to be helpful for me from now on.

Graph – figure 2 – MSD CV



References:

Sifan Liu and Edgar Dobriban. RIDGE REGRESSION: STRUCTURE, CROSSVALIDATION, AND SKETCHING. a conference paper at ICLR 2020. [https://arxiv.org/pdf/1910.02373v3](https://arxiv.org/pdf/1910.02373v3%20)

T. Bertin-Mahieux. Dataset Year Prediction MSD. UC Irvine Machine Learning Repository 2011. <https://archive.ics.uci.edu/dataset/203/yearpredictionmsd>

Note- At the time you're reading it I am probably recovering for a surgery (I'm fine don't worry) so I might not be reachable… In case that there's an unexpected difficulty with downloading the dataset or running the program:

* I couldn't upload the ZIP folder to the Moodle submission box, so I sent it to you via email, and uploaded it to a drive folder:

<https://drive.google.com/drive/folders/1oOMOF-Zqk_3kDhuA7lSlELnuM4PkgNcF?usp=sharing>

* I saved the msd.py as an executable file so when you double-click on the file it should produce the plot in ~30 seconds.

(msd.py should be in the same folder as the dataset file under the name YearPredictionMSD.txt to run)

I also created a git repository for the project- includes the python code and the output plot:

<https://github.com/RonyZafrir/RidgeRegression-MSD>

Thanks in advance! I've learned a lot from you in the past semester 😊