ORF 525: Statistical Foundations of Data Science

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Problem Set #2 Spring 2021

Choose any 5 problems

Due Friday, February 26, 2021.

1. Consider the Lasso problem $\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$, where $\lambda>0$ is a tuning parameter.

   (a) If $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ are both minimizers of the Lasso problem, show that they have the same prediction, i.e., $\mathbf{X}\widehat{\boldsymbol{\beta}}_1=\mathbf{X}\widehat{\boldsymbol{\beta}}_2$. **Hint:** Consider the vector $\alpha\widehat{\boldsymbol{\beta}}_1+(1-\alpha)\widehat{\boldsymbol{\beta}}_2$ for $\alpha\in(0,1)$.

   (b) Let $\widehat{\boldsymbol{\beta}}$ be a minimizer of the Lasso problem with j^{th} componenet $\widehat{\beta}_j$. Denote $\mathbf{X}_j$ to be the j-th column of $\mathbf{X}$. Show that

$$\begin{cases} \lambda = n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) & \text{if } \widehat{\beta}_j>0; \\ \lambda = -n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) & \text{if } \widehat{\beta}_j<0; \\ \lambda \geq |n^{-1}\mathbf{X}_j^T(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})| & \text{if } \widehat{\beta}_j=0. \end{cases}$$

   (c) If $\lambda>\|n^{-1}\mathbf{X}^T\mathbf{Y}\|_\infty$, prove that $\widehat{\boldsymbol{\beta}}_\lambda=\mathbf{0}$, where $\widehat{\boldsymbol{\beta}}_\lambda$ is the minimizer of the Lasso problem with regularization parameter λ.

2. Consider the elastic-net loss $p(\theta)=\lambda_1|\theta|+\lambda_2\theta^2$ with $\lambda_2>0$. Let $\widehat{\boldsymbol{\beta}}$ be the minimizer of $\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^p p(|\beta_j|)$.

   (a) Show that $\widehat{\boldsymbol{\beta}}$ is unique.

   (b) Give the necessary and sufficient conditions for $\widehat{\boldsymbol{\beta}}$ being the penalized least-squares solution.

   (c) If $\lambda_1>\|n^{-1}\mathbf{X}^T\mathbf{Y}\|_\infty$, show that $\widehat{\boldsymbol{\beta}}=\mathbf{0}$

3. Concentration inequalities.

   (a) The random vector $\boldsymbol{\varepsilon}\in\mathbb{R}^n$ is called σ-sub-Gaussian if $E\exp\left(\mathbf{a}^T\boldsymbol{\varepsilon}\right)\leq\exp\left(\|\mathbf{a}\|_2^2\sigma^2/2\right)$, $\forall\mathbf{a}\in\mathbb{R}^n$. Show that $E\boldsymbol{\varepsilon}=\mathbf{0}$ and $\text{var}(\boldsymbol{\varepsilon})\leq\sigma^2\mathbf{I}_n$. **Hint:** Expand exponential functions as infinite series (actually, you only need the condition for $\mathbf{a}$ in a small neighborhood around 0)

   (b) For $\mathbf{X}\in\mathbb{R}^{n\times p}$ with the j-th column denoted by $\mathbf{X}_j\in\mathbb{R}^n$, suppose that $\|\mathbf{X}_j\|_2^2=n$ for all j, and $\boldsymbol{\varepsilon}\in\mathbb{R}^n$ is a σ-sub-Gaussian random vector. Show that there exists a constant $C>0$ such that

$$P\Big(\|n^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\|_\infty>\sqrt{2(1+\delta)}\sigma\sqrt{\frac{\log p}{n}}\Big)\leq Cp^{-\delta},\qquad\forall\delta>0.$$

4. This problem intends to show that the gradient decent method for a convex function $f(\cdot)$ is a member of majorization-minimization algorithms and has a sublinear rate of convergence in terms of function values. From now on, the function $f(\cdot)$ is convex and let $\mathbf{x}^*\in\text{argmin}f(\mathbf{x})$. Here we implicitly assume the minimum can be attained at some point $\mathbf{x}^\star\in\mathbb{R}^p$.

(a) Suppose that $f''(\mathbf{x}) \leq L\mathbf{I}_p$ and $\delta \leq 1/L$. Show that the quadratic function $g(\mathbf{x}) = f(\mathbf{x}_{i-1}) + f'(\mathbf{x}_{i-1})^T(\mathbf{x} - \mathbf{x}_{i-1}) + \frac{1}{2\delta}\|\mathbf{x} - \mathbf{x}_{i-1}\|^2$ is a majorization of $f(\mathbf{x})$ at point $\mathbf{x}_{i-1}$, i.e., $g(\mathbf{x}) \geq f(\mathbf{x})$ for all $\mathbf{x}$ and also $g(\mathbf{x}_{i-1}) = f(\mathbf{x}_{i-1})$.

(b) Show that gradient step $\mathbf{x}_i = \mathbf{x}_{i-1} - \delta f'(\mathbf{x}_{i-1})$ is the minimizer of the majorized quadratic function $g(\mathbf{x})$ and hence the gradient descend method can be regarded as a member of MM-algorithms.

(c) Use (a) and the convexity of $f(\cdot)$ to show that

$$f(\mathbf{x}_i) \leq f(\mathbf{x}^*) + \frac{1}{2\delta}(\|\mathbf{x}_{i-1} - \mathbf{x}^*\|^2 - \|\mathbf{x}^* - \mathbf{x}_i\|^2).$$

(d) Conclude using (c) that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2/(2k\delta)$, namely gradient descent converges at a sublinear rate. (**Note**: The gradient descent method converges linearly if $f(\cdot)$ is strongly convex.)

5. Let us consider the 128 macroeconomic time series from Jan. 1959 to Dec. 2018, which can be downloaded from the course website (see the "transformed macroeconomic data" at the bottom of the page `https://orfe.princeton.edu/\%7Ejqfan/fan/classes/525.html`). In this problem, we will explore what macroeconomic variables are associated with the unemployment rate contemporarily and which macroeconomic variables lead the unemployment rates.

(a) Extract the data from Jan. 1960 to Oct. 2018 (in total 706 months) and remove the feature named "sasdate". Then, remove the features with missing entries and report their names.

(b) The column with name "UNRATE" measures the difference in unemployment rate between the current month and the previous month. Take this column as the response and take the remaining variables as predictors. To conduct contemporary association studies, do the following steps for Lasso (using R package `glmnet`) and SCAD (using R package `ncvreg`): Set a random seed by `set.seed(525)`; Plot the regularization paths as well as the mean squared errors estimated by 10-fold cross-validation; Choose a model based on cross-validation, report its in-sample R^2, and point out two most important macroeconomic variables that are associated with the change of unemployment rate in terms of largest regression coefficients for the standardized variables.

(c) In this sub-problem, we are going to study which macroeconomic variables are leading indicators for the changes of future unemployment rate. To do so, we will pair each row of predictors with the next row of response. The last row of predictors and the first element in the response are hence discarded. After this change, do the same exercise as (b).

6. Let us consider the Zillow data again. We drop the first 3 columns ("(empty)", "id", "date") and treat "zipcode" as a factor variable. Now, consider the variables

(a) "bedrooms", "bathrooms", "sqft_living", and "sqft_lot" and their interactions and the remaining 14 variables in the data, including "zipcode". (We can use *model.matrix* to expand factors into a set of dummy variables.)

(b) Add the following additional variables to (b): $X_{12} = I(view == 0)$, $X_{13} = L^2$, $X_{13+i} = (L - \tau_i)_+^2$, $i = 1, \cdots, 9$, where τ_i is $10 * i^{th}$ percentile and L is the size of living area ("sqft_living").

Compute and compare out-of-sample R^2 using ridge regression, Lasso, SCAD with regularization parameter chosen by 10 fold cross-validation.