ORF 525: Statistical Foundations of Data Science

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

Problem Set #5 Spring 2021

Due Friday, April 30, 2021.

Choose any of the 5 problems

1. Suppose $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix of a random vector $\mathbf{X} = (X_1, X_2, \cdots, X_p)^T$.

   (a) Prove the second part of Theorem 9.1: If $\max_{i \le p} \sigma_{ii} \le C_2$, uniformly over $\mathcal{C}_q(m_p)$

   $$p^{-1} \|\widehat{\boldsymbol{\Sigma}}_\lambda^\tau - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 = O_{\mathbb{P}}\left(m_p \Big(\frac{\log p}{n} \Big)^{1-q/2} \right).$$

   You may use the notation and results proved in the first part of Theorem 9.1.

   (b) Assume that $E(X_i) = 0$ for all i, and $\sigma = \max_{i,j} \mathrm{SD}(X_i X_j)$ is bounded. Consider the elementwise adaptive Huber estimator $\widehat{\boldsymbol{\Sigma}}_H = (\widehat{\sigma}_{ij}^\tau)$ where $\widehat{\sigma}_{ij}^\tau$ is the adaptive Huber estimator of the mean $E(X_i X_j) = \sigma_{ij}$ based on the data $\{X_{ki} X_{kj}\}_{k=1}^n$ with parameter τ. Show that when $\tau = \sqrt{n/(a \log p)}\sigma$, we have

   $$P\Big(\|\widehat{\boldsymbol{\Sigma}}_H - \boldsymbol{\Sigma}\|_{\max} \ge \sqrt{\frac{a\sigma^2 \log p}{n}} \Big) \le 4p^{2-a/16},$$

   for any constant $a > 0$. **Hint**: Use Theorem 3.2(d) of the book.

2. This exercise intends to give a least-squares interpretation of the "conditional linear expectation" of $\mathbf{X}_1$ given $\mathbf{X}_2$.

   (a) Without the normality assumption, find the best linear prediction of $\mathbf{X}_1$ by $\mathbf{a} + \mathbf{B}\mathbf{X}_2$. Namely, find $\mathbf{a}$ and $\mathbf{B}$ to minimize $E\|\mathbf{X}_1 - \mathbf{a} - \mathbf{B}\mathbf{X}_2\|^2$. Let $\mathbf{a}^*$ and $\mathbf{B}^*$ be the solution. Show that

   $$\mathbf{B}^* = \mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2)\,\mathrm{var}(\mathbf{X}_2)^{-1}, \qquad \mathbf{a}^* = E\mathbf{X}_1 - \mathbf{B}^* E\mathbf{X}_2,$$

   where $\mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2) = E(\mathbf{X}_1 - E\mathbf{X}_1)(\mathbf{X}_2 - E\mathbf{X}_2)^T$ and $\mathrm{var}(\mathbf{X}_2) = E(\mathbf{X}_2 - E\mathbf{X}_2)(\mathbf{X}_2 - E\mathbf{X}_2)^T$.

   (b) Let $\mathbf{U} = \mathbf{X}_1 - \mathbf{a}^* - \mathbf{B}^*\mathbf{X}_2$ be the residual. Show that $E\mathbf{U} = \mathbf{0}$ and the covariance between $\mathbf{X}_2$ and $\mathbf{U}$ is zero:

   $$E(\mathbf{X}_2 \mathbf{U}^T) = \mathbf{0}.$$

   (c) Show that

   $$\mathrm{var}(\mathbf{U}) = E\mathbf{U}\mathbf{U}^T = \mathrm{var}(\mathbf{X}_1) - \mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2)\,\mathrm{var}(\mathbf{X}_2)^{-1}\,\mathrm{cov}(\mathbf{X}_2, \mathbf{X}_1).$$

3. Consider the one-factor model:

   $$\mathbf{X}_i = \mathbf{b} f_i + \mathbf{u}_i, \qquad i = 1, 2, \cdots, n,$$

   where $\mathbf{u}_i$ and f_i are zero-mean and uncorrelated, $\mathrm{var}(f_i) = 1$ and $\mathrm{var}(\mathbf{u}_i) = \mathbf{I}_p$. Show that

   (a) The largest eigenvalue of $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X}_i)$ is $(1 + \|\mathbf{b}\|^2)$ with the associated eigenvector $\mathbf{b}/\|\mathbf{b}\|$.

(b) Let $\widehat{f}_i = \mathbf{b}^T\mathbf{X}_i/\|\mathbf{b}\|^2$. If $\mathbf{u}_i \sim N(0, \mathbf{I}_p)$, show that

$$\max_{i\le n}(\widehat{f}_i - f_i)^2 = O_P(\log n\|\mathbf{b}\|^{-2}).$$

4. Let $\mathbf{Z} = (\mathbf{X}^T, \mathbf{f}^T)^T$, where $\mathbf{X}$ follows the factor model $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{f} + \mathbf{u}$.

(a) Show that the covariance matrix of $\mathbf{Z}$ is given by

$$\boldsymbol{\Sigma}_z = \begin{pmatrix} \mathbf{B}\boldsymbol{\Sigma}_f\mathbf{B}^T + \boldsymbol{\Sigma}_u & \mathbf{B}\boldsymbol{\Sigma}_f \\ \boldsymbol{\Sigma}_f\mathbf{B}^T & \boldsymbol{\Sigma}_f \end{pmatrix} := \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

(b) Use (a) to prove that the covariance of the idiosyncratic component is given by

$$\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

(c) If $\widehat{\boldsymbol{\Sigma}}_z$ is obtained by the sample covariance matrix of the data $\mathbf{Z}_i = (\mathbf{X}_i^T, \mathbf{f}_i)^T$, show that $\widehat{\boldsymbol{\Sigma}}_u = \widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{21}$ is the same as the sample covariance matrix of $\{\widehat{\mathbf{u}}_i\}_{i=1}^n$, where $\widehat{\mathbf{u}}_i$ is the residual vector based on the least-squares fit. Note that you can use the fact that the least-squares estimate for $\mathbf{B}$ is just the empirical substitution of $\mathbf{B}^*$ in problem 2(a).

5. Download the mice protein expression data from UCI Machine Learning Repository:

`https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression.`

There are in total 1080 measurements. Use attributes 2–78 (the expression levels of 77 proteins), as the input variables for large covariance estimation. We clean the missing values by dropping 6 columns "BAD_N", "BCL2_N", "pCFOS_N", "H3AcK18_N", "EGR1_N", "H3MeK4_N" and then removing all rows with at least 1 missing value. For your convenience, you can also download directly the clean input data from our course website (see the "mice protein expression data (preprocessed)" at the page `https://fan.princeton.edu/fan/classes/525.html`). You should have an input data of size 1047 × 71.

(a) Compute the correlation-thresholded covariance estimate with $\lambda = 4$. Report the entries of zeros and the minimum and maximum eigenvalues of the resulting estimate. Compare them with those from the input covariance matrix.

(**Hint**: the entry-dependece thresholds for the covariance matrix are $\lambda_{ij} = \lambda\sqrt{\widehat{\sigma}_{ii}, \widehat{\sigma}_{jj}\frac{\log p}{n}}$.)

(b) Estimate the precision matrix by CLIME. (Hint: you can use the package "`flare`".)

(c) Construct the protein network by using the estimated sparsity pattern of the precision matrix. (Hint: you can use the package "`graph`" or "`qgraph`")

(d) Use the winsorized data to repeat part (a), where the winsorized parameter is taken to be 2.5 standard deviation.

6. 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://fan.princeton.edu/fan/classes/525.html`). As before, we extract the data from Jan. 1960 to Oct. 2018 (in total 706 months) and remove the feature named "sasdate" and the features with missing entries. We use $\mathbf{M}$ to denote the preprocessed data matrix.

   (a) Extract the first 120 months (i.e. the first 120 rows of $\mathbf{M}$), and standardize the data such that all the variables have zero mean and unit variance.

      i. For the standardized data, draw a scree plot of the 20 leading eigenvalues of the sample covariance matrix. Use the eigen-ratio method with $k_{\max} = 10$ to determine the number of factors K. Compare the result with that using the eigenvalue thresholding $\{j : \lambda_j(\mathbf{R}) > 1 + p/n\}$. Here, we do not do eigenvalue adjustments to facilitate the implementations for both methods.

      ii. Use the package `POET` to estimate the covariance matrix. Report the maximum and minimum eigenvalues of the estimated matrix. Print the sub-covariance matrix of the first 3 variables.

   (b) The column "UNRATE" of $\mathbf{M}$ corresponds to monthly changes of the unemployment rate. We will predict the future "UNRATE" using current values of the other macroeconomic variables.

      i. Pair each month's "UNRATE" with the other macroeconomic variables in the previous month (to do so, you need to drop the "UNRATE" in the first month and other variables in the last month). Let $\{(\mathbf{x}_t, y_t)\}_{t=1}^N$ denote the derived pairs of covariates and responses.

      ii. We next train the model once a year using the past 120 months' data and use the model to forecast the next 12 months' unemployment rates. More precisely, for each month $t \in \{121 + 12m : \ m = 0, 1, \cdots 40\}$, we want to forecast the next 12 monthly "UNRATE"s $\{y_{t+i}\}_{i=0}^{11}$ based on $\{\mathbf{x}_{t+i}\}_{i=0}^{11}$, using the past 120 months' data $\{(\mathbf{x}_{t-i}, y_{t-i})\}_{i=1}^{120}$ for training. Implement the FarmSelect method by following the steps: standardize the covariates; fit a factor model; fit lasso on augmented covariates; output the predicted values $\{y_{t+i}^{\text{farm}}\}_{i=0}^{11}$ for $\{y_{t+i}\}_{i=0}^{11}$. Also, compute the baseline predictions $\bar{y}_{t+i} = \frac{1}{120}\sum_{j=1}^{120} y_{t-j}$ for all $i = 0, 1, \cdots, 11$.

      iii. Report the out-of-sample R^2 using

$$R^2 = 1 - \frac{\sum_{t=121}^{612}(y_t - y_t^{\text{farm}})^2}{\sum_{t=121}^{612}(y_t - \bar{y}_t)^2}.$$

   Hint: Try to use R package `FarmSelect`.

   (c) Run principal component regression (PCR) with 5 components for the same task as Part (b) and report the out-of-sample R^2.

   (d) Treat $\{\mathbf{x}_i\}_{i=1}^{60}$ and $\{\mathbf{x}_{60+i}\}_{i=1}^{120}$ as i.i.d. samples from two distributions and let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ denote their expectations. Use the R package `FarmTest` to test

$$H_0 : \ \mu_{1j} = \mu_{2j}, \ \ \forall j \qquad \text{v.s.} \qquad H_1 : \ \mu_{1j} \neq \mu_{2j} \text{ for some } j.$$

Print the indices of the rejected hypotheses.