A Mathematical Introduction to Data Science **Sep. 22, 2016**

Homework 3. High Dimensional Statistics Models

Instructor: Yuan Yao *Due: Open Date*

The problem below marked by * is optional with bonus credits.

1. *Maximum Likelihood Method*: consider n random samples from a multivariate normal distribution, $X_i \in \mathbb{R}^p \sim \mathcal{N}(\mu, \Sigma)$ with $i = 1, \ldots, n$.

   (a) Show the log-likelihood function

   $$l_n(\mu, \Sigma) = -\frac{n}{2}\text{trace}(\Sigma^{-1}S_n) - \frac{n}{2}\log\det(\Sigma) + C,$$

   where $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T$, and some constant C does not depend on μ and Σ;

   (b) Show that $f(X) = \text{trace}(AX^{-1})$ with $A, X \succeq 0$ has a first-order approximation,

   $$f(X + \Delta) \approx f(X) - \text{trace}(X^{-1}A'X^{-1}\Delta)$$

   hence formally $df(X)/dX = -X^{-1}AX^{-1}$ (note $(I + X)^{-1} \approx I - X$. A typo in previous version missed '-' sign here.);

   (c) Show that $g(X) = \log\det(X)$ with $A, X \succeq 0$ has a first-order approximation,

   $$g(X + \Delta) \approx g(X) + \text{trace}(X^{-1}\Delta)$$

   hence $dg(X)/dX = X^{-1}$ (note: consider eigenvalues of $X^{-1/2}\Delta X^{-1/2}$);

   (d) Use these formal derivatives with respect to positive semi-definite matrix variables to show that the maximum likelihood estimator of Σ is

   $$\hat{\Sigma}_n^{MLE} = S_n.$$

   A reference for (b) and (c) can be found in Convex Optimization, by Boyd and Vandenbergh, examples in Appendix A.4.1 and A.4.3:

   `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`

2. *Shrinkage:* Suppose $y \sim \mathcal{N}(\mu, I_p)$.

   (a) Consider the Ridge regression

   $$\min_\mu \frac{1}{2}\|y - \mu\|_2^2 + \frac{\lambda}{2}\|\mu\|_2^2.$$

Show that the solution is given by

$$\hat{\mu}_i^{ridge} = \frac{1}{1+\lambda} y_i.$$

Compute the risk (mean square error) of this estimator. The risk of MLE is given when $C = I$.

(b) Consider the LASSO problem,

$$\min_{\mu} \frac{1}{2}\|y-\mu\|_2^2 + \lambda\|\mu\|_1.$$

Show that the solution is given by Soft-Thresholding

$$\hat{\mu}_i^{soft} = \mu_{soft}(y_i;\lambda) := \text{sign}(y_i)(|y_i|-\lambda)_+.$$

For the choice $\lambda = \sqrt{2\log p}$, show that the risk is bounded by

$$\mathbb{E}\|\hat{\mu}^{soft}(y)-\mu\|^2 \le 1 + (2\log p + 1)\sum_{i=1}^{p} \min(\mu_i^2, 1).$$

Under what conditions on μ, such a risk is smaller than that of MLE? Note: see Gaussian Estimation by Iain Johnstone, Lemma 2.9 and the reasoning before it.

(c) Consider the l_0 regularization

$$\min_{\mu} \|y-\mu\|_2^2 + \lambda^2\|\mu\|_0,$$

where $\|\mu\|_0 := \sum_{i=1}^{p} I(\mu_i \neq 0)$. Show that the solution is given by Hard-Thresholding

$$\hat{\mu}_i^{hard} = \mu_{hard}(y_i;\lambda) := y_i I(|y_i| > \lambda).$$

Rewriting $\hat{\mu}^{hard}(y) = (1-g(y))y$, is $g(y)$ weakly differentiable? Why?

(d) Consider the James-Stein Estimator

$$\hat{\mu}^{JS}(y) = \left(1 - \frac{\alpha}{\|y\|^2}\right) y.$$

Show that the risk is

$$\mathbb{E}\|\hat{\mu}^{JS}(y)-\mu\|^2 = \mathbb{E}U_\alpha(y)$$

where $U_\alpha(y) = p - (2\alpha(p-2) - \alpha^2)/\|y\|^2$. Find the optimal $\alpha^* = \arg\min_\alpha U_\alpha(y)$. Show that for $p > 2$, the risk of James-Stein Estimator is smaller than that of MLE for all $\mu \in \mathbb{R}^p$.

(e) In general, an odd monotone unbounded function $\Theta : \mathbb{R} \to \mathbb{R}$ defined by $\Theta_\lambda(t)$ with parameter $\lambda \ge 0$ is called *shrinkage* rule, if it satisfies

[shrinkage] $0 \le \Theta_\lambda(|t|) \le |t|$;
[odd] $\Theta_\lambda(-t) = -\Theta_\lambda(t)$;
[monotone] $\Theta_\lambda(t) \le \Theta_\lambda(t')$ for $t \le t'$;
[unbounded] $\lim_{t\to\infty} \Theta_\lambda(t) = \infty$.

Which rules above are shrinkage rules?

3. **Necessary Condition for Admissibility of Linear Estimators.* Consider linear estimator for $y \sim \mathcal{N}(\mu, \sigma^2 I_p)$

$$\hat{\mu}_C(y) = Cy.$$

Show that $\hat{\mu}_C$ is admissible only if

(a) C is symmetric;

(b) $0 \leq \rho_i(C) \leq 1$ (where $\rho_i(C)$ are eigenvalues of C);

(c) $\rho_i(C) = 1$ for at most two i.

These conditions are satisfied for MLE estimator when $p = 1$ and $p = 2$.

Reference: Theorem 2.3 in Gaussian Estimation by Iain Johnstone,
`http://statweb.stanford.edu/~imj/Book100611.pdf`

4. *James Stein Estimator for $p = 1$:*

From Theorem 3.1 in the lecture notes, we know that MLE $\hat{\mu} = Y$ is admissible when $p = 1$ or 2. However if we use SURE to calculate the risk of James Stein Estimator,

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) = \mathbb{E}U(Y) = p - \mathbb{E}_\mu \frac{(p-2)^2}{\|Y\|^2} < p = R(\hat{\mu}^{\mathrm{MLE}}, \mu)$$

it seems that for $p = 1$ James Stein Estimator should still has lower risk than MLE for any μ. Explain what violates the above calculation for $p = 1$.

5. *Phase transition in PCA "spike" model:* Consider a finite sample of n i.i.d vectors $x_1, x_2, \ldots, x_n$ drawn from the p-dimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 I_{p\times p} + \lambda_0 u u^T)$, where λ_0/σ^2 is the signal-to-noise ratio (SNR) and $u \in \mathbb{R}^p$. In class we showed that the largest eigenvalue λ of the sample covariance matrix S_n

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

pops outside the support of the Marcenko-Pastur distribution if

$$\frac{\lambda_0}{\sigma^2} > \sqrt{\gamma},$$

or equivalently, if

$$\mathrm{SNR} > \sqrt{\frac{p}{n}}.$$

(Notice that $\sqrt{\gamma} < (1 + \sqrt{\gamma})^2$, that is, λ_0 can be "buried" well inside the support Marcenko-Pastur distribution and still the largest eigenvalue pops outside its support). All the following questions refer to the limit $n \to \infty$ and to almost surely values:

(a) Find λ given $\mathrm{SNR} > \sqrt{\gamma}$.

(b) Use your previous answer to explain how the SNR can be estimated from the eigenvalues of the sample covariance matrix.

(c) Find the squared correlation between the eigenvector v of the sample covariance matrix (corresponding to the largest eigenvalue λ) and the "true" signal component u, as a function of the SNR, p and n. That is, find $|\langle u, v\rangle|^2$.

(d) Confirm your result using MATLAB or R simulations (e.g. set $u = e$; and choose $\sigma = 1$ and λ_0 in different levels. Compute the largest eigenvalue and its associated eigenvector, with a comparison to the true ones.)

6. *Exploring S&P500 Stock Prices:* Take the Standard & Poor's 500 data:
`http://math.stanford.edu/~yuany/course/data/snp452-data.mat`,
which contains the data matrix $X \in \mathbb{R}^{n\times p}$ of $n = 1258$ consecutive observation days and $p = 452$ daily closing stock prices, and the cell variable "stock" collects the names, codes, and the affiliated industrial sectors of the 452 stocks. Use Matlab or R for the following exploration.

(a) Take the logarithmic prices $Y = \log X$;

(b) For each observation time $t \in \{1, \ldots, 1257\}$, calculate logarithmic price jumps

$$\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}, \quad i \in \{1, \ldots, 452\};$$

(c) Construct the realized covariance matrix $\hat{\Sigma} \in \mathbb{R}^{452\times 452}$ by,

$$\hat{\Sigma}_{i,j} = \frac{1}{1257} \sum_{\tau=1}^{1257} \Delta Y_{i,\tau} \Delta Y_{j,\tau};$$

(d) Compute the eigenvalues (and eigenvectors) of $\hat{\Sigma}$ and store them in a descending order by $\{\hat{\lambda}_k, k = 1, \ldots, p\}$.

(e) *Horn's Parallel Analysis*: the following procedure describes a so-called Parallel Analysis of PCA using random permutations on data. Given the matrix $[\Delta Y_{i,t}]$, apply random permutations $\pi_i : \{1, \ldots, t\} \to \{1, \ldots, t\}$ on each of its rows: $\Delta \tilde{Y}_{i,\pi_i(j)}$ such that

$$[\Delta \tilde{Y}_{\pi(i),t}] = \begin{bmatrix} \Delta Y_{1,1} & \Delta Y_{1,2} & \Delta Y_{1,3} & \ldots & \Delta Y_{1,t} \\ \Delta Y_{2,\pi_2(1)} & \Delta Y_{2,\pi_2(2)} & \Delta Y_{2,\pi_2(3)} & \ldots & \Delta Y_{2,\pi_2(t)} \\ \Delta Y_{3,\pi_3(1)} & \Delta Y_{3,\pi_3(2)} & \Delta Y_{3,\pi_3(3)} & \ldots & \Delta Y_{3,\pi_3(t)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \Delta Y_{n,\pi_n(1)} & \Delta Y_{n,\pi_n(2)} & \Delta Y_{n,\pi_n(3)} & \ldots & \Delta Y_{n,\pi_n(t)} \end{bmatrix}.$$

Define $\tilde{\Sigma} = \frac{1}{t} \Delta \tilde{Y} \cdot \Delta \tilde{Y}^T$ as the null covariance matrix. Repeat this for R times and compute the eigenvalues of $\tilde{\Sigma}_r$ for each $1 \le r \le R$. Evaluate the p-value for each estimated eigenvalue $\hat{\lambda}_k$ by $(N_k+1)/(R+1)$ where N_k is the counts that $\hat{\lambda}_k$ is less than the k-th largest eigenvalue of $\tilde{\Sigma}_r$ over $1 \le r \le R$. Eigenvalues with small p-values indicate that they are less likely arising from the spectrum of a randomly permuted matrix and thus considered to be signal. Draw your own conclusion with your observations and analysis on this data. A reference is: Buja and Eyuboglu, "Remarks on Parallel Analysis", Multivariate Behavioral Research, 27(4): 509-540, 1992.

7. **Finite rank perturbations of random symmetric matrices:* Wigner's semi-circle law (proved by Eugene Wigner in 1951) concerns the limiting distribution of the eigenvalues of random symmetric matrices. It states, for example, that the limiting eigenvalue distribution of $n \times n$ symmetric matrices whose entries w_{ij} on and above the diagonal ($i \leq j$) are i.i.d Gaussians $\mathcal{N}(0, \frac{1}{4n})$ (and the entries below the diagonal are determined by symmetrization, i.e., $w_{ji} = w_{ij}$) is the semi-circle:

$$p(t) = \frac{2}{\pi}\sqrt{1-t^2}, \quad -1 \leq t \leq 1,$$

where the distribution is supported in the interval $[-1, 1]$.

   (a) Confirm Wigner's semi-circle law using MATLAB or R simulations (take, e.g., $n = 400$).

   (b) Find the largest eigenvalue of a rank-1 perturbation of a Wigner matrix. That is, find the largest eigenvalue of the matrix

$$W + \lambda_0 uu^T,$$

   where W is an $n \times n$ random symmetric matrix as above, and u is some deterministic unit-norm vector. Determine the value of λ_0 for which a phase transition occurs. What is the correlation between the top eigenvector of $W + \lambda_0 uu^T$ and the vector u as a function of λ_0? Use techniques similar to the ones we used in class for analyzing finite rank perturbations of sample covariance matrices.

[Some Hints about homework] For Wigner Matrix $W = [w_{ij}]_{n\times n}, w_{ij} = w_{ji}, w_{ij} \sim N(0, \frac{\sigma}{\sqrt{n}})$, the answer is

$$\begin{aligned} \text{eigenvalue is} \quad & \lambda = R + \tfrac{1}{R} \\ \text{eigenvector satisfies} \quad & (u^T\hat{v})^2 = 1 - \tfrac{1}{R^2} \end{aligned}$$