ES 201 Homework 1, Due 02/09/2017 (in class)

Reading: Secs. 2.1, 2,2, 2.3; Ch. 3, Secs. 6.1, 6.2, 6.5

Regression and optimization

Problem 1

(Variation on Problems 6.1 and 6.3 in the book). Let X be full-rank $n \times p$ matrix with $n > p$. Show that

(a) $X^T X$ is positive definite.

(b) XX^T is positive semi-definite.

(c) The eigenvalues of $X^T X$ are all positive, and hence so is its trace.

Problem 2

Consider the unconstrained least-squares problem with full-rank $\mathbf{X}$:

$$\min_{\theta} \frac{1}{2} ||\mathbf{y} - \mathbf{X}\theta||_2^2 , \tag{1}$$

Show that Newton's method applied to this problem converges in one step.

Problem 3

In this problem, we use linear regression to rank the 30 teams the NBA comprises based on the outcomes of match-ups over the course of the 2014–2015 season. There were a total of 1408 match-ups during this season. Suppose the teams are given indices 1 through 30, according to their alphabetical order. The data consists of the sequences $\{(H_i, A_i)\}_{i=1}^{1408}$ $\{(y_i^H, y_i^A)\}_{i=1}^{1408}$, where $H_i \in \{1, \cdots, 30\}$ and $A_i \in \{1, \cdots, 30\}$ are the respective indices of the home and away teams and y_i^H and y_i^A their respective scores in the i^{th} match-up.

For instance, the first match-up of the season occurred on October 28, 2014, between the San Antonio Spurs (team index 27) and the Dallas Mavericks (team index 7) in San Antonio. The Spurs beat the Mavericks 101 to 100. Therefore, the data for this match-up are $(101, 100)$ and $(27, 7)$.

Consider the following linear model of the match-up data

$$y_i^H - y_i^A = \theta_{H_i} - \theta_{A_i} + \mu + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \cdots, 1408. \tag{2}$$

The model expresses the point difference between the Home and Away teams as a function of an underlying skill level for each team, θ_{H_i} for the home team and θ_{A_i} for the away team, and a home-court advantage parameter μ. The vector of unknowns is the vector $\beta = (\theta_1, \theta_2, \cdots, \theta_{30}, \mu)^T \in \mathbb{R}^{31}$.

(a) Consider a fictitious league with 3 teams and make up 5 match-ups among the teams in your league where each team plays one of the other teams at least once. Write the model of Eq. 2 in the form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{3}$$

For your fictitious league, show that the null-space of $\mathbf{X}$ is the span of the vector $(1, 1, 1, 0)^T$. Use this to conclude that, for the model of Eq. 2, the null-space of the corresponding $\mathbf{X}$ is the span of the vector $\gamma = (1, 1, 1, \cdots, 1, 0)^T$.

(b) Consider the problem

$$\min_{\beta} \frac{1}{2} ||\mathbf{y} - \mathbf{X}\beta||_2^2 , \tag{4}$$

where $\mathbf{X}$ and β are as defined above. Suppose $\hat{\beta}$ minimizes the above objective function. Show that this solution is not unique.

(c) To identify a unique solution, we solve the *constrained* least-squares problem

$$\min_{\beta} \frac{1}{2} ||\mathbf{y} - \mathbf{X}\beta||_2^2 \text{ s.t. } \gamma^T \beta = 0. \tag{5}$$

Write down the Lagrangian function $L(\beta, \lambda)$ for this constrained optimization problem.

(d) The normal equations identify the solution of the unconstrained least-squares problem. The corresponding equations for the constrained case are

$$\nabla_\beta L(\beta, \lambda) = 0 \tag{6}$$

$$\nabla_\lambda L(\beta, \lambda) = 0. \tag{7}$$

Show that solving this system of equations is equivalent to solving a linear system of the form

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b}^T \\ \hline \mathbf{b} & 0 \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} \mathbf{z} \\ \hline 0 \end{array}\right],$$

where $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{z}$ depend on $\mathbf{X}$, $\mathbf{y}$ and γ.

(e) Solve the ranking problem using constrained lest-squares and the model of Eq. 4 applied to the data in `nba2014-2015-seasonWL-WL-data.json`.

Estimators

Problem 4

(Problem 2.8 in book) Let $(y_i)_{i=1}^n$ be n i.i.d. samples from a random variable with bounded variance. Show that the sample mean is an unbiased estimator with variance that tends to zero asymptotically as n tends to infinity.

Extra credit: Show that the sample mean converges to the true mean in probability.

Problem 5

(Problem 3.4 in book) Let $\hat{\theta}_u$ be an unbiased estimator of a deterministic, unknown, parameter θ_0. Define a family of biased estimators $\hat{\theta}_b = (1+\alpha)\hat{\theta}_b$. Show that the range of α for which the MSE of $\hat{theta_b}$ is smaller than than of $\hat{\theta}_u$ is

$$-2 < -\frac{2MSE(\hat{\theta}_u)}{MSE(\hat{\theta}_u) + \theta_0^2} < \alpha < 0$$

Problem 6

(Problem 3.24 in book) Show that for the linear regression model

$$\mathbf{y} = \mathbf{X}\theta + \epsilon, \tag{8}$$

the a-posteriori probability density function $p_\Theta(\theta|\mathbf{y})$ is Gaussian if the prior $p_\Theta(\theta) \sim \mathcal{N}(\theta_0, \Sigma_0)$ and ϵ follows a $\mathcal{N}(0, \Sigma_\epsilon)$ distribution. Compute the mean and covariance matrix of the posterior distribution.