Ec240a, Fall 2015

Professor Bryan Graham

Problem Set 2

Due: November 9th, 2015

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including individually commented and executed code if applicable). A copy of any computer code should also be e-mailed to the GSI prior to the due date.

1 Pencil and paper problems

1. Let $\mathbf{Z} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$. Show that $\mathbb{E}\left[g\left(\mathbf{Z}\right)\left(\mathbf{Z}-\theta\right)\right] = \frac{\sigma^2}{N}\mathbb{E}\left[\nabla_{\mathbf{Z}} g\left(\mathbf{Z}\right)\right]$ (HINT: Use integration by parts).

2. Consider the following estimate of the Risk $\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|^2\right]$ associated with the weakly differentiable estimate $\hat{\theta}$:

$$\hat{R}\left(\mathbf{Z}\right) = K\sigma^2 + 2\sigma^2 \sum_{k=1}^{K} \frac{\partial g\left(\mathbf{Z}\right)}{\partial Z_k} + \sum_{k=1}^{K}\left(\hat{\theta}_k - Z_k\right)^2.$$

   Show that this risk estimate is unbiased: $\mathbb{E}_\theta\left[\hat{R}\left(\mathbf{Z}\right)\right] = R\left(\hat{\theta}, \theta\right)$.

3. (Wasserman, 2006). Let $\mathbf{Z} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$ and consider the following "soft threshold" estimator of θ :

$$\hat{\theta}_k = \operatorname{sgn}\left(Z_k\right)\left(\left|Z_k\right| - \lambda\right)_+, \; k = 1, \ldots, K.$$

   In words this estimator shrinks the MLE of θ_k toward zero when it is large (in absolute value) and shrinks it exactly to zero when it is small (in absolute value). Use Stein's Unbiased Risk Estimate (SURE) to show that

$$\hat{R}_{\text{SURE}}\left(\mathbf{Z}, \lambda\right) = \frac{K}{N}\sigma^2 - \frac{2\sigma^2}{N}\sum_{k=1}^{K} \mathbf{1}\left(\left|Z_k\right| \leq \lambda\right) + \sum_{k=1}^{K} \min\left(Z_k^2, \lambda^2\right).$$

   Provide a concrete prediction problem where you would expect the risk properties of the soft threshold estimator to be attractive.

4. (Wasserman, 2006) Let $\mathbf{Z} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$ and $\mathcal{M}$ be the class of ordered subsets

$$\left\{\emptyset, \{1\}, \{1,2\}, \{1,2,3\}, \ldots, \{1, \ldots, K\}\right\}$$

   and consider the estimator $\hat{\theta}_k\left(M\right) = Z_k \mathbf{1}\left(k \in M\right)$ for $k = 1, \ldots, K$. Use SURE to show that

$$\hat{R}_{\text{SURE}}\left(\mathbf{Z}, M\right) = \frac{\sigma^2}{N}\left|M\right| + \sum_{k \in M^c}\left(Z_k^2 - \frac{\sigma^2}{N}\right)$$

   with $|M|$ denoting the cardinality of M and M^c the absolute complement of M in the universe $\mathcal{M}$.

2 Calorie demand

Download the **RPS_calorie_data.out** dataset from the course webpage. For this problem set you will require only two columns of the dataset, those with the headings **Y0tc** and **X0te**. The first equals the log of total calories consumed in a household and the latter equals the log of total expenditure. To learn more about the dataset see Section 4 of Graham and Powell (2012, *Econometrica*). Subramanian and Deaton (1996, *Journal of Political Economy*) provide more background on calorie demand analysis.

1. Let Y denote log calories and X denote log expenditure. Assume that

$$m(x) = \mathbb{E}[Y|X = x] = \sum_{k=1}^{K} \alpha_k g_k(x)$$

   where $g_k(x) = x^{k-1}$.

2. Using the power series basis described above and the Gram-Schmidt algorithm construct a new basis that is orthogonal to the design points (set $K = 12$). Let W_i denote the $K \times 1$ vector of orthonormal basis functions for the i^{th} household. Compute the least squares fit

$$m(X_i) = W_i'\hat{\theta}$$

   with

$$\hat{\theta} = \left[\sum_{i=1}^{N} W_i W_i'\right]^{-1} \times \left[\sum_{i=1}^{N} W_i Y_i\right].$$

   Plot this function onto a scatter of the unsmoothed data.

3. Now use the shrinkage estimator of Efromovich (1999) as described in lecture to estimate $m(X_i)$. Plot this function onto a scatter of the unsmoothed data. Comment on your findings.

4. Now compute the soft threshold estimate of $m(X_i)$ (as defined above). Plot this function onto a scatter of the unsmoothed data.

3 Income and Geography

Download the dataset used in Hall and Jones (1997, *Quarterly Journal of Economics*) from

$$\texttt{http://web.stanford.edu/}\sim\texttt{chadj/HallJones400.asc}$$

From the second table in this file extract the logYL and Latitude columns, retaining only complete cases. Repeat the analysis in Problem 2 above with Y equal to logYL and X equal to Latitude.