Assignment #4 STA410H1F/2102H1F

due Monday December 4, 2017

Instructions: Solutions to problems 1 and 2 are to be submitted on Blackboard (PDF files strongly preferred).

1. Consider the model

$$Y_i = \theta_i + \varepsilon_i \quad (i = 1, \cdots, n)$$

where $\{\varepsilon_i\}$ is a sequence of random variables with mean 0 and finite variance representing noise. As in Assignment #2, we will assume that $\theta_1, \cdots, \theta_n$ are dependent or "smooth" in the sense that the absolute differences $\{|\theta_i - \theta_{i-1}|\}$ are small for most values of i. Rather than penalizing a lack of smoothness by $\sum_i (\theta_i - \theta_{i-1})^2$ (as in Assignment #2), we will consider estimating $\{\theta_i\}$ given data $\{y_i\}$ by minimizing

$$\sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}| \tag{1}$$

where $\lambda > 0$ is a tuning parameter. The resulting estimates $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$, are called fusion estimates.

The Gauss-Seidel algorithm used in Assignment #2 is effectively a coordinate descent algorithm. However, in computing the fusion estimates minimizing (1), application of the coordinate descent algorithm is frustrated to some extent by the fact that the non-differentiable penalty in (1) is not separable. However, this can be easily fixed by defining $\phi_i = \theta_i - \theta_{i-1}$ for $i = 2, \cdots, n$ and then minimizing

$$\sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=2}^{n} |\phi_i| \tag{2}$$

where now each θ_i (for $i = 2, \cdots, n$) will be a function of $\theta_1, \phi_2, \cdots, \phi_i$.

(a) Show that $\theta_k = \theta_1 + \sum_{i=2}^{k} \phi_i$ for $k \geq 2$.

(b) For fixed values of $\phi_2, \cdots, \phi_n$, show that the objective function is minimized at

$$\theta_1 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=2}^{i} \phi_j \right).$$

(c) If we fix the values of θ_1 and $\{\phi_i : 2 \leq i \neq j \leq n\}$, show that the subgradient of the objective function (2) with respect to ϕ_j is

$$-2 \sum_{i=j}^{n} \left(y_i - \theta_1 - \sum_{k=2}^{i} \phi_k \right) + \lambda \partial |\phi_j| = 2(n - j + 1)\phi_j - 2 \sum_{i=j}^{n} \left(y_i - \theta_1 - \sum_{k=2; k \neq j}^{i} \phi_k \right) + \lambda \partial |\phi_j|$$

where

$$\partial|\phi_j| = \begin{cases} +1 & \text{if } \phi_j > 0 \\ [-1,1] & \text{if } \phi_j = 0 \\ -1 & \text{if } \phi_j < 0 \end{cases}$$

and hence the objective function is minimized over ϕ_j for fixed values of θ_1 and $\{\phi_i : 2 \leq i \neq j \leq n\}$ at

$$\begin{aligned}
\phi_j &= 0 \ \text{ if } \left|\sum_{i=j}^{n}\left(y_i - \theta_1 - \sum_{k=2;k\neq j}^{i} \phi_k\right)\right| \leq \frac{\lambda}{2} \\
\phi_j &= \frac{1}{n-j+1}\left\{\sum_{i=j}^{n}\left(y_i - \theta_1 - \sum_{k=2;k\neq j}^{i} \phi_k\right) - \frac{\lambda}{2}\right\} \ \text{ if } \sum_{i=j}^{n}\left(y_i - \theta_1 - \sum_{k=2;k\neq j}^{i} \phi_k\right) > \frac{\lambda}{2} \\
\phi_j &= \frac{1}{n-j+1}\left\{\sum_{i=j}^{n}\left(y_i - \theta_1 - \sum_{k=2;k\neq j}^{i} \phi_k\right) + \frac{\lambda}{2}\right\} \ \text{ if } \sum_{i=j}^{n}\left(y_i - \theta_1 - \sum_{k=2;k\neq j}^{i} \phi_k\right) < -\frac{\lambda}{2}
\end{aligned}$$

(d) [BONUS - but highly recommended!] Write a function in R implementing the coordinate descent algorithm suggested in parts (b) and (c) and test it on the data generated as in Assignment #2. (Convergence of the coordinate descent algorithm here is very slow but you should be able to produce good estimates of the underlying function.)

2. The Hidalgo stamp data is a (semi-)famous dataset containing thicknesses of 482 postage stamps from the 1872 Mexican "Hidalgo" issue. It is believed that these stamps were printed on different types of papers so that the data can be modeled as a "mixture" of several distributions. The data are available in a file `stamp.txt` on Blackboard.

There is some consensus that a good model for these data is a mixture of normals; that is, the density of the data is

$$f(x) = \sum_{k=1}^{m} \lambda_k f_k(x)$$

where f_k is the density of a normal distribution with unknown mean and variance μ_k and σ_k^2 respectively. $\{\lambda_k\}$ are non-negative with $\lambda_1 + \cdots + \lambda_m = 1$.

(a) The "complete" data likelihood here is

$$L(\mu_1, \cdots, \mu_m, \sigma_1^2, \cdots, \sigma_m^2, \lambda_1, \cdots, \lambda_m) = \prod_{i=1}^{n}\prod_{k=1}^{m}\{f_k(x_i)^{u_{ik}}\lambda_k^{u_{ik}}\}$$

where $u_{ik} = 0$ or 1 with $u_{i1} + \cdots + u_{im} = 1$ for $i = 1, \cdots, n$. Show that the complete data MLEs are

$$\widehat{\mu}_k = \left(\sum_{i=1}^{n} u_{ik}\right)^{-1} \sum_{i=1}^{n} u_{ik} x_i$$

$$\begin{aligned}\widehat{\sigma}_k^2 &= \left(\sum_{i=1}^{n} u_{ik}\right)^{-1} \sum_{i=1}^{n} u_{ik}(x_i - \widehat{\mu}_k)^2 \\ \widehat{\lambda}_k &= \frac{1}{n}\sum_{i=1}^{n} u_{ik}.\end{aligned}$$

For the EM algorithm, we need to estimate the unobserved $\{u_{ik}\}$; for given $\{\lambda_k\}$ and $\{f_k\}$ (which depend on the means and variances), these estimates are

$$\widehat{u}_{ik} = \frac{\lambda_k f_k(x_i)}{\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)}.$$

(You do not need to prove this.)

(b) Write an R function that uses the EM algorithm to compute the MLEs of $\lambda_1, \cdots, \lambda_m$, $\mu_1, \cdots, \mu_m$ and $\sigma_1^2, \cdots, \sigma_m^2$. Your function should take as inputs initial estimates of these unknown parameters. (A template function will be provided.)

(c) Using your function, estimate the parameters in the normal mixture model with $m = 5, 6,$ and 7 components. The success of the EM algorithm for mixture models depends on having reasonable initial estimates of the parameters. One simple *ad hoc* approach is to use a kernel density estimate where the bandwidth parameter is varied so that the estimate has the appropriate number of modes (corresponding to the different components. This can be done using the R function `density` — use the default kernel (which is a Gaussian kernel) and change the bandwidth using the parameter `bw`; for example, if the data are in `stamp` then `plot(density(stamp,bw=0.002))` will give a plot of the density estimate for a bandwidth of 0.002.

(d) Which of the 3 models considered in part (b) do you prefer and why?