**Assignment #4 STA410H1F/2102H1F**

due Friday November 29, 2019

**Instructions:** Solutions to problems 1 and 2 are to be submitted on Quercus (PDF files only).

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. 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 $\sum_i (\theta_i - 2\theta_{i-1} + \theta_{i-2})^2$ (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 and $\displaystyle\sum_{i=2}^{n} |\theta_i - \theta_{i-1}|$ represents the total variation of $\{\theta_i\}$. The resulting estimates $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$, are sometimes called fusion estimates and are useful if $\{\theta_i\}$ contain "jumps", that is, $\theta_i = g(i/n)$ where $g$ is a smooth function with a small number of discontinuities (i.e. jumps).

The non-differentiable part of the objective function in (1) can be made separable 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 $\theta_i$ (for $i = 2, \cdots, n$) will be a function of $\theta_1, \phi_2, \cdots, \phi_i$. The representation of the objective function in (2) can be used to compute the parameter estimates using coordinate descent although there is must faster algorithm. However, (2) is useful for deriving properties of the estimates.

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

(b) Show that if $\widehat{\theta}_1, \cdots, \widehat{\theta}_n$ minimize (1) (or (2)) then

$$\sum_{i=1}^{n} (y_i - \widehat{\theta}_i) = 0.$$

(Hint: Use the representation (2) and compute its partial derivative with respect to $\theta_1$.)

(c) Show that $|y_i - \widehat{\theta}_i| \le \lambda$ for all $i$. (Hint: Show that

$$\partial \left\{ \lambda \sum_{i=2}^{n} |\theta_i - \theta_{i-1}| \right\} \subset [-2\lambda, 2\lambda]^n$$

subgradient

for any $\theta_1, \cdots, \theta_n$.)

(d) For $\lambda$ sufficiently large, we will have $\widehat{\theta}_1 = \cdots = \widehat{\theta}_n = \bar{y}$ or equivalently $\widehat{\phi}_2 = \cdots = \widehat{\phi}_n = 0$. How large must $\lambda$ be in order to have $\widehat{\theta}_1 = \cdots = \widehat{\theta}_n = \bar{y}$? (Hint: Look at the sub-gradient of (2) with respect to $(\theta_1, \phi_2, \cdots, \phi_n)$; when is $(0, 0, \cdots, 0)$ an element of this sub-gradient at $(\bar{y}, 0, \cdots, 0)$?)

**Note:** An R function `tvsmooth` is available on Quercus in a file `tvsmooth.txt`. You may find it useful to simulate data from a discontinous function with additive noise and estimate the function using `tvsmooth` to gain some insight into this method.

2. Suppose that $X_1, \cdots, X_n$ are sampled from the following truncated Poisson distribution:

$$P_\lambda(X_i = x) = \frac{\exp(-\lambda)\lambda^x}{x!\kappa_\lambda(r)} \quad \text{for } x = r+1, r+2, \cdots$$

for some integer $r \ge 0$ where

$$\kappa_\lambda(r) = \sum_{x=r+1}^{\infty} \frac{\exp(-\lambda)\lambda^x}{x!} = 1 - \sum_{x=0}^{r} \frac{\exp(-\lambda)\lambda^x}{x!}.$$

Such a sample might arise if we were sampling from a Poisson population but were unable to observe data less than or equal to $r$.

The EM algorithm can be employed to estimate $\lambda$ from the observed $X_1, \cdots, X_n$. The key is to think of the observed data as a subset of some larger ("complete") data set $X_1, \cdots, X_n, X_{n+1}, \cdots, X_{n+M}$ where $M \ge 0$ is a random variable and $X_{n+1}, \cdots, X_{n+M} \le r$; given $M = m$, this complete data set is now assumed to be $m + n$ independent observations from a Poisson distribution with mean $\lambda$. The log-likelihood for the complete data is

$$\ln \mathcal{L}(\lambda) = \ln(\lambda) \sum_{i=1}^{n+m} x_i - (n+m)\lambda,$$

which depends on two unknowns $\displaystyle\sum_{i=n+1}^{n+m} x_i$ and $m$. To use the EM algorithm, we need to estimate these two unknowns.

(a) The probability distribution of $M$ is

$$P_\lambda(M = m) = \binom{n+m-1}{m} (1 - \kappa_\lambda(r))^m \kappa_\lambda(r)^n \quad \text{for } m = 0, 1, 2, \cdots$$

Show that $E_\lambda(M) = n(1-\kappa_\lambda(r))/\kappa_\lambda(r)$.

(b) Show that

$$E_\lambda\left(\sum_{i=n+1}^{n+M} X_i \;\middle|\; X_1 = x_1, \cdots, X_n = x_n\right) = E_\lambda(M)E_\lambda(X_i|X_i \le r).$$

(Hint: Note that (a) $X_{n+1}, \cdots, X_{n+M}$ are independent of $X_1, \cdots, X_n$ and (b) $X_{n+1}, \cdots, X_{n+M} \le r$.)

(c) Consider the data given in the table below. They represent the accident claims submitted to La Royale Belge Insurance Company during a single year. A crude model for the number of claims submitted for a given policy is Poisson. However, the data below does not provide the number of policies for which no claims were submitted. We want to estimate $\lambda$ as well as to impute (estimate) the number $M$ of policies with no claims.

| Number of claims | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Number of policies | 1317 | 239 | 42 | 14 | 4 | 4 | 1 |

Assume a truncated Poisson model for these data taking $r = 0$ and estimate $\lambda$ as well as $M$ using the EM algorithm (which in this case has a particularly simple form). Do you think the truncated Poisson model is useful for these data? (For example, do you think your estimate of $M$ is reasonable?)