Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation via an MM Algorithm

Chong Wang
Department of Statistics
North Carolina State University
Raleigh, NC 27695
cwang19@ncsu.edu

Abstract

The quantile regression explores the entire conditional distribution of a random variable by controlling target quantile levels. For time series or spatial data, due to the dependence among neighboring observations it is challenging to estimate quantile curves. The fused adaptive lasso (FAL) penalized the difference among quantiles to account for the dependence of data at neighboring time or space. The temporal and spatial FAL has been implemented on both simulation and real data with equally spaced observations when construct the L1 penalty term in (Sun et al., 2015). (Hunter and Lange, 2000) study on quantile regression algorithm is termed an MM algorithm, which entails majorizing the objective function by a quadratic function followed by minimizing that quadratic. In this project we develop an MM algorithm to estimate the FAL for spatial and temporal quantile function. Furthermore, we will modify the FAL penalty term by using unevenly spaced observations for the quantile function over time.

1 Introduction

In (Sun et al., 2015) uses linear programing to solve their problem, which may not scale as well as a smoothed alternative. We use the smoothing trick based on (Hunter and Lange, 2000) and then use an alternative solver for the smoothed problem.
1. Speed
2. Accuracy
3. Flexibility

2 Methodology

Let $\{Y(\boldsymbol{s}) : s \in S \subset R^2\}$ be a random field. Suppose we observe spatial data $\{Y(\boldsymbol{s_1}), ..., Y(\boldsymbol{s_n})\}$ at locations $\boldsymbol{s_1}, ..., \boldsymbol{s_n} \in S$. For each location $\boldsymbol{s_i}$, the selected set of neighbors, excluding $\boldsymbol{s_i}$, is denoted by $N_{\boldsymbol{s_i}}$ with $j \in N_{\boldsymbol{s_i}}$ representing $\boldsymbol{s_j}$ is a neighbor of $\boldsymbol{s_i}$. The size of $N_{\boldsymbol{s_i}}$ is denoted as $n_{\boldsymbol{s_i}}$. Define $\theta_i = q_\tau(\boldsymbol{s}_i) - \frac{1}{n_{\boldsymbol{s_i}}} q_\tau(\boldsymbol{s}_j)$ and let $\frac{1}{n_{\boldsymbol{s_i}}} \sum_{j \in N_{\boldsymbol{s_i}}} q_\tau(\boldsymbol{s}_j) = \sum_{j=1}^n \phi_{\boldsymbol{s_i},\boldsymbol{s_j}} q_\tau(\boldsymbol{s}_j)$, where $\phi_{\boldsymbol{s_i},\boldsymbol{s_j}} = 1/n_{\boldsymbol{s_i}}$ if $j \in N_{\boldsymbol{s_i}}$; $\phi_{\boldsymbol{s_i},\boldsymbol{s_j}} = 0$ otherwise. The τth quantile function is $\boldsymbol{q_\tau} = (q_\tau(\boldsymbol{s}_1), ..., q_\tau(\boldsymbol{s}_n))^T$ and $\boldsymbol{\phi} = \{\phi_{\boldsymbol{s_i},\boldsymbol{s_j}}\}_{i,j=1}^n$ indicates how $q_\tau(\boldsymbol{s}_i)$ depends on the rest of $q_\tau(\boldsymbol{s}_j)$. Therefore, $\boldsymbol{\theta} = (\boldsymbol{I}_n - \boldsymbol{\phi})\boldsymbol{q_\tau}$ and assume N_{s_i} is selected such that $rank(\boldsymbol{\phi}) = n$, then $\boldsymbol{q_\tau} = \boldsymbol{L}^T\boldsymbol{\theta}$ where $\boldsymbol{L}^T = (\boldsymbol{I}_n - \boldsymbol{\phi})^{-1}$. To estimate $\boldsymbol{\theta}$, define the objective function as

$$f(\boldsymbol{\theta}) = \sum_{i=1}^n \rho_\tau(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}) + \lambda \sum_{i=1}^n w_{\boldsymbol{s_i}} |\theta_i|, \tag{1}$$

where λ is the tuning parameter and $\boldsymbol{l_i}$ is the ith column of $\boldsymbol{L}$. The adaptive weights $w_{\boldsymbol{s_i}}$ are chosen to be equal to $min(|\tilde{\theta}_j|^{-1}, \sqrt{n})$, where the initial estimator $\tilde{\boldsymbol{\theta}}$ is the minimizer of (1) with $w_{\boldsymbol{s_i}} = 1$.

To simplify the notation, we denote the penalized term in (1) as $\lambda||\boldsymbol{W}^T\boldsymbol{\theta}||_1$, where $\boldsymbol{W}$ is a diagonal matrix with elements of $w_{\boldsymbol{s_i}}$ on the diagonal. Thus the objective function can be represented as

$$f(\boldsymbol{\theta}) = \sum_{i=1}^n \rho_\tau(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}) + \lambda||\boldsymbol{W}^T\boldsymbol{\theta}||_1. \tag{2}$$

2.1 Majorize Step

In order to minimize the objective function (2), we decide to use the MM algorithm applied in (Hunter and Lange, 2000). Therefore, we first need to find the majorizer of function (2). However, the first part $\sum_{i=1}^n \rho_\tau(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta})$ of the objective function (2) is difficult to minimize because it can admit multiple minima and because it is not differentiable everywhere. Inspired by the smoothing technique in (Hunter and Lange, 2000), our approach to this minimization problem is first to construct a function that approximates the first part in function (2) very closely and then to use an MM algorithm to minimize the approximating function.

For $\epsilon > 0$, define

$$\rho_{i,\tau}^\epsilon(\boldsymbol{\theta}^k) = \rho_\tau(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k) - \frac{\epsilon}{2} ln(\epsilon + |Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k|), \tag{3}$$

Then the sum $\sum_{i=1}^n \rho_{i,\tau}^\epsilon(\boldsymbol{\theta}^k)$ approximates the first part $\sum_{i=1}^n \rho_\tau(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta})$. A majorizer of $\rho_{i,\tau}^\epsilon(\boldsymbol{\theta}^k)$ is defined as

$$\xi_{i,\tau}^\epsilon(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = \frac{1}{4}\left[\frac{(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta})^2}{\epsilon + |Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k|} + (4\tau - 2)(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}) + c\right] \tag{4}$$

where c is a constant chosen so that $\xi_{i,\tau}^\epsilon(\boldsymbol{\theta}^k|\boldsymbol{\theta}^k) = \rho_{i,\tau}^\epsilon(\boldsymbol{\theta}^k)$.

Following the derivation in (Selesnick, 2012), a majorizer of $\lambda||\boldsymbol{W}^T\boldsymbol{\theta}||_1$ is given by

$$\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} + \frac{\lambda}{2}||\boldsymbol{W}^T\boldsymbol{\theta}^{(k)}||_1, \tag{5}$$

where matrix $\boldsymbol{\Lambda}_k$ is a diagonal matrix with elements of $\lambda/|\boldsymbol{W}^T\boldsymbol{\theta}^{(k)}|$ on the diagonal.

Therefore,

$$G_\epsilon(\boldsymbol{\theta}|\boldsymbol{\theta}^k) = \sum_{i=1}^{n} \xi_{i,\tau}^{\epsilon}(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + \frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{W}\boldsymbol{\Lambda}_k \boldsymbol{W}^T\boldsymbol{\theta} + \frac{\lambda}{2}||\boldsymbol{W}^T\boldsymbol{\theta}^{(k)}||_1, \tag{6}$$

majorizes

$$f_\epsilon(\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho_{i,\tau}^{\epsilon}(\boldsymbol{\theta}^k) + \lambda||\boldsymbol{W}^T\boldsymbol{\theta}^{(k)}||_1. \tag{7}$$

2.2 Minimization Step

The MM update (1) for $\boldsymbol{\theta}^k$ is

$$\boldsymbol{\theta}^{k+1} = \text{argmin}_{\boldsymbol{\theta}} \left[\textstyle\sum_{i=1}^{n} \xi_{i,\tau}^{\epsilon}(\boldsymbol{\theta}|\boldsymbol{\theta}^k) + \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} \right] \tag{8}$$

$$= (\boldsymbol{L}\boldsymbol{U}_\epsilon(\boldsymbol{\theta}^k)\boldsymbol{L}^T + 2\boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T)^{-1}\boldsymbol{L}\boldsymbol{V}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}^k) \tag{9}$$

where

$$\boldsymbol{V}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}^k) = \left(\frac{Y(\boldsymbol{s_1})}{\epsilon + |(Y(\boldsymbol{s_1}) - \boldsymbol{l}_1^T\boldsymbol{\theta}^k)|} + 2q - 1, ..., \frac{Y(\boldsymbol{s_n})}{\epsilon + |(Y(\boldsymbol{s_n}) - \boldsymbol{l}_n^T\boldsymbol{\theta}^k)|} + 2q - 1 \right)^T$$

$$\boldsymbol{U}_\epsilon(\boldsymbol{\theta}^k) = \text{diag}\left\{ \frac{1}{\epsilon + |(Y(\boldsymbol{s_1}) - \boldsymbol{l}_1^T\boldsymbol{\theta}^k)|}, ..., \frac{1}{\epsilon + |(Y(\boldsymbol{s_n}) - \boldsymbol{l}_n^T\boldsymbol{\theta}^k)|} \right\}$$

3 Analysis and Results

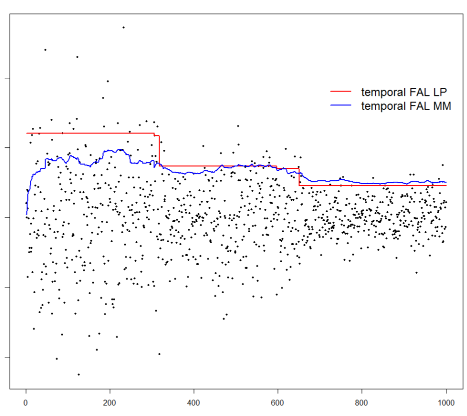

Figure 1: Sample figure caption.

4 Dicussion

References

Hunter, D. R. and Lange, K. (2000). Quantile regression via an mm algorithm, *Journal of Computational and Graphical Statistics* **9**(1): 60–77.

Selesnick, I. (2012). Sparse deconvolution (an mm algorithm), *Connexions* .

Sun, Y., Wang, H. J. and Fuentes, M. (2015). Fused adaptive lasso for spatial and temporal quantile function estimation, *Technometrics* (just-accepted): 00–00.

Appendix

The last term of $G_\epsilon(\boldsymbol{\theta}|\boldsymbol{\theta}^k)$ has been omitted because it does not depend on $\boldsymbol{\theta}$. Given the first differential

$$\begin{aligned}
dG_\epsilon(\boldsymbol{\theta}|\boldsymbol{\theta}^k) &= -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}}{\epsilon + |(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k)|} + 2q - 1\right]\boldsymbol{l}_i + \boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{Y(\boldsymbol{s_i})}{\epsilon + |(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k)|} + 2q - 1\right]\boldsymbol{l}_i + \frac{1}{2}\sum_{i=1}^{n}\frac{\boldsymbol{l}_i^T\boldsymbol{\theta}\boldsymbol{l}_i}{\epsilon + |(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k)|} + \boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{Y(\boldsymbol{s_i})}{\epsilon + |(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k)|} + 2q - 1\right]\boldsymbol{l}_i + \frac{1}{2}\sum_{i=1}^{n}\frac{\boldsymbol{l}_i\boldsymbol{l}_i^T\boldsymbol{\theta}}{\epsilon + |(Y(\boldsymbol{s_i}) - \boldsymbol{l}_i^T\boldsymbol{\theta}^k)|} + \boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} \\
&= -\frac{1}{2}\boldsymbol{L}\boldsymbol{V_\epsilon}(\boldsymbol{\theta}^k) + \frac{1}{2}\boldsymbol{L}\boldsymbol{U}_\epsilon(\boldsymbol{\theta}^k)\boldsymbol{L}^T\boldsymbol{\theta} + \boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T\boldsymbol{\theta} \overset{\text{set}}{=} 0
\end{aligned} \tag{A.1}$$

Then the updated θ in the new step is

$$\boldsymbol{\theta}^{k+1} = (\boldsymbol{L}\boldsymbol{U}_\epsilon(\boldsymbol{\theta}^k)\boldsymbol{L}^T + 2\boldsymbol{W}\boldsymbol{\Lambda}_k\boldsymbol{W}^T)^{-1}\boldsymbol{L}\boldsymbol{V_\epsilon}(\boldsymbol{\theta}^k) \tag{A.2}$$

where

$$\boldsymbol{V_\epsilon}(\boldsymbol{\theta}^k) = \left(\frac{Y(\boldsymbol{s_1})}{\epsilon + |(Y(\boldsymbol{s_1}) - \boldsymbol{l}_1^T\boldsymbol{\theta}^k)|} + 2q - 1, ..., \frac{Y(\boldsymbol{s_n})}{\epsilon + |(Y(\boldsymbol{s_n}) - \boldsymbol{l}_n^T\boldsymbol{\theta}^k)|} + 2q - 1\right)^T$$

$$\boldsymbol{U}_\epsilon(\boldsymbol{\theta}^k) = \text{diag}\left\{\frac{1}{\epsilon + |(Y(\boldsymbol{s_1}) - \boldsymbol{l}_1^T\boldsymbol{\theta}^k)|}, ..., \frac{1}{\epsilon + |(Y(\boldsymbol{s_n}) - \boldsymbol{l}_n^T\boldsymbol{\theta}^k)|}\right\}$$