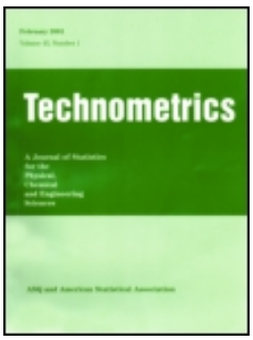

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation

Ying Sun, Huixia J. Wang & Montserrat Fuentes

To cite this article: Ying Sun, Huixia J. Wang & Montserrat Fuentes (2016) Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation, Technometrics, 58:1, 127-137, DOI: 10.1080/00401706.2015.1017115

To link to this article: https://doi.org/10.1080/00401706.2015.1017115

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation

Ying SUN

CEMSE Division
King Abdullah University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
(*ying.sun@kaust.edu.sa*)

Huixia J. WANG

Department of Statistics
George Washington University
Washington, DC 20052
(*judywang@gwu.edu*)

Montserrat FUENTES

Department of Statistics
North Carolina State University
Raleigh, NC 27695
(*fuentes@stat.ncsu.edu*)

Quantile functions are important in characterizing the entire probability distribution of a random variable, especially when the tail of a skewed distribution is of interest. This article introduces new quantile function estimators for spatial and temporal data with a fused adaptive Lasso penalty to accommodate the dependence in space and time. This method penalizes the difference among neighboring quantiles, hence it is desirable for applications with features ordered in time or space without replicated observations. The theoretical properties are investigated and the performances of the proposed methods are evaluated by simulations. The proposed method is applied to particulate matter (PM) data from the Community Multiscale Air Quality (CMAQ) model to characterize the upper quantiles, which are crucial for studying spatial association between PM concentrations and adverse human health effects.

KEY WORDS: Fused adaptive Lasso; Particulate matter; Quantile estimation; Spatial quantiles; Temporal quantiles.

1. INTRODUCTION

For a continuous univariate random variable Y, the quantile function is defined as $q_\tau = F^{-1}(\tau) = \{y \in \mathbb{R} : F(y) = \tau\}$, where F is the strictly increasing cumulative distribution function (cdf) and $\tau \in (0, 1)$ is the quantile level. Parzen (2004, 2008) gave comprehensive discussions on quantiles from theory to practice. The quantile function is important in characterizing the entire probability distribution, for example, the median is a robust location estimator and upper quantiles are useful in describing extreme events.

When data, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, are from a stochastic process, multivariate quantile function is not uniquely defined. We consider marginal quantiles and define the quantile function as $\mathbf{q}_\tau = (q_\tau(1), \ldots, q_\tau(n))^{\mathrm{T}}$, where $q_\tau(i) = F_i^{-1}(\tau)$ with F_i denoting the marginal cdf of Y_i. In practice, estimating $\mathbf{q}_\tau$ is difficult because Y_i's are correlated and repeated measurements of the random vector $\mathbf{Y}$ are usually not available. For example, for time series or spatial data, it is challenging to estimate the temporal quantile curves or spatial quantile surfaces due to the dependence among neighboring observations. Ignoring the dependence may underestimate the magnitude of the quantile function. In this article, we propose a method to tackle this problem via a smoothing procedure to estimate $\mathbf{q}_\tau$ locally. The advantage of this method is that the index-varying $q_\tau(i)$, $i = 1, \ldots, n$, describes the changing behavior in the marginal distribution, while a local smoothing technique accounts for the dependence among $q_\tau(i)$'s. Moreover, only one realization of $\mathbf{Y}$ is needed to estimate $\mathbf{q}_\tau$.

In the presence of right-skewed distributions, the usual Gaussian assumption will underestimate the upper tail probability and methods based on the extreme value theory or quantiles are more appropriate to characterize the tail distribution. Cooley et al. (2012) reviewed current methodology for the analysis of spatial extremes, focusing on methods that account for dependence in extremes, such as max-stable processes and copula models. Under these models, the statistical inference requires repeated measurements, which are not generally available for stochastic processes. In such a framework, the main goal of the article consists of proposing nonparametric methods for estimating upper quantile values in time or space without replicated observations. Our work is motivated by the need of better tools to characterize the tail distribution of the particulate matter (PM), a mixture of pollutants. Since most air quality data are right-skewed under stable meteorological conditions,

TECHNOMETRICS, FEBRUARY 2016, VOL. 58, NO. 1
DOI: 10.1080/00401706.2015.1017115
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/tech.

and are observed as a time series at a particular location, or a spatial surface at a certain time, we need to characterize the upper quantiles using one time series or one set of spatial data only.

For stochastic processes, one way to create replication is via a stationarity assumption. For example, the second-order stationarity means that the mean and the covariance of the process do not change when shifted in time or space. Specifically, a time series, $\{Y(t)\}$, is second-order stationary if, for all δ, and for all t, $E\{Y(t)\} = \mu$, and $\text{cov}\{Y(t), Y(t+\delta)\} = K(\delta)$. Under the second-order stationarity assumption, covariance functions play an important role in interpreting the local behavior, and Gaussian process models are widely used. For skewed distributions, methodologies based on multivariate extreme value theory are developed to model the tail dependence, but it becomes difficult for high or infinite dimensions (Cooley et al. 2012). Another way to model the local behavior is through nonparametric smoothing. In practice, it is often difficult to tell whether the true underlying process is nonstationary or stationary with long range dependence by only looking at the observed data. The local features can be either interpreted as a result of the nonstationarity or the long range dependence in a stationary process. For example, Furrer and Nychka (2007) presented a framework to understand the asymptotic properties of kriging and splines and proved the equivalent interpretation between kriging estimators and generalized splines based on Nychka (1995). Stein (1991) showed that the predictions that a generalized covariance function yields are identical to an order 2 thin plate smoothing spline. Draghicescu, Guillas, and Wu (2009) addressed time-varying quantile curve estimation for a class of nonstationary time series.

In this article, we propose nonparametric methods assuming nonstationarity to capture the local dependence by minimizing a loss function with a penalty term. For example, when observations are temporally inhomogeneous, we assume the marginal distribution of Y_i varies over time to characterize the temporal dependence, and improve the quantile estimation by penalizing the roughness. We use the fused Lasso technique (Tibshirani et al. 2005), but focus on estimating quantile functions in space or time, where the fused Lasso penalty enforces sparsity in the difference among neighboring quantiles. This method borrows neighboring information and accommodates the dependence for temporal and spatial quantile function estimation. Hence it is desirable for applications with features ordered in time or space without repeated measurements.

The rest of our article is organized as follows. In Section 2, we propose spatial and temporal fused adaptive Lasso estimators for quantile function estimation. We then present theoretical properties of the proposed estimators, and introduce a computationally efficient framework. For time series and spatial data, the performance of the proposed methods are justified by simulation studies in Section 3, and applications to $PM_{2.5}$ air quality data are presented in Section 4. Section 5 discusses possible extensions. The proof of the theorem and discussions on verification of the assumed conditions are provided in the supplementary material.

2. METHODOLOGY

2.1 Background

To estimate quantile functions, quantile regression, first proposed by Koenker and Bassett (1978), has become an important and widely used technique. Quantile regression can be used to estimate quantiles of a response variable conditional on the regressors or predictors, and thus provide a comprehensive picture of the conditional response. Suppose we observe $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$, where $\mathbf{X}_i \in \mathbb{R}^p$ is the vector of predictors and $Y_i \in \mathbb{R}$ is the continuous response. Let $q_\tau(\mathbf{X}_i)$ be the conditional τth quantile of Y_i, such that $P(Y_i \le q_\tau(\mathbf{X}_i)|\mathbf{X}_i) = \tau$, for $0 < \tau < 1$. The quantile $q_\tau(\mathbf{X}_i)$ is estimated by minimizing $\sum_{i=1}^n \rho_\tau(Y_i - q_\tau(\mathbf{X}_i))$, where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$.

Before introducing the regularization technique in quantile regression model, we briefly review two types of Lasso penalties in the least squares regression model. The Lasso methodology, as proposed by Tibshirani (1996), is a regularization technique for simultaneous estimation and variable selection. In a linear regression model $E(Y_i|\mathbf{X}_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, the coefficient $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is estimated by minimizing $\sum_{i=1}^n (Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|$, where λ is a tuning parameter. The L_1 norm penalty shrinks the coefficients toward 0 as λ increases. The fused Lasso penalty (Tibshirani et al. 2005) enforces sparsity in both the coefficients and their successive differences, which is desirable for applications with features ordered in some meaningful way. The penalty term can be written as $\lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=2}^p |\beta_j - \beta_{j-1}|$.

The Lasso-type penalties in the least squares regression model are widely used in many fields. For example, in comparative genomic hybridization (CGH) analysis, Eilers (2003), Huang et al. (2005), and Tibshirani and Wang (2008) employed the penalized least squares method by shrinking the distances between signals at adjacent clone locations, taking the spatial dependence into account. Nowak et al. (2011) considered multiple arrays of the CGH data. Moreover, Storlie et al. (2011) studied the Lasso penalty for surface estimation and variable selection in multivariate nonparametric regression. Das et al. (2012) applied the penalized least squares method with both L_1 and L_2 norm penalties to spatial precipitation extremes. Harchaoui and Lévy-Leduc (2010) used the fused Lasso idea for the estimation of the location of change-points in one-dimensional signals.

Just as in the ordinary least squares regression, it is also necessary to perform smoothing or variable selection in quantile regression model. For example, Wu and Liu (2009) proposed penalized quantile regression with adaptive Lasso penalties, where simultaneous estimation and variable selection can be done in the quantile regression model. Eilers and Menezes (2005), and Li and Zhu (2007) considered smoothing the quantiles of the log ratio CGH data. Wang and Hu (2011) proposed a new penalized regression approach with a fused Lasso penalty to accommodate the spatial dependence of clones. Jiang, Wang, and Bondell (2013) developed a joint modeling of multiple quantiles with fusion-type penalties, shrinking the differences of quantile slopes at two adjacent quantile levels toward zero.

In this article, we are interested in time series and spatial data. Unlike the usual quantile regression setting, we assume data are observed from $\{Y(\mathbf{r}) : \mathbf{r} \in \mathcal{I}\}$, where $\mathcal{I}$ is the index set: $\mathbf{r} = t$ represents time and $\mathcal{I} = \mathbb{R}$; $\mathbf{r} = \mathbf{s}$ represents space and $\mathcal{I} = \mathbb{R}^d$ with $d = 2$ although $d > 2$ could be considered as well. Then the τth quantile function $\mathbf{q}_\tau = (q_\tau(\mathbf{r}_1), \ldots, q_\tau(\mathbf{r}_n))^{\mathrm{T}}$ can be estimated by minimizing $\sum_{i=1}^n \rho_\tau(Y(\mathbf{r}_i) - q_\tau(\mathbf{r}_i))$. To capture the local dependence, we add an L_1 penalty term to penalize the difference among quantiles that are close to each other in time or space, called temporal or spatial fused adaptive Lasso that is introduced in Sections 2.2 and 2.3. The asymptotic properties of the proposed estimator are presented in Section 2.4, and the implementation and computational details are illustrated in Section 2.5.

2.2 Spatial Fused Adaptive Lasso

Let $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{S} \subset \mathbb{R}^2\}$ be a random field. Suppose we observe spatial data $\{Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\}$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{S}$. For each location $\mathbf{s}_i$, the selected set of neighbors, excluding $\mathbf{s}_i$, is denoted by $N_{\mathbf{s}_i}$ with $j \in N_{\mathbf{s}_i}$ representing that $\mathbf{s}_j$ is a neighbor of $\mathbf{s}_i$. We propose to estimate the τth quantile function $q_\tau(\mathbf{s})$ by a spatial fused adaptive Lasso (FAL) that minimizes

$$\sum_{i=1}^n \rho_\tau(Y(\mathbf{s}_i) - q_\tau(\mathbf{s}_i)) + \lambda \sum_{i=1}^n \omega_{\mathbf{s}_i} \left| q_\tau(\mathbf{s}_i) - \frac{1}{n_{\mathbf{s}_i}} \sum_{j \in N_{\mathbf{s}_i}} q_\tau(\mathbf{s}_j) \right|, \tag{1}$$

where λ is the tuning parameter, $n_{\mathbf{s}_i} = |N_{\mathbf{s}_i}|$ is the size of $N_{\mathbf{s}_i}$, and $\omega_{\mathbf{s}_i}$ are the adaptive weights. If $\lambda = \infty$, $q_\tau(\mathbf{s}) = q_\tau$, that is, the quantile function does not change over space; if $\lambda = 0$, $q_\tau(\mathbf{s}_i) = Y(\mathbf{s}_i)$, that is, the quantile function goes through each observation.

The spatial FAL method estimates $q_\tau(\mathbf{s})$ by minimizing the quantile loss function and penalizing the difference between neighboring quantiles in space, that is, $q_\tau(\mathbf{s}_i)$ and the average of its neighboring quantiles. The tuning parameter λ is important and determines how $q_\tau(\mathbf{s})$ changes over space. The following reparameterization provides a better understanding of the objective function (1). Define $\theta_i = q_\tau(\mathbf{s}_i) - \frac{1}{n_{\mathbf{s}_i}} \sum_{j \in N_{\mathbf{s}_i}} q_\tau(\mathbf{s}_j)$ and let $\frac{1}{n_{\mathbf{s}_i}} \sum_{j \in N_{\mathbf{s}_i}} q_\tau(\mathbf{s}_j) = \sum_{j=1}^n \phi_{\mathbf{s}_i,\mathbf{s}_j} q_\tau(\mathbf{s}_j)$, where $\phi_{\mathbf{s}_i,\mathbf{s}_j} = 1/n_{\mathbf{s}_i}$ if $j \in N_{\mathbf{s}_i}$; $\phi_{\mathbf{s}_i,\mathbf{s}_j} = 0$ otherwise. Therefore, the penalty terms can be written as $\boldsymbol{\theta} = (\mathbf{I}_n - \boldsymbol{\phi})\mathbf{q}_\tau$, where $\mathbf{q}_\tau = (q_\tau(\mathbf{s}_1), \ldots, q_\tau(\mathbf{s}_n))^{\mathrm{T}}$, and $\boldsymbol{\phi} = \{\phi_{\mathbf{s}_i,\mathbf{s}_j}\}_{i,j=1}^n$, indicating how $q_\tau(\mathbf{s}_i)$ depends on the rest of $q_\tau(\mathbf{s}_j)$. Assume $N_{\mathbf{s}_i}$ is selected such that $\operatorname{rank}(\boldsymbol{\phi}) = n$, then $\mathbf{q}_\tau = (\mathbf{I}_n - \boldsymbol{\phi})^{-1}\boldsymbol{\theta}$. Let $\mathbf{L}^{\mathrm{T}} = (\mathbf{I}_n - \boldsymbol{\phi})^{-1}$, and $\boldsymbol{\ell}_i$ be the ith column of $\mathbf{L}$, then $q_\tau(\mathbf{s}_i) = \boldsymbol{\ell}_i^{\mathrm{T}}\boldsymbol{\theta}$. The objective function (1) can be written as

$$\sum_{i=1}^n \rho_\tau(Y(\mathbf{s}_i) - \boldsymbol{\ell}_i^{\mathrm{T}}\boldsymbol{\theta}) + \lambda \sum_{i=1}^n \omega_{\mathbf{s}_i} |\theta_i|. \tag{2}$$

The adaptive weights $\omega_{\mathbf{s}_i}$ are chosen to be equal to $\min(|\tilde{\theta}_i|^{-1}, \sqrt{n})$, where $\tilde{\boldsymbol{\theta}}$ is the initial estimator that can be taken as any consistent estimator of the rate $\sqrt{n/p_0 \log n}$, where p_0 is the number of nonzero parameters in the truth of $\boldsymbol{\theta}$, denoted as $\boldsymbol{\theta}^*$ (see Section 2.4 for more details). Throughout our numerical studies, we calculate $\tilde{\boldsymbol{\theta}}$ as the consistent estimator from Belloni and Chernozhukov (2011), that is, the minimizer of (2) with $\omega_{\mathbf{s}_i} = 1$. The general idea of the adaptive weights is that instead of applying the same penalty, we assign larger penalties to adjacent observations that are more similar and discourage the changes in $q_\tau(\mathbf{s}_i)$, so that the differences are shrunk toward zero faster. If the neighboring observations are very different, a smaller weight is assigned, leading to a larger change in $q_\tau(\mathbf{s}_i)$. In practice, the tuning parameter λ can be selected by minimizing the Bayesian information criterion (BIC; Schwarz 1978):

$$\mathrm{BIC}(\lambda) = \log\left\{ \sum_{i=1}^n \rho_\tau\left(Y(\mathbf{s}_i) - \boldsymbol{\ell}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}\right)\right\} + k\log(n), \tag{3}$$

where the estimated quantile $\hat{q}_\tau(s) = \boldsymbol{\ell}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}$, and $k = \sum_{i=1}^n I\{\hat{\theta}_i \neq 0\}$ is a measure of the fitted model complexity for a given tuning parameter λ. The form of the penalty term in (1) is determined by the choice of neighbors, $N_{\mathbf{s}_i}$ for each $\mathbf{s}_i$, $i = 1, \ldots, n$. For example, if only one neighbor is selected for each $\mathbf{s}_i$, the penalty term reduces to $\sum_{i=1}^n |q_\tau(\mathbf{s}_i) - q_\tau(\mathbf{s}^i)|$, where $\mathbf{s}^i$ can be chosen as the nearest neighbor of $\mathbf{s}_i$. For data on a regular grid, each $\mathbf{s}_i$ has four neighbors with common borders. One example is to choose the four nearest neighbors in the penalty term, encouraging the smoothness over space. This neighborhood structure is analogous to the conditionally autoregressive (CAR) model for spatial data, which is the counterpart of time series AR(1) model, both having Markov properties. Other metrics can be adopted for the choice of neighbors as well, although the optimal number of neighbors depends on the underlying dependence structure.

2.3 Temporal Fused Adaptive Lasso

Suppose we observe a time series $\{Y(t) : t = 1, \ldots, n\}$. We propose to estimate the τth quantile function $q_\tau(t)$ by a temporal FAL that minimizes

$$\sum_{t=1}^n \rho_\tau(Y_t - q_\tau(t)) + \lambda \sum_{t=2}^n \omega_t |q_\tau(t) - q_\tau(t-1)|. \tag{4}$$

The temporal FAL can be viewed as a one-dimensional case of the spatial FAL with the set of neighbors for $q_\tau(t)$ chosen to be the past nearest neighbor $q_\tau(t-1)$, similar to AR(1) model. The objective function (4) is also of the form (2), where $\theta_t = q_\tau(t) - q_\tau(t-1)$ for $t = 1, \ldots, n$ with $q_\tau(0) = 0$, and $\mathbf{L}^{\mathrm{T}}$ is a $n \times n$ lower triangular matrix with 1 on and below the diagonal, which is invertible and provides an one-to-one transformation from $\boldsymbol{\theta}$ to $\mathbf{q}_\tau$. The tuning parameter λ can be selected by minimizing the BIC in (3), where the estimated quantile $\hat{q}_\tau(t) = \boldsymbol{\ell}_t^{\mathrm{T}}\hat{\boldsymbol{\theta}} = \sum_{i=1}^t \hat{\theta}_i$, and $k = \sum_{t=2}^n I\{\hat{q}_\tau(t) \neq \hat{q}_\tau(t-1)\}$ is the number of distinct $\hat{q}_\tau(t)$ associated with λ.

2.4 Asymptotic Properties

In this section we establish the oracle properties (Fan and Li 2001) of the spatial and temporal FAL estimators $\hat{q}_\tau(\mathbf{s})$ and $\hat{q}_\tau(t)$ as defined in Sections 2.2 and 2.3. For easy presentation, we consider a general setup. Suppose we observe data from the

linear model

$$y_i = \boldsymbol{\ell}_i^{\mathrm{T}}\boldsymbol{\theta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where $\{\epsilon_i\}$ are independent with τth quantile zero. Let $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_{\mathcal{A}}^{*\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{A}^c}^{*\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ denote the truth, where $\boldsymbol{\theta}_{\mathcal{A}}^{*\mathrm{T}}$ and $\boldsymbol{\theta}_{\mathcal{A}^c}^{*\mathrm{T}}$ are the vectors of nonzero and zero coefficients respectively, $\mathcal{A} = \{j : \theta_j \neq 0\}$, $|\mathcal{A}| = p_0$ and $|\mathcal{A}^c| = p_n - p_0$. For fixed p, Wu and Liu (2009) established the oracle properties of the adaptive Lasso estimator. For increasing dimension p_n, under general regularity conditions (Belloni and Chernozhukov 2011), Zheng, Gallagher, and Kulasekera (2013) demonstrated the oracle properties of adaptive Lasso estimator with three additional mild conditions on the behavior of the covariates and regression coefficients. In our setting for $\theta_i = q_\tau(\mathbf{s}_i) - \frac{1}{n_{\mathbf{s}_i}}\sum_{j\in N_{\mathbf{s}_i}} q_\tau(\mathbf{s}_j)$, or $\theta_t = q_\tau(t) - q_\tau(t-1)$, the dimension $p_n = n$ grows as n increases with $|\mathcal{A}| = p_0 \ll n$ and $|\mathcal{A}^c| = n - p_0$. We consider the case when there is a one-to-one transformation from $\mathbf{q}_\tau$ to $\boldsymbol{\theta}$, and thus present the theorem in terms of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}$ be the minimization solution of (1) or (4), and $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ be the estimates of the nonzero coefficients. The FAL estimator $\hat{\boldsymbol{\theta}}$ enjoys the following oracle properties:

Theorem 1. Suppose assumptions **A1–A5** stated in Appendix A hold. If $\lambda_n/(\sqrt{p_0}\log n) \to \infty$, and $\lambda_n p_0^{3/2}/\sqrt{n} \to \infty$, then the spatial or temporal FAL estimator $\hat{\boldsymbol{\theta}}$ has the following three properties as $n \to \infty$,

1. Consistency in selection: $P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = \mathbf{0}) \to 1$;
2. Estimation consistency: $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| = O_p(\sqrt{p_0/n})$;
3. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \Sigma_1^{-1}\tau(1-\tau)\Sigma_{\mathcal{A}}\Sigma_1^{-1})$, where $\Sigma_{\mathcal{A}} = n^{-1}\lim_{n\to\infty}\sum_{i=1}^n \boldsymbol{\ell}_{ia}\boldsymbol{\ell}_{ia}^{\mathrm{T}}$ is the top-left $p_0 \times p_0$ submatrix of a positive definite matrix $\Sigma = n^{-1}\lim_{n\to\infty}\sum_{i=1}^n \boldsymbol{\ell}_i\boldsymbol{\ell}_i^{\mathrm{T}}$, with $\boldsymbol{\ell}_{ia}$ indicating the elements of $\boldsymbol{\ell}_i = (\ell_{i1}, \ldots, \ell_{in})^{\mathrm{T}}$ corresponding to the nonzero $\boldsymbol{\theta}_{\mathcal{A}}^*$, $\Sigma_1 = n^{-1}\lim_{n\to\infty}\sum_{i=1}^n f_{\epsilon_i}(0)\boldsymbol{\ell}_{ia}\boldsymbol{\ell}_{ia}^{\mathrm{T}}$ is assumed to be positive definite, with $f_{\epsilon_i}(0) \equiv f_i(q_\tau(i))$ and f_i denoting the density of y_i.

The proof and examples on the verification of conditions are provided in the supplementary material.

2.5 Computation

For a given tuning parameter λ, the minimization of the proposed FAL objective function can be formulated as a linear programming problem and solved efficiently by any existing linear programming software. In our numerical studies, we use the R (R Core Team 2014) function "`rq.fit.sfn`" in the package `quantreg`. This function adopts the sparse Frisch–Newton interior-point algorithm and the computational time is proportional to the number of nonzero entries in the design matrix (Koenker and Ng 2003, 2005). For example, in the temporal FAL, when τ is equal to 0.5 and corresponds to the median, $\rho_\tau(Y_t - q_\tau(t)) = 0.5|Y_t - q_\tau(t)|$. The minimization of (4) is equivalent to the method of least absolute deviations that seeks to minimize $\|\mathbf{Y} - \mathbf{D}\mathbf{q}_{0.5}\|_1$, where $\mathbf{Y} = (Y_1, \ldots, Y_n, 0, \ldots, 0)^{\mathrm{T}}$ with $n-1$ zero elements, and $\mathbf{D}$ is a $(2n-1)\times n$ augmented design matrix for $\mathbf{q}_{0.5} = (q_{0.5}(1), \ldots, q_{0.5}(n))^{\mathrm{T}}$,

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_n \\ \lambda\mathbf{H}_{(n-1)\times n} \end{pmatrix},$$

with $\mathbf{I}_n$ denoting the $n \times n$ identity matrix and

$$\mathbf{H}_{(n-1)\times n} = \begin{pmatrix} -\omega_2 & \omega_2 & & & \\ & -\omega_3 & \omega_3 & & \\ & & \ddots & \ddots & \\ & & & -\omega_n & \omega_n \end{pmatrix}.$$

For a general quantile level $0 < \tau < 1$, ρ_τ can still be formulated as a linear programming problem and solved by the sparse interior point algorithm (Section 6.8 of Koenker 2005, Wang and Hu 2011). For the design matrix $\mathbf{D}$ with $3n-2$ nonzero entries, the minimization takes about 86 sec per λ value for n = 10,000 using R on a PC with 4.0 GB of RAM. Under the same setting but n = 10,000 observations on a 100×100 grid, the design matrix for the spatial FAL is denser than the temporal case and it takes about 90 sec per λ value to minimize (1).

3. SIMULATION STUDIES

In our simulation studies, we generate data from various nonstationary and stationary models, and focus on quantile estimation at high quantile levels. For nonstationary models where the true quantile function is known, we evaluate the mean and standard deviation of the mean integrated squared errors (MISE). For stationary models where the joint quantile function is unknown, we investigate whether the estimated quantile function leads to the nominal quantile level marginally, and evaluate how much dependence among quantiles can be captured by different methods.

3.1 Temporal Fused Adaptive Lasso

First, we simulate nonstationary time series $Y(t) = \sigma(t)Z(t)$, where $Z(t)$'s are iid $N(0,1)$, and $\sigma(t) = 0.5\sum_{i=1}^n\{I(t \le n/3) + I(t \le 2n/3) + I(t \le n)\}$. The τth quantile function is estimated at $t = 1, \ldots, n$, where $n = 1000$, by the temporal fused Lasso (FL) and fused adaptive Lasso (FAL). Figure 1 shows the observations and the estimated 95% quantile function estimated by the temporal FAL from one simulation. For each simulation, we compute the MISE of the quantile estimates and evaluate the mean and the standard deviation of the MISE with 1000 replications. Table 1 summarizes the mean and standard deviation of the MISE for $\tau = 0.93, 0.94, 0.95, 0.96, 0.97$. For these high quantile levels, both the FL and FAL have small MISEs, and the adaptive version performs a little better due to the more localized penalty term, especially for higher quantile levels where the change over t is relatively big.

Then, we generate data from an AR(1) process given by $Y(t) = \alpha Y(t-1) + \epsilon(t)$, $t = 1, \ldots, n,$, where $\epsilon(t)$ has a standard normal distribution. Larger values of the parameter $|\alpha| \in (0,1)$ indicate longer range of dependence in the time series. In our simulations, we set $\alpha = 0.1, 0.5, 0.9$ and $n = 1000$. Figure 2 shows the 95% quantile estimation for one realization as an example. It shows that the marginal quantile estimator tends to be smaller than the FL estimates, and that the FAL

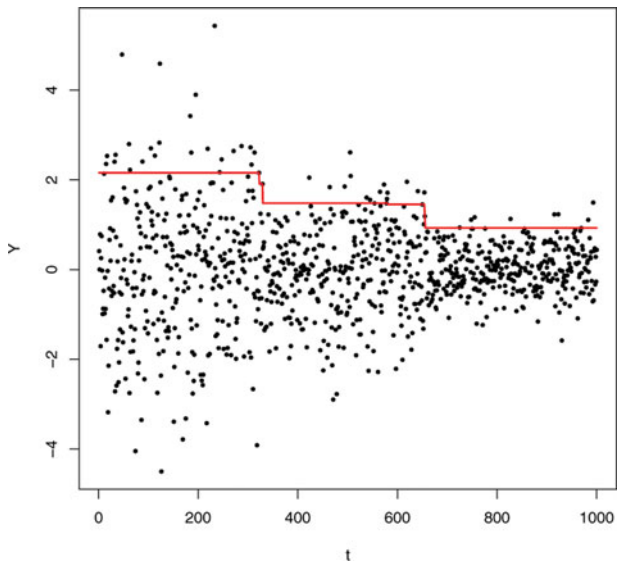

Figure 1. Nonstationary time series plot: black dots represent observations and solid lines indicate the 95% quantile function estimated by the temporal FAL.

captures more local dependence by applying different penalties more locally. To assess the performance of different methods, we define an indicator function $\delta(t) = I\{Y(t) \le q_\tau(t)\}$. Then, for a given $t = 1, \ldots, n$, we have $E(\delta(t)) = P(Y(t) \le q_\tau(t)) = \tau$, $\text{var}(\delta(t)) = \tau(1-\tau)$, and the correlation

$$\gamma(h) = \text{corr}(\delta(t), \delta(t+h)) \tag{6}$$

which measures the temporal dependence among quantiles for $t = 1, \ldots, n-h$.

In the simulation studies, $q_\tau(t)$ is estimated by the marginal quantile estimator (MQ), the sample quantile estimator (SQ), as well as the FL and the FAL methods. We then repeat the simulation $N = 1000$ times and calculate $\hat{\tau} = \frac{1}{nN}\sum_{i=1}^{N}\sum_{t=1}^{n}\hat{\delta}(t)$, where $\hat{\delta}(t) = I\{Y(t) \le \hat{q}_\tau(t)\}$. Let s_δ denote the sample standard deviation of $\{\hat{\delta}(t),\ t = 1, \ldots, n\}$. When a quantile estimator captures the dependence well, $\hat{\delta}(t)$'s are correlated and the sample standard deviation s_δ calculated from the correlated $\hat{\delta}(1), \ldots, \hat{\delta}(n)$ is larger than the true value $\sqrt{\tau(1-\tau)}$. The values of $\hat{\tau}$ and s_δ, for $\alpha = 0.1, 0.5, 0.9$, are shown in Table 2, with $\tau = 0.95, 0.97$, and thus $\sqrt{\tau(1-\tau)} = 0.218, 0.171$. We can see that $\hat{\tau}$ estimated by the MQ, SQ, FL, and FAL methods are all close to the true values of τ. Therefore, all the four methods preserve the nominal quantile levels marginally. However, for the MQ and SQ methods, s_δ is equal to the true value $\sqrt{\tau(1-\tau)}$ and does not change with α, which indicates that $\hat{\delta}(1), \ldots, \hat{\delta}(n)$ are close to independence and the temporal dependence is not captured well by these two methods. In contrast, the FL and FAL methods show slightly larger s_δ when the correlation α gets stronger.

Next, we estimate the correlation, $\gamma(h)$, N times by the sample correlation of $\{\delta(t), \delta(t+h)\}_{t=1,\ldots,n-h}$, $h = 1, 2, 3, 4$. For high quantile values, the correlation $\gamma(h)$ defined in (6) is generally smaller than α^h, the correlation between $Y(t)$ and $Y(t+h)$. A good quantile estimator should be able to capture the correlation in $Y(t)$ as much as possible, in addition to preserving the nominal quantile level τ. Therefore, smaller values of $\alpha^h - \hat{\gamma}(h)$ indicate better performance when $\hat{\tau}$ is close to τ. The mean and standard deviation of $\alpha^h - \hat{\gamma}(h)$, for $\alpha = 0.1, 0.5, 0.9$ and $h = 1, 2, 3, 4$, are shown in Table 2, with $\tau = 0.95, 0.97$. Compared to the FL and FAL methods, the larger values of $\alpha^h - \hat{\gamma}(h)$ for the MQ and SQ imply that these two methods do not capture the dependence in the upper quantiles well, especially when α is large. This is not surprising because the MQ method only estimates quantiles marginally, and the SQ method does not penalize the difference among neighboring quantiles. Therefore, the FL and FAL outperform the MQ and SQ by adding the penalty term to account for the local dependence. Moreover, the values of $\alpha^h - \hat{\gamma}(h)$ for $\tau = 0.97$ tend to be larger than those for $\tau = 0.95$, suggesting that it is more difficult to capture the dependence for quantiles at higher levels.

3.2 Spatial Fused Adaptive Lasso

We consider spatial data generated from a mixture of max-stable processes on a grid. Max-stable processes arise from an infinite-dimensional generalization of extreme value theory. They form a natural class of processes when sample block maxima are observed at each site of a spatial process and characterize the dependence among extremes. The Schlather model (Schlather 2002) is a flexible class of max-stable processes by taking $Z_i(\mathbf{s})$ to be any stationary Gaussian random field with finite expectation. A stationary max-stable process with unit-Fréchet margins can be obtained by $Y(\mathbf{s}) = \max_i[\zeta_i \max\{0, Z_i(\mathbf{s})\}]$, where $\mu = E\{\max(0, Z_i(\mathbf{s}))\} < \infty$ and $\{\zeta_i\}$ denote the points of a Poisson process on $(0, \infty)$ with intensity measure $\mu^{-1}\zeta^{-2}d\zeta$. If the random process is specified for a stationary isotropic Gaussian random field $Z_i(\cdot)$ with unit variance, correlation $K(\cdot)$ and $\mu^{-1} = \sqrt{2\pi}$, then the bivariate marginal distributions are given by

$$P\{Y(\mathbf{s}_1) \le y_1, Y(\mathbf{s}_2) \le y_2\} = \exp\left\{-\frac{1}{2}\left(\frac{1}{y_1}+\frac{1}{y_2}\right)\left(1+\sqrt{1-2(K(h)+1)\frac{y_1y_2}{(y_1+y_2)^2}}\right)\right\},$$

Table 1. The mean and standard deviation (in parentheses) of the MISE for different τ

	τ				
	0.93	0.94	0.95	0.96	0.97
FL	0.0405 (0.0240)	0.0469 (0.0272)	0.0600 (0.0327)	0.0770 (0.0415)	0.1383 (0.0813)
FAL	0.0398 (0.0194)	0.0434 (0.0213)	0.0480 (0.0225)	0.0542 (0.0259)	0.0629 (0.0312)

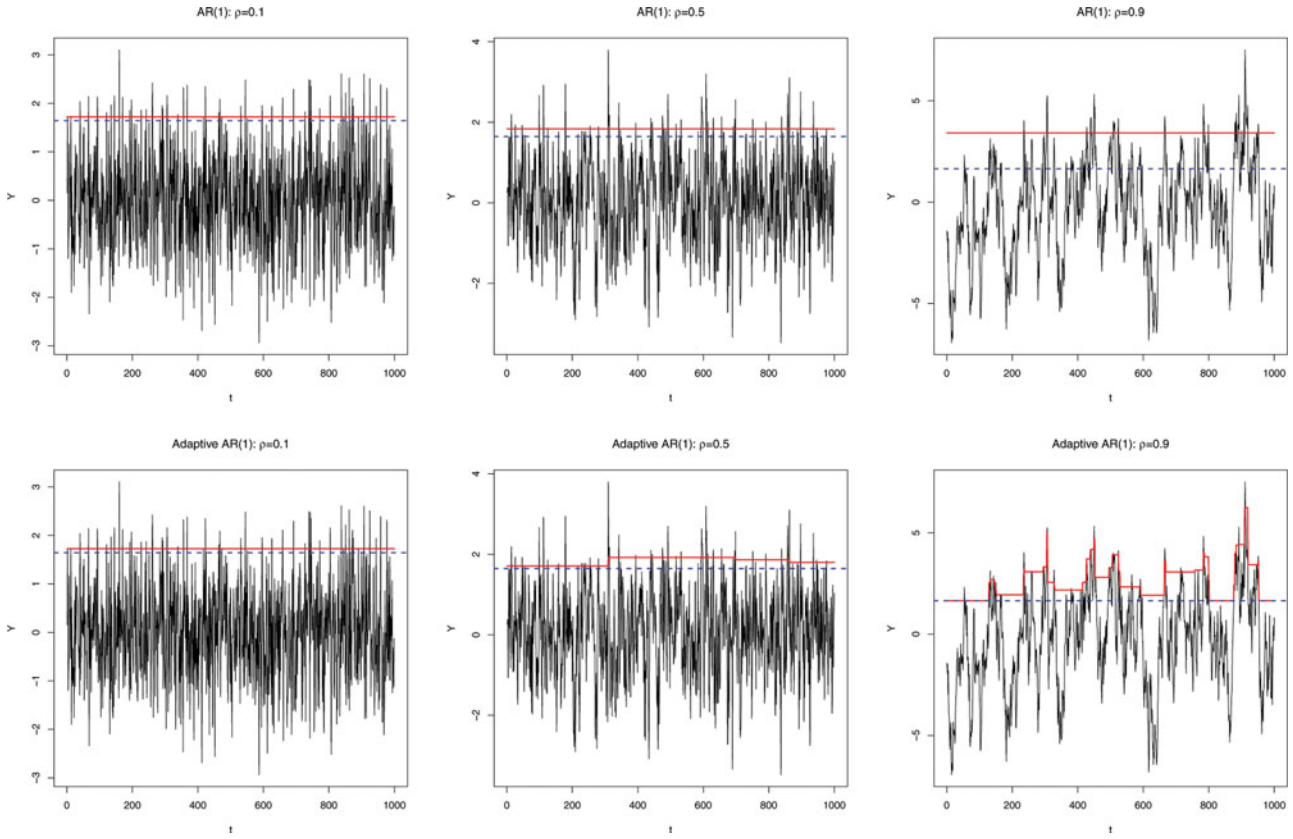

Figure 2. Dashed lines denote the 95% marginal quantile function estimation. Solid lines indicate the temporal FL estimation of the 95% quantile function in the top panels, and present the FAL estimation in the bottom panels.

Table 2. The estimated τ, s_δ, the mean and standard deviation (in parentheses) of $\alpha^h - \hat{\gamma}(h)$, for $\alpha = 0.1, 0.5, 0.9$ and $h = 1, 2, 3, 4$, with $\tau = 0.95, 0.97$

α	Method	$\tau = 0.95$	s_δ	$h = 1$	2	3	4
0.1	MQ	0.953	0.218	0.076 (0.037)	0.010 (0.032)	0.001 (0.032)	0.000 (0.031)
	SQ	0.953	0.218	0.078 (0.037)	0.011 (0.032)	0.002 (0.032)	0.000 (0.032)
	FL	0.953	0.218	0.037 (0.032)	0.006 (0.031)	0.001 (0.030)	0.000 (0.032)
	FAL	0.953	0.218	0.037 (0.032)	0.005 (0.031)	0.000 (0.030)	0.000 (0.032)
0.5	MQ	0.951	0.218	0.303 (0.062)	0.177 (0.049)	0.092 (0.043)	0.049 (0.036)
	SQ	0.951	0.218	0.303 (0.059)	0.178 (0.048)	0.092 (0.042)	0.049 (0.036)
	FL	0.951	0.219	0.170 (0.032)	0.095 (0.033)	0.048 (0.036)	0.024 (0.035)
	FAL	0.950	0.220	0.171 (0.032)	0.089 (0.035)	0.042 (0.038)	0.017 (0.037)
0.9	MQ	0.951	0.218	0.307 (0.099)	0.367 (0.124)	0.368 (0.124)	0.372 (0.123)
	SQ	0.951	0.218	0.304 (0.077)	0.359 (0.097)	0.362 (0.103)	0.369 (0.102)
	FL	0.950	0.230	0.222 (0.039)	0.260 (0.065)	0.262 (0.073)	0.257 (0.075)
	FAL	0.950	0.252	0.267 (0.031)	0.206 (0.047)	0.202 (0.052)	0.190 (0.056)
α	Method	$\tau = 0.97$	s_δ	$h = 1$	2	3	4
0.1	MQ	0.970	0.171	0.085 (0.039)	0.009 (0.033)	0.002 (0.031)	0.000 (0.032)
	SQ	0.970	0.171	0.084 (0.038)	0.008 (0.034)	0.002 (0.032)	0.000 (0.032)
	FL	0.970	0.171	0.038 (0.032)	0.006 (0.031)	0.002 (0.031)	0.000 (0.031)
	FAL	0.970	0.171	0.037 (0.032)	0.004 (0.032)	0.000 (0.031)	0.000 (0.031)
0.5	MQ	0.970	0.171	0.335 (0.074)	0.193 (0.054)	0.102 (0.044)	0.053 (0.037)
	SQ	0.970	0.171	0.335 (0.069)	0.193 (0.053)	0.102 (0.044)	0.053 (0.036)
	FL	0.970	0.172	0.180 (0.032)	0.102 (0.035)	0.052 (0.036)	0.025 (0.035)
	FAL	0.970	0.175	0.178 (0.033)	0.093 (0.037)	0.044 (0.038)	0.018 (0.038)
0.9	MQ	0.970	0.171	0.354 (0.142)	0.402 (0.152)	0.416 (0.155)	0.422 (0.144)
	SQ	0.970	0.171	0.342 (0.098)	0.404 (0.118)	0.407 (0.125)	0.414 (0.118)
	FL	0.968	0.201	0.237 (0.045)	0.287 (0.081)	0.293 (0.089)	0.287 (0.088)
	FAL	0.968	0.219	0.274 (0.033)	0.209 (0.049)	0.204 (0.054)	0.193 (0.059)

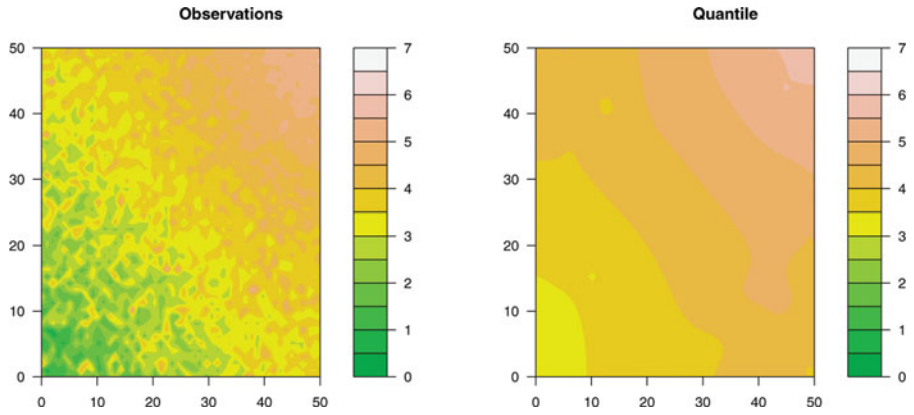

Figure 3. Left panel: the simulated observations. Right panel: the estimated 95% quantile by the spatial FAL.

where h is the Euclidean distance between locations $\mathbf{s}_1$ and $\mathbf{s}_2$, and the correlation $K(\cdot)$ is chosen from one of the valid families of correlations for Gaussian processes.

Among many available covariance models, the Matérn family (Matérn 1986) has gained widespread interest in recent years. Handcock and Stein (1993) introduced the Matérn family of spatial correlations into statistics as a flexible parametric class with one parameter determining the smoothness of the underlying spatial random field. The Matérn class of covariance functions is defined as $C(h) = \frac{2\sigma^2}{\Gamma(\nu)}(\frac{h}{2\eta})^\nu K_\nu(\frac{h}{\eta})$, where $\nu > 0$ is a smoothness parameter which measures the smoothness of the random field, $\eta > 0$ is a range parameter which measures how quickly the correlation decays with distance, and $K_\nu(\cdot)$ denotes the modified Bessel function of order ν.

In our simulations, we generate spatial data from a mixture of two Schlather processes F_1 and F_2 on a 50×50 grid. We choose the Matérn covariance structure for $K(\cdot)$ in the Schlather model with $\sigma = 1$, and let $Y(\mathbf{s}) = h_0^{-1}F_1(\mathbf{s}) + h_1^{-1}F_2(\mathbf{s})$, where the process F_1 has parameters $\nu = 0.5$, $\eta = 0.01$, and $F_2(\mathbf{s}) = F_1(\mathbf{s}) + 5$. The observations are the weighted average of data generated from F_1 and F_2, where the weight h_0^{-1} is the inverse distance of $\mathbf{s}$ to $(0, 0)$ and h_1^{-1} is the inverse distance of $\mathbf{s}$ to (n, n). Figure 3 shows the simulated spatial observations and the spatial FAL estimation of the 95% quantile function. Comparing the observations and the estimated 95% quantile surface, we can see that the FAL estimator captures the increasing pattern from $(0, 0)$ to (n, n), and obtains the 95% quantile surface by smoothing out the original observations. Similar to the time series case, we repeat the simulation 1000 times, and the summary of $\hat{\tau}$ is shown in Table 3 with $\tau = 0.95$. Since the mean and median are close to the true value 95%, the spatial FAL preserves the nominal quantile level reasonably well.

Table 3. The summary statistics of $\hat{\tau}$ for $\tau = 0.95$

	Min	Q_1	Median	Mean	Q_3	Max
Spatial FAL	0.9260	0.9450	0.9500	0.9498	0.9550	0.9740

In Figure 4, we generate "hot spots" in simulated time series or spatial data by adding abrupt changes which can be found in many applications. A good quantile function estimator should be able to correctly characterize the changing tail probabilities due to extreme large observations. We choose the 95% quantile level to examine how the FAL quantile estimates change across time or space. Figure 4 shows that the estimated 95% quantile functions successfully capture the abrupt changes in time (left panel) and space (middle and right panels), so that the changing tail behavior of the distribution is characterized and the hot spot regions are identified.

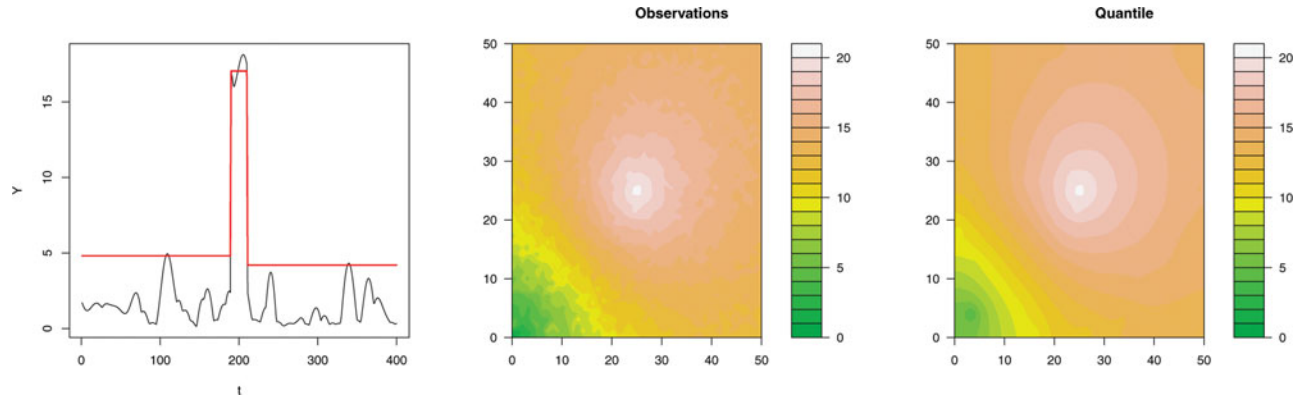

Figure 4. Left panel: the estimated 95% quantile (solid line) by the temporal FAL imposed on the simulated time series. Middle panel: the simulated spatial observations. Right panel: the estimated 95% quantile by the spatial FAL for the simulated spatial observations.

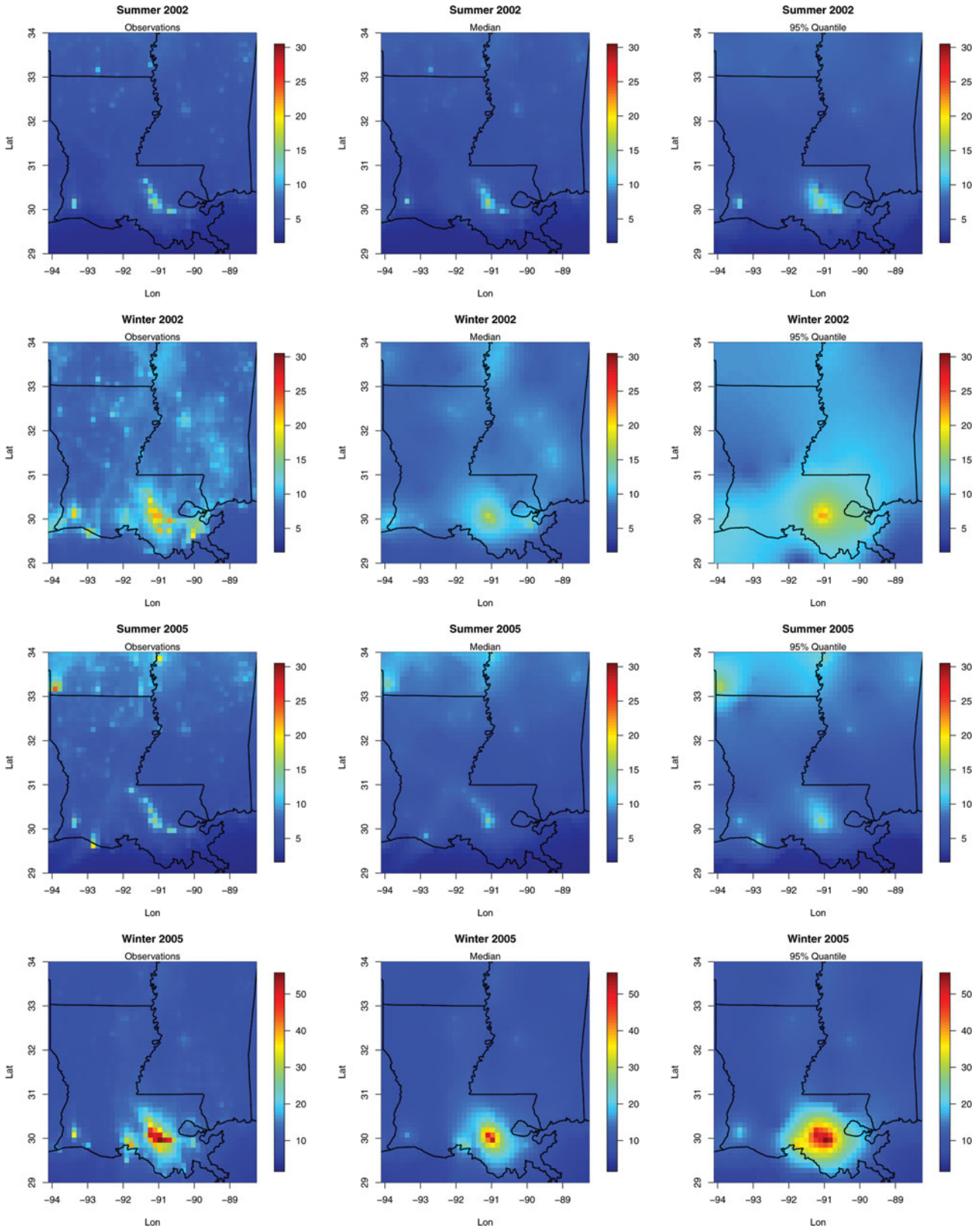

Figure 5. Observations, median, and 95% quantile of the $PM_{2.5}$ concentrations for Louisiana and Mississippi estimated by spatial FAL in the winter and summer of 2002 and 2005.

4. APPLICATIONS TO AIR QUALITY DATA

Concentrations of the particle matter (PM) that exceed regulated levels are believed to contribute significantly to adverse health effects. Particles less than 2.5 μmin diameter ($PM_{2.5}$) are referred to as "fine" particles. PM can be suspended in the air for long periods of time and has been known as a major risk factor of human health, among which $PM_{2.5}$ are believed to pose the greatest health risks since they are so small and can be inhaled into, and accumulate in, the respiratory system. Numerous

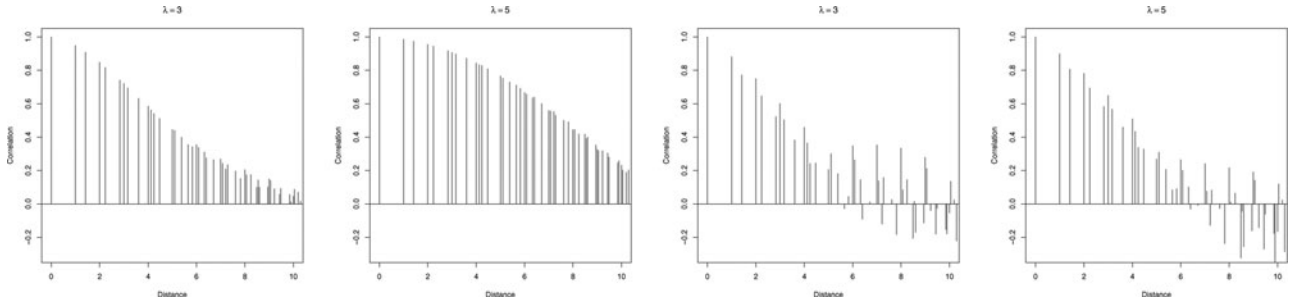

Figure 6. Top panel: the correlation of $\hat{q}_\tau$ as a function of distance for $\lambda = 1, 5$ under the penalty of type (1). Bottom panel: the correlation of $\hat{q}_\tau$ as a function of distance for $\lambda = 1, 5$ under the penalty of type (4).

epidemiological studies have indicated that exposure to $PM_{2.5}$ is associated with asthma, respiratory infections, lung cancer, cardiovascular problems, and premature death (Peters et al. 2001; Pope et al. 2002; Hu, Liebens, and Rao 2008).

Hurricane Katrina, a Category 5 hurricane of the 2005 Atlantic hurricane season on August 29, caused catastrophic damage along the coastlines of Louisiana, Mississippi, and Alabama. The U.S. Environmental Protection Agency (EPA) examined the effects of Katrina to determine whether associated flooding of Louisiana and Mississippi and subsequent cleanup efforts caused the changes in emissions of some air pollutants in the affected areas. Post-Katrina and Pre-Katrina concentrations (5th, 25th, 50th, 75th, 95th percentiles) from monitoring sites were compared. The EPA technical report (2008) pointed out that among pollutants with concentrations that showed statistically significant differences, concentrations of $PM_{2.5}$ were higher than previously measured values at the Gulfport/Pascagoula sites and in New Orleans. We are interested in estimating the median surface and the 95% quantile surface of the averaged daily $PM_{2.5}$ total mass in Louisiana and Mississippi Pre-Katrina and Post-Katrina from the Community Multiscale Air Quality (CMAQ) model developed by EPA for air quality management. More detailed information about CMAQ and its development can be found at the EPA website (*http://www.epa.gov/AMD/CMAQ/*).

We consider four time periods: 2002 Summer (June 1, 2002 to August 31, 2002), 2002 Winter (December 1, 2002 to February 29, 2002), 2005 Summer (June 1, 2005 to August 31, 2005), and 2005 Winter (December 1, 2005 to February 28, 2005), under 12 km × 12 km spatial resolution with 2500 pixels for the region of study. Then, we average the daily $PM_{2.5}$ at each location within each time period.

For the year 2002 and the year 2005, the observations from CMAQ, median, and 95% quantile estimated by the spatial FAL are shown in Figure 5. For both years, the $PM_{2.5}$ total mass was higher in winter than in summer for most parts of Louisiana and Mississippi. For the year 2002, we can see that the difference between the summer and winter medians across space is much smaller than that between the 95% quantiles, which suggests that the main difference between summer and winter is in the upper tail. In other words, there are more extreme values in the $PM_{2.5}$ total mass in winter. For the year 2005, the winter period clearly shows a significant increase in the $PM_{2.5}$ total mass after hurricane Katrina, which agrees with the conclusion from the monitoring sites in the EPA technical report (2008). As illustrated by the simulation studies in Section 3, our method with the FAL penalty accommodates the spatial dependence in the high concentrations of the $PM_{2.5}$ total mass, and the "hot spots" regions are also identified.

Table 4. The estimated quantile level, MAE and VAR by leave-one-out cross-validation and the mean ($\bar{x}$) and standard deviation (s) of MAE by four-fold cross-validation

$\tau = 0.5$	2002 Summer	2002 Winter	2005 Summer	2005 Winter
$\hat{\tau}$	0.5373	0.5056	0.5208	0.5074
MAE	0.1283	0.2910	0.1974	0.3665
VAR	0.0780	0.2906	0.2269	1.0300
$\bar{x}$	0.1147	0.3267	0.1951	0.3282
s	0.0117	0.0215	0.0665	0.0273
$\tau = 0.95$	2002 Summer	2002 Winter	2005 Summer	2005 Winter
$\hat{\tau}$	0.9352	0.9388	0.9258	0.9232
MAE	0.0921	0.1785	0.2777	0.2341
VAR	0.1399	0.2704	0.8402	1.2454
$\bar{x}$	0.0795	0.1979	0.1431	0.2580
s	0.0127	0.0161	0.0600	0.0315

To quantify the uncertainties of the spatial FAL estimator, we use cross-validation by leaving out one observation at a time, predicting the left-out observation by the average of its neighboring estimates. Besides estimating the nominal τ, we define the mean absolute error as $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} e_i$, where $e_i = \rho_\tau(y_i - \hat{q}_\tau(i))$ and $\hat{q}_\tau(i)$ is the estimated τth quantile. For $\tau = 0.5$, it can be written as $\text{MAE} = \frac{1}{2n}\sum_{i=1}^{n} |y_i - \hat{q}_\tau(i)|$, where $2 \times \text{MAE}$ is the mean of the absolute prediction errors, which measures the difference between the prediction and the observed values. To assess the variability of the prediction errors, we compute the variance $\text{VAR} = \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2$, where $\bar{e}$ is the average of e_i's. The results are summarized in Table 4. For all cases, the estimated quantile levels are close to the nominal ones with reasonably small MAE and VAR.

To further consider the standard deviation of the MAE, we conduct a four-fold cross-validation by partitioning the domain into four subregions with equal size. The mean and standard deviation of the MAE are summarized in Table 4. Since the distribution of $PM_{2.5}$ is right-skewed with extreme values at certain locations, the MAE of the estimated median is larger than that of the estimated 95% quantile across space. Median is

a robust estimator for the center of the distribution, whereas the upper quantile should be considered when the tail behavior is of interest.

5. DISCUSSION

In this article, we proposed temporal and spatial FAL penalized quantile function estimators that can be easily adopted when temporal or spatial quantiles are of interest. The proposed method provides a nonparametric way to estimate quantile functions for time series or spatial data, which is essentially an optimization problem with an L_1 loss function plus an L_1 norm penalty. By penalizing the difference among quantiles in time or space, the local features are captured, assuming the nonstationarity of the underlying process. Our theory shows that the temporal and spatial FAL enjoy the oracle properties, that is, the consistency in selection, estimation consistency, and the asymptotic normality under mild conditions. From the parameterization, we can see that the temporal FAL can be viewed as a special case of the spatial FAL with a specific choice of neighbors. The proposed methods do not need replicates but borrow information from neighboring observations via nonparametric smoothing. In this article, we use smoothing to account for the dependence of data at neighboring time or space, and assume that after accounting for the spatial/temporal quantile function the residuals are independent. The theoretical study of penalized quantile regression with dependent errors is challenging especially in a high-dimensional setup; the topic itself deserves separate treatment. In the optimization procedure, the tuning parameter plays an important role. In this article, BIC is used to choose λ_n in (1). The theoretical properties for BIC were studied by Wang, Li, and Jiang (2007) for fixed dimensionality in least squares regression. Zheng, Gallagher, and Kulasekera (2013) discussed the challenges of the choice of λ_n in quantile regression when the dimension increases as n grows, and gave one possible construction based on the initial estimator used for the adaptive weights. Nonetheless, the theoretical study of BIC for penalized quantile regression in the high-dimensional case with $p \geq n$ requires future research. It is also worth noting that the temporal and spatial FAL have an L_1 loss function plus an L_1 norm penalty which facilitate the optimization. For the quantile function over space, an L_2 norm penalty can be considered if we want to encourage the smoothness rather than sparseness, but this requires more complex optimization; see similar discussions in Liu, Yuan, and Ye (2010).

To take spatio-temporal interaction into account, the FAL can be generalized to space–time data, $\{Y(\mathbf{s}_i, t) : i = 1, \ldots, m; t = 1, \ldots, n\}$, by defining a cylinder to determine the neighbor $N_{\mathbf{s}_i,t}$ for each observation $Y(\mathbf{s}_i, t)$. The fused Lasso penalty term becomes $\lambda \sum_{t=1}^{n} \sum_{i=1}^{m} |q_\tau(\mathbf{s}_i, t) - \frac{1}{n_{t,\mathbf{s}_i}} \sum_{\{t',j\} \in N_{\mathbf{s}_i,t}} q_\tau(\mathbf{s}_j, t')|$. Although it is not trivial to describe the interaction between space and time, we discuss some possible construction of the penalty term. For example, suppose we observe data $\{Y(i, j, t)\}$ at evenly spaced time points $t = 1, \ldots, n$, on a grid with grid points $\{(i, j)\}_{i=1,\ldots,m_1, j=1,\ldots,m_2}$ and $m = m_1 m_2$. Then, for interior locations, $N_{t,i,j}$ can be defined as a set consisting of four nearest neighbors of location (i, j) at time t, plus two adjacent time points $t - 1$ and $t + 1$ at location (i, j). The penalty for (t, i, j) is

$$\lambda \Big| q_\tau(i, j, t) - \frac{1}{6}\{q_\tau(i, j, t-1) + q_\tau(i, j, t+1) + \sum_{k=-1,1} \sum_{\ell=-1,1} q_\tau(i+k, j+\ell, t)\} \Big|.$$

We can also consider more neighbors by adding the four nearest neighbors at time $t - 1$ and $t + 1$ with a total number of 14 neighbors for (t, i, j). As discussed in Section 2.5, more selected neighbors will definitely require more computations due to the increasing number of nonzero entries in the design matrix. In this article, we have only considered equally spaced observations when constructing the penalty term. In fact, for unevenly spaced times series, the penalty term can be modified to accommodate the distance. For example, the penalty term in the temporal fused Lasso can be modified to $\lambda \sum_i \frac{1}{t_i - t_{i-1}} |q(t_i) - q(t_{i-1})|$. Moreover, our methods can also be generalized to quantile function estimation over multiple quantile levels by constructing penalty terms consisting of quantile parameters from different quantile levels.

The spatial and temporal fused Lasso methods can be viewed in a more general framework with possible predictors. Let $\mathbf{X}_i = (\, X_0 \; X_1 \; \ldots \; X_p \,)_i^{\mathrm{T}}$ be the vector consisting of p predictors with an intercept term X_0 for the ith observation. The corresponding coefficient vector is defined as $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$. Then, the fused Lasso criterion can be written as $\sum_i \rho_\tau(Y(\mathbf{r}_i) - q_\tau(\mathbf{X}_i)) + \lambda \sum_i |\theta_i|$, subject to $\boldsymbol{\theta} = \mathbf{H}\mathbf{q}_\tau$, where $\mathbf{q}_\tau = (q_\tau(\mathbf{X}_1), \ldots, q_\tau(\mathbf{X}_n))^{\mathrm{T}}$ and $\mathbf{H}$ defines different types of penalties. For the temporal fused Lasso without predictors, $\mathbf{r}_i = t$, $q_\tau(\mathbf{X}_i) = X_{0,t}\beta_0 = q_\tau(t)$, $\theta_t = (X_{0,t} - X_{0,t-1})\beta_0 = q_\tau(t) - q_\tau(t-1)$, and $\mathbf{H}$ is an $(n-1) \times n$ matrix as in Section 2.5. The spatial fused Lasso can be defined in a similar way.

Obviously, the quantile estimation depends on the choice of $\mathbf{H}$ defined in Section 2.5, and the optimal $\mathbf{H}$ should be determined by the underlying true process. To illustrate the effect of the penalty term, we apply the spatial FAL with two different penalties to a Gaussian white-noise process on a regular grid of size 30×30 for $\tau = 0.95$, then we examine the correlation of $\hat{q}_\tau$. We first consider the penalty defined in (1), where the nearest neighbors are selected for each $q_\tau(i)$, then we construct a penalty term similar to the temporal FAL defined in (4) by ordering $q(i, j)$ on the grid in a zigzag fashion, that is, $q(1, 1), \ldots, q(30, 1), q(30, 2), \ldots, q(1, 2), \ldots, q(1, 30)$, and penalizing the difference between two adjacent quantiles. Figure 6 shows the correlation of $\hat{q}_\tau$ as a function of distance for $\lambda = 1, 5$ under the penalties of type (1) (top) and of type (4) (bottom). For both penalties, larger values of λ lead to larger correlations. However, the correlation introduced by the penalty of type (4) has larger values at integer distances and is not as smooth as that for the penalty of type (1). In other words, the penalty of type (1) encourages the smoothness in space by penalizing the difference among nearest neighbors.

ACKNOWLEDGMENTS

The authors thank the valuable suggestions from the associate editor and referees for contributing to the noticeable improvement of the article. This research was partially supported by the US National Science Foundation (NSF)

grants DMS-1106862, 1106974, and 1107046, the NSF CAREER award DMS-1149355, and the STATMOS research network on Statistical Methods in Oceanic and Atmospheric Sciences.

[Received May 2013. Revised January 2015.]

REFERENCES

Belloni, A., and Chernozhukov, V. (2011), "L_1-Penalized Quantile Regression in High-Dimensional Sparse Models," *Annals of Statistics*, 39, 82–130. [129,130]

Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O., and Sun, Y. (2012), "A Survey of Spatial Extremes: Measuring Spatial Dependence and Modeling Spatial Effects," *REVSTAT*, 10, 135–165. [127,128]

Das, D., Ganguly, A., Chatterjee, S., Kumar, V., and Obradovic, Z. (2012), "Spatially Penalized Regression for Dependence Analysis of Rare Events: A Study in Precipitation Extremes," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International* 1948–1951, 22–27 July 2012. [128]

Draghicescu, D., Guillas, S., and Wu, W. (2009), "Quantile Curve Estimation and Visualization for Non Stationary Time Series," *Journal of Computational and Graphical Statistics*, 18, 1–20. [128]

Eilers, P. H. C. (2003), "A Perfect Smoother," *Analytical Chemistry*, 75, 3631–3636. [128]

Eilers, P., and de Menezes, R. (2005), "Quantile Smoothing of Array CGH Data," *Bioinformatics*, 21, 1146–1153. [128]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [129]

Furrer, E. M., and Nychka, D. W. (2007), "A Framework to Understand the Asymptotic Properties of Kriging and Splines," *Journal of the Korean Statistical Society*, 36, 57–76. [128]

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410. [133]

Harchaoui, Z., and Lévy-Leduc, C. (2010), "Multiple Change-Point Estimation With a Total Variation Penalty," *Journal of the American Statistical Association*, 105, 1480–1493. [128]

Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005), "Detection of DNA Copy Number Alterations Using Penalized Least Squares Regression," *Bioinformatics*, 21, 3811–3817. [128]

Hu, Z., Liebens, J., and Rao, K. R. (2008), "Linking Stroke Mortality With Air Pollution, Income, and Greenness in Northwest Florida: An Ecological Geographical Study," *International Journal of Health Geographics*, 7, 7–20. [135]

Jiang, L., Wang, H., and Bondell, H. (2013), "Interquantile Shrinkage in Regression Models," *Journal of Computational and Graphical Statistics*, 22, 970–986. [128]

Koenker, R. (2005), *Quantile Regression*, London: Cambridge University Press. [130]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [128]

Koenker, R., and Ng, P. (2003), "SparseM: A Sparse Matrix Package for R," *Journal of Statistical Software*, 8, 1–9. [130]

Koenker, R., and Ng, P. (2005), "A Frisch-Newton Algorithm for Sparse Quantile Regression," *Journal Acta Mathematicae Applicatae Sinica (English Series)*, 21, 225–236. [130]

Li, Y., and Zhu, J. (2007), "Analysis of Array CGH Data for Cancer Studies Using Fused Quantile Regression," *Bioinformatics*, 23, 2470–2476. [128]

Liu, J., Yuan, L., and Ye, J. (2010), "An Efficient Algorithm for a Class of Fused Lasso Problems," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 323–332. [136]

Matérn, B. (1986), "Spatial Variation," in *Lecture Notes in Statistics* (Vol. 36, 2nd ed.), Berlin, New York: Springer-Verlag, p. 151. [133]

Nowak, G., Hastie, T., Pollack, J. R., and Tibshirani, R. (2011), "A Fused Lasso Latent Feature Model for Analyzing Multi-Sample aCGH Data," *Biostatistics*, 12, 776–791. [128]

Nychka, D. W. (1995), "Splines as Local Smoothers," *Annals of Statistics*, 23, 1175–1197. [128]

Parzen, E. (2004), "Quantile Probability and Statistical Data Modeling," *Statistical Science*, 19, 652–662. [127]

——— (2008), "United Statistics, Confidence Quantiles, Bayesian Statistics," *Journal of Statistical Planning and Inference*, 138, 2777–2785. [127]

Peters, A., Dockery, D. W., Muller, J. E., and Mittleman, M. A. (2001), "Increased Particulate Air Pollution and the Triggering of Myocardial Infarction," *Circulation*, 103, 2810–2815. [135]

Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002), "Lung Cancer, Cardiopulmonary Mortality, and Long-Term Exposure to Fine Particulate Air Pollution," *The Journal of the American Medical Association*, 287, 1132–1141. [135]

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org/*. [130]

Schlather, M. (2002), "Models for Stationary Max-Stable Random Fields," *Extremes*, 5, 33–44. [131]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464. [129]

Stein, M. L. (1991), "A Kernel Approximation to the Kriging Predictor of a Spatial Process," *Annals of the Institute of Statistical Mathematics*, 43, 61–75. [128]

Storlie, C., Bondell, H., Reich, B., and Zhang, H. (2011), "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property," *Statistica Sinica*, 21, 679–705. [128]

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [128]

Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [128]

Tibshirani, R. J., and Wang, P. (2008), "Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso," *Biostatistics*, 9, 18–29. [128]

U.S. EPA Air Quality Data Analysis Group (2008), "Ambient Air Quality After Hurricane Katrina," Technical Report, Contract No. EP-D-05-004. [135]

Wang, H., and Hu, J. (2011), "Identification of Differential Aberrations in Multiple-Sample Array CGH Studies," *Biometrics*, 67, 353–362. [128,130]

Wang, H., Li, G., and Jiang, G. (2007), "Robust Regression Shrinkage and Consistent Variable Selection via the LAD-Lasso," *Journal of Business and Economic Statistics*, 25, 347–355. [136]

Wu, Y., and Liu Y. (2009), "Variable Selection in Quantile Regression," *Statistica Sinica*, 19, 801–817. [128,130]

Zheng, Q., Gallagher, C., and Kulasekera, K. B. (2013), "Adaptive Penalized Quantile Regression for High-Dimensional Data," *Journal of Statistical Planning and Inference*, 143, 1029–1038. [130,136]