Nonparametric Estimation of Extreme Conditional Quantiles

Jan Beirlant *
Tertius de Wet †
Yuri Goegebeur ‡

Abstract

The estimation of extreme conditional quantiles is an important issue in different scientific disciplines. Up to now, the extreme value literature focused mainly on estimation procedures based on i.i.d. samples. On the other hand, quantile regression based procedures work well for estimation within the data range - i.e. the estimation of nonextreme quantiles - but break down when main interest is in extrapolation. Our contribution is a two-step procedure for estimating extreme conditional quantiles. In a first step nonextreme conditional quantiles are estimated nonparametrically using a local version of the Koenker and Bassett (1978) regression quantile methodology. Next, these nonparametric quantile estimates are used as analogues of univariate order statistics in well known extreme quantile estimators. The performance of the method is evaluated for both heavy tailed distributions and distributions with a finite right endpoint using a small sample simulation study. A bootstrap procedure is developed to guide in the selection of an optimal local bandwidth. Finally the procedure is illustrated in three cases.
Keywords: local polynomial estimation, quantile regression, extreme value index, extrapolation.

*Department of Mathematics, K.U.Leuven, Celestijnenlaan 200B, 3001 Heverlee, Belgium

†Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa. This author's research was partially supported by National Research Foundation grant 2046922 and a research grant from the University of Stellenbosch

‡Postdoctoral researcher, Fund for Scientific Research - Flanders, Department of Applied Economics, K.U.Leuven, Naamsestraat 69, 3000 Leuven and University Centre for Statistics, K.U.Leuven, de Croylaan 52B, 3001 Heverlee, Belgium

1 Introduction

In this paper we study the estimation of conditional quantiles with emphasis on the range beyond the data. This problem manifests itself in many areas e.g. an insurance company facing a fire insurance portfolio is typically interested in the claim size that will be exceeded once in, say, 10000 cases given additional factors such as sum insured and type of building. The extreme value literature on the estimation of extreme quantiles focuses mainly on the univariate case see for instance Weissman (1978), Pickands (1975), Dekkers *et al.* (1989), Dekkers and de Haan (1989), Beirlant and Matthys (2001a, 2001b). On the other hand, the literature in case of covariates is very sparse. Recent contributions are the work of Chernozhukov (1998, 2001), Gijbels and Peng (2000), Charnes *et al.* (1995). We apply the regression quantile methodology of Koenker and Bassett (Koenker and Bassett, 1978) in a nonparametric fashion using a two step procedure. In a first step we locally approximate the conditional quantile function using a polynomial yielding the regression analogues of univariate order statistics. Next, these nonparametric regression quantiles are used in a fashion analogous to the use of order statistics in the i.i.d. case for extrapolating beyond the data. This extrapolation is based on recent extreme value techniques see e.g. Beirlant *et al.* (1999) and Beirlant and Matthys (2001a, 2001b). This two-stage procedure generalizes the method based on local maxima proposed in Gijbels and Peng (2000). Chernozhukov (1998, 2001) gave a theoretical study of modern extreme value methods based on extreme regression quantiles in order to extrapolate outside the sample range. Further, these papers by Chernozhukov assume an additive error structure. In the present paper a fully nonparametric approach is pursued, using recent extreme value methods on local polynomial quantile regression estimates. Recently, Hall and Tajvidi (2000) and Davison and Ramesh (2000) proposed nonparametric estimates of some tail characteristics by means of extreme value modelling based on the generalized extreme value distribution and the generalized Pareto distribution, combined with local curve fitting for the corresponding parameter functions.

In section 2 the nonparametric estimation of regression quantiles on the basis of local polynomial approximations to the true conditional quantile function is discussed. Further it is shown how univariate extreme value methods can be used in the extrapolation step i.e. the estimation of extreme conditional quantiles. The performance of this two step method is evaluated in section 3 with a small sample simulation and this for both heavy tailed and right bounded distributions. Next, in section 4, we discuss a bootstrap procedure to guide in the selection of an optimal local bandwidth. In a final section the procedure is illustrated with three practical examples.

2 Methodology

Consider a random variable Y whose distribution depends on the covariate x. Without loss of generality we restrict ourselves to the single covariate case. We denote the conditional distribution of Y given x by $F_{Y|x}$ and the associated quantile function by $Q(\theta; x)$. More precisely

$$Q(\theta; x) = \inf\{y : F_{Y|x}(y) \geq \theta\} \qquad 0 < \theta < 1.$$

Interest lies in estimating $Q(\theta; x)$ with θ close to 1 or sometimes $Q(x) := \lim_{\theta \uparrow 1} Q(\theta; x)$ (finite right endpoint case). We assume no knowledge about $F_{Y|x}$ or Q and follow a nonparametric approach to estimating the latter. The estimation problem is considered in its full generality

without any particular structure on the model (e.g. linear additive/multiplicative models).

Suppose a dataset $(Y_1, x_1), \ldots, (Y_n, x_n)$ of independent observations according to $F_{Y|x}$ is given and we want to use these to estimate $Q(\theta; x^*)$ for a fixed x^*. Since Q is unknown we approximate it locally by a polynomial of degree p centered around x^* in the following way

$$Q(\theta; x) \approx \sum_{j=0}^{p} \beta_j (x - x^*)^j$$

for x sufficiently close to x^*, where $\beta_j = \beta_j(\theta) = \frac{1}{j!} \left. \frac{d^j Q(\theta;x)}{dx^j} \right|_{x=x^*}$, $j = 0, \ldots, p$. In particular note that $\beta_0 = Q(\theta; x^*)$. Consider a window of size $2h$ centered at x^* in which we apply the above polynomial approximation to $Q(\theta; x)$. Within this window we estimate the coefficients of this approximation as follows:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} I_{[x^*-h, x^*+h]}(x_i) f_\theta(Y_i - \sum_{j=0}^{p} \beta_j (x_i - x^*)^j) \tag{1}$$

with $f_\theta(x) = \theta x^+ + (1-\theta)x^-$, $x^+ = \max(0, x)$, $x^- = \max(0, -x)$ and $I_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}, 0$ otherwise. This optimization problem is a local version of the Koenker-Bassett approach to estimating regression quantiles (Koenker and Bassett, 1978). The maximum estimator for $Q(x^*)$ proposed by Gijbels and Peng (2000) can be obtained as a special case of (1) by setting $\theta = 1$ and $p = 0$. Noting that the first component of the solution to (1) estimates $Q(\theta; x^*)$, we denote this component by $\hat{Q}(\theta; x^*)$.

Before considering further the regression case let us first look at the i.i.d. case in order to motivate our approach.

Consider $Y_1, \ldots, Y_n$ independent and identically distributed random variables according to distribution function F_Y and associated quantile function $Q(\theta)$. Denote the ascending order statistics corresponding to $Y_1, \ldots, Y_n$ by $Y_{1,n} \leq \ldots \leq Y_{n,n}$. Further assume that F_Y is in the domain of attraction of the Generalized Extreme Value distribution i.e. there exist sequences of constants $(a_n > 0)$ and (b_n) such that for some $\gamma \in \mathbb{R}$

$$\lim_{n \to \infty} P\left(\frac{Y_{n,n} - b_n}{a_n} \leq y \right) = \exp\left(-(1 + \gamma y)^{-\frac{1}{\gamma}} \right) \qquad 1 + \gamma y > 0. \tag{2}$$

The parameter γ, called the extreme value index, gives important information about the tail of the underlying distribution function F_Y, where tails become heavier with increasing γ. Consequently, γ will play an important role when estimating $Q(\theta)$ for θ close to 1. The literature on estimating γ on the basis of an i.i.d. sample is very elaborate, see for instance Hill (1975), Pickands (1975), Dekkers *et al.* (1989), Beirlant *et al.* (1999), Feuerverger and Hall (1999).

In case $\gamma > 0$, the following approximate representation holds for log-spacings of successive order statistics (Beirlant *et al.*, 1999)

$$j \left(\log Y_{n-j+1,n} - \log Y_{n-j,n} \right) \overset{\mathcal{D}}{\approx} \left(\gamma + b_{n,k} \left(\frac{j}{k+1} \right)^{-\rho} \right) F_j \qquad j = 1, \ldots, k \tag{3}$$

with $b_{n,k} \in I\!R$, $\rho < 0$ and F_j, $j = 1, \ldots, k$, independent standard exponential random variables, from which γ can be estimated jointly with $b_{n,k}$ and ρ using the maximum likelihood method. Based on this, Beirlant and Matthys (2001a) proposed the following estimator for extreme quantiles

$$\hat{Q}_k^{(1)}(\theta) = Y_{n-k,n} \left(\frac{k+1}{(n+1)(1-\theta)} \right)^{\hat{\gamma}_k^{(1)}} \exp \left(\hat{b}_{n,k} \frac{1 - \left(\frac{(n+1)\theta}{k+1} \right)^{-\hat{\rho}_k}}{-\hat{\rho}_k} \right) \tag{4}$$

with $\hat{\gamma}_k^{(1)}$, $\hat{b}_{n,k}$ and $\hat{\rho}_k$ denoting the maximum likelihood estimators for respectively γ, $b_{n,k}$ and ρ under (3) using k log-spacings of order statistics, $k \in \{3, \ldots, n-1\}$.

For the more general case where γ can be positive or negative Beirlant and Matthys (2001b) derived the following approximate model for log-ratios of spacings

$$j \log \frac{Y_{n-j+1,n} - Y_{n-k,n}}{Y_{n-j,n} - Y_{n-k,n}} \overset{\mathcal{D}}{\approx} \frac{\gamma}{1 - \left(\frac{j}{k+1} \right)^{\gamma}} F_j \qquad\qquad j = 1, \ldots, k-1 \tag{5}$$

with $\gamma \in I\!R$, F_j, $j = 1, \ldots, k$, independent standard exponential random variables and γ is estimated using the maximum likelihood method. Model (5) can now be used as a basis for estimating extreme quantiles, see Beirlant and Matthys (2001b):

$$\hat{Q}_{k+1}^{(2)}(\theta) = Y_{n-k,n} + \hat{a}_{n,k+1} \frac{\left(\frac{k+1}{(n+1)(1-\theta)} \right)^{\hat{\gamma}_{k+1}^{(2)}} - 1}{\hat{\gamma}_{k+1}^{(2)}} \tag{6}$$

where

$$\hat{a}_{n,k+1} = \frac{1}{k} \sum_{j=1}^{k} j (Y_{n-j+1,n} - Y_{n-j,n}) \left(\frac{j}{k+1} \right)^{\hat{\gamma}_{k+1}^{(2)}} \tag{7}$$

and $\hat{\gamma}_{k+1}^{(2)}$ denotes the maximum likelihood estimator for γ under (5) based on the $k+1$ upper order statistics.

Remarks

- By imposing a second order condition on the tail of F_Y a model more complicated than (5) can be derived for log-ratios of spacings (Beirlant and Matthys, 2001b). Even though the more complicated model leads to a smaller bias for the estimators, it gives a substantially larger variance and thus a larger mean squared error. Therefore, in our work we only consider the reduced model (5).

- All the above estimators involve the choice of k, the number of upper order statistics used in the estimation. We will return to this issue when discussing the estimation of extreme conditional quantiles.

- In the special case where $\gamma < 0$ the endpoint $\lim_{\theta \uparrow 1} Q(\theta)$ is finite and can be estimated by setting θ equal to 1 in (6).

We now return to the case of estimating the conditional quantiles $Q(\theta, x^*)$ for θ close to 1. The following two-step procedure is proposed for this estimation problem:

1. Compute $\hat{Q}(\theta; x^*)$ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ with $n^* = \sum_{i=1}^{n} I_{[x^*-h, x^*+h]}(x_i)$. To ensure monotonicity, the constraints $\hat{Q}(\frac{i}{n^*+1}; x^*) \geq \hat{Q}(\frac{i-1}{n^*+1}; x^*)$, $i = 2, \ldots, n^*$, were imposed.

2. Estimate γ and extreme conditional quantiles based on (3) and (4), or (5) and (6), using $\hat{Q}(\frac{i}{n^*+1}; x^*)$, $i = 1, \ldots, n^*$, taking over the role of univariate order statistics. These estimates will be denoted by $\hat{\gamma}_k^{(1,RQ)}$ and $\hat{Q}_k^{(1,RQ)}(\theta)$, respectively $\hat{\gamma}_{k+1}^{(2,RQ)}$ and $\hat{Q}_{k+1}^{(2,RQ)}(\theta)$.

Remarks

- Note that in (1) the estimator $\hat{Q}(\theta; x^*)$ remains at $\hat{Q}(\frac{n^*}{n^*+1}; x^*)$ for $\frac{n^*}{n^*+1} < \theta < 1$. This does not yield sensible estimators beyond the data range. This is illustrated in Figure 1 using a small sample simulation from the Burr(β, τ, λ) distribution with distribution function given by

$$F_Y(y) = 1 - \left(\frac{\beta}{\beta + y^\tau}\right)^\lambda \qquad y > 0;\ \beta, \lambda, \tau > 0$$

  for which $\gamma = 1/(\lambda\tau)$ and $\rho = -1/\lambda$. The dependence on the covariate x was obtained by setting $\tau(x) = \exp(1 - x)$; further we took $\beta = \lambda = 1$ so $\gamma(x) = \exp(-1 + x)$. The values for the covariate x were drawn from the $U(-2, 2)$ distribution. In Figure 1 we show $Q(0.9999; x)$ (broken line) and the quartiles of $\hat{Q}(0.9999; x)$ (solid lines) computed over 500 simulated datasets of size $n = 500$ using $p = 2$ and $h = 0.5$. Clearly, a one-step procedure based on local quantile regression underestimates these extreme conditional quantiles. Step 2 in the above proposed procedure is a way of overcoming this shortcoming.

- Monotonicity of $\hat{Q}(\theta; x^*)$ can also be obtained using local versions of restricted regression quantiles; see He (1997) and Zhao (2000). We experimented with this using as common slope estimator both the 20% and 50% regression trimmed means but these were found not to perform well.

3 Simulation design

The methodology described in section 2 was studied extensively through simulation. The distributions from which we simulated are:

- the Burr$(1, \exp(1 - x), 1)$ distribution introduced above,

- the GPD(σ, γ) distribution for which the distribution function is given by

$$F_Y(y) = 1 - \left(1 + \frac{\gamma y}{\sigma}\right)^{-\frac{1}{\gamma}} \qquad \gamma \in I\!R, \sigma > 0$$

  with support $y > 0$ if $\gamma \geq 0$ and $0 < y < -\sigma/\gamma$ if $\gamma < 0$. In this study we take $\sigma = 1$ and $\gamma(x) = -\exp(x)$, so the conditional endpoint is given by $\exp(-x)$.

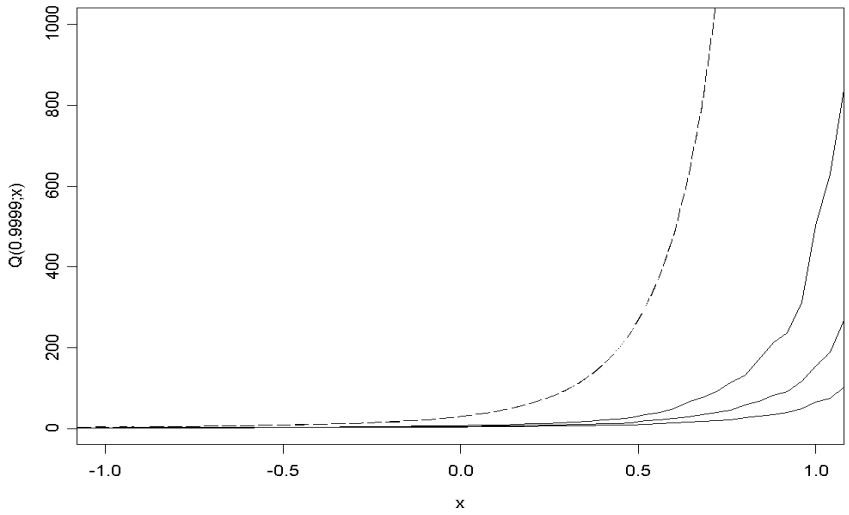

Figure 1: Burr$(1, \exp(1-x), 1)$ simulation: $Q(0.9999; x)$ (broken line) and the quartiles of $\hat{Q}(0.9999; x)$ (solid lines) computed over 500 simulated datasets of size $n = 500$.

For both distributions 500 datasets of size $n = 500$ were generated. The values for the covariate x were drawn from the $U(-2, 2)$ distribution. We considered the estimation problem at $x^* = -1$, $x^* = 0$ and $x^* = 1$. In case $\gamma \geq 0$ the main interest is in estimating the $1 - \frac{1}{10n^*}$ conditional quantile, $Q(1 - \frac{1}{10n^*}; x^*)$, for $\gamma < 0$ we attempt to estimate the conditional endpoint $\lim_{\theta \uparrow 1} Q(\theta; x^*)$. Results are reported for local quadratic approximations to $Q(\theta; x)$, i.e. $p = 2$, using bandwidths $h = 0.25$, $h = 0.5$ and $h = 0.75$.

Figures 2, 3 and 4 show the results for the Burr$(1, \exp(1-x), 1)$ simulation. In Figure 2 the quartiles (computed over the 500 simulated datasets) of $\hat{Q}(\theta; x^*)$ are shown as a function of θ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using bandwidths (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$. As h increases the interquartile range decreases as the estimation is based on a larger number of observations. On the other hand, increasing h leads to a larger bias. The quartiles of the maximum likelihood estimates for γ using model (3) and regression quantiles are shown as a function of k, $k = 5, \ldots, n^* - 1$, in Figure 3 for (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ and (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$. Finally, in Figure 4 we give the quartiles of the $1 - \frac{1}{10n^*}$ conditional quantile estimates, $\hat{Q}_k^{(1,RQ)}(1 - \frac{1}{10n^*})$, as a function of k, $k = 5, \ldots, n^* - 1$, at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ and (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$. Note that accurate estimation is harder as the tail of the underlying conditional distribution becomes heavier.

The corresponding results for the GPD$(1, -\exp(x))$ simulation are shown in Figures 5, 6 and 7. Here, using $p = 2$ yields only for $h = 0.25$ sensible estimates for $\gamma(1)$ and $\lim_{\theta \uparrow 1} Q(\theta; 1)$. The poor performance at bandwidths $h = 0.5$ and $h = 0.75$ is caused by the bias in the upper regression quantiles, see Figure 8. Applying a local cubic approximation gives good estimates for all 3 bandwidths considered. From this it can be concluded that the framework of local polynomial

quantile regression offers extra flexibility which is needed in the given regression context.

Overall we conclude that the choice of the bandwidth is quite important. A bootstrap algorithm to choose this parameter adaptively is given in the next section.

4 Selection of the bandwidth h

When applying nonparametric techniques, a common issue is the selection of the bandwidth parameter h. On the one hand, h should be taken sufficiently large in order to have a substantial number of observations in the interval $[x^* - h, x^* + h]$. On the other hand, h should not be too large since a large h value will increase the bias due to the fact that the local polynomial approximation to $Q(\theta; x)$ is getting worse. An asymptotic mean squared error (AMSE) criterion could be constructed to guide in the selection of a *local* optimal h value. Of course, this requires knowledge of the expressions for the asymptotic variance and bias of $\hat{Q}(\theta; x^*)$. Further, these expressions will depend on the unknown model parameters, making a plug-in of estimates necessary.

In this paper, the bootstrap estimate of $\mathrm{MSE}(\hat{Q}(\theta; x^*))$ defined as

$$\hat{\mathrm{MSE}}(\hat{Q}(\theta; x^*)) = \mathrm{E}(\hat{Q}^*(\theta; x^*) - \hat{Q}(\theta; x^*)|\hat{F})^2 \tag{8}$$

with $\hat{Q}^*(\theta; x^*)$ the bootstrap estimate of $Q(\theta; x^*)$ and $\hat{F}$ the empirical distribution function, will be used to guide the selection of the bandwidth parameter h, see also Hall (1990) and Efron and Tibshirani (1993). Given bootstrap samples $(Y^*_{1,b}, x^*_{1,b}), \ldots, (Y^*_{n,b}, x^*_{n,b})$, $b = 1, \ldots, B$, drawn with replacement from $(Y_1, x_1), \ldots, (Y_n, x_n)$, (8) can be estimated by

$$\hat{\hat{\mathrm{MSE}}}(\hat{Q}(\theta; x^*)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{Q}^*_b(\theta; x^*) - \hat{Q}(\theta; x^*))^2$$

where $\hat{Q}^*_b(\theta; x^*)$ denotes the estimate of $Q(\theta; x^*)$ from bootstrap sample b. The optimal h value obtained by the bootstrap is then defined as

$$\hat{h}^*_{opt} = \arg\min \hat{\hat{\mathrm{MSE}}}(\hat{Q}(\theta; x^*)).$$

In Figure 9 the bootstrap procedure is illustrated on the basis of a small sample simulation from the $\mathrm{GPD}(1, -\exp(x))$ distribution introduced above. The values for the covariate x were drawn from the $U(-2, 2)$ distribution. We considered the selection of an optimal local h for the estimation of $Q(0.999; 1)$ using $p = 2$. The results are based on 100 samples of size $n = 500$. The bootstrap estimates of $\mathrm{MSE}(\hat{Q}(0.999; 1))$ at a particular h value are based on $B = 500$ bootstrap samples. In Figure 9 (a) the histogram of the 100 values of $\hat{h}^*_{opt}$ for estimating $Q(0.999; 1)$ are shown. Figure 9 (b) shows the boxplots of the 100 realisations of $\hat{Q}(0.999; 1)$ obtained with the bootstrap procedure and for some fixed h-values.

Next to the above described bootstrap procedure, used to guide in the selection of an optimal local h value, we also tried a cross validation procedure developed to obtain a data driven *global* h value i.e. a h that performs well over the whole covariate range. Similarly to the cross-validation

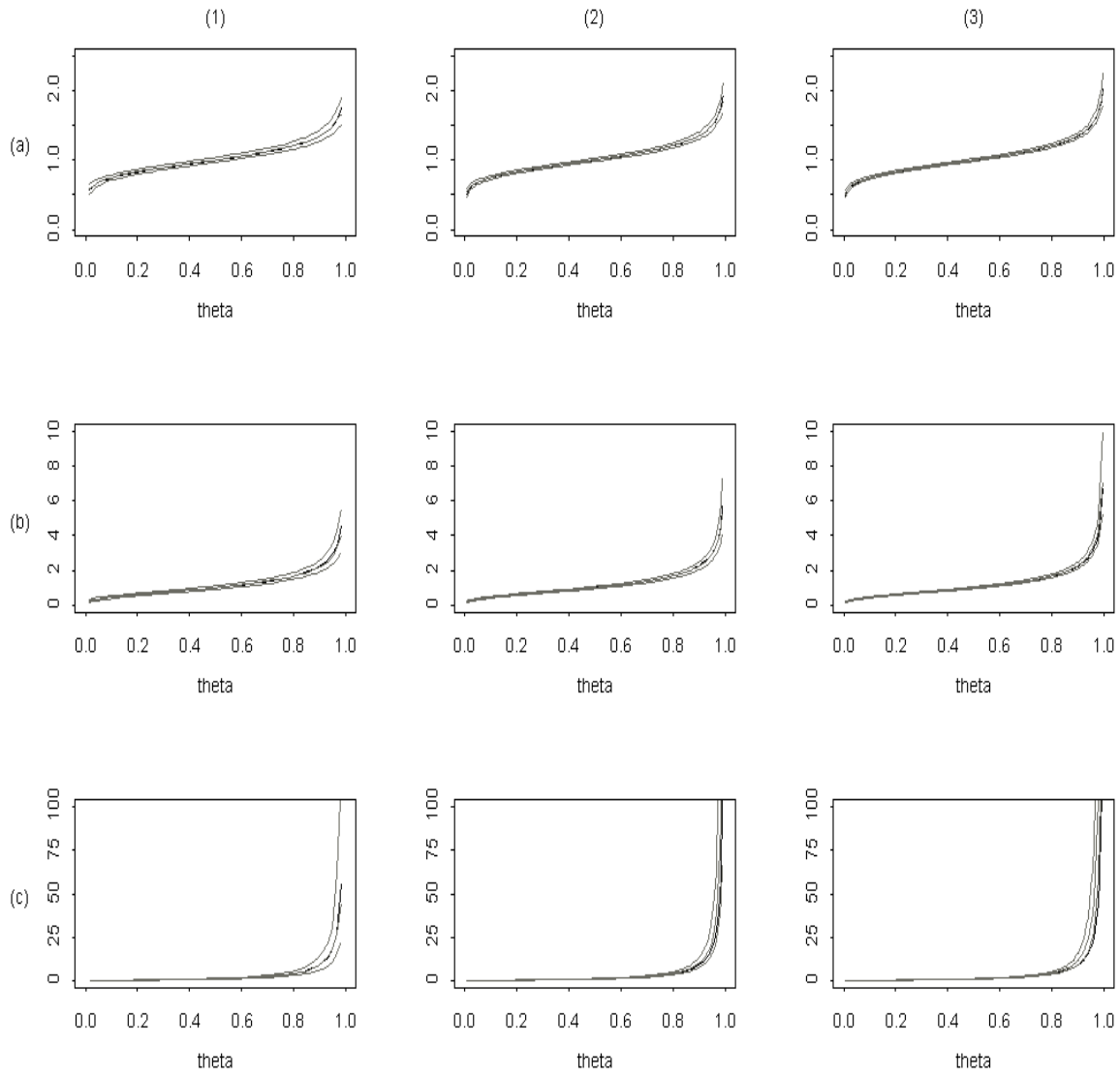

Figure 2: Burr$(1, \exp(1-x), 1)$ simulation: $Q(\theta; x^*)$ (black line) and quartiles of $\hat{Q}(\theta; x^*)$ (grey lines) as a function of θ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

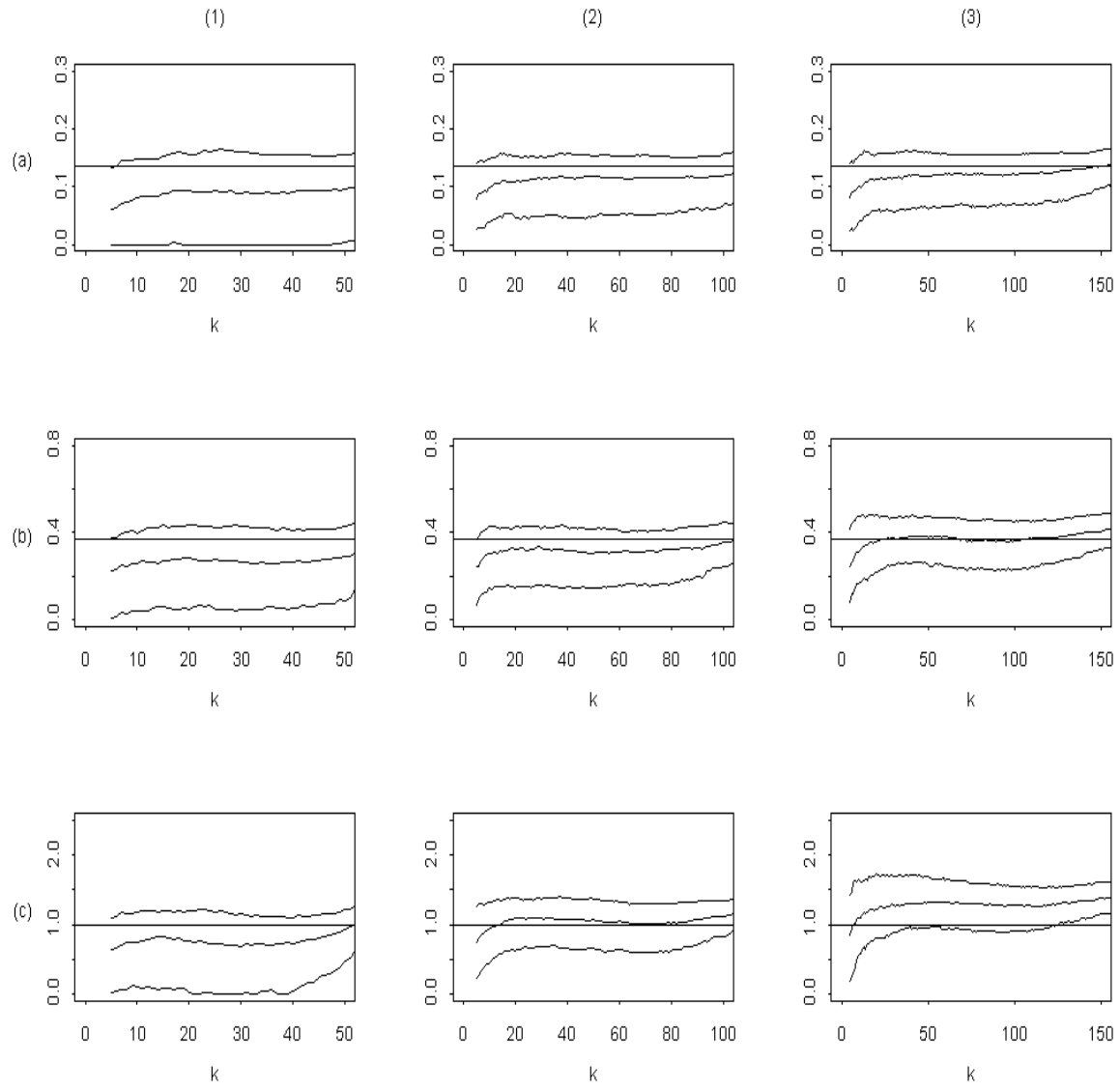

Figure 3: Burr$(1, \exp(1-x), 1)$ simulation: quartiles of $\hat{\gamma}_k^{(1,RQ)}$ as a function of k, $k = 5, \ldots, n^* - 1$, at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

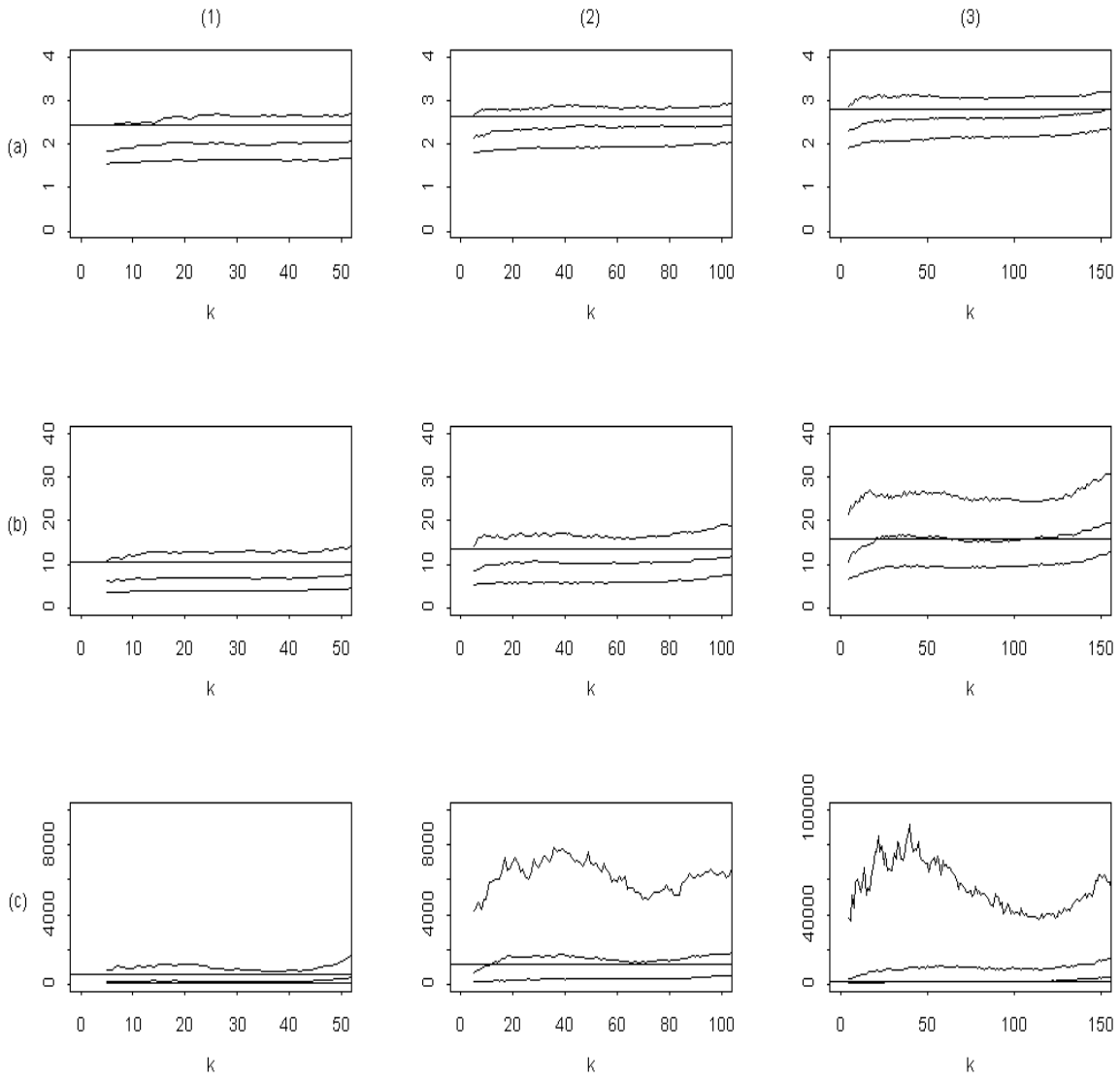

Figure 4: Burr$(1, \exp(1-x), 1)$ simulation: quartiles of $\hat{Q}_k^{(1,RQ)}(1-\frac{1}{10n^*})$ as a function of k, $k = 5, \ldots, n^* - 1$, at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

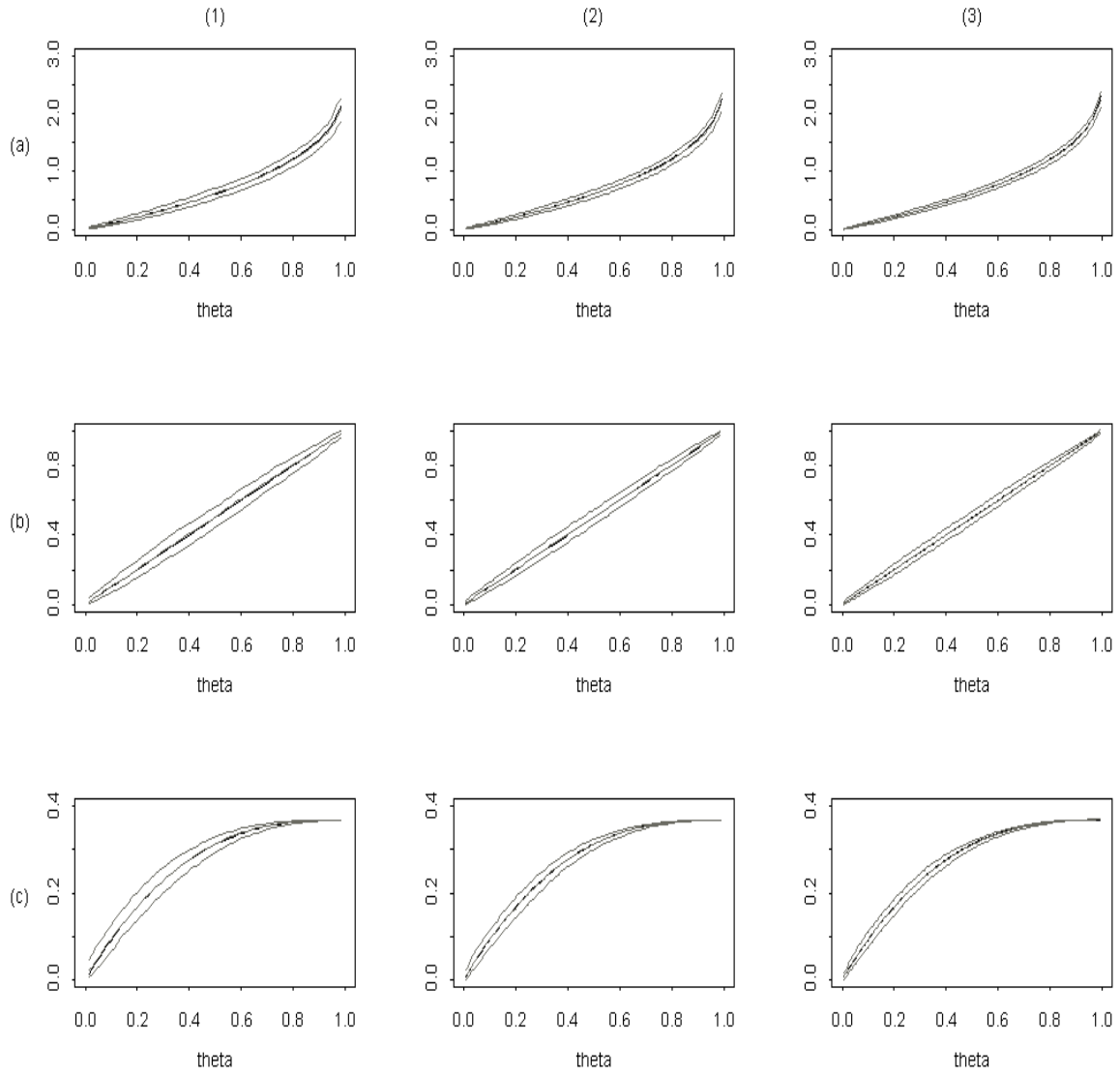

Figure 5: GPD$(1, -\exp(x))$ simulation: $Q(\theta; x^*)$ (black line) and quartiles of $\hat{Q}(\theta; x^*)$ (grey lines) as a function of θ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

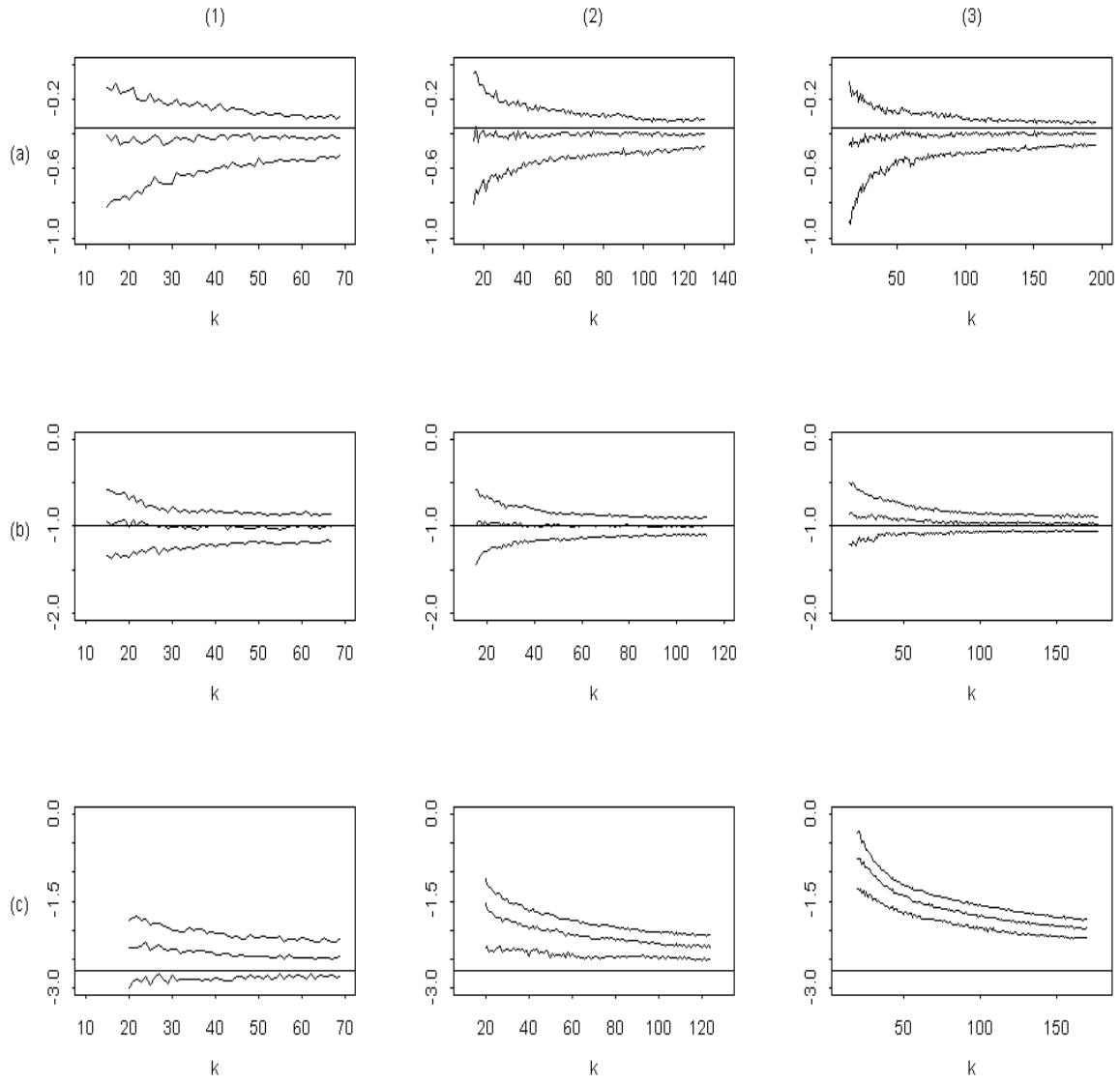

Figure 6: GPD$(1, -\exp(x))$ simulation: quartiles of $\hat{\gamma}_{k+1}^{(2,RQ)}$ as a function of k, $k = 15, \ldots, n^* - 1$, at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

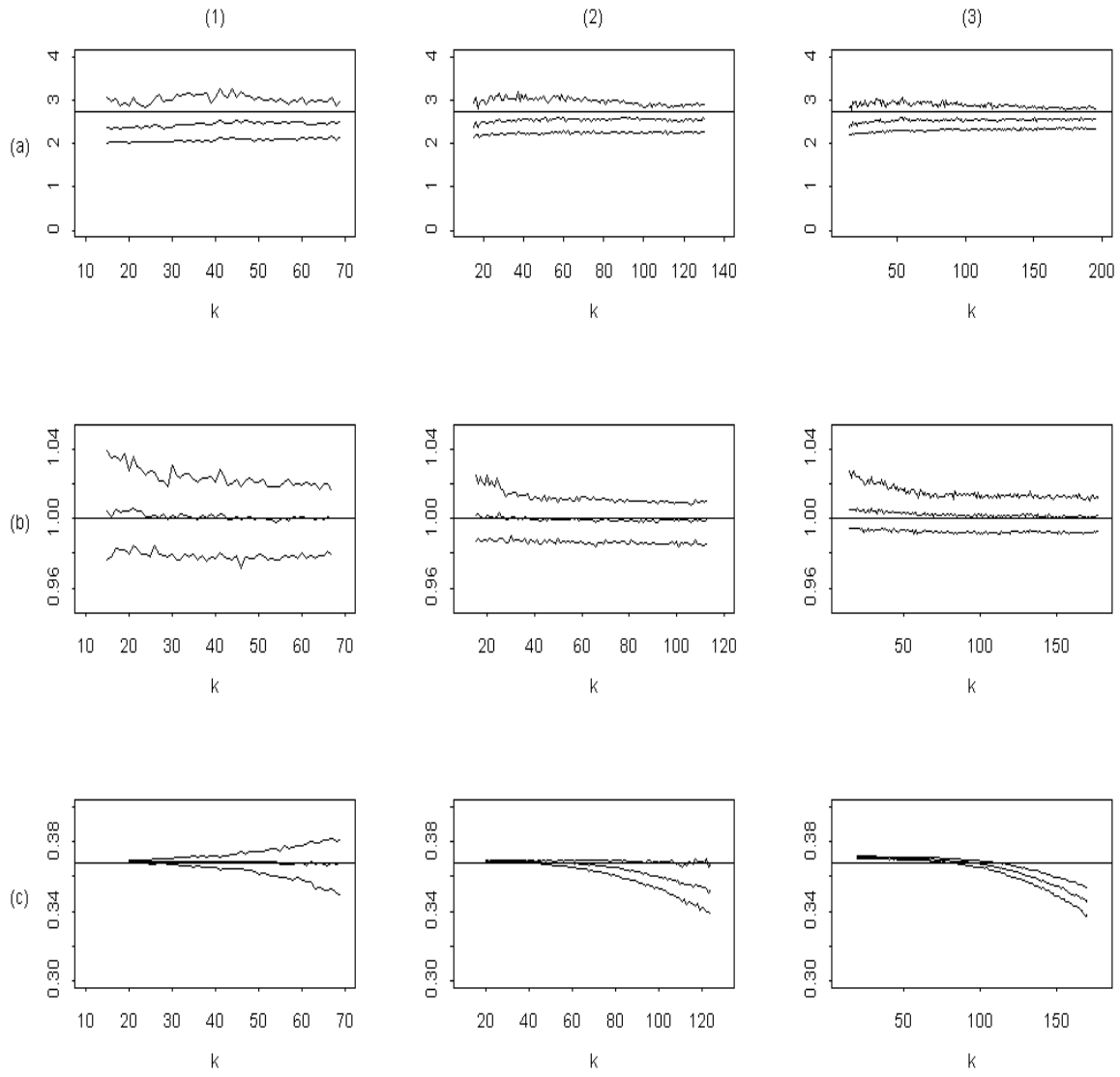

Figure 7: GPD$(1, -\exp(x))$ simulation: quartiles of $\hat{Q}_k^{(2,RQ)}(1)$ as a function of k, $k = 15, \ldots, n^* - 1$, at (a) $x^* = -1$, (b) $x^* = 0$, (c) $x^* = 1$ using (1) $h = 0.25$, (2) $h = 0.5$, (3) $h = 0.75$.

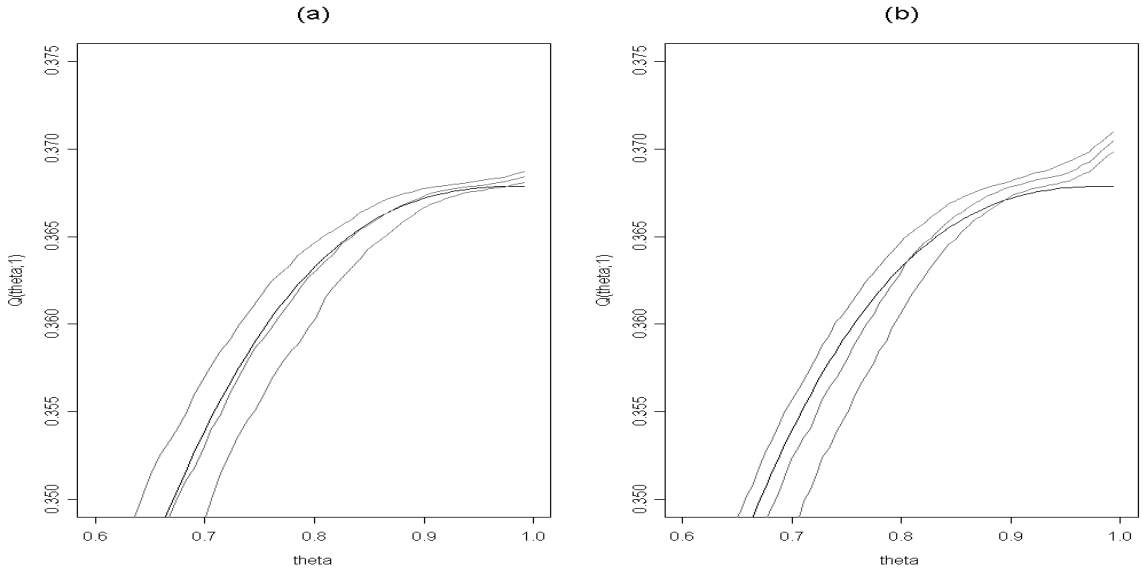

Figure 8: GPD$(1, -\exp(x))$ simulation: $Q(\theta; x^*)$ (black line) and quartiles of $\hat{Q}(\theta; x^*)$ (grey lines) as a function of θ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ at $x^* = 1$ using (a) $h = 0.5$ and (b) $h = 0.75$.

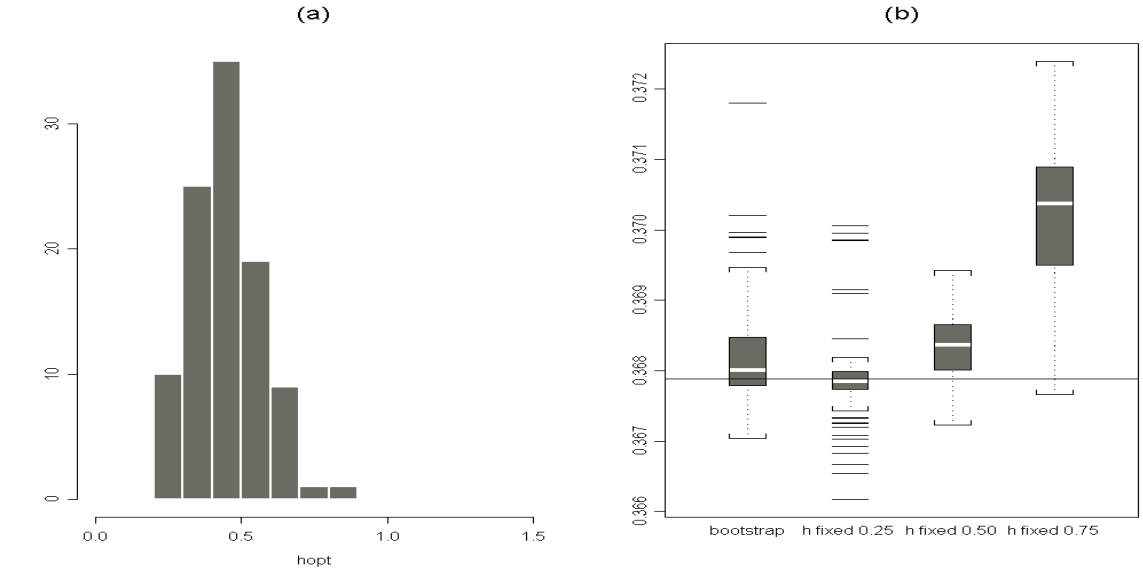

Figure 9: GPD$(1, -\exp(x))$ simulation: (a) histogram of $\hat{h}^*_{opt}$-values for the estimation of $Q(0.999; 1)$ and (b) boxplots of $\hat{Q}(0.999, 1)$ resulting from using $\hat{h}^*_{opt}$ and some fixed h-values (the horizontal line represents the true value of $Q(0.999; 1)$).

procedure proposed by Aerts and Claeskens (1997) in the context of local polynomial maximum likelihood estimation, the optimal h obtained by cross-validation is here defined as

$$\hat{h}_{CV} = \arg\min \sum_{i=1}^{n} f_\theta(Y_i - \hat{\beta}_{0,(i)}) \tag{9}$$

where $\hat{\beta}_{0,(i)}$ is the local estimator for $Q(\theta, x_i)$ based on the sample without the i-th observation. Based on the results obtained from a small sample simulation study, this procedure did not prove to be useful for practical purposes and hence will not be further considered in this paper.

5 Case studies

In this section we illustrate the proposed two step procedure using three case studies.

5.1 Condroz data

Our first example comes from the pedochemical context. The database contains measurements on the variables Calcium (Ca) content and pH level for soil samples taken in different cities in the Condroz (a geographical region in the southern part of Belgium). These data were already analyzed in Goegebeur *et al.* (2002) with emphasis on the development of an automatic procedure for highlighting suspicious points. Here we concentrate on the related problem of estimating extreme quantiles of the conditional distribution of the variable Ca given pH level. In Figure 10 (a) we show the scatterplot of Ca content versus pH level for one of the cities. Next to the global upwards trend of the point cloud, extreme Ca measurements tend to occur more often for the larger pH levels. Figure 10 (b) shows the bootstrap estimate of MSE$(\hat{Q}(0.9;6))$ for $p = 1$ as a function of h. This estimated bootstrap MSE attains its mimum at $h = 0.44$. Next we applied the exponential regression model (3) to $\hat{Q}(\theta;\text{pH})$ for $\theta = \frac{1}{n^*+1}, \ldots, \frac{n^*}{n^*+1}$ and this on a grid of pH values. The median of $\hat{\gamma}_k^{(1,RQ)}$, $k = 1, \ldots, n^* - 1$, is given as a function of pH in Figure 10 (c). Finally, in Figure 10 (d) we show the median of $\hat{Q}_k^{(1,RQ)}(0.9995)$, $k = 1, \ldots, n^*-1$, as a function of pH on the Ca versus pH scatterplot.

5.2 Coca Cola data

Our second example comes from the financial context. Figure 11 (a) shows the daily volumes of Coca Cola quotes, adjusted for dividends and splits, between January 1, 1980 and December 31, 2000. The data are taken from the web site http://table.finance.yahoo.com/k?s=ko&g=d. We applied the local estimation procedure discussed above with $p = 1$ and $h = 560$. This value for h minimizes the bootstrap estimate of the mean squared error of $\hat{Q}(0.9; x^*)$, where x^* is taken as day 1000 of the dataset (December 13, 1983), cf Figure 11 (b). In Figure 11 (c) we show the median (over k) of the $\hat{\gamma}_k^{(1,RQ)}$ as a function of time. Note that, except for the small downwards trend during the last years, the gamma estimate is quite stable over time. Finally, Figure 11 (d) shows the median (over k) of the estimated 5 year return level $\hat{Q}_k^{(1,RQ)}(1 - \frac{1}{1250})$ as a function of time. Exactly 2 observations are above this estimated five year return level which corresponds approximately to what can be expected over a period of 20 years.

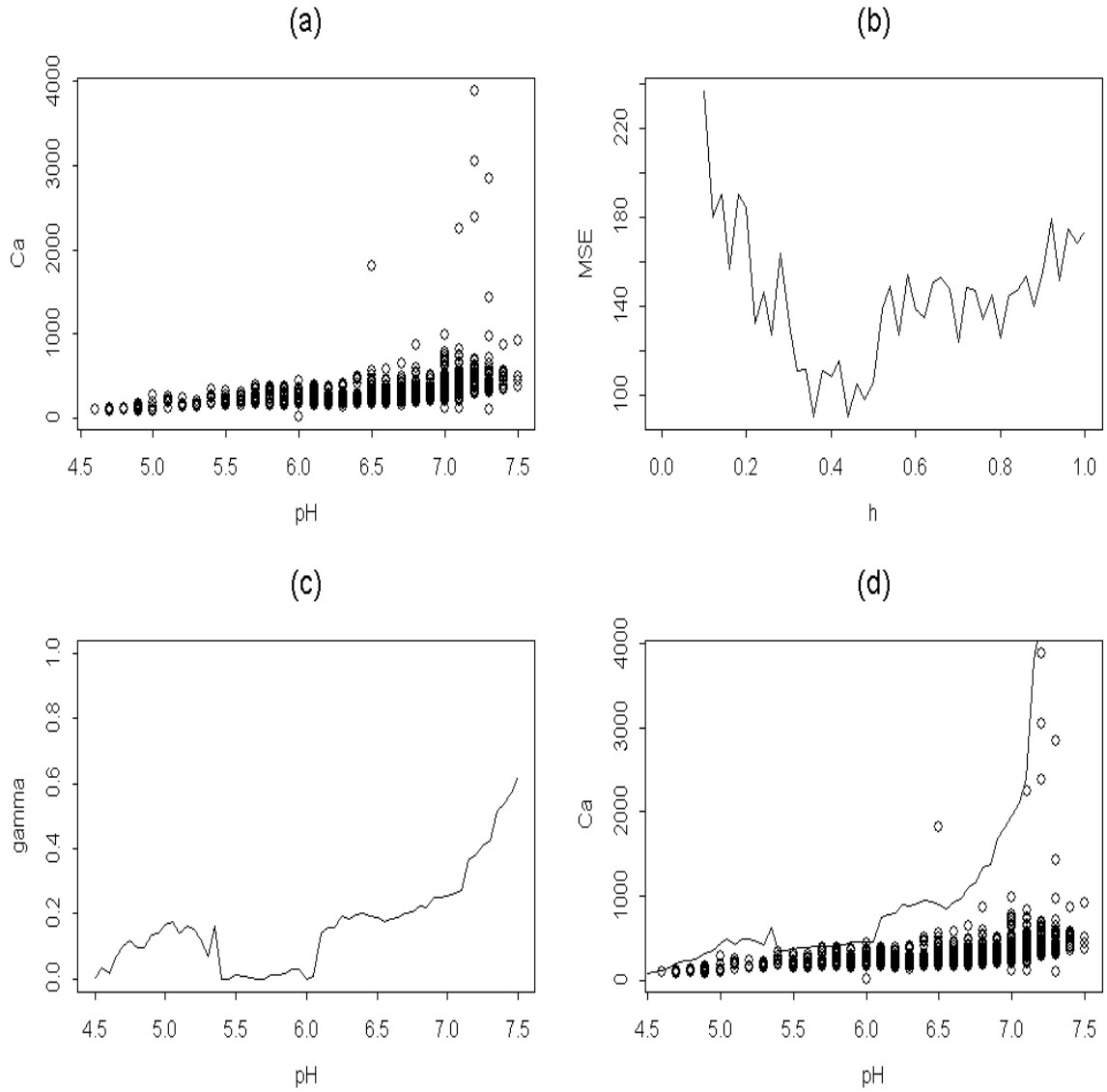

Figure 10: Condroz data: (a) Ca versus pH scatterplot, (b) $\hat{MSE}(\hat{Q}(0.9;6))$ versus h for $p=1$, (c) med$\{\hat{\gamma}_k^{(1,RQ)};\ k=5,\ldots,n^*-1\}$ as a function of pH and (d) scatterplot of Ca versus pH with med$\{\hat{Q}_k^{(1,RQ)}(0.9995);\ k=5,\ldots,n^*-1\}$ superimposed.

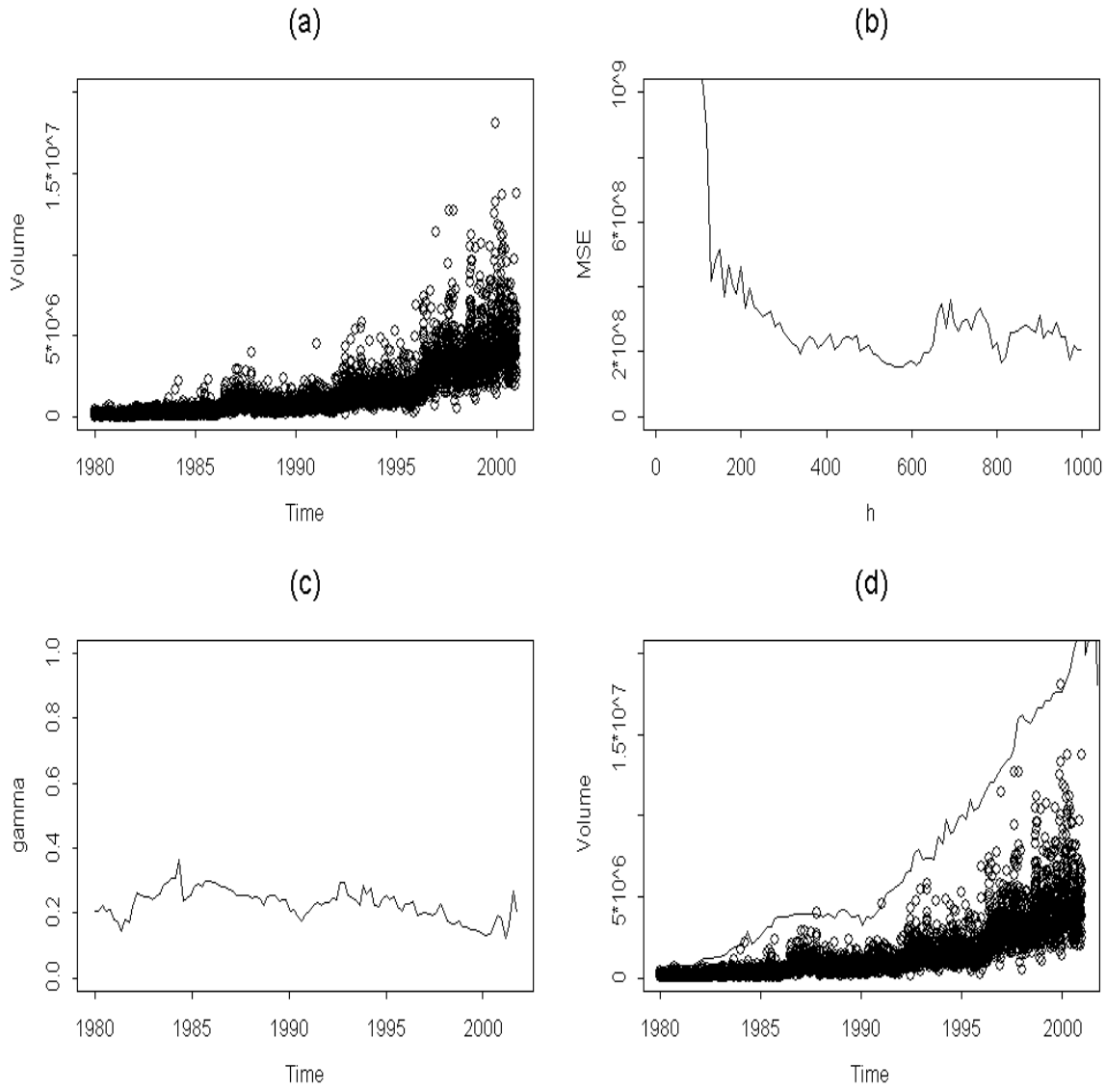

Figure 11: Coca Cola data: (a) Volume versus time scatterplot, (b) $\hat{\hat{\text{MSE}}}(\hat{Q}(\theta; x^*))$ as a function of h, (c) $\text{med}\{\hat{\gamma}_k^{(1,RQ)};\ k = 5, \ldots, n^* - 1\}$ as a function of time and (d) scatterplot of Volume versus time and $\text{med}\{\hat{Q}_k^{(1,RQ)}(1 - \frac{1}{1250});\ k = 5, \ldots, n^* - 1\}$ (5 year return level) as a function of time.

5.3 Electric utility data

Endpoint or boundary estimation is often encountered when studying the relation between input (e.g. labor, capital) and output (e.g. goods produced) of firms in a productivity analysis. Clearly, given a certain amount of input, the possible output is bounded above. This upper bound is the so-called efficiency curve. We use a data set on 123 American electric utility companies to illustrate this aspect of the methodology given above. This data set was also used in Gijbels and Peng (2000). Figure 12 (a) shows the Output versus log(Cost) scatterplot. Because of the sparseness of the observations with $\log(\text{Cost}) < 1$ we restricted the analysis to the observations with $1 \leq \log(\text{Cost}) \leq 6$. In Figure 12 (b) we show $\hat{\hat{\text{MSE}}}(\hat{Q}(0.9;3))$ as a function of h. Application of the bootstrap procedure at different log(Cost)-values indicated the need for a covariate dependent bandwidth parameter. The γ estimates at $k = n^* - 1$ shown in Figure 12 (c) were obtained by using $h = 0.5$ if $\log(\text{Cost}) < 3.5$ and $h = 1$ otherwise. This choice was clearly indicated by the bootstrap algorithm. To avoid difficulties in the estimation of the conditional endpoints, the γ estimates were bounded away from 0. However, as can be seen in the plot, this constraint is active only at a small number of log(Cost)-values. Finally, in Figure 12 (d) we give the Output versus log(Cost) scatterplot with the estimated conditional endpoint superimposed.

6 Conclusion

We have proposed a flexible nonparametric method for estimating extreme quantiles in a regression setting. In this two-stage procedure we combine the merits of local polynomial quantile regression (first step) and recent extreme value methods (second step) in order to obtain a general and effective technique in comparison with earlier attempts from the literature. In future work we intend to complete this study with inferential tools based on the given quantile estimates. To this end asymptotic results based on the work of Chernozhukov (1998, 2001) could be pursued.

References

[1] Aerts, M. and Claeskens, G., 1997. Local polynomial estimators in multiparameter likelihood models. *Journal of the American Statistical Association*, **92**, 1536-1545.

[2] Beirlant, J., Dierckx, G., Goegebeur, Y. and Matthys, G., 1999. Tail index estimation and an exponential regression model. *Extremes*, **2**, 177-200.

[3] Beirlant, J. and Matthys, G., 2001a. Extreme quantile estimation for heavy tailed distributions. Technical report, Department of Mathematics, K.U.Leuven.

[4] Beirlant, J. and Matthys, G., 2001b. Estimating the extreme value index and high quantiles with exponential regression models. Technical report, Department of Mathematics, K.U.Leuven.

[5] Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M. (Eds), 1995. *Data Envelopment, Theory, Methodology and Applications*. Kluwer Academic, Boston.

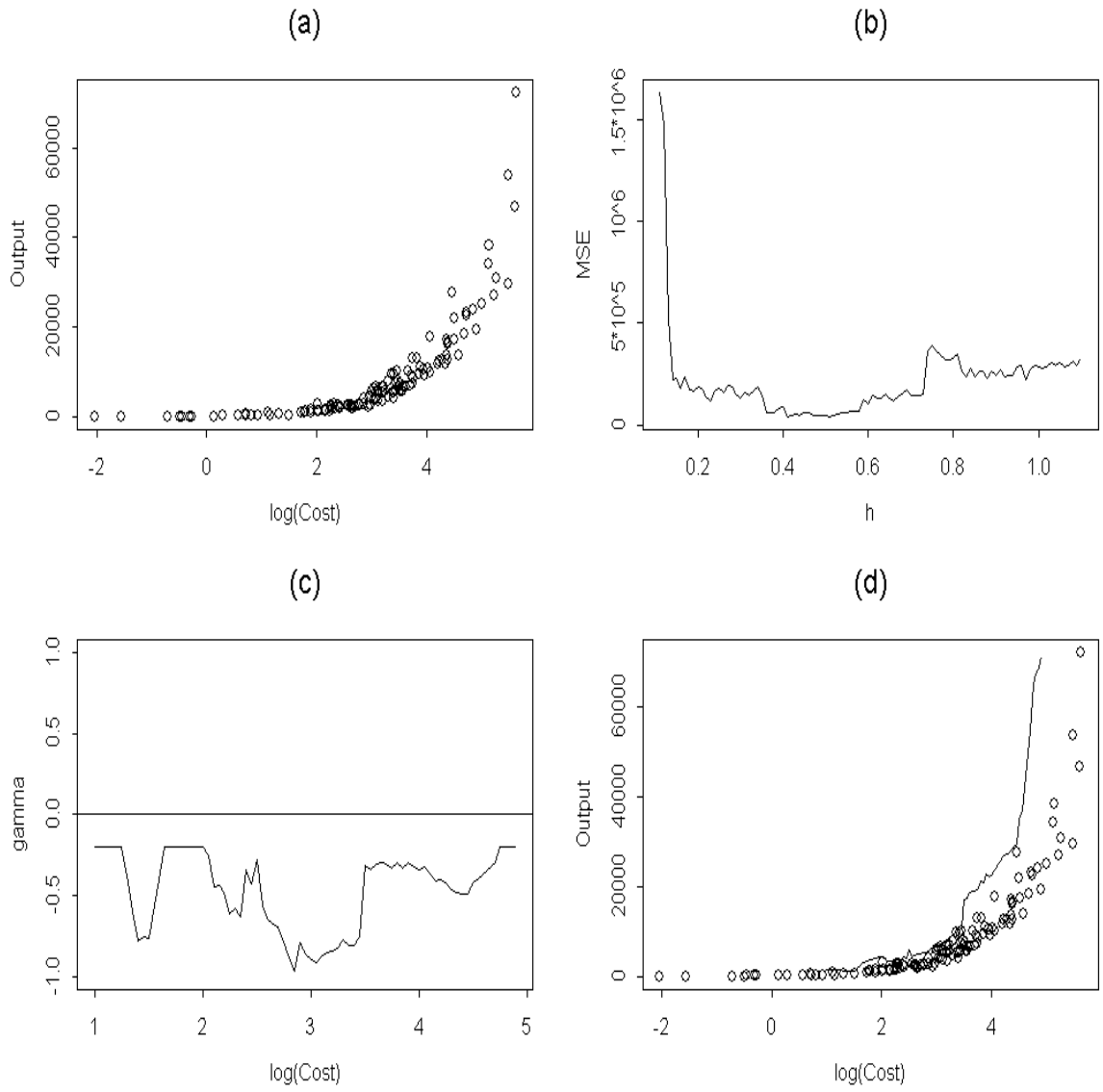

Figure 12: Electric utility data: (a) Output versus log(Cost) scatterplot, (b) $\hat{\hat{\text{MSE}}}(\hat{Q}(0.9;3))$ as a function of h, (c) $\hat{\gamma}_{k+1}^{(2,RQ)}$ as a function of log(Cost) and (d) $\hat{Q}_{k+1}^{(2,RQ)}(1)$ as a function of log(Cost).

[6] Chernozhukov, V., 1998. Nonparametric extreme regression quantiles. Technical report, Department of Economics, Stanford.

[7] Chernozhukov, V., 2001. Conditional extremes and near extremes. Technical report, Department of Economics, MIT, Cambridge.

[8] Davison, A.C. and Ramesh, N.I., 2000. Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society B*, **62**, 191-208.

[9] Dekkers, A.L.M and de Haan, L., 1989. On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics*, **17**, 1795-1832.

[10] Dekkers, A.L.M., Einmahl, J.H.J. and de Haan, L., 1989. A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, **17**, 1833-1855.

[11] Efron, B. and Tibshirani, R.J., 1993. *An introduction to the bootstrap*. Chapman & Hall.

[12] Feuerverger, A. and Hall, P., 1999. Estimating a tail exponent by modelling departure from a Pareto distribution. *Annals of Statistics*, **27**, 760-781.

[13] Gijbels, I. and Peng, L., 2000. Estimation of a support curve via order statistics. *Extremes*, **3**, 251-277.

[14] Goegebeur, Y., Planchon, V., Beirlant, J. and Oger, R., 2002. Quality assessment of pedochemical data using extreme value methodology. Technical report, K.U.Leuven.

[15] Hall, P., 1990. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, **32**, 177-203.

[16] Hall, P. and Tajvidi, N., 2000. Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, **15**, 153-167.

[17] He, X., 1997. Quantile curves without crossing. *The American Statistician*, **51**, 186-192.

[18] Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, **3**, 1163-1174.

[19] Koenker, R. and Bassett, G., 1978. Regression quantiles. *Econometrica*, **46**, 33-50.

[20] Pickands III, J., 1975. Statistical inference using extreme order statistics. *Annals of Statistics*, **3**, 119-131.

[21] Weissman, I., 1978. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, **73**, 812-815.

[22] Zhao, Q., 2000. Restricted regression quantiles. *Journal of Multivariate Analysis*, **72**, 78-99.