Variable data driven bandwidth choice in nonparametric quantile regression

Klaus Abberger, University of Konstanz, Germany

Abstract:

The choice of a smoothing parameter or bandwidth is crucial when applying nonparametric regression estimators. In nonparametric mean regression various methods for bandwidth selection exists. But in nonparametric quantile regression bandwidth choice is still an unsolved problem. In this paper a selection procedure for local varying bandwidths based on the asymptotic mean squared error (MSE) of the local linear quantile estimator is discussed. To estimate the unknown quantities of the MSE local linear quantile regression based on cross-validation and local likelihood estimation is used.

Key Words: quantile regression, nonparametric regression, conditional quantile estimation, local linear estimation, local bandwidth selection, local likelihood, generalized logistic distribution

1 Introduction

It is an interesting problem in a study of the interdependence between a random variable Y and a covariate X is how estimate the quantiles of Y for a given value of X. For fixed $\alpha \in (0,1)$, the quantile regression function gives the α th quantile $q_\alpha(x)$ in the conditional distribution of Y given $X = x$. Quantile regression can be

used to measure the effect of covariates not only in the center of the distribution, but also in its lower and upper tails.

Various nonparametric estimation methods for quantile regression have been discussed. These methods include spline smoothing, kernel estimation, nearest-neighbour estimation and local weighted polynomial regression. Yu and Jones (1998) propose two kinds of local linear quantile regression.

In this paper the local weighted linear quantile regression estimator is used. the estimator is defined by setting $\hat{q}_\alpha(x) = \hat{a}$, where $\hat{a}$ and $\hat{b}$ minimize

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - a - b(X_i - x))K\left(\frac{x - X_i}{h}\right), \tag{1}$$

with kernel function $K(\cdot)$, bandwidth h and loss function

$$\rho_\alpha = \alpha 1_{\{u \geq 0\}}(u) \cdot u + (\alpha - 1)1_{\{q < 0\}}(u) \cdot u \tag{2}$$

introduced by Koenker and Basset (1978) in connection with parametric quantile regression. For a discussion of this nonparametric estimator see Heiler (2000), or Yu and Jones (1998), who also derives the mean squared error (MSE) of this estimator. The considerations in Sec. 2 of this paper are based on this MSE.

The practical performance of $\hat{q}_\alpha(x)$ depends strongly on the bandwidth h. Yu and Jones (1998) develop a rule-of-thumb bandwidth choice procedure based on the plug-in idea. Starting point is the asymptotically optimal bandwidth minimizing the MSE. Since this bandwidth depends on unknown quantities the authors introduce some simplifying assumptions. These assumptions result in the bandwidth selection strategy

$$h_\alpha = h_{mean}\{\alpha(1 - \alpha)/\phi(\Phi^{-1}(\alpha))^2\}^{1/5}. \tag{3}$$

ψ and Ψ are standard normal density and distribution function and h_{mean} is a bandwidth choice for regression mean estimation with one of various existing methods.

As it can be seen this procedure leads to identical bandwidths for the α and $(1-\alpha)$ quantiles.

Abberger (1998) adapts the cross-validation idea to kernel quantile regression and presents some simulation examples.

In contrast to the above two bandwidth selection strategies where one global bandwidth is chosen, in this paper a method for locally varying bandwidth choice is developed. An algorithm based on the MSE optimal bandwidth is discussed in Sec. 2 and some simulation examples are presented in Sec. 3.

2 Variable bandwidth choice

For local linear quantile regression, the asymptotic form of the mean squared error is

$$MSE(\hat{q}_\alpha(x)) \approx \frac{1}{4}h^4\mu_2(K)^2 q_\alpha''(x)^2 + \frac{R(K)\alpha(1-\alpha)}{nhg(x)f(q_\alpha(x)|x)^2}, \tag{4}$$

where $\mu_2(K) = \int u^2 K(u)du$, $R(K) = \int K^2(u)du$, and g is the "design density", the marginal density of X. f denotes the conditional density $f(y|x)$ of Y given $X = x$ and $q_\alpha''(x)$ the second derivative of the conditional α-quantile (see Yu and Jones (1998)).

From (4) follows the asymptotically optimal bandwidth

$$h_\alpha^5(x) = \frac{R(K)\alpha(1-\alpha)}{n\mu_2(K)^2 q_\alpha''(x)^2 g(x) f(q_\alpha(x)|x)^2}. \tag{5}$$

This bandwidth depends on the unknown quantities $g(x)$, $q_\alpha(x)$ and $f(y|x)$. Plug-in estimates for $h_\alpha(x)$ use formula (5), replacing the unknown quantities by some estimates. Before calculating the local bandwidths it is necessary to estimate:

(i) the design density $g(x)$

(ii) the conditional quantile and its second derivative

(iii) the conditional density $f(y|x)$ at $y = q_\alpha(x)$.

An algorithm is needed which gives estimates for these quantities. In this paper the following procedure is chosen:

(i) $g(x)$ is easiest to estimate. Various nonparametric density estimators can be applied. Bandwidth choice procedures also exist. In equidistant designs $g(x)$ is uniform.

(ii) A prior estimate of $q_\alpha(x)$ and its second derivative is estimated by local quadratic quantile regression

$$\min_{a,b,c} \left\{ \sum_{i=1}^{n} \rho_\alpha(Y_i - a - b(X_i - x) - c(X_i - x)^2) K\left(\frac{x - X_i}{h}\right) \right\}, \qquad (6)$$

with $\hat{q}_\alpha(x) = a$ and $\hat{q}_\alpha''(x) = c$ (see Fan and Gijbels (1996) for local polynomial estimation in general). These estimates are based on a global bandwidth chosen by cross-validation. That is the bandwidth minimizing

$$\min_{h} \left\{ \sum_{i=1}^{n} \rho_\alpha(Y_i - \hat{q}_\alpha^{(-i)}(X_i)) \right\}, \qquad (7)$$

with $\hat{q}_\alpha^{(-i)}(X_i)$, the so called leave-one-out estimator. That means that the estimator of the conditional quantile at X_i is calculated without using the observation (Y_i, X_i) (see Abberger (1998) for details).

(iii) The most crucial point is the estimation of the conditional density $f(\cdot|x)$ at $q_\alpha(x)$. To estimate this density we use local likelihood estimation similar to Staniswalis (1989). With presumed density $\tilde{f}_\omega$, parameter vector ω and parameter space Ω the parameters are estimated locally as maximizers of the weighted likelihood criterion

$$\hat{\omega}(x) = \max_{\omega \in \Omega} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \log f(Y_i, \omega). \qquad (8)$$

Having estimated the parameters they are plugged into the assumed density $\tilde{f}_{\hat{\omega}}$ and the value of $\tilde{f}_{\hat{\omega}}(\hat{q}_\alpha(x)|x)$ is calculated. Doing this a primer bandwidth and a density $\tilde{f}_\omega$ has to be chosen. As discussed by Staniswalis (1989) a global bandwidth selection procedure is cross-validation similar to step (ii) of the present algorithm. It remains the presumption of a family of densities. Therefore, the location-scale-shape model of the generalized logistic distribution is used. The generalized logistic distribution with location (θ), scale (σ) and shape (b) parameters has the density

$$f(x) = \frac{\frac{b}{\sigma} e^{-\frac{(x-\theta)}{\sigma}}}{(1 + e^{-\frac{(x-\theta)}{\sigma}})^{b+1}},\ b > 0,\ \sigma > 0,\ \theta \in R,\ x \in R. \tag{9}$$

This distribution and the maximum likelihood estimation of its parameters is discussed in detail by Abberger and Heiler (2000). For $b = 1$ the distribution is symmetric, for $b < 1$ the distribution is skewed to the left and for $b > 1$ it is skewed to the right.

The logistic distribution and its various generalizations are discussed in Johnson, Kotz and Balakrishnan (1995). The logistic distribution is one of the most important statistical distributions because of its simplicity and also its historical importance as growth curve. The generalized logistic distributions are very useful classes of densities as they possess a wide range of indices of skewness and kurtosis. Therefore, an important application of these distributions is their use in studying robustness of estimators. In bandwidth choice the flexibility of the generalized logistic distribution is used to approximate a wide range of possibly underlying distributions. Obviously other distributions might be used and for any special problem at hand there may be natural other choices. But the generalized logistic seems to be a suitable choice in general.

After estimation of the parameters the value of $f(q_\alpha(x)|x)$ can be estimated

and thus we have estimators for all unknown quantities in the formula of the asymptotically optimal bandwidth.

The above three steps build a framework of the bandwidth choice selector which clearly could be varied at several stages. So for global bandwidth choice in steps (ii) and (iii) other procedures might be used. In step(iii) the local likelihood might be based on an different distribution family. If there is further information about the underlying data generating process available, e.g. symmetry of the conditional distribution or heavy tails, this can be considered in the selection of the distribution family. The above used settings are very general. Let us demonstrate their applicability in some simulation examples in the next section.

3 Simulation examples

In this section some simulation results are presented. Two different densities are chosen. In one example the true underlying distribution is exponential with density

$$f(y) = se^{-sy-1}1_{\{y>-1/a\}}(y),\ s > 0. \tag{10}$$

This distribution is asymmetric and has expectation Zero for all $a > 0$. With $x = 1, 2, ..., 600$ we chose

$$s = 1.5 + \sin(\frac{x}{100}\pi) \tag{11}$$

Thus for $g(x)$ an equidistant design is used. The second distribution under study is the lognormal distribution also with scale parameter s as defined in (11). The generalized logistic distribution is intentionally not used as data generating distribution so that the flexibility of the above algorithm is demonstrated.

The two data setting are quite extreme as Figure 1 shows. This figure presents two data sets generated by the two distributions. The exponential data are very smooth and not really exciting. In contrast to the lognormal data where strong

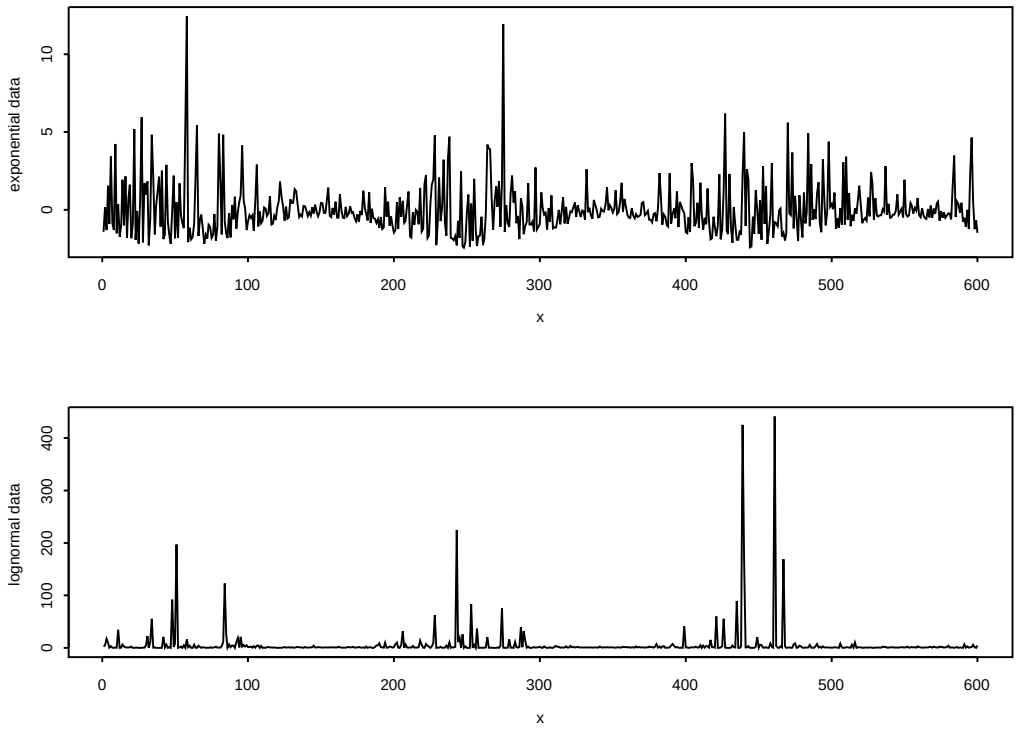

Figure 1: Two simulated data sets with scale function as defined in equation (11)

swings can be observed.

In both settings our aim is to estimate the conditional $0.75-$quantiles. The true quantile functions are presented in Figure 2 and 3. They both look identical but mind the different scales on the ordinates.

To evaluate the resulting quantile estimates for each setting 100 repetitions are calculated. Local linear quantile estimation with locally chosen bandwidths is used and compared with the local linear quantile estimation based on a global bandwidth chosen by cross-validation. The resulting local MSE are shown in Figure 4 and 5.

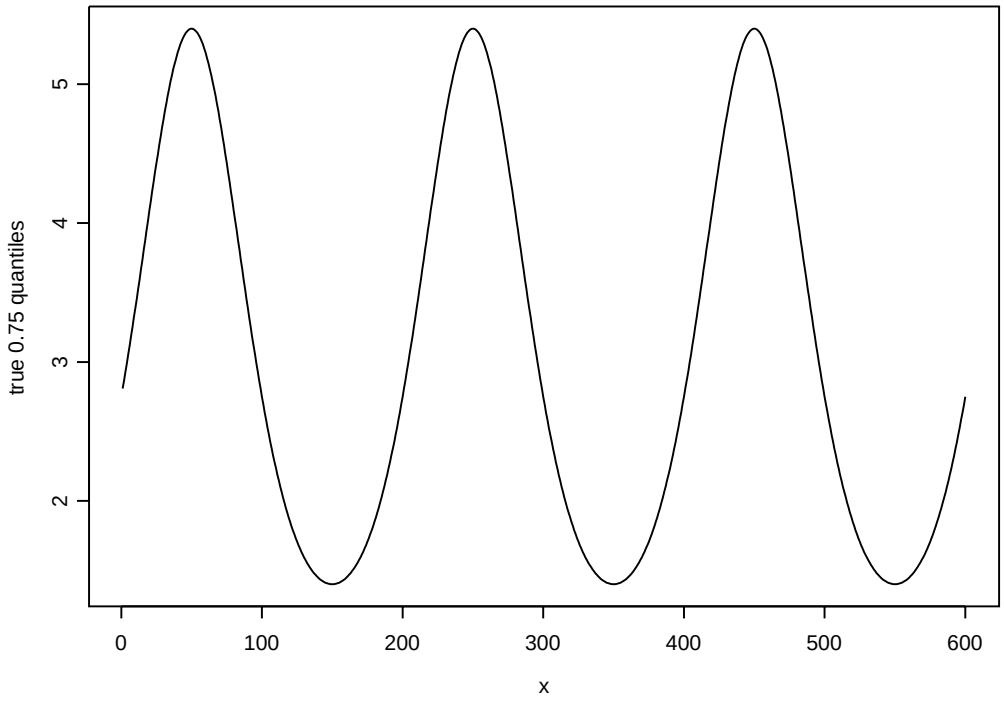

Figure 2: True 0.75−quantiles for the lognormal distribution

Figure 4 contains the "brave" case of exponential data. It can bee seen that relating to the MSE, estimation based on local bandwidth choice and estimation with a global bandwidth selected by cross-validation perform almost identical. Although, there are changes in the components of the MSE. Compared to the global procedure local bandwidth choice using the above algorithm leads to an increase in the bias but to an decrease in the variance part. But local bandwidth choice seems to be not really necessary in this case. On the other hand there is also no disadvantage using it.

A different situation presents Figure 5. In this more extreme data situation local bandwidth choice clearly beats the global method. In the peaks of the quantile function local bandwidth choice leads to a considerable reduction of the MSE.

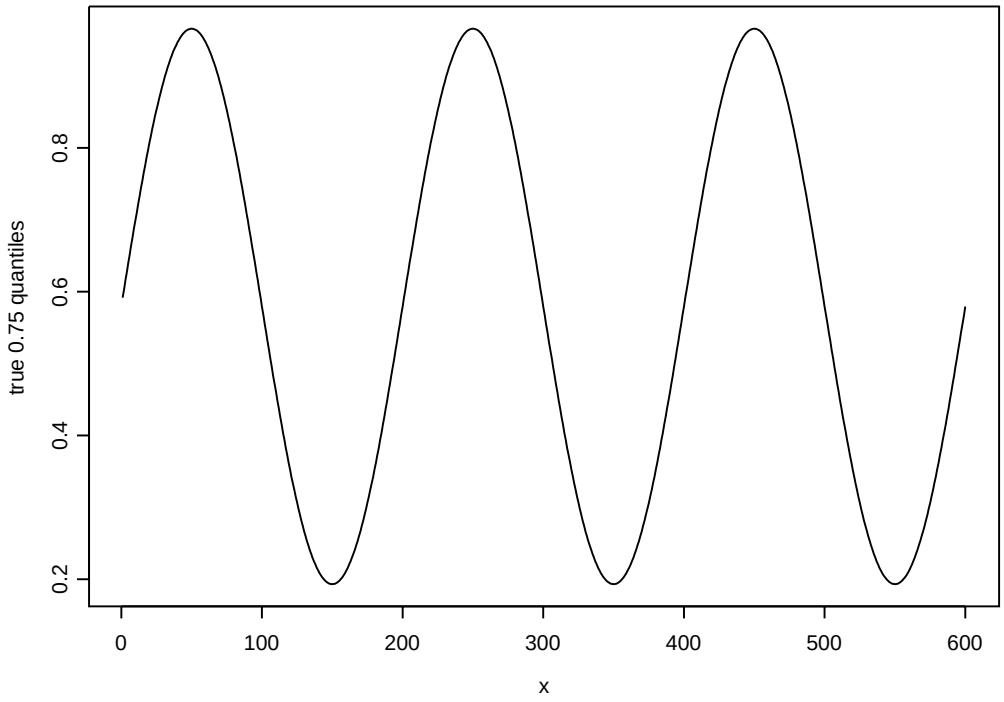

Figure 3: True 0.75−quantiles for the exponential distribution

Finally the ability of the local likelihood approach based on the generalized logistic distribution to approximate the behaviour of the underlying lognormal is demonstrated . Figure 6 shows for one example the difference between the local likelihood based density estimation in step (iii) and the values using the true underlying lognormal distribution. The conditional quantiles are estimated as described in step (ii) of the algorithm. The figure shows that the local likelihood estimate is quite reasonable.

To sum up the two examples it can be stated that the presented algorithm works well. Local bandwidth choice is not needed in general. But there are data situations

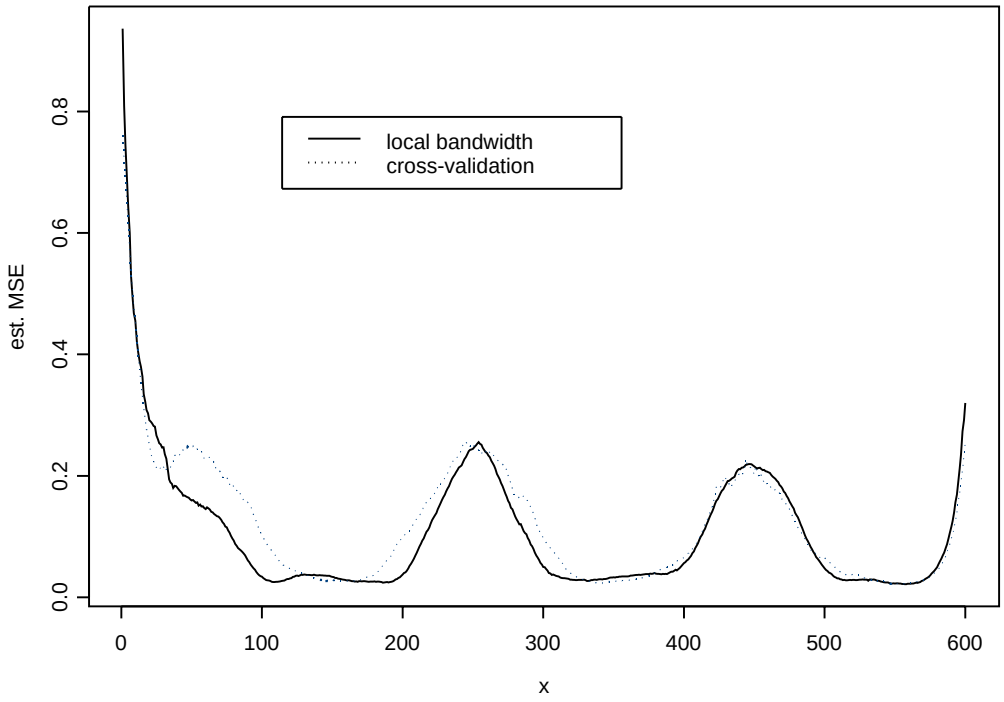

Figure 4: Simulated MSE for local and cross-validation bandwidth choice with exponential data

as demonstrated in the lognormal example, where local bandwidth choice leads to remarkable improvements about the global choice.

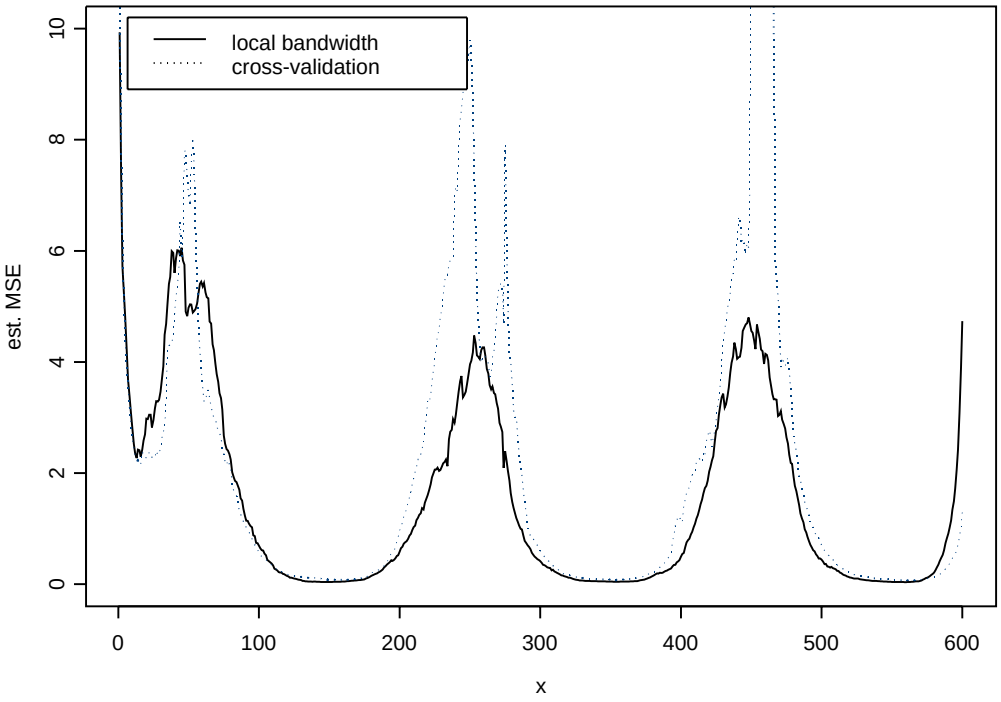

Figure 5: Simulated MSE for local and cross-validation bandwidth choice with log-normal data

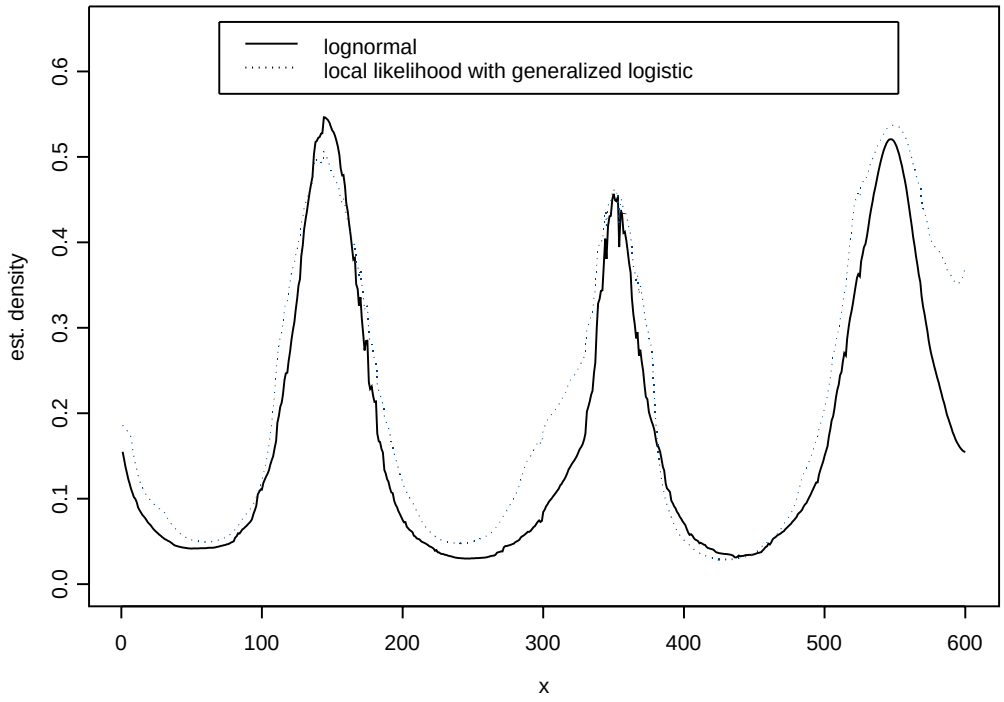

Figure 6: Example of the calculated conditional density at $\hat{q}_{0.75}(x)$ (first using the true underlying lognormal density and second using approximation with estimated generalized logistic density

4 References

Abberger K. (1998). Cross-validation in nonparametric quantile regression. *Allgemeines Statistisches Archiv* 82, 149-161.

Abberger K., Heiler S. (2000). Simulataneous estimation of parameters for a generalized logistic distribution and application to time series models. *Allgemeines Statistisches Arichv* 84, 41-50.

Fan J., Gijbels I. (1996). *Local polynomial modeling and its applications.* Chapman and Hall, London.

Heiler S. (2000). Nonparametric time series analysis. In: *A course in time series analysis*, edited by D. Pena and G.C. Tiao. John Wiley, New York.

Johnson N.L., Kotz S., Balakrishnan N. (1995). *Continuous univariate distributions, volume 2.* John Wiley, New York.

Koenker R. Basset G. (1978). Regression quantiles. *Econometrica* 46, 33-50.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* 84, 276-283.

Yu K., Jones M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association* 93, 228-237.