Penalizing function based bandwidth choice in nonparametric quantile regression

Klaus Abberger, University of Konstanz, Germany

Abstract:

In nonparametric mean regression various methods for bandwidth choice exist. These methods can roughly be divided into plug-in methods and methods based on penalizing functions. This paper uses the approach based on penalizing functions and adapt it to nonparametric quantile regression estimation, where bandwidth choice is still an unsolved problem. Various criteria for bandwitdth choice are defined and compared in some simulation examples.

Key Words: nonparametric quantile regression, bandwidth choice, cross-validation, penalizing functions

1 Introduction

Although most regression investigations are concerned with the regression mean function other aspects of the conditional distribution of Y given X are also often of interest. For fixed $\alpha \in (0,1)$, the quantile regression function gives the αth quantile $q_\alpha(x)$ in the conditional distribution of a response variable Y given the value $X = x$. It can be used to measure the effect of covariates not only in the center of a population, but also in the lower and upper tails. Especially of interest is the case where the data pattern shows heteroscedasticities and asymmetries.

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

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

ϕ and Φ are the standard normal density and distribution function and h_{mean} is a bandwidth choice for regression mean estimation with one of the several existing methods. As it can be seen this procedure leads to identical bandwidths for the α and $(1-\alpha)$ quantiles. Although this strategy might work very well in some situations our special interest lies in asymmetric data patterns where the above rule is to restrictive.

Abberger (1998) adapts the cross-validation idea to kernel quantile regression and presents some simulation examples. Also asymmetric data patterns based on the lognormal distribution are included.

This paper tries to use penalizing function based criteria to choose the bandwidth in nonparametric quantile regression. In the next section these criteria are presented and simulation examples are discussed in Section 3.

2 Quantile estimation and bandwidth choice

A locally weighted linear quantile regression estimator is defined by setting $\hat{q}_\alpha(x) = \hat{a}$, where $\hat{a}$ and $\hat{b}$ minimize

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

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

$$\rho_\alpha(u) = \alpha 1_{\{u \geq 0\}}(u) \cdot u + (\alpha - 1) 1_{\{u < 0\}}(u) \cdot u \tag{3}$$

introduced by Koenker and Basset (1978) in connection with the parametric quantile regression. For a discussion of this estimator see Heiler (2000) or Yu and Jones (1998), which also derive the MSE of this estimator. To calculate $\hat{q}_\alpha$ we use an iteratively reweighted least squares algorithm. Initial estimates are conditional quantiles calculated with a kernel estimator of the Nadaraya-Watson type (see Heiler (2000)).

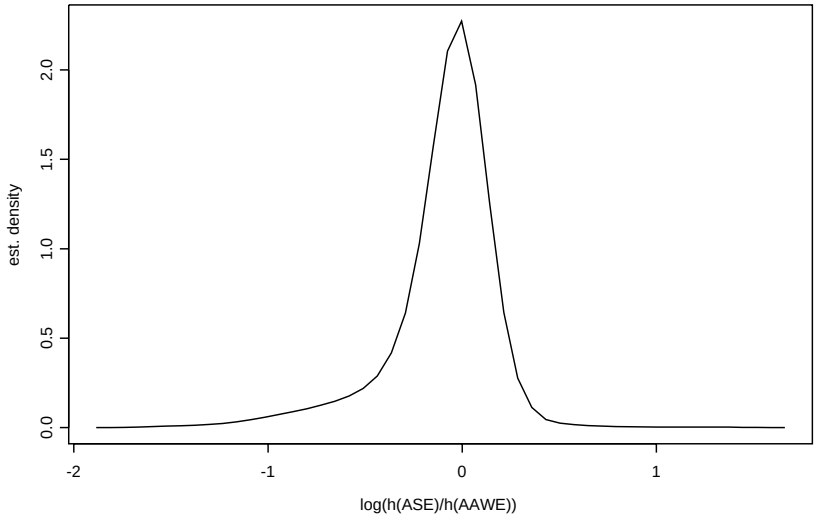

Figure 1: Estimated density of log differences between ASE and AAWE optimal bandwidths for simulated data

Estimation by minimizing equation (2) can be interpreted as M-estimator or in the notation of Bickel and Doksum (2001) as minimum contrast estimate with contrast function ρ_α. In general they define a discrepancy function

$$D(\theta_0, \theta) \equiv E_{\theta_0} \rho(Y, \theta) \tag{4}$$

as a function of θ which measures the (population) discrepancy between θ and the true value θ_0.

Nonparametric estimation of quantile regression requires the choice of a bandwidth. In nonparametric mean regression procedures for bandwidth choice usually ground on the MSE. Various definitions of the optimal bandwidth are available. One candidate is the bandwidth that minimizes MISE (mean integrated squared error) for the given sample size and design. This bandwidth is optimal with respect to the average performance over all possible data sets for a given population, rather than for the performance for the observed data set. Another choice is the bandwidth that minimizes the average squared error (ASE) for the observed data set. Between these two concepts we chose the later one. For further discussion of this issue see e.g. Mammen (1990), Grund et al. (1994), Härdle (1988).

Another natural choice in quantile regression is based on the discrepancy function (4). It is

$$E[\rho_\alpha(Y - m(x))] = \alpha(\mu_Y(x) - m(x)) + \int_{-\infty}^{m(x)} F(y|x)dy \tag{5}$$

and thus the optimal bandwidth is that one for which the corresponding quantile estimator minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{-\infty}^{\hat{q}_\alpha(X_i)} F(y|X_i)dy - \alpha\hat{q}_\alpha(X_i)\right\}. \tag{6}$$

In the sequel this criterion will be called average alpha weighted error (AAWE).

The difference between ASE (which in quantile regression is $1/n\sum(q_\alpha(X_i) - \hat{q}_\alpha(X_i))^2$) and AAWE is demonstrated for a data pattern which is further considered in the simulation examples section. The true underlying distribution is exponential with density

$$f(y) = ae^{-ay-1}1_{\{y>-1/a\}}(y), a > 0. \tag{7}$$

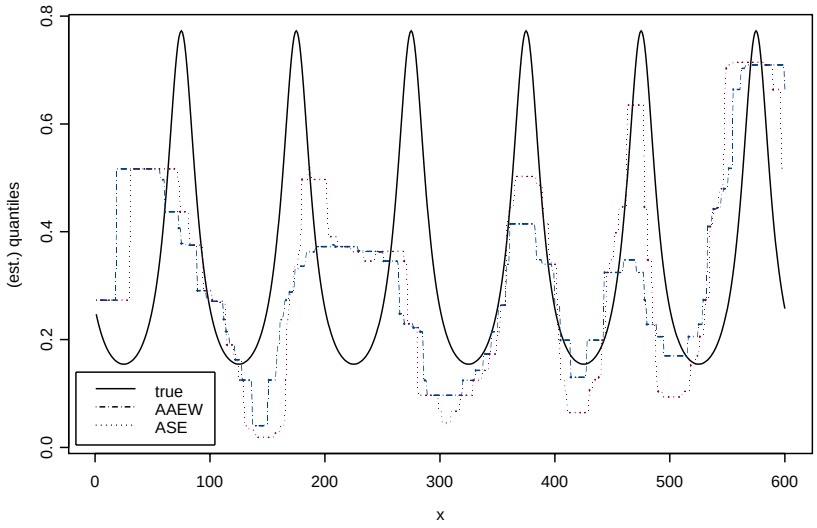

Figure 2: Estimated 0.75 quantiles with ASE and AAWE optimal bandwidths for a simulated data set

This density is asymmetric and has expectation Zero for all $a > 0$. With $x = 1, ..., 600$ we chose $a = 1.5 + \sin(\frac{x}{100} 2\pi)$. For 1000 repetitions the bandwidths minimizing the average errors estimating the 0.75-quantiles with a kernel estimator are calculated. Figure 1 shows the density of $\log(h_{ASE}/h_{AAWE})$. There is a high peak at 0 indicating that the chosen bandwidths coincide quite often. But there is also a slight left skewness observable. This indicates a tendency of the AAWE method to smooth stronger than the ASE procedure. Figure 2 shows a "typical" example where the ASE bandwidth is smaller than the AAWE bandwidth. Since the conditional distribution in the peaks is much flatter than in the valleys where the conditional distribution is very steep, derivations in the valleys are in the AAWE sense more important than errors in the peaks. This perspective is quite natural for quantile estimation, especially when we think of doing quantile forecasts.

In local linear mean regression the estimators are usually linear, that means they are of the form $\hat{\mathbf{y}} = \hat{m}(\mathbf{x}) = H\mathbf{y}$, where the matrix H is commonly called the smoother matrix and depends on $\mathbf{x}$ but not on $\mathbf{y}$. The trace of H can be interpreted as the effective number of parameters used in the smoothing (e.g. Hastie and Tibshirani (1990), sec. 3.5). One possible strategy to find a suitable smoothing parameter is to chose the bandwidth which is the minimizer of

$$\log(\hat{\sigma}^2) + \Psi(H), \text{ where} \tag{8}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{m}_h(X_i)\}^2 \tag{9}$$

and $\Psi(\cdot)$ a penalty function designed to decrease with increasing smoothness of $\hat{m}_h$. Common choices of Ψ lead to GCV ($\Psi(H) = -2\log\{1 - tr(H)/n\}$), Rice's T ($\Psi(H) = -\log\{1 - 2tr(H)/n\}$) and AIC_c (Hurvich et al. (1998)) ($\Psi(H) = \{1 + tr(H)/n\}/\{1 - [tr(H) + 2]/2\}$).

These smoothing parameter selectors can be adapted to quantile regression estimation. The first modification concerns $\log(\hat{\sigma}^2)$. Since the quantile estimator (2) falls into the class of M-estimators we can proceed as usual in M-estimation (see e.g. Hampel et al. (1986)) and interpret the ρ_α function as "-loglikelhood=ρ_α ". So the AIC criterion and all the other above mentioned criteria can be adapted by using $\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(y_i - \hat{q}_\alpha(x_i))$ instead of $\hat{\sigma}$.

The second modification concerns the smoother matrix H. Estimator (2) does not lead to a linear estimator $\hat{\mathbf{y}} = H\mathbf{y}$. Because the actual estimator is carried out by iteratively reweighted least squares the smoother matrix H can be approximated by the implied smoother matrix from the last iteration of the iteratively reweighted least squares fit of the model.

With these modifications we arrive at the following strategy to find a suitable smoothing parameter for local linear quantile regression: choose the bandwidth

to be the minimizer of

$$2\log\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(y_i-\hat{q}_\alpha(x_i))\right)+\Psi(H), \tag{10}$$

where $\Psi(\cdot)$ is one of the above mentioned penalizing functions and H the approximative smoother matrix.

3 Simulation examples

In this section, some simulation results are presented. The underlying density functions were of the exponential type shown in equation (7). The two models

$$\text{Model I:} \quad a = 1.5+\sin(\frac{x}{100}2\pi) \tag{11}$$

$$\text{Model II:} \quad a = 10\cdot exp(-1/200\cdot x) \tag{12}$$

with $x = 1,...,400$ are considered. For each setting 100 repetitions were calculated. The 0.25- and 0.75- quantiles were estimated for both models. Bandwidths are chosen with the help of the above discussed methods based on penalizing functions and in addition with the cross-validation method which chooses

$$h_{CV} = \min_h\left\{\sum_{i=1}^{n}\rho_\alpha(Y_i-\hat{q}_\alpha^{(-i)}(X_i))\right\}, \tag{13}$$

with $\hat{q}_\alpha^{(-i)}(X_i)$ the so called leave-one-out estimator. This is the estimator for the conditional quantile at X_i which is calculated without the observation (Y_i, X_i).

To avoid boundary effects only the 200 observations $x = 101,...,300$ in the middle are used for bandwidth choice.

The estimated densities of $\log(h_{CV}/h_{AAWE})$, $\log(h_{GCV}/h_{AAWE})$ and $\log(h_{AIC_c}/h_{AAWE})$ for the two quantiles and both models are shown in the Figures 3-6. We also calculated Rice's T but the results are quite similar to the AIC_c criterion so these results are not shown in the graphs.

With Model I the arithmetic mean of h_{AAWE} for the 0.25 quantiles is 19.95 and for the 0.75 quantiles the mean is 58.1. This difference in the means confirms the need of methods which can handle asymmetric data patterns.

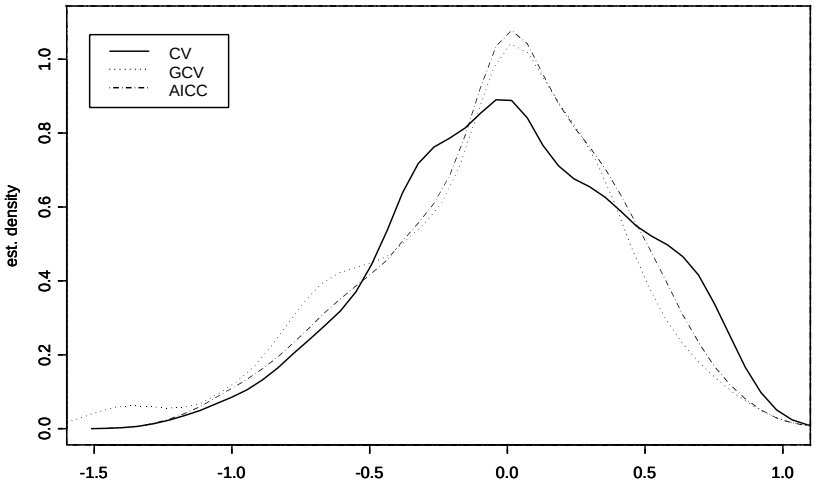

Figure 3: Estimated densities of $\log(h./h_{AAWE})$ for 0.25 quantiles of Model I

Figure 3 shows the results for the 0.25 quantiles of Model I. The three estimated densities have all modi around Zero. But the peaks for the penalizing methods are higher and sharper than for the cross-validation method where the density is flatter.

Also in Figure 4 which presents the results for the 0.75 quantiles of Model I the cross-validation density is relatively flat. But in this case it is the only density with modus around Zero. The penalizing methods tend to undersmooth.

A similar behaviour can be obtained for Model II visualized in Figure 6 and 7. The mean of h_{AAWE} for $\alpha = 0.25$ is 109.6 and for $\alpha = 0.75$ the mean is

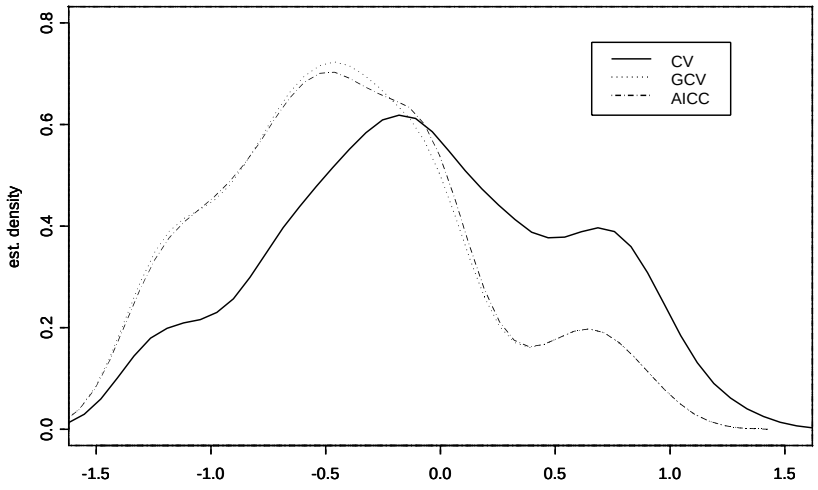

Figure 4: Estimated densities of $\log(h./h_{AAWE})$ for 0.75 quantiles of Model I

245.6. This difference is again a result of the asymmetric density. And just as for Model I the penalizing methods tend to undersmooth the upper quantile.

These results remain unchanged when h_{ASE} is used as reference bandwidth instead of h_{AAWE} because in these examples the difference between h_{ASE} and h_{AAWE} is not such large.

Figure 7 shows the estimated densities for the 0.75 quantiles of Model I but now 100 observations are used for bandwidth choice instead of 200. The penalizing methods still undersmooth but the smaller sample size leads to stronger differences between the methods. The AIC_C method undersmooth less than the GCV method.

To sum up the simulation results it can be stated that the penalizing func-

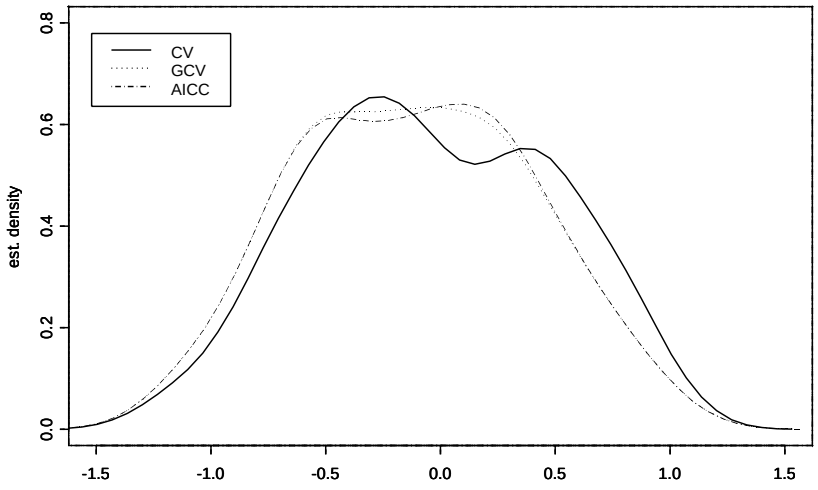

Figure 5: Estimated densities of $\log(h./h_{AAWE})$ for 0.25 quantiles of Model II

tion based methods for bandwidth choice can lead to a reduction in variability compared with the cross-validation method. But for this we have to take into account the tendency of penalizing methods to undersmooth when large bandwidths are appropriate. Maybe this disadvantage can be get under control with the development of adapted penalizing functions. Simulations based on smaller sample sizes show that the AIC_C penalizing function undersmooth less than some other penalizing functions.

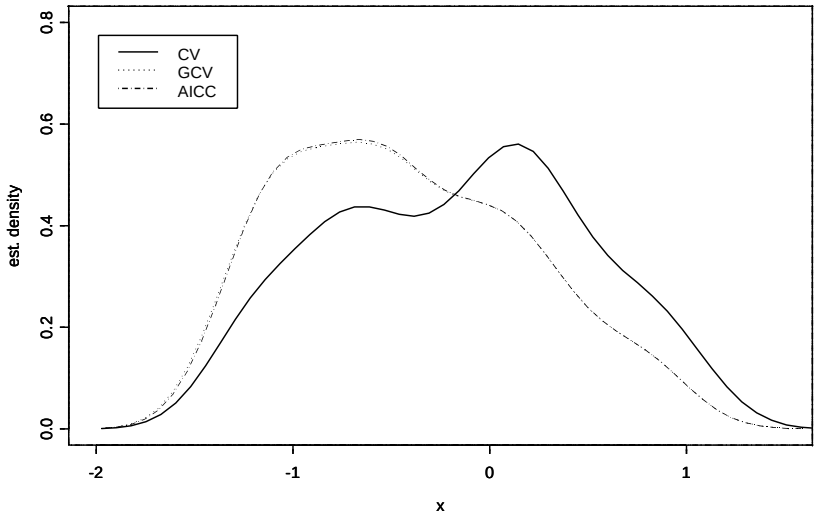

Figure 6: Estimated densities of $\log(h./h_{AAWE})$ for 0.75 quantiles of Model II

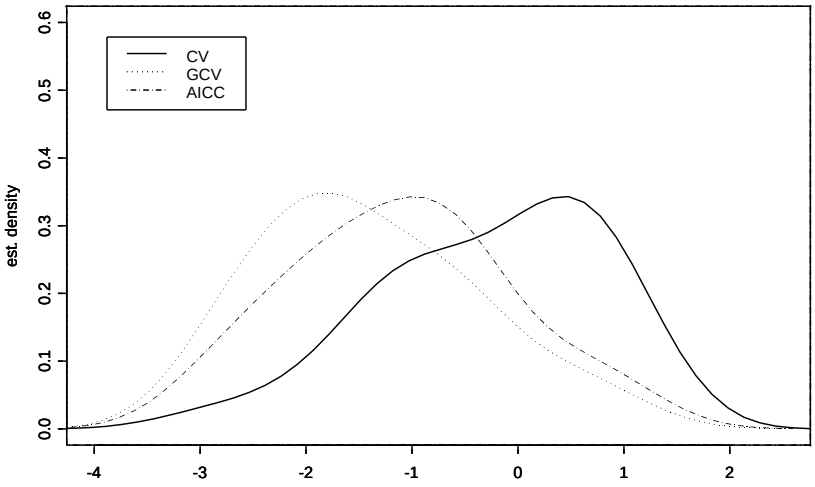

Figure 7: Estimated densities of $\log(h./h_{AAWE})$ for 0.75 quantiles of Model I (bandwidth choice with 100 observations)

4 Literature

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

Bickel P. J. , Doksum K. A. (2001): Mathematical Statistics. Longman Higher Education, New Jersey.

Grund B., Hall P., Marron J.S (1994): Loss and risk in smoothing parameter selection. Journal of Nonparametric Statistics, 4, 107-132.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986): Robust Statistics. Wiley, New York.

Härdle W., Hall P., Marron J:S. (1988): How far are automatically chosen regression smoothing parameters from their optimum? Journal of the American Statistical Association, 83, 86-101.

Hastie T.J., Tibshirani R.J. (1990): Generalized Additive Models. Chapman and Hall, New York.

Heiler S. (2000): Nonparametric Time Series Analysis. In: A Course in Time Series Analysis, edited by D. Pena and G.C. Tiao. John Wiley, London.

Hurvich C.M., Simonoff J.S., Tsay C.L. (1998): Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. journal of the Royal Statistical Society, Ser. B, 60, 271-293.

Hurvich C.M., Tsay C.L. (1989): Regression and time series model selection in small samples. Biometrika, 76, 297-307.

Koenker R., Bassett G. (1978): Regression quantiles. Econometrica, 46, 33-50.

Koenker R., Portnoy S., Ng P. (1992): Nonparametric estimation of conditional quantile functions. In: L_1-Statistical Analysis and Related methods (ed. Y. Dodge), North-Holland, New York.

Mammen E. (1990): A short note on optimal bandwidth selection for kernel estimators. Statistics and Probability Letters, 9, 23-25.

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