Comment on Article by Spokoiny, Wang and Härdle

by Keming Yu[a,b], Wei Dang[b], Huiming Zhu[c] and Rahim Al-Hamzawi[a]

[a]Brunel University, UK [b]Shihezi University, China and [c]Hunan University, China

1 Introduction

We congratulate the authors of Spokoiny, Wang and Haerdle (henceforth SWH) for their important and interesting contribution to the development of local quantile regression. Quantile regression has been used in a number of disciplines to explore the relationship between the response and covariates at both the center and extremes of the conditional distribution. Since Koenker and Bassett (1978) first introduced the linear quantile regression, nonparametric kernel smoothing quantile regression has attracted much attention in literature (Fan and Gijbels, 1996; Yu and Jones, 1998; Hall, Wolff and Yao, 1999; Cai and Xu, 2008; Dette and Volgushev, 2008; Chen and Mller, 2012; among others). An important issue in nonparametric smoothing techniques is the selection of smoothing parameter or bandwidth. Bandwidth selection in nonparametric smoothing quantile regression requires not only data-driven but also quantile-driven. The main contribution of SWH's paper lies in their adaptive bandwidth selection rule for kernel smoothing quantile regression. That is, their bandwidth selection rule is adaptive and novel, although the regression estimator named qMLE in their equation (8) is simply equivalent to a local polynomial quantile regression or a type of kernel-based weighting 'check function' approach, such as the local linear single-kernel approach of Yu and Jones (1998).

Our discussion is organized as follows. We first comment the asymmetric Laplace distribution (henceforth ALD) based quantile regression approach and bandwidth selection in Sections 2 and 3 respectively. In particular, we point out that SHW's bandwidth selection rule is well-adaptive for smoothing moderate or central quantile curves but may loss adaptation for smoothing extreme quantile curves. We then propose an alternatively adaptive bandwidth selection rule based on a normal scale-mixture representation of ALD and show that this alternative version is well-adaptive for smoothing extreme quantile curves. Finally in Section 4 we point out that adaptive bandwidth selection rules may be able to avoid the problem of crossing quantile curves (calculated for various $\tau \in (0,1)$).

2 ALD

All likelihood functions used may never be the exact ones in practice but some are useful. ALD-based likelihood function unlikely matches many real situations, but it is working in practice.

The local likelihood function in SWH's paper is a local ALD-based likelihood function. This local likelihood function is a nonparametric extension of the parametric ALD-based likelihood function (Yu and Moyeed, 2001). ALD-based inference has nowadays become a powerful tool for formulating different quantile regression techniques, including nonparametric extension of ALD for nonparametric Bayesian quantile regression (Gelfand and Kottas 2003; Kottas and Krnjajic, 2009; Thompson et al. 2010). Moreover, another recent discussion paper in *Bayesian Statistics* by Lum and Gelfand (2012) makes an excellent spatial extension of ALD for spatial data analysis. This is interesting especially given the fact the true underlying distribution in practical problems is almost never ALD. SWH also note on their pages 12 and 14 that likelihood in their method is not necessarily coincide with ALD. In fact, their extensively numerical studies conclude that "mis-specification of (model) error distribution from ALD would not contaminate our results significantly." This is consistent with early attempts carried out by Koenker and Machado (1999) and Yu and Moyeed (2001). Koenker and Machado (1999) develop goodness of fit inference processes for quantile regression based on ALD and show that asymptotic work even if the underlying distribution is not ALD. Yu and Moyeed (2001) and Yu and Stander (2007) argue based on empirical results that even if the underlying distribution is not ALD, the results would be reasonable. This is also consistent to a recently theoretic justification by Sriram, Ramamoorthi and Ghosh (2011). DWH may need to cite some of these publications while they introduce their ALD-based local likelihood function. The local ALD-based likelihood approach in the paper plays an important role of consistency check of estimators and the derivation of the adaptive bandwidth selection rule.

3 Bandwidth selection

There are several methodologies for automatic smoothing parameter selection. One class of methods chooses the smoothing parameter value to minimize a criterion that incorporates both the tightness of the fit and model complexity. Such a criterion can usually be written as a function of the error mean square, and a penalty function designed to decrease with increasing smoothness of the fit. Examples of specific criteria are generalized cross-validation (Craven and Wahba 1979) and the Akaike information criterion (AIC: Akaike 1973). These classical selectors have two undesirable properties when used with local polynomial and kernel estimators: they tend to undersmooth and tend to be non-robust in the sense that small variations of the input data can change the choice of smoothing parameter value significantly. Hurvich, Simonoff, and Tsai (1998) obtained several bias-corrected AIC criteria that limit these unfavorable properties and perform comparably with the plug-in selectors (Ruppert, Sheather, and Wand 1995).

The adaptive bandwidth selection rule in SWH's paper is different from the rule-of-thumb rule of Yu and Jones (1998) and AIC rule of Cai and Xu (2008). It does add a nice option to the bandwidth selection menu for practitioners. But, comparing to a rule-of-thumb rule, the method needs special care to implement in practice. In fact, besides fixing a finite ordered set of candidates of bandwidth, the implementation of the method is subject to three-type of parameter selection: significant level α of test, test power r and a set of critical values. Critical values are simulated via their condition (12) but this condition is not easy to check. Some of us may question: is it worth using sophisticated procedures for choosing the smoothing parameter in quantile regression when one often requires to present or estimate several quantile curves together?

Furthermore, the method may loss adaptation for smoothing extreme quantile curves. Taking the Lidar data analysis as an example, DWH's paper displays the smoothed 90% quantile curve for this data in their Figure 1, which looks good. And this is also true for other moderate or central quantile curves. However, we seem to see from smoothing extreme quantile curves in Figure 1 here, the proposed bandwidth selection rule results in over-smoothing phenomenon, whatever the selection of α and r. Figure 1 displays the smoothed 99% and 1% quantile curves using DWH's method, and shows that when the curves start to switch smoothness, the rule is not adaptive so that the estimated curves are too smoothing out of the data ranges. A possibly theoretical interpretation for this problem is: when $\tau \to 0$, the weighted 'check function' $\rho_\tau(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta}) w_i$ takes constant 0 if $Y_i > \boldsymbol{\psi}_i^T \boldsymbol{\theta}$ (also, when $\tau \to 1$ and if $Y_i < \boldsymbol{\psi}_i^T \boldsymbol{\theta}$). This may result in that the proposed significant test always picks constant bandwidth for smoothing extreme quantile curves although this is not a problem for the local quantile regression estimation equation. We want to point out that this over-smoothing problem will be solved by a new version of adaptive bandwidth selection rule. See the details from the Section 4 below.

3.1 An alternative qMLE and pointwise bandwidth selection

Reed and Yu (2009) and Kozumi and Kobayashi (2011) note that, under the assumption of ALD-based 'working likelihood', the quantile regression model error ϵ can be represented as a scale mixture of normal variable, that is,

$$\epsilon = \mu z + \tau \sqrt{z} e,$$

where $\mu = \frac{1-2\tau}{\tau(1-\tau)}$, $\delta^2 = \frac{2}{\tau(1-\tau)}$, $z \sim Exp(1)$ and $e \sim N(0,1)$, and z and e are independent. Hence, DWH's mode (1) $\Big(Y_i = f(X_i) + \epsilon_i\Big)$ could be re-written as

$$Y_i = f(X_i) + \mu z_i + \delta\sqrt{z_i}e_i.$$

That is, for given $\boldsymbol{z} = (z_1, z_2, ..., z_n)$ and $\boldsymbol{x} = (X_1, X_2, ..., X_n)$,

$$Y_i \sim N\Big(f(X_i) + \mu z_i,\ \delta^2 z_i\Big),$$

i.e., the joint conditional density of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is given by

$$f(\boldsymbol{Y}|\boldsymbol{z}, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\delta\sqrt{z_i}} \exp\{-\frac{(Y_i - f(X_i) - \mu z_i)^2}{2\delta^2 z_i}\}.$$

Clearly, if $\boldsymbol{z}$ is fixed in advance, then the local log-likelihood (DWH's equation (7)) can be replaced by a Gaussian-type of local likelihood function:

$$L_{New}(W, \boldsymbol{\theta}) \equiv -\log(\sqrt{2\pi}\delta)\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\log(z_i)\, w_i -$$

$$-\frac{1}{2\delta^2}\sum_{i=1}^{n}\frac{(Y_i - f(X_i) - \mu z_i)^2}{z_i} w_i - \sum_{i=1}^{n} z_i w_i.$$

Now, once a local pth-degree polynomial $\psi_i^T\boldsymbol{\theta}$ is used to approximate $f(x)$ at $X = x$, the corresponding local qMLE at x could be defined via maximization of $L_{New}(W, \boldsymbol{\theta})$ above:

$$\begin{aligned}\tilde{\theta}(x) &\equiv \Big(\tilde{\theta}_0(x), \tilde{\theta}_1(x), ..., \tilde{\theta}_p(x)\Big) \\ &= argmax_{\boldsymbol{\theta}\in\Theta} L_{New}(W, \boldsymbol{\theta}) \\ &= argmin_{\boldsymbol{\theta}\in\Theta}\sum_{i=1}^{n}\frac{(Y_i - \psi_i\boldsymbol{\theta} - \mu z_i)^2}{\delta^2 z_i} w_i,\end{aligned}$$

where $\tilde{\theta}_0(x)$ estimates $f(x)$, and $\tilde{\theta}_m(x)$ estimates the derivatives of $f(x)$. Further, let $\boldsymbol{\psi} = (\psi_1, .., \psi_n)^T$ and $\boldsymbol{w}_k = diag\Big(\frac{w_1^{(k)}}{\delta^2 z_1}, ..., \frac{w_n^{(k)}}{\delta^2 z_n}\Big)$, we have

$$\tilde{\theta}_k(x) = \Big(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\Big)^{-1}\boldsymbol{\psi}\boldsymbol{w}_k(\boldsymbol{Y} + \mu\boldsymbol{z}).$$

This clearly shows that an alternatively adaptive 'local quantile regression' based on a 'mean regression model' can be developed. In particular, a new adaptive bandwidth selection rule for local quantile regression is proposed. Then, the localized likelihood ratio test for consistent check of estimators will be given by a simple quadratic-type of test statistic:

$$T_{lk} = \Big(\tilde{\theta}_l(x) - \tilde{\theta}_k(x)\Big)^T\Big(\boldsymbol{\psi}\boldsymbol{w}_l\boldsymbol{\psi}^T\Big)\Big(\tilde{\theta}_l(x) - \tilde{\theta}_k(x)\Big).$$

Further, according to Serdyukova (2012), the propagation conditions now become:

for a given $\alpha \in (0,1]$ and $r > 0$ the critical values $\zeta_1, ..., \zeta_K$ satisfy

$$E|(\tilde{\theta}_k(x) - \hat{\theta}_k(x))^T \, (\boldsymbol{\psi} \boldsymbol{w}_k \boldsymbol{\psi}^T) \, (\tilde{\theta}_k(x) - \hat{\theta}_k(x))| \leq \alpha C(p,r),$$

for all $k = 2, ..., K$, where $C(p,r) = 2^r \Gamma(r + p/2)/\Gamma(p/2)$.

We note that the $L_{New}(W, \boldsymbol{\theta})$ involves in a specification of vector $\boldsymbol{z}$, and we point out that $\boldsymbol{z}$ could be fixed in advance via a sample from a data-driven inverse Gaussian distribution. In fact, note that the joint likelihood function of $(\boldsymbol{Y}, \boldsymbol{z})$ is give by

$$f(\boldsymbol{Y}, \boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\tau\sqrt{z_i}} \exp\{-\frac{(Y_i - f(X_i) - \mu z_i)^2}{2\tau^2 z_i}\} \prod_{i=1}^{n} \exp(-z_i).$$

Therefore, the conditional density of $f(\boldsymbol{z}|\boldsymbol{Y})$ is given by

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{Y}) &\propto f(\boldsymbol{Y}, \boldsymbol{z}) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{z_i}} \exp\Big(-\frac{1}{2}[\frac{(Y_i - f(X_i) - \mu z_i)^2}{\delta^2} z_i^{-1} + (\frac{\mu^2}{\delta^2} + 2) z_i]\Big). \end{aligned} \quad (1)$$

That is, $z_i, z_2, ..., z_n$ are iid with a generalized inverse Gaussian (GIG) distribution:

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{Y}) &\propto z_i^{\frac{1}{2}-1} \exp\Big(-\frac{1}{2}[\frac{(Y_i - f(X_i) - \mu z_i)^2}{\delta^2} z_i^{-1} + (\frac{\mu^2}{\delta^2} + 2) z_i]\Big) \\ &\sim GIG\Big(\frac{1}{2}, \frac{(Y_i - f(X_i) - \mu z_i)^2}{\delta^2}, (\frac{\mu^2}{\delta^2} + 2)\Big). \end{aligned}$$

An advantage of this local Gaussian conditional likelihood function over DWH's method is that the derived bandwidth has better adaptation when τ tends to zero or 1. Figure 2 display the bandwidth sequence (upper panel) and smoothed quantile curves for quantiles 99% (left) and 1% (right) based on the Lidar data set, which provide much better fitting than those curves presented in Figure 1. The dependency structure change on smoothness is more adaptive than the bandwidth sequence in Figure 1. This local Gaussian conditional likelihood function method also works well for other moderate or central quantile curves. Figure 3 shows that the method gives quite similar estimates to SWH's method for $\tau = 0.5$ and 0.9 quantile curves.

4 Quantile crossing

Nonparametric quantile regression methods, including kernel smoothing quantile regression, sometimes suffers quantile crossing. But the proposed bandwidth selection rule in SWH's method seems to have no quantile crossing phenomenon

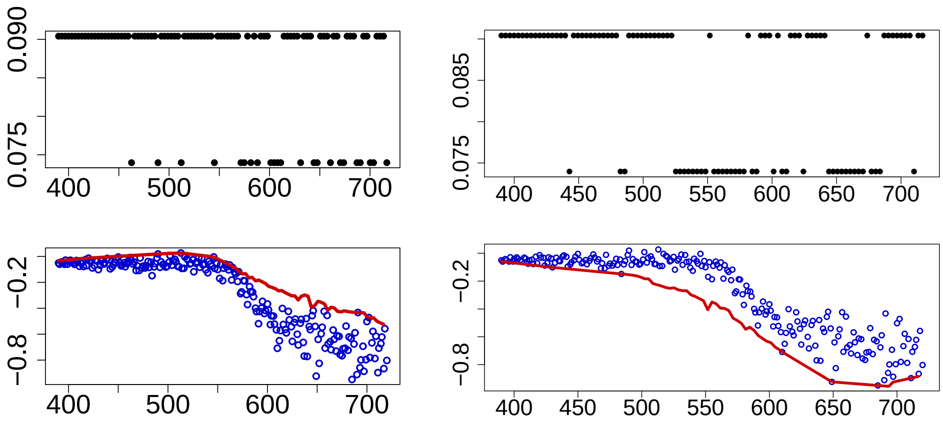

Fig. 1. Smoothed quantile curves for Lidar data with $\tau = 0.99$ (left) and $\tau = 0.01$ (right) by DWH's local likelihood function

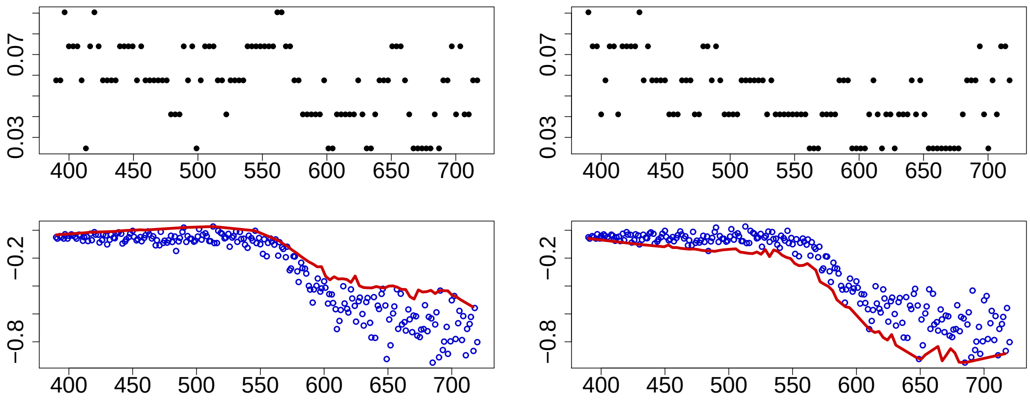

Fig. 2. Smoothed quantile curves for Lidar data with $\tau = 0.99$ (left) and $\tau = 0.01$ (right) by local Gaussian conditional likelihood function

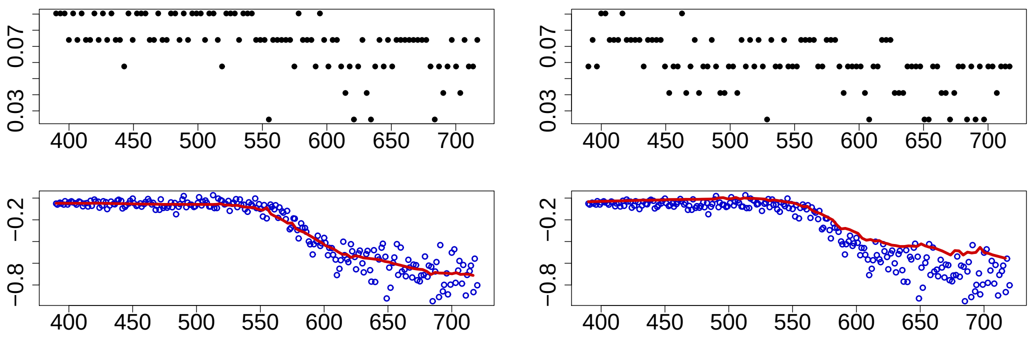

Fig. 3. Smoothed quantile curves for Lidar data with $\tau = 0.5$ (left) and $\tau = 0.9$ (right) by local Gaussian conditional likelihood method

when several smoothed quantile curves are provided together. See the top figure of Figure 4, which displays five quantile curves with $\tau = c(0.05, 0.25, 0.50, 0.75, 0.95)$ for the Lidar data, using DWH's adaptive bandwidth and proposed Guaussian likelihood based adaptive bandwidth respectively. This indicates the advantage of local bandwidth selection rule. Whereas most of published articles on the topic, which include constrained smoothing spline (He, 1997; Bondell, Reich and Wang, 2010), double-kernel smoothing (Yu and Jones, 1998; Jones and Yu, 2007) and monotone constraint on conditional distribution function (Hall, Wolff and Yao, 1999; Dette and Volgushev, 2008) among others, focus on development of new methods rather than adaptive bandwidth selection for avoiding quantile crossing. DWH show, even working with 'local constant' kernel smoothing quantile regression via

$$\hat{q}_\tau(x) = argmin_a \sum_{i=1}^{n} \rho_\tau(Y_i - a) K_h(x - X_i),$$

adaptive bandwidth selection rule may not have quantile crossing either.

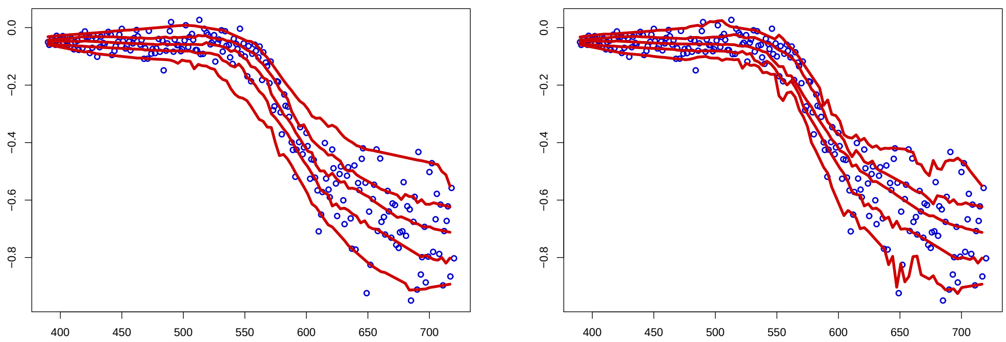

Fig. 4. Smoothed quantile curves for Lidar data with $\tau = c(0.05, 0.25, 0.5, 0.75, 0.95)$ with SWH's method (left) and with local Gaussian conditional likelihood method (right)

The right figure of Figure 4 shows that non-quantile crossing is also true for the rule in Section 3.1, which is based on local Gaussian conditional likelihood function.

Acknowledgments: The research was partially supported by the National Science Foundation of China (No 11261048).

5 References

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In: B. N. PETROV and F. CSAKI, eds. Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267–281.

Bondell, H.D., Reich, B.J. and Wang, H., 2010. Noncrossing quantile regression curve estimation. Biometrika 97, 825-838.

Cai, Z. and Xu, X., 2008. Nonparametric quantile estimations for dynamic smooth coefficient models. Journal of the American Statistical Association 103, 1595–1608.

Chen, K. and Mller, H.G., 2012. Conditional quantile analysis when covariates are functions, with application to growth data. Journal of the Royal Statistical Society: B, 74, 67–89.

Craven, P. and Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerical Mathematics 31, 377–403.

Dette, H. and Volgushev, S., 2008. Non-crossing non-parametric estimates of quantile curves. Journal of the Royal Statistical Society B 70, 609-627.

Fan, J. and Gijbels, I., 1996. Local polynomial modelling and its applications. Chapman and Hall.

Gelfand, A. E. and Kottas, A., 2003. Bayesian Semiparametric Regression for Median Residual Life. Scandinavian Journal of Statistics 30, 651–665.

Hall, P., Wolff, R. C. and Yao, Q., 1999. Methods for estimating a conditional distribution function. Journal of the American Statistical Association 94, 154–163.

He, X., 1997. Quantile curves without crossing. The American Statistician 51, 186–92.

Hurvich, C.M., Simonoff, J. S. and Tsai, C., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society: B 60, 271–293.

Kottas, A. and Krnjajic, M., 2009. Bayesian nonparametric modeling in quantile regression. Scandinavian Journal of Statistics 36, 297–319.

Jones, M.C. and Yu, K., 2007. Improve Double Kernel Local Linear Quantile Regression. Statistical Modelling 7, 377–389.

Koenker, R. and Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc 94, 1296–1310.

Koenker, R. and Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Kozumi, H. and Kobayashi, G., 2011. Gibbs sampling methods for Bayesian quantile regression. Journal of Statistical Computation and Simulation 81, 1565–1578.

Lepski, O., Mammen, E. and Spokoiny, V., 1997. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. The Annals of Statistics 25, 929–947.

Lum, K. and Gelfand, A.E., 2012. Spatial Quantile Multiple Regression Using the Asymmetric Laplace Process (with discussion). Bayesian Analysis 7, 235–258.

Reed, C. and Yu, K., 2010. Efficient Gibbs Sampling for Bayesian quantile regression. Brunel University Mathematics Technical Report.

Ruppert, D., Sheather, S.J. and Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association 90, 1257–1270.

Serdyukova, N., 2012. Spatial adaptation in heteroscedastic regression: Propagation approach. Electron. J. Stat. 6, 861–907.

V. Spokoiny, C. Vial, Parameter tuning in pointwise adaptation using a propagation approach, Ann. Statist., 37 (2009) pp. 2783–2807.

Sriram, K., Ramamoorthi, R.V., Ghosh, P., 2011. Posterior consistency of Bayesian quantile regression under a mis-specified likelihood based on asymmetric Laplace density. Indian Institute of Management Bangalore and Michigan State University.

Thompson, P., Cai, Y., Moyeed, R., Reeve, D., and Stander, J., 2010. Bayesian nonparametric quantile regression using splines. Computational Statistics and Data Analysis 54, 1138–1150.

Yu, K. and Jones, M. C., 1998. Local linear quantile regression. Journal of the American Statistical Association 93, 228–237.

Yu, K. and Moyeed, R., 2001. Bayesian Quantile Regression. Statistics and Probability Letters 54, 437–447.

Yu, K. and Stander, J., 2007. Bayesian Analysis of a Tobit Quantile Regression Model. Journal of Econometrics 137, 260–276.