Issues on quantile autoregression *

Jianqing Fan and Yingying Fan

We congratulate Koenker and Xiao on their interesting and important contribution to the quantile autoregression (QAR). The paper provides a comprehensive overview on the QAR model, from probabilistic aspects, to model identification, statistical inferences, and empirical applications. The attempt to integrate the quantile regression and the QAR process is intriguing. It demonstrates surprisingly that nonparametric coefficient functions can be estimated at root-n rate for the QAR processes. The authors then put forward some useful tools for testing significance of lag-variables and asymmetric dynamics of the time series. We appreciate the opportunity to comment several aspects of this article.

1 Connections with varying coefficient models

QAR is closely related to the functional-coefficient autoregressive (FCAR) model. In the time series context, Cai *et al.* (2000b) proposed the following FCAR model for capturing the nonlinearity of a time series:

$$Y_t = \alpha_0(U_t) + \alpha_1(U_t)Y_{t-1} + \cdots + \alpha_p(U_t)Y_{t-p} + \varepsilon_t, \quad (1)$$

where U_t is a thresholding variable and $\{\varepsilon_t\}$ is a sequence of independent innovations. In particular, when $U_t = Y_{t-d}$ for some lag d, the model is called a functional autoregressive model (FAR) by Chen and Tsay (1993). Varying coefficient models have been widely used in many aspects of statistical modeling. See, for example,

*Jianqing Fan is Professor (E-mail: *jqfan@princeton.edu*) and Yingying Fan is Ph.D. Candidate (E-mail:*yingying@princeton.edu*), Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. The authors thank the AE for helpful comments. The work was partially supported by the NSF grant DMS-0532370 and NIH grant R01 GM072611.

Hastie and Tibshirani (1993), Carroll *et al.* (1998), Cai *et al.* (2001) for applications to generalized linear models, Brumback and Rice (1998), Fan and Zhang (2000), and Chiang *et al.* (2001) for analysis of functional data, Lin and Ying (2001) and Fan and Li (2004) for analysis of longitudinal data, Tian *et al.* (2005) and Fan *et al.* (2006) for applications to the Cox hazards regression model, and Fan *et al.* (2003), Hong and Lee (2003), Mercurio and Spokoiny (2004) for applications to financial modeling. These are just a few examples that testify the flexibility, popularity and explanatory power of the varying coefficient models. In the same vein, it reflects the importance of the QAR model.

What makes QAR differ from the FCAR model or more generally the varying coefficient model is that the variable U_t is unobservable and $\varepsilon_t = 0$. This makes estimating techniques completely different. For example, in the varying-coefficient model, the coefficient functions in (1) are estimated via localizing on U_t (which are observable), while in the QAR model, the coefficient functions are estimated via quantile regression techniques. As a result, two completely different sets of rates of convergence are obtained. The former model admits a nonparametric rate, while the latter reveals the parametric one.

Despite their differences in statistical inferences, QAR is a subfamily of models of FCAR as far as the probabilistic aspects are concerned. Hence, the stochastic properties established in FCAR are applicable directly to QAR. Chen and Tsay (1993) have given sufficient conditions for the solution to (1) to admit a stationary and ergodic solution. With some modifications of their proof, it can be shown that if $\alpha_j(\cdot)$ is bounded by c_j for all j and all roots of the characteristic function

$$\lambda^p - c_1\lambda^{p-1} - \cdots - c_p = 0$$

are inside the unit circle, then there exists a stationary solution that is geometrically ergodic.

2 Identifiability of the model

An important observation made by Koenker and Xiao is that if given $Y_{t-p}, \cdots, Y_{t-1}$, the function

$$\beta_t(u) = \theta_0(u) + \theta_1(u)Y_{t-1} + \cdots + \theta_p(u)Y_{t-p} \tag{2}$$

is strictly increasing in u, then $\beta_t(\tau)$ is the conditional τ-quantile of Y_t given $Y_{t-p}, \cdots, Y_{t-1}$. Since the conditional τ-quantile is identifiable under some mild conditions, the identifiability condition becomes that with probability 1, the QAR model generates at least $(p+1)$ linearly independent vectors of form $\mathbf{Y}_t = (1, Y_{t-1}, \cdots, Y_{t-p})^T$. In other words, letting

$$\mathcal{T} = \{t : \beta_t(u) \text{ is strictly increasing in } u\}, \tag{3}$$

there are at least $(p+1)$ distinct time points $t_i \in \mathcal{T}$ such that $\mathbf{Y}_{t_i}$ are linearly independent for each realization. A natural and open question is what kind of population would generate, with probability one, the samples that satisfy the above condition.

The aforementioned identifiability conditions are hard to check. However, they are needed for not only connections to the quantile regression, but also the issue of identifiability. To see this, let us look at specific case where $p = 0$, in which $Y_t = \theta_0(U_t)$. Clearly, $\theta_0(\cdot)$ is the quantile function of Y_t only when $\theta_0(\cdot)$ is monotone increasing. When this condition is violated, the model is not necessarily identifiable. For example, $Y_t = |U_t - 0.5|$ and $Z_t = U_t - 0.5I(U_t > 0.5)$ have identically the same distribution, but have very different $\theta_0(\cdot)$.

We would like to note that the QAR(p) model

$$Y_t = \theta_0(U_t) + \theta_1(U_t)Y_{t-1} + \cdots + \theta_p(U_t)Y_{t-p} \equiv \theta(U_t)^T X_t$$

is indifferentiable from the model

$$Y_t = \theta(1 - U_t)^T X_t,$$

where $X_t = (Y_{t-1}, \cdots, Y_{t-p})^T$. Thus, if $\theta(\tau)$ is a solution, so is $\theta(1 - \tau)$.

3 Fitting and diagnostics

Koenker and Xiao estimate the coefficient functions $\theta(\tau)$ with the quantile regression:

$$\min_\theta \sum_t \rho_\tau(Y_t - X_t^T\theta). \tag{4}$$

This convex optimization usually exists. The resulting estimates $\hat{\theta}(\tau)$ are consistent estimate of the parameter

$$\theta^*(\tau) = \mathrm{argmin}_\theta E\rho_\tau(Y_t - X_t^T\theta) \tag{5}$$

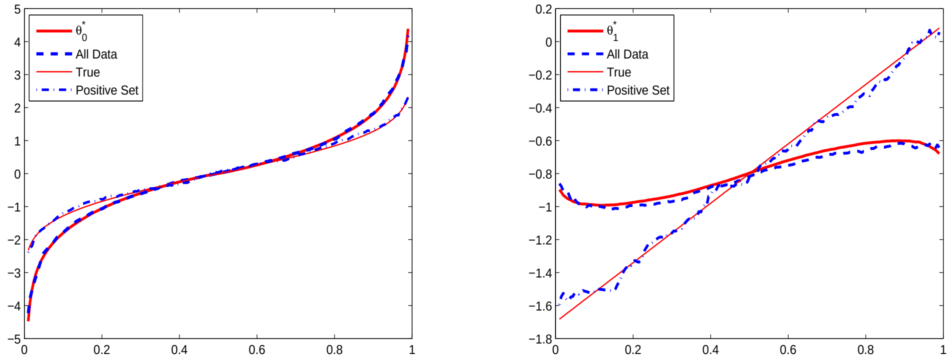

Figure 1: Estimates of $\theta_0(\tau) = \Phi^{-1}(\tau)$ (left panel) and $\theta_1(\tau) = 1.8\tau - 1.7$ (right panel) in model (7). The thin curves are the true coefficient functions $\theta_0(\tau)$ and $\theta_1(\tau)$, the dashed curves are the estimates obtained by using conditional quantile regression (4), the dotted curves are the estimates obtained by using restricted conditional quantile regression (6) and the thick solid curves are $\theta_0^*(\tau)$ and $\theta_1^*(\tau)$.

under some mild conditions. Without some technical conditions, $\theta^*(\tau)$ and $\theta(\tau)$ are not necessarily the same. This is evidenced by the example given in the last section in which $\theta^*(\tau) = \tau/2$ with no ambiguity, while $\theta(\tau) = |\tau - 0.5|$ or $\tau - 0.5I(\tau > 0.5)$.

The above argument suggests that the results in the main article should replace $\theta(\tau)$ by $\theta^*(\tau)$ unless the conditions under which they are identical are clearly imposed. If the primary interest is really on $\theta(\tau)$, then the conditional quantile regression should be replaced by the restricted conditional quantile regression (RCQR)

$$\min_{\theta} \sum_{t} I(t \in \mathcal{T}) \rho_\tau(Y_t - X_t^T \theta). \tag{6}$$

This avoids some samples, where the monotonicity condition is violated, that create inconsistent estimators. However, the set $\mathcal{T}$ is unknown and depends on the value at other quantile τ. This makes some difficulties in the implementation.

One possible way out is to replace $\mathcal{T}$ by one of its subsets. For example, if all $\theta_j(\cdot)$ is monotonically increasing, then we can replace $\mathcal{T}$ by the subset that all components of X_t are non-negative. Another possibility is to use (4) to get an initial estimate and then to check if the functions $\{\hat{\beta}_t(\tau), t = 1, \cdots, T\}$ are strictly increasing at some percentiles (e.g. $\tau = 0.05, 0.1, 0.15, \cdots, 0.95$). Delete the cases where the monotonicity is violated and use RCQR (6). The process can be iterated.

To illustrate the problem using the conditional quantile regression (4) and to address the issue of identifiability, we generate 2000 data points from the QAR(1) model

$$Y_t = \Phi^{-1}(U_t) + (1.8U_t - 1.7)Y_{t-1}. \tag{7}$$

Hence we have $\theta_0(\tau) = \Phi^{-1}(\tau)$ and $\theta_1(\tau) = 1.8\tau - 1.7$. Fit the data using (4) and (6). The resulting estimates are depicted in Figure 1. The estimates (dot-dashed curve) obtained by using RCQR (6) are very close to the true coefficient functions (thin solid curve), while the conditional quantile method (4) results in the estimates (dashed curve) that are far away from the true functions. Indeed, the latter estimates are for the functions $\theta_0^*(\tau)$ and $\theta_1^*(\tau)$ defined by (5), which were computed numerically and depicted in Figure 1 (thick solid curve). This example shows that even if monotonicity conditions are not fulfilled at all t, the coefficient functions can still be identifiable and consistently estimated, but the conditional quantile regression estimate, defined by maximizing (4), can be inconsistent.

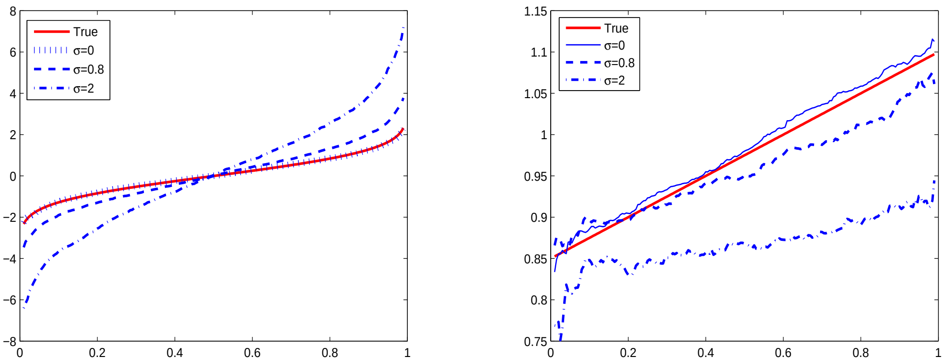

Figure 2: The influences of the error ε_t on the estimation of $\theta_0(\tau) = \Phi^{-1}(\tau)$ (left panel) and the estimation of $\theta_1(\tau) = 0.85 + 0.25\tau$ (right panel) in model (8) for different noise level σ.

A related question is how robust the fitting techniques are to the model misspecification. For example, if the data generating process is FCAR (1) without observing U_t, but we still use the conditional quantile regression (4) or its modification (6) to fit the data, how robust the quantile estimate is? To quantify this, we simulate the 2000 data from the model

$$Y_t = \Phi^{-1}(U_t) + (0.85 + 0.25U_t)Y_{t-1} + \varepsilon_t, \tag{8}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Figure 2 shows the plots for small noise $\sigma = 0$, moderate noise $\sigma = 0.8$ and relatively large noise $\sigma = 2$. The fitting techniques are very sensitive to the noise level. The estimates differ substantially from the true coefficient functions even for moderate $\sigma = 0.8$.

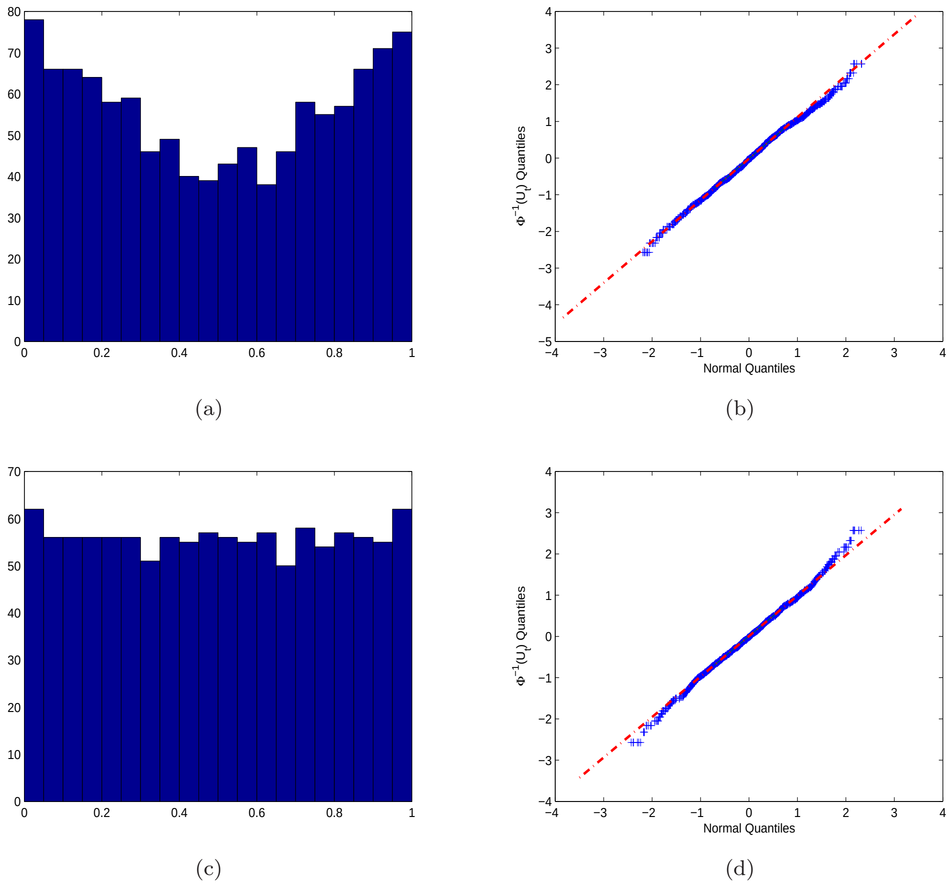

Figure 3: (a) Histogram plot of $\hat{U}_t = \hat{\beta}_t^{-1}(Y_t)$ for $t \in \mathcal{T}$, where $\hat{\beta}_t(\tau)$ is estimated by using the conditional quantile regression (4). (b) Quantile-quantile plot of $\Phi^{-1}(\hat{U}_t)$ versus the standard normal distribution, where $\hat{U}_t$ is the same as in (a). (c) and (d): The same as in (a) and (b) except that the estimation method is the restricted conditional quantile regression (6).

Checking monotonicity of $\hat{\beta}_t(\tau)$ is one aspect of model diagnostics. Another aspect is to check if the distribution of $\hat{U}_t = \hat{\beta}_t^{-1}(Y_t)$ for $t \in \mathcal{T}$ is uniform. There are many approaches to this kind of testing problem such as the Kolmogorov-Smirnov test or the visual inspection of the estimated density. For example, one can create the normally transformed data $\hat{Z}_t = \Phi^{-1}(\hat{U}_t)$ and then use the normal-reference rule of the kernel density estimate to see if the transformed residuals $\{\hat{Z}_t\}$ are normally distributed. Alternatively, one can use the quantile-quantile plot to accomplish a similar task.

For the data generated from (7) used in Figure 1, the histograms of $\hat{U}_t$ and the quantile-quantile plots of $\hat{Z}_t$ are presented in Figure 3. From these plots we can see that the distribution of $\hat{U}_t = \hat{\beta}_t^{-1}(Y_t)$ for $t \in \mathcal{T}$ by using the conditional quantile regression method (4) is not uniform, while $\hat{U}_t$ obtained by using the RCQR method (6) is uniformly distributed. These results are also supportive to our previous points.

Note that when the data are generated from model (1) without observing U_t, the model can still be identifiable when the density f_ε of innovation ε_t is known. The problem is more complicated but similar to a deconvolution problem. The estimation procedure can be quite complicated. To see this, note that

$$\begin{aligned} P(Y_t \leq a | Y_{t-1}, \cdots, Y_{t-p}) &= P(\beta_t(U_t) + \varepsilon_t \leq a) \\ &= \int \beta_t^{-1}(a-x) f_\varepsilon(x) dx. \end{aligned}$$

Thus, letting $Q_t(\tau)$ be the conditional τ-quantile, we have

$$\int \beta_t^{-1}(Q_t(\tau) - x) f_\varepsilon(x) dx = \tau.$$

Let us denote the solution by $Q_t(\tau) = h(\beta_t(\cdot), \tau)$ for some function h. Then the coefficient functions can be estimated by minimizing a quantity similar to (6):

$$\min_\theta \int_0^1 \sum_t I(t \in \mathcal{T}) \rho_\tau(Y_t - h(\beta_t(\cdot), \tau)) d\tau,$$

where $\theta(\cdot)$ and $\beta_t(\cdot)$ are related through (2). This is indeed a complicated optimization problem. The method of Koener and Xiao is a specific case of this method with $\varepsilon_t = 0$.

4 Concluding Remarks

Koenker and Xiao have developed a nice scheme for conditional quantile inference and made insightful connections with the QAR model. However, the issues of identifiability and possible misspecification of models suggest that extra care should be made in making this kind of links. In particular, the conditional quantile method does not always produce a consistent estimate for the random coefficient functions $\theta(\cdot)$ when the monotonicity conditions are not satisfied. Further studies are needed.

Additional References

Brumback, B. and Rice, J. A. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves" (with discussion), *J. Amer. Statist. Assoc.*, 93, 961-994.

Cai, Z., Fan, J. and Li, R. (2000a), " Efficient estimation and inferences for varying-coefficient models", *Jour. Ameri. Statist. Assoc.*, 95, 888-902.

Cai, Z., Fan, J. and Yao, Q. (2000b), "Functional-coefficient regression models for nonlinear time series", *Journal of American Statistical Association*, 95, 941-956.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998), "Nonparametric estimation via local estimating equations", *Jour. Ameri. Statist. Assoc*, 93, 214-227.

Chen, R. and Tsay, R.J. (1993), "Functional-coefficient autoregressive models", *J. Amer. Statist. Assoc.*, **88**, 298–308.

Chiang, C.-T., Rice, J.A. and Wu, C.O. (2001), "Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables", *Jour. Amer. Statist. Assoc.*, 96, 605-619.

Fan, J., Jiang, J., Zhang, C. and Zhou, Z. (2003), "Time-dependent diffusion models for term structure dynamics and the stock price volatility", *Statistica Sinica*, 13, 965-992.

Fan, J. and Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis", *Jour. Ameri. Statist. Assoc.*, **99**, 710–723.

Fan, J., Lin, H. and Zhou, Y. (2006), "Local partial-likelihood estimation for life time data", *The Annals of Statistics*, to appear.

Fan, J. and Zhang, J.T. (2000), "Functional linear models for longitudinal data", *Jour. Roy. Statist. Soc. B*, **62**, 303-322

Hastie, T.J. and Tibshirani, R. J. (1993), "Varying-coefficient models" (with discussion), *Jour. Roy. Statist. Soc. B.*, 55, 757-796.

Hong, Y. and Lee, T.-H. (2003), "Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models", *Review of Economics and Statistics*, 85, 1048-1062.

Lin, D.Y. and Ying, Z. (2001), "Semiparametric and nonparametric regression analysis of longitudinal data (with discussions)", *Jour. Amer. Statist. Assoc.*, 96, 103-126.

Mercurio, D. and Spokoiny, V. (2004), "Statistical inference for time-inhomogeneous volatility models", *The Annals of Statistics*, 32, 577–602.

Tian, L., Zucker, D., Wei, L.J. (2005), "On the Cox model with time-varying regression coefficients", *Jour. Amer. Statist. Assoc.*, **100** (2005), 69, 172–183.