DISCUSSION: POSTERIOR INFERENCE IN BAYESIAN QUANTILE REGRESSION WITH ASYMMETRIC LAPLACE LIKELIHOOD

ROGER KOENKER

1. Introduction

To bake a Bayesian π (posterior) I was taught that you needed an $\mathcal{L}$ (likelihood) and a p (prior) – Oh yeah, and probably some data, don't forget the data! So it comes as something of a shock to discover that there are 5,240 web documents employing the phrase "Bayesian quantile regression," as of September 1, 2015, according to Google. Quantile regression would seem to be the very antithesis of a likelihood based procedure, committing the investigator to a parametric model for one paltry conditional quantile function, while professing total ignorance, even indifference, about the rest of the *Deus ex machina*, aka data generating mechanism.

2. A Most Perplexing Paradox

So what is the attraction? What brings Bayesians to quantile regression like bears to honey? Is it that sweet smell of sin, always so powerful for the priesthood? Or is it that jihadist spirit of the Crusades, intent to recapture Jerusalem from the infidels? Maybe. But more likely it is simply that "Anything you can do, I can do better" confidence immortalized by Ethel Merman in the musical *Annie Get Your Gun*. If you listen to the song carefully, e.g. Berlin (1966), you will hear that the only thing that both parties to this competitive duet agree upon is that neither one can bake a π.

The usual knock on Bayesian methods focuses on the difficulty of coming up with sensible priors. I've never quite understood this complaint; of course it isn't easy especially in high dimensional problems to elicit a prior, anyone who thinks it is should consult the recent exchange between Larry Wasserman and Chris Sims. But everyone is entitled to the courage of their own convictions, I suppose, provided that they are not too dogmatic. It is just this last proviso that really worries me about the other crucial ingredient of the Bayesian paradigm: how is it that one can be so ignorant about model parameters but so confident about the specification of the likelihood? Likelihoods have proven to be especially problematic for quantile regression. There are, of course, several proposals, but by far the most commonly applied is the asymmetric Laplace distribution (ALD) employed in Yu and Moyeed (2001), which simply exponentiates the usual quantile regression objective function, introduces a scale parameter and computes a normalizing constant. *Voilà* we have magically

Version: September 24, 2015. This is an invited comment on the paper: Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood, by Yunwen Yang, Huixia Judy Wang and Xuming He to appear in the *International Statistical Review*.

transformed a local model for a single quantile into a global model for the entire data cloud. But does this make any sense? Isn't it paradoxical that we would extrapolate a local model that was totally agnostic about the probabilistic behavior of Y except for its conditional quantiles, $Q_Y(\tau|x)$ at one particular τ, to make a global model that assumed iid error, thus parallel conditional quantile functions for all $\tau \in (0,1)$? And what if we now consider the likelihood for another τ? Don't we have a completely different global model? Don't they conflict with one another?

3. Second Thoughts

I would like to express my profound gratitude to the authors of this paper for their cogent unraveling of this paradox. The authors have done a great service to the research community by clarifying this rather murky situation. Not only do they explain why naive implementations of Bayesian MCMC methods using the ALD approach lead to poor inference, they also reveal how to modify standard MCMC posterior inference to achieve good performance. Their results, building on earlier work of Chernozhukov and Hong (2003), provide a general approach to Bayesian inference for situations in which likelihoods are potentially misspecified. The quantile regression setting is especially well-suited to this analysis since the limiting (sandwich) covariance matrix, $V(\tau) = \tau(1-\tau)D_1^{-1}D_0D_1^{-1}$ has a rather complicated form for $D_1 = \lim_{n\to\tau} \sum_{i=1}^n f_i(F_i^{-1}(\tau))x_i^\top x_i$, an estimate of which is delivered automatically by the MCMC iterations. This avoids the necessity of estimating the local conditional density $f_i(F_i^{-1}(\tau))$. Estimating the meat of the sandwich, $D_0 = \lim_{n\to\infty} n^{-1}X^\top X$ is trivial, so building the modified covariance matrix for the estimator is easy. In other misspecified Bayesian settings the situation is unlikely to be quite so straightforward, nevertheless it is valuable to see the general framework.

It is worth reemphasizing that the usual claims in the earlier ALD literature that "the posterior is consistent" should not be taken as a justification for using the unmodified MCMC posterior for inference. It is reassuring of course to know that the posterior eventually converges to point mass at the true parameter, but without further information on rates and the behavior of the normalized estimator it is impossible offer any reliable advice on how to construct credible sets, or confidence regions. This is precisely what the emerging literature on Bernstein von Mises results is intended to resolve. Again, quantile regression offers a nice environment since we have a semiparametric estimator that is known to have relatively simple asymptotic behavior and convergence at the standard $1/\sqrt{n}$ rate.

There are many new and challenging questions raised by this important paper. An obvious motivation for the Bayesian formulation of quantile regression is the desire to impose prior information on some form of the local conditional quantile model. Even in the simplest case of a single quantile of interest with ℓ_1 or ℓ_2, i.e. Laplacian or Gaussian prior, it would be nice to know more about how to adapt inference under various regimes for the selection of the penalty parameter λ. For fixed "strength" of the prior, the usual Bernstein von Mises theory implies that the prior has no impact asymptotically, but this is a highly unsatisfactory conclusion. Often we would like to consider prior (penalties) that allow the effective dimension of the model to grow with sample size, and this is more challenging from an inferential perspective. Another important motivation for Bayesian formulations of quantile regression involves so-called composite quantile regression models in which one

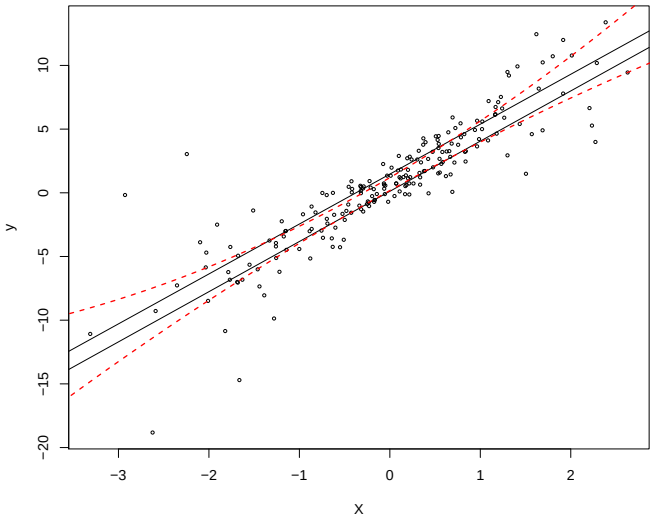

FIGURE 1. One Realization of Simulation 2: Black lines depict the linear fit for $\tau = 0.5 \pm 0.166$, and red lines show the quadratic fit at these quantiles.

wishes to estimate several conditional quantile models simultaneously; in these settings it will be often advantageous to consider prior smoothness restrictions across coefficients. Effectively we can "borrow strength" in the terminology of Tukey across quantiles and produce more reliable estimates, but formal inference in such settings is still a challenge.

4. A Homework Exercise

One disturbing aspect of the Yang, Wang and He (YWH) results, at least for me, was the poor performance of the Wald based "nid" confidence intervals in their second simulation setting. See their Table 1. I wanted to understand what was going wrong. The basic idea of the "nid" method is to estimate the limiting covariance matrix using a difference quotient for the local conditional density terms,

$$\hat{f}_i(F_i^{-1}(\tau)) = 2h/(x_i^\top \beta(\tau + h_n) - x_i^\top \beta(\tau - h_n))$$

where h_n is a bandwidth computed by default as in Hall and Sheather (1988). The second simulation model of Section 4.1 takes the form:

$$y_i = \tfrac{2}{3} + 4x_{1i} + 4x_{2i} + (1 + 0.6x_{1i}^2)u_i,$$

where the x_{ij} and u_i are independent standard normal random variables. Thus, at the median where the simulation results are focused, the model is linear in the covariates. However, at any quantile other than the median, the model is quadratic in the first covariate. Therein lies the difficulty, since the $\pm h$ estimation of a quadratic model is in such a case replaced by a more restrictive linear estimate.

TABLE 1. Coverage frequency in 10,000 trials Model 2: Linear Median Model

	n = 200				**n = 500**			
	rank	nid	ker	boot	rank	nid	ker	boot
b_0	0.886	0.912	0.975	0.909	0.892	0.905	0.951	0.904
b_1	0.889	0.716	0.948	0.897	0.892	0.672	0.929	0.899
b_2	0.892	0.894	0.973	0.911	0.895	0.904	0.956	0.907

The situation is illustrated for one realization of the simulation model in Figure 4. We have sample size $n = 200$, and we have dropped the second covariate to facilitate the visualization. The black lines illustrate the fitted linear model for this data at $\tau = 0.5 \pm 0.166$, and the (red) dashed lines illustrate the fitted quadratic model at the same quantiles. What is the consequence of using the linear fit rather than the quadratic? The linear fitting understates the quantile differences in the difference quotient above, especially for the extreme x_i's, and this tends to overstate the precision (Hessian) matrix $\hat{D}_1$, relative to what it would have been under the quadratic model where the outlying x_i's would have received less weight. Now when we make the sandwich, $\hat{V}(\tau) = \tau(1-\tau)\hat{D}_1^{-1}\hat{D}_0\hat{D}_1^{-1}$, the estimated precision of the $\hat{\beta}$'s is exaggerated, confidence intervals are too small and coverage is substantially less than the nominal level.

To explore this further I have replicated the experiment of YWH for this particular simulation setting using four standard methods for constructing confidence intervals available from the R package `quantreg`, Koenker (2015). In addition to the rank inversion and "nid" methods studied by YWH, I've added the Powell (1991) kernel method and a conventional implementation of the xy-bootstrap. We see in Tables 1 and 2 that coverage at the nominal 0.9 level, is quite good for all the methods with the exception of the "nid" method for the first slope coefficient confirming the YWH finding. Note that the Powell "ker" is quite conservative in this setting, but the xy-bootstrap performs quite well. Can we rescue the performance of the "nid" method by fitting quadratic models?

In Table 2 I report results of the same experiment except that now I've fitted a quadratic model for both of the covariates so the b3 and b4 rows of the table correspond to the quadratic coefficients of the median model, which are both zero at the median. As conjectured coverage performance for the "nid" intervals are now quite good, not only for the problematic slope coefficient, b_1, but also for the two quadratic coefficients. The lesson I would draw from this exercise is that linearity assumptions can be dangerous even when they are correct. One sometimes needs to be more flexible, at least for inferential purposes.

5. Conclusion

Finally, in light of the proceeding simulation exercise, I would like to raise a general question about the trade-off between statistical and computational efficiency. This is a topic that has gradually come to the forefront as statistics and machine learning have needed to confront larger datasets. But it is also a question that can be relevant in more moderate data settings. This was brought home to me recently by an email inquiry I received from someone interested in using the confidence intervals produced by rank test inversion method mentioned above on a regression problem with about half a million observations

TABLE 2. Coverage frequency in 10,000 trials Model 2: Quadratic Median Model

	n = 200				**n = 500**			
	rank	nid	ker	boot	rank	nid	ker	boot
b_0	0.877	0.887	0.974	0.908	0.877	0.893	0.953	0.905
b_1	0.886	0.888	0.945	0.912	0.894	0.899	0.928	0.909
b_2	0.895	0.884	0.972	0.925	0.897	0.890	0.954	0.914
b_3	0.878	0.874	0.901	0.894	0.887	0.890	0.896	0.898
b_4	0.888	0.856	0.964	0.919	0.892	0.865	0.952	0.913

and 50 parameters. My correspondent wondered if the `quantreg` computation of the rank inversion confidence intervals had "gone into an infinite loop." He had decided to terminate the process after a couple of hours without a solution. I assured him to the contrary, that on problems of this size the parametric linear programming problem that needed to be solve (twice!) for each coefficient must step through a very large number of matrix pivots and consequently was going to take an egregious effort. He wrote back sceptically saying that it was hard to "prove" the existence of an infinite loop, but he was going to let the job run and see what happened. The next day I got a somewhat sheepish note saying, "Well, you were right. My large job took over 27 hours of cpu time, but it did finally complete." So although the rank inversion intervals are quite reliable statistically, they do not "scale up" in the usual machine learning jargon, and one needs to find alternatives.

The Wald "nid" method of estimating the covariance matrix is usually also pretty reliable, but we saw that it can perform poorly in terms of coverage when the model specification ignores nonlinearities, rendering it "quick, but wrong." The standard "xy-bootstrap" seems to be a reasonable compromise, usually performing reliably from a coverage and mean length viewpoint, and it is reasonably computationally efficient. Of course when sample sizes become large one would have to reconsider, and at that point one might want to resort to the "m-out-n" bootstrap. All of this is leading up to the question: Can the authors provide any specific, or general, advice about this statistical vs computational trade-off for their modified MCMC procedure and its comparison with any of the available resampling methods?

REFERENCES

BERLIN, I. (1966): "Anything You Can Do, I Can Do Better," `https://www.youtube.com/watch?v=I-bgYxMO5cY`, As interpreted by Ethel Merman and Bruce Yarnell.

CHERNOZHUKOV, V., AND H. HONG (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.

HALL, P., AND S. SHEATHER (1988): "On the distribution of a studentized quantile," *J. of Royal Stat. Society (B)*, 50, 381–391.

KOENKER, R. (2015): *quantreg: Quantile Regression* R package version 5.19, available from: `https://CRAN.R-project.org/package=quantreg`.

POWELL, J. L. (1991): "Estimation of monotonic regression models under quantile restrictions," in *Nonparametric and Semiparametric Methods in Econometrics*, ed. by W. Barnett, J. Powell, and G. Tauchen. Cambridge U. Press: Cambridge.

SIMS, C. (2012): "Robins-Wasserman, Round-N," `http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR4a.pdf`.

Yu, K., and R. A. Moyeed (2001): "Bayesian Quantile Regression," *Statistics and Probability Letters*, 54, 437–447.