Approximate Bayesian Computation (ABC) in Quantile Regression

Antonio Pulcini[1]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 36

[1] Dipartimento SEFEMEQ. Università di Roma TOR VERGATA, Rome, Italy
antoniopulcini@gmail.com

Abstract

We propose an approximate Bayesian approach to estimate the joint distribution of the response variable Y and the set of covariates $\mathbf{X}$ based on the notion of quantile distribution. We focus on cases where the quantile regression framework is necessary, but the unknown form of the regression function and the large number of quantiles suggests to directly estimate the conditional distribution. In these cases the use of very flexibly-shaped distributions may be of interest. In this context we adopt the multivariate **g**-and-**h** distribution, a member of the quantile distribution family. Due to the lack of the likelihood function in a closed and manageable form, the estimation proceed via an Approximate Bayesian Computation (ABC) algorithm that allows us to easily estimate all the parameters. The performance of the proposed approach is evaluated via simulated data sets.

Keywords: Approximate Bayesian Computation; Joint Distribution; Quantile Distribution; Quantile Regression.

1 Summary

The usual assumptions of the standard linear model imply that the conditional distribution of the response variable is, at least approximately, Gaussian. In practice, this assumption is rarely acceptable. In many observational studies, the conditional distribution is not symmetric and, even worse, its shape depends on the value of the covariates (Yu et al. (2003)). For these situations, methods based on quantile regression are common alternatives (Koenker (2005)). Nevertheless when influence concerns more than one quantile, it may be convenient to consider the problem of directly estimating the conditional distribution of

the response variable given the explanatory variables (Peracchi (2002)). In this context quantile distributions, due to their flexibility and the small number of parameters, may represent a valid choice. Field & Genton (2006) have proposed a generalization of the univariate g-and-h distribution to the multivariate case. We exploit the use of the multivariate **g**-and-**h** distribution for the estimation of the joint distribution of the response variable Y given a vector of covariates $\mathbf{X} = (X_1, \ldots, X_K)$.
A drawback of quantile distributions, which has represented an obstacle to their use, is the lack of a closed form expression of the likelihood function. On the other hand, the problem of generating random values from them is an easy task. These issues, suggest the estimation via the Approximate Bayesian Computation (ABC) approach (Allingham et al. (2009), Tavaré et al. (1997)).
ABC allows to produce a sample from an approximate version of the posterior distribution. No likelihood evaluation is required, only a way to sample from the model distribution.
Briefly, we assume data, $\mathcal{D}$, arise from the multivariate **g**-and-**h** distribution:

$$\mathbf{W} = \boldsymbol{\Sigma}^{1/2}\mathbf{R}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) + \boldsymbol{\mu}$$

where:
- $\boldsymbol{\mu} \in \mathbb{R}^{K+1}$ is the location
- $\boldsymbol{\Sigma}$ is the variance covariance matrix
- $\mathbf{g} = (g_1, g_2, \ldots, g_{K+1}) \in \mathbb{R}^{K+1}$ controls the skewness
- $\mathbf{h} = (h_1, h_2, \ldots, h_{K+1}) \in {\mathbb{R}_+}^{K+1}$ controls the kurtosis
- $\mathbf{Z} \sim N_{K+1}(0, \mathbf{I})$
- $\mathbf{R}_{\mathbf{g},\mathbf{h}}(\mathbf{Z}) = (R_{g_1,h_1}(Z_1), R_{g_2,h_2}(Z_2), \ldots, R_{g_{K+1},h_{K+1}}(Z_{K+1}))^T$,
with $R_{g,h}(z) = \left(\frac{\exp(gz)-1}{g}\right)\exp\left(\frac{hz^2}{2}\right)$.

In order to estimate all the parameters in the model we propose the following ABC-MCMC algorithm:

1. Being at $\boldsymbol{\theta}_t$, propose a move to $\boldsymbol{\theta}'$ according to a Normal transition kernel $q(\boldsymbol{\theta}_t \to \boldsymbol{\theta}')$
2. Generate M samples, $\mathcal{D}'_1, \ldots, \mathcal{D}'_M$, from the model with parameters $\boldsymbol{\theta}'$
3. Calculate $\alpha = \min\left(1, \frac{\frac{1}{M}\sum_{m=1}^{M} K_\epsilon(\rho(S(\mathcal{D}),S(\mathcal{D}'_m)))\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}'\to\boldsymbol{\theta}_t)}{\frac{1}{M}\sum_{m=1}^{M} K_\epsilon(\rho(S(\mathcal{D}),S(\mathcal{D}'_{m;t})))\pi(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_t\to\boldsymbol{\theta}')}\right)$
4. Accept $\boldsymbol{\theta}'$ with probability α, otherwise stay at $\boldsymbol{\theta}_t$
5. Update the variance of $q(\cdot)$ through an Adaptive MCMC scheme; then return to 1.

Specifically, $S(\cdot)$ is a multivariate quantile (Chaudhuri (1996)), $\rho(\cdot)$ the Euclidean norm, and $K_\epsilon(\cdot)$ a Multivariate Gaussian kernel centered on $S(\mathcal{D}') =$

$S(\mathcal{D})$ with variances ϵ's.
The performance of the proposed method is evaluate with simulated datasets. Secondly we apply our approach to estimate the joint distribution of price and demand in the Italian Day-Ahead electricity market.

References

Allingham, D., King, R. A. R. & Mengersen, K. L. (2009). Bayesian estimation of quantile distributions. *Stat. Comput.* 19 189–201.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* 91 862–872.

Field, C. & Genton, M. G. (2006). The multivariate **g**-and-**h** distribution. *Technometrics* 48 104–111.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Peracchi, F. (2002). On estimating conditional quantiles and distribution functions. *Comput. Statist. Data Anal.* 38 433–447. Nonlinear methods and data mining (Rome, 2000).

Tavaré, S., Balding, D., Griffiths, R. c. & Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics* 145 505–518.

Yu, K., Lu, Z. & Stander, J. (2003). Quantile regression: applications and current research areas. *The Statistician* 52 331–350.