DOI:10.13546/j.cnki.tjyjc.2016.19.006

变量惩罚效应在贝叶斯分位数回归模型的应用

郭俊峰
（厦门大学 经济学院，福建 厦门 361005）

摘　要：尽管贝叶斯分位数回归方法能够有效克服经济金融数据的尖峰厚尾、结构突变等问题，充分借鉴已有研究成果信息，但是其并不能很好解决多维变量模型的维数灾难问题。为此，文章在贝叶斯分位数回归基础上，结合自适应Lasso变量惩罚作用，构建了基于MH抽样的自适应Lasso惩罚贝叶斯分位数回归模型。通过仿真模拟实验以及MCMC链条检验，证明上述模型具有优良拟合性质，尤其是在小样本情形下。

关键词：维数灾难；自适应Lasso惩罚；贝叶斯；分位数回归
中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2016)19-0020-03

0　引言

伴随着计算机技术和计量模型的发展，学者们开始将分位数回归（Quantile Regression，QR）方法运用于经济金融、卫生统计等领域的研究，它能够有效克服数据的尖峰厚尾以及结构突变等问题，还对极端异常值有很强的鲁棒性，因此该方法日益受到研究人员重视。分位数回归方法本身也不断扩展延伸，其中一个重要方向是与贝叶斯估计结合，通过不对称Laplace分布来构建贝叶斯分位数回归（Bayesian Quantile Regression，BQR）模型[1]，从而有效利用以往研究成果信息、提高样本数据较少时的参数估计精度。

可是在多维变量模型中，BQR方法平等估计每个解释变量而不考虑变量作用显著与否，换句话说，BQR模型不能解决维数灾难问题，即使Tibshirani在1996年[2]提出了Lasso变量惩罚方法，也不能很好处理多维变量模型的维数灾难问题，因为该方法对所有自变量都施以相同惩罚，而这显然与不同自变量对因变量影响各异的规律相悖。

基于此，本文在贝叶斯分位数回归模型基础上，尝试着结合自适应Lasso变量（Adaptive Lasso）惩罚作用[3]，对不同自变量给予不同惩罚系数。经过理论推导，最终构建了基于MH抽样的自适应Lasso惩罚贝叶斯分位数回归（Adaptive Lasso Bayesian Quantile Regression，ALBQR）模型。仿真模拟分析表明，相比于OLS模型、QR模型及BQR模型，ALBQR模型有更好的拟合效果。

1　模型构建与贝叶斯分析推导

1.1　贝叶斯分位数回归BQR模型

Koenker和Bassett（1978）[4]率先提出分位数回归方法。给定自变量 X 信息后，Y 的第 τ 分位数水平线性条件分位数模型表达式为

$$Q_\tau(Y|X)=X^T\beta \tag{1}$$

也就是

$$y_t=x_t^T\beta+\varepsilon_t,\ t=1,2,\cdots,n \tag{2}$$

其中含常数项解释变量 $x_t=(1,x_{1t},\ldots,x_{kt})^T$，$\beta=(\beta_0,\beta_1,\cdots,\beta_k)^T$，$\varepsilon_t$ 是模型误差项。通过极小化

$$\min_\beta\sum_{t=1}^n\rho_\tau(y_t-x_t^T\beta) \tag{3}$$

得到QR模型系数 β 的估计值，其中 $\rho_\tau(u)=u(\tau-I(u<0))$，$I(\bullet)$ 为示性函数。

实际研究中，我们往往还可以参照以前相关成果。然而，普通QR模型并没有借鉴这些经验，所以下面对该模型进行贝叶斯分析推导，构建贝叶斯分位数回归BQR模型。为了将贝叶斯方法纳入到分位数回归框架，本文需要运用不对称拉普拉斯先验分布（Asymmetric Laplace Distribution，ALD）。给定 $x\sim ALD(\mu,\sigma,p)$，μ 是位置参数，σ 是尺度参数，p 是偏度参数，那么其密度函数如下：

$$f(x|\mu,\sigma,p)=\sigma p(1-p)\exp\{-\sigma\rho_p(x-\mu)\} \tag{4}$$

令 $X=(x_1,\cdots,x_t,\cdots,x_n)$，若 $x_t\sim ALD(\mu,\sigma,p)$，有似然函数

$$L(X|\mu,\sigma,p)=\sigma^n p^n(1-p)^n\exp\{-\sigma\sum_{t=1}^n\rho_p(x_t-\mu)\} \tag{5}$$

Tsionas（2003）[5]证明，如果 $x\sim ALD(\mu,\sigma,p)$，那么 x 可以等价表示为：

$$x=\mu+\theta z+\phi\xi\sqrt{\sigma^{-1}z} \tag{6}$$

在式（6），$\theta=\frac{1-2p}{p(1-p)}$、$\phi^2=\frac{2}{p(1-p)}$、$z\sim E(\frac{1}{\sigma})$、$\xi\sim N(0,1)$。式（6）说明，服从ALD分布的变量可以表示为服从正态分布变量和指数分布变量的组合。设式（2）中

基金项目：国家自然科学基金面上项目（71373219）；国家自然科学基金青年项目（71103150）；中央高校基本科研业务费专项资金资助项目（2013221012）
作者简介：郭俊峰（1988—），男，江西赣州人，博士研究生，研究方向：金融计量经济学。

$\varepsilon_t \sim ALD(0,\sigma,p)$，根据ALD分布线性变换性质，有$y_t \sim ALD\left(x_t^T\beta,\sigma,p\right)$，由式(5)得到$Y=\left(y_1,\cdots,y_t,\cdots,y_n\right)$的似然函数：

$$L(Y|X^T\beta,\sigma,p)=\sigma^n p^n(1-p)^n \exp\{-\sigma\sum_{t=1}^{n}\rho_p(y_t-x_t^T\beta)\} \quad (7)$$

比较式(3)与式(7)，看出极小化式(3)等价于极大化式(7)，分位数水平τ等同于ALD分布的偏度系数P。根据式(6)，将因变量y_t表示成：

$$y_t=x_t^T\beta+\theta z_t+\phi\xi_t\sqrt{\sigma^{-1}z_t} \quad (8)$$

相应地，BQR模型的参数估计值为：

$$\hat{\beta}=\arg\min_{\beta}\sum_{t=1}^{n}\rho_p\left(y_t-x_t^T\beta-\theta z_t\right) \quad (9)$$

1.2 带有变量惩罚效应的贝叶斯分位数回归模型

尽管BQR模型可以很好地解决数据的尖峰厚尾、结构突变等问题，也充分利用了已有先验信息。但在参数估计时，该方法却不加选择地平等对待每个解释变量。由于多维变量模型普遍存在“维数灾难”难题，所以Tibshirani(1996)[2]提出了Lasso变量惩罚方法。可是Lasso惩罚方法没有Oracle估计性质，其对所有变量的回归系数都施以相同惩罚。这显然与现实规律相违背。为此，对于BQR模型，我们借助自适应Lasso惩罚方法，通过选择适当权重，对不同变量给予不同惩罚系数，从而得到自适应Lasso惩罚贝叶斯分位数回归(Adaptive Lasso Bayesian Quantile Regression，ALBQR)模型，其具有Oracle性质的参数估计值为：

$$\hat{\beta}=\arg\min_{\beta}\sum_{t=1}^{n}\rho_p\left(y_t-x_t^T\beta\right)+\sum_{j=1}^{k}\lambda_j|\beta_j| \quad (10)$$

其中λ_j是非负的可变惩罚系数。

1.3 ALBQR模型参数估计与算法设计

假设系数$\beta_j \sim ALD\left(0,\frac{\sigma}{\lambda_j},\frac{1}{2}\right)$，由式(5)，其密度函数

$$p\left(\beta_j|\sigma,\lambda_j\right)=\frac{\sigma}{4\lambda_j}\exp\left(-\frac{\sigma|\beta_j|}{\lambda_j}\right) \quad (11)$$

令$v_j=\frac{\sigma}{\lambda_j}$，根据文献[6]，参数$\beta$先验密度可以表示成：

$$p\left(\beta_j|\sigma,\lambda_j\right)=\frac{1}{2}\frac{v_j}{2}\exp\left(-v_j|\beta_j|\right)$$

$$=\frac{1}{2}\int_0^{\infty}\frac{1}{\sqrt{2\pi s_j}}\exp\left(-\frac{\beta_j^2}{2s_j}\right)\frac{v_j^2}{2}\exp\left(-\frac{v_j^2 s_j}{2}\right)ds_j \quad (12)$$

进而

$$p\left(\beta_j|\sigma^2,\lambda_j^2\right)=\frac{1}{2}\int_0^{\infty}\frac{1}{\sqrt{2\pi s_j}}\exp\left(-\frac{\beta_j^2}{2s_j}\right)\frac{\sigma^2}{2\lambda_j^2}\exp\left(-\frac{\sigma^2 s_j}{2\lambda_j^2}\right)ds_j \quad (13)$$

对于参数λ_j^2，令其先验分布服从逆伽玛分布，有密度函数：

$$p\left(\lambda_j^2|\delta,\psi\right)=\frac{\psi^{\delta}}{\Gamma(\delta)}\left(\lambda_j^2\right)^{-1-\delta}\exp\left(-\frac{\psi}{\lambda_j^2}\right) \quad (14)$$

式(14)中，δ、ψ为超参数。综上所述，本文通过假设参数β_j和误差项ε_t都服从ALD先验分布，并对参数β_j施以可变惩罚作用$\frac{\sigma}{\lambda_j}$。参数先验分布分别为：

$\varepsilon_t \sim ALD(0,\sigma,p)$，$\beta \sim ALD\left(0,\frac{\sigma}{\lambda_j},\frac{1}{2}\right)$，$z_t \sim E(\frac{1}{\sigma})$，$\xi_t \sim N(0,1)$

$\sigma \sim Ga(c_0,d_0)$，$\lambda_j^2 \sim IG(\delta,\psi)$，$\delta \propto 1$，$\psi \sim unif\left(\psi^{-1}\right)$

贝叶斯估计参数时，后验分布密度函数较难求解并且形式复杂，一般很难得到后验分布密度的明确表达式，所以只能借助模拟抽样技术。MCMC是一种简单有效的数值模拟计算方法，包括Gibbs抽样和MH抽样，Gibbs抽样本质是接受概率恒为1的MH抽样特例，本文用MH抽样算法来进行贝叶斯参数估计。MH抽样从建议分布$q\left(\theta,\theta'\right)$中抽样得到候选样本$\theta'$，然后以概率$a\left(\theta,\theta'\right)$决定是否接受由$\theta \to \theta'$，形成转移核$p\left(\theta,\theta'\right)$，具体如下：

$$p\left(\theta,\theta'\right)=q\left(\theta,\theta'\right)a\left(\theta,\theta'\right) \quad (15)$$

设第k步马尔可夫链的状态向量为$\theta^{(k)}$，根据建议分布$q\left(\theta^{(k)},\theta'\right)$产生另一状态向量$\theta'$，然后随机从均匀分布$U(0,1)$中抽取$a$，如果$a<a\left(\theta^{(k)},\theta'\right)$，就接受$\theta^{(k+1)}=\theta'$，否则$\theta^{(k+1)}=\theta$。

2 仿真模拟分析

2.1 数据来源

我们接下来进行仿真模拟，以检验ALBQR模型的合理性和优越性，尤其在小样本情形下。简单起见，设定123456为随机数种子，生成6个在不同区段的均匀分布变量$X=\left(X_1,X_2,X_3,X_4,X_5,X_6\right)$，变量个数用N表示，本文取N为20、50及100。然后根据下列方程式生成因变量Y：

$$Y_t=1+2\cdot X_{1,t}+3\cdot X_{2,t}+4\cdot X_{3,t}+5\cdot X_{4,t}+6\cdot X_{5,t}+7\cdot X_{6,t}+\varepsilon_t \quad (16)$$

上式中，误差项ε_t被设为服从零均值、异方差的正态分布。很明显，对于7×N个模拟数据而言，式(16)就是多维变量模型回归方程，并且样本数量N也有大有小，因此这些数据符合仿真模拟的要求。

2.2 仿真结果分析

假定ALBQR模型的先验参数$\sigma \sim Gamma(0.001,0.001)$，步长是1。进行MH抽样50000次，预烧30000次，剩下数据用于估计上述6个模拟变量的系数。表1—表3分别提供了样本量N为20、50及100时的参数后验均值。为便于比较，我们还列出OL和BQR模型的相应结果。

根据表1至表3，我们发现如下规律：第一，普通最小二乘法OLS的参数估计值的确介于不同分位数水平的BQR(或者ALBQR)估计值之间，这是由于OLS方法估计的是条件均值方程，注重平均角度，而分位数模型通过变动分位数水平，还可以研究两端尾部极端情况下的变量关系，所以OLS能够挖掘出的信息量最少。第二，就同一模

表1　仿真模拟结果(样本量N=20)

	真实值	OLS	BQR			ALBQR		
分位数			0.25	0.50	0.75	0.25	0.50	0.75
常数项	1.000	12.220 (11.220)	3.498 (2.498)	13.698 (12.698)	15.740 (14.740)	1.840 (0.840)	8.598 (7.598)	10.585 (9.585)
X_1	2.000	1.752 (−0.124)	0.943 (−0.529)	2.009*** (0.004)	1.947** (−0.027)	1.074 (−0.463)	1.801* (−0.100)	1.921** (−0.040)
X_2	3.000	2.951** (−0.016)	3.220* (0.073)	2.912** (−0.029)	3.133** (0.044)	3.177* (0.059)	2.915** (−0.028)	2.927** (−0.024)
X_3	4.000	3.669* (−0.083)	3.649* (−0.088)	4.403 (0.101)	4.271* (0.068)	3.623* (−0.094)	4.008*** (0.002)	4.433 (0.108)
X_4	5.000	4.903** (−0.019)	4.922** (−0.016)	5.142** (0.028)	4.915** (−0.017)	5.018*** (0.004)	5.029*** (0.006)	5.042*** (0.008)
X_5	6.000	5.023 (−0.163)	5.554* (−0.074)	4.436 (−0.261)	4.705 (−0.216)	5.716** (−0.047)	5.251 (−0.125)	5.144 (−0.143)
X_6	7.000	6.558* (−0.063)	7.222* (0.032)	6.444* (−0.079)	6.133 (−0.124)	7.166** (0.024)	6.695** (−0.044)	6.501* (−0.071)

注:限于篇幅,本表仅列示BQR模型和ALBQR模型的参数后验均值。括号内是变量估计值与真实值之间的误差百分比。***、**、*表示在1%、5%和10%水平下该比值显著。下同。

表2　仿真模拟结果(样本量N=50)

	真实值	OLS	BQR			ALBQR		
分位数			0.25	0.50	0.75	0.25	0.50	0.75
常数项	1.000	4.887 (3.887)	9.153 (8.153)	1.436 (0.436)	3.819 (2.819)	1.112 (0.112)	0.272 (−0.728)	1.272 (0.272)
X_1	2.000	2.375 (0.188)	2.209 (0.105)	2.389 (0.195)	2.901 (0.451)	2.055** (0.028)	2.433 (0.217)	2.799 (0.400)
X_2	3.000	3.065** (0.022)	2.856** (−0.048)	3.081** (0.027)	3.113** (0.038)	2.962** (−0.013)	3.029*** (0.010)	3.098** (0.033)
X_3	4.000	4.431 (0.108)	4.612 (0.153)	4.038** (0.010)	4.062** (0.016)	4.210* (0.053)	4.120** (0.030)	3.850** (−0.038)
X_4	5.000	4.872** (−0.026)	5.036*** (0.007)	5.171** (0.034)	4.903** (−0.019)	5.087** (0.017)	5.191** (0.038)	4.994*** (−0.001)
X_5	6.000	5.220 (−0.130)	4.671 (−0.222)	5.588* (−0.069)	5.734** (−0.044)	5.576* (−0.071)	5.687* (−0.052)	6.005*** (0.001)
X_6	7.000	6.974*** (−0.004)	6.703** (−0.042)	7.047*** (0.007)	6.754** (−0.035)	7.068*** (0.010)	7.081** (0.012)	6.896** (−0.015)

表3　仿真模拟结果(样本量N=100)

	真实值	OLS	BQR			ALBQR		
分位数			0.25	0.50	0.75	0.25	0.50	0.75
常数项	1.000	1.676 (0.767)	0.154 (−0.846)	0.017 (−0.983)	1.938 (0.938)	0.116 (−0.884)	0.207 (−0.793)	1.288 (0.288)
X_1	2.000	1.910** (−0.045)	1.562 (−0.219)	1.914** (−0.043)	2.529 (0.265)	1.733 (−0.134)	2.015*** (0.008)	2.482 (0.241)
X_2	3.000	2.970*** (−0.010)	2.997*** (−0.001)	2.949** (−0.017)	2.965** (−0.012)	2.984*** (−0.005)	2.966** (−0.011)	3.011*** (0.004)
X_3	4.000	3.915** (−0.021)	4.023*** (0.006)	3.950** (−0.013)	4.102** (0.026)	4.028*** (0.007)	3.986*** (−0.003)	3.994*** (−0.001)
X_4	5.000	5.061** (0.012)	5.014*** (0.003)	5.025*** (0.005)	4.934** (−0.013)	5.028*** (0.006)	5.011*** (0.002)	4.986*** (−0.003)
X_5	6.000	5.901** (−0.017)	5.977*** (−0.004)	6.043*** (0.007)	5.853** (−0.025)	5.966*** (−0.006)	6.007*** (0.001)	5.920** (−0.013)
X_6	7.000	7.046*** (0.007)	7.092** (0.013)	7.106** (0.015)	7.025*** (0.004)	7.088** (0.013)	7.104** (0.015)	7.050*** (0.007)

型来说,随着样本量N增大,所有估计值都越来越显著,这说明误差百分比逐渐降低,参数估计精度都得到提高。同时,OLS、BQR与ALBQR模型之间的估计精度差别也不断缩小。第三,在同一样本量下,OLS方法最不准确,相比而言,ALBQR的参数估计系数最接近各个模拟变量的真实值。尤其是在样本量很小(N=20)时,ALBQR模型的优势更加明显。

采用贝叶斯方法估计参数后,需要检验变量MCMC链条的收敛性,本文使用Geweke检验方法。限于篇幅,我们只列出样本量N为100时的MCMC链条(tau=0.25、0.5、0.75)收敛性判断结果。表4汇报了检验情况。

表4　MCMC链条收敛性判断(样本量N=100)

	BQR			ALBQR		
分位数	0.25	0.5	0.75	0.25	0.5	0.75
常数项	0.023	2.036	0.143	0.355	−1.738	−1.009
X_1	0.807	0.582	−0.782	−1.859	−1.121	0.151
X_2	0.299	0.871	−0.197	−0.052	−0.639	0.937
X_3	−1.892	−0.919	0.377	−0.963	0.275	0.515
X_4	−0.356	−0.873	−0.334	−1.564	−0.699	1.132
X_5	0.693	−1.068	−0.453	0.599	1.577	0.794
X_6	0.435	−0.960	0.337	0.541	0.245	−0.003

注:Geweke检验中取初始10%和后面50%两个子链。

在表4,样本量为100时,BQR模型和ALBQR模型所有链条的Z统计量绝对值都小于2,均通过Geweke收敛性检验,因此判断这些MCMC链条收敛稳定,从而侧面印证前文关于ALBQR模型的分析结论是合理有根据的。

3　结束语

虽然贝叶斯分位数回归模型可以解决数据普的尖峰厚尾、结构突变等问题,也充分利用先验信息,但该方法没有很好地处理多维变量模型的维数灾难问题,本文在贝叶斯分位数回归方法基础上,采用自适应Lasso惩罚进行变量选择,构建了基于MH抽样算法的自适应Lasso惩罚贝叶斯分位数回归模型。仿真模拟实验表明,在小样本时,ALBQR模型的拟合性能更优也更稳健。

参考文献:

[1]陈耀辉,郭俊峰,殷文超. 人民币升值对中小板市场波动的影响——基于贝叶斯分位数回归的分析[J]. 系统工程,2015,(1).

[2]Tibshirani R. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society(Series B), 1996, 58(1).

[3]Zou H. The Adaptive Lasso and Its Oracle Properties [J]. Journal of the American Statistical Association, 2006, 101(476).

[4]Koenker R, Bassett G. Regression Quantiles [J]. Econometrica: Journal of the Econometric Society, 1978, 46(1).

[5]Tsionas E G. Bayesian Quantile Inference [J]. Journal of Statistical Computation and Simulation, 2003, 79(3).

[6]Andrews D F, Mallows C L. Scale Mixtures of Normal Distributions [J]. Journal of the Royal Statistical Society(Series B), 1974, 36(1).

(责任编辑/易永生)