计算机科学
COMPUTER SCIENCE

基于spike-and-slab先验的贝叶斯时间序列模型

郭晨蕾, 李东喜

引用本文

郭晨蕾, 李东喜. 基于spike-and-slab先验的贝叶斯时间序列模型[J]. 计算机科学, 2023, 50(11A): 221200131-6.

GUO Chenlei, LI Dongxi. Bayesian Time-series Model Based on spike-and-slab Prior[J]. Computer Science, 2023, 50(11A): 221200131-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221200131

基于 spike-and-slab 先验的贝叶斯时间序列模型

郭晨蕾[1]　　李东喜[2]
1 太原理工大学数学学院　山西 晋中 030600
2 太原理工大学大数据学院　太原 030024
(1044650626@qq.com)

摘　要　贝叶斯方法通过引入先验信息并结合似然的方法进行参数估计和变量选择,使模型估计和预测结果更为精确。在贝叶斯框架下考虑时间序列之间的相关性,将偏自相关系数融合先验信息,提出基于 spike-and-slab 先验的贝叶斯层次时间序列模型(Spike-and-slab Prior with Partial Autocorrelation Coefficients,SS-PAC)。SS-PAC 模型采用 spike-and-slab 先验并结合偏自相关系数,实现时间序列滞后阶数的选择、参数估计和预测。基于模拟数据和真实数据的实证研究表明,该模型相较于以往模型在变量选择和预测结果上表现更优。

关键词:时间序列预测;spike-and-slab 先验;贝叶斯方法;偏自相关系数;变量选择
中图法分类号　O212.8

Bayesian Time-series Model Based on spike-and-slab Prior

GUO Chenlei[1] and LI Dongxi[2]
1 College of Mathematics,Taiyuan University of Technology,Jinzhong,Shanxi 030600,China
2 College of Data Science,Taiyuan University of Technology,Taiyuan 030024,China

Abstract　Bayesian method makes the results of estimation and prediction more accurate by introducing prior information and combining with likelihood for parameter estimation and variable selection. ABayesian hierarchical time-series model based on spike-and-slab prior with partial autocorrelation coefficients(SS-PAC) is proposed under the Bayesian framework,considering the correlation between time series,fusing with the partial autocorrelation coefficient and prior information,the SS-PAC model uses spike-and-slab prior and partial autocorrelation coefficient to realize the selection,parameter estimation and prediction of time series lag order. Empirical research through simulated data and real data shows that the model performs better than previous models in variable selection and prediction results.

Keywords　Time-series prediction,Spike-and-slab prior,Bayesian method,Partial autocorrelation coefficient,Variable selection

1　引言

对数据建立模型时,参数的估计以及模型的预测都离不开准确的变量选择。时间序列预测作为一个经典的研究方向,基于过去时刻的状态信息能够预测未来的信息。最早应用于时间序列预测的是线性模型,其代表模型是自回归模型(Autoregressive,AR)[1]。将 AR 和 MA 模型结合起来就形成了自回归滑动平均模型(Autoregressive Moving Average,ARMA),在此基础上引入差分法就形成了差分自回归移动平均模型(Autoregressive Integrated Moving Average,ARIMA)——一个十分经典的时间序列预测模型。随着数据量的不断增加,一些学者考虑利用更具稳定性的惩罚方法解决时间序列变量选择及预测问题。常见的变量选择方法有统计学教授 Tibshirani 首次运用的 Lasso 方法[2],通过加入ℓ_1范数构造惩罚函数,从而达到变量选择的效果;Fan 提出的 SCAD 方法具备无偏性、稀疏性及连续性特点[3],使得变量选择过程更加精细;Zou 和 Hastie 将ℓ_1和ℓ_2范数进行加权提出的弹性网(Elastic-Net)方法可使关键变量不轻易被筛掉,降低了信息损失的风险[4];Lasso 是 Zou 对 Lasso 估计法的进一步改进[5],在ℓ_1惩罚项前增加权重,从而达到自适应效果。之后有学者将这些方法应用于时间序列数据,Verbesselt 等将 Lasso 用于树木死亡率的预测中,结果表明基于三年卫星数据的 Lasso 回归的连续子集选择模型拥有最佳预测能力[6];Zhang 等在功能磁共振成像时间序列中运用了 SCAD,Adaptive Lasso 以及 Elastic-Net,仿真研究表明,在稀疏恢复方面,正则化方法优于传统的非正则化方法[7]。

Tibshirani 提出 Lasso 方法的同时,也给出了该方法在贝叶斯角度的解释,如果参数的先验分布为拉普拉斯分布,那么其后验分布所得众数估计与 Lasso 方法所估一致,这就将贝叶斯方法与变量选择联系起来。贝叶斯学派最基本的观点是:任何一个未知量都可看作一个随机变量,应该用一个概率分布去描述对未知量的未知状况。这个概率分布是在抽样前就有的关于未知量先验信息的概率陈述,这个概率被称为先验分布。之后使用观测数据确定似然函数,以确定后验分布,最后通过后验分布得到未知量的估计量。其中,先验分布的选取关乎参数估计结果的精确度。

Mitchell 和 Beauchamp[8]、George 和 McCulloch[9] 以及 Kuo 和 Mallick[10] 都提出了采用 spike-and-slab 先验进行变量

通信作者:李东喜(dxli0426@126.com)

选择的方法。该先验的独特之处在于,对于变量选择,它有天然适应的概率结构,通过控制每个系数的指示变量,为每个预测值做出选入或者剔除的解释。时间序列数据就是按照时间顺序所留下的有序数据,即时间序列是由多个滞后阶数引起的,因此时间序列之间的相关性将影响变量选择与参数估计的结果,但许多滞后阶数与结果几乎没有关系,而且以往使用的变量选择模型并没有考虑到时间次序的相关性,忽略了滞后阶数对模型的影响。贝叶斯方法则提供了一种自然的解决方案,通过强加先验来缩小参数,从而克服维数问题[11]。Peter将时域中的频率识别问题重新表达为变量选择模型,其中每个变量对应于不同的频率,在每个频率的权重上放置收缩先验分布,并包括先验信息[12]。Li等依据主观经验构造先验进行时间序列的变点检测任务[13]。Cathy为每年的登革热病例提供了有效的贝叶斯估计和模型选择[14]。

综上所述,本文将时间序列的相关表达与先验融合,充分利用时间序列数据信息提出一种稀疏模型,即基于spike-and-slab先验的贝叶斯层次时间序列模型,在贝叶斯层次结构框架内,将偏自相关系数与spike-and-slab先验融合,通过MCMC算法,对模拟和真实时间序列数据进行变量选择、参数估计以及预测。同时,将所提模型与传统变量选择方法进行对比,结果表明,所提模型优于其他方法。

2 模型建立及算法推断

2.1 准备知识

设有给定协变量$\{x_{ij}\}_{j=1}^{p}$,输出为y_i的多元线性回归模型的一般形式:

$$y_i=\sum_{j=1}^{p}x_{ij}\beta_j+\varepsilon_i \tag{1}$$

其中,$i=1,2,\cdots,n$和$\beta_j(j=1,2\cdots,p)$为待求的回归系数,$\varepsilon_i\sim N_n(0,\sigma^2)$表示具独立同分布的噪声。为了避免出现截距$\beta_0$,假设响应向量$y$以零为中心;此外,为了便于根据回归系数的大小评估单个预测因子对模型的影响,对预测因子也进行中心化和标准化处理,使得$\sum_{j=1}^{p}x_{ij}=0$,$\|x_j\|^2=n-1(1\leqslant j\leqslant p)$。

假设每个回归系数$\beta_j(j=1,2,\cdots,p)$由一个spike-and-slab先验控制,即:

$$\beta_j|\gamma_j\sim(1-\gamma_j)N(0,\xi_{j1})+\gamma_jN(0,\xi_{j2})$$

$$\gamma_j|\delta_j\sim Bernoulli(\delta_j)$$

其中,ξ_{j1}较小而ξ_{j2}较大,γ_j为二元指示变量,用来诱导β_j的两种分布:如果$\gamma_j=1$,表明第j个预测因子应当选入模型中,此时β_j由slab分布决定:$N(0,\xi_{j2})$生成,较大的ξ_{j2}使得在0附近产生分散效应,表现为$\beta_j\neq0$;如果$\gamma_j=0$,则表明第j个预测因子不应当选入模型中,此时β_j由spike分布决定:$N(0,\xi_{j1})$生成,较小的ξ_{j1}使得在0附近产生集中效应,表现为$\beta_j=0$。

在统计研究中,我们把一定时间段内同一统计指标的数值按时间先后顺序排列而成的一组数列$x_1,x_2,\cdots,x_t$称为时间序列。平稳时间序列指各阶统计特征(如均值、方差、协方差等)不随时间的变化而变化,这样就能够通过建立模型来探究时间序列,并以此对未来时刻进行预测。

时间序列$x_1,x_2,\cdots,x_t$,x_t与x_{t-k}的偏自相关系数,是指去掉$x_{t-1},x_{t-2},\cdots,x_{t-k+1}$间接影响后的简单相关系数。

考虑如下模型:

$$x_t=\varphi_{11}x_{t-1}+\alpha_{1t}$$

$$x_t=\varphi_{21}x_{t-1}+\varphi_{22}x_{t-2}+\alpha_{2t}$$

$$x_t=\varphi_{31}x_{t-1}+\varphi_{32}x_{t-2}+\varphi_{33}x_{t-3}+\alpha_{3t}$$

$$\cdots$$

$$x_t=\varphi_{k1}x_{t-1}+\varphi_{k2}x_{t-2}+\varphi_{k3}x_{t-3}+\cdots+\varphi_{kk}x_{t-k}+\alpha_{kt}$$

上述模型中φ_{kk}就是偏自相关系数,即只考虑第k阶滞后对当期的影响。

贝叶斯统计中,关键是使用后验分布对未知参数作估计,本文采取的是目前应用较为广泛的MCMC方法的Gibbs算法。假设有p维待估参数$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_p)^{\mathrm{T}}$,其中$\pi(\alpha|Y)$为观察到数据集后的参数联合后验分布,其基本的抽样步骤如下:

步骤1 任意选取初值$\alpha^{(0)}=(\alpha_1^{(0)},\alpha_2^{(0)},\cdots,\alpha_p^{(0)})^{\mathrm{T}}$,设置第$l$次迭代值为$\alpha^{(l)}$,$l=1$;

步骤2 根据$\pi(\alpha_1|\alpha_2^{(l-1)},\cdots,\alpha_p^{(l-1)},Y)^{\mathrm{T}}$抽取$\alpha_1^{(l)}$;

根据$\pi(\alpha_2|\alpha_3^{(l-1)},\cdots,\alpha_p^{(l-1)},\alpha_1^{(l)},Y)^{\mathrm{T}}$抽取$\alpha_2^{(l)}$。

$$\cdots$$

根据$\pi(\alpha_p|\alpha_1^{(l)},\alpha_2^{(l)},\cdots,\alpha_{p-1}^{(l)},Y)^{\mathrm{T}}$抽取$\alpha_p^{(l)}$。

步骤3 令$l=l+1$,返回步骤2,直到算法收敛。

其中,$\pi(\alpha_p|\alpha_1^{(l)},\alpha_2^{(l)},\cdots,\alpha_{p-1}^{(l)},Y)$称为$\alpha_p$的满条件分布。当$l\to\infty$时,按照以上步骤得到的随机序列$\alpha^{(1)},\alpha^{(2)},\cdots,\alpha^{(l)},\cdots$,就是来自$\pi(\alpha|Y)$的随机样本。

对所有未知参数进行收敛性检验是很关键的步骤,只有$\alpha^{(1)},\alpha^{(2)},\cdots,\alpha^{(l)},\cdots$构成的Markov链收敛才能够实现贝叶斯推断的准确性。MCMC算法的收敛性可以通过观察未知参数模拟得到的Markov链的迭代图[15]判断,如果该链能够快速远离初始值,且始终在某一固定值附近随机波动,没有明显趋势和周期,说明算法是收敛的。

2.2 基于spike-and-slab先验的贝叶斯层次时间序列模型

我们可以对平稳时间序列建立如下结构的模型,并用ADF单位根检验法来检验时间序列的平稳性,同时得到模型阶数p。

时间序列是由多个滞后阶数引起的,许多滞后阶数与预测结果几乎没有关系,因此我们假设只有一些回归系数不等于零,由此变量选择问题就归结为非零回归系数的识别。我们采用一种稀疏模型,在层次结构框架内进行变量选择和参数估计。

将预测的时间变量看作响应变量y_t,将滞后的时间变量看作包含p个潜在预测因子的解释变量$X=(y_{t-1},y_{t-2},\cdots,y_{t-p})$。

根据式(1),可得到响应变量的多元线性回归模型为:

$$y_{ti}=\sum_{j=1}^{p}y_{(t-j)i}\beta_j+\varepsilon_i \tag{2}$$

则可知:

$$y_t|\beta,\sigma^2\sim N(X\beta,\sigma^2I_n) \tag{3}$$

考虑到时间序列之间的相关性并基于偏自相关系数,将结合偏自相关系数的spike-and-slab先验分布定义如下:

$$\beta_j|\gamma_j\sim(1-\gamma_j)N(\beta_j;\tau_1)+\gamma_jN(\beta_j;\tau_2) \tag{4}$$

其中,$N(\beta_j;\tau_c)=\left(\frac{2\pi}{\tau_c}\right)^{-\frac{1}{2}}\mathrm{e}^{-\frac{\tau_c\beta_j^2}{2}}$,$c=1,2$且$\tau_1=\frac{1}{\eta}$,$\tau_2=\frac{1}{|\varphi_{kk}|}$($|\varphi_{kk}|>\eta$,$0<\eta<1$)。另外,指示变量$\gamma_j$对于协变量$y_{(t-j)i}$是否进入预测模型具有不可忽略的贡献,将其扩展到包含概率:

$$\gamma_j|\delta_j\sim Bernoulli(\delta_j) \tag{5}$$

$$\delta_j \sim U(0,1) \tag{6}$$

先验通过 spike 鼓励稀疏性，并将 slab 之下的系数收缩为零。模型选择中的不确定性很容易通过每个γ_j的后验概率来解释。如果$\gamma_j=1$，表明$y_{(t-j)}$应当进入模型中，满足$|\varphi_{kk}|>\eta$，表明滞后 k 阶的时间变量对响应变量有显著影响，模型此时由 slab 分布 $N(\beta_j;\tau_2)$控制，$\tau_2=\frac{1}{|\varphi_{kk}|}$使得 $N(\beta_j;\tau_2)$拥有较大的方差，在 0 附近产生分散效应，表现为$\beta_j\neq 0$，从而将对应的预测因子选入模型，并且$|\varphi_{kk}|$对于参数的估计有着自适应的效果；如果$\gamma_j=0$，表明$y_{(t-j)}$不应当进入模型中，模型此时由 spike 分布 $N(\beta_j;\tau_1)$控制，$\tau_1=\frac{1}{\eta}$使得 $N(\beta_j;\tau_1)$拥有较小的方差，从而产生 0 附近的集中效应，对应的β_j估计为 0，从而将无关预测因子剔除出模型。另外，η的选择也可以有效控制选入模型的变量个数，较大的 η 可能会剔除更多时间变量，这里的超参数 η的确定采用五折交叉验证方法。

除此之外，我们将σ^2先验分布确定为逆伽马分布，即：

$$\sigma^2 \sim IG(a,b) \tag{7}$$

众多研究表明，此分层模型下，逆伽马分布在最小后验均方误差准则下得到的参数估计值十分逼近真值。其中的超参数确定为 Samorodnitsky S 建议的值：$a=b=0.01$[16]。

由于β_j的二值隐变量γ_j决定了$y_{(t-j)}$是否应该被选入模型，因此γ_j的分布函数对β_j至关重要。由于γ_j在宏观意义上表示系数是否取零值，为保证系数稀疏性的准确度，规定只有当取γ_j取零值的条件后验概率不小于 0.5 时，系数估计为零的信息才可以被接受，即$\hat{\gamma}_j=0<=>P(\gamma_j=0|\beta_j,\sigma^2,\delta_j,y_t)\geqslant 0.5$。

综上，将本文的模型总结如下：

$$y_t|\beta,\sigma^2 \sim N(X\beta,\sigma^2 I_n)$$

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(\beta_j;\tau_1)+\gamma_j N(\beta_j;\tau_2)$$

$$\gamma_j|\delta_j \sim Bernoulli(\delta_j)$$

$$\delta_j \sim U(0,1)$$

$$\sigma^2 \sim IG(a,b)$$

2.3 基于 MCMC 算法的后验推断

得到上述贝叶斯层次模型后，关键一步就是使用后验分布对未知参数进行推断。MCMC 方法提供了一种直接、直观的方法来模拟未知分布的值，并使用这些模拟值进行后续分析。根据 2.1 节中 MCMC 方法的基本理论，可以得到具体的抽样步骤如下。

由各参数的先验分布可得到$(\beta,\gamma,\sigma^2,\delta)$的联合后验分布为：

$$\pi(\beta,\gamma,\sigma^2,\delta|y_t)\propto\pi(y_t|\beta,\sigma^2)\pi(\beta|\gamma)\pi(\gamma|\delta)\pi(\delta)\pi(\sigma^2) \tag{8}$$

根据联合后验分布可得以下参数的满条件分布：

$$(1)\pi(\delta|\beta,\gamma,\sigma^2,y_t)\propto\frac{\prod_{j=1}^{p}\delta_j^{\gamma_j+1}(1-\delta_j)^{(1-\gamma_j)}}{[\Gamma(\gamma+2)\Gamma(2-\gamma)]^p}$$

$$(2)\pi(\gamma|\beta,\delta,\sigma^2,y_t)\propto\frac{\prod_{j=1}^{p}\left[(1-\gamma_j)\mathrm{e}^{-\frac{\tau_1\beta_j^2}{2}}+\gamma_j\mathrm{e}^{-\frac{\tau_1\beta_j^2}{2}}\right]\cdot\delta_j^{\gamma_j}(1-\delta_j)^{(1-\gamma_j)}}{\prod_{j=1}^{p}\left[\mathrm{e}^{-\frac{\tau_1\beta_j^2}{2}}(1-\delta_j)+\mathrm{e}^{-\frac{\tau_1\beta_j^2}{2}}\cdot\delta_j\right]}$$

$$(3)\pi(\beta|\gamma,\delta,\sigma^2,y_t)\propto\frac{\exp\left\{-\frac{1}{2\sigma^2}\prod_{i=1}^{n}(y_{ti}-\sum_{k=1}^{p}y_{(t-k)i}\beta_k)^2\right\}\cdot\prod_{j=1}^{p}\left[(1-\gamma_j)\mathrm{e}^{-\frac{\tau_1\beta_j^2}{2}}+\gamma_j\mathrm{e}^{-\frac{\tau_2\beta_j^2}{2}}\right]}{\prod_{j=1}^{p}\left[\frac{(1-\gamma_j)\sigma}{\sqrt{\tau_1\sigma^2+\sum_{i=1}^{n}[y_{(t-j)i}]^2}}+\frac{\gamma_j\sigma}{\sqrt{\tau_2\sigma^2+\sum_{i=1}^{n}[y_{(t-j)i}]^2}}\right]\cdot\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}y_{ti}^2}}$$

$$(4)\pi(\sigma^2|\beta,\delta,\sigma^2,y_t)\propto\frac{\exp\left\{-\frac{1}{2\sigma^2}\prod_{i=1}^{n}(y_{ti}-\sum_{k=1}^{p}y_{(t-k)i}\beta_k)^2\right\}\cdot(\sigma^2)^{-1.01}\mathrm{e}^{-\frac{0.01}{\sigma^2}}}{\left[\frac{1}{2}\prod_{i=1}^{n}(y_{ti}-\sum_{k=1}^{p}y_{(t-k)i}\beta_k)^2+0.01\right]^{-0.01}}$$

根据上述求得的满条件分布可进行 Gibbs 抽样：首先选取初值$\beta^{(0)},\sigma^{2(0)},\delta^{(0)},\gamma^{(0)}$，令 $l=1$；根据 $\pi(\delta|\beta^{(l-1)},\gamma^{(l-1)},\sigma^{2(l-1)},y_t)$抽取$\delta^{(l)}$；根据 $\pi(\gamma|\beta^{(l-1)},\sigma^{2(l-1)},\delta^{(l)},y_t)$抽取$\gamma^{(l)}$；根据 $\pi(\beta|\sigma^{2(l-1)},\gamma^{(l)},\delta^{(l)},y_t)$抽取$\beta^{(l)}$；根据 $\pi(\sigma^2|\beta^{(l)},\gamma^{(l)},\delta^{(l)},y_t)$更新$\sigma^{2(l)}$；令 $l=l+1$，返回第 2 步，直到算法收敛。

3 模拟分析

为了验证该模型的性能，首先在模拟数据上进行了实验。首先，根据以下模型产生一列时间序列数据：

$$y_t=0.053y_{t-1}-0.072y_{t-2}+0.131y_{t-4}-0.142y_{t-5}+0.310y_{t-7}-0.252y_{t-8}+\varepsilon_t$$

$$\varepsilon_i \sim N(0,0.01^2)$$

由于本文模型是基于平稳时间序列提出的，所以需要判断该模拟时间序列是否平稳。将模拟出的数据可视化，初步判断其是否平稳，根据生成的模拟数据得到时序图，如图 1 所示。

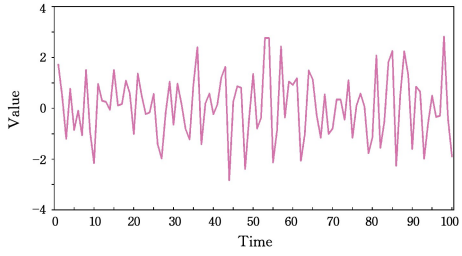

图 1 模拟数据时序图

Fig. 1 Analog data sequence diagram

曲线围绕 0 值上下波动，波动幅度前后、上下一致，因此初步判定其为平稳序列。采用 ADF 单位根检验方法进一步验证，得到 $p<0.01$ 且模型阶数为 8，因此经 ADF 单位根检验后可确定该序列是平稳的，可以对该时间序列进行进一步分析。

对该时间序列建立本文提出的模型，在本文中，模型中存在的超参数 $\eta=\{0.01,0.05\}$，经由五折交叉验证方法得到 $\eta=0.05$时的预测误差更小，因此在模拟实验中，取 $\eta=0.05$。

为了实验结果更直观，取数据的前 80% 作为训练集，其余的 20% 作为测试集，进行变量选择和参数估计。

下面根据 2.3 节的步骤进行样本的抽取，令迭代次数为 10^4，图 2 为 MCMC 算法的参数抽样痕迹图。

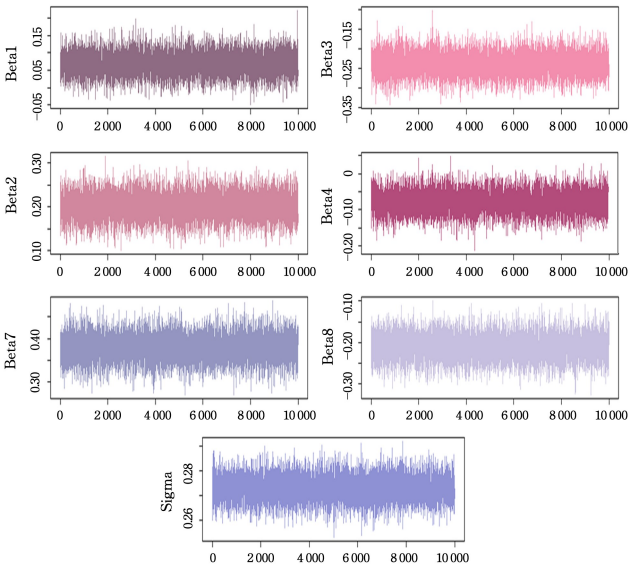

图 2 模拟数据参数 Markov 链迭代图

Fig. 2 Markov chain iteration diagram of simulation data parameters

根据各参数的抽样迭代图可清晰看到各参数的 Markov 链能够在某一固定值附近随机波动，没有明显趋势和周期，从迭代初期开始就是较为稳定的，说明算法是收敛的，即得到的贝叶斯估计是有效的。

为了说明结果的有效性和可靠性，以 RMS 来说明估计参数的估计精度，该值越小表明贝叶斯估计的结果越精确。RMS 的计算公式如下：

$$\hat{d}=\frac{1}{m}\sum_{l=1}^{m}d^{(l)} \tag{9}$$

$$RMS=\sqrt{\frac{1}{m}\sum_{l=1}^{m}(d^{(l)}-d)^2} \tag{10}$$

其中，d 为参数的真实值，$\hat{d}$ 为 d 的贝叶斯估计值，在本文中以均值作为估计 $d^{(l)}$ 为第 l 次迭代时对应参数生成的随机样本。表 1 为具体的模拟结果。

表 1 模拟结果

Table 1 Simulation results

β	真值	均值	*RMS*	3%分位数	97%分位数
β_1	0.053	0.069	0.032	0.010	0.130
β_2	0.072	−0.238	0.029	−0.293	−0.182
β_4	0.131	0.202	0.030	0.147	0.258
β_5	−0.142	−0.079	0.029	−0.135	−0.025
β_7	0.310	0.378	0.029	0.324	0.433
β_8	−0.252	−0.213	0.031	−0.272	−0.155

从表 1 中可以看到，置信区间较窄，各参数的 *RMS* 均在 0.035 以下，表明估计值与真值较为接近，估计精度和可信度较高。根据表中的估计值可以得到最后的训练模型如下：

$$y_t=0.069y_{t-1}-0.238y_{t-2}+0.202y_{t-4}-0.079y_{t-5}+0.378y_{t-7}-0.213y_{t-8}$$

为了更清楚地看出估计值趋势和与真实值的差异，根据训练所得的模型对模拟数据进行拟合并得到拟合图，如图 3 所示。

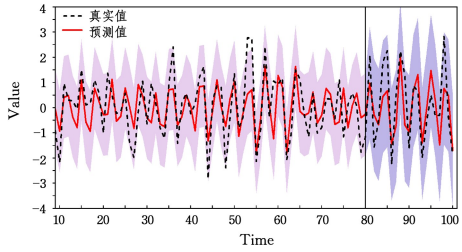

图 3 模拟数据拟合图

Fig. 3 Simulation data fitting diagram

可以从图 3 中看出，得到的贝叶斯估计值较真实值曲线虽有差异，但该曲线上升与下降趋势基本与真实值吻合；同时，许多估计值与真实值非常接近，真实值几乎都落在 95% 置信区间内。总体来说，所提模型拟合效果良好。

为了更清楚地看到本文模型优势，我们还采用 Lasso，SCAD，Elastic-Net，Adaptive Lasso 方法对模拟的时间序列进行了变量选择和参数估计来与本文所提出的模型进行对比，并且为了更直观地体现出本文模型的优势，我们根据训练所得的模型对模拟数据进行了预测，所得结果如表 2 所列。

表 2 不同方法的变量选择和参数估计结果

Table 2 Variable selection and parameter estimation results of different methods

模型	变量选择和参数估计结果
Lasso	$y_t=-0.205y_{t-2}+0.131y_{t-4}-0.016y_{t-5}+0.368y_{t-7}-0.094y_{t-8}$
SCAD	$y_t=-0.209y_{t-2}+0.134y_{t-4}-0.033y_{t-5}+0.336y_{t-7}-0.113y_{t-8}$
Elastic-Net	$y_t=-0.215y_{t-2}+0.152y_{t-4}-0.034y_{t-5}+0.364y_{t-7}-0.119y_{t-8}$
Adaptive Lasso	$y_t=-0.233y_{t-2}+0.134y_{t-4}-0.142y_{t-5}+0.310y_{t-7}-0.252y_{t-8}$
SS-PAC	$y_t=0.069y_{t-1}-0.238y_{t-2}+0.202y_{t-4}-0.079y_{t-5}+0.378y_{t-7}-0.213y_{t-8}$

从表 2 中数据得知，Lasso，SCAD，Elastic-Net 和 Adaptive Lasso 都只选择了 5 个预测因子，与设定的模型不符，只有本文模型正确选择了所有的预测因子并且参数的估计也较为准确。另外由表 3 各模型的拟合误差和预测误差数据可得知，所提模型无论是在训练集还是测试集中都是表现最好的。

表 3 不同方法的误差

Table 3 Errors of different methods

模型	拟合误差	预测误差
Lasso	0.870082919	1.276530 26
SCAD	0.869474324	1.245179976
Elastic-Net	0.856727368	1.223350144
Adaptive Lasso	0.848093978	1.242355896
SS-PAC	0.832970889	1.138161 83

4 实证分析

本节将本文所提出的模型应用到医学、环境以及经济 3 方面进行了实证分析，选取的数据分别为北京市肺结核发病数、北京市 pm2.5 指数以及白银价格数据。由于实证过程类似，接下来仅以北京肺结核发病数为例介绍实证过程。关于该数据，我们收集了从 2010 年 1 月至 2018 年 12 月的逐月数据，其时序图如图 4 所示。

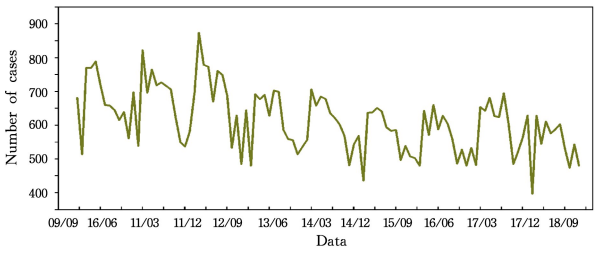

图 4　北京市肺结核发病数时序图

Fig. 4　Time series chart of pulmonary tuberculosis incidence in Beijing

根据时序图可以初步认为其是平稳时间序列。采用 ADF 单位根检验方法进一步进行平稳性验证，得到 $p<0.01$ 且模型阶数为 12，因此该序列确实为平稳时间序列，可进行下一步分析。对此序列建立模型，模型中超参数取 $\eta=\{0.01,0.05\}$，经五折交叉验证方法确定 $\eta=0.01$。为了实验结果更直观，同样取数据的前 80% 作为训练集，其余的 20% 作为测试集，进行变量选择和参数估计。

下面根据 2.3 节的步骤进行样本的抽取，令迭代次数为 10^4。图 5 为 MCMC 算法的参数抽样痕迹图。

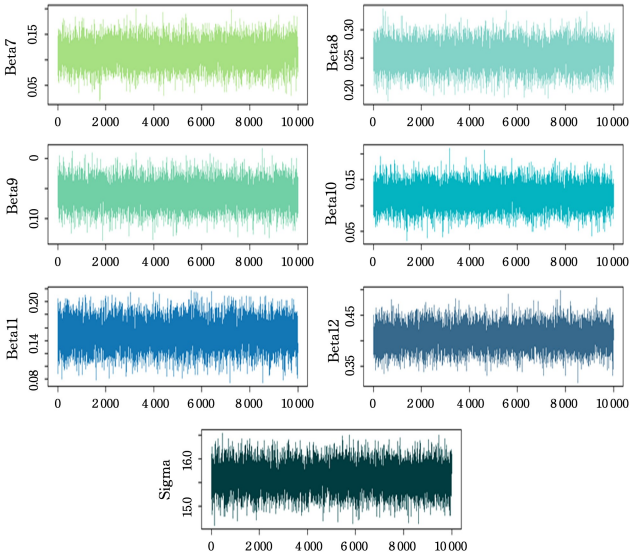

图 5　北京肺结核发病数参数 Markov 链迭代图

Fig. 5　Markov chain iterative chart of incidence parameters of pulmonary tuberculosis in Beijing

根据各参数的抽样迭代图可清晰看到各参数的 Markov 链能够在某一固定值附近随机波动，没有明显趋势和周期，从迭代初期开始迭代过程就是较为稳定的，说明算法是收敛的，即得到的贝叶斯估计是有效的。具体的参数估计结果如表 4 所列。

表 4　北京肺结核发病数参数估计

Table 4　Parameter estimation of pulmonary tuberculosis incidence in Beijing

β	均值	3%分位数	97%分位数
β_1	0.112	0.067	0.155
β_2	0.253	0.215	0.298
β_3	−0.059	−0.098	−0.021
β_{10}	0.118	0.075	0.159
β_{11}	0.149	0.110	0.188
β_{12}	0.407	0.364	0.449

从表 4 中可以看到，本文模型在北京肺结核发病数时间序列的 12 阶滞后阶数里选择了 1 阶、2 阶、3 阶、10 阶、11 阶和 12 阶这 6 个滞后阶数，大大减少了变量个数。根据表中的估计值可以得到最后的训练模型如下：

$$y_t=0.112y_{t-1}+0.253y_{t-2}-0.059y_{t-3}+0.118y_{t-10}+0.149y_{t-11}+0.407y_{t-12}$$

将根据训练所得模型得到的估计曲线与真实数据曲线进行拟合，拟合结果如图 6 所示。

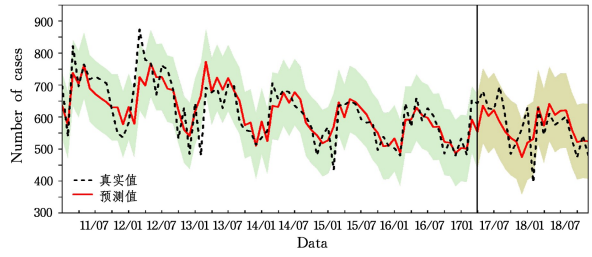

图 6　北京肺结核发病数拟合图

Fig. 6　Fitting chart of incidence of pulmonary tuberculosis in Beijing

从拟合图可以看出估计曲线整体波动趋势与真实数据曲线基本一致，没有过拟合的情况，对于后续的预测更为有益，真实值几乎都落在 95% 置信区间内。总体来说，本文模型关于北京肺结核发病数的拟合效果良好。

同样地，使用 LASSO，SCAD，Elastic-Net 和 Adaptive Lasso 方法对该时间序列数据进行了变量选择和参数估计并与本文所提出的模型进行对比，所得结果如表 5 所列。

表 5　不同方法的变量选择和参数估计结果

Table 5　Variable selection and parameter estimation results of different methods

模型	变量选择和参数估计结果
Lasso	$y_t=0.005y_{t-1}+0.234y_{t-2}+0.060y_{t-10}+0.168y_{t-11}+0.512y_{t-12}$
SCAD	$y_t=0.229y_{t-2}+0.023+0.153y_{t-11}+0.533y_{t-12}$
Elastic-Net	$y_t=0.007y_{t-1}+0.233y_{t-2}+0.062y_{t-10}+0.169y_{t-11}+0.509y_{t-12}$
Adaptive Lasso	$y_t=0.228y_{t-2}+0.001y_{t-10}+0.153y_{t-11}+0.590y_{t-12}$
SS-PAC	$y_t=0.112y_{t-1}+0.253y_{t-2}-0.059y_{t-3}+0.118y_{t-10}+0.149y_{t-11}+0.407y_{t-12}$

根据训练所得的模型对本文的 3 个实证时间序列数据进行预测，结果如表 6 所列。

表 6　不同方法的误差

Table 6　Error of different methods

模型	误差	医学	环境	经济
		北京市肺结核发病数	北京市 pm2.5 指数	白银价格
Lasso	Train	53.220650 11	24.746954 78	0.142876628
	Test	59.889079 29	10.663422 1	0.093531986
SCAD	Train	59.515186 13	28.498182 19	0.138558983
	Test	65.597525 47	14.935870 77	0.103819533
Elastic-Net	Train	53.223522 2	24.691041 7	0.136850524
	Test	59.900785 88	10.475069 04	0.107001598
Adaptive Lasso	Train	53.854394 46	28.326864	0.138290066
	Test	60.005963 75	14.713560 67	0.103467083
SS-PAC	Train	53.077095 06	24.637627 32	0.140139947
	Test	59.665301 13	10.285357 3	0.092591948

由表 5 和表 6 第三列可知，文中模型选取了 6 个预测

因子，无论是在训练集还是测试集上，其在所有方法中效果都是最好的。Lasso 和 Elastic-Net 方法选取了 5 个预测因子，模型误差稍大。Adaptive Lasso 和 SCAD 虽然选取了更少的 4 个预测因子，但误差也随之增大，在 5 种模型中拟合效果最差。

表中第 2，3 列为其余两个数据集的变量选择与参数估计结果，可以明显看出，在北京市 pm2.5 指数这一数据集中我们的模型无论是在训练集还是测试集都取得了最佳效果。在白银价格数据集中，文中模型虽然在训练集中误差值没有达到最小，但在测试集上达到最小误差，而且在训练集中的误差和误差值最小的相比差值极小。由此可知，我们的模型在实际的时间序列数据中的应用也是可行的。

结束语 本文提出一种基于 spike-and-slab 先验的贝叶斯层次时间序列模型(SS-PAC)。模型将贝叶斯变量选择方法引入时间序列预测研究，使用具有稀疏特性的 spike-and-slab 先验并结合时间序列数据的偏自相关系数建立贝叶斯时间序列模型。同时，使用后验分布对未知参数进行推断，采用 Gibbs 抽样算法模拟未知分布并进行分析，结果表明算法收敛，估计结果有效。通过实证研究发现，SS-PAC 模型在医学、环境以及经济等多种时间序列数据上，与传统方法 Lasso，SCAD，Elastic-Net，Adaptive Lasso 相比，SS-PAC 模型更能准确地选取变量，趋势预测效果较好，时间序列预测结果更优。

参 考 文 献

[1] DE GOOIJER J G, HYNDMAN R J. 25 years of time series forecasting[J]. International Journal of Forecasting, 2006, 22(3): 443-473.

[2] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B(Methodological), 1996, 58(1): 267-288.

[3] FAN J, LI R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[4] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2005, 67(2): 301-320.

[5] ZOU H. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[6] VERBESSELT J, ROBINSON A, STONE C, et al. Forecasting tree mortality using change metrics derived from MODIS satellite data[J]. Forest Ecology and Management, 2009, 258(7): 1166-1173.

[7] ZHANG C M, ZHANG Z J. Regularized estimation of hemodynamic response function for fMRI data[J]. Statistics and Its Interface, 2010, 3(1): 15-31.

[8] MITCHELL T J, BEAUCHAMP J J. Bayesian variable selection in linear regression[J]. Am Stat Assoc, 1988, 83(404): 1023-1032.

[9] GEORGE E I, MCCULLOCH R E. Variable selection via gibbs sampling[J]. Am Stat Assoc, 1993, 88(423): 881-889.

[10] KUO L, MALLICK B. Variable selection for regression models [J]. Sankhyā Indian J Stat Ser B, 1998, 66(1): 65-81.

[11] YANG A, XIANG J, SHU L, et al. Sparse Bayesian Variable Selection with Correlation Prior for Forecasting Macroeconomic Variable using Highly Correlated Predictors[J]. Computational Economics, 2017, 51(2): 323-338.

[12] FRANKE P M, HUNTLEY B, PARNELL A C. Frequency selection in paleoclimate time series: A model-based approach incorporating possible time uncertainty[J]. Environmetrics, 2018, 29(2): 1-19.

[13] LI Y, LUND R, HEWAARACHCHI A. Multiple changepoint detection with partial information on changepoint times[J]. Electronic Journal of Statistics, 2019, 13(2): 2462-2520.

[14] CHEN C W S, LIU F C, PINGAL A C. Integer-valued transfer function models for counts that show zero inflation[J]. Statistics & Probability Letters, 2023, 193(1): 109701.

[15] NELDER J, WEDDERBURN R. Generalized Linear Models[J]. Journal of the Royal Statistical Society, 1972(1): 370-384

[16] SAMORODNITSKY S, HOADLEY K A, LOCK E F. A hierarchical spike-and-slab model for pan-cancer survival using panomic data[J]. BMC Bioinformatics, 2022, 23(1): 235.

[17] LIU J S, XIA Q. Bayesian Statistical Method Based on MCMC Algorithm [M]. Beijing: Science Press, 2016.

[18] JOSHUAS. A Conceptual Introduction to Markov Chain Monte Carlo Methods[J]. arXiv: Other Statistics, 2019.

[19] POSCH K, ARBEITER M, PILZ J. A novel Bayesian approach forvariable selection in linear regression models[J]. Computational Statistics and Data Analysis, 2020, 144(1): 106881.

[20] OUYANG L, PARK C, MA Y, et al. Bayesian hierarchical modelling for process optimization[J]. International Journal of Production Research, 2020, 59(15): 4649-4669.

[21] LIN Z, VEERABHADRAN B. Bayesian hierarchical structured variable selectionmethods with application to molecular inversionprobe studies in breast cancer[J]. J R Stat Soc Ser C Appl Stat, 2014, 63(4): 595-620.

[22] MITCHELL T J, BEAUCHAMP J J. Bayesian Variable Selection inLinear Regression[J]. Journal of the American Statistical Association, 1988, 83(404): 1023-1032.

[23] LIU Z, ZHOU J L, DONG C L. Bayesian estimation of multivariate linear regression change point model based on MCMC algorithm [J]. Henan Science, 2020, 38(8): 1210-1214.

[24] VERONIKA R, EMMANUEL L, JOLANDA L, et al. Hierarchical Bayesian formulations for selecting variables in regression models[J]. Statistics in Medicine, 2012, 31(11/12): 1221-1237.

GUO Chenlei, born in 1997, postgraduate. Her main research interests include variable selection and so on.

LI Dongxi, born in 1982, Ph.D, associate professor, postgraduate supervisor. His main research interests include high dimensional data analysis, data mining, machine learning, biostatistics and biological mathematics.