Statistics and Application 统计学与应用, 2020, 9(5), 900-908
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.95094

Hans汉斯

纵向数据下均值协方差模型的贝叶斯统计诊断

徐登可[1]，赵远英[2]

[1]浙江农林大学统计系，浙江 杭州
[2]贵阳学院数学与信息科学学院，贵州 贵阳
Email: 175384319@qq.com

收稿日期：2020年10月7日；录用日期：2020年10月22日；发布日期：2020年10月29日

摘 要

研究了纵向数据下均值协方差模型的贝叶斯统计诊断。通过应用Gibbs抽样和Metropolis-Hastings (MH)算法相结合的混合算法获得模型贝叶斯数据删除影响诊断统计量来识别数据异常点。模拟研究和实例分析都表明所提出的诊断方法是可行有效的。

关键词

纵向数据，数据删除，Gibbs抽样，MH算法，贝叶斯诊断

Bayesian Statistical Diagnosis of Joint Mean and Covariance Models with Longitudinal Data

Dengke Xu[1], Yuanying Zhao[2]

[1]Department of Statistics, Zhejiang Agriculture and Forestry University, Hangzhou Zhejiang
[2]College of Mathematics and Information Science, Guiyang University, Guiyang Guizhou
Email: 175384319@qq.com

Received: Oct. 7th, 2020; accepted: Oct. 22nd, 2020; published: Oct. 29th, 2020

Abstract

Bayesian statistical diagnosis of joint mean and covariance models with longitudinal data is stu-

文章引用: 徐登可, 赵远英. 纵向数据下均值协方差模型的贝叶斯统计诊断[J]. 统计学与应用, 2020, 9(5): 900-908.
DOI: 10.12677/sa.2020.95094

died. By combining the Gibbs sampler and Metropolis-Hastings algorithm, the Bayesian case deletion diagnosis statistic is obtained to identify data outliers. Simulation study and a real data analysis show that the proposed diagnosis method is feasible and effective.

Keywords

Longitudinal Data, Case Deletion, Gibbs Sampler, Metropolis-Hastings Algorithm, Bayesian Diagnosis

1. 引言

纵向数据常常出现在经济学，生物学，环境科学，医学等领域。此类数据是指同一组受试个体在不同时间点上的重复观测数据，它具有组间独立，组内相关的特性。因此，在纵向数据分析中协方差矩阵的估计是很重要的，并且对协方差矩阵进行估计也是统计学家感兴趣的。众所周知，一个好的协方差矩阵估计是提高回归系数估计效率的常用方法。而对协方差矩阵进行估计的方法一般都是基于 Cholesky 分解建模展开的。目前已有很多作者基于此分解研究了协方差矩阵的估计问题。其中，Pourahmadi [1] [2]对协方差矩阵进行改进的 Cholesky 分解且对分解后的部分进行广义线性建模，然后考虑了模型中的参数估计问题。这种建模的主要优点包括便于统计意义上的解释和参数估计中的方便计算。类似的研究还可见文献[3] [4]。另外，Rothman 等[5]提出了通过新的改进的 Cholesky 分解来参数化协方差矩阵自身，给出了分解后因子的新的回归解释，并且也确保了估计得到的协方差矩阵的正定性。进一步地，基于这种新的改进的 Cholesky 分解，Zhang 和 Leng [6]针对联合均值协方差建模提出了一种有效的极大似然估计方法。Xu 等[7]基于惩罚极大似然方法提出了一种有效的变量选择方法。

另一方面，异常点的识别是统计诊断的主要内容，而基于数据删除影响进行统计诊断是识别异常点的一个重要方法。自从 Cook [8]在这一领域做了开创性的工作之后，数据删除影响诊断方法，尤其是贝叶斯数据删除影响诊断方法引起越来越多的关注。例如，Cho 等[9]在一般模型的框架下，以复杂生存数据模型为例提出一种简单可行的贝叶斯数据删除 K-L 距离计算公式。其它相关内容还可以参见文献[10] [11] [12]。尽管已经有众多的作者研究贝叶斯数据删除影响诊断方法，也做出了卓有成效的工作。然而，很少有文献基于新的改进的 Cholesky 分解研究纵向数据下均值协方差模型的贝叶斯数据删除影响诊断问题。因此针对纵向数据下均值协方差模型，本文基于 K-L 距离研究贝叶斯数据删除影响诊断方法，以识别模型的异常点。

2. 模型与符号

2.1. 纵向数据下均值协方差模型

假设有 n 个独立样本和对第 i 个样本进行 m_i 次重复观测。具体地，记第 i 个个体在时间 $t_i=\left(t_{i1},\cdots,t_{im_i}\right)^{\mathrm{T}}$ 的响应变量向量为 $Y_i=\left(Y_{i1},\cdots,Y_{im_i}\right)^{\mathrm{T}}, i=1,\cdots,n$，并且假设响应变量服从正态分布 $Y_i\sim N\left(\mu_i,\Sigma_i\right)$，其中 $\mu_i=\left(\mu_{i1},\cdots,\mu_{im_i}\right)^{\mathrm{T}}$ 是一个$(m_i\times 1)$向量和 Σ_i 是一个$(m_i\times m_i)$正定矩阵。

其中基于改进的 Cholesky 分解，Pourahmadi [1]首先提出分解 $\sum_i$ 为 $T_i \sum_i T_i^{\mathrm{T}} = D_i$，其中 T_i 是下三角矩阵，对角线元素为 1，下三角元素是自回归模型 $Y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} \left(Y_{ik} - \mu_{ik} \right) + \varepsilon_{ij}$ 中的自回归参数 ϕ_{ijk} 的负数。D_i 的对角元素是新息方差 $\sigma_{ij}^2 = Var\left(\varepsilon_{ij} \right)$。

而 Rothman 等[5]提出了新的分解思想，令 $L_i = T_i^{-1}$，它是对角线元素为 1 的下三角矩阵，因此可以写成 $\sum_i = L_i D_i L_i^{\mathrm{T}}$。实际上也就是利用 Rothman 等[5]提出的改进的 Cholesky 分解来表示协方差矩阵，这样就可以用新的统计意义解释。L_i 中的元素 l_{ijk} 可以解释为以下滑动平均模型的滑动平均系数：$Y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} l_{ijk} \varepsilon_{ij} + \varepsilon_{ij}, j = 2, \cdots, m_i$。其中 $\varepsilon_{i1} = Y_{i1} - \mu_{i1}$ 和 $\varepsilon_i \sim N\left(0, D_i \right)$，且 $\varepsilon_i = \left(\varepsilon_{i1}, \cdots, \varepsilon_{im_i} \right)^{\mathrm{T}}$。注意到这里的 l_{ijk} 和 $\log\left(\sigma_{ij}^2 \right)$ 是没有限制的。基于改进的 Cholesky 分解和受到 Pourahmadi [1] [2]，Ye 和 Pan [13]的启发，无限制的参数 μ_{ij}, l_{ijk} 和 $\log\left(\sigma_{ij}^2 \right)$ 可以用以下线性模型来建模

$$\mu_{ij} = X_{ij}^{\mathrm{T}} \beta, l_{ijk} = Z_{ijk}^{\mathrm{T}} \gamma, \log\left(\sigma_{ij}^2 \right) = H_{ij}^{\mathrm{T}} \lambda \tag{1}$$

其中 X_{ij}, Z_{ijk} 和 H_{ij} 分别是 $p \times 1, q \times 1$ 和 $d \times 1$ 协变量向量。协变量 X_{ij} 和 H_{ij} 是回归分析中的一般协变量，Z_{ijk} 一般可以取成时间差 $t_{ij} - t_{ik}$ 的多项式，即，$Z_{ijk} = \left(1, \left(t_{ij} - t_{ik} \right), \cdots, \left(t_{ij} - t_{ik} \right)^{q-1} \right)^{\mathrm{T}}$。另外，记 $X_i = \left(X_{i1}, \cdots, X_{im_i} \right)^{\mathrm{T}}$ 和 $H_i = \left(H_{i1}, \cdots, H_{im_i} \right)^{\mathrm{T}}$。进一步记 γ 为滑动平均系数和 λ 为新息系数。

由模型(1)，可以获得如下似然函数

$$\begin{aligned} & L\left(\beta, \gamma, \lambda \mid Y, X, Z, H, T \right) \\ & = \left(2\pi \right)^{-\frac{n}{2}} \prod_{i=1}^{n} \left| \sum_i \right|^{-\frac{1}{2}} \exp\left(\frac{1}{2} \sum_{i=1}^{n} \left(Y_i - X_i \beta \right)^{\mathrm{T}} \sum_i^{-1} \left(Y_i - X_i \beta \right) \right) \\ & = \left(2\pi \right)^{-\frac{n}{2}} \prod_{i=1}^{n} \left| D_i \right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \varepsilon_i^{\mathrm{T}} D_i^{-1} \varepsilon_i \right) \end{aligned} \tag{2}$$

其中 $Y = \left(Y_1^{\mathrm{T}}, \cdots, Y_n^{\mathrm{T}} \right)^{\mathrm{T}}, X = \left(X_1^{\mathrm{T}}, \cdots, X_n^{\mathrm{T}} \right)^{\mathrm{T}}, H = \left(H_1^{\mathrm{T}}, \cdots, H_n^{\mathrm{T}} \right)^{\mathrm{T}}$ 且 $Z = \left(Z_{ijk}, i = 1, \cdots, n, j = 1, \cdots, m_i, k = 1, \cdots, j-1 \right)$。

2.2. K-L 距离

K-L 距离也叫 K-L 信息，具有距离和信息的某些性质，在统计上以反映两个模型或分布的差异而著称。根据韦博成[14]对 K-L 距离的介绍，密度函数为 $f(x)$ 与 $g(x)$ 的两个分布的 K-L 距离 $K(f, g)$ 被定义为：$K(f, g) = E_f \left\{ \log \frac{f(x)}{g(x)} \right\}$，其中 E_f 表示对 $f(x)$ 求期望。

3. 贝叶斯统计诊断

3.1. 先验分布

为了应用贝叶斯方法来估计模型(1)中的未知参数，需要具体化未知参数的先验分布。为了简便，假设 β, γ 和 λ 相互独立且具有正态先验分布，分别为 $\beta \sim N\left(\beta_0, \Sigma_\beta \right), \gamma \sim N\left(\gamma_0, \Sigma_\gamma \right)$ 和 $\lambda \sim N\left(\lambda_0, \Sigma_\lambda \right)$，其中假设超参数 $\beta_0, \gamma_0, \lambda_0, \Sigma_\beta, \Sigma_\gamma$ 和 Σ_λ 是已知的。

3.2. Gibbs 抽样和条件分布

基于式子(2)，我们可以按照以下过程用 Gibbs 抽样从后验分布 $p\left(\theta|Y,X,Z,H,T\right)$ 中进行抽样，其中 $\theta=\left(\beta^{\mathrm{T}},\gamma^{\mathrm{T}},\lambda^{\mathrm{T}}\right)^{\mathrm{T}}$。

步骤1. 令参数的初值 $\theta^{(0)}=\left(\beta^{(0)^{\mathrm{T}}},\gamma^{(0)^{\mathrm{T}}},\lambda^{(0)^{\mathrm{T}}}\right)^{\mathrm{T}}$。

步骤2. 基于 $\theta^{(l)}=\left(\beta^{(l)^{\mathrm{T}}},\gamma^{(l)^{\mathrm{T}}},\lambda^{(l)^{\mathrm{T}}}\right)^{\mathrm{T}}$，计算 $D_i^{(l)}=diag\left\{\sigma_{i1}^{2(l)},\cdots,\sigma_{im_i}^{2(l)}\right\}$ 和 $\sum_i^{(l)}=L_i^{(l)}D_i^{(l)}L_i^{(l)\mathrm{T}}$。

步骤3. 基于 $\theta^{(l)}=\left(\beta^{(l)^{\mathrm{T}}},\gamma^{(l)^{\mathrm{T}}},\lambda^{(l)^{\mathrm{T}}}\right)^{\mathrm{T}}$ 按照以下抽取 $\theta^{(l+1)}=\left(\beta^{(l+1)^{\mathrm{T}}},\gamma^{(l+1)^{\mathrm{T}}},\lambda^{(l+1)^{\mathrm{T}}}\right)^{\mathrm{T}}$；

- 抽样 $\beta^{(l+1)}$：

$$p\left(\beta|Y,X,Z,H,T,\gamma,\lambda\right)\propto\exp\left\{-\frac{1}{2}\left(\beta-b^*\right)^{\mathrm{T}}B^{*-1}\left(\beta-b^*\right)\right\},\tag{3}$$

其中 $b^*=B^*\left(\sum_{i=1}^n X_i^{\mathrm{T}}\sum_i^{(l)^{-1}}Y_i+\sum_\beta^{-1}\mu_\beta\right)$ 和 $B^*=\left(\sum_\beta^{-1}+\sum_{i=1}^n X_i^{\mathrm{T}}\sum_i^{(l)^{-1}}X_i\right)^{-1}$。

- 抽样 $\gamma^{(l+1)}$：

$$p\left(\gamma|Y,X,Z,H,T,\beta,\lambda\right)\propto\exp\left\{-\frac{1}{2}\sum_{i=1}^n\varepsilon_i^{\mathrm{T}}D_i^{(l)^{-1}}\varepsilon_i-\frac{1}{2}\left(\gamma-\mu_\gamma\right)^{\mathrm{T}}\sum_\gamma^{-1}\left(\gamma-\mu_\gamma\right)\right\}\tag{4}$$

- 抽样 $\lambda^{(l+1)}$：

$$\begin{aligned}&p\left(\lambda|Y,X,Z,H,T,\beta,\gamma\right)\\&\propto\exp\left\{-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^{m_i}H_{ij}^{\mathrm{T}}\lambda-\frac{1}{2}\sum_{i=1}^n\varepsilon_i^{(l+1)^{\mathrm{T}}}D_i^{-1}\varepsilon_i^{(l+1)}-\frac{1}{2}\left(\lambda-\mu_\lambda\right)^{\mathrm{T}}\sum_\lambda^{-1}\left(\lambda-\mu_\lambda\right)\right\}\end{aligned}\tag{5}$$

步骤4. 重复步骤2和3。

那么这样就通过以上算法产生了样本序列 $\left(\beta^{(l)},\gamma^{(l)},\lambda^{(l)}\right),l=1,2,\cdots$，从(3)~(5)式中很容易发现，条件分布 $p\left(\beta|Y,X,Z,H,T,\gamma,\lambda\right)$ 是熟悉的正态分布。从正态分布抽取随机数是比较容易的。但是条件分布 $p\left(\gamma|Y,X,Z,H,T,\beta,\lambda\right)$ 和 $p\left(\lambda|Y,X,Z,H,T,\beta,\gamma\right)$ 是一些不熟悉且相当复杂的分布，如何从这些分布中抽取随机数也变得相当困难。这样，MH 算法就被应用来从这些分布中抽取随机数。选择正态分布 $N\left(\gamma^{(l)},\sigma_\gamma^2\Omega_\gamma^{-1}\right)$ 和 $N\left(\lambda^{(l)},\sigma_\lambda^2\Omega_\lambda^{-1}\right)$ 作为建议分布，其中通过选择 σ_γ^2 和 σ_λ^2，来使得接受概率在 0.25 与 0.45 之间，且取

$$\Omega_\gamma=\sum_\gamma^{-1}+\sum_{i=1}^n\frac{\partial\varepsilon_i^{\mathrm{T}}}{\partial\gamma}D_i^{-1}\frac{\partial\varepsilon_i}{\partial\gamma},\ \Omega_\lambda=\sum_\lambda^{-1}+\frac{1}{2}\sum_{i=1}^n H_iH_i^{\mathrm{T}}.$$

3.3. 贝叶斯估计

利用以上提出的计算过程来产生观测值来获得参数 β,γ 和 λ 的贝叶斯估计。令 $\left\{\theta^{(j)}=\left(\beta^{(j)},\gamma^{(j)},\lambda^{(j)}\right):j=1,2,\cdots,J\right\}$ 是通过上述混合算法从联合条件分布 $p\left(\beta,\gamma,\lambda|Y,X,Z,H,T\right)$ 中产生的观测值，那么 β,γ 和 λ 的贝叶斯估计为：

$$\hat{\beta}=\frac{1}{J}\sum_{j=1}^J\beta^{(j)},\hat{\gamma}=\frac{1}{J}\sum_{j=1}^J\gamma^{(j)},\hat{\lambda}=\frac{1}{J}\sum_{j=1}^J\lambda^{(j)}$$

类似于 Geyer [15]中展示的一样，当 J 趋于无穷时，$\hat{\theta}=\left(\hat{\beta},\hat{\gamma},\hat{\lambda}\right)$ 是对应后验均值向量的相合估计。

3.4. 贝叶斯数据删除影响诊断

在贝叶斯统计分析中，目前已存在许多诊断统计量用以评价个体观测对参数后验分布的影响，本文主要基于 K-L 距离研究贝叶斯数据删除影响的统计诊断方法。对任意的 $i=1,\cdots,n$，记 $\{Y_i,X_i,Z_i,H_i\}$ 是第 i 个个体观测数据点，$D=\{Y,X,Z,H\}$ 为完全数据集，D_{-i} 为完全数据集 D 删除第 i 个个体观测数据点得到的数据集，$L(\theta|D)$ 与 $L(\theta|D_{-i})$ 分别表示基于数据 D 与 D_{-i} 的似然函数，则 θ 于数据 D 与 D_{-i} 的后验分布分别为 $p(\theta|D)\propto L(\theta|D)p(\theta)$，$p(\theta|D_{-i})\propto L(\theta|D_{-i})p(\theta)$。根据 Cho 等[9]的讨论，定义 K-L 距离为：

$$K(P,P_{-i})=\int p(\theta|D)\log\left\{\frac{p(\theta|D)}{p(\theta|D_{-i})}\right\}\mathrm{d}\theta, \tag{6}$$

其中 P 与 P_{-i} 分别表示 θ 基于数据 D 与 D_{-i} 的后验分布，注意到 K-L 距离 $K(P,P_{-i})$ 是完全数据集 D 删除第 i 个数据点前后对参数 θ 后验分布影响的一种很好的度量。经过简单的计算(6)式变为：

$$K(P,P_{-i})=\log E_\theta\left[\frac{L(\theta|D_{-i})}{L(\theta|D)}|D\right]+E_\theta\left[\log\frac{L(\theta|D)}{L(\theta|D_{-i})}|D\right], \tag{7}$$

其中 $E_\theta[.|D]$ 表示 θ 基于数据 D 的后验期望。由 Gibbs 抽样算法抽取的随机观测序 $\{\theta^{(j)}:j=1,2,\cdots,J\}$，可以得到 K-L 距离 $K(P,P_{-i})$ 的估计为：

$$K(P,P_{-i})=\log\left[\frac{1}{J}\sum_{j=1}^{J}\frac{L(\theta^{(j)}|D_{-i})}{L(\theta^{(j)}|D)}\right]+\frac{1}{J}\sum_{j=1}^{J}\log\left[\frac{L(\theta^{(j)}|D)}{L(\theta^{(j)}|D_{-i})}\right]. \tag{8}$$

对任意的 $i=1,\cdots,n$，当 $K(P,P_{-i})$ 很大时，可以诊断第 i 个个体观测数据点为异常点。

4. 模拟研究

在这部分通过模拟研究来说明前面提出的贝叶斯统计诊断方法的有效性。我们选择均值参数，滑动平均参数和新息参数的真实值分别为 $\beta=(1,-0.5,1)^{\mathrm{T}}$，$\gamma=(-0.3,0.3)^{\mathrm{T}}$ 和 $\lambda=(0,0.5,-0.4)^{\mathrm{T}}$。在联合模型中 $X_{ij},H_{ij}\,(i=1,\cdots,n;\ j=1,\cdots,10)$ 分别是 $p\times1$ 和 $d\times1$ 的协变量向量，其中的元素独立的产生于标准正态分布 $N(0,1)$。选取 $Z_{ijk}=\left(1,t_{ij}-t_{ik},\left(t_{ij}-t_{ik}\right)^2\right)^{\mathrm{T}}$，其中测量时间 t_{ij} 产生于均匀分布 $U[0,2]$。利用这些值，均值 μ_i 和协方差矩阵 $\sum_i$ 可以通过前面描述的改进的 Cholesky 分解来构建。响应变量 Y_i 就可以从多元正态分布 $N(\mu_i,\sum_i)(i=1,\cdots,n)$ 中产生。在这里取样本量 $n=30,60$。

为了调查贝叶斯诊断方法对先验分布的敏感程度，考虑以下有关未知参数 β,γ,λ 的先验分布中超参数值的设置的两种情形：

TypeI：$\beta_0=(1,-0.5,1)^{\mathrm{T}}$，$\sum_\beta=0.25\times\mathrm{I}_3$，$\gamma_0=(-0.3,0.3)^{\mathrm{T}}$，$\sum_\gamma=0.25\times\mathrm{I}_2$，$\lambda_0=(0,0.5,-0.4)^{\mathrm{T}}$，$\sum_\lambda=0.25\times\mathrm{I}_3$。这种设置具有很好的先验信息。

TypeII：$\beta_0=(0,0,0)^{\mathrm{T}}$，$\sum_\beta=1000\times\mathrm{I}_3$，$\gamma_0=(0,0)^{\mathrm{T}}$，$\sum_\gamma=1000\times\mathrm{I}_2$，$\lambda_0=(0,0,0)^{\mathrm{T}}$，$\sum_\lambda=1000\times\mathrm{I}_3$。这些超参数值的设置代表的是没有先验信息的情况。

为了在数据集中产生异常点，分别在第 3 和 19 个个体观测数据点的响应变量 $(y_{ij},i=3,19;\ j=1,2,\cdots,10)$ 都加 1.2 构成人工数据集 D。然后对人工数据集 D 应用本文介绍的影响诊断方法来检测影响观测。其中 MCMC 算法的收敛性可以通过 EPSR 值来检验，并且发现在 3000 次迭代以后 EPSR 值都小于 1.2。因此在计算中丢掉前 3000 次迭代以后再收集 J = 2000 个随机样本来通过(8)式计算

K-L 距离 $K\left(P,P_{-i}\right)$。图 1 和图 2 报道了相应的诊断结果。正如预期的一样，通过图 1 和图 2 很容易就发现，第 3 和 19 个个体观测数据点被诊断为异常点，且诊断方法对先验分布超参数取值的选取不敏感。

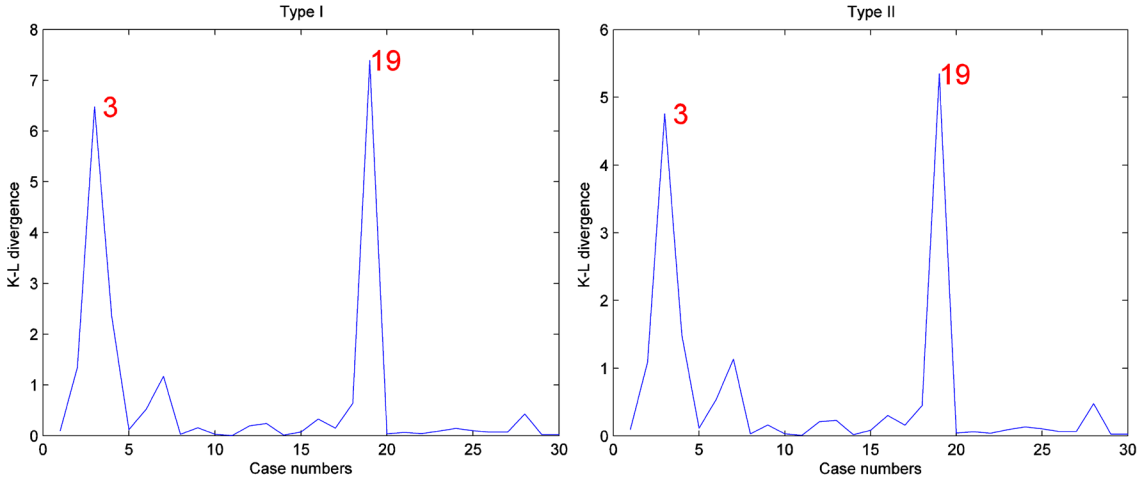

Figure 1. Results based on Bayesian case deletion diagnosis when n = 30
图 1. 当 n = 30 时，贝叶斯数据删除影响诊断的数值结果

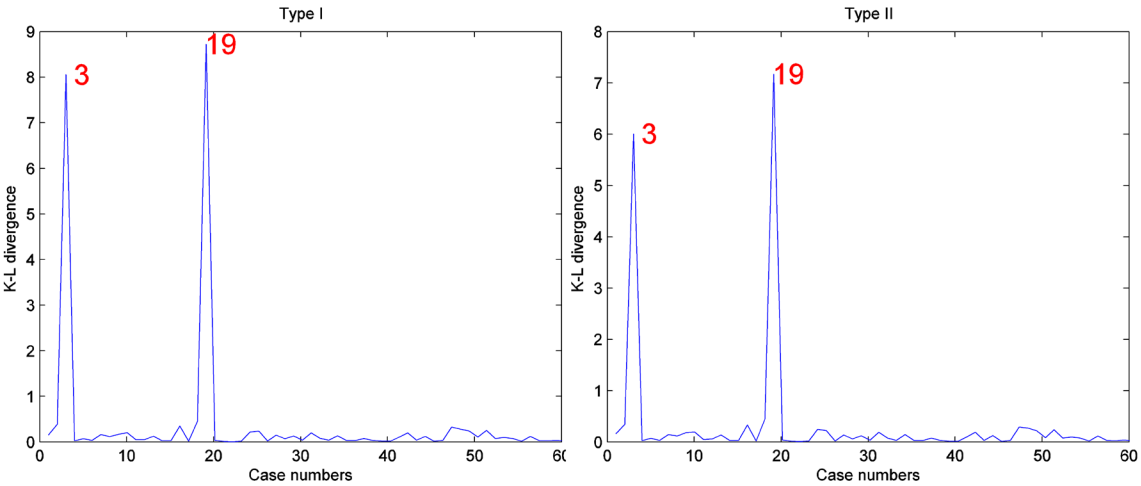

Figure 2. Results based on Bayesian case deletion diagnosis when n = 60
图 2. 当 n = 60 时，贝叶斯数据删除影响诊断的数值结果

5. 实例分析

在这部分把提出的方法应用到牛生长数据。Kenward [16]将牛实验随机的分为 A，B 两组，并分别记录它们的体重。组 A 中的 30 只动物接受 A 治疗，另外 B 组的 30 只动物接受 B 治疗。在 133 天的实验过程中，记录了 11 次动物的体重。采用 Pourahmadi [1] [2]中提出的联合建模模型，运用提出的贝叶斯方法分析 A 组的数据。图 3 显示了 A 组牛数据的曲线和多项式拟合曲线图。从图 3 中可观测在整个实验过程中，相应变量的均值(体重)都在增加，特别是在研究最初的几个星期，增长速度较快。从图 3 进一步可以看出，均值变量与时间变量不是线性的，因此牛群均值生长模型使用关于时间变量的二次或者三次模型更加合理。基于上述分析以及 Pourahmadi [1] [2]的研究，提出采用关于时间变量的三次模型分别对均值 μ_{ij}，移动平均参数 l_{ijk} 和对数新息方差 $\log\left(\sigma_{ij}^2\right)$ 进行建模。具体如下：

$$\begin{cases}\mu_{ij}=\beta_0+\beta_1 t_{ij}+\beta_2 t_{ij}^2+\beta_3 t_{ij}^3\\ \log\left(\sigma_{ij}^2\right)=\lambda_0+\lambda_1 t_{ij}+\lambda_2 t_{ij}^2+\lambda_3 t_{ij}^3\\ l_{ijk}=\gamma_0+\gamma_1\left(t_{ij}-t_{ik}\right)+\gamma_2\left(t_{ij}-t_{ik}\right)^2+\gamma_3\left(t_{ij}-t_{ik}\right)^3\\ i=1,2,\cdots,30;\ j=1,2,\cdots,11\end{cases}$$

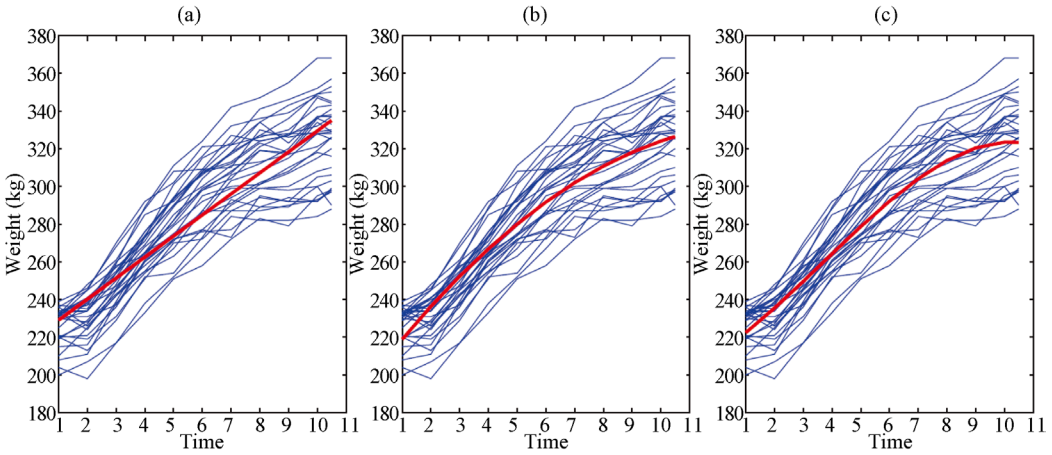

Figure 3. Plot for the cattle data and the thicker line is the polynomial fitted curve: (a) linear polynomial fitted curve; (b) quadratic polynomial fitted curve and (c) cubic polynomial fitted curve
图 3. 牛数据和粗线表示多项式拟合曲线：(a) 线性多项式拟合曲线；(b) 二次多项式拟合曲线；(c)三次多项式拟合曲线

由于贝叶斯统计方法对超参数取值的选取并不敏感，因此在此选取无信息先验信息分布。在 MH 算法中，我们令建议分布中的$\sigma_\gamma^2=\sigma_\lambda^2=1.5$，并且使得接受概率分别为 29.60%和 30.22%。为了测试算法的收敛性，画出了所有未知参数的 EPSR 值的图，具体在图 4 中展示，从图中也能看出 3000 次迭代以后所有参数的 EPSR 值都小于 1.2，这也表示 3000 次迭代以后算法都收敛了。因此我们摒弃所有参数的前 3000 次迭代的迭代值，收集 3000 次迭代以后的 2000 个迭代样本，然后通过(8)估计贝叶斯诊断统计量 K-L 距离$K\left(P,P_{-i}\right)$，图 5 给出了相应的诊断结果。由图 5 知第 7 个和第 25 个个体观测数据点可诊断为异常点。

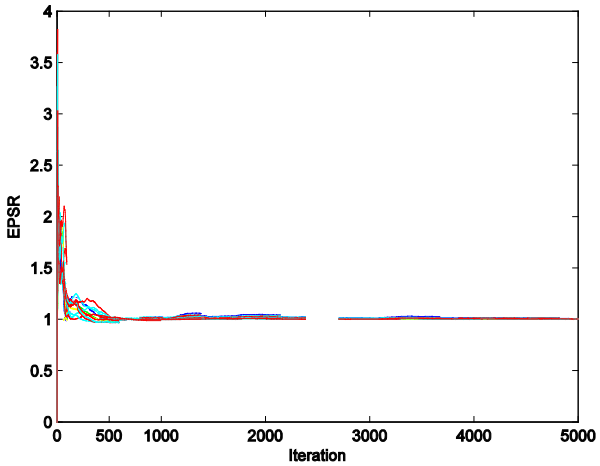

Figure 4. EPSR values of all parameters in the cattle data
图 4. 牛数据中所有未知参数的 EPSR 值

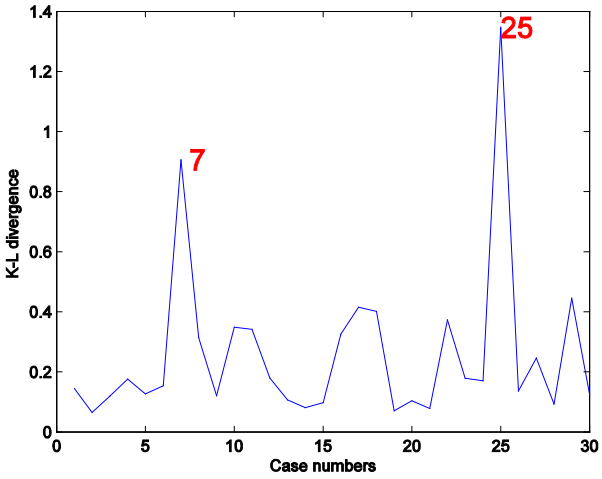

Figure 5. Statistical diagnosis results of the cattle data
图 5. 牛数据统计诊断的数值结果

6. 结论

本文针对纵向数据下均值协方差模型，基于 Gibbs 抽样和 MH 算法相结合的混合算法，以及后验分布之间的 K-L 距离研究贝叶斯数据删除影响的统计诊断方法。模拟研究和实例分析都显示了模型与方法的可行性和有效性。

基金项目

浙江省高校重大人文社科攻关计划项目资助(2018QN037)。

参考文献

[1] Pourahmadi, M. (1999) Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterization. *Biometrika*, **86**, 677-690. https://doi.org/10.1093/biomet/86.3.677

[2] Pourahmadi, M. (2000) Maximum Likelihood Estimation for Generalised Linear Models for Multivariate Normal Covariance Matrix. *Biometrika*, **87**, 425-435. https://doi.org/10.1093/biomet/87.2.425

[3] Pan, J.X. and MacKenzie, G. (2003) On Modelling Mean-Covariance Structures in Longitudinal Studies. *Biometrika*, **90**, 239-244. https://doi.org/10.1093/biomet/90.1.239

[4] Mao, J. and Zhu, Z.Y. (2011) Joint Semiparametric Mean-Covariance Model in Longitudinal Study. *Science China Mathematics*, **54**, 145-164. https://doi.org/10.1007/s11425-010-4078-4

[5] Rothman, A.J., Levina, E. and Zhu, J. (2010) A New Approach to Cholesky-Based Covariance Regularization in High Dimensions. *Biometrika*, **97**, 539-550. https://doi.org/10.1093/biomet/asq022

[6] Zhang, W.P. and Leng, C.L. (2012) A Moving Average Cholesky Factor Model in Covariance Modeling for Longitudinal Data. *Biometrika*, **99**, 141-150. https://doi.org/10.1093/biomet/asr068

[7] Xu, D.K., Zhang, Z.Z. and Wu, L.C. (2013) Joint Variable Selection of Mean-Covariance Model for Longitudinal Data. *Open Journal of Statistics*, **3**, 27-35. https://doi.org/10.4236/ojs.2013.31004

[8] Cook, R.D. (1977) Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15-18. https://doi.org/10.1080/00401706.1977.10489493

[9] Cho, H., Ibrahim, J.G., Sinha, D. and Zhu, H.T. (2009) Bayesian Case Influence Diagnostics for Survival Models. *Biometrics*, **65**, 116-124. https://doi.org/10.1111/j.1541-0420.2008.01037.x

[10] 赵远英, 徐登可, 庞一成. 联合均值与方差模型的 Bayes 分析[J]. 高校应用数学学报, 2018, 33(2): 241-252.

[11] 戴琳, 陶治, 吴刘仓. 联合均值与方差模型的统计诊断[J]. 统计与信息论坛, 2017, 32(1): 14-19.

[12] Tang, N.S. and Duan, X.D. (2012) A Semiparametric Bayesian Approach to Generalized Partial Linear Mixed Models for Longitudinal Data. *Computational Statistics and Data Analysis*, **56**, 4348-4365.

https://doi.org/10.1016/j.csda.2012.03.018

[13] Ye, H.J. and Pan, J.X. (2006) Modelling of Covariance Structures in Generalized Estimating Equations for Longitudinal Data. *Biometrika*, **93**, 927-941. https://doi.org/10.1093/biomet/93.4.927

[14] 韦博成. 参数统计教程[M]. 北京: 高等教育出版社, 2006.

[15] Geyer, C.J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473-511. https://doi.org/10.1214/ss/1177011137

[16] Kenward, M.G. (1987) A Method for Comparing Profiles of Repeated Measurements. *Applied Statistics*, **36**, 296-308. https://doi.org/10.2307/2347788