[文章编号] 1003－4684(2022)01-0115-06

函数型 Logistic 回归模型研究与应用

邓 楠，罗幼喜

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 作为一种新型高维数据，函数型数据重在研究数据的内在本质而不是外在结构，通过非参数方法将数据拟合为函数型数据以捕捉更多信息。针对响应变量为二分类情形，建立贝叶斯框架下的函数型 Logistic 回归模型，引入适当的先验信息并利用 MCMC 算法获得参数的条件后验分布。具体解决流程为：选取由数据驱动的主成分基函数对回归系数函数和回归函数型自变量进行展开，对展开项数进行截断，利用主成分基函数的正交性，将高维数据进行低维表示；再利用 Polya-Gamma 变换，建立易于获得参数后验的 Gibbs 抽样算法，从而得到回归函数展开项系数的后验分布。蒙特卡洛模拟结果显示，该方法具有较好的分类性能。将该方法应用于 Tecator 实际数据，发现其分类效果优于别的方法。

[关键词] 函数型数据；主成分分析；Logistic 回归；polya-gamma

[中图分类号] O212 [文献标识码] A

随着在许多领域对数据质量的要求都越来越高，对数据的分析也从低频数据分析向高频数据分析进行跨越，但在很多情形，我们获得的数据都为离散的数据，无法完全捕捉数据的信息。基于此，Ramsay 于 1982 年提出了函数型数据分析(FDA)[1]。与传统数据分析相比，FDA 具有更多优越性，它通过对数据进行曲线性质的分析进而挖掘出更多重要的信息。在函数型数据分析中，函数型 Logistic 回归是函数型线性回归模型的一个重要应用。它针对响应变量为二分类数据，协变量为函数型数据建立回归模型，利用样本曲线的信息来预测某件事情发生的可能性，通过函数型变量随时间的变化预测二元响应变量的变化。在国外，Ratcliffe 等[2]基于模拟的胎儿心率轨迹构建了函数型 Logistic 回归模型，将函数协变量和回归函数用傅里叶基函数进行展开，对极大似然估计的计算使用改进的 Fisher 评分算法，并将此模型应用到胎儿出生风险预测。Kim 等[3]考虑若函数数据高度混合，则基于整个区域的分类是无效的，因此提出了基于区间的函数型数据分类方法。该方法利用融合的 Lasso 惩罚自动选择函数数据中信息最丰富的片段，同时利用函数逻辑回归对选择的片段进行分类。Denhere[4]考虑了当存在异常曲线时对未加处理的数据进行函数型主成分 Logistic 回归不能得到良好的结果，提出了一种基于稳健主成分的函数型 Logistic 回归模型。Mousavi 等[5]则对许多情况下对函数协变量(作为输入)和二元响应(作为输出)之间的关系感兴趣，由此通过 3 种方法对该模型的参数估计结果进行比较，并判断这些方法正确分类的能力。在国内，王惠文等[6]针对同时包含数值型多元变量和函数型协变量的广义线性回归模型，采用非参数方法得到了参数部分和非参数部分的估计量，并给出了一种重加权算法进行参数求解，解决了含数值型和函数型混合数据类型自变量的回归问题，由此扩展了函数型线性模型的应用范围。孟银凤等[7]针对传统函数 Logistic 模型泛化性能不高的问题，通过求解优化问题提出了线性正则化的函数 Logistic 回归模型。梳理文献发现，尽管已有文献给出了函数型 Logistic 回归模型的不同分析方法和应用实例，但通过贝叶斯方法对其分类性能的研究还较少。Crainiceanu 等[8]曾介绍了在贝叶斯框架下函数型数据的分析方法，使用 WinBugs 对函数型数据进行分析，但未研究 Logistic 回归模型的分类性能，Zhu 等[9]则提出了针对二元响应变量和多元函数型协变量的贝叶斯变量选择模型，并将其应用于宫颈癌诊断，但其对函数型 Logistic 回归模型进行 Probit 变换时，未考虑 Logit 变换，因此本文考虑在贝叶斯框架下对函数型 Logistic 回归模型进行 Logit 变换并

[收稿日期] 2020－12－26

[基金项目] 国家社科基金项目(17BJY210)

[第一作者] 邓 楠(1996－)，女，四川泸州人，湖北工业大学硕士研究生，研究方向为应用统计

[通信作者] 罗幼喜(1979－)，男，湖北红安人，经济学博士，湖北工业大学教授，研究方向为数据挖掘，计量经济建模

对其分类性能进行研究。

1 函数型 Logistic 回归模型

函数型 Logistic 回归模型是针对响应变量为二分类数据、协变量为函数型数据的回归模型[10]。该模型假设协变量 $x(t)(t \in T)$ 为在 L^2 上的平方可积函数，即 $\int_T x^2(t) < \infty$，响应变量 $y \in (0,1)$，函数型 Logistic 回归模型可以表示为[10]：

$$y_i = \pi_i + \varepsilon_i, i = 1,2,\cdots,N \tag{1}$$

其中：

$$\pi_i = P[Y = 1 \mid]x_i(t); t \in T\}] = \frac{\exp\{\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t\}}{1 + \exp\{\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t\}}, \quad i = 1,\cdots,N \tag{2}$$

α 为实数参数，$\beta(t)$ 为参数函数，$\varepsilon_i (i = 1,2,\cdots,N)$ 为 N 个独立且均值为零的随机扰动项。等价地，通过 Logit 变换，式(2)可以表示为：

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \int_T x_i(t)\beta(t)\mathrm{d}t, \quad i = 1,\cdots,N \tag{3}$$

因此，函数型 Logistic 回归模型可以看作为以 Logit 变换为链接函数的广义函数线性模型。由于数据的高维特性，需要对数据进行低维表示，常用的方法是选取基函数对数据进行基展开，常用的基函数有 B 样条基函数、傅里叶基函数、小波基函数等，这里选取主成分基函数对回归系数函数 $\beta(t)$ 和函数回归变量 $x(t)$ 进行基展开。主成分基函数为由数据驱动的基函数，其本质为函数型数据 $x(t)$ 的自协方差算子 A_{Σ_X} 的特征函数，其中 $(A_{\Sigma_X}g)(t) = \int_\varphi \sum_X (s,t)g(s)\mathrm{d}s$，$g$ 为任意平方可积函数。由于自协方差算子是对称且非负定的，因此具有非负的特征值 λ_k 和特征函数 $\theta_k(t), k = 1,2,\cdots$[11]，从而由 Mercer 引理[12]，协方差函数 $\sum_X (s,t)$ 可分解为 $\sum_X (s,t) = \sum_{k=1}^{\infty} \lambda_k \theta_k(s)\theta_k(t)$，其中特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$，$\theta_k$ 为对应特征函数，即主成分基函数。

假设选取 K 个主成分基函数对回归系数函数 $\beta(t)$ 和函数数据 $x(t)$ 进行展开，则

$$\beta(t) = \sum_{k=1}^{K} b_k\theta_k(t), x_i(t) = \sum_{k=1}^{K} c_{ik}\theta_k(t) \tag{4}$$

其中，K 可通过累积方差贡献率对其进行选择，一般选择使累积方差贡献率大于等于 85% 的截断个数。累积方差贡献率 $FVE = \sum_{k=1}^{K}\lambda_i / \sum_{j=1}^{N}\lambda_j$。由主成分基函数的正交性，函数回归模型(3)可表示为：

$$l_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \sum_{k=1}^{K} b_k c_{ik} \tag{5}$$

写成矩阵形式表示为：$l = \alpha_1 + \boldsymbol{Cb}$，其中 $\boldsymbol{b} = (b_1,\cdots,b_K)^{\mathrm{T}}$，$\boldsymbol{1} = (1,1,\cdots,1)^{\mathrm{T}}$，$\boldsymbol{C} = (c_{ik})_{N\times K}$ 为函数主成分得分，其计算方法为：

$$c_{ik} = \int (x_i(t) - \mu(t))\theta_k(t)\mathrm{d}t$$

且满足

$$E(c_{ik}) = 0, E(c_{ik}^2) = \lambda_k, \text{其中 } \mu(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$$

从而在独立条件下，模型的似然函数可以表示为：

$$L(\alpha,\boldsymbol{b}) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \prod_{i=1}^{N} \frac{\exp(\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t)\, y_i}{1 + \exp(\alpha + \int_T x_i(t)\beta(t)\mathrm{d}t)} = \prod_{i=1}^{N} \frac{\exp(\alpha + c_i b)^{y_i}}{1 + \exp(\alpha + c_i b)} \tag{6}$$

2 基于 Polya-Gamma 变换的条件后验分布推导

虽求得函数型 Logisic 回归模型的似然函数，但由于一般先验和模型似然函数的非共轭性较难求得参数后验，因此考虑通过引入 Polson[13] 等提出的 Polya-Gamma 数据增强算法。Polya-Gamma 数据增强算法对于不同模型都求得了更简单且有效的后验分布。该数据增强算法表示为：

记 $\omega \sim PG(\boldsymbol{b},0)$，$\boldsymbol{b} > 0$ 表示服从参数为 $(\boldsymbol{b},0)$ 的 Polya-Gamma 分布，其密度函数

$$f(x \mid b,0) = \frac{2^{b-1}}{\Gamma(b)}\sum_{n=0}^{\infty}(-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \cdot \frac{(2n+b)}{\sqrt{2\pi x^3}}\exp -\frac{(2n+b)^2}{8x}$$

则对于所有 $a \in R$，有下列恒等式成立：

$$\frac{(\exp\psi)^a}{(1+\exp\psi)^b} = 2^{-b}\exp\kappa\psi\int_0^{\infty}\exp -\omega\psi^2/2p(\omega)\mathrm{d}\omega \tag{7}$$

其中，$\kappa = a - b/2$，且 $p(\omega \mid \psi) \sim PG(b,\psi)$。该数据增强算法有效规避了常用先验分布与函数型 Logistic 回归模型似然函数的非共轭性，从而在 Polya-Gamma 变换下，函数型 Logistic 回归模型的似然函数可以改写为：

$$L(\alpha,\boldsymbol{b}) = \prod_{i=1}^{N}\frac{(\exp(\eta_i))^{y_i}}{1+\exp(\eta_i)} = \prod_{i=1}^{N} 2^{-1}\exp\{\kappa_i\eta_i\}\int_0^{\infty}\exp\{-\omega_i\eta_i{}^2/2\}\, p(\omega_i \mid 1,0)\mathrm{d}\omega_i \tag{8}$$

其中 $\eta_i = \boldsymbol{\alpha} + \boldsymbol{c}_i\boldsymbol{b}$，选取先验 $b \sim N(0,\sigma_b^2\Sigma_b)$，$\alpha \sim N(0,\sigma_\alpha^2\Sigma_\alpha)$，$\sigma_b^2 \sim IG(u,v)$，$\sigma_\alpha^2 \sim IG(p,q)$，则 $\boldsymbol{b}$，α，ω，σ_b^2，σ_α^2 的联合后验分布为：

$$P(\boldsymbol{b},\alpha,\sigma_b^2,\sigma_\alpha^2 \mid Y) \propto P(Y \mid \alpha,\boldsymbol{b})P(\boldsymbol{b} \mid \sigma_b^2)$$
$$P(\alpha \mid \sigma_\alpha^2)P(\sigma_b^2)P(\sigma_\alpha^2) \tag{9}$$

则 b 的条件后验可表示为：

$$P(\boldsymbol{b} \mid \cdot) \propto P(Y \mid \alpha,\boldsymbol{b})P(\boldsymbol{b} \mid \sigma_b^2)$$
$$\propto \prod_{i=1}^{N} \frac{(\exp(\eta_i))^{y_i}}{1+\exp(\eta_i)} \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma_b^2}\boldsymbol{b}^{\mathrm{T}}\Sigma_b\boldsymbol{b}\right\}$$
$$\propto \prod_{i=1}^{N} 2^{-1}\exp\{\kappa_i\eta_i\} \int_0^{\infty} \exp\{-\omega_i\eta_i^2/2\}p(\omega_i \mid 1,0) \cdot$$
$$\mathrm{d}\omega_i \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma_b^2}b^{\mathrm{T}}\Sigma_b\boldsymbol{b}\right\} \tag{10}$$

即 b,ω 的联合后验为：

$$P(b,\omega \mid \cdot) \propto \prod_{i=1}^{N} 2^{-1}\exp\{\kappa_i\eta_i\} \cdot$$
$$\int_0^{\infty} \exp\{-\omega_i{\eta_i}^2/2\}p(\omega_i \mid 1,0)\mathrm{d}\omega_i \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma_b^2}b^{\mathrm{T}}\Sigma_b\boldsymbol{b}\right\} \tag{11}$$

从而由 $p(\omega \mid \cdot) \propto \prod_{i=1}^{N} \exp\{\kappa_i(\alpha_i + c_i^{\mathrm{T}}\boldsymbol{b}) - \omega_i(\alpha_i + c_i^{\mathrm{T}}\boldsymbol{b})^2/2\}p(\omega_i \mid 1,0)$ 可得条件后验为：

$$P(\omega_i \mid \cdot) = PG(1,\eta_i) \tag{12}$$

由
$$P(\boldsymbol{b} \mid \cdot) \propto \prod_{i=1}^{N} \exp\begin{Bmatrix}\kappa_i(\alpha_i + c_i\boldsymbol{b}) - \\ \omega_i(\alpha_i + c_i^{\mathrm{T}}\boldsymbol{b})^2/2\end{Bmatrix} \cdot$$
$$p(\omega_i \mid 1,0) \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma_b^2}\boldsymbol{b}^{\mathrm{T}}\Sigma_b\boldsymbol{b}\right\} \propto \prod_{i=1}^{N} \cdot$$
$$\exp\left\{-\frac{\omega_i}{2}\left[(\alpha_i + c_i\boldsymbol{b}) - \frac{\kappa_i}{\omega_i}\right]^2\right\} \prod_{k=1}^{K} \exp\left\{-\frac{1}{2\sigma_b^2}b^{\mathrm{T}}\Sigma_b\boldsymbol{b}\right\}$$
$$\propto \exp\begin{Bmatrix}-\frac{1}{2}(\boldsymbol{b}^{\mathrm{T}}(C^{\mathrm{T}}D_\omega C + \frac{1}{\sigma_b^2}\Sigma_b)\boldsymbol{b} + \\ 2\boldsymbol{b}^{\mathrm{T}}(C^{\mathrm{T}}\boldsymbol{\kappa} - D_\omega C^{\mathrm{T}}\alpha)\end{Bmatrix} \tag{13}$$

则 b 的条件后验为：$b \sim N(\widetilde{B}\tilde{b},\widetilde{B})$，其中 $\widetilde{B} = (C^{\mathrm{T}}D_\omega C + \sigma_b^{-2}\Sigma_b^{-1})^{-1}$，$\tilde{b} = C^{\mathrm{T}}\boldsymbol{\kappa} - C^{\mathrm{T}}D_\omega\alpha$。同理可得：

$$P(\alpha \mid \cdot) \propto \prod_{i=1}^{N} \exp\{\kappa_i(\alpha_i + c_i\boldsymbol{b}) -$$
$$\omega_i(\alpha_i + c_i^{\mathrm{T}}\boldsymbol{b})^2/2\}p(\omega_i \mid 1,0)\exp\left\{-\frac{1}{2\sigma_\alpha^2}\alpha^{\mathrm{T}}\Sigma_\alpha\alpha\right\}$$
$$\propto \prod_{i=1}^{N} \exp\left\{-\frac{\omega_i}{2}\left[(\alpha_i + c_i\boldsymbol{b}) - \frac{\kappa_i}{\omega_i}\right]^2\right\} \exp\left\{-\frac{1}{2\sigma_\alpha^2}\alpha^{\mathrm{T}}\Sigma_\alpha\boldsymbol{b}\right\}$$
$$\propto \exp\left\{\alpha^{\mathrm{T}}(D_\omega + \frac{1}{\sigma_\alpha^2}\Sigma_\alpha)\alpha + 2\alpha^{\mathrm{T}}(\boldsymbol{\kappa} - D_\omega C\tilde{b})\right\} \tag{14}$$

则 α 得条件后验为：

$$\alpha \sim N(\widetilde{A}\tilde{a},\tilde{a})$$

其中 $\widetilde{A} = (D_\omega + \sigma_\alpha^{-2}\Sigma_\alpha^{-1})^{-1}$，$\tilde{a} = \boldsymbol{\kappa} - D_\omega C\tilde{b}$

由 $P(\sigma_b^2 \mid \cdot) \propto P(b \mid \sigma_b^2)P(\sigma_b^2)$，可得：

$$P(\sigma_b^2 \mid \cdot) \sim IG(\frac{k}{2} + u, v + \frac{1}{2}b^{\mathrm{T}}\Sigma_b b) \tag{15}$$

由 $P(\sigma_\alpha^2 \mid \cdot) \propto P(\alpha \mid \sigma_\alpha^2)P(\sigma_\alpha^2)$，可得：

$$P(\sigma_\alpha^2 \mid \cdot) \sim IG(\frac{N}{2} + p, q + \frac{1}{2}\alpha^{\mathrm{T}}\Sigma_\alpha\alpha) \tag{16}$$

则 $b,\alpha,\omega,\sigma_\beta^2,\sigma_\alpha^2$ 的满条件分布为：

1) $\omega_i \mid \text{else} \sim PG(1,\eta_i)$，其中 $\eta_i = \alpha_i + c_i b$；

2) $b \mid \text{else} \sim N(\widetilde{B}\tilde{b},\widetilde{B})$，其中 $\widetilde{B} = (C^{\mathrm{T}}D_\omega C + \sigma_b^{-2}\Sigma_b^{-1})^{-1}$，$\tilde{b} = C^{\mathrm{T}}\boldsymbol{\kappa} - C^{\mathrm{T}}D_\omega\alpha$；

3) $\tilde{\alpha} \mid \text{else} \sim N(\widetilde{A}\tilde{a},\widetilde{A})$，其中 $\widetilde{A} = (D_\omega + \sigma_\alpha^{-2}\Sigma_\alpha^{-1})^{-1}$，$\tilde{a} = \boldsymbol{\kappa} - D_\omega Cb$；

4) $\sigma_b^2 \mid \text{else} \sim IG(s,t)$，其中 $s = \frac{k}{2} + u$，$t = v + \frac{1}{2}b^{\mathrm{T}}\Sigma_b^{-1}b$；

5) $\sigma_\alpha^2 \mid \text{else} \sim IG(m,n)$，其中 $m = \frac{r}{2} + p$，$n = q + \frac{1}{2}\alpha^{\mathrm{T}}\Sigma_\alpha^{-1}\alpha$

根据以上满条件分布，可构建 Gibbs 抽样算法为在给定各参数初始值后。重复以上 1)～5)步骤直至收敛，抽样算法每次重复抽样 10 000 次，舍弃前面 5000 次预烧期，取后面 5000 次作为计算各参数的估计值，在抽样中取先验信息 $\sigma_b^2 = \sigma_\alpha^2 = 1$，$\mu = v = 1$，$p = q = 10$。同时为了保证算法收敛，通过样本路径图、样本密度图和样本自相关函数图来监测样本链的收敛性。

3　数值模拟

3.1　数据生成

首先生成独立同分布的函数型随机变量 x_i，再根据函数型 Logistic 回归模型生成响应变量 y_i。该数据生成方法仿照文献[5]设计，具体数据生成为：

$$x_i(t_j) = \sum_{k=1}^{13} c_{ik}\varphi_k(t_j),$$
$$i = 1,2,\cdots,150, j = 1,2,\cdots,256, t_{ij} \in [0,10] \tag{18}$$

其中：$\{\varphi_k(t)\}_{k=1}^{13}$ 为在区间 $T = [0,10]$ 上内部节点为 9 的自然三次样条基函数，$c_{ik} = \boldsymbol{ZU}$，$\boldsymbol{Z}_{150\times13}$ 和 $\boldsymbol{U}_{13\times13}$ 分别为为服从标准正态分布 $N(0,1)$ 和均匀分布 $[0,1]$ 上的随机变量，并考虑测量误差，即 $W_i(t_j) = X_i(t_j) + \delta_i(t_j)$，其中 $\delta_i(t_j) \sim N(0,\sigma_X^2)$，$\sigma_X$ 分别为 0 或 0.5。第二步生成响应变量 Y_i，生成方式为：

$$LogitPr\{Y_i = 1 \mid X_i\} = \int_0^{10} X_i(t)\beta(t)\mathrm{d}t,$$
$$i = 1,2,\cdots,150 \tag{19}$$

其中 $\beta(t)$ 为区间 $T = [0,10]$ 上的已知函数，考虑 $\beta_1(t) = \sin(t\pi/3)$，$\beta_2(t) = -d(t \mid 2,0.3) + 3d(t \mid 5,0.4) + d(t \mid 7.5,0.5)$，其中 $d(\cdot \mid \mu,\sigma)$ 为服从均值为 μ 方差为 σ 的正态分布，采用主成分基函数进行拟合，模拟结果如图 1 所示。在这里 α 设为 0.5，使用截断点 0.5 作为分割，即

$$\hat{\pi}(x)=\frac{\exp\{\alpha+\int_T\beta(t)x(t)\mathrm{d}t\}}{1+\exp\{\alpha+\int_T\beta(t)x(t)\mathrm{d}t\}}\geqslant 0.5$$

则 $Y=1$，否则 $Y=0$[14]，图 2 为参数函数为 $\beta_1(t)$ 时模拟生成的 150 条曲线中的 40 条样本曲线。

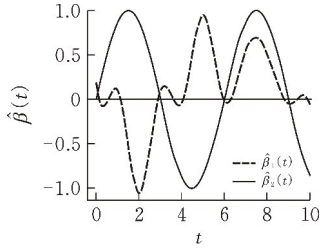

图 1 模拟参数函数曲线

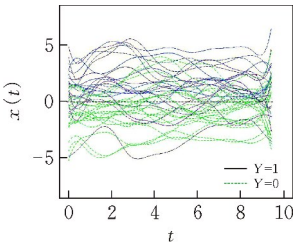

图 2 模拟函数曲线

3.2 模拟结果

为了检验该方法的分类能力，在测量误差分别为 0 和 0.5 的情况下对模型进行验证。由于为二分类问题，根据样本的实际标签与分类器给出的预测标签，可将样本分为 4 种，分别为 TruePositive（正类预测为正类的个数为 TP）、FalseNegative（正类预测为负类的个数为 FN）、FalsePositive（负类预测为正类的个数为 FP）、TrueNegative（负类预测为负类的个数为 TN）。根据上述定义，可对模拟生成的 100 个数据集给出 4 个分类指标，分别是精度（Acc）、准确率（Pre）、召回率（Rec）、F_1 得分（F_1），其计算公式分别为[7]：

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{F1}=\frac{2\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$$

同时将此方法（Bayesian Fuctional Logistic Regression，BFLR）与普通 Logistic 回归（Logistic Regression，LR）、支持向量机（Support Vector Machine，SVM）、决策树（Decision Tree，DT）、条件推断树（Conditonal Inference Tree，CIT）方法进行比较。

通过对比函数 Logistic 回归模型与其他分类方法在模拟数据上的分类性能，发现基于 BPLR 模型的方法对于数据的分类情况明显优于其他方法，在 4 个分类性能指标上都有更高的准确率。样本路径图、样本密度图和样本自相关函数图表明，在经过预烧期后算法已趋于稳定达到收敛，证明该抽样算法在数据分类上的有效性。

表 1 模拟数据分类性能

	Classifier	Acc	Pre	Rec	F1
$\hat{\beta}_1$ $\sigma_X^2=0$	BFLR	0.978	0.996	0.965	0.981
	LR	0.917	0.924	0.929	0.927
	SVM	0.925	0.933	0.935	0.933
	DT	0.913	0.941	0.902	0.920
	CIT	0.898	0.901	0.928	0.911
$\hat{\beta}_1$ $\sigma_X^2=0.5$	BFLR	0.965	0.988	0.951	0.968
	LR	0.893	0.900	0.910	0.905
	SVM	0.905	0.905	0.927	0.915
	DT	0.904	0.915	0.915	0.914
	CIT	0.889	0.898	0.909	0.902
$\hat{\beta}_2$ $\sigma_X^2=0$	BFLR	0.981	0.998	0.972	0.985
	LR	0.942	0.949	0.956	0.957
	SVM	0.935	0.93	0.968	0.948
	DT	0.923	0.937	0.935	0.935
	CIT	0.911	0.927	0.925	0.924
$\hat{\beta}_2$ $\sigma_X^2=0.5$	BFLR	0.969	0.995	0.951	0.978
	LR	0.911	0.924	0.923	0.924
	SVM	0.913	0.925	0.926	0.926
	DT	0.919	0.945	0.915	0.929
	CIT	0.906	0.918	0.923	0.919

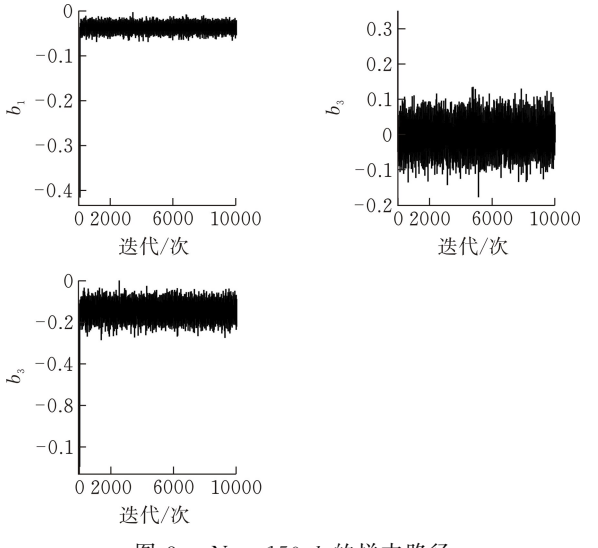

图 3 $N=150$，b 的样本路径

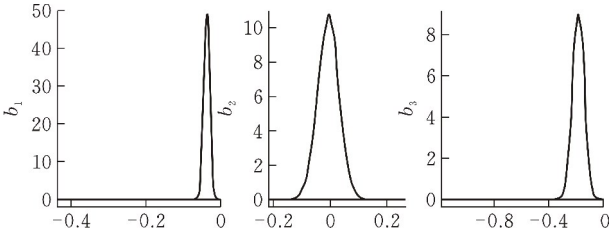

图 4 $N=150$，b 的样本密度图

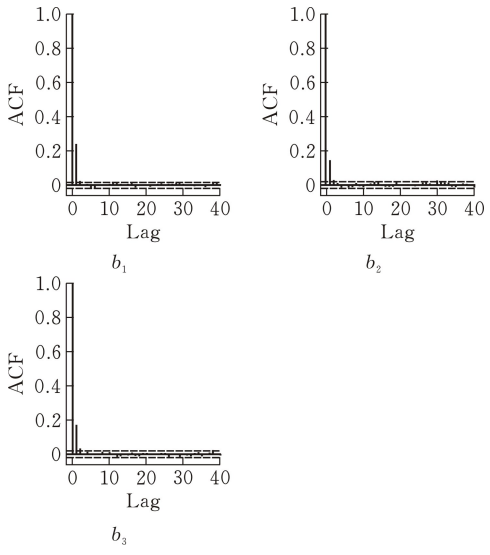

图 5　$N=150$，b 的自相关函数

4　实际数据分析

以 Tecator 数据为例，该数据可在 R 软件包“fda.usc”[15]中进行下载。Tecator 数据集由 215 个碎肉样本对波长为 850～1050 nm 的近红外吸收光谱曲线及其脂肪含量构成，每条吸收光谱曲线观测了 100 个通道，其中有 138 块碎肉样本的脂肪含量 Fat 低于 20%，77 块碎肉样本的脂肪含量 Fat 高于 20%。以此将 Tecator 数据集分为两类，图 6 给出了每类的各 30 条样本曲线。通过函数主成分分析发现 T，ecator 数据集前 3 个主成分已经达到 99% 的累积方差贡献率，因此选取前三个主成分基函数构建函数型 Logistic 回归模型。该模型可以表示为：

$$l_i=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\alpha+\sum_{k=1}^{3}b_kC_{ik},i=1,2,\cdots,215$$

其中初始值 α 设为 0.5，$b_k=(0,0,0)$，$k=1,2,3$，c_{ik} 为前三个主成分得分。

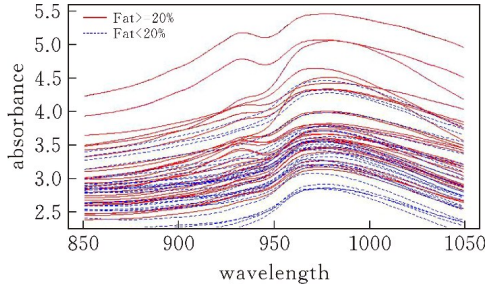

图 6　Tecator 数据集

为检验模型的分类能力，画出模型的 ROC 曲线。结果显示，基于贝叶斯分析的函数型 Logistic 回归模型对 Tecator 数据集的分类效果最优，其 AUC 面积达到了 0.984，说明模型具有较高的分类准确率。与其他方法在 4 个指标上的分类性能相比，尽管 BFLR 方法在准确率上表现不如普通 Logistic 回归、决策树和条件推断树，但在精度、召回率和 F1 得分上都显著优于其他方法，因此总体来说与其他模型相比拥有更好的分类能力。

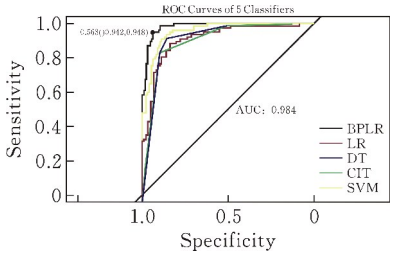

图 7　各分类器 ROC 曲线

表 2　Tecator 数据集分类性能

Classifier	Acc	Pre	Rec	F_1
BFLR	0.912	0.809	0.987	0.889
LR	0.856	0.829	0.753	0.789
SVM	0.879	0.859	0.792	0.824
DT	0.879	0.787	0.909	0.843
CIT	0.874	0.821	0.831	0.826

5　结束语

本文面向函数型数据的二分类问题，提出一种基于 Logit 变换的函数型 Logistic 回归模型，并通过模拟数据和实际数据分析验证了其分类能力。与其他模型的分类性能相比，在该模型上的分类结果均优，但不足是本文考虑的是单变量函数型回归变量的情形，针对多元函数型回归变量以及包括普通数据的函数型 Logistic 回归模型可为后续研究。

[参　考　文　献]

[1] RAMSAY J O, SILVERMAN B W. Functional data analysis[M]. New York: Springer-Verlag, 2005.

[2] RATCLIFFE S J, HELLER G Z, Leader L R. Functional data analysis with application to periodically stimulated foetal heart rate data. Ⅱ: functional logistic regression[J]. Statistics in medicine, 2002, 21(8): 1115-1127.

[3] KIM H, KIM H. Functional logistic regression with fused lasso penalty[J]. Journal of Statistical Computation and Simulation, 2018, 88(15): 2982-2999.

[4] DENHERE M, BILLOR N. Robust principal component functional logistic regression[J]. Communications in Statistics-Simulation and Computation, 2016, 45(1): 264-281.

[5] MOUSAVI S N, SøRENSEN H. Functional logistic regression: a comparison of three methods[J]. Journal of Statistical Computation and Simulation, 2018, 88(2): 250-268.

[6] 王惠文，黄乐乐，王思洋.基于函数型数据的广义线性

回归模型[J].北京航空航天大学学报,2016,42(1):8-12.

[7] 孟银凤,梁吉业.线性正则化函数 Logistic 模型[J].计算机研究与发展,2020,57(8):1617-1626.

[8] CRAINICEANU C M, Goldsmith A J. Bayesian functional data analysis using WinBUGS[J]. Journal of statistical software, 2010, 32(11):1-33.

[9] ZHU H, VANNUCCI M, COX D D. Functional data classification in cervical pre-cancer diagnosis—A bayesian variable selection model[C]. Proc. Jt Statist. Meet, 2007.

[10] ESCABIAS M, AGUILERA A M, VALDERRAMA M J. Principal component estimation of functional logistic regression: discussion of two different approaches[J]. Journal of Nonparametric Statistics, 2004, 16(3-4): 365-384.

[11] 丁辉,许文超,朱汉兵,等.函数型数据回归分析综述[J].应用概率统计,2018,34(6):630-654.

[12] HSING T, EUBANK R. Theoretical foundations of functional data analysis, with an introduction to linear operators[M].Hoboken:John Wiley & Sons, 2015.

[13] POLSON N G, SCOTT J G, WINDLE J. Bayesian inference for logistic models using pólya - gamma latent variables[J]. Journal of the American statistical Association, 2013, 108(504): 1339-1349.

[14] JAMES G M. Generalized linear models with functional predictors[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2002, 64(3): 411-432.

[15] FEBRERO-BANDE M, OVIEDO DE LA FUENTE M. Statistical computing in functional data analysis: the R package fda.usc[J]. J Stat Softw,2012,51:1-28.

Research and Application of Functional Logistic Regression Model

DENG Nan, LUO Youxi

(*School of Sciences, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: As a new kind of high-dimensional data, functional data focuses on the intrinsic nature of the data rather than the external structure. More information can be captured by fitting the data into functional data through non-parametric methods. In the case that response variables are binary classified, this paper considers establishing a functional Logistic regression model under the Bayesian framework, and uses the MCMC algorithm to obtain the conditional posterior distribution of parameters by introducing appropriate prior information. The concrete solution process is as follows: firstly, the regression coefficient function and regression function type independent variable are expanded by selecting data-driven principal component basis function, and the number of expanded items is truncated. The high-dimensional data are represented in low dimension by utilizing the orthogonality of principal component basis function. Then Poyla-Gamma transformation is used to establish the Gibbs sampling algorithm that is easy to obtain parameter posterior. The posterior distribution of regression function expansion term coefficient is obtained. Monte Carlo simulation results show that this method has good classification performance. Finally, this paper applies the method to the actual data of Tecator and finds that its classification effect is better than other methods.

Keywords: functional data; principal component analysis; logistic regression; polya-gamma

[责任编校:张 众]