The Bayesian Lasso

潘子瑞 [1)]

$PB20010370$

[1)](School of Mathematics, University of Science and Technology of China

摘 要 本文基于 Park, T., & Casella, G. (2008)[1]的研究，对其提出的 Bayesian Lasso 方法进行了深入分析和代码复现。原论文较为简略，未对许多重要结论进行详细证明。因此，本文对该方法的一些关键结论进行了补充证明，并通过代码复现了原论文的一些主要结论。通过本文的研究，读者可以更全面地理解 Bayesian Lasso 方法的原理和实现细节，为相关领域的研究提供更深入的参考。

关键词 Empirical Bayes; Gibbs sampler; Hierarchical model; Inverse Gaussian; Linear regression; Penalized regression; Scale mixture of normals

1 引言

Tibshirani(1996)[2]引入了 LASSO 模型，该模型以以下形式为人熟知：

$$\min_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{1}$$

其中，β 是待确定的系数向量，y_i 是观测到的响应变量 (因为截距项均不影响 Lasso 和 Bayesian Lasso 的推导，所以本文中考虑的是消除截距项后的 $\mathbf{y}$)，X_i 是相应的预测变量矩阵，λ 是控制稀疏性的正则化参数。在 Tibshirani（1996）第五节中，Tibshirani 提到当回归参数具有独立且相同的拉普拉斯（即双指数）先验时，Lasso 估计可以被解释为最大后验估计 ($Maximum\ A\ Posteriori\ Estimation$)。

在 Lasso 估计器被提出后，学界已经提出了多种对于 Lasso 标准差估计方法, 如 Fan &Li(2001)[3]在似然设置中引入了夹心估计法 ($sandwich\ formula$)，作为估计协方差的一种途径。在此基础上，Zou(2006)[4]推导了自适应 Lasso 的夹心估计法。然而，所有上述的近似协方差矩阵都表现出对于估计值中 $\beta_j = 0$ 的预测变量给出一个估计的方差为 0 的特性，这导致 Lasso 估计无法给出所有变量的置信区间。

Park, T., & Casella, G. (2008) 给出了贝叶斯 Lasso 的详细格式，并提出了一个方便的 Gibbs 采样方法来估计系数矩阵 β，对比了三种方法对系数估计的路径：LASSO、贝叶斯 Lasso 和岭回归。此外，Park, T., & Casella, G. (2008) 介绍了选择最优参数 λ 的方法。

在本文中，我们补充了 Park, T., & Casella, G. (2008) 中提出的贝叶斯 Lasso 模型的推导过程，然后推导了 Gibbs 采样所需的各参数的满条件分布 ($full\ condition\ distribution$)，针对 Park, T., & Casella, G.

(2008) 中一些重要结论做了实验复现。论文的其余部分组织如下。在下一节中，我们将讨论贝叶斯 Lasso 的原理和推导。在第 3、4 节中，我们考虑如何从贝叶斯 Lasso 中进行抽样，推导出层次模型和 Gibbs 采样器。在第 5 节中，我们介绍结合 EM 算法选择适当参数的方法。第 6 节介绍实验采取的数据集并展示实验结果。最后，在最后一节中，我们做了简单总结并进行深入讨论。

2 贝叶斯 Lasso

在本节中，我们详细阐述了如何将 Lasso 模型转化为贝叶斯后验估计模型：

我们考虑最常见的线性回归模型：

$$Y = X\beta + E \tag{2}$$

其中，Y 是响应变量，X 为设计矩阵，β 是回归系数，而 E 表示噪音。我们假设噪音项 E 为 $\{e_1, e_2, \ldots, e_n\}$，且各 e_i 之间相互独立，服从 $N(0, \sigma^2)$ 分布。

在贝叶斯方法中，我们引入系数 β 的先验分布 $\pi(\beta)$，并假设 $\mathbf{y}|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n)$。因此，$\beta$ 的后验分布可以写为：

$$\pi(\beta|\mathbf{y}, X, \sigma^2) \propto f(y|\beta, \sigma^2)\pi(\beta) = e^{-\frac{||\mathbf{y}-X\beta||_2^2}{2\sigma^2}}\pi(\beta)$$

其最大后验估计问题可以表述为：

$$\max e^{-\frac{||\mathbf{y}-X\beta||_2^2}{2\sigma^2}}\pi(\beta) \tag{3}$$

注意到：

$$\max e^{-\frac{||\mathbf{y}-X\beta||_2^2}{2\sigma^2}}\pi(\beta) \Leftrightarrow \max \ln(3) \Leftrightarrow \min -\ln(3) = \min(||\mathbf{y} - X\beta||_2^2 - \ln(\pi(\beta))) \tag{4}$$

结合 Lasso 的惩罚项 $\lambda^2 \sum_{j=1}^{p} |\beta_j|$(这里写 λ^2 是为了确保前面系数为正值)，在公式 (4) 中，我们选择：

$$-\ln(\pi(\beta)) \propto \lambda^2 \sum_{j=1}^{p} |\beta_j| \tag{5}$$

此时，我们可以推导出 $\pi(\beta) \propto e^{-\lambda^2 \sum_{j=1}^{p} |\beta_j|} \propto \prod_{j=1}^{p} e^{-\lambda^2|\beta_j|}$，这正是 Laplace 分布的核函数，并且由于联合密度等于各自密度的乘积，我们推出 β_i、β_j 之间是相互独立的。当 $\beta_i \sim Laplace(\lambda^2)$ 且相互独立时，在给定 y 条件下的后验分布可以表示为：

$$\pi(\beta, \sigma^2|\mathbf{y}) \propto \pi(\sigma^2)\sigma^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) - \lambda^2 \sum_{j=1}^{p} |\beta_j|\right) \tag{6}$$

对于任意固定的 $\sigma^2 > 0$，最大化 (6) 得到的 β 正是 Lasso 估计对应的 β，而得到最大化 (5) 对应的 β 可以通过我们上面构造的贝叶斯模型进行后验众数估计，这也就是 The Bayesian Lasso 的基本想法。

在 Park, T., & Casella, G. 有关 The Bayesian Lasso 更详细的一篇文章中，他们指出，实际应用过程中，由于先验 (5) 可能会导致后验估计函数 (6) 为一个多峰函数 (见附录 A)，这对我们进行贝叶斯后验估计会产生误导。所以需要将先验函数调整为：

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda^2}{2\sqrt{\sigma^2}} e^{-\frac{\lambda^2|\beta_j|}{\sqrt{\sigma^2}}} \tag{7}$$

除此之外，Park, T., & Casella, G. 还指出：尽管有时最大化后验概率很方便，但这并不是特别自然获取点估计的贝叶斯方式。例如，后验众数在边缘化下不一定被保留。一个完全贝叶斯的分析会建议使用后验的均值或中位数来估计 β。尽管这些估计缺乏 Lasso 模型的变量选择功能，但它们确实产生了类似的收缩系数。完全贝叶斯方法还为估计提供置信区间。

3 层次模型

在确认了贝叶斯方法与 Lasso 估计的等价性后，如何对这一贝叶斯模型进行采样成为了摆在我们面前的问题。在本学期学过的方法中，Metropolis-Hasting 算法、HMC 算法等多种采样方法都可以解决这个问题。然而，通过利用 Laplace 分布的特性，我们可以将其分解为一个正态分布和指数分布的乘积：

$$\underbrace{\frac{a}{2}e^{-a|z|}}_{Laplace分布} = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi s}}e^{-\frac{z^2}{2s}}}_{N(0,s)} \underbrace{\frac{a^2}{2}e^{-\frac{a^2 s}{2}}}_{Exp(\frac{a^2}{2})} ds, \quad a > 0. \tag{8}$$

在原论文中，作者并未给出公式 (8) 的证明，这里我们补充上:

proof of (8) :

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-\frac{z^2}{2s}}\frac{a^2}{2}e^{-\frac{a^2}{2}s}ds = \frac{a^2}{2}\int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{(-\frac{1}{2}(\frac{z^2}{s}+a^2 s))}ds$$

而注意到：

$\frac{1}{\sqrt{2\pi s}}e^{(-\frac{1}{2}(\frac{z^2}{s}+a^2 s))}$为广义逆高斯分布(*Generalized inverseGaussian distribution*)的核函数

于是我们有：

$$\frac{a^2}{2}\int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-\frac{1}{2}\left(\frac{z^2}{s}+a^2 s\right)}ds = \frac{a^2}{2}\times\frac{e^{-a|z|}}{a} = \frac{a}{2}e^{-a|z|} \quad \square$$

通过这一巧妙的拆解，我们得以推导出一个更易于进行抽样的层次模型：

$$\begin{aligned}
\mathbf{y} \mid \mathbf{X}, \beta, \sigma^2 &\sim \mathcal{N}_n\left(\mathbf{X}\beta, \sigma^2 I_n\right), \\
\beta \mid \tau_1^2, \ldots, \tau_p^2, \sigma^2 &\sim \mathcal{N}_p\left(0_p, \sigma^2 \mathbf{D}_\tau\right), \mathbf{D}_\tau = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2), \\
\tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \tau_j^2}{2}} d\tau_j^2, \quad \tau_1^2, \ldots, \tau_p^2 > 0, \\
\sigma^2 &\sim \pi(\sigma^2) d\sigma^2.
\end{aligned} \tag{9}$$

其中 $\tau_1^2, \ldots, \tau_p^2$和$\sigma^2$相互独立, 在对 $\tau_1^2, \ldots, \tau_p^2$ 积分后，β 的条件先验为公式 (7)

先验 (5) 也可以通过以下层次结构获得，如果将 (9) 中 β 的后验替换为：

$$\beta|\tau_1^2, \ldots, \tau_p^2 \sim \mathcal{N}_p(0_p, \mathbf{D}_\tau), \quad \mathbf{D}_\tau = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) \tag{10}$$

在这种情况下，β 是独立于 σ^2 的。在第 3 节中，我们会详细介绍对于模型 (9) 的贝叶斯层次结构的 Gibbs 采样实现，利用了与逆高斯分布相关的共轭性质。虽然采用模型 (10) 的层次结构也可以在 Gibbs 采

样中轻松实现，但正如前面提到的，模型 (10) 可能面临一些由非单峰后验引起的困难，所以我们在实际操作中考虑模型 (8)。

4 Gibbs 采样器

为了对上述推导出的层次模型进行 Gibbs 抽样, 我们需要推导出各参数的满条件分布 (*full conditional distribution*)

在这之前，我们先给出 σ^2 的先验分布;

$$\pi(\sigma^2) = \frac{\gamma^a}{\Gamma(a)\cdot(\sigma^2)^{(a+1)}}\cdot e^{-\frac{\gamma}{\sigma^2}}, \quad \sigma^2 > 0 \quad (a>0, \gamma>0) \tag{11}$$

在确定各参数先验后，我们可以得到整体的似然函数:

$$\begin{aligned} &f(\mathbf{y}|\mu,\beta,\sigma^2)\pi(\sigma^2)\pi(\mu)\prod_{j=1}^{p}\pi(\beta_j|\tau_j^2,\sigma^2)\pi(\tau_j^2) \propto \\ &\frac{1}{(\sigma^2)^{\frac{n-1}{2}}}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)}(\sigma^2)^{-a-1}e^{-\frac{\gamma}{\sigma^2}}\prod_{j=1}^{p}\frac{1}{(\sigma^2\tau_j^2)^{1/2}}e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2}e^{-\frac{\lambda^2\tau_j^2}{2}}. \end{aligned} \tag{12}$$

从整体的似然函数中，我们可以推导出 Gibbs 抽样中各参数的满条件分布:

对β 我们有:

$$\begin{aligned} \pi(\beta|\sim) &\propto exp(-\frac{1}{2\sigma^2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta) - \frac{1}{2\sigma^2}\beta^T\mathbf{D}_\tau^{-1}\beta) \\ &\propto exp(-\frac{1}{2\sigma^2}\beta^T\left(X^TX+\mathbf{D}_\tau^{-1}\right)\beta + 2\mathbf{y}^TX\beta) \end{aligned}$$

而这正与正态分布核函数相对应, 对于一般多元$X\sim N(\mu,\Sigma)$:

$$f(x) \propto exp(\frac{-(X-\mu)^T\Sigma^{-1}(X-\mu)}{2}) \propto exp(\frac{-X^T\Sigma^{-1}X}{2} + \mu\Sigma^{-1}X)$$

从而β 服从均值为$(X^TX+\mathbf{D}_\tau^{-1})^{-1}X^T\mathbf{y}$，方差为$\sigma^2(X^TX+\mathbf{D}_\tau^{-1})^{-1}$ 的正态分布

对σ^2 我们有:

$$\pi(\sigma^2|\sim) \propto (\sigma^2)^{(-\frac{n+p-1}{2}-a-1)}\exp\left(-\frac{1}{\sigma^2}\left((\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)/2 + \frac{\beta^T\mathbf{D}_\tau^{-1}\beta}{2}+\gamma\right)\right)$$

从而σ^2 服从形状参数为$\frac{n+p-1}{2}+a$，尺度参数为$\left(\frac{1}{2}(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta) + \frac{\beta^TD_\tau^{-1}\beta}{2}+\gamma\right)$ 的逆高斯分布

对τ_j，注意到β_j 只和 τ_j 有关，于是:

$$\pi(\tau_j^2) \propto (\tau_j^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\frac{\beta_j^2}{\sigma^2\tau_j^2}+\lambda^2\tau_j^2\right)\right)$$

从而τ_j^2 服从参数为$\frac{1}{2}, \frac{\beta_j^2}{\sigma^2}, \lambda^2$ 的广义逆高斯分布$GIG(\frac{1}{2}, \frac{\beta_j^2}{\sigma^2}, \lambda^2)$

5 调参算法

如果将第 2 节中的层次结构视为参数模型，则参数 λ 具有一个似然函数，于是 Park, T., & Casella, G 提出了对参数 λ 的更新算法, 他们结合 Casella (2001) 提出的与 Gibbs 抽样配合的蒙特卡洛 EM 算法，推导出了对贝叶斯 Lasso 的迭代公式，步骤如下：

(1) 令 $k = 0$ 并选择初始 $\lambda^{(0)}$。

(2) 使用第 4 节的 Gibbs 采样器生成来自 β、σ^2、$\tau_1, ..., \tau_p$ 的后验分布样本，其中 λ 设为 $\lambda^{(k)}$。

(3) **(E-步骤:)** 通过用前一步 Gibbs 样本的平均值替代涉及 β、σ^2 或 $\tau_1, ..., \tau_p$ 期望值的任何项来近似 λ 的期望“完整数据”对数似然。

(4) **(M-步骤:)** 令 $\lambda^{(k+1)}$ 为最大化前一步的期望对数似然的 λ 的值。

(5) 返回到第二步，并迭代直到达到所需的收敛水平。

基于第 3 节的层次结构和共轭先验 (11) 的“完整数据”对数似然为

$$-\left(\frac{n+p-1}{2}+a+1\right)\ln\left(\sigma^2\right)-\frac{1}{\sigma^2}((\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)/2+\gamma)-$$
$$\frac{1}{2}\sum_{j=1}^{p}\ln\left(\tau_j^2\right)-\frac{1}{2}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma^2\tau_j^2}+p\ln\left(\lambda^2\right)-\frac{\lambda^2}{2}\sum_{j=1}^{p}\tau_j^2+\text{不涉及}\lambda\text{的项}$$

第 k 此迭代中，E 步骤对在参数 λ 取条件期望得到:

$$Q(\lambda|\lambda^{(k)})=p\ln\left(\lambda^2\right)-\frac{\lambda^2}{2}\sum_{j=1}^{p}E_{\lambda^{(k)}}\left[\tau_j^2|\mathbf{y}\right]+\text{不涉及}\lambda\text{的项}$$

最大化此表达式，我们得到 M 步的迭代过程：

$$\lambda^{(k+1)}=\sqrt{\frac{2p}{\sum_{j=1}^{p}E_{\lambda^{(k)}}\left[\tau_j^2|\mathbf{y}\right]}}.$$

结合 Gibbs 采样得到的样本，我们用样本平均值替换条件期望 $E_{\lambda^{(k)}}\left(\tau_j^2|\mathbf{y}\right)$。

6 复现结果

在理论部分的详细介绍之后，本节将呈现我们的论文复现结果，并与原论文结果进行对比。与原文不同的是，我们按照计算参数的顺序展示复现结果。

6.1 数据集介绍

为了便于对比，我们采用了和原文一样的糖尿病数据集 (*Diabetes dataset*)。这一数据集包含 442 个患者的数据，数每个患者包括 10 个生理学特征，这些特征经过均值中心化和标准化处理。包括患者的年龄、性别、体质指数（BMI)、平均血压等。目标变量为一年后疾病进展的定量指标。

6.2 参数设置

首先，我们使用第五节介绍的调参方法来选取最优的 λ。在选择初始值为 1 的情况下，通过蒙特卡洛 EM 算法迭代 50 次后，我们发现 λ 的值迭代到 0.2367438321583834。这个结果与原文得到的 0.237 近似相符。在后续的数据展示中，我们采用 $\lambda = 0.237$ 作为参数。同时，我们设置 $\pi(\sigma^2)$ 中 $a = 0, \gamma = 0$，这与原文的设置相同。

6.3 系数估计

Table 1 Estimates of the linear regression parameters for the diabetes data.

Variable	Bayesian Lasso	Bayesian 置信区间 (95%)	Lasso(n-fold c.v.)	Lasso (L_1 范数相同)	Least Squares
(1) age	-4.73	(-112.02, 103.62)	0.00	0.00	-10.01
(2) sex	-213.57	(-334.42, -94.24)	-193.11	-217.40	-239.82
(3) bmi	521.63	(393.07, 653.82)	521.71	525.44	519.85
(4) map	308.41	(180.26, 436.70)	294.73	309.07	324.38
(5) tc	-172.18	(-579.33, 128.54)	-98.31	-166.72	-792.18
(6) ldl	-1.98	(-274.62, 341.48)	0.00	0.00	476.74
(7) hdl	-152.56	(-381.60, 69.75)	-222.41	-174.87	101.04
(8) tch	92.97	(-129.48, 349.82)	0.00	73.19	177.06
(9) ltg	521.12	(332.11, 732.75)	511.37	525.20	751.27
(10) glu	63.08	(-51.22, 188.75)	52.46	61.50	67.63

Table1 展示了糖尿病数据的贝叶斯 Lasso 估计的边际后验分布的中位数和 95% 可信区间。为了比较，我们还加入了最小二乘和两种 Lasso 估计（包括通过交叉验证选择的估计，以及具有与贝叶斯后验中位数相同 L_1 范数的估计）的结果。在下图中，我们展示了这些结果的直观对比：

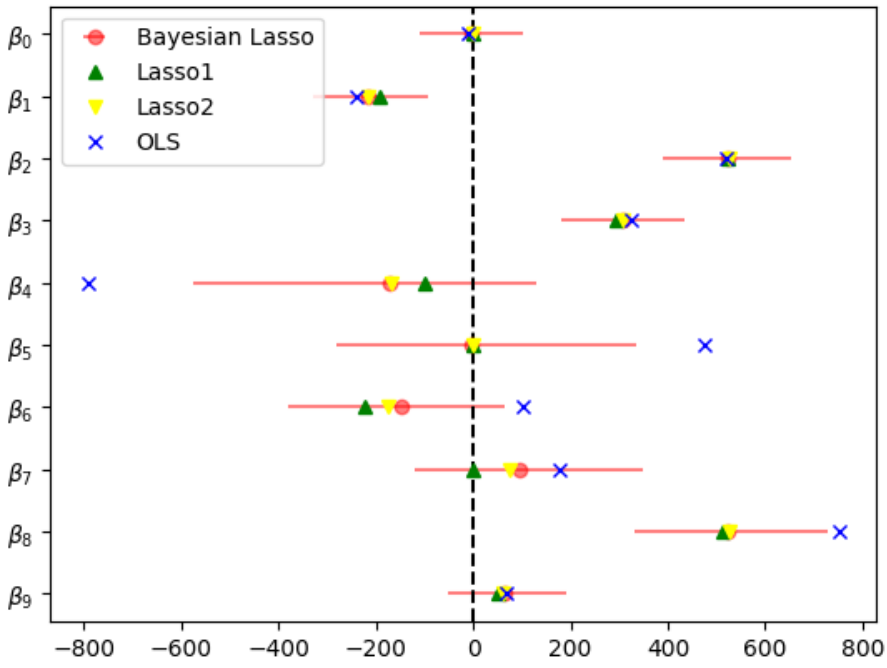

Fig. 1 系数估计图

可以看到，Bayesian Lasso 中位数估计和 Lasso 估计的结果相当接近 (特别是和 Bayesian Lasso 结果具有相同 L_1 范数的 Lasso)，两种 Lasso 估计的结果都落在 Bayesian Lasso 的 95% 置信区间内，而最小二乘估计有 4 个参数不在 95% 置信区间内, 其中一个还是重要参数 $Itg(\beta_8)$，**这与原文结果相同**。

6.4 解路径

与原文一样，我们给出 Lasso、Bayesian Lasso、岭回归的路径图：

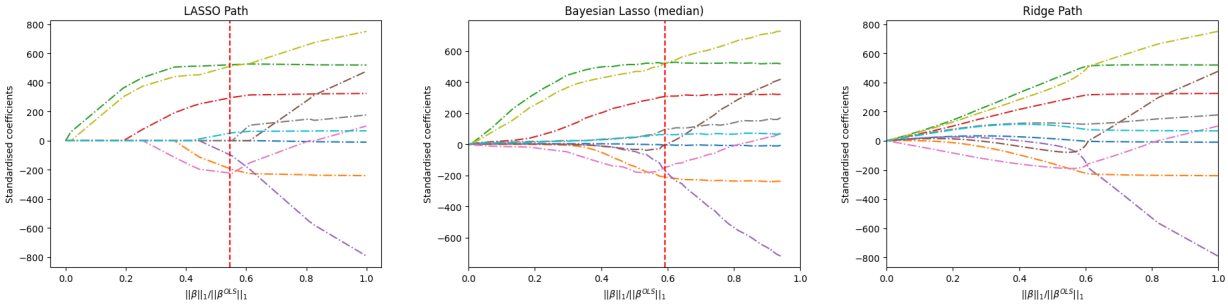

Fig. 2 路径图比较

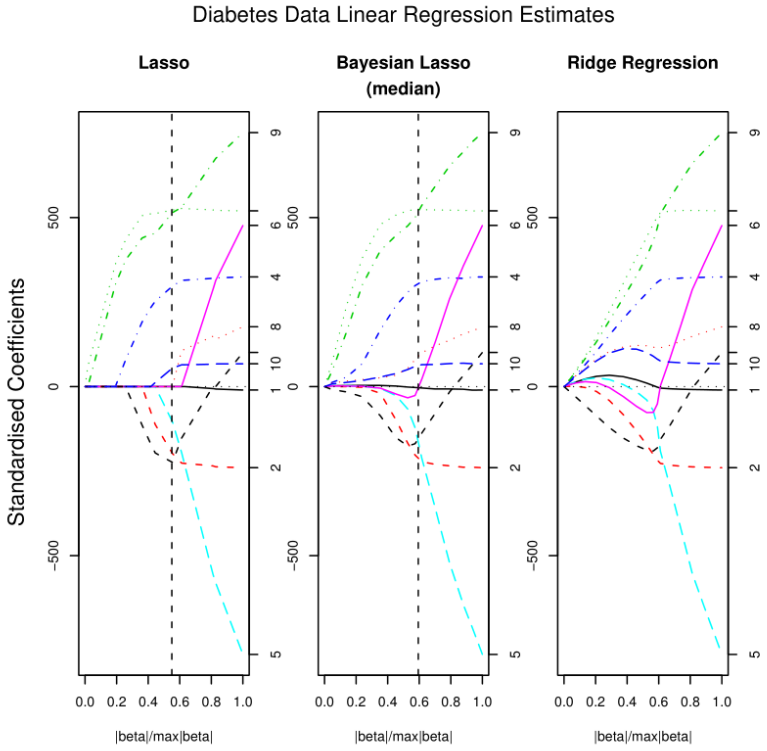

Fig. 3 原文路径图

Lasso 路径图、Bayesian Lasso 路径图中的垂线分别表示通过交叉验证选择出的最优参数对应的解和用第五节算法最大化边际似然得到的参数对应的解，复现结果与原文近似，通过竖线可以看到二者对应的解很接近。

从路径图中，我们可以观察到：贝叶斯 Lasso 估计是 Lasso 和岭回归之间的一种折中：它的解路径是平滑的，类似于岭回归，但在形状上更类似于 Lasso 解路径，特别是当 L_1 范数相对较小时。但与 Lasso 结果显著不同的一点是，贝叶斯 Lasso 估计并不会将系数压缩到 0，即贝叶斯 Lasso 估计不具备变量选择的作用。但其相比于 Lasso, 贝叶斯 Lasso 估计可以给出所有系数的置信区间 (表 1)。

6.5 参数 λ 对收敛速度的影响

原文在介绍蒙特卡洛 EM 选择参数时，指出 λ 的初始值选择对 EM 算法的收敛速度具有显著影响，特别是选择较大的 λ 值可能导致 EM 算法的收敛变得非常缓慢。但 Park, T., & Casella, G. (2008) 中并未给出这个结论的实验展示，这里我们补充上：

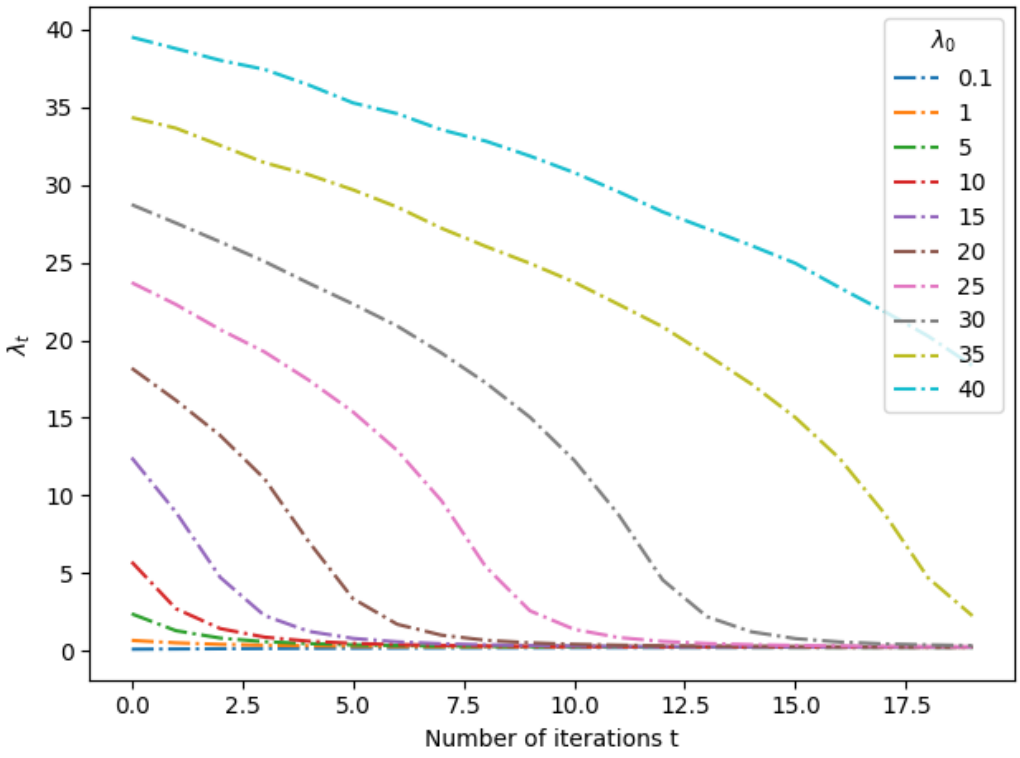

Fig. 4 不同 λ 对收敛速度的影响

可以看到，选择较大的 λ 时。EM 算法的收敛速度变得非常缓慢 (斜率较小)，为了解决这个问题，Park, T., & Casella, G. 指出，我们可以选择初值 $\lambda^{(0)} = p\sqrt{\hat{\sigma}^2_{LS}}/\sum_{j=1}^{p}\left|\hat{\beta}_j^{\mathrm{LS}}\right|$，其中，$\hat{\sigma}^2_{LS}$ 和 $\hat{\beta}_j^{LS}$ 是来自常规最小二乘的估计，这个经验估计往往比最大化的 λ 要小。

7 总结

本文对 Park, T., & Casella, G. (2008) 中一些重要结论来源进行了补充说明，并且对主要结论进行了复现，**复现结果与原论文十分接近**。通过上面的理论分析和实验验证，我们可以看出:

7.1 贝叶斯 Lasso 的优缺点

- **优点**:

  - 相较于 Lasso 的各种优化算法（例如 LARS 等），贝叶斯 Lasso 具有抽样简单、算法编程容易、更易于理解的特点。
  - 得到的是后验分布，可用于进行推断，例如假设检验、置信区间等。

- **缺点**：
  - 尽管理论上推导出的贝叶斯 Lasso 在选择最大后验分布时与 Lasso 等价，但由于抽样得到的 β 是一个分布，难以明确说明该分布中的元素是否为零。因此，贝叶斯 Lasso 很难进行变量选择。

7.2 未来可探究的方向

- 贝叶斯 Lasso 本质上是一种贝叶斯层次模型。除了与 Lasso 方法结合，我们可以探索将其思想与其他惩罚函数（例如岭回归的 L_2 范数）相结合，以期获得更好的性能。

- Lasso 方法的一个显著特性是其在变量选择方面的有效性。深入研究如何改进贝叶斯 Lasso，使其也能够实现变量选择，是一个值得探讨的问题。

- 在面对高维数据时，贝叶斯 Lasso 的适用性是一个有待研究的方向，我们需要考察其在这种情境下的表现和潜在的改进方法。

8 附录 A

如果我们使用非条件 Laplace 先验

$$\pi(\beta) = \prod_{j=1}^{p} \lambda e^{-\lambda|\beta_j|} \tag{A.1}$$

并且给 σ^2 一些与其它参数独立的先验 $\pi(\sigma^2)$，那么 β 和 σ^2 的联合后验分布与下式成正比：

$$\pi(\sigma^2)(\sigma^2)^{-(n-1)/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) - \lambda \sum_{j=1}^{p} |\beta_j|\right) \tag{A.2}$$

其中 y 是观测数据，X 是设计矩阵，β 是参数向量，σ^2 是方差参数，λ 是 Laplace 先验中的超参数。

形如（A.2）的后验分布往往具有不止一个峰值。例如，图 A.1 展示了当 p = 1 且 $\pi(\sigma^2)$ 是尺度不变先验 $1/\sigma^2$ 时，β 和 $ln(\sigma^2)$ 的双峰联合密度的等高线。(即使 $\pi(\sigma^2)$ 是适当的，类似的双峰性也可能发生。) 图 A.1 特定的例子是在取 $p = 1, n = 10, X^TX = 1, X^T\widehat{\mathbf{y}} = 5, \widehat{\mathbf{y}}^T\widehat{\mathbf{y}} = 26$ = 3 的情况下生成的。右下方的峰接近最小二乘解 β = 5,$\sigma^2 = 1/8$，而左上方的峰接近 $\beta = 0$ 和 σ^2 = –26/9 的数值，这接近将 β 设为 0 时的估计。

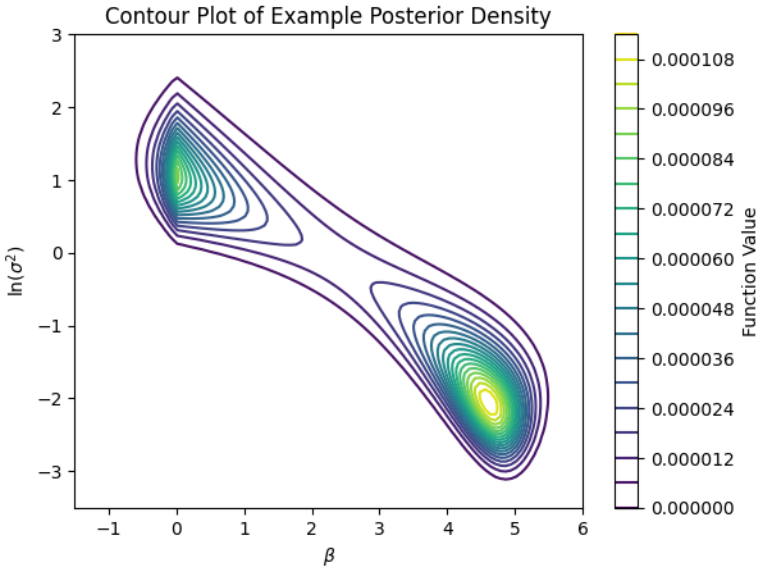

Fig. A.1 多峰分布

参 考 文 献

[1] PARK T, CASELLA G. The bayesian lasso[J]. Journal of the American Statistical Association, 2008, 103(482): 681-686.

[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: 1348-1360.

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