Advances in Applied Mathematics 应用数学进展, 2023, 12(1), 292-307
Published Online January 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.121032

Hans汉斯

基于预测校正算法的Spike and Slab Lasso逻辑回归模型

齐　琪，张　齐*

青岛大学，山东　青岛

收稿日期：2022年12月28日；录用日期：2023年1月24日；发布日期：2023年1月31日

摘　要

尽管Spike and Slab方法广泛应用于贝叶斯变量选择，但其惩罚似然估计的潜力在很大程度上被忽视了。通过在贝叶斯模态中引入惩罚化似然观点，本文提出了新的Spike and Slab Lasso逻辑回归模型，将两个拉普拉斯密度的混合先验置于单个坐标上，可以自适应地收缩系数，即弱收缩重要预测量，强收缩不相关预测量，从而可以得到准确的估计和预测。同时，我们使用了预测－校正算法来求解Spike and Slab Lasso逻辑回归模型，并将该算法扩展到不可分离惩罚的情况。该算法利用凸优化的预测校正算法，沿着整个正则化路径有效的计算解，方便了模型选择，避免了正则化参数值不同时的独立优化。最后，模拟学习和实证结果表明本文所提模型比Lasso逻辑回归模型具有更优的性能。

关键词

Spike and Slab Lasso，逻辑回归，惩罚似然，预测校正算法，贝叶斯先验

The Spike and Slab Lasso Logistic Regression Model Based on Prediction Correction Algorithm

Qi Qi, Qi Zhang*

Qingdao University, Qingdao Shandong

Received: Dec. 28th, 2022; accepted: Jan. 24th, 2023; published: Jan. 31st, 2023

Abstract

Although the Spike and Slab method is widely used in Bayesian variable selection, its potential for

*通讯作者。

文章引用: 齐琪, 张齐. 基于预测校正算法的 Spike and Slab Lasso 逻辑回归模型[J]. 应用数学进展, 2023, 12(1): 292-307. DOI: 10.12677/aam.2023.121032

penalized likelihood estimation is ignored. By introducing the penalized likelihood into the Bayesian case, this paper proposes a new Spike and Slab Lasso logistic regression model, which places the mixed priors of two Laplace densities on a single coordinate, and can adaptively adjust the shrinkage coefficient, that is, the weak shrinkage important predictor and the strong shrinkage irrelevant predictor, so that accurate estimation and prediction can be obtained. At the same time, we use the prediction-correction algorithm for the Spike and Slab Lasso logistic regression model, and extend the algorithm to the case of non-separable penalty. The algorithm uses the prediction correction algorithm of convex optimization to effectively calculate the solution along the entire regularization path, which facilitates model selection and avoids independent optimization when the regularization parameter values are different. Finally, the simulation learning and empirical results show that the proposed model has better performance than lasso logistic regression model.

Keywords

Spike and Slab Lasso, Logistic Regression, Penalized Likelihood, Predictor-Corrector Algorithm, Bayesian Prior

1. 引言

随着现代科技的飞速发展，空前规模的大数据，如基因和蛋白质组学数据，经济学和金融中的面板数据等，在当代的许多应用中遇到。在这些应用中，维数 p 可以与样本容量 n 相当，甚至比 n 大得多，这使得惩罚稀疏模型很适合分析这些数据，自从 Tibshirani 提出 lasso [1]以来，高维环境中变量选择的惩罚回归已经吸引了相当多的学者[2] [3]。lasso 和它的延伸是变量选择中常用的方法，这些方法对系数施加 l_1 惩罚，惩罚的似然可以通过极快的优化算法来解决，例如 LARS 和循环坐标下降算法[4] [5]。然而，lasso 对所有系数使用单一惩罚，因此可以包括许多无关的预测因子，或过度收缩大系数。理想的方法是在大效应时诱导弱收缩，在不相关效应时诱导强收缩。几乎所有的回归都容易受到两个问题的影响：过度拟合和多重共线性。前者，参数优化到统计噪声，而不是适合的兴趣关系；后者，协变量组相互关联，致其难以识别个别影响，这些问题在高维协变量集的分析中尤其突出。一种流行的策略是具有先验分布的贝叶斯方法，通过在回归系数上放置某种先验，将系数缩小到零，这提高了模型的稳定性，并防止过拟合。特别地，一种吸引人的选择是对每个回归系数使用拉普拉斯先验分布，该回归系数对应目标函数中的一个 l_1 惩罚，因此具有拉普拉斯先验分布的贝叶斯回归模型的 MAP 估计等价于 Lasso 回归的估计。在缺乏有力证据的情况下，拉普拉斯先验分布产生了精确的 $\beta_j = 0$ 和对应于重要变量的子集的非零 β_j 的惩罚点估计。通过这种方式，使用 l_1 惩罚自然允许变量选择，其中只有协变量的子集被选择为对结果变量具有实质性预测作用。

也有人提出了变量选择的 Spike and Slab 方法[6]，该方法直接产生于概率的考虑，涉及到线性回归的参数和模型空间上的设计先验层次。Gibbs 抽样[7]用于识别具有高后验概率的有希望的模型，先验的选择通常是棘手的，尽管经验贝叶斯方法可以用来处理这个问题[8]。最早的关于连续 Spike and Slab 先验的理论分析之一是由 Ishwaran 和 Rao 在 2005 年进行的[9]，作者提出并研究了一类连续的双峰先验(两点 student-t 混合)，建立了后验均值对变量选择和多组分类的拟 oracle 误分类性能，并创造了“选择性收缩”一词来表示后验均值的渐近行为。到了 2011 年，Ishwaran 和 Rao 在非正交低维设计的假定下进一步建立

了两点高斯混合下后验均值的 oracle 性质[10]。在另一个发展中，Narisetty 和 He 在 2014 年在协变量数目分散的更一般设计中建立了高斯混合先验下贝叶斯因子的模型选择一致性[11]。后来，到了 2017 年，V. Rockova 和 E. I. George 在单变量线性回归背景下引入了一种新的稀疏正态均值框架[12]，弥补了常用的频率主义策略(Lasso)和常用的贝叶斯策略(Spike and Slab)之间的差距，介绍了拉普拉斯先验和点质量的 Spike and Slab 先验之间的连续统一体——Spike and Slab Lasso (SSL)先验。在高维回归的背景下释放了由连续 SSL 先验产生的惩罚函数的潜力，超越了独立先验的框架。到了 2018 年，V. Rockova 又在另一篇文章中证明了 SSL 全局后验模式在平方误差损失下是(接近)最大最小速率最优的[13]，与 Lasso 相似。同时还证明了整个后验与全局模式保持同步，并集中(近似)于极大极小率，这是一致的单一拉普拉斯先验不具备的性质，利用适当的一类独立积先验(针对已知的稀疏度)和依赖的混合先验(针对未知的稀疏度)，可以获得极小极大速率的运算性。2020 年，Ray Bai 等人在 SSL 的基础上考虑了协变量的组结构[14]，引入了 Spike and Slab Group Lasso (SSGL)用于分组变量线性回归中的贝叶斯估计和变量选择，并进一步将 SSGL 扩展到稀疏广义可加模型。然而，SSL、SSGL 和以往的大多数方法都是基于正态线性模型发展起来的，不能直接应用于其他模型，因此，将混合拉普拉斯先验的高维方法扩展到常规线性模型之外的框架为方法学和应用研究提供了重要的新方向[15] [16]。

本文章节安排如下：在第 2 节中，我们介绍了 Spike and Slab Lasso (SSL)逻辑回归模型，展示了如何将 SSL 方法应用于逻辑模型；在第 3 节中，我们介绍了预测－校正算法，并利用该算法求解 Spike and Slab Lasso 逻辑回归模型；第 4 节将逻辑回归模型下的 SSL 方法扩展到不可分离情况，并相应的将预测－校正算法扩展到不可分离情况；在第 5 节中，我们使用仿真实验来验证所提出的 Spike and Slab Lasso 方法，并与常见的 Lasso 方法进行比较。在第 6 节中，我们利用了南非心脏病数据来说明 SSL 逻辑回归模型的使用，建立了南非西开普省地区缺血性心脏病危险因素的强度，通过报告预测效果的均方误差(MSE)，曲线下面积(AUC)和误判率(misclassification)，证实了该方法的可行性与有效性。最后在第 7 节中对本文进行了总结。

2. Spike and Slab Lasso

2.1. 经典线性 Spike and Slab Lasso 模型

对于以下线性模型

$$Y = X\beta_0 + \varepsilon \tag{1}$$

其中 Y 为 n 维响应变量，$X_{n\times p} = \left(X_1, \cdots, X_p\right)$ 是固定的 p 个潜在预测的回归矩阵，$\beta_0 = \left(\beta_{01}, \cdots, \beta_{0P}\right)'$ 是未知的 p 维回归系数向量，$\varepsilon \sim N_n\left(0, \sigma^2 I_P\right)$ 为噪声向量。一般的 Spike and Slab Lasso 先验是这样的形式：

$$\pi\left(\beta|\gamma\right) = \prod_{i=1}^{P}\left[\gamma_i\varphi_1\left(\beta_i\right) + \left(1-\gamma_i\right)\varphi_0\left(\beta_i\right)\right],\ \gamma \sim \pi\left(\gamma\right) \tag{2}$$

其中 $\gamma = \left(\gamma_1, \cdots, \gamma_P\right)'$，$\gamma_i \in \{0,1\}$ 是二进制向量的中间向量，索引 2^P 个可能的模型。这里，$\varphi_0\left(\beta\right)$ 作为建模无关(零)系数的 Spike 分布，$\varphi_1\left(\beta\right)$ 作为建模大效应的 Slab 分布。对于 Spike and Slab Lasso (SSL)，在上面的特殊变式(2)中引入拉普拉斯密度，即 $\varphi_0\left(\beta\right) = \frac{\lambda_0}{2}\mathrm{e}^{-\lambda_0|\beta|}$，$\varphi_1\left(\beta\right) = \frac{\lambda_1}{2}\mathrm{e}^{-\lambda_1|\beta|}$。则在稀疏正态均值的背景下，这些拉普拉斯分布的两点混合将被称为 Spike and Slab Lasso (SSL)先验。模型空间 $\pi\left(\gamma\right)$ 的灵活性极大地扩大了这些先验的范围，对我们来说，我们集中注意力于可交换的模型空间先验形式：

$$\pi\left(\gamma|\theta\right) = \prod_{j=1}^{P}\theta^{\gamma_j}\left(1-\theta\right)^{1-\gamma_j},\ \theta \sim \pi\left(\theta\right) \tag{3}$$

其中 $\theta = P\left(\gamma_i = 1 \middle| \theta\right)$ 为较大的 β_{0j} 的先验期望分数。在 θ 条件下，SSL 先验(2)可归结为混合物的独立乘积：

$$\pi\left(\beta \middle| \theta\right) = \prod_{i=1}^{P}\left[\theta\varphi_1\left(\beta_i\right) + \left(1-\theta\right)\varphi_0\left(\beta_i\right)\right] \tag{4}$$

其中 $\theta \in (0,1)$ 为混合比例。SSL 先验是将两个拉普拉斯密度的混合先验置于单个坐标 β_j 上。选择一个质点峰值 $\varphi_0\left(\beta_j\right) = \infty I\left(\beta_j = 0\right)$，(当 $\lambda_0 \to \infty$ 时获得)，$\varphi_1\left(\beta_j\right) \propto C > 0$ (当 $\lambda_1 \to 0$ 时获得)，$\log\pi\left(\beta \middle| \theta\right)$ 坍缩为 l_0 惩罚；在另一端，选择 $\varphi_1\left(\beta_j\right) = \varphi_0\left(\beta_j\right)$ 产生参数 $\lambda_1 = \lambda_0$ 的 lasso 惩罚。因此，SSL 先验的一个特征是它们能够在这两种处理中间形成非凹连续体。

2.2. 正则化逻辑回归

当响应变量为二元时，通常采用线性逻辑回归模型，设数据集为 $D = \left\{\left(x_1, y_1\right), \cdots, \left(x_i, y_i\right), \cdots, \left(x_n, y_n\right)\right\}$，其中 $X \in R^{n\times p}$ 是数据矩阵，其中 p 是特征(参数或属性)的数量[17]，对于每个 $x_i \in R^p$，y 是二进制结果向量，结果是 $y_i = 0$ 或 $y_i = 1$，逻辑回归模型通过预测因子的线性函数来表示类条件的期望：

$$E\left(y_i = 1 \middle| x_i, \beta\right) = p_i = \frac{\mathrm{e}^{\beta_0 + x_i'\beta}}{1+\mathrm{e}^{\beta_0 + x_i'\beta}} = \frac{1}{1+\mathrm{e}^{-\left(\beta_0 + x_i'\beta\right)}},\quad i = 1, \cdots, n$$

Logit 变换是响应为 1 的概率的对数，定义为：

$$\eta_i = g\left(p_i\right) = \log\frac{p_i}{1-p_i} = \beta_0 + x_i'\beta$$

在矩阵形式下，logit 转换函数表示为：

$$\eta = X\beta$$

Logit 转换函数很重要，因为它是线性的，故具有线性回归模型的许多特性，在逻辑回归中，函数 $g(\bullet)$ 也称为正则链接函数。现在，假设观测值是独立的，则似然函数为：

$$L(\beta) = \prod_{i=1}^{n}\left(p_i\right)^{y_i}\left(1-p_i\right)^{1-y_i} = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{\beta_0 + x_i'\beta}}{1+\mathrm{e}^{\beta_0 + x_i'\beta}}\right)^{y_i}\left(\frac{1}{1+\mathrm{e}^{\beta_0 + x_i'\beta}}\right)^{1-y_i} \tag{5}$$

因此，对数似然函数为

$$l(\beta) = \sum_{i=1}^{n}\left(y_i\log\left(\frac{\mathrm{e}^{\beta_0 + x_i'\beta}}{1+\mathrm{e}^{\beta_0 + x_i'\beta}}\right) + \left(1-y_i\right)\log\left(\frac{1}{1+\mathrm{e}^{\beta_0 + x_i'\beta}}\right)\right) \tag{6}$$

已经有学者证明了逻辑回归的极大似然估计量满足极大似然的期望性质[18]，但不幸的是，关于 β 最大化 $l(\beta)$ 没有封闭解。因此，逻辑回归的极大似然估计是通过数值优化方法得到的，在本文中，使用了预测－校正算法，该算法在预测步中使用交替方向法得到预测点，然后对得到的部分预测点进行校正，从而保证了算法的收敛性。

2.3. 逻辑回归的可分离 SSL 惩罚

我们的方案的一个关键因素是利用 Spike and Slab Lasso 模态估计(变量选择程序的基础)和广义 Lasso 估计之间的联系，一个重要的第一步是理解可分离的 SSL 惩罚的结构。假设混合比例 θ 是固定的，这简化了 SSL 惩罚的操作，因为它是可根据 θ 条件分离的，即此时 β_j 具有条件独立的先验，这导致了可分离的 SSL 惩罚。

定义 1：给定$\theta\in(0,1)$，可分离 SSL 惩罚被定义为：

$$pen_s(\beta|\theta)=\log\left[\frac{\pi(\beta|\theta)}{\pi(0_p|\theta)}\right]=\sum_{i=1}^{p}\log\left[\frac{\theta\varphi_1(\beta_j)+(1-\theta)\varphi_0(\beta_j)}{\theta\varphi_1(0)+(1-\theta)\varphi_0(0)}\right] \tag{7}$$

为了便于惩罚操作，我们将其居中，使$pen_s(0_p|\theta)=0$。由于β与θ的条件无关，惩罚函数由单次函数建立：

$$p(\beta_j|\theta)=-\lambda_1|\beta_j|+\log\left[\frac{p_\theta^*(0)}{p_\theta^*(\beta_j)}\right] \tag{8}$$

$$pen_s(\beta|\theta)=\sum_{j=1}^{p}p(\beta_j|\theta)=-\lambda_1|\beta|+\sum_{j=1}^{p}\log\left[\frac{p_\theta^*(0)}{p_\theta^*(\beta_j)}\right] \tag{9}$$

其中

$$p_\theta^*(\beta_j)=\frac{\theta\varphi_1(\beta_j)}{\theta\varphi_1(\beta_j)+(1-\theta)\varphi_0(\beta_j)} \tag{10}$$

描述(9)将可分离的 SSL 惩罚写成 Lasso 惩罚和非凹惩罚的自适应和，使其最终是非凹的。与在$|\beta_j|$和λ中都是线性的 Lasso 惩罚不同，SSL 惩罚是$|\beta_j|$和$(\lambda_0,\lambda_1,\theta)$的非线性函数，尽管有明显的差异，但两个惩罚之间具有一定的联系，求导后，这个联系就暴露出来了，导数对应于一个隐偏差项，在估计中起着至关重要的作用，因此我们有以下引理：

引理 1：可分离的 SSL 惩罚满足：

$$\frac{\partial pen_s(\beta|\theta)}{\partial|\beta_j|}\equiv-\lambda_\theta^*(\beta_j) \tag{11}$$

其中，

$$\lambda_\theta^*(\beta_j)=\lambda_1 p_\theta^*(\beta_j)+\lambda_0\left[1-p_\theta^*(\beta_j)\right] \tag{12}$$

另一方面，拉普拉斯先验求导为：$\frac{\partial}{\partial|\beta_j|}\log\varphi(\beta_j|\lambda)=-\lambda$。因此 SSL 偏差项是两个 Lasso 偏差项的凸组合，重要的是，这种组合是有适应性的，因为$p_\theta^*(\beta_j)$依赖于β_j，这是 Spike and Slab 惩罚的独特特征，它对每个系数分别加权 Spike 和 Slab 的贡献。与之形成鲜明对比的是，无论系数大小如何，Lasso 惩罚对每个系数的偏差都是相同的，这通常是收缩和偏差之间冲突的根源，$\lambda_\theta^*(\beta_j)$的额外灵活性极大地缓解了这种冲突。

$\lambda_\theta^*(\beta_j)$是一个自适应的线性组合，由$p_\theta^*(\beta_j)$加权，影响较大的系数$p_\theta^*(\beta_j)$接近 1 并且收缩的更小，这是因为$\lambda_\theta^*(\beta_j)$主要由$\lambda_1$驱动，为了避免过度收缩，$\lambda_1$被设置的很小。影响较小的系数则相反，它们具有较小的包含概率，因此$\lambda_\theta^*(\beta_j)$被较大的惩罚$\lambda_0$所接管，也就是说，当$|\beta_j|$较大时，倾向于收缩少量，当$|\beta_j|$较小时，倾向于收缩大量，这是 SSL 估计器背后选择性收缩特性的一种表现。混合比例$p_\theta^*(\beta_j)$可以看做条件包含概率$p(\gamma_i=1|\beta_i,\theta)$，从而有：

$$p_\theta^*(\beta_j)=p(\gamma_i=1|\beta_i,\theta)=\frac{1}{1+\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\exp\left[-|\beta_i|(\lambda_0-\lambda_1)\right]}$$

3. 预测校正算法

预测校正算法是实现数值延拓的基本策略之一[19]。在许多方法中，预测校正算法通过实用初始条件(参数一个极值的解)显式的求出一系列解，并根据当前解继续求出相邻解，我们详细说明了如何使用预测校正算法跟踪曲线 $H\left(\beta,\lambda_\theta^*\right)=0$ 到我们的问题设置。为了找到 $\beta=\left(\beta_0,\beta'\right)'$，我们的准则使用 λ_θ^* 从一个固定的值开始减少，这等价于最小化以下问题：

$$l\left(\beta,\lambda_\theta^*\right)=\left\{-\sum_{i=1}^{n}\left[y_i\log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right]-pen_s\left(\beta\middle|\theta\right)\right\} \tag{13}$$

假设 β 的分量全都不为 0，且 $l\left(\beta,\lambda_\theta^*\right)$ 关于 β 可导，定义函数 H 为：

$$H\left(\beta,\lambda_\theta^*\right)=\frac{\partial l}{\partial\beta}=-X^{\mathrm{T}}\left(y-p\right)+\lambda_\theta^*\left(\beta\right)Sgn\begin{pmatrix}0\\ \beta\end{pmatrix} \tag{14}$$

其中假设 β 的所有分量都不为零，X 是 $n\times\left(p+1\right)$ 阶矩阵，包括列为 1。即使已经假设了 β 的所有分量都不为零，但 β 的非零分量集随 λ_θ^* 的变化而变化，λ_θ^* 中，$\lambda_1\in\left(0,\infty\right)$，$\lambda_2$ 是很小的，固定的正常数，$\theta\in\left(0,1\right)$ 是给定的。令 $\hat{\beta}$ 表示可分离 SSL 惩罚 $pen_s\left(\beta\middle|\theta\right)$ 下的全局后验模式，即

$$\hat{\beta}\left(\lambda_\theta^*\right)=\underset{\beta\in R^p}{\arg\min}\left\{-\sum_{i=1}^{n}\left[y_i\log p_i+\left(1-y_i\right)\log\left(1-p_i\right)\right]-pen_s\left(\beta\middle|\theta\right)\right\} \tag{15}$$

引理 2：$\hat{\beta}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)'$ 为 $pen_s\left(\beta\middle|\theta\right)$ 下的全局后验模式(15)的 KKT 条件为：

$$X_j'\left(y-\hat{p}\right)=\lambda_\theta^*\left(\hat{\beta}_j\right)sign\left(\hat{\beta}_j\right)\quad\text{for}\quad\hat{\beta}_j\neq0\quad j=1,\cdots,p \tag{16}$$

$$\left|X_j'\left(y-\hat{p}\right)\right|\le\lambda_\theta^*\left(\hat{\beta}_j\right)\quad\text{for}\quad\hat{\beta}_j=0\quad j=1,\cdots,p \tag{17}$$

对于任意 $j=1,\cdots,p$，当 $\hat{\beta}_j=0$ 时，KKT 条件暗示了 $1'\left(y-\hat{p}\right)=0$，即

$$\hat{p}=\bar{y}1=g^{-1}\left(\hat{\beta}_0\right)1 \tag{18}$$

引理 3：当 λ_θ^* 超过某个阈值时，截距是唯一的非零系数，即

$$\tilde{\beta}_0=g\left(\bar{y}\right)，且\ H\left(\left(\hat{\beta}_0,0,\cdots,0\right)',\lambda_\theta^*\right)=0\ 对于\ \lambda_\theta^*>\max_{j\in\{1,\cdots,p\}}\left|x_j'\left(y-\bar{y}1\right)\right|。$$

我们把 λ_θ^* 的这个阈值表示为 $\lambda_{\theta,\max}^*$，从变量 $j_0=\arg\max_j\left|x_j'\left(y-\bar{y}1\right)\right|$ 开始，随着 λ_θ^* 的进一步减少，其他变量加入活动集。从 $\lambda_{\theta,\max}^*$ 开始减少 λ_θ^*，我们在预测步和校正步之间交替，第 k 次迭代的步骤如下：

1) 步长：确定下降量 λ_θ^*，给定 $\lambda_{\theta,k}^*$，我们近似第二大 λ_θ^*，其中活动集发生变化，命名为 $\lambda_{\theta,k+1}^*$。
2) 预测步：用 λ_θ^* 的下降量线性近似 β 的相应改变，称为 $\hat{\beta}^{k+}$。
3) 校正步：找到 β 中和 $\lambda_{\theta,k+1}^*$ 配对的精确解(即 $\beta\left(\lambda_{\theta,k+1}^*\right)$)，用作为初始值，称为 $\hat{\beta}^{k+1}$。
4) 活动集：测试以查看当前活动集是否必须是改进的，如果是，用更新后的活动集重复校正步。

3.1. 预测步

定义 $f\left(\lambda_\theta^*\right)=H\left(\beta\left(\lambda_\theta^*\right),\lambda_\theta^*\right)$，则在第 k 个预测步中，在生成的当前活动集的范围内，$f\left(\lambda_\theta^*\right)$ 对于所有的 λ_θ^* 都是 0，通过对 $f\left(\lambda_\theta^*\right)$ 关于 λ_θ^* 求导有 $f'\left(\lambda_\theta^*\right)=\frac{\partial H}{\partial\lambda_\theta^*}+\frac{\partial H}{\partial\beta}\frac{\partial\beta}{\partial\lambda_\theta^*}=0$，从上式我们计算 $\frac{\partial\beta}{\partial\lambda_\theta^*}$，有

$$\frac{\partial \beta}{\partial \lambda_\theta^*}=-\left(\frac{\partial H}{\partial \beta}\right)^{-1}\frac{\partial H}{\partial \lambda_\theta^*}=-\left(X_A' W_k X_A\right)^{-1} Sgn\begin{pmatrix}0\\ \hat{\beta}^k\end{pmatrix} \tag{19}$$

其中 W 是 n 阶对角阵，元素为 $p_i(1-p_i)$, $i=1,\cdots,n$， W_k 和 X_A 表示当前活动集中 X 的列。从而在第 k 个预测步中， $\beta\left(\lambda_{\theta,k+1}^*\right)$ 被近似通过

$$\begin{aligned}\hat{\beta}^{k+} &= \hat{\beta}^k+\left(\lambda_{\theta,k+1}^*-\lambda_{\theta,k}^*\right)\frac{\partial \beta}{\partial \lambda_\theta^*}\\ &= \hat{\beta}^k-\left(\lambda_{\theta,k+1}^*-\lambda_{\theta,k}^*\right)\left(X_A' W_k X_A\right)^{-1} Sgn\begin{pmatrix}0\\ \hat{\beta}^k\end{pmatrix}\end{aligned} \tag{20}$$

上式中 β 仅由当前非零系数组成，这种线性化相当于对对数似然进行二次逼近，并通过加权 Lasso 步骤来扩展当前解 $\hat{\beta}^k$。下列定理说明，预测步近似可以通过使 $\left(\lambda_{\theta,k}^*-\lambda_{\theta,k+1}^*\right)$ 小来任意接近真实解。

定理 1：令 $h_{\theta,k}^*=\lambda_{\theta,k}^*-\lambda_{\theta,k+1}^*$，假设 $h_{\theta,k}^*$ 足够小，活动集在 $\lambda_\theta^*=\lambda_{\theta,k}^*$ 和 $\lambda_\theta^*=\lambda_{\theta,k+1}^*$ 是相同的，则近似解 $\hat{\beta}^{k+}$ 和真实解 $\hat{\beta}^{k+1}$ 有 $O\left(h_{\theta,k}^2\right)$ 的区别。

证明：因为 $\frac{\partial \beta}{\partial \lambda_\theta^*}=-\left(X_A' W_k X_A\right)^{-1} Sgn\begin{pmatrix}0\\ \hat{\beta}^k\end{pmatrix}$ 是连续可微的对于 $\lambda_\theta^*\in\left(\lambda_{\theta,k+1}^*,\lambda_{\theta,k}^*\right]$

$$\hat{\beta}^{k+1}=\hat{\beta}^k-h_{\theta,k}^*\left.\frac{\partial \beta}{\partial \lambda_\theta^*}\right|_{\lambda_{\theta,k}^*}+O\left(h_{\theta,k}^2\right)=\hat{\beta}^{k+}+O\left(h_{\theta,k}^2\right)$$

3.2. 校正步

在第 k 次校正步中，我们使用先前的预测步中的近似作为初始值来计算精确解 $\beta\left(\lambda_{\theta,k+1}^*\right)$。

定理 2：如果将在 $\lambda_{\theta,k}^*$ 和 $\lambda_{\theta,k+1}^*=\lambda_{\theta,k}^*-h_{\theta,k}^*$ 处的解记为 $\hat{\beta}^k$ 和 $\hat{\beta}^{k+1}$，他们被连接，使得对某个 $\alpha\in[0,1]$，在 $\lambda_\theta^*=\lambda_{\theta,k}^*-\alpha h_{\theta,k}^*$ 处，我们的估计为：

$$\hat{\beta}\left(\lambda_\theta^*-\alpha h_{\theta,k}^*\right)=\hat{\beta}^k+\alpha\left(\hat{\beta}^{k+1}-\hat{\beta}^k\right) \tag{21}$$

则 $\hat{\beta}\left(\lambda_\theta^*-\alpha h_{\theta,k}^*\right)$ 和真实解 $\beta\left(\lambda_\theta^*-\alpha h_{\theta,k}^*\right)$ 有 $O\left(h_{\theta,k}^*\right)$ 的区别。

证明：因为 $\frac{\partial \beta}{\partial \lambda_\theta^*}$ 是连续可微的对于 $\lambda_\theta^*\in\left(\lambda_{\theta,k+1}^*,\lambda_{\theta,k}^*\right]$，则下式成立：

$$\hat{\beta}\left(\lambda_\theta^*-\alpha h_{\theta,k}^*\right)=\hat{\beta}^k-\alpha h_{\theta,k}^*\left(\frac{\hat{\beta}^{k+1}-\hat{\beta}^k}{-h_{\theta,k}^*}\right)=\hat{\beta}^k-\alpha h_{\theta,k}^*\left.\frac{\partial \beta}{\partial \lambda_\theta^*}\right|_{\lambda_{\theta,\ k}^*}+O\left(h_{\theta,k}^2\right)$$

类似的，在 $\lambda_\theta^*=\lambda_{\theta,k}^*-\alpha h_{\theta,k}^*$ 处的真实解为：

$$\beta\left(\lambda_\theta^*-\alpha h_{\theta,k}^*\right)=\hat{\beta}^k-\alpha h_{\theta,k}^*\left.\frac{\partial \beta}{\partial \lambda_\theta^*}\right|_{\lambda_{\theta,k}^*}+O\left(h_{\theta,k}^2\right)$$

3.3. 活动集

活动集 A 从引理 3 中的截距开始，在每个校正步之后，应用下面的检查程序去检查 A 是否应该被扩充：

$$\left|X_j^{\mathrm{T}}(y-p)\right|>\lambda_\theta^*(\beta)，\text{对于任意 } j\in A\Rightarrow A\leftarrow A\cup\{j\}$$

我们用改进后的活动集重复校正步直到活动集不再扩充。然后我们从活动集中删除带有零系数的变

量，即$\left|\hat{\beta}_j\right|=0$对于任意$j\in A\Rightarrow A\leftarrow A\backslash\{j\}$。

KKT最优条件(16)~(17)暗示了$\hat{\beta}_j\neq 0\Rightarrow\left|X_j'\left(Y-\hat{P}\right)\right|=\lambda_\theta^*\left(\hat{\beta}_j\right)$，结合(16)式和(18)式，我们有以下引理：

引理4：令$\hat{p}$是y的来自一个校正步的估计，$\hat{c}$表示因子和当前残差相关权重：

$$\hat{c}=X'\left(y-\hat{p}\right) \tag{22}$$

A中各因子(除截距外)的绝对相关系数为λ_θ^*，而A^c中各因子的值小于λ_θ^*。也就是说，在每个校正步之后，如果对于任意$l\in A^c$有$\left|\hat{c}_l\right|>\lambda_\theta^*$，我们通过增加$x_l$来扩大活动集。直到活动集不再进一步扩大，校正步停止重复。如果对于任意$l\in A$，$\hat{\beta}_l=0$，则我们从活动集中估计x_l。

3.4. 步长

如果$\lambda_\theta^*=0$，算法停止，如果$\lambda_\theta^*>0$，我们在λ_θ^*中近似最小的减量，活动集将会被改进。随着λ_θ^*减少h_θ^*，变化量表示为a，在当前校正步的近似变化量如下：

$$\begin{aligned}c\left(h_\theta^*\right)&=\hat{c}-h_\theta^*(a)\\&=\hat{c}-h_\theta^*X'\hat{W}X_A\left(X_A'\hat{W}X_A\right)^{-1}Sgn\begin{pmatrix}0\\\hat{\beta}\end{pmatrix}\end{aligned} \tag{23}$$

其中$h_\theta^*>0$是λ_θ^*中给定的减少量，对于A中的因子，a的值为$Sgn\begin{pmatrix}0\\\hat{\beta}\end{pmatrix}$的值，求$A^c$的任意因子与$A$中任意因子的绝对相关系数相同的$h_\theta^*$，我们解以下方程：

$$\left|c_j\left(h_\theta^*\right)\right|=\left|\hat{c}_j-h_\theta^*a_j\right|=\lambda_\theta^*-h_\theta^*,\ \forall j\in A^c \tag{24}$$

方程给出了λ_θ^*中的步长估计为

$$h_\theta^*=\min_{j\in A^c}{}^{+}\left\{\frac{\lambda_\theta^*-\hat{c}_j}{1-a_j},\frac{\lambda_\theta^*+\hat{c}_j}{1+a_j}\right\} \tag{25}$$

另外，检查活动集中是否有变量在λ_θ^*减少h_θ^*之前达到0，我们解出方程：

$$\beta_j\left(\tilde{h}_\theta^*\right)=\hat{\beta}_j+\tilde{h}_\theta^*\left(X_A'\hat{W}X_A\right)^{-1}Sgn\begin{pmatrix}0\\\hat{\beta}\end{pmatrix}=0,\ \forall j\in A \tag{26}$$

若$\forall j\in A$，$0<\tilde{h}_\theta^*<h_\theta^*$，我们期望相应的变量在任何其他变量加入活动集之前被消去，因此$\tilde{h}_\theta^*$而不是h_θ^*作为下一个步长。

4. 不可分离情况

如前所述，惩罚由参数$(\lambda_0,\lambda_1,\theta)$的三元组索引，它们串联工作以产生理想的性质。在整篇论文中，我们假设λ_1被设为一个很小的值，因此不需要调谐，两个参数(λ_0,θ)将驱动惩罚的性能，它们的调谐将是最重要的。在可分离情况中，我们假设$\theta\in(0,1)$是给定的，然而θ控制着大系数的期望比例，在缺乏真稀疏度水平q的先验信息的情况下，任意预先指定θ可能会无意识的高估\低估真稀疏度分数q/p，从而影响性能。

本节采用贝叶斯策略，赋予θ一个合适的先验$\theta\sim\pi(\theta)$，使惩罚可以在不需要设置θ接近q/p的情况下，实现一定的自适应和升压性能。假设$\pi(\theta)$是一个一般的先验，β中的坐标是根据$\pi(\theta)$的边际相依分布的。

$$\begin{aligned}\pi(\beta)&=\int_0^1\prod_{j=1}^{p}\left[\theta\varphi_1(\beta_j)+(1-\theta)\varphi_0(\beta_j)\right]\pi(\theta)\mathrm{d}\theta\\&=\left(\frac{\lambda_1}{2}\right)^p\mathrm{e}^{-\lambda_1|\beta|}\int_0^1\frac{\theta^p}{\prod_{j=1}^{p}p_\theta^*(\beta_j)}\pi(\theta)\mathrm{d}\theta\end{aligned}\tag{27}$$

重写(27)作为惩罚函数，我们得到以下不可分离的 SSL 惩罚变量。

4.1. 不可分离 SSL 惩罚

定义 2：不可分离的 Spike and Slab Lasso (NSSL)惩罚在 $\theta\sim\pi(\theta)$ 下定义为：

$$pen_{NS}(\beta)=\log\left[\frac{\pi(\beta)}{\pi(0_p)}\right]=-\lambda_1|\beta|+\log\left[\frac{\int\frac{\theta^p}{\prod_{j=1}^{p}p_\theta^*(\beta_j)}\pi(\theta)\mathrm{d}\theta}{\int\frac{\theta^p}{\prod_{j=1}^{p}p_\theta^*(0)}\pi(\theta)\mathrm{d}\theta}\right]\tag{28}$$

再一次，我们把惩罚中心化，使 $pen_{NS}(0)=0$。将(28)与(11)相比，NSSL 惩罚仍然是(可分离的) Lasso 部分和非凹部分的相加组成，但现在非凹部分将是不可分的。一般来说，(28)中的积分没有一个封闭形式的解，这似乎使惩罚的可处理型变得复杂化了，然而，当意识到先验(内含偏差项)的得分函数可以写成简单且非常直观的形式后，操作就变得非常简单了，这种形式出现在下面引理 5 的不可分离类比中。令 $\beta_{\setminus j}$ 表示去掉第 j 个分量之外的其余分量所组成的 β 的子向量。

引理 5：不可分离的 Spike and Slab Lasso 惩罚(28)的导数满足：

$$\frac{\partial pen_{NS}(\beta)}{\partial|\beta_j|}\equiv-\lambda^*(\beta_j;\beta_{\setminus j})\tag{29}$$

其中，

$$\lambda^*(\beta_j;\beta_{\setminus j})=p^*(\beta_j;\beta_{\setminus j})\lambda_1+\left[1-p^*(\beta_j;\beta_{\setminus j})\right]\lambda_0\tag{30}$$

$$p^*(\beta_j;\beta_{\setminus j})\equiv\int_0^1p_\theta^*(\beta_j)\pi(\theta\mid\beta)\mathrm{d}\theta\tag{31}$$

我们现在稍微停顿一下，来看(30)和(12)的区别，不可分离惩罚的混合概率为 $p_\theta^*(\bullet)$ 在 $\pi(\theta|\beta)$ 上平均得到的聚合概率 $p^*(\beta_j;\beta_{\setminus j})$。正是通过这种平均，惩罚才有机会了解 β 的稀疏水平。对不可分离惩罚的初步认识表明，它的自适应机制通过条件分布在概率域内运行，这一点在可分离惩罚中完全缺失了。由于 $p_\theta^*(\beta_j)$ 是 θ 的非线性函数，因此在(17)中撇去 θ 对平均混合权重 $p_\theta^*(\beta_j;\beta_{\setminus j})$ 的影响尚不明显。这个谜题在以下引理中展开，它为 NSSL 惩罚的实施和理论研究提供了简化。

引理 6：给定 $\beta\in\mathbb{R}^p$ 以及先验 $\pi(\theta)$，则有：

$$p^*(\beta_j;\beta_{\setminus j})=p_{\theta_j}^*(\beta_j)\tag{32}$$

其中，

$$\theta_j=E\left[\theta\middle|\beta_{\setminus j}\right]\tag{33}$$

上述引理的意义在于，通过简单的代入 $\theta=E\left[\theta\middle|\beta_{\setminus j}\right]$，我们可以将对可分离情形的认识转化为不可分离情形的认识，上述两个式子也说明了完全贝叶斯公式如何将概率意义赋予 NSSL 惩罚元素。

4.2. 自适应权重

我们现在考虑条件均值 $E\left[\theta\middle|\hat{\beta}_{\setminus j}\right]$ 的形式，当 p 充分大时，后验期望 $E\left[\theta\middle|\hat{\beta}_{\setminus j}\right]$ 将非常相似且接近于 $E\left[\theta\middle|\hat{\beta}\right]$。对于 θ 的先验，我们将使用标准 beta 先验 $\theta\sim B(a,b)$。现在我们检验条件分布 $\pi\left(\theta\middle|\hat{\beta}\right)$，这种条件分布将受到非零系数 $\hat{q}=\left\|\beta\right\|_0$ 的数量及其大小的影响。假设 $\hat{\beta}$ 中的前 $\hat{q}$ 项是非零的，这个分布的密度为：

$$\pi\left(\theta\middle|\hat{\beta}\right)\propto\theta^{a-1}\left(1-\theta\right)^{b-1}\left(1-\theta z\right)^{p-\hat{q}}\prod_{j=1}^{\hat{q}}\left(1-\theta x_j\right) \tag{34}$$

其中，$z=1-\dfrac{\lambda_1}{\lambda_0}$，$x_j=\left(1-\dfrac{\lambda_1}{\lambda_0}\mathrm{e}^{\left|\hat{\beta}_j\right|\left(\lambda_0-\lambda_1\right)}\right)$。这个分布是高斯超几何分布的一种推广，归一化常数写为多变量超几何函数的欧拉积分表示，因此它的期望可以写成：

$$E\left[\theta\middle|\hat{\beta}\right]=\frac{\int_0^1\theta^a\left(1-\theta\right)^{b-1}\left(1-\theta z\right)^{p-\hat{q}}\prod_{j=1}^{\hat{q}}\left(1-\theta x_j\right)\mathrm{d}\theta}{\int_0^1\theta^{a-1}\left(1-\theta\right)^{b-1}\left(1-\theta z\right)^{p-\hat{q}}\prod_{j=1}^{\hat{q}}\left(1-\theta x_j\right)\mathrm{d}\theta} \tag{35}$$

虽然上面的表达式(35)似乎很难计算，但当 λ_0 非常大时，它可以有一个简单的多的形式，根据[12]，我们在引理 7 中获得了这个简单得多的形式：

引理 7：假设 $\pi\left(\theta\middle|\hat{\beta}\right)$ 是根据 $\pi\left(\theta\middle|\hat{\beta}\right)\propto\theta^{a-1}\left(1-\theta\right)^{b-1}\left(1-\theta z\right)^{p-\hat{q}}\prod_{j=1}^{\hat{q}}\left(1-\theta x_j\right)$ 来分布的。令 $\hat{q}=\left\|\hat{\beta}\right\|_0$，当 $\lambda_0\to\infty$ 时，

$$E\left[\theta\middle|\hat{\beta}\right]=\frac{a+\hat{q}}{a+b+G} \tag{36}$$

我们注意到这个表达式本质上是在 beta 先验下的 θ 的通常的后验均值。直观的说，当 λ_0 发散时，权重 $p_\theta^*\left(\beta_j\right)$ 集中在 0 和 1 处，得到熟悉的 $E\left(\theta\middle|\hat{\beta}\right)$ 的形式。

4.3. 不可分离情况的预测校正算法

在本节中，我们将预测－校正算法扩展到不可分离惩罚的情况。具有不可分离惩罚的逻辑回归模型通过最小化下列损失函数来寻找系数：

$$l\left(\beta,\lambda_{\theta_j}^*\right)=\left\{-\sum_{i=1}^{n}\left[y_i\log\left(p_i\right)+\left(1-y_i\right)\log\left(1-p_i\right)\right]-pen_{NS}\left(\beta\right)\right\} \tag{37}$$

$l\left(\beta,\lambda_{\theta_j}^*\right)$ 关于 β 的一阶导和二阶导函数分别如下所示：

$$H\left(\beta,\lambda_{\theta_j}^*\right)=\frac{\partial l}{\partial\beta}=-X^{\mathrm{T}}\left(y-p\right)+\lambda_{\theta_j}^*\left(\beta\right)Sgn\begin{pmatrix}0\\\beta\end{pmatrix} \tag{38}$$

$$\frac{\partial H}{\partial\beta}=\frac{\partial^2 l}{\partial\beta\partial\beta'}=X'WX \tag{39}$$

其中 W 是 n 阶对角阵，元素为 $W_{ii}=p_i\left(1-p_i\right),\ i=1,\cdots,n$。在贝叶斯框架中，通过将 θ 作为一个附加的模型参数，可将预测－校正算法扩展到不可分离惩罚的情况，现在只需要通过算法来更新 θ，而不是使用 θ 的固定值。因此，在第 k 次迭代中，使用 $\theta=\theta^{(k)}$ 来更新 $\beta^{(k+1)}$，其中是 θ 的第 k 次迭代值，是在引理 7 中得到的。

5. 模拟学习

在本节中，我们使用仿真实验来验证所提出的 Spike and Slab Lasso 方法，并与常见的 Lasso 方法进行比较。我们模拟了两个数据集，用第一个数据集作为训练数据来拟合模型，第二个数据集作为测试数据来评估预测值，对于每个模拟设置，我们重复模拟 50 次，并总结这些重复的结果。

我们使用所提出的 SSL 逻辑回归方法分析了每个模拟数据集，并与 Lasso 逻辑回归方法进行了对比，以说明我们的方法相比 Lasso 方法的优越性。对于 Lasso 方法，我们通过十折交叉验证选择 λ 的最优值，λ 决定了最优的 Lasso，我们还报告了最优的 Lasso 模型的结果(见图 1)。

对于每个数据集，令 $n=100$， $p=1000$，我们从多元高斯分布中生成一个数据矩阵 X，均值为 0_p，$\Sigma=\left(\sigma_{ij}\right)_{i,j=1}^{p}$，其中当 $i\neq j$ 时，$\sigma_{ij}=0.6$，否则，$\sigma_{ii}=1$。真向量 β_0 通过指定回归系数 $\frac{1}{\sqrt{3}}\{-2.5,-2,-1.5,-1,1,1.5,2,2.5\}$ 到 $q=8$ 个随机方向，其余系数设为 0，响应由 $N\left(\eta_i,1.6^2\right)$ 生成，其中 $\eta_i=\beta_0+\sum_{j=1}^{m}x_{ij}\beta_j$。我们考虑以下三种设置，其均方误差(MSE)图分别见图 2，图 3，图 4：

1) 非自适应选择 $\theta=0.5$，明显高估了真稀疏分数 8/1000；
2) 非自适应 oracle 选择 $\theta=8/1000$；
3) 自适应选择 $\theta\sim B(1,1000)$；

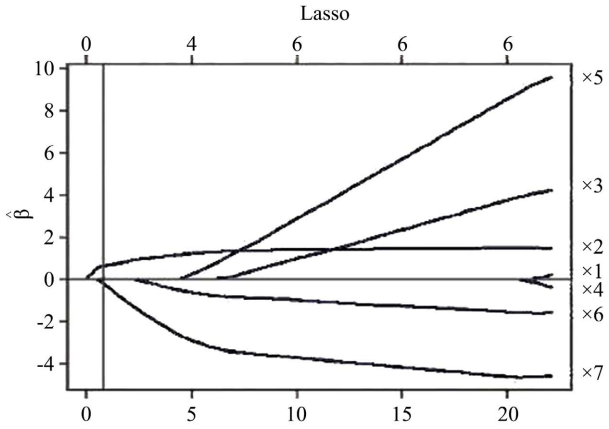

Figure 1. Estimator of the Lasso coefficient. The vertical lines in the figure correspond to the best Lasso model (cross validation)
图 1. Lasso 系数的估计值。图中的垂直线对应最佳 Lasso 模型(交叉验证)

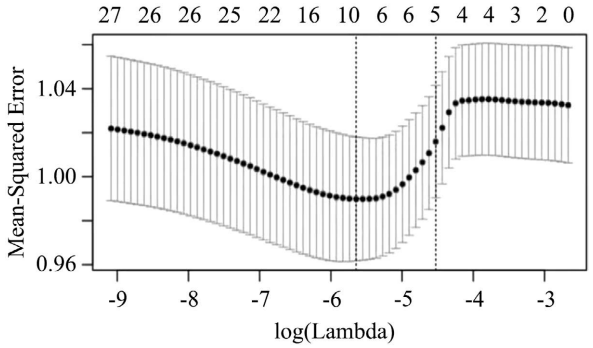

Figure 2. The MSE in case 1
图 2. 设置 1 情况下的 MSE

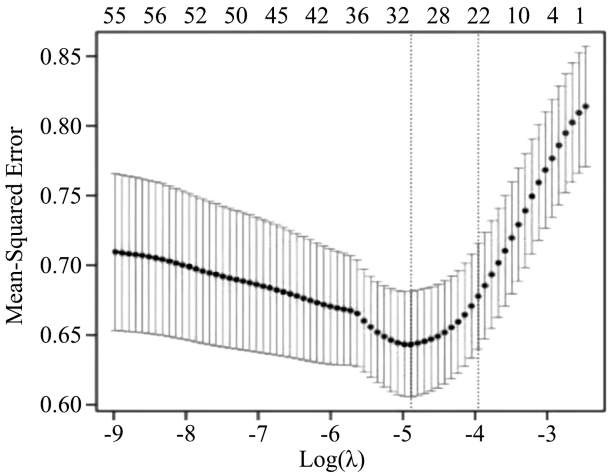

Figure 3. The MSE in case 2
图 3. 设置 2 情况下的 MSE

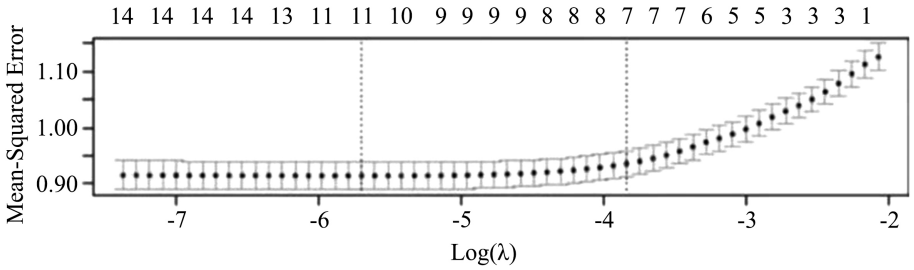

Figure 4. The MSE in case 3
图 4. 设置 3 情况下的 MSE

Table 1. Comparison of MSE between three Settings and Lasso under different values of Lambda0
表 1. Lambda0 的不同取值下，三种设置和 Lasso 情况的 MSE 对比

Lambda0	设置 1	设置 2	设置 3	Lasso
−3	1.332	0.96	0.105	1.032
−4	1.358	0.96	0.104	1.028
−5	1.356	0.96	0.096	1.001
−6	1.391	0.95	0.338	0.952
−7	1.353	0.96	0.359	1.013
−8	1.358	0.96	0.350	1.016
−9	1.402	0.90	0.301	1.021

从表 1 可以看出，在 $n = 100$，$p = 1000$ 的设置下，设置 3 是最优的，其次是设置 2，因为设置 2 是系数的真实稀疏度。说明当我们设置权重 θ 自适应于数据时，我们的模型能更好的适应数据的稀疏性。上述表中结果也表明 SSL 逻辑回归方法与 Lasso 进行比较时表现良好。

6. 南非心脏病数据集

在这里，我们提出了一个二元数据的分析，以说明 SSL 逻辑回归模型的使用效果。图 5 中的数据是冠状动脉危险因素研究(CORIS)基线调查的一个子集，该调查在南非西开普省的三个农村地区进行。该研究的目的是建立该高发地区缺血性心脏病危险因素的强度。数据代表 15~64 岁的白人男性，响应变量为调查时是否存在心肌梗死(MI)，(该地区 MI 的总体患病率为 5.1%)。在我们的数据集中有 160 个病例以及 302 个对照样本。

图 5 是南非心脏病数据的散点图矩阵。每个图显示一对风险因素，病例和对照用颜色编码(红色表示病例)。变量心脏病的家族史(famhist)是二元的(是或否)。使用疾病/非疾病变量，在不可分离惩罚的情况下，我们可以拟合一个 SSL 逻辑回归模型。系数的精确路径图在图 6 到图 8 中给出。

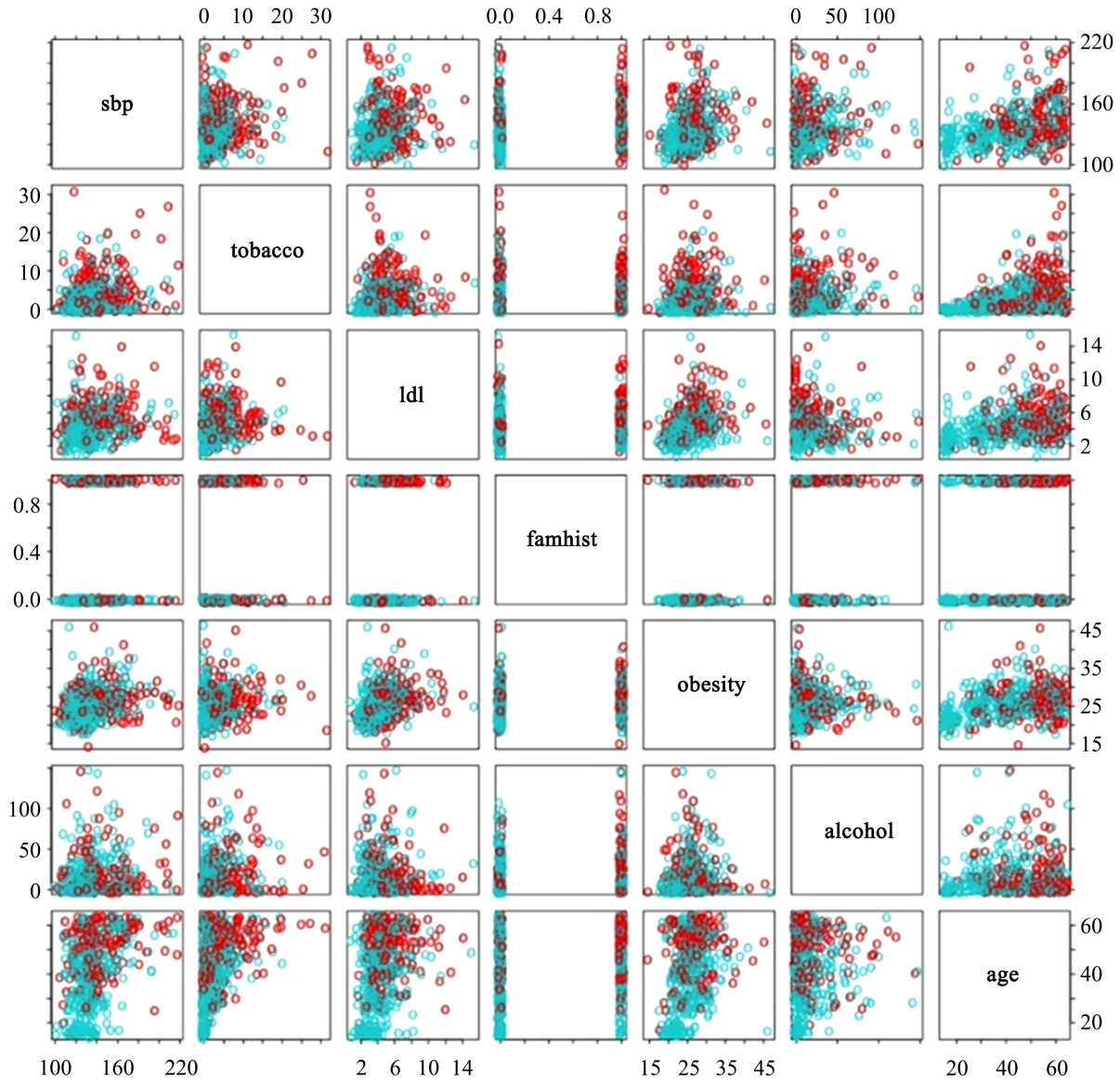

Figure 5. Scatter plot matrix of heart disease data in South Africa
图 5. 南非心脏病数据的散点图矩阵

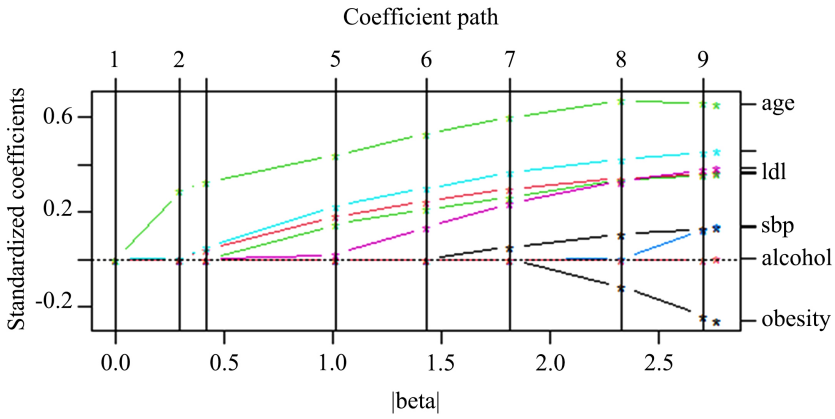

Figure 6. Exact path set of coefficients, where the x-axis is the l_1 norm of the coefficient, and the vertical discontinuity represents the modified position of the active set
图 6. 系数的精确路径集，其中 x 轴为系数的 l_1 范数，垂直间断表示活动集的修改位置

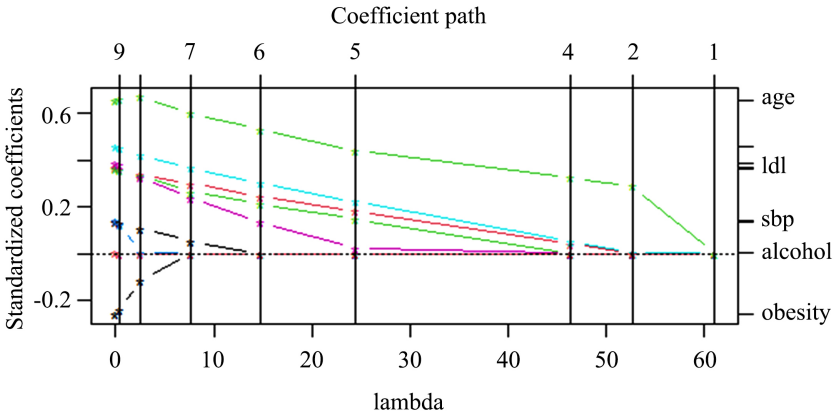

Figure 7. Exact path set of coefficients, where the x-axis is λ_θ^*, and the vertical discontinuity represents the modified position of the active set
图 7. 系数的精确路径集，其中 x 轴为 λ_θ^*，垂直间断表示活动集的修改位置

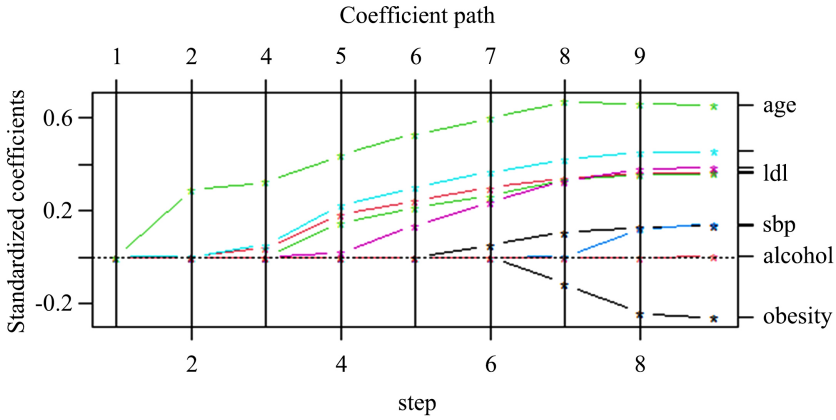

Figure 8. Exact path set of coefficients, where the x-axis is the number of steps, and the vertical discontinuity represents the modified position of the active set
图 8. 系数的精确路径集，其中 x 轴为步骤数，垂直间断表示活动集的修改位置

Table 2. Comparison of three methods for Heart Disease Data in South Africa
表 2. 南非心脏病数据使用三种方法的对比

	SSL 逻辑回归 $\theta \sim B(1,p)$	SSL 逻辑回归 $\theta = 0.5$	Lasso 逻辑回归
MSE	0.2282609	0.2580645	0.2783626
Bias	0.01086957	0.02150538	0.08695652
AUC	0.705	0.687	0.669
Misclassification	0.27841982	0.27120152	0.2783626
Num.steps	32	78	14

对于该数据集，我们分别建立 Spike and Slab Lasso 逻辑回归模型和 Lasso 逻辑回归模型，并且均采用预测校正算法来求解，其中 λ_1 被设置为 0.1，λ_0 由十折交叉验证获得，结果如表 2 所示。可以看出，尽管 Lasso 逻辑回归模型的步骤数少于 SSL 逻辑回归，但无论 θ 是否固定，SSL 逻辑回归模型的 MSE，Bias，AUC，Misclassification 均优于 Lasso 逻辑回归。这正是因为(12)式，我们所提出的 SSL 逻辑回归模型是自适应于数据的。另外，可以看出，不可分离惩罚的情况下也就是 $\theta \sim B(1,p)$ 时，和可分离惩罚也就是 $\theta = 0.5$ 的情况相比，后者的误分类率略优于前者，而前者的 MSE，Bias，AUC 和步骤数均优于后者。该表说明了在心脏病数据集下 spike and slab lasso 方法的性能优于 Lasso 方法。

7. 结论

在本文中，我们提出了 Spike and Slab Lasso 逻辑回归模型，将两个拉普拉斯密度的混合先验置于单个坐标上，可以自适应地收缩系数，分别讨论了权重 θ 固定和不固定两种情况。当 θ 固定时，导致可分离的惩罚，当 θ 不固定时，令 θ 服从 beta 分布，此时导致不可分离的惩罚，并利用预测校正算法分别求解两种惩罚模型。在模拟学习和南非心脏病数据集的研究中，将这两种模型与 Lasso 逻辑回归模型做对比，说明了本文所提出的 Spike and Slab Lasso 逻辑回归模型的有效性。

基金项目

本文由国家社会科学基金项目(21BTJ045)资助。

参考文献

[1] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning. Springer, New York, 33. https://doi.org/10.1007/978-0-387-84858-7

[3] Wainwright, M. (2015) Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC, London.

[4] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[5] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[6] George, E.I. and McCulloch, R.E. (1993) Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881-889. https://doi.org/10.1080/01621459.1993.10476353

[7] Popovic, M. (2022) Strain Wars 4-Darwinian Evolution through Gibbs' Glasses: Gibbs Energies of Binding and Growth Explain Evolution of SARS-CoV-2 from Hu-1 to BA.2. *Virology*, **575**, 36-42.

https://doi.org/10.1016/j.virol.2022.08.009

[8] Chipman, H., George, E.I., McCulloch, R.E., Clyde, M., Foster, D.P. and Stine, R.A. (2001) The Practical Implementation of Bayesian Model Selection. *IMS Lecture Notes-Monograph Series*, **38**, 65-134. https://doi.org/10.1214/lnms/1215540964

[9] Ishwaran, H. and Rao, J.S. (2005) Spike and Slab Gene Selection for Multigroup Microarray Data. *Journal of the American Statistical Association*, **100**, 764-780. https://doi.org/10.1198/016214505000000051

[10] Ishwaran, H. and Rao, J.S. (2011) Consistency of Spike and Slab Regression. *Statistics & Probability Letters*, 81(12), 1920-1928. https://doi.org/10.1016/j.spl.2011.08.005

[11] Narisetty, N.N. and He, X. (2014) Bayesian Variable Selection with Shrinking and Diffusing Priors. *Annals of Statistics*, **42**, 789-817. https://doi.org/10.1214/14-AOS1207

[12] Ročková, V. and George, E.I. (2018) The Spike-and-Slab Lasso. *Journal of the American Statistical Association*, **113**, 431-444. https://doi.org/10.1080/01621459.2016.1260469

[13] Ročková, V. (2018) Bayesian Estimation of Sparse Signals with a Continuous Spike-and-Slab Prior. *Annals of Statistics*, **46**, 401-437. https://doi.org/10.1214/17-AOS1554

[14] Bai, R., Moran, G.E., Antonelli, J.L., Chen, Y. and Boland, M.R. (2020) Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models. *Journal of the American Statistical Association*, **117**, 184-197. https://doi.org/10.1080/01621459.2020.1765784

[15] Hu, H.-Y., Wu, D., You, Y.-Z., Olshausen, B. and Chen, Y. (2022) RG-Flow: A Hierarchical and Explainable Flow Model Based on Renormalization Group and Sparse prior. *Machine Learning*: *Science and Technology*, **3**, Article ID: 035009. https://doi.org/10.1088/2632-2153/ac8393

[16] Ji, Y. and Shi, H. (2022) Constrained Bayesian Doubly Elastic Net Lasso for Linear Quantile Mixed Models. *Journal of Statistical Computation and Simulation*, **92**, 579-609. https://doi.org/10.1080/00949655.2021.1968398

[17] Muhammedrisaevna, T.M., Shukrullaevich, A.F. and Bakhriddinovna, A.N. (2021) The Logistics Approach in Managing a Tourism Company. *ResearchJet Journal of Analysis and Inventions*, **2**, 231-236.

[18] Park, M.Y. and Hastie, T. (2007) L_1-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **69**, 659-677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

[19] Lim, M. and Hastie, T. (2015) Learning Interactions via Hierarchical Group-Lasso Regularization. *Journal of Computational and Graphical Statistics*, **24**, 627-654. https://doi.org/10.1080/10618600.2014.938812