涉密论文 □　公开论文 □

浙 江 大 學

本科生毕业论文

题目　单指数模型的贝叶斯直接估计方法

姓名与学号　孙寅瑞 3160104871

指导教师　蒋杭进

年级与专业　2016级数学与应用数学

所在学院　数学科学学院

递交日期　2020 年 5 月 17 日

浙江大学本科生毕业论文（设计）承诺书

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2. 本人在毕业论文（设计）中除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得浙江大学或其他教育机构的学位或证书而使用过的材料。

3. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文（设计）工作过程中没有伪造数据等行为。

5. 若在本毕业论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6. 本人完全了解浙江大学有权保留并向有关部门或机构送交本论文（设计）的复印件和磁盘，允许本论文（设计）被查阅和借阅。本人授权浙江大学可以将本论文（设计）的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文（设计）。

作者签名：　　导师签名：

签字日期：　　年　月　日　　签字日期　　年　月　日

致谢

感谢我所有的家人对我的照顾和关心，家人的支持与陪伴是我不断前行的坚实后盾。

感谢蒋杭进老师在这近半年的时间里对我的悉心指导。这篇论文的完成，离不开蒋老师在选题上的引导和在我遇到问题时提供的帮助。向蒋杭进老师致以衷心的谢意。

转眼间，本科生涯即将结束。4 年的时间里，我在浙里经历过酸甜苦辣，我想多年以后这都将成为美好的回忆。感谢在浙里遇到的每一位老师，正是老师们的细心教导，帮助我在这 4 年里不断成长。怀念居住在同一屋檐下的室友们：王焕、张宇然、林弋皓、黄宗玮、张泽奇、陈威涛、王哲辰、陶淳涛，宿舍里的集体生活总是那么有趣。感谢一路以来帮助过我、一起努力的朋友们：徐浩森、金寅、习卓凡、王卓凡、徐天仪、代汶利、李嘉文、卫艺萌、张睿……感谢何呈栎学长一直以来对我的帮助和关心。感谢我的老友们：赵中源、于洋欢、张喆、韩仪韬。以及我的小伙伴们：王伟钊、牛梓豪。祝愿每一个人都能够前程似锦，在追梦的路上不断向前。

丈夫志四海，万里犹比邻。祝愿每个人都能拥有自己的海阔天空。

摘要

单指数模型是广义线性模型的一般化推广，它利用未知函数 g 将自然参数与解释变量的线性组合连接起来。在本篇论文中，我们发展了单指数回归模型和单指数 Logistic 模型的贝叶斯估计及其变量选择方法。我们应用高斯过程先验模型来处理单指数模型中的未知函数 g。结合增加服从 Polya-Gamma 分布的辅助变量的方法，我们得到了单指数 Logistic 模型的贝叶斯估计方法。考虑到高斯过程中协方差矩阵的奇异性问题，我们使用 Metropolis-Hasting within Partially Collapsed Gibbs 抽样的方法来得到参数的估计。

针对单指数 β，我们考虑了不同的先验模型及其抽样算法来得到它的贝叶斯估计，其中包括单位球面先验、放松先验和无约束先验。并且基于这些结果，结合经典的 Spike-Slab 先验模型，我们发展了单指数模型的贝叶斯变量选择方法。每一种先验模型的后验分布、抽样算法的条件分布的推导及其抽样细节列在了附录中。

我们分别在模拟数据集上和真实数据集上应用了我们的估计方法，并且将部分结果和文献结果作比较。数值结果显示我们的基于高斯过程模型的贝叶斯估计方法得到了良好的结果。

关键词：单指数模型；高斯过程；Logistic 模型；变量选择；贝叶斯推断

Abstract

Single Index Model is the generalization of the Generalized Linear Model. The model link the natural parameter with the linear combination of explanatory variables through the unknown function g. In the thesis, we develope Bayesian estimation and variable selection approaches for single index regression model and single index logistic model. We apply the Gaussian process model to deal with the unknown function g in the model. Combined with the data augmentation method that adds auxiliary random variables following the Polya-Gamma distribution, we obtain the Bayesian estimation approach for the single index Logistic model. Considering the singularity of the gram matrix with Gaussian process, we apply the Metropolis-Hasting within Partially Collapsed Gibbs sampler for the Bayesian inference.

For the single index β, we consider different prior models and sampling algorithms to obtain its Bayesian estimation, including unit sphere prior, relaxed prior and unconstrained prior. And based on these results, combined with the classic Spike-Slab prior, we develope Bayesian variable selection method for the single index model. Each model's posterior distributions and the samplers' conditional posterior distributions with its sampling details are derived and listed in the appendix.

We apply our method on the simulation data as well as real data and compare some of the results with the results in the literature. The numerical results show that our Bayesian estimation approaches based on Gaussian process model work well.

Key Words: Single Index Model; Gaussian Process; Logistic Model; Variable Selection; Bayesian Inference

目录

第一部分　毕业论文

第一部分

毕业论文

1 绪论

1.1 引言

回归模型假设 $Y = f(X) + \epsilon$ 的形式，其中 Y 是 1 维响应变量，X 是 p 维解释变量，ϵ 是误差项。单指数回归模型假设 $f(X) = g(X^T\beta)$ 的形式，其中称回归系数 $\beta \in \mathbb{R}^p$ 为单指数，g 是未知连接函数。通常情况下，为了可以唯一地确定单指数 β，我们假设 $\beta \in \{\theta \in S^{p-1} : \theta$的第一个非 0 分量的符号为正$\}$。从形式上可见，单指数回归模型是一般线性模型 $Y = X^T\beta + \epsilon$ 更一般的推广，同时是投影寻踪回归 $f(X) = \sum_{m=1}^{M} g_m\left(X^T\beta_m\right)$ 在 $M = 1$ 时的特殊形式[1]。由此可见，单指数回归模型不仅具有拟合 Y 与 X 之间可能存在的非线性关系的能力，同时通过估计解释变量 X 的线性组合 $X^T\beta$ 中的单指数 β，此模型还具有一定的解释能力。

不限于回归问题，我们可以将单指数回归模型推广至广义线性模型中。广义线性模型假设响应变量 Y 服从指数族分布，其密度可以写成 $p(Y) = b(Y)\exp\{\eta^T T(Y) - a(\eta)\}$ 的形式，其中 η 称为自然参数，$T(Y)$ 为充分统计量，并且假设 $\eta = X^T\beta$。常见的广义线性模型包括一般线性模型、Logisitc 模型等。我们可以将单指数模型推广至广义线性模型，假设自然参数 $\eta = g(X^T\beta)$，其中 g 为未知函数。通过未知函数 g 将自然参数 η 与 $X^T\beta$ 连接起来，可以拓展广义线性模型的适用范围，我们称此模型为广义单指数模型。由此可见，单指数模型适用于任何可使用广义线性模型的情景中。当响应变量分别是连续、二分类变量时，我们可以应用单指数回归模型与单指数 Logistic 回归模型来代替一般线性模型与线性 Logistic 模型。

单指数模型中的一项重要的工作是单指数 β 和未知函数 g 的估计。由于函数 g 未知，因此与广义线性模型的估计方法相比，单指数模型的估计更加困难与复杂。以往的文献主要集中于单指数回归模型的估计工作，其中包括频率方法与贝叶斯方法。频率方法常常利用核密度估计的方法来处理未知连接函数 g，而贝叶斯方法则需要给予函数 g 一定的先验信息。在本文中，我们聚焦于单指数模型的贝叶斯估计方法，在单指数回归模型与单指数 Logistic 模型下，基于高斯过程模型，提出了新的关于 β 的估计及其变量选择方法，并且实现了多种关于单指

数 β 的抽样方法。

本文文章结构如下：小节 1.2中介绍了关于单指数模型估计方法的一些文献成果；节 2中介绍了本文工作中所需要的一些基础知识，其中包括高斯过程模型与 MCMC 抽样算法；节 3展示了单指数回归模型的先验模型、抽样方法与数值结果；节 4中展示了单指数 Logistic 模型的先验模型、估计方法与数值结果；节 5展示了回归模型与 Logistic 模型下单指数 β 的变量选择方法及其数值结果；节 6总结了我们的估计方法的优缺点，提出了未来工作的一些展望。

1.2 单指数模型

在广义线性模型下，当服响应变量 Y 从指数分布族时，我们可以计算得到 $\mathbb{E}(Y|\eta) = a'(\eta)$。常见的正态分布、伯努利分布关于其期望都属于指数分布族。例如，当 Y 服从正态分布时，$\mathbb{E}(Y|\eta) = \eta = X^T\beta$，这与一般线性模型相对应；当 Y 服从伯努利分布时，$\mathbb{E}(Y|\eta) = \exp(\eta)/\{1+\exp(\eta)\} = \exp(X^T\beta)/\{1+\exp(X^T\beta)\}$，这与 Logistic 模型相对应。因此，单指数回归模型与单指数 Logistic 模型可以分别假设为 $Y = g(X^T\beta)+\epsilon$，$\epsilon \sim \mathrm{N}(0,\sigma)$ 与 $Y \sim \text{Bernoulli}(\pi)$，$\pi = \exp\{g(X^T\beta)\}/\{1+\exp\{g(X^T\beta)\}\}$ 的形式，其中函数 g 未知。记 $x_i, i = 1,2,\cdots,n$ 是来自 p 维总体 X 的独立同分布样本，$x = (x_1, x_2, \cdots, x_n)^T$；$y_i, i = 1,2,\cdots,n$ 是来自总体 Y 的响应变量样本，$y = (y_1, y_2, \cdots, y_n)^T$。我们的目标是在给定 (x, y) 下，估计单指数回归与 Logistic 模型中的单指数 β，并发展变量选择方法。

单指数模型的估计方法主要包括频率估计方法与贝叶斯估计方法两类。频率估计方法主要包括 M 估计、平均导数估计与充分降维方法三类。M 估计方法主要在单指数回归模型 $Y = g(X^T\beta)+\epsilon$ 下，通过优化 $\min_\beta \sum_{i=1}^n \psi\left(y_i, g(x_i^T\beta)\right)$ 来得到 β 的估计值 $\hat{\beta}$。由于函数 g 未知，因此需要事先利用核密度估计方法来得到其估计量 $\hat{g}$。通常 ψ 选择为平方损失函数。在此基础上，Ichimura 提出利用加权最小二乘方法来处理误差项异方差问题[2]。Hardle 等考虑了利用核密度估计方法估计 g 时的带宽选择问题，将带宽 h 加入优化函数中，将单指数 β 的估计与带宽 h 的选择转变为解决 $\min_{\beta,h} \Sigma_{i=1}^n\{y_i - \hat{g}_i(x_i^T\beta, h)\}^2$ 这一优化问题[3]。Xia 利用局部线性函数逼近未知函数 g，解决优化问题 $\min_\beta \sum_{j=1}^n \sum_{i=1}^n \left\{y_i - a_j - b_j\beta^T(x_i - x_j)\right\}^2 w_{ij}$，其中 (a_j, b_j) 需要提前估计，w_{ij} 是对应权重[4]。但是，此类 M 估计方法通常面临

核密度估计时的带宽选择问题与数值优化问题。

与 M 估计方法相比，平均导数估计方法更加巧妙地处理了未知函数 g 与单指数 β 之间的关系[5]。在 $\mathbb{E}(Y|X) = m(X) = g(X^T\beta)$ 模型假设下，我们可以得到 $m'(X) = \partial g(X^T\beta)/\partial X = g'(X^T\beta)\beta$，$\mathbb{E}(m') = \mathbb{E}\left(g'(X^T\beta)\right)\beta$，即 $m'(X)$ 的期望与单指数 β 成正比。因此我们只需要估计函数 m 的导数的期望。但是这篇文章中的方法只适用于 X 均为连续变量情形，Horowitz 和 Hardle 将平均导数方法推广至 X 含有离散变量的情形[6]。Hristache 等使用迭代的方法交替估计 $\hat{\beta}$ 与 $\hat{m}$[7]。

充分降维方法估计方法在假设 $Y \perp\!\!\!\perp X|X^T\beta$ 下估计由 β 张成的线性子空间 Span(β)。事实上不仅限于单指数，充分降维方法聚焦于估计满足 $Y \perp\!\!\!\perp X|X^TB$ 条件的矩阵 B 的列空间 Span(B)，其中 B 是 $p \times d$ 维矩阵。Span(B) 称为降维子空间，所有满足条件的降维子空间的交称为中心子空间。目前已经有许多方法用来估计中心子空间，包括经典的切片逆回归[8] 与切片平均方差估计[9] 等，但是此类方法通常有较强的假设条件。除此以外，在条件独立 $Y \perp\!\!\!\perp X|X^T\beta$ 的假设下，Yin 和 Cook 还提出了一种一般性的单指数估计方法[10]。

相比于频率估计方法，贝叶斯单指数估计方法的研究相对较少。与频率方法利用核密度方法或局部线性逼近方法来处理未知函数 g 的思路不同，贝叶斯方法通常赋予函数 g 一定的先验信息。Antoniadis 等用 B 样条函数来逼近未知函数 g，此时单指数模型成为了一种特殊的线性回归模型[11]。这里样条基函数的次数与节点位置需要事先确定。与样条逼近方法不同，Choi 等利用高斯过程方法来对未知函数 g 设定先验过程，并且提出用经验贝叶斯的方法来处理高斯过程先验中协方差函数中的超参数[12]。Gramacy 和 Lian 提出在高斯过程先验模型下，可以取消掉单指数 β 在单位球面上的约束，使得 β 的先验分布的选择更加广泛[13]。

除此以外，单指数回归模型中结合变量选择的估计方法也得到了一定的发展。Kong 和 Xia 在交叉验证准则下，发展了一种类似逐步子集选择的参数估计与变量选择方法[14]。其他的的变量选择方法大多建立在压缩惩罚思想的基础上。比如，Wang 和 Yin 在最小平均方差估计方法[4] 的基础上，在损失函数中加上单指数的 l_1 范数作为惩罚项来对 β 进行估计，这种 Lasso 型的变量选择方法被称为稀疏最小平均方差估计[15]。Zeng 等改进了稀疏最小平均方差估计方法，将 β 与 g 的线性逼近斜率项相结合，更改了损失函数中惩罚项的形式[16]。Peng 和 Huang

将局部线性回归与 SCAD 型惩罚函数[17] 相结合，针对单指数模型提出了新的惩罚最小二乘方法及求解算法[18]。但是，基于在目标函数中添加惩罚项的变量选择方法，一直以来都具有求解过程复杂、很难选择最优参数的问题。在贝叶斯单指数回归模型变量选择方面，Wang 在样条先验逼近未知函数 g 的基础上，对每个回归系数引入指示变量来表示对应解释变量是否包含在模型中，并发展了对应的抽样方法[19]。

本文中我们将基于高斯过程模型，发展单指数模型的贝叶斯估计及其变量选择方法。目前单指数 β 的贝叶斯估计方法在回归模型上已经有一定的研究[12, 13]，我们将在这些研究的基础上比较关于 β 的不同抽样方法，并且将这些方法与变量选择方法结合，推广至单指数 Logistic 模型当中。

2 基础知识

2.1 高斯过程模型

高斯过程是任意有限维分布均为联合正态分布的随机过程，其理论在贝叶斯机器学习中应用广泛。对于函数 $f:\mathcal{X}\to\mathbb{R}$ 服从高斯过程先验，意味着对于任意有限个输入数据 $x=\{x_i\in\mathcal{X}\}_{i=1}^n$，作为多维随机变量，$(f(x_1),\cdots,f(x_n))$ 服从多元正态分布，其均值函数与协方差函数分别记为：

$$m(x_i)=\mathbb{E}\left(f(x_i)\right),$$

$$k(x_i,x_j)=\mathbb{E}\left(f(x_i)-m(x_i)\right)\left(f(x_j)-m(x_j)\right).$$

所以 f 服从高斯过程可记为 $f\sim\mathrm{GP}(m,k)$，在给定数据 $x=\{x_i\}_{i=1}^n$ 下，$f(x)\sim N\left(m(x),K(x)\right)$，其中 $K(i,j)=k(x_i,x_j)$。

联合正态分布由均值函数和协方差函数唯一确定，因此当应用高斯过程模型于机器学习任务中，我们需要选择合适的均值函数 m 和协方差函数 k。在实际应用中，常常选择均值函数 $m\equiv 0$。协方差函数的选择的基本要求是要使得协方差矩阵 K 满足半正定条件，满足此条件的函数 k 称为正定核函数，而对应的协方差矩阵 K 称为 Gram 矩阵。不同的核函数具有不同的特点，一个常用的协方差函数是高斯型核函数：

$$k(x_i,x_j)=\tau\exp\{-\frac{1}{2}\sum_{d=1}^{p}\left(\frac{x_i^{(d)}-x_j^{(d)}}{l_d}\right)^2\},\tag{2-1}$$

其中 $\tau,\{l_d\}_{d=1}^p$ 是协方差函数中的超参数。除此以外还有很多其他协方差函数，如点积函数、Matern 类函数等[20]。不同的核函数适用于不同的数据集。

高斯过程在贝叶斯回归、分类、模型选择等任务上都有良好表现，其作用可以从基函数回归的角度进行解释[20]。一般线性回归模型 $\mathbb{E}(Y|X)=X^T\beta$ 不具有拟合 Y 与 X 之间可能存在的非线性关系的能力，对此我们可以将解释变量 X 投影到更高维空间，用高维空间中的线性关系拟合当前空间中的非线性关系。这里记投影函数为 $\phi:\mathbb{R}^p\to\mathbb{R}^d$，其中 $d<\infty$ 或 $d=\infty$，函数 ϕ 称为基函数。这样线性回归可以写作 $\mathbb{E}(Y|X)=f(X)=\phi(X)^T\beta$，为表示方便我们把 β 写作有限维

的形式。如果对 β 给定正态分布先验：$\beta \sim N(0,\Sigma)$，那么 $f|X = f(X) = \phi(X)^T\beta$ 同样服从正态分布，并且有：

$$\begin{aligned}
\mathbb{E}\left(f(X)\right) &= \phi(X)^T\mathbb{E}(\beta) = 0, \\
\mathrm{Cov}\left(f(X), f(X')\right) &= \mathbb{E}\left(f(X)\cdot f(X')\right) \\
&= \mathbb{E}\left(\phi(X)^T\beta\cdot\phi(X')^T\beta\right) \\
&= \phi(X)^T\mathbb{E}(\beta\beta^T)\phi(X') \\
&= \phi(X)^T\Sigma\phi(X') \\
&= \psi(X)^T\psi(X') = k(X,X'),
\end{aligned}$$

其中 $\psi(X) = \Sigma^{1/2}\phi(X)$。从这个角度看，在高斯过程中协方差函数 k 的选择就意味着基函数 ϕ 与 β 先验协方差矩阵 Σ 的选择，并且协方差 $k(X,X')$ 可以看作是 $\psi(X)$ 和 $\psi(X')$ 的内积。尤其是当 $\Sigma = I$ 时，$k(X,X') = \phi(X)^T\phi(X')$ 是有限或无限维高维空间中经过 ϕ 投影的解释变量的内积。这里体现的思想与支持向量机模型有很多相通之处。

高斯过程在应用中取得了良好的效果，但是它同样具有一定的问题，其中最令人困扰的是应用高斯过程先验带来的计算问题。比如，在高斯过程回归模型 $Y = f(X) + \epsilon$ 中，在给定先验 $f \sim \mathrm{GP}(0,k), \epsilon \sim N(0,\sigma)$ 且 ϵ 与 X 独立下，给定样本数据 $y = \{y_i\}_{i=1}^n, x = \{x_i\}_{i=1}^n$ 时，容易计算得到 $y|x \sim N(0, K+\sigma I)$。在给定新的输入数据 $\tilde{x}$ 时，对应的预测结果 $\tilde{f}|y,x,\tilde{x} \sim N(E_{\tilde{f}}, Cov_{\tilde{f}})$，其中

$$\begin{aligned}
E_{\tilde{f}} &= k(\tilde{x},x)\left(\sigma I + K\right)^{-1} y, \\
Cov_{\tilde{f}} &= k(\tilde{x},\tilde{x}) - k(\tilde{x},x)\left(\sigma I + K\right)^{-1} k(x,\tilde{x}).
\end{aligned}$$

由上式可见，我们需要做关于 $(\sigma I + K)^{-1}$ 的计算。常规操作是利用诸如 Cholesky 分解的方法解线性方程 $(\sigma I + K)\, s = y$，此类操作的计算量大多是 $O(n^3)$，其中 n 是训练集的样本容量。当解决大规模推断问题，或者样本量很大时，高斯过程模型会产生相当大的计算负担。为此，有一系列方法用来解决此问题，较为常见的是选择 m 个子样本数据来逼近原结果[21-24]。其中子样本可能来自样本数据，也可能是不限于样本数据的“伪输入”。比如 Nystrom 逼近方法选择用 $\tilde{K} = K_{nm}K_{mm}^{-1}K_{mn}$ 代替上述线性方程组中的 K，此时方程组可用特定公式求解，其计算量为 $O(m^2n)$，其中 m 是选择训练集中子集的个数[21]。除此以外，有时我

们需要计算 $\log p(y|x)$ 的值，由于 $y|x \sim N(0, K+\sigma I)$，所以行列式 $\log|\sigma I + K|$ 的计算同样成为计算负担，为此有不同的随机逼近的方法用来解决此问题[25-27]。

除了回归与分类在内的监督学习外，高斯过程还广泛应用于无监督学习与半监督学习，如高斯过程潜变量模型[28]。总的来说，高斯过程模型是一种灵活且有效的贝叶斯工具，解释力强，并且目前已经有相当丰富的理论研究，适用场景广泛。

2.2 MCMC 抽样

马尔可夫链蒙特卡洛方法，简称 MCMC 方法，是贝叶斯统计中用于从后验分布中进行抽样的有效计算工具。MCMC 抽样的原理是通过迭代的方法构造目标参数的一个马尔可夫链，使得其平稳分布与我们的目标后验分布一致。在这一部分中，我们不再叙述关于 MCMC 抽样的理论性质与收敛性证明，只介绍本文中需要的具体的抽样方法。

Gibbs 抽样是一种用于高维参数抽样的 MCMC 算法。假设我们将参数 θ 分为 d 个部分 $\theta = (\theta_1, \cdots, \theta_d)$，每一个参数的条件分布 $p(\theta_j|\theta_{-j})$ 都可以直接抽样，其中 $\theta_{-j} = (\theta_1, \cdots, \theta_{j-1}, \theta_{j+1}, \cdots, \theta_d)$。记 θ_j^t 为参数 θ_j 在第 t 次迭代的抽样结果，如果我们已经得到参数的第 t 次估计 $\theta^t = (\theta_1^t, \cdots, \theta_d^t)$，那么第 $t+1$ 的迭代按照 $j = 1, 2, \cdots, d$ 的顺序通过 $\theta_j^{t+1} \sim p(\theta_j|\theta_{-j}^t)$ 进行抽样，得到第 $t+1$ 次估计，其中 $\theta_{-j}^t = (\theta_1^{t+1}, \cdots, \theta_{j-1}^{t+1}, \theta_{j+1}^t, \cdots, \theta_d^t)$。在给定初始值 θ^0 下，迭代 N 次，可以得到一条马尔可夫链 $\theta^0, \theta^1, \theta^2, \cdots, \theta^N$。

另外一种常见的 MCMC 抽样方法是 Metropolis-Hasting 算法。Metropolis-Hasting 算法是一种接受/拒绝规则的抽样算法，假设 $p(\theta)$ 是我们的目标分布，θ^t 是第 t 次迭代后得到的抽样结果，那么 Metropolis-Hasting 算法从某个条件分布 $q(x|y)$ 中抽取 $\theta^* \sim q(\theta|\theta^t)$，计算

$$r = \min\{\frac{p(\theta^*)q(\theta^t|\theta^*)}{p(\theta^t)q(\theta^*|\theta^t)}, 1\},$$

并以 r 的概率 $\theta^{t+1} = \theta^*$，以 $1-r$ 的概率 $\theta^{t+1} = \theta^t$。这里分布 $q(x|y)$ 称为建议分布或跳跃分布。不同类型的参数 θ 应该选择不同的建议分布，常见的建议分布有正态分布随机游走 $\theta^* \sim \mathrm{N}(\theta^t, \sigma)$ 等。

在 Gibbs 抽样的实际应用中，常常出现的情况是并非每一个条件分布 $p(\theta_j|\theta_{-j})$ 都可以直接的快速抽样。在这种情况下，我们可以使用 Metropolis-Hasting 抽样与 Gibbs 抽样的混合抽样方法。假设第 j 个分量 θ_j 很难从 $p(\theta_j|\theta_{-j})$ 中直接抽样，在第 $t+1$ 次迭代更新 θ_j 时，我们已经得到 $\theta^t_{-j} = (\theta^{t+1}_1, \cdots, \theta^{t+1}_{j-1}, \theta^t_{j+1}, \cdots, \theta^t_d)$，为了从 $p(\theta_j|\theta^t_{-j})$ 中进行抽样，我们可以取 $p(\theta_j|\theta^t_{-j})$ 为 Metropolis-Hasting 中的目标分布 $p(\theta)$，利用 Metropolis-Hasting 抽样的方法更新 θ_j，得到 θ^{t+1}_j。这样的混合抽样方法也称为 Metropolis-Hasing within Gibbs 抽样。

Gibbs 抽样中，分块 (Blocking) 和坍塌 (Collapsing) 一种有效的技巧[29]。以参数 $\theta = (\theta_1, \cdots, \theta_d)$ 为例，将参数 θ_{d-1} 和 θ_d 整合为同一个参数 $\theta_{d-1,d} = (\theta_{d-1}, \theta_d)$，在 Gibbs 抽样中从后验分布 $p(\theta_{d-1,d}|\theta_{-(d-1,d)})$ 中抽取 $\theta_{d-1,d}$ 的过程称为 Blocked Gibbs 抽样；如果参数 θ_d 可以在目标分布 $p(\theta)$ 中被积分得到 θ_{-d} 边际分布 $p(\theta_{-d})$，并且我们并不需要参数 θ_d 的估计，那么我们只需要从目标分布 $p(\theta_{-d})$ 中对 θ_{-d} 进行抽样，这样的过程称为 Collapsed Gibbs 抽样。Partially Collapsed Gibbs(简记为 PCG) 抽样是分块与坍塌抽样的更一般化推广，它有利于提高马尔可夫链的收敛速度，并且方便计算[30]。以三个参数 $(\theta_1, \theta_2, \theta_3)$ 为例，Gibbs 抽样需要从条件分布 $p(\theta_1|\theta_2, \theta_3)$，$p(\theta_2|\theta_1, \theta_3)$，$p(\theta_3|\theta_1, \theta_2)$ 中迭代抽样。如果我们很难直接从 $p(\theta_1|\theta_2, \theta_3)$ 中抽样，但是可以从联合条件分布 $p(\theta_1, \theta_2|\theta_3)$ 中抽样，那么我们可以从 $p(\theta_1, \theta_2|\theta_3)$，$p(\theta_2|\theta_1, \theta_3)$，$p(\theta_3|\theta_1, \theta_2)$ 中迭代抽样，并简化为从 $p(\theta_1|\theta_3)$，$p(\theta_2|\theta_1, \theta_3)$，$p(\theta_3|\theta_1, \theta_2)$ 中迭代抽样，这样的抽样过程称为 PCG 抽样。但是，PCG 抽样中每次迭代时参数的抽样顺序需要特别注意，不恰当的 PCG 抽样方法可能导致平稳分布与目标分布并不一致，尤其是当使用 Metropolis-Hasting 算法与 PCG 抽样结合时，更应该格外注意马尔可夫链的平稳分布是否被破坏[31]。

应用 MCMC 方法从目标分布中进行抽样时，我们需要检验得到的马尔可夫链是否收敛。常用的诊断方法是从不同的初始值出发，对参数进行多次抽样，得到多条马尔可夫链，做出参数值关于迭代次数的图。如果随着迭代次数的增加，不同初始值得到的参数的估计结果汇聚在一起，那么我们可以认为得到的马尔可夫链收敛。另外一种常用的诊断方法基于不同初始值下得到的马尔可夫链的链内与链间方差的计算，由组内与组间方差的计算，可以得到 Potential Scale Reduction Factor(简记为 PSRF)，随着抽样次数趋于无穷，PSRF 减小趋于 1[32, 33]。这里不再详细介绍此方法的内容，通常来说，PSRF<1.2 意味着马尔可夫链收

敛。

总的来说，MCMC 方法的发展解决了贝叶斯统计中令人困扰的计算问题，不同的抽样方法适用于不同的贝叶斯方法与模型。在本文中，我们将利用上述 MCMC 方法，从我们的模型的后验分布中进行抽样，得到参数的估计。

3 单指数回归模型估计

3.1 基本模型

单指数回归模型可以假设为：

$$Y = g\left(X^T\beta\right) + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma),$$

其中 Y 是 1 维响应变量，X 是 p 维解释变量，ϵ 是服从 0 期望的正态分布的误差项，g 为未知函数，单指数 $\beta \in \mathbb{R}^p$，$\|\beta\|_2 = 1$ 且其第一个非 0 分量符号为正。设

$$y_i = g(x_i^T\beta) + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma), i = 1, 2, \cdots, n,$$

基于小节 2.1中所描述的高斯过程模型内容，给定 g 高斯过程先验，$g \sim GP$，选择期望函数 $m \equiv 0$，协方差函数如式 2-1所示，则在 $x = (x_1, \cdots, x_n)$ 下，函数 g 的有限维分布为：

$$\left(g(x_1^T\beta), \cdots, g(x_n^T\beta)\right) \sim \mathrm{N}(0, K),$$

其中

$$K(i,j) = \tau \exp\{-\frac{\left(x_i^T\beta - x_j^T\beta\right)^2}{l}\}. \tag{3-1}$$

由于为了 β 的可识别性，我们通常需要给 β 设定单位球面 S^{p-1} 上的先验分布。一个简单的先验选择是单位球面上的均匀分布，即 $p(\beta) \propto I\left(\beta \in S^{p-1}\right)$。如果我们有关于 β 有一定的先验信息，那么可以选择 von-Mises Fisher 分布，给定先验 $\beta \sim \mathrm{vMF}(\kappa, \beta_{prior})$。这是以 β_{prior} 为中心的单位球面分布，超参数 κ 决定了分布的离散程度，当 $\kappa = 0$ 时即为球面上均匀分布[34, 35]。这里我们选择单位球面上的均匀分布，即 $\beta \sim \mathrm{Uniform}(S^{p-1})$。因此，我们可以得到下述先验模型：

$$\epsilon|\sigma \sim \mathrm{N}(0, \sigma I_n),$$

$$\begin{aligned} y|g,\beta,\sigma &\sim \mathrm{N}(g,\sigma I_n), \\ g|\beta,\tau,l &\sim \mathrm{N}\left(0,K\right), \\ \beta &\sim \mathrm{Uniform}(S^{p-1}). \end{aligned} \tag{3-2}$$

关于超参数 σ,τ,l 的设定，Choi 等提出利用经验 Gibbs 的方法，在每一步抽样中利用数值优化的方法更新参数的值，但是这样的方法大大增加了计算量[12]。这里我们选择对超参数 σ,τ,l 选择先验分布，计算其后验分布并利用 MCMC 抽样的方法更新这些超参数。这里我们选择逆 Gamma 分布先验，即 $\sigma \sim \mathrm{InvGamma}(a_\sigma,b_\sigma),\tau \sim \mathrm{InvGamma}(a_\tau,b_\tau),l \sim \mathrm{InvGamma}(a_l,b_l)$。

当高斯过程应用于单指数模型，由式 3-1 中协方差矩阵中的元素的形式可见，尽管在识别 β 的问题上我们需要约束 $\|\beta\|_2=1$，但事实上在此模型中 $\beta/\sqrt{l}$ 作为整体来影响最终的结果。Gramacy 和 Lian 提出在计算过程中可舍弃 $\|\beta\|_2=1$ 的约束，同时取消协方差函数中的超参数 l，这样可以为 β 提供更多的先验选择[13]。在无约束先验下，选择相互独立的正态分布是常见的先验选择。但是，正态分布的密度在尾部快速下降，属于先验信息较为集中的先验分布。Gelman 等提出在先验信息较少的情况下，选择厚尾的 t 分布族作为弱信息先验更为合适[36]。为了表示与抽样上的简洁，t 分布先验通常表示成正态分布-逆 Gamma 分布混合先验的形式，即 $\beta|\sigma_\beta \sim \mathrm{N}(0,\sigma_\beta)$，$\sigma_\beta \sim \mathrm{InvGamma}(a_{\sigma_\beta},b_{\sigma_\beta})$。在此混合先验下，我们可以计算 β 的边际分布为：

$$\begin{aligned} p\left(\beta\right) &\propto \int p\left(\beta,\sigma_\beta\right)d\sigma_\beta \propto \int p\left(\beta|\sigma_\beta\right)p\left(\sigma_\beta\right)d\sigma_\beta \\ &\propto \left(1+\frac{1}{2a_{\sigma_\beta}}\frac{\beta^2}{b_{\sigma_\beta}/a_{\sigma_\beta}}\right)^{-\left(1+2a_{\sigma_\beta}\right)/2}, \end{aligned} \tag{3-3}$$

因此我们可以得到边际分布 $\beta \sim \mathrm{t}_{2a_{\sigma_\beta}}\left(0,\sqrt{b_{\sigma_\beta}/a_{\sigma_\beta}}\right)$。特别地，当 $a_{\sigma_\beta}=0.5$ 时，β 服从 Cauchy 分布。在这个想法下，我们可以给出另外一种先验选择：

$$\begin{aligned} \epsilon|\sigma &\sim \mathrm{N}(0,\sigma I_n), \\ y|g,\beta,\sigma &\sim \mathrm{N}(g,\sigma I_n), \\ g|\beta,\tau, &\sim \mathrm{N}\left(0,K\right), \\ \beta_j|\sigma_{\beta_j} &\sim \mathrm{N}(0,\sigma_{\beta_j}), j=1,2,\cdots,p, \end{aligned} \tag{3-4}$$

$$\sigma_{\beta_j} \sim \text{InvGamma}(a_{\sigma_{\beta_j}}, b_{\sigma_{\beta_j}}), j = 1, 2, \cdots, p.$$

这里我们选择独立的 t 分布作为 β 的先验分布来代替式 3-2中的球面均匀分布，先验模型中的超参数仍然选择逆 Gamma 分布：$\sigma \sim \text{IG}(a_\sigma, b_\sigma)$，$\tau \sim \text{IG}(a_\tau, b_\tau)$。这样，我们得到了两种关于单指数回归模型的先验模型。

3.2 抽样与计算

在先验模型式 3-2、式 3-4下，我们利用 Metropolis-Hasting within PCG 的方法对参数进行抽样，模型后验分布的计算以及各个参数的条件后验分布展示在了附录小节 8.1中。从后验分布的结果中可见，式 3-2模型的参数中，g, σ 可以分别从正态分布和逆 Gamma 分布中直接抽样，β, τ, l 需要分别利用 Metropolis-Hasting 抽样方法实现抽样。但是由于高斯过程中的协方差矩阵 K 经常表现为病态矩阵，因此为了避免关于 K 的行列式与求逆的计算，我们分别将 β, τ, l 与参数 g 分组 (Blocking) 在一起，从组合的两个参数的联合条件分布中进行抽样。但是，为了避免关于 g 的重复抽样，我们对联合条件分布中的 g 积分 (Collapsing)，得到参数 β, τ, l 的边际条件后验分布，并从中对 β, τ, l 抽样。在 PCG 抽样的应用中，迭代抽样时参数的抽样顺序需要额外注意，否则可能会破坏马尔可夫链的平稳分布。在式 3-4模型中，同样 $g, \sigma, \sigma_{\beta_j}$ 可直接从相应的分布中抽样，β, τ 需要利用 Metropolis-Hasting 抽样实现。

关于 Metropolis-Hasting 抽样时建议分布的选择，在模型式 3-2中，单指数限制在 $\|\beta\|_2 = 1$ 上，新的 β 可以从 vMF 分布中抽取，即 $\beta^{new} \sim \text{vmf}(\kappa, \beta^t)$，其中 κ 是调节参数，β^t 是第 t 次迭代中得到的 β 的抽样结果。在模型式 3-4中没有 $\|\beta\|_2 = 1$ 的约束，我们可以直接利用正态分布随机游走来抽取新的 β 值，即 $\beta^{new} \sim \text{N}(\beta^t, \epsilon I_p)$。当抽取 τ 或者 l 时，由于其取值为正，所以这里我们取对数正态分布作为建议分布，即 $\log(x^{new}) \sim \text{N}(\log(x^t), \epsilon)$ 的方法抽取，其中 x 代表 τ 或者 l。

Patra 和 Dunson 提出在贝叶斯推断中，当参数限制在全空间中的某个子空间时，利用 MCMC 抽样的难度可能会大大增加，我们可以放松或者取消参数在子空间上的限制，在更大的的空间中选择更一般的先验分布代替原限制分布，然后从对应的后验分布中抽样，并在 MCMC 的每一步迭代中将结果投影到限制子空

间中[37]。利用此方法，在式 3-2先验模型中，当对 β 进行抽样时，我们可以取消先验分布 $\beta \sim \text{Uniform}(S^{p-1})$ 的限制，选择新的放松先验分布，从对应的后验分布中抽样，并在每一步迭代中将结果投影到单位球面上 S^{p-1} 上。在本文的实验里，我们选择单位立方体 $[-1,1]^p$ 上的均匀分布和 t 分布作为更大空间上的先验分布，这分别意味着选择无信息先验与弱信息先验。

抽样的过程中涉及许多具体的矩阵计算，但是高斯过程中的协方差矩阵 K 常常表现为病态矩阵，条件数大且接近奇异，因此在计算中我们需要避免关于 K 的直接计算。比如，当利用式 8-1抽取 g 的后验分布时，不直接计算 $(I_n/\sigma+K^{-1})^{-1}$，而是利用 Cholesky 分解的方法来解方程 $(K+\sigma I_n)x = K$ 或者 $(K+\sigma I_n)x = Ky$。其次，当处理其他正定矩阵的计算，如计算正态密度式 8-2中指数项里的 $y^T(K+\sigma I_n)^{-1}y$ 时，可以利用 Cholesky 分解 $K+\sigma I_n = LL^T$，解方程 $Lx = y$，最后 $y^T(K+\sigma I_n)^{-1}y = x^Tx$; 或者利用共轭梯度法解方程 $(K+\sigma I_n)x = y$，$y^T(K+\sigma I_n)^{-1}y = y^Tx$。当需要计算矩阵的行列式时，如式 8-2中的 $|K+\sigma I_n|$，可以通过 Cholesky 分解得到 $K+\sigma I_n = LL^T$，进而 $\log(|K+\sigma I_n|) = 2\sum_{i=1}^n L_{ii}$; 或者利用级数展开随机逼近的方法计算相应的行列式的 log 值[25, 38]。

3.3 数值结果

在本节中，我们将在模拟数据集和真实数据集上应用单指数回归模型式 3-2与式 3-4。在实际的单指数估计中，对于模型中超参数的先验逆 Gamma 分布中的参数，我们均选择 $(0.5, 0.5)$，即 $\theta \sim \text{InvGamma}(0.5, 0.5)$，其中 θ 为式 3-2中的 σ、τ 和 l，或者式 3-4中的 σ 和 τ。对于每个数据集，我们将样本 y 与 x 标准化，使其每个变量的样本均值为 0，样本方差为 1。在标准化变量下得到的单指数 β 的估计，只需要经过与标准化尺度相同的线性变换，再经过归一化处理，即可得到原变量下的回归系数。

在数值实验中，我们将用三种方法来对单指数 β 进行估计，分别是：

- 方法 1：基于先验模型式 3-2，利用 vMF 分布作为建议分布，来对 β 进行抽样。
- 方法 2：基于先验模型式 3-2，按照 Patra 和 Dunson 的方法，放松先验约束，利用抽样-投影的方法来对 β 进行抽样[37]。

- 方法 3：基于先验模型式 3-4来对 β 进行估计。

其中在方法 2 中，我们需要将 $\beta \sim \text{Uniform}(S^{p-1})$ 的先验分布放松到更大的空间中，这里我们分别选择单位立方体上的均匀分布先验 Uniform $([-1,1]^p)$ 和 t 分布族先验 $\beta_j|\sigma_{\beta_j} \sim \text{N}(0,\sigma_{\beta_j})$，$\sigma_{\beta_j} \sim \text{InvGamma}(a_{\sigma_{\beta_j}}, b_{\sigma_{\beta_j}})$，并查看不同先验选择对结果的影响。在实验中，我们均选择 $a_{\sigma_{\beta_j}} = b_{\sigma_{\beta_j}} = 0.5$，即尺度参数为 1 的 Cauchy 分布先验。

从后验分布中进行抽样，我们会得到模型中参数的抽样结果，每个参数都对应一条马尔可夫链。我们取每条链的后半部分抽样结果作为参数的估计，可以得到对应的后验期望估计。为了判断利用 MCMC 抽样方法得到的 β 的马尔可夫链是否收敛，我们将从三个独立的不同初始值出发，在模拟数据集上画图诊断其结果是否收敛，并在真实数据集上计算 PSRF 来帮助判断是否收敛。最大后验密度区间 (简记为 HPDI) 是贝叶斯推断中的长度最短的后验置信区间，在真实数据集中，我们还将计算每个单指数的 95%HPDI。

3.3.1 数值模拟

示例 1

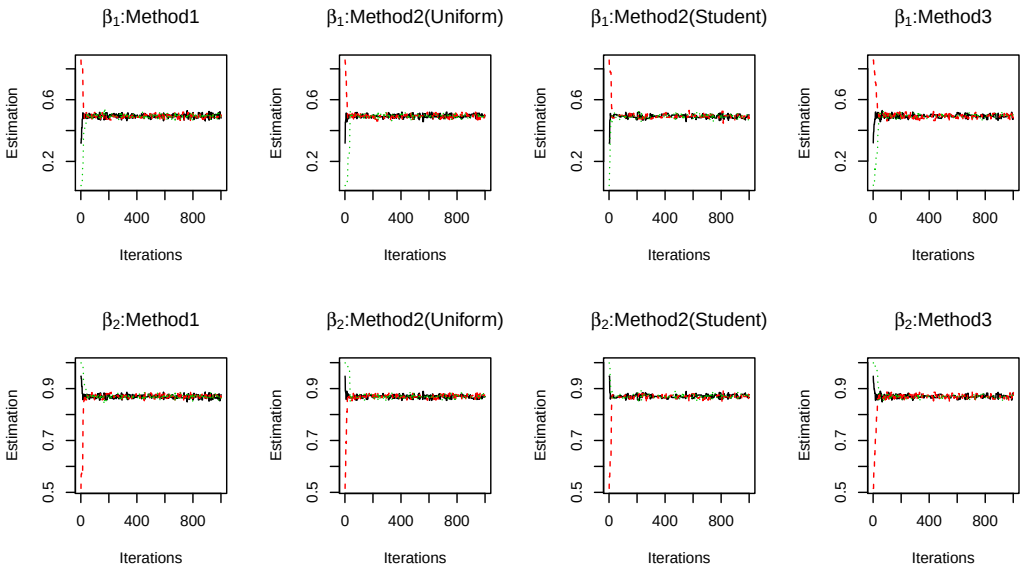

图 3.1: 示例 1 在样本容量 n=100 下的 β 估计结果。真实值 $\beta = (0.5, \sqrt{3}/2)^T$。

单指数回归模型的第一个数值模拟模型为：

$$g(t) = 2t^3 + 3\sin(0.5t)^3, t = X^T\beta,$$

$$Y = g(t) + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma),$$

$$\beta = (0.5, \frac{\sqrt{3}}{2})^T, \sigma = 0.01,$$

$$X_1 \sim \mathrm{Uniform}(-3, 5), X_2 \sim \mathrm{N}(0, 9).$$

n=100		示例 1		示例 2			
		β_1	β_2	β_1	β_2	β_3	β_4
方法 1	均值	.4930	.8699	.7995	.3982	-.3844	-.1989
	标准误	.0011	.0006	.0097	.0352	.0651	.0195
方法 2	均值	.4931	.8699	.7995	.3986	-.3987	-.2006
(Uniform)	标准误	.0013	.0008	.0104	.0151	.0073	.0081
方法 2	均值	.4930	.8700	.8020	.3993	-.3957	-.2011
(Student)	标准误	.0015	.0008	.0056	.0080	.0068	.0093
方法 3	均值	.4931	.8699	.7998	.3977	-.3984	-.2000
	标准误	.0015	.0009	.0084	.0213	.0085	.0107

表 3.1: 单指数回归模型示例 1 与示例 2 的数值结果：重复 100 次抽样得到 β 的后验期望估计的均值及其标准误差。

利用此模型，我们生成数据 $(y_i, x_i^T), i = 1, 2, \cdots, n$。在此模拟数据集上，我们利用不同方法来对单指数 β 进行估计。图 3.1是在样本容量 $n = 100$ 下，三个不同初始值的情况下，MCMC 抽样迭代 1000 次得到的 β 的估计结果。由其结果可见，每一个参数在不同初始值下得到的马尔可夫链都都快速收敛，并且其估计值与真实值 $(0.5, \sqrt{3}/2)^T$ 非常接近。不同的抽样方法与先验选择都表现良好。

然后，我们在 $n = 100$ 的样本容量下，随机生成不同的初始值，重复抽样过程 100 次，每个参数得到 100 条对应的马尔可夫链，其中每条链迭代 1000 次。我们取每条链的后 500 次抽样结果作为参数的估计，计算 β 的每个分量的后验期

望估计，并利用这 100 个后验期望估计值来计算估计量的样本期望与标准误差，其结果列在了表 3.1中。由结果可见，我们得到的后验期望估计与真实值非常接近，并且标准误差很小，每一种抽样方法都得到了良好的结果。

示例 2

单指数回归模型的第二个模拟数据集从下述模型中生成：

$$g(t) = 5\cos(t) + \exp(-t^2), t = X^T\beta,$$

$$Y = g(t) + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma),$$

$$\beta = (0.8, 0.4, -0.4, -0.2)^T, \sigma = 0.1,$$

$$X_1, X_2 \sim \mathrm{Uniform}(-3, 5), X_3 \sim \mathrm{N}(0, 9), X_4 \sim \chi^2(3).$$

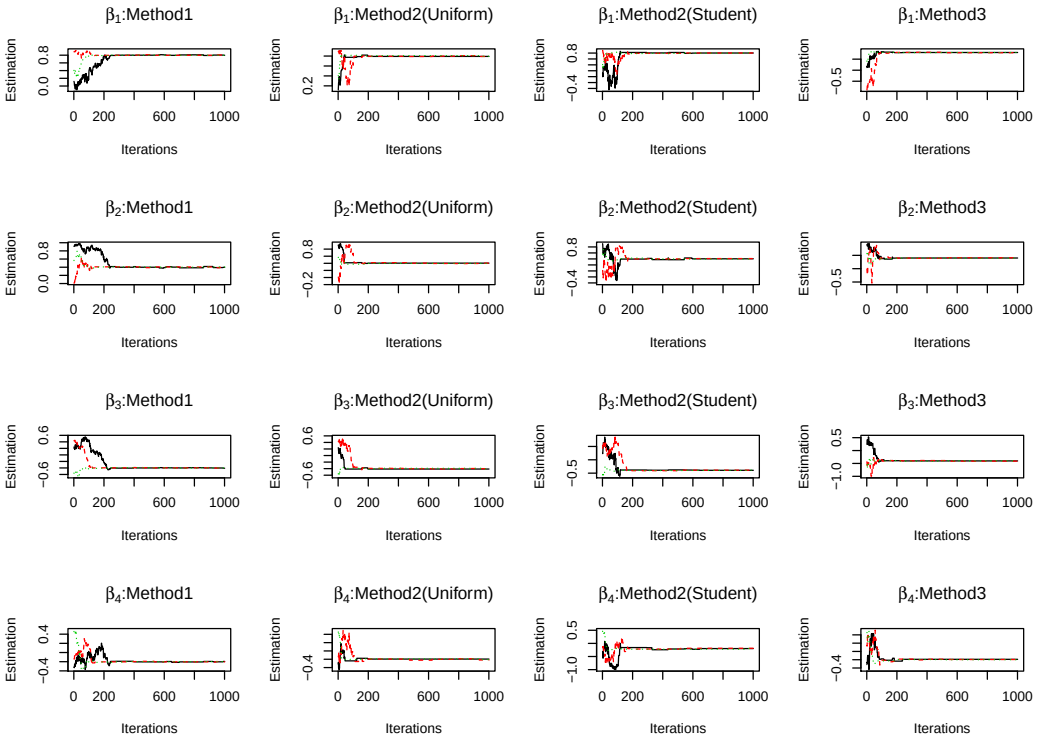

图 3.2: 示例 2 在样本容量 n=100 下的 β 估计结果。真实值 $\beta = (0.8, 0.4, -0.4, -0.2)^T$。

同样地，我们从此模型中生成样本容量为 $n = 100$ 的模拟数据，并在此数据集中应用我们的估计方法。图 3.2是在三个不同初始值下，MCMC 抽样迭代 1000

次得到的 β 抽样结果。由图上体现的结果可见，每个参数的每一条马尔可夫链都收敛，并且收敛速度快。不同的抽样方法与先验得到的结果并未体现出明显差异。表 3.1中列出了重复 100 次抽样得到的后验期望估计的均值与标准误差，由结果可见，每一种抽样方法都得到了良好的估计结果。

3.3.2 实例分析

示例 3

在本例中，我们在空气污染数据上应用我们的估计方法。此数据来源于 R 语言 lattice 包，包含 111 天中的平均臭氧浓度 (Ozone)、太阳辐射 (Radiation)、日最高气温 (Temperature) 和平均风速数据 (Wind)，其中臭氧浓度是响应变量。很多单指数回归模型的估计方法都应用于此数据集中[11, 14, 16, 19]。

		Radiation	Temperature	Wind
方法 1	估计	.304	.733	-.586
	HPDI(95%)	(.111 , .512)	(.580 , .888)	(-.777 , -.425)
	PSRF	1	1	1.02
方法 2 (Uniform)	估计	.285	.707	-.629
	HPDI(95%)	(.096 , .449)	(.561 , .859)	(-.786 , -.441)
	PSRF	1.02	1.03	1.03
方法 2 (Student)	估计	.316	.722	-.598
	HPDI(95%)	(.135 , .495)	(.550 , .866)	(-.777 , -.430)
	PSRF	1.03	1	1.01
方法 3	估计	.272	.744	-.593
	HPDI(95%)	(.119 , .451)	(.605 , .888)	(-.763 , -.433)
	PSRF	1.02	1	1

表 3.2: 示例 3 中空气污染数据的估计结果。

我们将每个变量标准化后，利用 MCMC 抽样方法迭代 1000 次，取后 500 次抽样结果作为参数的估计，取其样本均值作为后验期望估计，计算最大后验密度区间，并在三个不同初始值下得到不同的马尔可夫链，计算对应的 PSRF 的

值。数值结果总结在表 3.2中，由结果可见每一种方法都得到了收敛的马尔可夫链。Kong 和 Xia 得到的估计结果为 (0.3443 , 0.7051 , -0.6199)[14]，Zeng 等得到的估计结果为 (0.356 , 0.798 , -0.487)[16]，与我们得到的结果一致。由变量对应的单指数的符号可见，太阳辐射和温度对臭氧浓度的作用与风速相反。

		Sex1	Sex2	Length	Diameter	Height
方法 1	估计 (PSRF)	.215 (1.08)	.251 (1.1)	.188 (1.06)	-.115 (1.02)	-.039 (1.03)
	HPDI(95%)	(-.136 , .453)	(-.051 , .515)	(-.206 , .614)	(-.442 , .286)	(-.308 , .249)
方法 2	估计 (PSRF)	.128 (1.02)	.155 (1.02)	.047 (1.05)	-.017 (1.13)	-.070 (1.16)
(Uniform)	HPDI(95%)	(-.221 , .420)	(-.124 , .516)	(-.354 , .515)	(-.412 , .666)	(-.458 , .272)
方法 2	估计 (PSRF)	.139 (1.02)	.172 (1.02)	.037 (1.1)	-.040 (1.02)	-.090 (1.07)
(Student)	HPDI(95%)	(-.215 , .427)	(-.181 , .474)	(-.354 , .500)	(-.482 , .579)	(-.406 , .294)
方法 3	估计 (PSRF)	.177 (1.13)	.158 (1.12)	.172 (1.01)	-.338 (1.06)	.095 (1.2)
	HPDI(95%)	(-.070 , .312)	(-.062 , .302)	(-.045 , .507)	(-.505 , .331)	(-.440 , .390)

		Whole Wt.	Shucked Wt.	Viscera Wt.	Shell Wt.
方法 1	估计 (PSRF)	.357 (1.07)	.141 (1.03)	-.374 (1.07)	.535 (1.01)
	HPDI(95%)	(-.300 , .681)	(-.199 , .607)	(-.618 , -.007)	(.177 , .827)
方法 2	估计 (PSRF)	.387 (1.04)	.115 (1.08)	-.290 (1.02)	.586 (1.06)
(Uniform)	HPDI(95%)	(-.100 , .786)	(-.281 , .535)	(-.624 , .202)	(.354 , .908)
方法 2	估计 (PSRF)	.309 (1.13)	.150 (1.07)	-.268 (1.03)	.621 (1.07)
(Student)	HPDI(95%)	(-.155 , .618)	(-.158 , .481)	(-.551 , .171)	(.323 , .941)
方法 3	估计 (PSRF)	.280 (1.93)	.133 (1.16)	-.398 (1.33)	.600 (1.81)
	HPDI(95%)	(-.482 , .734)	(-.023 , .496)	(-.571 , .237)	(.269 , .858)

表 3.3: 示例 4 中鲍鱼数据的估计结果。

示例 4

在本例中，我们在鲍鱼数据上应用我们的单指数回归模型的单指数估计方法，此数据集来源于 UCI 机器学习数据库。此数据集共包含 9 个变量，样本容量 $n = 4177$，其中响应变量是鲍鱼的环数，代表着鲍鱼的年龄，解释变量是鲍鱼的 8 个物理指标测量，分别是性别 (M，F，Infant)，长度 (Length)，宽度 (Diameter)，厚度 (Height)，鲍鱼重量 (Whole Wt.)，鲍鱼肉的重量 (Shucked Wt.)，内脏重量 (Viscera Wt.)，外壳重量 (Shell Wt.)。由其物理意义可以看到，各个变量间相关性

很高。事实上，各个变量间的样本相关系数都在 $0.8-1$ 的范围内，具有很强的线性相关性，这增加了估计单指数的难度。

在实际应用中，我们取两个哑变量 (Sex1,Sex2) 代表性别变量，其中 (1,0)、(0,1)、(0,0) 分别代表 M、F、Infant，并将所有变量标准化。我们取样本容量 $n=100$ 的子集，利用我们的估计与抽样方法来估计该数据集的单指数。表 3.3是利用 MCMC 抽样方法迭代 2000 次，取后 1000 次抽样结果作为参数的估计得到的结果，其中 PSRF 是利用三个不同初始值下得到的马尔可夫链计算得到的结果。

由结果可见，参数估计的后验最大密度区间的长度较大，其原因可能在于解释变量间较高的相关性导致后验分布的方差较大。在不同的初始值下，由 PSRF 的计算结果可见，方法 1 和方法 2 的结果表现出良好的收敛性，而方法 3 的结果的收敛性较差。从单指数对应的解释变量的现实意义出发，个别回归系数的符号为负，尤其是宽度与厚度变量，这似乎与常识相悖。上述结果显示，在此数据集上，方法 1 和方法 2 优于方法 3，并且当样本相关系数很高时，我们可能需要对数据做预处理来降低解释变量间的强相关性对结果带来的不利影响。

4 单指数 Logistic 模型估计

4.1 基本模型

单指数 Logistic 回归模型假设：

$$Y \sim \text{Ber}(\pi),\ \pi = \frac{\exp\{g(X^T\beta)\}}{1+\exp\{g(X^T\beta)\}},$$

其中 X 是 p 维解释变量，单指数 $\beta \in \mathbb{R}^p$。假设 Y 利用 $\{1,-1\}$ 编码，且 $P(Y=1)=\pi$，那么在样本 $(y_i, x_i^T), i=1,2,\cdots,n$ 下，似然函数为：

$$p(y|g,\beta) = \prod_{i=1}^{n} \frac{\exp\{y_i g(x_i^T\beta)\}}{1+\exp\{y_i g(x_i^T\beta)\}}.$$

这样的似然函数为贝叶斯后验分布的计算与抽样增加了困难。在一般线性 Logistic 模型中，Polson 等提出了一种增加服从 Polya-Gamma 分布的辅助变量的方法，

利用 Gibbs 抽样来对 β 进行抽样[39]。其核心方法是利用如下等式：

$$\begin{aligned}\frac{(\exp\{\psi\})^a}{(1+\exp\{\psi\})^b} &= 2^{-b}\exp\{\kappa\psi\}\mathbb{E}\left(\exp\{-\omega\psi^2/2\}\right)\\ &= 2^{-b}\exp\{\kappa\psi\}\int_0^\infty \exp\{-\omega\psi^2/2\}p(\omega)d\omega,\end{aligned}$$

其中 $b>0$，$a\in\mathbb{R}$，$\kappa=a-b/2$，ω 服从参数为 $(b,0)$ 的 Polya-Gamma 分布，以下简记为 $\omega\sim \text{PG}(b,0)$。因此，在增加辅助变量 $\{\omega_i\}_{i=1}^n$ 的情况下，Logistic 模型中的似然函数可以写成正态密度的形式，这使 β 的抽样方法得以简化，同时新增辅助变量 ω_i 的条件后验分布依然服从 Polya-Gamma 分布。在本节中，我们采用 Polson 等的方法来处理单指数 Logistic 模型中的抽样问题，并且像节 3中那样，对未知函数 g 选择高斯过程先验，即 $g\sim \text{GP}(0,k)$，其中协方差函数 k 仍然选择式 2-1的形式。给定 β 单位球面上的均匀分布，这样我们可以得到单指数 Logistic 模型的先验模型：

$$\begin{aligned} y_i|g,\beta &\sim \text{Bernoulli}\left(\frac{\exp\{g(x_i^T\beta)\}}{1+\exp\{g(x_i^T\beta)\}}\right), i=1,2,\cdots,n,\\ g|\beta,\tau,l &\sim \text{N}\left(0,K\right),\\ \beta &\sim \text{Uniform}(S^{p-1}),\\ \omega_i &\overset{i.i.d.}{\sim} \text{PG}(1,0), i=1,2,\cdots,n,\end{aligned} \tag{4-1}$$

这里的超参数 τ,l 仍然选择逆 Gamma 分布的先验，即：$\tau\sim \text{InvGamma}(a_\tau,b_\tau)$，$l\sim \text{InvGamma}(a_l,b_l)$。

同样地，类似单指数回归模型中的先验模型式 3-4，将 $\beta/\sqrt{l}$ 视为整体，取消球面分布的约束，对 β 采用 t 分布族先验，我们可以得到另外一类先验模型：

$$\begin{aligned} y_i|g,\beta &\sim \text{Bernoulli}\left(\frac{\exp\{g(x_i^T\beta)\}}{1+\exp\{g(x_i^T\beta)\}}\right), i=1,2,\cdots,n,\\ g|\beta,\tau &\sim \text{N}\left(0,K\right),\\ \beta_j|\sigma_{\beta_j} &\sim \text{N}(0,\sigma_{\beta_j}), j=1,2,\cdots,p,\\ \sigma_{\beta_j} &\sim \text{InvGamma}(a_{\sigma_{\beta_j}},b_{\sigma_{\beta_j}}), j=1,2,\cdots,p,\\ \omega_i &\overset{iid}{\sim} \text{PG}(1,0), i=1,2,\cdots,n,\end{aligned} \tag{4-2}$$

其中超参数 $\tau\sim \text{InvGamma}(a_\tau,b_\tau)$。

4.2 抽样与计算

像小节 3.2中描述的那样，这里我们利用 Metropolis-Hasting within PCG 抽样的方法来对参数进行估计，其中式 4-1和式 4-2的后验分布和条件后验分布的计算过程展示在小节 8.2中。在先验模型式 4-1中，参数为 $(g, \beta, \omega, \tau, l)$，由条件分布可以得到，在抽样的过程中，$g$ 和 ω 可以直接从相应的条件分布中抽取，β, τ 和 l 需要和 g 分组在一起并对 g 积分，利用 Metropolis-Hasting 的方法来抽样。先验模型式 4-2中参数的抽样方法与小节 3.2中内容类似。而 Metropolis-Hasting 抽样中建议分布的选择同样采取小节 3.2中所描述的方式。

同理，在抽样中涉及矩阵计算的部分与小节 3.2中一致。首先，我们最好避免关于协方差矩阵 K 的直接求逆或行列式计算。其次，当需要处理正定矩阵求逆计算时，我们可以首先对正定矩阵做 Cholesky 分解，再解相关的方程组；或者利用共轭梯度优化的方法直接解关于正定矩阵的线性方程组。当需要计算正定矩阵的行列式时，我们可以采用矩阵分解方法或级数展开随机逼近的方法。

但是，在单指数 Logistic 模型中，并非所有的矩阵求逆的对象均是正定矩阵。在式 8-6中正态分布的期望与协方差矩阵中的 $K\Omega + I_n$ 并非是正定矩阵。事实上，由于 $\Omega = \text{diag}(\omega_1, \cdots, \omega_n)$，一般情况下 $K\Omega + I_n$ 并非对称矩阵。我们可以通过一般方法，如 LU 分解计算得到 $(K\Omega + I_n)x = K$ 的解。但是因为其特殊形式，我们有另外一种分解方法来直接计算得到协方差矩阵：

$$\begin{aligned}\Sigma_g =& (\Omega + K^{-1})^{-1} \\ =& K - K\Omega^{1/2}\left(I_n + \Omega^{1/2}K\Omega^{1/2}\right)^{-1}\Omega^{1/2}K \\ =& K - K\Omega^{1/2}B^{-1}\Omega^{1/2}K,\end{aligned}$$

其中矩阵 $B = I_n + \Omega^{1/2}K\Omega^{1/2}$ 是正定矩阵。这样的计算可以直接得到协方差矩阵 Σ_g，进而得到期望 μ_g。

4.3 数值结果

在本节中，同小节 3.3中内容一样，我们将分别在模拟数据集与真实数据集上应用我们的单指数 Logistic 模型的估计方法。在这里超参数逆 Gamma 分布的参数选择与小节 3.3一致，并且对解释变量做标准化处理。这里我们的分类响应

变量 Y 使用 $\{1,-1\}$ 编码。

同样地，这里我们分别应用小节 3.3中所描述的三种不同抽样方法来对 β 进行估计，其中 t 分布先验仍选择尺度参数为 1 的 Cauchy 分布。在抽样时，我们在不同的初始值下对 β 进行估计，得到多条 β 估计的链，由此利用作图或者计算 PSRF 的方式帮助判断我们得到的马尔可夫链是否收敛。在模拟数据集中，我们在不同初始值下重复多次抽样，利用每一条链的结果得到 β 的后验期望估计计算均值及其标准误差，并在真实数据集上计算参数的 95%HPDI。

4.3.1 数值模拟

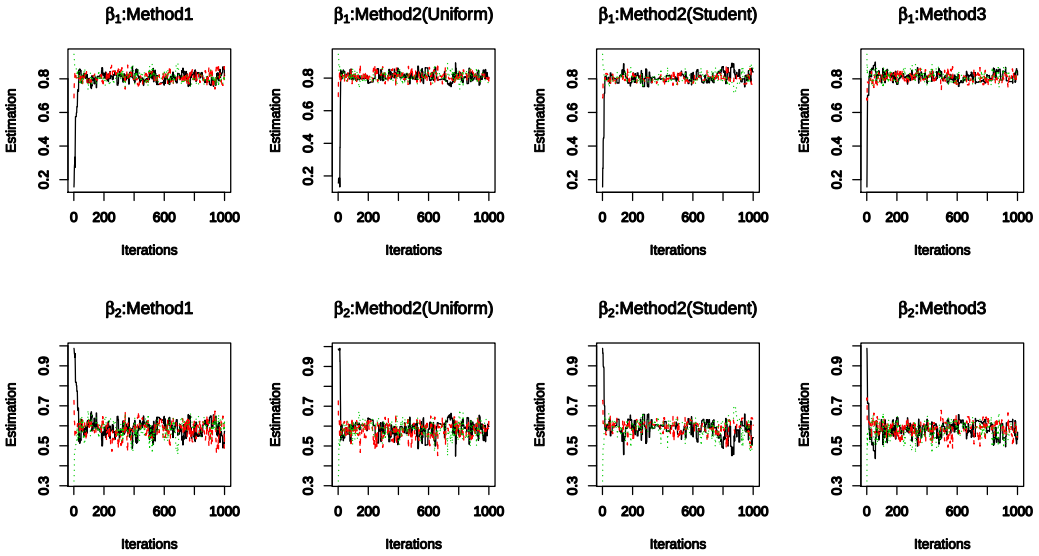

图 4.1: 示例 5 在样本容量 n=100 下 β 的估计结果。真实值 $\beta=(0.8,0.6)^T$。

示例 5

单指数 Logisitic 模型的第一个模拟来自于模型：

$$P(Y=1)=\pi, P(Y=-1)=1-\pi,$$

$$\pi=\frac{\exp\{g(t)\}}{1+\exp\{g(t)\}}, t=X^T\beta,$$

$$g(t)=t^2\sin t,$$

$$\beta=(0.8,0.6)^T, X_1\sim \text{Uniform}(0,8), X_2\sim\chi^2(4).$$

我们将从此模型中生成样本容量为 $n = 100$ 的模拟数据集。图 4.1是不同抽样方法在三个不同初始值下迭代 1000 次对 β 抽样的结果。由图中结果可见，每一条链均收敛并且收敛速度很快，并且估计结果良好。然后在 100 个不同的初始值下，用不同的抽样方法重复对 β 进行抽样，每一条链迭代 1000 次，取后 500 次抽样作为估计结果，得到参数的后验期望估计，并由 100 个后验估计计算其均值和标准误差，结果列在了表 4.1中。由结果可见，每一种抽样方法在此模拟数据集上都得到了很好的估计结果。

n=100		示例 5		示例 6			
		β_1	β_2	β_1	β_2	β_3	β_4
方法 1	均值	.8090	.5861	.5243	.4640	.5266	-.4612
	标准误	.0050	.0071	.0105	.0876	.0829	.0174
方法 2	均值	.8087	.5866	.5201	.4415	.4937	-.4370
(Uniform)	标准误	.0057	.0080	.0241	.1610	.1988	.1221
方法 2	均值	.8084	.5871	.5243	.4564	.5130	-.4441
(Student)	标准误	.0063	.0089	.0147	.1232	.1461	.0993
方法 3	均值	.8070	.5889	.5228	.4557	.5061	-.4401
	标准误	.0063	.0087	.0188	.1009	.1574	.1219

表 4.1: 单指数 Logistic 模型示例 5 和示例 6 的数值模拟结果：重复 100 次的后验期望的均值及其标准误差。

示例 6

单指数 Logistic 模型的第二个模拟模型来自于：

$$P(Y = 1) = \pi, P(Y = -1) = 1 - \pi,$$

$$\pi = \frac{\exp\{g(t)\}}{1 + \exp\{g(t)\}}, t = X^T\beta,$$

$$g(t) = t\cos t,$$

$$\beta = (0.5, 0.5, 0.5, -0.5)^T,$$

$$X_1 \sim \text{Uniform}(0, 8), X_2 \sim \text{N}(3, 9), X_3 \sim \chi^2(4), X_4 \sim \text{F}(2, 5).$$

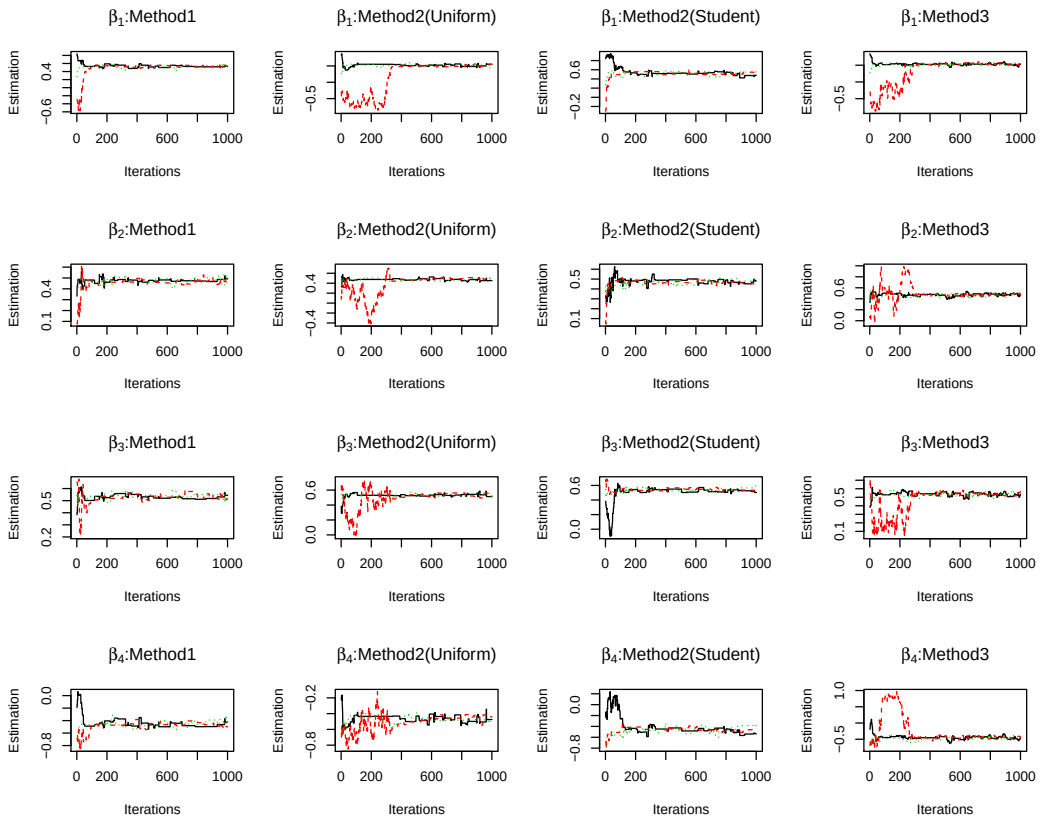

图 4.2: 示例 6 在样本容量 n=100 下的 β 估计结果。真实值 $\beta = (0.5, 0.5, 0.5, -0.5)^T$。

与前面的模拟数据集一样，我们在此模型上生成样本容量 $n = 100$ 的模拟数据集。图 4.2是在模拟数据集上，在三个不同初始值下利用不同方法得到的 β 的估计。由结果可见，尽管有些马尔可夫链的收敛速度相对较慢，但是仍在 500 次迭代以内达到收敛状态。然后同样地，我们在 100 个不同初始值下重复抽样 100 次，每个参数得到 100 条估计的马尔可夫链，基于每一条链的结果计算参数的后验期望估计值，并利用这 100 个结果计算其均值和标准误差，表 4.1中列出了这些结果。尽管相对于示例 1、2、5 的结果来说，此例中有些参数估计的标准误差相对较大，但是每一种方法仍得到了良好的估计。

4.3.2 实例分析

示例 7

在本例中我们在一个汽车碰撞数据上应用我们的单指数 Logistic 模型的估

计方法，此数据集最初被 Hardle 和 Stoker 所研究[5]。此数据集包含 4 个变量，样本容量为 $n = 58$，解释变量分别是汽车车龄 (AGE)、车速 (VEL) 和最大加速度 (ACL)，响应变量是汽车事故是否是致命的。在将解释变量标准化后，我们用不同的方法，利用 MCMC 抽样方法迭代 1000 次，取后 500 次抽样结果作为参数的估计，计算其后验期望估计和 95% 最大后验密度区间，并在三个不同初始值下重复抽样，利用 3 条马尔可夫链的抽样结果计算 PSRF 的值来帮助判断链的收敛性。表 4.2是此数据集的估计结果，由结果可见，每一种方法都得到了良好的估计结果。Hardle 和 Stoker 基于平均导数估计的方法得到的 β 的估计结果为 $(0.89, 0.34, 0.30)$[5]，Hardle 等基于 M 估计方法得到的估计结果为 $(0.3, 0.3, 0.9)$[3]，可见我们的结果与平均导数估计方法得到的估计值更接近。事实上，基于 M 估计方法的结果更加依赖优化过程中的初始值和优化算法的选择，因此在这个方面我们的贝叶斯估计方法更具有优势。

		AGE	VEL	ACL
方法 1	估计	.888	.227	.289
	HPDI(95%)	(.762 , .987)	(-.146 , .564)	(-.092 , .639)
	PSRF	1.01	1.01	1.01
方法 2 (Uniform)	估计	.875	.299	.255
	HPDI(95%)	(.734 , .972)	(-.112 , .628)	(-.102 , .688)
	PSRF	1.04	1.03	1.02
方法 2 (Student)	估计	.902	.243	.252
	HPDI(95%)	(.788 , 972.)	(-.074 , .610)	(-.116 , .502)
	PSRF	1.12	1.04	1.04
方法 3	估计	.911	.169	.283
	HPDI(95%)	(.746 , .997)	(-.123 , .574)	(-.060 , .680)
	PSRF	1.01	1.07	1.07

表 4.2: 示例 7 中汽车碰撞数据的估计结果。

5 单指数模型变量选择

5.1 基本模型

在实际应用中，我们常常需要筛选与响应变量 Y 有关的解释变量。在本节中，我们聚焦于单指数回归模型和单指数 Logistic 模型的稀疏单指数估计方法。

Spike-Slab 先验模型是经典的一般线性回归中的贝叶斯变量选择方法。对于 p 维回归系数 $\beta \in \mathbb{R}^p$，Spike-Slab 模型通过引入指标变量 $\delta = (\delta_1, \cdots, \delta_p)$ 来表示对应的解释变量是否应包含在模型中，而 β_j 的先验分布取决于其对应变量是否包含在模型当中。通常，$\delta_j = 1$ 表示 β_j 应包含在模型中，此时 β_j 的先验分布称为 Slab 分布，$\delta_j = 0$ 表示 β_j 不包含在模型中，此时 β_j 的先验分布称为 Spike 分布。因此，我们可以将 Spike-Slab 模型写成以下形式：

$$p(\beta_j|\delta_j) = (1-\delta_j)p_{Spike}(\beta_j) + \delta_j p_{Slab}(\beta_j).$$

很多文章研究了 Spike 分布与 Slab 分布的选择问题。由于当 $\delta_j = 0$ 时，β_j 不包含在模型中，此时得到的回归系数的估计应该为 $\beta_j = 0$，因此一个直接的 Spike 先验是 $P(\beta_j = 0|\delta_j = 0) = 1$。记 $\delta_0(x)$ 为 Dirac 函数，则 Spike Dirac 先验的密度可以表示成 $p_{Spike}(\beta_j) = \delta_0(\beta_j)$ 的形式，而正态分布 $\mathrm{N}(\mu, \Sigma)$ 是 Slab 分布的常见选择[40, 41]。与 Dirac 先验分布不同，为了抽样方法的简洁，George 和 Mcculloch 提出用不同方差的零期望的正态分布族表示 Spike-Slab 先验分布，其中 $\mathrm{Var}_{Spike}(\beta_j)/\mathrm{Var}_{Slab}(\beta_j) \ll 1$[42]。这种先验模型可以写成 $\beta_j|\delta_j, \sigma_j \sim \mathrm{N}(0, r(\delta_j)\sigma_j)$ 的形式，其中 $r(0) = c \ll 1$，$r(1) = 1$。

在节 3和节 4中，我们用不同的方法来实现对 $\beta \in S^{p-1}$ 的抽样，并且大多数的数值实验结果表明这三种方法并没有明显的优劣。当我们需要结合变量选择方法估计单指数时，我们需要在 β 的先验分布中添加一定的先验信息，使得我们可以得到 β 的稀疏估计。这里我们采用 Patra 和 Dunson 的方法，放松参数的先验约束，采用更大空间上的先验分布，得到的后验分布并利用 MCMC 方法抽样，同时在每一步抽样后将抽样结果投影到 S^{p-1} 上[37]。出于对实现单指数模型中 β 的抽样方法的考虑，我们采用 George 和 Mcculloch 的先验模型，来代替原 S^{p-1} 上的某种具有稀疏信息的先验分布，对 β 设定同分布族异方差的 Spike-Slab 先验，来实现 β 的变量选择[42]。与节 3和节 4中的内容保持一致，我们利用

正态分布-逆 Gamma 分布的混合先验来实现设定 t 分布族先验。利用与式 3-3类似的推导，先验模型 $\beta_j|\delta_j,\sigma_{\beta_j}\sim \mathrm{N}(0,r(\delta_j)\sigma_{\beta_j})$，$\sigma_{\beta_j}\sim \mathrm{InvGamma}(a_{\sigma_{\beta_j}},b_{\sigma_{\beta_j}})$ 与 $\beta_j|\delta_j\sim \mathrm{t}_{2a_{\sigma_{\beta_j}}}\left(0,\sqrt{r(\delta_j)b_{\sigma_{\beta_j}}/a_{\sigma_{\beta_j}}}\right)$ 的先验模型是等价的。

一些研究中假设指标变量 δ 服从相同的先验分布，即 $\delta_j\sim \mathrm{Bernoulli}(\pi), j=1,2,\cdots,p$。在本文中，我们假设 $\delta_j\sim \mathrm{Ber}(\pi_j), j=1,2,\cdots,p$。由上述模型与方法，我们可以得到单指数回归模型的变量选择先验模型：

$$
\begin{aligned}
\epsilon|\sigma &\sim \mathrm{N}(0,\sigma I_n),\\
y|g,\beta,\sigma &\sim \mathrm{N}(g,\sigma I_n),\\
g|\beta,\tau,l &\sim \mathrm{N}\left(0,K\right),\\
\beta_j|\sigma_{\beta_j},\delta_j &\sim \mathrm{N}(0,r(\delta_j)\sigma_{\beta_j}), j=1,2,\cdots,p\\
\sigma_{\beta_j} &\sim \mathrm{InvGamma}(a_{\sigma_{\beta_j}},b_{\sigma_{\beta_j}}), j=1,2,\cdots,p\\
\delta_j|\pi_j &\sim \mathrm{Bernoulli}(\pi_j), j=1,2,\cdots,p,
\end{aligned}
\tag{5-1}
$$

其中 $r(0)=c\ll 1$，$r(1)=1$，超参数选择 $\pi_j\sim \mathrm{Beta}(a_{\pi_j},b_{\pi_j}), j=1,2,\cdots,p$，$\sigma\sim \mathrm{InvGamma}(a_\sigma,b_\sigma)$，$\tau\sim \mathrm{InvGamma}(a_\tau,b_\tau)$，$l\sim \mathrm{InvGamma}(a_l,b_l)$ 的先验模型。

同理，结合 Spike-Slab 先验与节 4中描述的 Logistic 模型抽样方法，我们可以得到单指数 Logistic 模型的变量选择先验模型：

$$
\begin{aligned}
y_i|g,\beta &\sim \mathrm{Bernoulli}\left(\frac{\exp\{g(x_i^T\beta)\}}{1+\exp\{g(x_i^T\beta)\}}\right), i=1,2,\cdots,n,\\
g|\beta,\tau,l &\sim \mathrm{N}\left(0,K\right),\\
\beta_j|\sigma_{\beta_j},\delta_j &\sim \mathrm{N}(0,r(\delta_j)\sigma_{\beta_j}), j=1,2,\cdots,p,\\
\sigma_{\beta_j} &\sim \mathrm{InvGamma}(a_{\sigma_{\beta_j}},b_{\sigma_{\beta_j}}), j=1,2,\cdots,p,\\
\delta_j|\pi_j &\sim \mathrm{Bernoulli}(\pi_j), j=1,2,\cdots,p,\\
\omega_i &\overset{i.i.d.}{\sim} \mathrm{PG}(1,0), i=1,2,\cdots,n,
\end{aligned}
\tag{5-2}
$$

其中 $r(0)=c\ll 1$，$r(1)=1$，超参数选择 $\pi_j\sim \mathrm{Beta}(a_{\pi_j},b_{\pi_j}), j=1,2,\cdots,p$，$\tau\sim \mathrm{InvGamma}(a_\tau,b_\tau)$，$l\sim \mathrm{InvGamma}(a_l,b_l)$。

由上述单指数模型的变量选择先验模型式 5-1和式 5-2，我们可以得到对应的后验分布，利用 Metropolis-Hasting within PCG 抽样的方法，我们可以对 β 和

δ 进行抽样。后验分布和条件分布的推导展示在了小节 8.3中。在抽样过程中的抽样细节和计算问题上的处理与节 3和节 4中的内容中一致，因此这里不再赘述，具体的实现过程可以参考小节 3.2与小节 4.2的内容。

5.2 数值结果

我们将分别在模拟数据集和真实数据集上估计模型中的 β 和 δ，并利用 δ 的估计结果判断对应的变量是否包含在模型当中。一般来说，在得到 δ 的估计后，如果 $\delta_j = 1$ 的频率大于 0.5，可以认为对应的变量 X_j 应该选入模型中。在实际应用中，我们仍对 σ，τ，l 的逆 Gamma 先验分布的参数设定 $(0.5, 0.5)$，并对 Slab 先验设定尺度参数为 1 的 Cauchy 分布。π_j 服从无信息 Beta 先验分布，即 $a_{\pi_j} = b_{\pi_j} = 1$。初始值 $\delta_j = 1$，$\pi_j = 0.5$，$j = 1, 2, \cdots, p$。

参数 $c = r(0)$ 的设定时需要根据数据集的不同来选择，如果选择的 c 过小，那么 δ 的马尔可夫链很难在 0，1 状态间跳跃，如果选择的 c 过大，那么很难起到变量选择的作用。在实际操作中，我们发现相对于单指数回归变量选择模型，单指数 Logistic 变量选择模型对 c 的取值更加敏感。在模拟数据集上，选择 $c = 1/1000$ 都取得了良好的估计结果。

5.2.1 数值模拟

示例 8

单指数回归模型的变量选择方法的模拟数值实验来自下述模型：

$$g(t) = 2\sin(t) + \exp(-t/10), t = X^T\beta,$$

$$Y = g(t) + \epsilon, \epsilon \sim \mathrm{N}(0, \sigma), \sigma = 0.01,$$

$$\beta^T = (2, 2, 1, -1, 0, 0, 0, 0, 0, 0)/\sqrt{10},$$

$$X \sim \mathrm{N}(0, I_{10}).$$

此模型中，β_1 和 β_2 对应的变量意味着强效应变量，β_3 和 β_4 对应的变量意味着弱效应变量，其他变量为零效应变量。我们基于此模型，生成样本容量 $n = 100$ 的模拟数据集，在此数据集上应用我们的变量选择先验模型式 5-1，利用 MCMC

抽样算法迭代抽样 2000 次，并取后 1000 次抽样结果作为参数的估计。然后我们在 100 个不同的初始值下，对 β 和 δ 重复抽样 100 次，统计每一次抽样得到的变量选择的结果。图 5.1是上述抽样过程的结果，为节省文章篇幅，这里不再展示马尔可夫链的收敛性。由结果显示，在此模拟数据集上，每一次抽样都得到了良好的估计结果，可以区别有效应的变量和无效应的变量。由图 5.1中第三张图 δ_j 的估计结果可见，强效应的估计比弱效应的估计更加稳健，但是仍然每次都正确选择了这些变量。

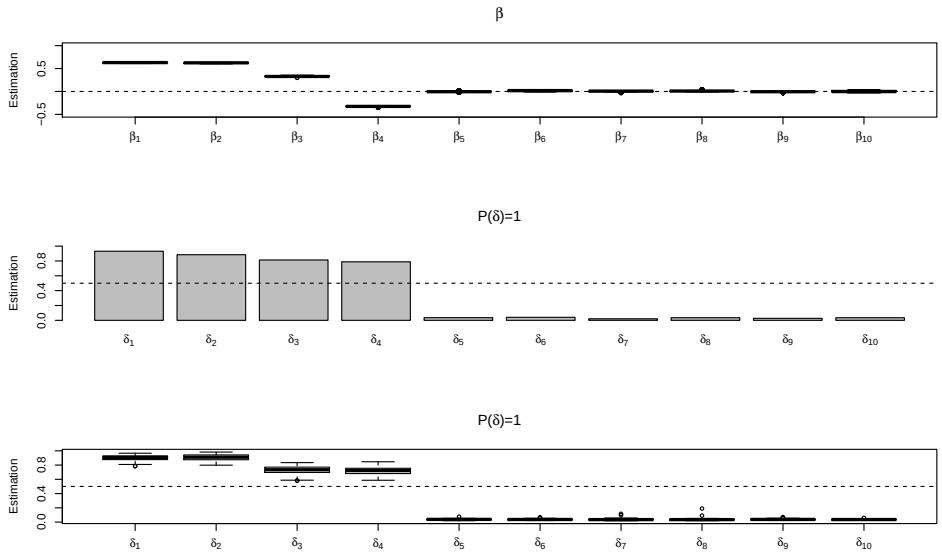

图 5.1: 示例 8 在样本容量 n=100 下的估计结果。第一张图是一次抽样得到的 β 的估计，第二张图是一次抽样得到的 $\delta = 1$ 的频率，第三张图是重复抽样 100 次得到的 $\delta = 1$ 的频率的箱线图。

示例 9

单指数 Logistic 变量选择模型的模拟数据集从下述模型中生成：

$$g(t) = 4(t + \sin(0.5t^3)), t = X^T\beta,$$

$$Y \sim \text{Bernoulli}\left(\frac{\exp\{g(t)\}}{1+\exp\{g(t)\}}\right),$$

$$\beta^T = (2, 2, 1, 1, 0, 0, 0, 0, 0, 0)/\sqrt{10},$$

$$X \sim \text{N}(0, I_{10}).$$

我们按上述单指数 Logistic 模型生成 $n = 100$ 个样本，应用变量选择模型式 5-2，得到单指数 β 和指标 δ 的估计结果。本例中我们利用 MCMC 抽样算法迭代 2000 次得到参数的抽样结果，取后 1000 次结果作为参数的估计，并计算 $\delta = 1$ 的频率。然后在 100 个不同的初始值下，重复抽样实验 100 次，得到 100 个 $\delta = 1$ 的频率。

关于 β 和 δ 的估计结果展示在图 5.2中，其中第一张图是一次抽样得到的 β 的估计，第二张图是一次抽样得到的 $\delta = 1$ 的频率，第三张图是重复抽样 100 次得到的 $\delta = 1$ 的频率的箱线图。由第一张图和第二张图的结果显示，β 和 δ 都得到了良好的估计。在重复抽样 100 次后，可以看到少数抽样结果未正确选择合适的变量，并且强效应变量的选择比弱效应变量的选择更加稳健。从总体上看，单指数 Logistic 变量选择模型在此模拟数据集上得到了良好的估计和变量选择结果。

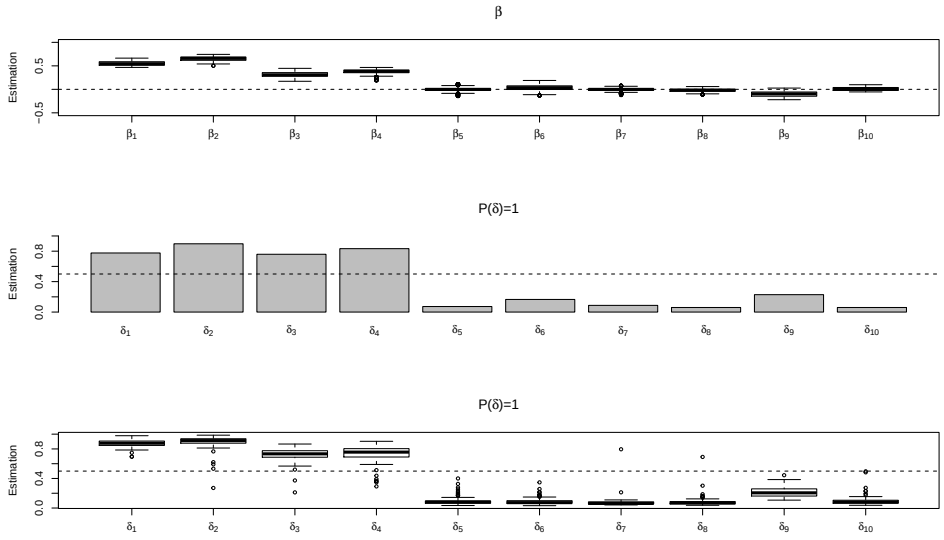

图 5.2: 示例 9 在样本容量 n=100 下的估计结果。第一张图是一次抽样得到的 β 的估计，第二张图是一次抽样得到的 $\delta = 1$ 的频率，第三张图是重复抽样 100 次得到的 $\delta = 1$ 的频率的箱线图。

5.2.2 实例分析

示例 10

在此例中，我们应用单指数回归的变量选择方法于前列腺癌数据集，此数据集包含在 R 语言 bayesQR 包中。此数据集包含 9 个变量，样本容量 $n = 97$，其中响应变量是前列腺特异性抗原的水平的对数 (lpsa)，解释变量是一系列临床测量数据，分别是 X_1(lcavol)，X_2(lweight)，X_3(age)，X_4(lbph)，X_5(svi)，X_6(lcp)，X_7(gleason)，X_8(pgg45)。这里不再详细描述这些指标的具体现实含义，其现实含义可见 bayesQR 包中的叙述。图 5.3是此数据集各个变量间的样本相关系数，由结果可以看到，一些解释变量间，如 X_1 和 X_6、X_5 和 X_6、X_7 和 X_8 之间具有较强的相关性，并且变量 X_1、X_5 和 X_6 和响应变量 lpsa 之间相关性较强。

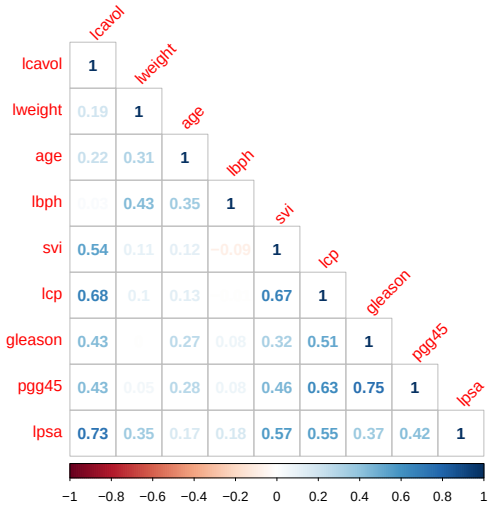

图 5.3: 示例 10 中前列腺癌数据变量间的样本相关系数

变量 X_j	1,2,5	1,4,5	1,5	1,2,5,7	1,2,4,5	1,2	1,2,5,6	1,2,5,8	其它
次数	39	9	7	7	6	6	5	4	17

表 5.1: 示例 10 中重复 100 次抽样的变量选择结果。

此例中，我们选择参数 $c = 1/10000$，在 100 个不同的初始值下，利用 MCMC 方法重复抽样 100 次，每次抽样迭代 4000 次，并取后 2000 次抽样结果作为参数的估计。利用这 100 次抽样结果，基于 $P(\delta_j = 1) > 0.5$ 的准则来判断对应变量 X_j 是否包含在模型当中。表 5.1是重复抽样后变量选择的结果。由结果可见，变量 X_1(lcavol)、X_2(lweight) 和 X_5(svi) 是被选择次数最多的模型，并且变量选择

并非直接根据响应变量与解释变量之间的相关性的强弱来选择，同时可以将一些相关强的解释变量分别开来。我们的变量选择方法在此数据集上展现了良好的结果。

示例 11

在此例中，我们在 Boston 房价数据集上应用我们的单指数 Logistic 模型变量选择模型。此数据集包含在 R 语言 MASS 包中，共包含 14 个变量，样本容量 $n = 506$，其中响应变量为街区房价的中位数 (medv)，解释变量分别是 X_1(crim)，X_2(zn)，X_3(indus)，X_4(chas)，X_5(nox)，X_6(rm)，X_7(age)，X_8(dis)，X_9(rad)，X_{10}(tax)，X_{11}(ptratio)，X_{12}(black)，X_{13}(lstat)，这里不再详细介绍各个解释变量的具体现实含义。此数据集中 medv 变量可以看作是连续变量，因此我们按照 medv 变量是否高于中位数将其转变为二分类变量。图 5.4是各个变量间的样本相关系数，由结果可见一些变量间具有很强的线性相关性。

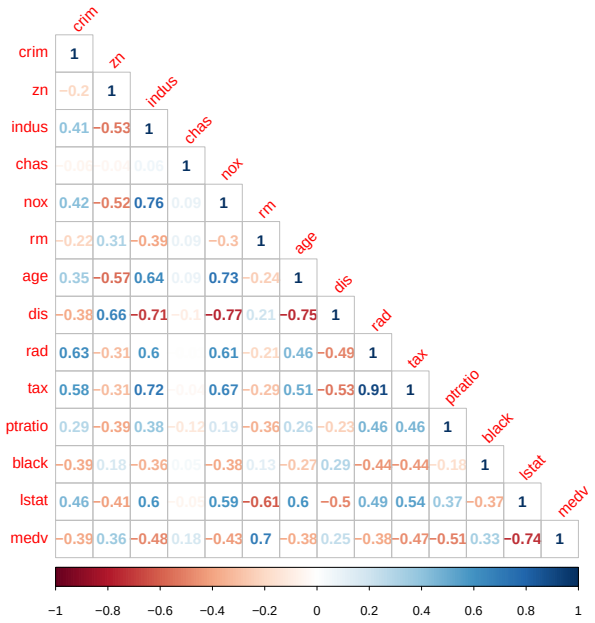

图 5.4: 示例 11 中 Boston 数据变量间的样本相关系数

在此例中，我们取 $n = 150$ 的子集，取 $c = 1/1000$，在 100 个不同的初始值下，利用 MCMC 抽样算法重复抽样 100 次，每次迭代抽样 4000 次，并取每条链的后 2000 次作为后验抽样结果。表 5.2是 100 次抽样变量选择的结果，变量 X_6(rm)，X_{11}(ptratio)，X_{13}(lstat) 和 X_6(rm)，X_{13}(lstat) 是选择次数最多的变量组

合。由结果可见，在此数据集上，我们的单指数 Logistic 变量选择方法得到了良好的结果。

变量 X_j	6,11,13	6,13	6,9,11,13	4,13	6,8,13	9,13	6,10,13	其它
次数	30	29	11	9	6	2	2	11

表 5.2: 示例 11 中重复 100 次抽样的变量选择结果。

6 总结

在本文中，聚焦于单指数模型中的回归模型和 Logistic 模型，我们发展了单指数的贝叶斯估计方法。对于单指数模型中的未知函数 g，我们应用高斯过程先验模型，假设在给定数据下，函数 g 作为随机过程，其有限维分布服从联合正态分布，并选择高斯型核函数作为协方差函数。一些文献工作中已经涉及利用高斯过程回归的方法来解决单指数回归中的参数估计问题。在我们的工作中，我们通过对高斯过程中的参数设定先验分布，来解决超参数的取值问题。并且我们实现并比较了多种关于单指数 β 的先验模型及其抽样方式，其中包括单位球面均匀分布先验、单位立方体均匀分布和 t 分布族的放松先验，以及无约束先验。

Logistic 模型的贝叶斯估计方法一直以来是一个具有挑战性的问题。基于增加服从 Polya-Gamma 分布的辅助变量的 Gibbs 抽样方法极大地方便了一般线性 Logistic 模型中回归系数的抽样过程，我们结合基于高斯过程的估计方法与增加辅助变量的方式，发展了单指数 Logistic 模型的贝叶斯估计方法。同样地，我们也考虑了 β 的不同先验模型和抽样方法。

Spike-Slab 先验是一种经典的用于一般线性模型变量选择的贝叶斯方法。此模型假设当变量包含在模型当中时，对应的系数服从 Slab 先验分布；如果不包含在模型中，对应的系数服从 Spike 先验分布。我们将 Spike-Slab 变量选择方法与上述单指数贝叶斯估计方法相结合，发展了单指数回归模型和单指数 Logistic 模型的变量选择方法。

高斯过程中的协方差矩阵在数值上常常表现病态。为避免矩阵计算上可能出现的数值问题，我们利用 Metropolis-Hasting within PCG 抽样的方法，来对单

指数模型中的参数进行抽样。当使用 PCG 抽样方法时，在一次迭代中参数的抽样顺序不是可以随意调整的，不合适的抽样顺序可能会破坏马尔可夫链的平稳分布，导致其平稳分布与我们的目标分布并不一致。

我们在一系列模拟数据集上和真实数据集上应用了我们的估计方法，并且表现出了良好的结果。但是，由于高斯过程方法涉及到求解规模为 $n \times n$ 的矩阵的线性方程组和行列式的值，其中 n 是数据集的样本量，因此这给我们在计算上带来了一定的负担。常规的求解此类问题的方法的算法复杂度为 $O(n^3)$，因此当样本容量变得比较大时，我们的贝叶斯估计方法在时间上的效率将会变得比较低。一系列快速高斯过程的研究降低了处理协方差矩阵相关问题的算法复杂度，但是在本文中我们并未将此类研究成果与我们的估计方法相结合。在将来的应用中，我们可以尝试将快速高斯过程的方法与我们的贝叶斯估计相结合，使得我们的方法在计算上更加具有效率。

7 参考文献

[1] FRIEDMAN J H, STUETZLE W. Projection pursuit regression[J]. Journal of the American statistical Association, 1981, 76(376): 817-823.

[2] ICHIMURA H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models[J]. Journal of Econometrics, 1993, 58(1-2): 71-120.

[3] HARDLE W, HALL P, ICHIMURA H. Optimal smoothing in single-index models[J]. Annals of Statistics, 1993: 157-178.

[4] XIA Y, TONG H, LI W K, et al. An adaptive estimation of dimension reduction space[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2002, 64(3): 363-410.

[5] HÄRDLE W, STOKER T M. Investigating smooth multiple regression by the method of average derivatives[J]. Journal of the American statistical Association, 1989, 84(408): 986-995.

[6] HOROWITZ J L, HÄRDLE W. Direct semiparametric estimation of single-index models with discrete covariates[J]. Journal of the American Statistical Association, 1996, 91(436): 1632-1640.

[7] HRISTACHE M, JUDITSKY A, SPOKOINY V. Direct estimation of the index coefficient in a single-index model[J]. Annals of Statistics, 2001: 595-623.

[8] LI K C. Sliced inverse regression for dimension reduction[J]. Journal of the American Statistical Association, 1991, 86(414): 316-327.

[9] COOK R D, WEISBERG S. Sliced inverse regression for dimension reduction: Comment[J]. Journal of the American Statistical Association, 1991, 86(414): 328-332.

[10] YIN X, COOK R D. Direction estimation in single-index regressions[J]. Biometrika, 2005, 92(2): 371-384.

[11] ANTONIADIS A, GRÉGOIRE G, MCKEAGUE I W. Bayesian estimation in single-index models[J]. Statistica Sinica, 2004: 1147-1164.

[12] CHOI T, SHI J Q, WANG B. A Gaussian process regression approach to a single-index model[J]. Journal of Nonparametric Statistics, 2011, 23(1): 21-36.

[13] GRAMACY R B, LIAN H. Gaussian process single-index models as emulators for computer experiments[J]. Technometrics, 2012, 54(1): 30-41.

[14] KONG E, XIA Y. Variable selection for the single-index model[J]. Biometrika, 2007, 94(1): 217-229.

[15] WANG Q, YIN X. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE[J]. Computational Statistics & Data Analysis, 2008, 52(9): 4512-4520.

[16] ZENG P, HE T, ZHU Y. A lasso-type approach for estimation and variable selection in single index models[J]. Journal of Computational and Graphical Statistics, 2012, 21(1): 92-109.

[17] 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.

[18] PENG H, HUANG T. Penalized least squares for single index models[J]. Journal of Statistical Planning and Inference, 2011, 141(4): 1362-1379.

[19] WANG H B. Bayesian estimation and variable selection for single index models[J]. Computational Statistics & Data Analysis, 2009, 53(7): 2617-2627.

[20] WILLIAMS C K, RASMUSSEN C E. Gaussian processes for machine learning[M]. Cambridge, MA: MIT press, 2006.

[21] WILLIAMS C K, SEEGER M. Using the Nyström method to speed up kernel machines[C]//Advances in neural information processing systems. Cambridge, MA: MIT press, 2001: 682-688.

[22] HERBRICH R, LAWRENCE N D, SEEGER M. Fast sparse Gaussian process methods: The informative vector machine[C]//Advances in neural information processing systems. Cambridge, MA: MIT press, 2003: 625-632.

[23] SNELSON E, GHAHRAMANI Z. Sparse Gaussian processes using pseudo-inputs[C]//Advances in neural information processing systems. Cambridge, MA: MIT press, 2006: 1257-1264.

[24] QUIÑONERO-CANDELA J, RASMUSSEN C E. A unifying view of sparse approximate Gaussian process regression[J]. Journal of Machine Learning Research, 2005, 6(Dec): 1939-1959.

[25] ZHANG Y, LEITHEAD W E. Approximate implementation of the logarithm of the matrix determinant in Gaussian process regression[J]. Journal of Statistical Computation and Simulation, 2007, 77(4): 329-348.

[26] HAN I, MALIOUTOV D, SHIN J. Large-scale log-determinant computation through stochastic Chebyshev expansions[C]//International Conference on Machine Learning. New York, NY: ACM, 2015: 908-917.

[27] DONG K, ERIKSSON D, NICKISCH H, et al. Scalable log determinants for Gaussian process kernel learning[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT press, 2017: 6327-6337.

[28] LAWRENCE N D. Gaussian process latent variable models for visualisation of high dimensional data[C]//Advances in neural information processing systems. Cambridge, MA: MIT press, 2004: 329-336.

[29] LIU J S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem[J]. Journal of the American Statistical Association, 1994, 89(427): 958-966.

[30] VAN DYK D A, PARK T. Partially collapsed Gibbs samplers: Theory and methods[J]. Journal of the American Statistical Association, 2008, 103(482): 790-796.

[31] VAN DYK D A, JIAO X. Metropolis-Hastings within partially collapsed Gibbs samplers[J]. Journal of Computational and Graphical Statistics, 2015, 24(2): 301-327.

[32] GELMAN A, RUBIN D B, et al. Inference from iterative simulation using multiple sequences[J]. Statistical science, 1992, 7(4): 457-472.

[33] BROOKS S P, GELMAN A. General methods for monitoring convergence of iterative simulations[J]. Journal of computational and graphical statistics, 1998, 7(4): 434-455.

[34] ULRICH G. Computer Generation of Distributions on the M-Sphere[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1984, 33(2): 158-163.

[35] WOOD A T. Simulation of the von Mises Fisher distribution[J]. Communications in statistics-simulation and computation, 1994, 23(1): 157-164.

[36] GELMAN A, JAKULIN A, PITTAU M G, et al. A weakly informative default prior distribution for logistic and other regression models[J]. The annals of applied statistics, 2008, 2(4): 1360-1383.

[37] PATRA S, DUNSON D B. Constrained bayesian inference through posterior projections[J]. ArXiv preprint arXiv:1812.05741, 2018.

[38] BOUTSIDIS C, DRINEAS P, KAMBADUR P, et al. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix[J]. Linear Algebra and its Applications, 2017, 533: 95-117.

[39] 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.

[40] O'HAGAN A. Fractional Bayes factors for model comparison[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1995, 57(1): 99-118.

[41] ZELLNER A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions[J]. Bayesian inference and decision techniques, 1986.

[42] GEORGE E I, MCCULLOCH R E. Variable selection via Gibbs sampling[J]. Journal of the American Statistical Association, 1993, 88(423): 881-889.

附录

8 后验分布与抽样

8.1 单指数回归模型

这一小节展示了小节 3.1中单指数回归模型在先验模型式 3-2，式 3-4下的后验分布计算。

首先在先验模型式 3-2下，我们可以计算得到参数 $(\beta, g, \sigma, \tau, l)$ 的后验分布为：

$$
\begin{aligned}
p(g,\beta,\sigma,\tau,l|y) &\propto p(y|g,\beta,\sigma,\tau,l)p(g,\beta,\sigma,\tau,l) \\
&\propto p(y|g,\sigma)p(g|\beta,\tau,l)p(\beta)p(\sigma)p(\tau)p(l) \\
&\propto \frac{1}{\sigma^{n/2}}\exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\} \times \frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\} \\
&\times I(\beta \in S^{p-1}) \times \frac{1}{\sigma^{a_\sigma+1}}\exp\{-\frac{b_\sigma}{\sigma}\} \\
&\times \frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\} \times \frac{1}{l^{a_l+1}}\exp\{-\frac{b_l}{l}\}.
\end{aligned}
$$

为了从该后验分布中抽样，我们需要计算下述条件分布：

1. $p(g|y,\beta,\sigma,\tau,l)$

$$
\begin{aligned}
p(g|y,\beta,\sigma,\tau,l) &\propto \exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\}\exp\{-\frac{1}{2}g^TK^{-1}g\} \\
&= N(g;y,\sigma I_n)N(g;0,K) \\
&\propto \exp\{-\frac{1}{2}g^T(\frac{1}{\sigma}I_n+K^{-1})g+\frac{1}{\sigma}y^Tg\} \\
&= \exp\{-\frac{1}{2}g^T(\frac{1}{\sigma}I_n+K^{-1})g+\frac{1}{\sigma}g^T(\frac{1}{\sigma}I_n+K^{-1})(\frac{1}{\sigma}I_n+K^{-1})^{-1}y\},
\end{aligned}
$$

因此我们可以得到后验分布

$$
g|y,\beta,\sigma,\tau,l \sim \mathrm{N}(g;\mu_g,\Sigma_g), \tag{8-1}
$$

其中协方差矩阵

$$
\Sigma_g = (\frac{1}{\sigma}I_n+K^{-1})^{-1} = \sigma K(K+\sigma I_n)^{-1} = \sigma(K+\sigma I_n)^{-1}K,
$$

期望

$$\mu_g = \frac{1}{\sigma}\Sigma_g y = K(K+\sigma I_n)^{-1}y = (K+\sigma I_n)^{-1}Ky.$$

2. $p(\beta|y,g,\sigma,\tau,l)$

$$p(\beta|y,g,\sigma,\tau,l) \propto |K|^{-\frac{1}{2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} I(\beta \in S^{p-1}),$$

由于协方差矩阵 K 在数值上常是病态的，条件数很大且接近奇异矩阵，因此我们很难处理关于 K^{-1} 与 $|K|$ 的计算。为避免与 K 相关直接的计算，我们使用 PCG 抽样的方法，考虑 β, g 的联合后验分布 $\beta, g|y,\sigma,\tau,l$：

$$\begin{aligned} p(\beta,g|y,\sigma,\tau,l) &\propto \left(\exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\} \times \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\}\right) \times I(\beta \in S^{p-1}) \\ &= N(y;g,\sigma I_n) N(g;0,K) \times I(\beta \in S^{p-1}) \\ &= p(g|y,\beta,\sigma,\tau,l)\, p(\beta|y,\sigma,\tau,l), \end{aligned}$$

并且为了减少 g 的抽样次数，对 $\beta, g|y,\sigma,\tau,l$ 中的 g 做积分，得到：

$$\begin{aligned} p(\beta|y,\sigma,\tau,l) &= \int p(\beta,g|y,\sigma,\tau,l)dg \\ &\propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times I(\beta \in S^{p-1}). \end{aligned} \tag{8-2}$$

因此在这一步中，从 $\beta|y,\sigma,\tau,l$ 的分布中抽取 β 可以有效避免处理协方差矩阵 K 的数值问题。

3. $p(\sigma|y,\beta,g,\tau,l)$

$$\begin{aligned} p(\sigma|y,\beta,g,\tau,l) &\propto \frac{1}{\sigma^{n/2}} \exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\} \times \frac{1}{\sigma^{a_\sigma+1}} \exp\{-\frac{b_\sigma}{\sigma}\} \\ &= \frac{1}{\sigma^{a_\sigma+1+n/2}} \exp\{-\left(\frac{1}{2}(y-g)^T(y-g)+b_\sigma\right)/\sigma\} \end{aligned}$$

因此我们可以得到

$$\sigma|y,\beta,g,\tau,l \sim \text{InvGamma}\left(a_\sigma + n/2, (y-g)^T(y-g)/2 + b_\sigma\right).$$

4. $p(\tau|y,\beta,g,\sigma,l)$

$$\begin{aligned} p(\tau|y,\beta,g,\sigma,l) &\propto \frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\} \\ &\propto \frac{1}{\tau^{n/2}|K_0|^{1/2}}\exp\{-\frac{g^TK_0^{-1}g}{2\tau}\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\} \\ &\propto \frac{1}{\tau^{a_\tau+1+n/2}}\exp\{-\left(\frac{1}{2}g^TK_0^{-1}g+b_\tau\right)/\tau\}, \end{aligned}$$

其中 $K_0=K/\tau$，即 $K_0(i,j)=\exp\{-\frac{\left(x_i^T\beta-x_j^T\beta\right)^2}{l}\}$。因此可以得到

$$\tau|y,\beta,g,\sigma,l\sim\text{InvGamma}(a_\tau+n/2,g^TK_0^{-1}g/2+b_\tau). \tag{8-3}$$

为避免直接关于 K_0 的计算，记 $g_\tau=g^Tg/(2b_\tau)=\|g\|_2^2/(2b_\tau)$，我们可以做下述变换：

$$\begin{aligned} \frac{1}{2}g^TK_0^{-1}g+b_\tau &= \frac{1}{2}g^TK_0^{-1}g+\frac{1}{g_\tau}\frac{1}{2}g^Tg \\ &= \frac{1}{2}g^T(K_0^{-1}+\frac{1}{g_\tau}I_n)g \\ &= \frac{1}{2}g^T\left((K_0+g_\tau I_n)^{-1}K_0\right)^{-1}g. \end{aligned} \tag{8-4}$$

同样地，我们可以采用 PCG 抽样的方式，像式 8-2一样，在 $p(\tau,g|y,\beta,\sigma,l)$ 中对 g 积分，得到：

$$p(\tau|y,\beta,\sigma,l)\propto|K+\sigma I_n|^{-\frac{1}{2}}\exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\},$$

并从此分布中抽取 τ。

5. $p(l|y,\beta,g,\sigma,\tau)$

$$p(l|y,\beta,g,\sigma,\tau)\propto\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\times\frac{1}{\tau^{a_l+1}}\exp\{-\frac{b_l}{l}\}.$$

这里同样像式 8-2一样，在 $p(l,g|y,\beta,\sigma,\tau)$ 中对 g 积分，得到：

$$p(l|y,\beta,\sigma,\tau)\propto|K+\sigma I_n|^{-\frac{1}{2}}\exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\}\times\frac{1}{l^{a_l+1}}\exp\{-\frac{b_l}{l}\},$$

并从此分布中抽取 l。

当我们利用单位立方体 $[-1,1]^p$ 上的均匀分布和 t 分布作为更大空间上的先验分布来替代单位球面上的均匀分布时，我们分别设定 $\beta \sim \text{Uniform}([-1,1]^p)$ 和 $\beta_j|\sigma_{\beta_j} \sim \text{N}(0,\sigma_{\beta_j})$，$\sigma_{\beta_j} \sim \text{InvGamma}(a_{\sigma_{\beta_j}}, b_{\sigma_{\beta_j}}), j = 1,2,\cdots,p$ 先验，因此与式 8-2不同，我们需要从

$$p(\beta|y,\sigma,\tau,l) \propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times I(\beta \in [-1,1]^p),$$

$$p(\beta|y,\sigma,\sigma_\beta,\tau,l) \propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times \exp\{-\frac{1}{2}\beta^T\Sigma_\beta^{-1}\beta\},$$

中进行 Metropolis-Hasting 抽样，并将抽样结果投影到单位球面上，其中 $\sigma_\beta = (\sigma_{\beta_1},\cdots,\sigma_{\beta_p})$，$\Sigma_\beta^{-1} = \text{diag}(1/\sigma_{\beta_1},\cdots,1/\sigma_{\beta_p})$。而 σ_β 则需要从 $\sigma_{\beta_j}|y,\beta,g,\sigma,\tau,l \sim \text{InvGamma}(a_{\sigma_{\beta_j}}+1/2, \beta_j^2/2+b_{\sigma_{\beta_j}})$ 中进行抽样，这里的推导可以参考式 3-4模型下式 8-5的计算。

由上述条件分布，我们可以利用 Metropolis-Hasting within PCG 的抽样方法，按照 β,τ,l,g,σ 或者 $\beta,,\tau,l\sigma_\beta,g,\sigma$ 的顺序进行迭代抽样。这个顺序不是唯一的，但是也并不像一般 Gibbs 抽样可以任意更改抽样顺序，这里的抽样顺序需要保证马尔可夫链的平稳分布与目标后验分布一致。

在先验模型式 3-4下，记 $\sigma_\beta = (\sigma_{\beta_1},\cdots,\sigma_{\beta_p})$，我们可以计算得到参数 $(\beta,g,\sigma,\tau,\sigma_\beta)$ 的后验分布为：

$$\begin{aligned}
p(g,\beta,\sigma,\tau,\sigma_\beta|y) &\propto p(y|g,\beta,\sigma,\tau,\sigma_\beta)p(g,\beta,\sigma,\tau,\sigma_\beta) \\
&\propto p(y|g,\sigma)p(g|\beta,\tau)p(\beta|\sigma_\beta)p(\sigma)p(\tau)p(\sigma_\beta) \\
&\propto \frac{1}{\sigma^{n/2}}\exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\} \times \frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\} \\
&\times \prod_{j=1}^{p}\frac{1}{\sigma_{\beta_j}^{1/2}}\exp\{-\frac{\beta_j^2}{2\sigma_{\beta_j}}\} \times \frac{1}{\sigma^{a_\sigma+1}}\exp\{-\frac{b_\sigma}{\sigma}\} \\
&\times \frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\} \times \prod_{j=1}^{p}\frac{1}{\sigma_{\beta_j}^{a_{\sigma_{\beta_j}}+1}}\exp\{-\frac{b_{\sigma_{\beta_j}}}{\sigma_{\beta_j}}\}.
\end{aligned}$$

按照先验模型式 3-2的计算过程，同理我们可以得到：

1. $p(g|y,\beta,\sigma,\tau,\sigma_\beta)$

$$g|y,\beta,\sigma,\tau,\sigma_\beta \sim N(g;\mu_g,\Sigma_g),$$

其中

$$\Sigma_g = (\frac{1}{\sigma}I_n + K^{-1})^{-1} = \sigma K(K+\sigma I_n)^{-1} = \sigma(K+\sigma I_n)^{-1}K,$$

$$\mu_g = \frac{1}{\sigma}\Sigma_g y = K(K+\sigma I_n)^{-1}y = (K+\sigma I_n)^{-1}Ky.$$

2. $p(\beta|y,g,\sigma,\tau,\sigma_\beta)$

$$p(\beta|y,\sigma,\tau,\sigma_\beta) \propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times \exp\{-\frac{1}{2}\beta^T\Sigma_\beta^{-1}\beta\},$$

其中 $\Sigma_\beta^{-1} = \text{diag}(1/\sigma_{\beta_1},\cdots,1/\sigma_{\beta_p})$。

3. $p(\sigma|y,\beta,g,\tau,\sigma_\beta)$

$$\sigma|y,\beta,g,\tau,\sigma_\beta \sim \text{InvGamma}(A_\sigma, B_\sigma),$$

其中 shape $A_\sigma = a_\sigma + \frac{n}{2}$, scale $B_\sigma = \frac{1}{2}(y-g)^T(y-g) + b_\sigma$。

4. $p(\sigma_\beta|y,\beta,g,\sigma,\tau)$

对每个 $j = 1,2,\cdots,p$，

$$\sigma_{\beta_j}|y,\beta,g,\sigma,\tau \sim \text{InvGamma}(A_{\sigma_{\beta_j}}, B_{\sigma_{\beta_j}}), \tag{8-5}$$

其中 shape $A_{\sigma_{\beta_j}} = a_{\sigma_{\beta_j}} + \frac{1}{2}$，scale $B_{\sigma_{\beta_j}} = \frac{1}{2}\beta_j^2 + b_{\sigma_{\beta_j}}$。

5. $p(\tau|y,g,\beta,\sigma,\sigma_\beta)$

$$p(\tau|y,\beta,\sigma,\sigma_\beta) \propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times \frac{1}{\tau^{a_\tau+1}} \exp\{-\frac{b_\tau}{\tau}\}.$$

这里我们可以按照 $\beta,\tau,\sigma_\beta,g,\sigma$ 的顺序进行抽样，这个顺序不唯一，但是需要保证马尔可夫链的平稳分布。

8.2 单指数 Logistic 模型

这一节中我们展示了节 4中关于单指数 Logistic 模型在先验模型式 4-1、式 4-2下的后验分布和条件分布的计算过程。

在先验模型式 4-1下，当 $y_i \in \{1,-1\}$，参数 (g,β,ω,τ,l) 的后验分布为：

$$\begin{aligned}
p(g,\beta,\omega,\tau,l|y) &\propto p(y|g,\beta,\omega,\tau,l)p(g,\beta,\omega,\tau,l) \\
&\propto p(y|g,\omega)p(g|\beta,\tau,l)p(\beta)p(\omega)p(\tau)p(l) \\
&\propto \prod_{i=1}^{n} \exp\{\frac{1}{2}y_i g_i\} \exp\{-\frac{1}{2}\omega_i (y_i g_i)^2\} \times \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \\
&\times I(\beta \in S^{p-1}) \times \prod_{i=1}^{n} p(\omega_i) \times \frac{1}{\tau^{a_\tau+1}} \exp\{-\frac{b_\tau}{\tau}\} \times \frac{1}{l^{a_l+1}} \exp\{-\frac{b_l}{l}\} \\
&\propto \exp\{\frac{1}{2}y^T g - \frac{1}{2}g^T \Omega g\} \times \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \\
&\times I(\beta \in S^{p-1}) \times \prod_{i=1}^{n} p(\omega_i) \times \frac{1}{\tau^{a_\tau+1}} \exp\{-\frac{b_\tau}{\tau}\} \times \frac{1}{l^{a_l+1}} \exp\{-\frac{b_l}{l}\},
\end{aligned}$$

其中 $\Omega = \text{diag}(\omega_1,\cdots,\omega_n), \omega = (\omega_1,\cdots,\omega_n)^T$。为利用 Metropolis-Hasting within PCG 抽样的方法来对这些参数进行抽样，我们需要计算以下条件分布：

1. $p(g|y,\beta,\omega,\tau,l)$

$$\begin{aligned}
p(g|y,\beta,\omega,\tau,l) &\propto \exp\{\frac{1}{2}y^T g - \frac{1}{2}g^T \Omega g\} \times \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \\
&\propto N(g; \frac{1}{2}\Omega^{-1} y, \Omega^{-1}) N(g; 0, K) \\
&\propto \exp\{-\frac{1}{2}(g - \frac{1}{2}(\Omega + K^{-1})^{-1} y)^T (\Omega + K^{-1})(g - \frac{1}{2}(\Omega + K^{-1})^{-1} y)\},
\end{aligned}$$

因此我们可以得到后验分布：

$$g|y,\beta,\omega,\tau,l \sim \text{N}(g; \mu_g, \Sigma_g), \tag{8-6}$$

其中协方差矩阵

$$\Sigma_g = (\Omega + K^{-1})^{-1} = (K\Omega + I_n)^{-1} K,$$

期望

$$\mu_g = \frac{1}{2}(\Omega + K^{-1})^{-1} y = \frac{1}{2}(K\Omega + I_n)^{-1} K y.$$

2. $p(\beta|y,g,\omega,\tau,l)$

$$p(\beta|y,g,\omega,\tau,l) \propto |K|^{-\frac{1}{2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} I(\beta \in S^{p-1})$$

正如小节 8.1中关于协方差矩阵 K 的叙述，K 在数值上常是病态的，接近奇异矩阵。为避免与 K 相关的直接计算，与第小节 8.1中的处理方法相同，这里我们使用 PCG 抽样的方法，考虑 β, g 的联合后验分布 $\beta, g|y, \omega, \tau, l$：

$$\begin{aligned} p(\beta, g|y, \omega, \tau, l) &\propto \exp\{\frac{1}{2}y^T g - \frac{1}{2}g^T \Omega g\} \times \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \times I(\beta \in S^{p-1}) \\ &\propto N(g; \frac{1}{2}\Omega^{-1} y, \Omega^{-1}) N(g; 0, K) \times I(\beta \in S^{p-1}), \end{aligned}$$

为减少对 g 的重复抽样，对 $p(\beta, g|y, \omega, \tau, l)$ 中的 g 积分，得到：

$$\begin{aligned} p(\beta|y, \omega, \tau, l) &= \int p(\beta, g|y, \omega, \tau, l) dg \\ &\propto |K + \Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8} y^T \Omega^{-1} (K + \Omega^{-1})^{-1} \Omega^{-1} y\} \times I(\beta \in S^{p-1}). \end{aligned}$$

我们从上述条件分布中，利用 Metropolis-Hasting 抽样的方法来对 β 进行抽样。

3. $p(\omega|y, \beta, g, \tau, l)$

由 Polya-Gamma 分布的性质，我们可以得到 $p(\omega|b, \psi) \propto \exp\{-\omega\psi^2/2\} p(\omega|b, 0)$，其中 $p(\omega|b, \psi)$ 和 $p(\omega|b, 0)$ 分别是 PG(b, ψ) 和 PG$(b, 0)$ 分布的密度函数[39]。由此，我们可以得到下述密度：

$$\begin{aligned} p(\omega|y, \beta, g, \tau, l) &\propto \prod_{i=1}^{n} \exp\{-\frac{1}{2}\omega_i (y_i g_i)^2\} p(\omega_i) \\ &\propto \prod_{i=1}^{n} p(\omega_i; 1, g_i), \end{aligned}$$

因此有条件分布 $\omega_i|y, \beta, g, \tau, l \sim$ PG$(1, g_i), i = 1, 2, \cdots, n$。

4. $p(\tau|y, \beta, g, \omega, l)$

$$\begin{aligned} p(\tau|y, \beta, g, \omega, l) &\propto \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \times \frac{1}{\tau^{a_\tau + 1}} \exp\{-\frac{b_\tau}{\tau}\} \\ &\propto \frac{1}{\tau^{n/2} |K_0|^{1/2}} \exp\{-\frac{g^T K_0^{-1} g}{2\tau}\} \times \frac{1}{\tau^{a_\tau + 1}} \exp\{-\frac{b_\tau}{\tau}\} \\ &\propto \frac{1}{\tau^{a_\tau + 1 + n/2}} \exp\{-\left(\frac{1}{2} g^T K_0^{-1} g + b_\tau\right)/\tau\} \end{aligned}$$

其中 $K_0 = K/\tau$，即 $K_0(i,j) = \exp\{-\frac{\left(x_i^T\beta - x_j^T\beta\right)^2}{l}\}$。因此像小节 8.1 中式 8-3、式 8-4的结果一样，我们可以得到

$$\begin{aligned}\tau|y,\beta,g,\omega,l &\sim \text{InvGamma}\left(a_\tau + n/2, g^T K_0^{-1} g/2 + b_\tau\right) \\ &= \text{InvGamma}\left(a_\tau + n/2, \frac{1}{2} g^T \left((K_0 + g_\tau I_n)^{-1} K_0\right)^{-1} g\right),\end{aligned}$$

其中 $g_\tau = g^T g/(2b_\tau) = \|g\|_2^2/(2b_\tau)$。同样地，这里我们可以使用 PCG 抽样的方法，在 $p(\tau, g|y,\beta,\omega,l)$ 中对 g 积分，得到：

$$p(\tau|y,\beta,\omega,l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8} y^T \Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1} y\} \times \frac{1}{\tau^{a_\tau+1}} \exp\{-\frac{b_\tau}{\tau}\}.$$

5. $p(l|y,\beta,g,\omega,\tau)$

$$p(l|y,\beta,g,\omega,\tau) \propto \frac{1}{|K|^{1/2}} \exp\{-\frac{1}{2} g^T K^{-1} g\} \times \frac{1}{\tau^{a_l+1}} \exp\{-\frac{b_l}{l}\}.$$

像上述过程关于 K 的处理方式一样，在 $p(l, g|y,\beta,\omega,\tau)$ 中对 g 积分得到条件密度：

$$p(l|y,\beta,\omega,\tau) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8} y^T \Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1} y\} \times \frac{1}{l^{a_l+1}} \exp\{-\frac{b_l}{l}\}.$$

同样地，当我们利用单位立方体 $[-1,1]^p$ 上的均匀分布和 t 分布作为更大空间上的先验分布来替代单位球面上的均匀分布时，我们分别设定 $\beta \sim \text{Uniform}([-1,1]^p)$ 和 $\beta_j|\sigma_{\beta_j} \sim \text{N}(0, \sigma_{\beta_j})$，$\sigma_{\beta_j} \sim \text{InvGamma}(a_{\sigma_{\beta_j}}, b_{\sigma_{\beta_j}}), j = 1, 2, \cdots, p$ 先验，我们需要从

$$p(\beta|y,\omega,\tau,l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8} y^T \Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1} y\} \times I(\beta \in [-1,1]^p),$$

$$p(\beta|y,\omega,\sigma_\beta,\tau,l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8} y^T \Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1} y\} \times \exp\{-\frac{1}{2}\beta^T \Sigma_\beta^{-1} \beta\},$$

中对 β 进行 Metropolis-Hasting 抽样，并将抽样结果投影到单位球面上，其中 $\sigma_\beta = (\sigma_{\beta_1}, \cdots, \sigma_{\beta_p})$，$\Sigma_\beta^{-1} = \text{diag}(1/\sigma_{\beta_1}, \cdots, 1/\sigma_{\beta_p})$。而 σ_β 则需要从 $\sigma_{\beta_j}|y,\beta,g,\omega,\tau,l \sim \text{InvGamma}(a_{\sigma_{\beta_j}} + 1/2, \beta_j^2/2 + b_{\sigma_{\beta_j}})$ 中进行抽样。

式 4-1可以按照 $\beta, \tau, l, g, \omega$ 或者 $\beta, \tau, l, \sigma_\beta, g, \omega$ 的顺序进行抽样。

在先验模型式 4-2下，可以得到后验分布：

$$
\begin{aligned}
p(g,\beta,\omega,\tau,\sigma_\beta|y) &\propto p(y|g,\beta,\omega,\tau,\sigma_\beta)p(g,\beta,\omega,\tau,\sigma_\beta)\\
&\propto p(y|g,\omega)p(g|\beta,\tau)p(\beta|\sigma_\beta)p(\omega)p(\tau)p(\sigma_\beta)\\
&\propto \prod_{i=1}^{n}\exp\{\frac{1}{2}y_ig_i\}\exp\{-\frac{1}{2}\omega_i(y_ig_i)^2\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\\
&\times\frac{1}{\sigma_\beta^{p/2}}\exp\{-\frac{1}{2\sigma_\beta}\beta^T\beta\}\times\prod_{i=1}^{n}p(\omega_i)\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\}\times\frac{1}{\sigma_\beta^{a_{\sigma_\beta}+1}}\exp\{-\frac{b_{\sigma_\beta}}{\sigma_\beta}\}\\
&\propto \exp\{\frac{1}{2}y^Tg-\frac{1}{2}g^T\Omega g\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\\
&\times\frac{1}{\sigma_\beta^{p/2}}\exp\{-\frac{1}{2\sigma_\beta}\beta^T\beta\}\times\prod_{i=1}^{n}p(\omega_i)\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\}\times\frac{1}{\sigma_\beta^{a_{\sigma_\beta}+1}}\exp\{-\frac{b_{\sigma_\beta}}{\sigma_\beta}\},
\end{aligned}
$$

与上述关于式 4-1的推导类似，同理可以得到：

1. $p(g|y,\beta,\omega,\tau,\sigma_\beta)$

$$g|y,\beta,\omega,\tau,\sigma_\beta\sim N(g;\mu_g,\Sigma_g),$$

其中

$$\Sigma_g=(\Omega+K^{-1})^{-1}=(K\Omega+I_n)^{-1}K,$$
$$\mu_g=\frac{1}{2}(\Omega+K^{-1})^{-1}y=\frac{1}{2}(K\Omega+I_n)^{-1}Ky.$$

2. $p(\beta|y,g,\omega,\tau,\sigma_\beta)$

$$p(\beta|y,\omega,\tau,\sigma_\beta)\propto|K+\Omega^{-1}|^{-\frac{1}{2}}\exp\{-\frac{1}{8}y^T\Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1}y\}\times\exp\{-\frac{1}{2\sigma_\beta}\beta^T\beta\}.$$

3. $p(\omega|y,\beta,g,\tau,\sigma_\beta)$

$$\omega_i|y,\beta,g,\tau,\sigma_\beta\sim\text{PG}(1,g_i),i=1,2,\cdots,n.$$

4. $p(\sigma_\beta|y,\beta,g,\omega,\tau)$

对每个 $j=1,2,\cdots,p$，

$$\sigma_{\beta_j}|y,\beta,g,\omega,\tau\sim\text{InvGamma}(A_{\sigma_{\beta_j}},B_{\sigma_{\beta_j}}),$$

其中 shape $A_{\sigma_{\beta_j}}=a_{\sigma_{\beta_j}}+\frac{1}{2}$，scale $B_{\sigma_{\beta_j}}=\frac{1}{2}\beta_j^2+b_{\sigma_{\beta_j}}$。

5. $p(\tau|y,g,\beta,\omega,l)$

$$p(\tau|y,\beta,\omega,l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8}y^T\Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1}y\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\}.$$

这里式 4-2可以按照 $\beta,\sigma_\beta,g,\omega,\tau$ 的顺序进行抽样。

8.3 单指数模型变量选择

这一节中展示了节 5中变量选择先验模型的后验分布和条件分布的计算过程。记 $\delta=(\delta_1,\cdots,\delta_p)$，$\sigma_\beta=(\sigma_{\beta_1},\cdots,\sigma_{\beta_p})$，$\pi=(\pi_1,\cdots,\pi_p)$，则由单指数回归模型的先验模型式 5-1，我们可以得到参数 $(g,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l)$ 的后验分布为：

$$\begin{aligned}
p(g,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l|y) &\propto p(y|g,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l)p(g,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l)\\
&\propto p(y|g,\sigma)p(g|\beta,\delta,\tau,l)p(\beta|\delta,\sigma_\beta)p(\delta|\pi)p(\sigma_\beta)p(\sigma)p(\pi)p(\tau)p(l)\\
&\propto \frac{1}{\sigma^{n/2}}\exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\\
&\times\prod_{j=1}^{p}\frac{1}{(r(\delta_j)\sigma_{\beta_j})^{1/2}}\exp\{-\frac{\beta_j^2}{2r(\delta_j)\sigma_{\beta_j}}\}\times\prod_{j=1}^{p}\frac{1}{\sigma_{\beta_j}^{a_{\sigma_{\beta_j}}+1}}\exp\{-\frac{b_{\sigma_{\beta_j}}}{\sigma_{\beta_j}}\}\\
&\times\prod_{j=1}^{p}\pi_j^{\delta_j}(1-\pi_j)^{1-\delta_j}\times\prod_{j=1}^{p}\pi_j^{a_{\pi_j}-1}(1-\pi_j)^{b_{\pi_j}-1}\\
&\times\frac{1}{\sigma^{a_\sigma+1}}\exp\{-\frac{b_\sigma}{\sigma}\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\}\times\frac{1}{l^{a_l+1}}\exp\{-\frac{b_l}{l}\},
\end{aligned}$$

为利用 Metropollis-Hasting within PCG 抽样的方法，我们计算下列条件分布：

1. $p(g|y,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l)$

$$p(g|y,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l)\propto\exp\{-\frac{1}{2\sigma}(y-g)^T(y-g)\}\exp\{-\frac{1}{2}g^TK^{-1}g\},$$

因此我们可以得到与式 8-1相同的结果，即：

$$g|y,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l\sim\mathrm{N}(g;\mu_g,\Sigma_g),$$

其中协方差矩阵

$$\Sigma_g=(\frac{1}{\sigma}I_n+K^{-1})^{-1}=\sigma K(K+\sigma I_n)^{-1}=\sigma(K+\sigma I_n)^{-1}K,$$

期望

$$\mu_g=\frac{1}{\sigma}\Sigma_g y=K(K+\sigma I_n)^{-1}y=(K+\sigma I_n)^{-1}Ky.$$

2. $p(\beta|y,g,\delta,\sigma_\beta,\pi,\sigma,\tau,l)$

$$p(\beta|y,g,\delta,\sigma_\beta,\pi,\sigma,\tau,l) \propto |K|^{-\frac{1}{2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \times \exp\{-\frac{1}{2}\beta^T \Sigma_\beta^{-1}\beta\},$$

其中 $\Sigma_\beta = \text{diag}(r(\delta_1)\sigma_{\beta_1}, \cdots, r(\delta_p)\sigma_{\beta_p})$。同样考虑到协方差矩阵 K 在数值上的问题，考虑 β, g 的联合后验分布 $\beta, g|y,\delta,\sigma_\beta,\pi,\sigma,\tau,l$ 并对 g 积分，得到：

$$\begin{aligned} p(\beta|y,\delta,\sigma_\beta,\pi,\sigma,\tau,l) &= \int p(\beta,g|y,\delta,\sigma_\beta,\pi,\sigma,\tau,l)dg \\ &\propto |K+\sigma I_n|^{-\frac{1}{2}} \exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\} \times \exp\{-\frac{1}{2}\beta^T\Sigma_\beta^{-1}\beta\}. \end{aligned}$$

3. $p(\delta|y,g,\beta,\sigma_\beta,\pi,\sigma,\tau,l)$

$$p(\delta|y,g,\beta,\sigma_\beta,\pi,\sigma,\tau,l) \propto \prod_{j=1}^{p} \frac{1}{(r(\delta_j)\sigma_{\beta_j})^{1/2}} \exp\{-\frac{\beta_j^2}{2r(\delta_j)\sigma_{\beta_j}}\} \pi_j^{\delta_j}(1-\pi_j)^{1-\delta_j},$$

所以对每个 $\delta_j, j = 1, 2, \cdots, p$，可以得到

$$p(\delta_j = 1|y,g,\beta,\sigma_\beta,\pi,\sigma,\tau,l) = C \cdot \sigma_{\beta_j}^{-1/2} \exp\{-\beta_j^2/2\sigma_{\beta_j}\}\pi_j,$$

$$p(\delta_j = 0|y,g,\beta,\sigma_\beta,\pi,\sigma,\tau,l) = C \cdot (r(0)\sigma_{\beta_j})^{-1/2} \exp\{-\beta_j^2/2r(0)\sigma_{\beta_j}\}(1-\pi_j),$$

因此可以得到

$$\delta_j|y,g,\beta,\sigma_\beta,\pi,\sigma,\tau,l \sim \text{Bernoulli}\left(\left(1 + \frac{p_{Spike}(\beta_j)(1-\pi_j)}{p_{Slab}(\beta_j)\pi_j}\right)^{-1}\right).$$

4. $p(\pi|y,g,\beta,\delta,\sigma_\beta,\sigma,\tau,l)$

$$p(\pi|y,g,\beta,\delta,\sigma_\beta,\sigma,\tau,l) \propto \prod_{j=1}^{p} \pi_j^{\delta_j}(1-\pi_j)^{1-\delta_j} \times \prod_{j=1}^{p} \pi_j^{a_{\pi_j}-1}(1-\pi_j)^{b_{\pi_j}-1},$$

所以对每个 $j = 1, 2, \cdots, p$，有：

$$\pi_j|y,g,\beta,\delta,\sigma_\beta,\sigma,\tau,l \sim \text{Beta}(a_{\pi_j} + \delta_j, b_{\pi_j} + 1 - \delta_j).$$

5. $p(\sigma_\beta|y,g,\beta,\delta,\pi,\sigma,\tau,l)$

$$p(\sigma_\beta|y,g,\beta,\delta,\pi,\sigma,\tau,l) \propto \prod_{j=1}^{p} \frac{1}{\sigma_{\beta_j}^{1/2}} \exp\{-\frac{\beta_j^2}{2r(\delta_j)\sigma_{\beta_j}}\} \times \prod_{j=1}^{p} \frac{1}{\sigma_{\beta_j}^{a_{\sigma_{\beta_j}}+1}} \exp\{-\frac{b_{\sigma_{\beta_j}}}{\sigma_{\beta_j}}\},$$

所以对每个 $j = 1, 2, \cdots, p$，有：

$$\sigma_{\beta_j}|y,g,\beta,\delta,\pi,\sigma,\tau,l \sim \text{InvGamma}(\frac{1}{2} + a_{\sigma_{\beta_j}}, \frac{\beta_j^2}{2r(\delta_j)} + b_{\sigma_{\beta_j}}).$$

6. 关于 σ,τ,l 的条件分布与小节 8.1中内容一致，我们可以得到：

$$\sigma|y,g,\beta,\delta,\sigma_\beta,\pi,\tau,l \sim \text{IG}(a_\sigma+n/2,(y-g)^T(y-g)/2+b_\sigma),$$

$$p(\tau|y,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l) \propto |K+\sigma I_n|^{-\frac{1}{2}}\exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\}\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\},$$

$$p(l|y,\beta,\delta,\sigma_\beta,\pi,\sigma,\tau,l) \propto |K+\sigma I_n|^{-\frac{1}{2}}\exp\{-\frac{1}{2}y^T(K+\sigma I_n)^{-1}y\}\times\frac{1}{l^{a_l+1}}\exp\{-\frac{b_l}{l}\}.$$

这里可以按照 $\beta,\tau,l,\sigma_\beta,\delta,\pi,g,\sigma$ 的顺序进行抽样。

在单指数Logistic模型的变量选择式5-2下，我们可以得到参数 $(g,\beta,\delta,\omega,\sigma_\beta,\pi,\tau,l)$ 的后验分布为：

$$\begin{aligned}
p(g,\beta,\delta,\omega,\sigma_\beta,\pi,\tau,l|y) &\propto p(y|g,\beta,\delta,\omega,\sigma_\beta,\pi,\tau,l)p(g,\beta,\delta,\omega,\sigma_\beta,\pi,\tau,l)\\
&\propto p(y|g,\omega)p(g|\beta,\tau,l)p(\beta|\delta,\sigma_\beta)p(\delta|\pi)p(\sigma_\beta)p(\pi)p(\omega)p(\tau)p(l)\\
&\propto \prod_{i=1}^{n}\exp\{\frac{1}{2}y_ig_i\}\exp\{-\frac{1}{2}\omega_i(y_ig_i)^2\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\\
&\times\prod_{j=1}^{p}\frac{1}{(r(\delta_j)\sigma_{\beta_j})^{1/2}}\exp\{-\frac{\beta_j^2}{2r(\delta_j)\sigma_{\beta_j}}\}\times\prod_{j=1}^{p}\frac{1}{\sigma_{\beta_j}^{a_{\sigma_{\beta_j}}+1}}\exp\{-\frac{b_{\sigma_{\beta_j}}}{\sigma_{\beta_j}}\}\\
&\times\prod_{j=1}^{p}\pi_j^{\delta_j}(1-\pi_j)^{1-\delta_j}\times\prod_{j=1}^{p}\pi_j^{a_{\pi_j}-1}(1-\pi_j)^{b_{\pi_j}-1}\times\prod_{i=1}^{n}p(\omega_i)\\
&\times\frac{1}{\tau^{a_\tau+1}}\exp\{-\frac{b_\tau}{\tau}\}\times\frac{1}{l^{a_l+1}}\exp\{-\frac{b_l}{l}\}\\
&\propto \exp\{\frac{1}{2}y^Tg-\frac{1}{2}g^T\Omega g\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}.
\end{aligned}$$

同样地，我们计算下列条件分布：

1. $p(g|y,\beta,\delta,\omega,\sigma_\beta,\pi,\tau,l)$

$$\begin{aligned}
p(g|y,\beta,\omega,\tau,l) &\propto \prod_{i=1}^{n}\exp\{\frac{1}{2}y_ig_i\}\exp\{-\frac{1}{2}\omega_i(y_ig_i)^2\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\}\\
&\propto \exp\{\frac{1}{2}y^Tg-\frac{1}{2}g^T\Omega g\}\times\frac{1}{|K|^{1/2}}\exp\{-\frac{1}{2}g^TK^{-1}g\},
\end{aligned}$$

其中 $\Omega=\text{diag}(\omega_1,\cdots,\omega_n)$。因此我们可以得到与小节 8.2中式 8-6一致的结果：

$$g|y,\beta,\omega,\tau,l \sim N(g;\mu_g,\Sigma_g),$$

其中协方差矩阵

$$\Sigma_g = (\Omega + K^{-1})^{-1} = (K\Omega + I_n)^{-1}K,$$

期望

$$\mu_g = \frac{1}{2}(\Omega + K^{-1})^{-1}y = \frac{1}{2}(K\Omega + I_n)^{-1}Ky.$$

2. $p(\beta|y, g, \delta, \omega, \sigma_\beta, \pi, \tau, l)$

$$p(\beta|y, g, \delta, \omega, \sigma_\beta, \pi, \tau, l) \propto |K|^{-\frac{1}{2}} \exp\{-\frac{1}{2}g^T K^{-1} g\} \times \exp\{-\frac{1}{2}\beta^T \Sigma_\beta^{-1}\beta\},$$

其中 $\Sigma_\beta = \text{diag}(r(\delta_1)\sigma_{\beta_1}, \cdots, r(\delta_p)\sigma_{\beta_p})$。同样考虑到协方差矩阵 K 在数值上的问题，考虑 β, g 的联合后验分布 $\beta, g|y, \delta, \omega, \sigma_\beta, \pi, \tau, l$ 并对 g 积分，得到：

$$\begin{aligned} p(\beta|y, \delta, \sigma_\beta, \pi, \sigma, \tau, l) &= \int p(\beta, g|y, \delta, \sigma_\beta, \pi, \sigma, \tau, l)dg \\ &\propto |K + \Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8}y^T\Omega^{-1}(K + \Omega^{-1})^{-1}\Omega^{-1}y\} \times \exp\{-\frac{1}{2}\beta^T\Sigma_\beta^{-1}\beta\}. \end{aligned}$$

3. 关于 $\delta, \sigma_\beta, \pi$ 的条件分布，与单指数回归模型变量选择部分的结果一致，对每个 $j = 1, 2, \cdots, p$，我们可以得到：

$$\delta_j|y, g, \beta, \sigma_\beta, \pi, \omega, \tau, l \sim \text{Bernoulli}\left(\left(1 + \frac{p_{Spike}(\beta_j)(1-\pi_j)}{p_{Slab}(\beta_j)\pi_j}\right)^{-1}\right),$$

$$\sigma_{\beta_j}|y, g, \beta, \delta, \pi, \omega, \tau, l \sim \text{InvGamma}(\frac{1}{2} + a_{\sigma_{\beta_j}}, \frac{\beta_j^2}{2r(\delta_j)} + b_{\sigma_{\beta_j}}),$$

$$\pi_j|y, g, \beta, \delta, \sigma_\beta, \omega, \tau, l \sim \text{Beta}(a_{\pi_j} + \delta_j, b_{\pi_j} + 1 - \delta_j).$$

4. 关于 ω, τ, l 的条件分布，与小节 8.2中的内容一致，我们可以得到：

$$\omega_i|y, g, \beta, \delta, \sigma_\beta, \pi, \tau, l \sim \text{PG}(1, g_i), i = 1, 2, \cdots, n,$$

$$p(\tau|y, \beta, \delta, \sigma_\beta, \pi, \omega, \tau, l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8}y^T\Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1}y\} \times \frac{1}{\tau^{a_\tau+1}} \exp\{-\frac{b_\tau}{\tau}\},$$

$$p(l|y, \beta, \delta, \sigma_\beta, \pi, \omega, \tau, l) \propto |K+\Omega^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{8}y^T\Omega^{-1}(K+\Omega^{-1})^{-1}\Omega^{-1}y\} \times \frac{1}{l^{a_l+1}} \exp\{-\frac{b_l}{l}\}.$$

这里可以按照 $\beta, \tau, l, \sigma_\beta, \delta, \pi, g, \omega$ 的顺序进行抽样。

本科生毕业论文（设计）任务书

一、题目：

二、指导教师对毕业论文（设计）的进度安排及任务要求：

（1）1 月 1 日—1 月 14 日：导师下达任务书，对进度、文献和开题提出要求。

（2）1 月 15 日——1 月 23 日：学生确认任务书，对确定的课题搜集相关文献资料，了解问题的背景、应用、研究历史与现状。从中确定论文最终题目。

（3）1 月 24 日——2 月 24 日：对确定的题目进一步展开学习，包括所必需的基础知识及近几年涉及此问题的文章。初步撰写并完成开题报告、文献综述，并提交导师审核。

（4）2 月 25 日——3 月 1 日：组织开题，每位学生准备 10 分钟左右的答辩。

（5）3 月 2 日——4 月 2 日：将定稿的开题报告、文献综述、外文翻译稿上传至教务系统。做中期检查报告。

（6）4 月 3 日——5 月 12 日：完成论文初稿，进行论文稿的修改并最终完成，向导师提交论文终稿。

（7）5 月 13 日——5 月 15 日：导师评阅，学生提交导师填写评语和签字的“毕业论文考核表”及符合规范格式要求的送审论文。

（8）5 月 16 日——5 月 21 日：毕业论文专家评阅。

（9）5 月 22 日——5 月 24 日：评阅结果有修改意见的，根据评阅意见对论文进行修改。

（10）5 月 24 日——5 月 31 日：组织毕业论文答辩。提交最终版毕业论文，并将论文上传至教务系统。

起讫日期 20　年　月　日　至　20　年　月　日

指导教师（签名）_____ 职称 _____

三、系或研究所审核意见：

负责人（签名）_____

年　月　日

本科生毕业论文（设计）考核

一、指导教师对毕业论文（设计）的评语：

指导教师（签名）_____

年　月　日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述（10%）	开题报告（15%）	外文翻译（5%）	毕业论文质量及答辩（70%）	总评成绩
分值					

负责人（签名）_____

年　月　日