2020 届本科生学士学位论文　　学校代码：10269

華東師範大學

East China Normal University

本科生毕业论文

基于 INLA 算法的贝叶斯推断及应用

Bayesian Inference and Application Using INLA Method

姓　　名:	吴文韬
学　　号:	10165000225
学　　院:	统计学院
专　　业:	统计学
指导教师:	汤银才
职　　称:	教授
完成时间:	2020 年 5 月

目 录

插 图 目 录

表 格 目 录

摘　要

在贝叶斯理论中，我们将参数视为随机变量并具备先验分布，将先验与现有数据的似然结合得到后验分布并进行参数推断，因此计算后验分布的积分是关键。然而实证中的后验分布往往是高维的，使得传统计算方法无法得到理想结果。随着近代计算机科学的发展，马尔科夫链蒙特卡洛 (MCMC) 方法产生，MCMC 方法通过构造马尔科夫链，使其收敛后的平稳分布为待积后验分布，得到后验分布的经验样本，进而可以准确的得到参数估计。然而，当前的世界已经逐步进入了大数据的时代，数据在时间和空间上的结构复杂性大大增加了 MCMC 方法的时间成本，这也限制了 MCMC 方法在实际中的应用。

在此之前，拉普拉斯近似也可以计算后验积分，对被积函数在峰值点进行二阶泰勒展开并作高斯近似，以得到被积函数的积分近似。但随着 MCMC 方法出现，此方法逐渐失去了关注度。本文引用 Rue 等基于拉普拉斯近似提出的进行贝叶斯推断的新方法——集成嵌套拉普拉斯近似 (INLA) 方法。INLA 方法作用于属于高斯马尔科夫随机场的隐高斯模型，将拉普拉斯近似与数值积分结合，超参数分布近似使用拉普拉斯近似，再利用数值积分得到参数的后验边际，进而做出较为精确的推断，且运算效率远高于 MCMC 方法。

本文使用预测研究生录取的数据集，建立 Logistic 回归模型，并对参数分别进行基于 MCMC 和 INLA 方法的贝叶斯推断。使用 SAS 和 R 分别完成 MCMC 方法和 INLA 方法的实现，并从参数估计结果、运行时间和 DIC 值三个方面对两种方法进行比较。结果显示 INLA 方法可以得到精确的参数估计，且在运行时间方面明显优于 MCMC 方法，这说明了 INLA 方法对参数的贝叶斯推断有十分理想的效果，对 MCMC 方法也有很好的的替代性。本文最后使用参数估计结果代入模型对测试集进行预测，同时使用 KNN、随机森林和 XGboost 三种机器学习算法建模，并从混淆矩阵、衍生指标 (准确率、Kappa 值和 AUC 值等) 和曲线 (ROC 曲线、提升图等) 三个方面对比模型预测效果，结果显示基于 INLA 方法估计参数的 Logistic 回归模型具备较好的预测效果，并不逊于这三种机器学习算法。

关键词: INLA，MCMC，拉普拉斯近似，Logistic 回归，DIC，机器学习

Abstract

In Bayesian theory, we treat the parameters that have a prior distribution as random variables. The inference of parameters is based on posterior distribution, which is obtained by combining likelihood from existing data and prior. Therefore, the key is to calculate the integral of posterior distribution. However, in the past, statistical analysis based on the Bayes theorem was often daunting as the posterior distribution is often in a space of high dimension. With the development of modern computer science, for many years, Bayesian inference has relied upon Markov chain Monte Carlo method to compute the posterior distribution. The MCMC method constructs a Markov chain so that the stationary distribution of this chain when converged is the posterior distribution, then we get empirical samples of posterior, which can accurately obtain the estimates of parameters. However, with the advent of big data era, the structural complexity of data in time and space has greatly increased the time cost of MCMC method, which also limits its application.

Before that, there is a Laplace approximation method that can also calculate posterior integral, by performing a second-order Taylor expansion of integrand at the peak point and make a Gaussian approximation to obtain an integral approximation of the integrand. However, with the emergence of MCMC method, this method gradually lost focus. This paper cites a novel approach for Bayesian inference proposed by Rue et al: the integrated nested Laplace approximation (INLA) method. The INLA method focuses on latent Gaussian model that can be expressed as latent Gaussian Markov random fields (GMRF). It combines the Laplace approximation with numerical integration. The approximation of hyperparameter distribution uses the Laplace approximation, and we use numerical integration to obtain the posterior marginal distribution for further inference of parameters, which

is more efficient than the MCMC method.

In this paper, we use a dataset for prediction of Graduate Admissions to establish a Logistic regression model, and perform Bayesian inference for the parameters based on MCMC and INLA method respectively. Use SAS and R to complete the implementation of MCMC and INLA method respectively, and compare the two methods from parameter estimation results, program running time and DIC value. The result shows that the INLA method can obtain an accurate parameter estimation, and is significantly better than the MCMC method in terms of running time. This shows that the INLA method can be regarded as a valid alternative to the MCMC method. Finally, we model a Logistic regression using the parameter estimation results from INLA method to forecast the test set. Furthermore, we use three machine learning algorithms (KNN, random forest and XGboost) to model and forecast, and compare these four methods from derived indicators (accuracy rate, Kappa value and AUC value, etc.) calculated from confusion matrix and curves (ROC curve, lifting chart, etc.). Results show that the Logistic regression model with parameters estimated by INLA method also has good predictive effect, which is not inferior to these three machine learning algorithms.

Key Words: INLA, MCMC, Logistic regression, DIC, Machine Learning

第一章　序言

§1.1　研究背景

贝叶斯方法最早来源于英国数学家托马斯. 贝叶斯 (1702-1761)，提出贝叶斯公式，将参数先验分布和似然函数合并得到参数的后验分布。但主要因为在处理先验概率方面存在争议，加之理论并不完善且实际应用中也频繁出现问题，该理论在很长一段时间内不被接受，发展缓慢。20 世纪 90 年代前，贝叶斯方法在统计分析中的应用十分有限，因为针对复杂模型的数学计算很难完成。不过，近代用于拟合贝叶斯模型的算法不断创新，同时伴随着计算机的普及进步，贝叶斯统计得到了极大的推动。如今贝叶斯统计被广泛地用于农业、工业、商业以及学术研究领域。

在贝叶斯理论中，模型参数被视为随机变量，拥有先验分布，先验分布结合似然得到参数的后验分布，从而可以提供待估参数的后验均值、标准差、分位数以及置信区间等，这也使得参数可以拥有更加合理的解释。然而，在过去，这种方法往往很难进行，因为这一过程中需要计算后验分布的积分，而后验分布在实证中往往是高维且具有非标准的形式，使用基础方法无法进行计算。不过随着贝叶斯理论的进步和完善，对参数估计的方法不断出新，近代信息技术革新也大大推动了贝叶斯的进步，同时也为 MCMC 方法提供了技术可能，对参数后验分布的模拟效率大大提升。

马尔可夫链蒙特卡洛 (MCMC) 本质是使用马尔可夫链进行蒙特卡洛模拟，其思想是通过取样构造一条收敛的马尔可夫链，使得该链的平稳分布与待积的后验分布一致，从而可作进一步推断。相关理论可以参看 Metropolis 等 (1953)[1] 以及 Hastings(1970)[2]。构造马尔科夫链的方法有很多，但大多数方法 (Metropolis 抽样、独立抽样、随机游动抽样等) 包括目前最常见的、使用最普遍的 Gibbs 抽样 (Geman and Geman, 1984)[3] 其实是 Metropolis-Hastings 算法的特殊形式。描述 MCMC 应用于贝叶斯模型拟合的开创性参考文献是 Gelfand 和 Smith(1990)[4]。

MCMC 花了近 40 年时间才渗透到主流统计实践中。它起源于统计物理文献，并已被用于空间统计和图像分析十年。在过去的几年中，MCMC 方法极大的推动了贝叶斯统计的发展。起初 MCMC 方法也并非普及，因其对计算机也有极高的要求，但随着近些年计算机的进步以及各种软件的开发，基于 MCMC 方法的贝叶斯统计推断也就愈加方便起来，其中尤其是 WinBUGS (Lunn et al. 2000)[5] 和 OpenBUGS (Lunn et al. 2009)[6] 软件的开发，以及可实现 MCMC 的 R 软件包的出现，实现 MCMC 也愈加简易方便起来。然而，当前的世界已经逐步进入了大数据的时代，MCMC 方法的缺点也随之产生：考虑到数据收集规模的扩大，导致大数据集的可用性增加，以及考虑到数据集空间和时间结构的复杂性，迭代取样使得收敛速度过于缓慢，大大增加了时间成本。

起初在针对后验分布的积分问题中存在着使用拉普拉斯近似计算的一种思路，但随着 MCMC 方

法的出现，这种思路也就失去了关注度。Havard Rue, Martino 和 Chopin(2009)[7] 结合拉普拉斯近似和数值积分，提出了一种新的计算方法: **集成嵌套拉普拉斯方法 (Integrated Nested Laplace Approximations, INLA)**。这种方法对参数的估计精度与 MCMC 方法近似，但在运算时间上远优于 MCMC 方法，对于 MCMC 方法通常需要几小时甚至几天去处理的情形，INLA 仅需要几分钟或数小时。INLA 所能拟合的模型仅限于属于潜高斯马尔可夫随机场 (Rue and Held,2005)[8] 的一类模型，但这已经包括了大量常用的模型，包括本文使用到的 Logistic 回归模型。此外，作者开发了 R-INLA 包以便于进行模型拟合，目前 R-INLA 程序包已比较成熟。

§1.2　国内外研究现状

因 INLA 方法在参数估计精度上与 MCMC 方法一致，但运算速度上远优于后者的特性逐渐得到学者们的关注。随着 Rue 等人的推广以及 R 语言中 INLA 包的完善，INLA 方法已经应用到广义线性模型、半参数模型、样条模型、时空模型和地域模型等多类模型中。例如，Martino 等 (2010)[9] 建立金融时间序列领域中的随机波动模型 (SV 模型) 并使用 INLA 方法进行推断，对 SP500 指数和微软日收盘数据建立 SV 模型进行实证分析，并表明 INLA 方法可以有效的实现模型的参数贝叶斯估计，且运行时间较短；Ugarte(2014)[10] 等应用 R-INLA 软件包进行疾病绘图；Fong 和 Rue 等 (2010)[11] 运用 INLA 方法对广义线性混合模型 (GLMM) 的参数进行了贝叶斯推断；Riebler 和 Held(2017)[12] 将 INLA 算法应用到年龄-周期-队列模型中；Natário 等 (2014)[13] 建立时空分层模型分析葡萄牙森林火灾数据，并使用 INLA 算法对模型进行推断；Kandt(2016)[14] 应用 INLA 算法以研究在城市中获得住房、健康和幸福之间的关系；Gómez-Rubio 和 Bivand(2017)[15] 利用 INLA 方法估计空间计量模型的参数，并基于波士顿房产数据进行实证分析；Santermans(2016)[16] 将其应用到一项关于埃博拉病毒进化的研究；Tsiko(2016) 运用此算法研究非洲人类的暴力行为的普遍性和相关性；Niemi(2015)[17] 将此算法应用到基因表现杂种优势的研究中；Martino 和 Akerkar(2011)[18] 将 INLA 算法应用到生存分析中；Wang, XF(2013)[19] 运用 INLA 算法到非参数回归模型中；Yu 和 Rue(2011)[20] 针对慕尼黑房屋租赁数据，运用 INLA 方法对租金的 0.25，0.50 和 0.75 分位数分别进行贝叶斯参数估计；Cameletti. M(2019)[21] 将 INLA 算法和随机偏微分方程结合应用于存在空间偏差的空气污染数据，Poggio L(2016)[22] 也用了这种思路针对苏格兰有机土壤数据提供了一个数字土壤测绘的新方法，其具有可比的验证结果和重要的计算收益；Selle M L 等 (2019)[23] 配合 R-INLA 包拟合了不同空间模型来分析农业田间试验。

本文介绍 INLA 算法在 Logistic 回归模型中的应用，同时也使用 MCMC 方法并比较两种方法对参数估计的精度以及运行时间等，最后利用建立好的模型进行预测。从查询到的文献中显示，目前对 Logistic 回归模型进行贝叶斯方法研究时基本使用 MCMC 方法：周翔等 (2016)[24] 利用 MCMC 方法估计 Logistic 回归模型中的参数，研究了丁酸梭菌株对于给定辐照区间剂量的应答趋势；付志慧等 (2019)[25] 使用了 Rstan 包进行了四参数 Logistic 模型 (4PLM) 参数估计；Zucknick(2014)[26] 使用基于不同抽样的 MCMC 方法对 logistic 回归模型进行变量选择。

§1.3　本文主要内容

本文的研究安排为：

第一章介绍贝叶斯理论研究背景以及 MCMC 方法的发展，与 MCMC、INLA 相关的国内外研究现状以及论文全篇的结构安排。

第二章介绍预备知识，包括贝叶斯方法的内容与发展、Logistic 回归模型以及在实证分析中所使用的数据集。本文使用一个预测学生能否被研究生院校录取的数据集，响应变量为可被录取的概率，解释变量为该学生的各项成绩和材料评分等。本章最后对此数据集进行可视化和预处理。

第三章主要介绍马尔科夫链蒙特卡洛方法，包含基本内容、蒙特卡洛积分和马尔科夫链，最后介绍使用 MCMC 时所采用的抽样方法。

第四章主要介绍集成嵌套的拉普拉斯近似方法。首先介绍拉普拉斯近似，其次介绍 INLA 方法适用的模型：隐高斯模型，并介绍高斯马尔科夫随机场，最后总结 INLA 方法，基于前面内容提出处理参数后验的新计算方式，使用一系列近似与数值积分得到参数后验。

第五章进行实证分析，对实证所用的数据进行 Logistic 回归模型建模，并分别基于 MCMC 方法和 INLA 方法进行参数的贝叶斯估计，得到参数的后验均值、标准差和各分位数估计，以及近似的后验分布。本文使用 SAS 软件实现 MCMC 方法，使用 R 软件的 INLA 包实现 INLA 方法。

第六章将第五章得到的结果进行对比，主要从参数估计结果、DIC 值和运行时间三个方面进行对比，观察 INLA 的估计结果是否精确且高效。并将 INLA 方法估计的参数结果代回模型进行预测，并与 KNN、随机森林和 XGboost 三种机器学习方法进行混淆矩阵及衍生指标和曲线的对比。

第二章　预备内容

§2.1　贝叶斯方法

2.1.1　贝叶斯方法定义

目前我们所熟知的最常用的统计方法来自于频率学派 (或经典学派)。频率学派主张未知参数是固定的常数，他们通过限制相关的频率来定义概率。他们从这些假设中得出概率是客观的，即不能用概率性的陈述对参数做出解释，因为参数是固定的。而贝叶斯方法提供了另一种看法：他们将参数视为随机变量，并将概率定义为“置信度”，即事件的概率为我们相信事件是正确的程度。从这些假设可以得出贝叶斯派认为概率是主观的，我们可以对参数做出概率陈述。自贝叶斯定理被提出后，以十八世纪长老会牧师托马斯·贝叶斯的名字命名的术语“贝叶斯”被广泛使用。贝叶斯意在解决参数概率的问题，即在观察了一系列事件之后，一个事件的概率是多少。

假设我们需要通过数据 $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ 来估计参数 θ，利用分布 $p(\boldsymbol{x}|\theta)$ 所描述的统计模型。贝叶斯学说认为 θ 是不能被准确的确定的，参数的不确定性通过对概率的表述和分布来表达。例如我们认为 θ 服从标准正态分布，是因为我们认为这个分布能够最好的解释与参数有关的不确定性。下面的步骤描述了贝叶斯推断的基本要素：

1. 参数 θ 的概率分布被表述为 $\pi(\theta)$，叫做先验分布，简称为先验。先验分布表示我们在处理数据之前对参数 θ 的看法。本文仅考虑 θ 的取值是连续的，其先验分布用密度表示。

2. 已知数据 $\boldsymbol{x}$，我们选择一个已知 θ 下 $\boldsymbol{x}$ 的条件分布 $p(\boldsymbol{x}|\theta)$。

3. 我们使用先验分布与已有数据来更新对参数 θ 的认知，即综合二者信息得到后验分布 $p(\theta|\boldsymbol{x})$。

其中第三步利用了贝叶斯公式，可得到后验分布与先验和条件分布 $p(\boldsymbol{x}|\theta)$ 的关系：

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta, \boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\int p(\boldsymbol{x}|\theta)\pi(\theta)d\theta}, \tag{2.1}$$

其中

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)\pi(\theta)d\theta \tag{2.2}$$

是后验分布的归一化常数，同样是 $\boldsymbol{x}$ 的边际分布，因此有时被称为数据边际分布；任何与 $p(\boldsymbol{x}|\theta)$ 成比例的函数可称为 θ 的似然函数，记为

$$H(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta),$$

于是上式可写为

$$p(\theta|\boldsymbol{x}) = \frac{H(\theta|\boldsymbol{x})\pi(\theta)}{\int H(\theta|\boldsymbol{x})\pi(\theta)d\theta}.$$

边际分布是一个积分，因此只要此积分有限，积分的特定值就不会为后验分布提供额外信息。故我们可将上式写为

$$p(\theta|\boldsymbol{x}) \propto \pi(\theta)H(\theta|\boldsymbol{x}).$$

简单来说，贝叶斯理论通过使用先验 $\pi(\theta)$ 和来自于数据 $\boldsymbol{x}$ 的似然信息通过贝叶斯公式得到后验 $p(\theta|\boldsymbol{x})$。理论上，贝叶斯方法为统计推断提供了简单的替代方法——所有的推断都遵循后验分布。然而，在实践中，我们仅能在最基本的问题中通过直接计算解得到后验分布。大多数贝叶斯分析需要复杂的计算，例如使用模拟法，这种方法旨在生成服从后验分布的样本，作为经验样本对后验分布做出推断，估计相应的统计量。

2.1.2 先验分布的选取

参数的先验是一个概率分布，表示在处理数据前我们对参数的看法。2.1.1节提到，贝叶斯理论中的概率为衡量我们相信一个随机事件的程度。在此定义下，概率是主观的，这也意味着先验应该同样是主观先验，也被称为有信息先验，例如共轭先验。然而并非所有人都会同意这种概念，特别在指定先验分布时，我们也认为结果应该更加客观，即使用对后验分布影响最小的先验分布，也被称为无信息先验，例如均匀先验和 Jeffreys 先验。

1. 共轭 (Conjugate) 先验属于有信息先验，指先验和后验来自同一分布族，意味着先验与后验有相同类型的分布。例如，当似然为二项分布 $Bin(n,\theta)$，若先验为 Beta 分布，则后验也具有 Beta 分布的核。其他的组合例子有：正态分布/正态分布，伽马分布/泊松分布，伽马分布/伽马分布以及伽马分布/贝塔分布等。

2. 均匀 (Uniform) 先验属于无信息先验。此时先验 $\pi \sim U(a,b)$，即参数在 (a,b) 内的取值等可能。其中比较特殊的一种均匀先验为非正常先验，即参数在实数内取值等可能，有

   $$\int \pi(\theta)d\theta = \infty,$$

   在实证中常用 $N(0,10000)$ 近似表示。由于没有查阅到任何历史经验，本文使用非正常先验进行贝叶斯推断。

3. Jeffreys 先验是一种无信息先验，它取决于似然而非当前的观测数据。Jeffreys 先验定义为

   $$\pi(\theta) \propto |I(\theta)|^{1/2},$$

   其中 $I(\theta)$ 为基于似然的 Fisher 信息矩阵

   $$I(\theta) = -E\left[\frac{\partial^2 \log H(\theta|\boldsymbol{y})}{\partial\theta^2}\right].$$

§2.2 Logistic 回归模型

2.2.1 广义线性模型

广义线性模型理论起源于 Nelder 和 Wedderburn(1972)[27]，Wedderburn(1974)[28]，随后在 McCullagh 和 Nelder(1989)[29] 的专著中流行。这类模型将线性模型的理论和方法扩展到具有非正态特征的数据。在此理论发展之前，非正态数据的建模通常依赖于数据转换，而选择转换是为了使数据具备对称性、正态性和方差齐性，但这种做法通常会引起误差。

广义线性模型使用链接函数进行变换，但它应用于确定性成分，即数据平均值。此外，广义线性模型考虑了数据的分布，而不是假设数据的转换会得到满足标准线性建模条件的正态分布数据。其模型设置过程如下：

- 响应部分是一个线性预测器 η，满足 $\eta = \boldsymbol{x}'\boldsymbol{\beta}$。$\eta$ 为参数的线性函数，且与普通线性模型不同，其并不代表数据的均值。
- 链接函数 $g(\cdot)$ 将 η 与均值联系起来，有 $g(\mu) = \eta$。链接函数是单调可逆的，均值可由链接函数的逆表示，即 $\mu = g^{-1}(\eta)$。

2.2.2 logit 链接函数

对于二分类或二项分布的数据，最常见的链接函数为 logit 链接，$g(t) = \log\{t/(1-t)\}$。因此带有 logit 链接和单回归变量的广义线性模型可表示为

$$\log\left\{\frac{\mu}{1-\mu}\right\} = \beta_0 + \beta_1 x.$$

这就是我们所熟知的 logistic 回归模型。

§2.3 数据介绍

研究生教育是世界上各个国家高等教育的重要组成部分之一，国家培养的研究生是社会高层人才的主要来源。研究生招生是研究生教育的第一个重要环节，也是不容忽视的一个环节。美国名校众多，在世界上极具名气，其培养出的研究生也表现优异、广受称赞。美国的研究生教育起初借鉴英国的教育方式，后续转向借鉴德国，在此过程中衍生出独特的教育特色，成为研究生教育强国。美国的硕士研究生招生制度在发展的过程中，形成以申请者的 GRE、GPA 成绩等认知能力作为录取标准，同时注重考察非认知能力，独具特色的招生制度。

本文选取的数据集是 Mohan S Acharya 等[1]为预测研究生入学录取情况所建立的数据集，为帮助学生在他们凭借个人履历筛选大学时，可以使他们对自己进入某所大学的机会有一个公平的认知。数据集的灵感来自于 UCLA 的毕业生数据，数据集包含多个变量，这些变量在申请硕士生项目中被认为是重要的 (考试分数和 GPA 都是以前的格式)。变量情况如表2.1：

[1]Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

表 2.1　数据变量介绍.
Table 2.1　Variables introduction.

变量名	含义	变量类型	变量名	含义	变量类型
GRE	GRE 分数 (满分 340)	数值型	TOEFL	TOEFL 分数 (满分 120)	数值型
UR	申请的大学评级	数值型	SOP	目的陈述分数	数值型
LOR	推荐信分数	数值型	CGPA	本科时期 GPA	数值型
RE	是否有科研经历	因子型	COA(y)	录取概率	数值型

§2.4　数据可视化与预处理

数据集共有 500 条观测，8 个变量 (包括响应变量)。首先检查数据集是否具有缺失值。使用 R 软件对数据集进行处理，结果显示数据并无缺失值。对每个变量整合，计算各变量的统计特征，显示结果如表2.2

表 2.2　各变量统计特征.
Table 2.2　Summary of variables.

变量名	Mean	Sd	Median	Min	Max	偏度	峰度	下四分位数	上四分位数
GRE	316.47	11.30	317.00	290.00	340.00	-0.04	-0.73	308.00	325.00
TOEFL	107.19	6.08	107.00	92.00	120.00	0.10	-0.67	103.00	112.00
UR	3.11	1.14	3.00	1.00	5.00	0.09	-0.82	2.00	4.00
SOP	3.37	0.99	3.50	1.00	5.00	-0.23	-0.72	2.50	4.00
LOR	3.48	0.93	3.50	1.00	5.00	-0.14	-0.76	3.00	4.00
CGPA	8.58	0.60	8.56	6.80	9.92	-0.03	-0.58	8.13	9.04
RE	0.56	0.50	1.00	0.00	1.00	-0.24	-1.95	0.00	1.00
COA(y)	0.72	0.14	0.72	0.34	0.97	-0.29	-0.47	0.63	0.82

偏度定义为

$$\mathrm{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{k_3}{\sigma^3} = \frac{k_3}{k_2^{3/2}},$$

它表示统计数据分布偏斜方向和程度，k_2, k_3 分别表示样本二阶、三阶中心矩。偏度大于 0 具有右偏态，即右尾长，且值越大程度越高；偏度小于 0 具有左偏态，即左尾长，且值越大程度越高。偏度为 0 时显示为对称分布。通过表2.2可看出除 TOEFL 和 UR 变量外其余变量均有左偏倾向，但可以看到偏度值均较小，尾部并不长。

峰度定义为

$$Kurt = \frac{k_4}{k_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2} - 3,$$

它表示概率密度峰部的尖度。正态分布峰度为 0。峰度大于 0 则称为尖峰态，为负则称为低峰态。通过表2.2可以看出各变量峰度均为低峰，较为平坦。

将数据散点图、各变量密度曲线和变量相关系数放到同一图中，如图2.1所示。

由图2.1可知，观测中没有明显离群值。响应变量 COA 为学生被录取的概率，这里将其转换为分类变量，取概率大于 0.8 的观测为可以成功录取，记为 1，低于 0.8 的观测不会被录取，记为 0。分类

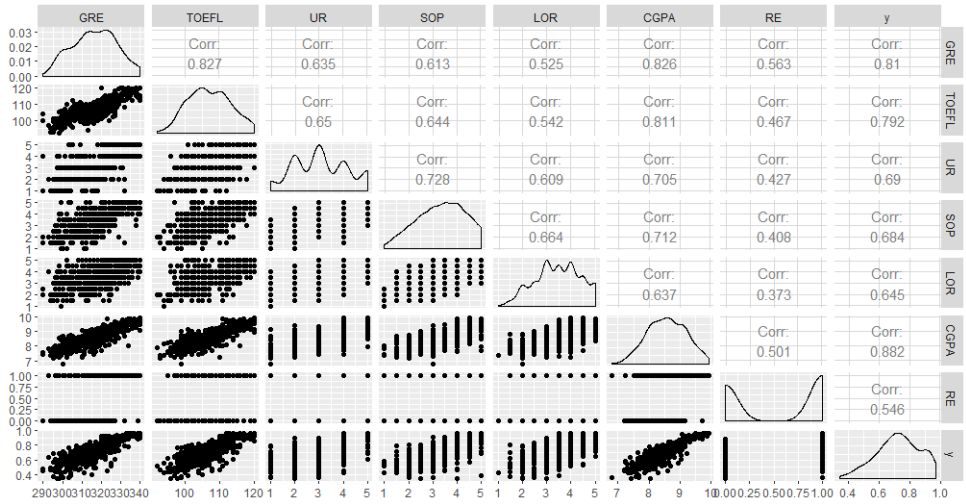

图 2.1　数据散点图.

Figure 2.1　Scttorplot of data.

后结果为 1 的占比为 28.4%。按训练集 : 测试集 = 7 : 3 比例划分数据集，响应变量在训练集和测试集的情况如表2.3

表 2.3　划分数据集.

Table 2.3　Partition dataset.

划分 / COA	Training	Testing
0	68	29
1	283	120

考虑到量纲的影响，对划分后的训练集的数值型变量进行标准化，并应用到测试集，标准化后的数据前六行如表2.4所示：

表 2.4　标准化训练集.

Table 2.4　Standardize training set.

INDEX	GRE	TOEFL	UR	SOP	LOR	CGPA	RE	y0
2	0.68	-0.02	0.77	0.64	1.09	0.49	1	0.00
3	-0.03	-0.52	-0.10	-0.35	0.04	-0.94	1	0.00
4	0.50	0.49	-0.10	0.14	-1.00	0.16	1	0.00
5	-0.21	-0.69	-0.97	-1.34	-0.48	-0.59	0	0.00
6	1.22	1.33	1.64	1.13	-0.48	1.25	1	1.00
8	-0.74	-1.03	-0.97	-0.35	0.57	-1.10	0	0.00

第三章　马尔科夫链蒙特卡洛方法 (MCMC)

马尔可夫链蒙特卡洛方法是一种从后验分布中采样并计算后验性质的模拟方法。MCMC 方法从目标分布中采样，每个样本都依赖于前一个样本，因此有马尔可夫链的概念。Metropolis & Ulam(1949)[30] 以及 Metropolis 等 (1953)[1] 提出了所谓的 Metropolis 算法，该算法可以从多变量的联合分布中生成样本序列，是 MCMC 方法的基础。MCMC 方法最早出现在 Tanner 和 Wong(1987)[31] 的主流统计文献中。

§3.1　蒙特卡洛积分

假设现在需要计算某个积分

$$H = \int h(x) f(x) \mathrm{d}x,$$

其中 $h(\cdot)$ 是密度函数为 $f(\cdot)$ 的随机变量 X 的函数。在很多时候我们无法用普通积分方法计算结果，这时可选择使用数值积分的方法进行近似模拟。其中一种方法为蒙特卡洛方法，使用经验均值进行估计：

$$\hat{H} = \frac{\sum_{i=1}^{m} h\left(x^{(i)}\right)}{m}, \tag{3.1}$$

其中 $\left\{x^{(1)}, \ldots, x^{(m)}\right\}$ 是一组取自密度函数 $f(\cdot)$ 的独立随机样本。由强大数定律可知经验均值 $\mathcal{H} = \left\{\sum_{i=1}^{m} h\left(X^{(i)}\right)\right\}/m$ 收敛到真值，且随样本 m 取值越大，结果越趋近于真值，即 $\mathcal{H} \xrightarrow{as} H,\ m \to +\infty.$

贝叶斯推断主要围绕在后验分布 $p(\theta|\boldsymbol{y})$ 上，$\boldsymbol{y}$ 为样本数据，由式 (2.1) 和 (2.2) 可得

$$p(\theta|\boldsymbol{y}) \propto p(\boldsymbol{\theta}) \times p(\boldsymbol{y}|\theta).$$

我们用后验的期望 $E(\theta|\boldsymbol{y}) = \int_{\theta\in\Theta} \theta p(\theta|\boldsymbol{y}) \mathrm{d}\theta$ 来估计参数 θ 的后验均值，这需要进行积分操作。然而后验分布的积分大多时候为复杂形式，无法用常规计算方法得出，于是采用上述的蒙特卡洛方法进行估计。假设我们得到来自后验分布的独立的参数样本 $\left\{\theta^{(1)}, \ldots, \theta^{(m)}\right\}$，就可以根据式 (3.1) 得到参数估计 $(\sum_{i=1}^{m} \theta^{(i)})/m.$

但在很多实际情况下，很难从后验分布 $p(\theta|\boldsymbol{y})$ 中抽取样本，MCMC 方法因此诞生。

§3.2 马尔科夫链

假设现有随机变量序列 $\{Y_0, Y_1, Y_2, \dots\}$，若对任意时间点 $t > 0$，状态 Y_{t+1} 取样于分布 $P(Y_{t+1}|Y_t)$，这个条件分布仅由当前状态 Y_t 决定，即有

$$P\left(Y_{t+1}|Y_t, \dots, Y_1\right) = P\left(Y_{t+1}|Y_t\right). \tag{3.2}$$

由式3.2可看出，当给定 Y_t，Y_{t+1} 的状态只取决于前一步 Y_t 的状态，而不受历史状态 $\{Y_0, Y_1, \dots, Y_{t-1}\}$ 的影响。满足上述性质的序列称为**马尔科夫链**[1]，简称为马氏链，$P(\cdot|\cdot)$ 被称为**转移核**。这里假设讨论的马氏链是时齐的 (或平稳的)，即 $P(\cdot|\cdot)$ 不依赖于时间 t。

下面给出与马氏链相关的定义[32]

定义 3.1. 若对任意状态 $m, n \in S$，由状态 m 到状态 n 的概率为正，即有 $P_{mn}^{(k)} = P(Y_k = n|Y_0 = m) > 0$，此时马氏链 Y 的所有状态是互通的，也称 Y 是**不可约**的。

定义 3.2. 对马氏链状态 k，现有正整数集 $\left\{n \geq 1;\, p_{k,k}^{(n)} > 0\right\}$ 非空，集合中元素的最大公因数 d_i 叫做状态 k 的周期。若对 $\forall n \geq 1$，有 $p_{k,k}^{(n)} = 0$，则状态 k 的周期为 ∞。称 k **非周期**，若 k 的周期为 1。

定义 3.3. 令 $\tau_i = \inf\{n \geq 1; X_n = i\}$ 为状态 i 的首次回访时间，若平均返回时间小于无穷，即 $E(\tau_i) < \infty$，则称状态 i 为**正常返**。

定义 3.4. 若 $\mathbf{u} = (u_i\,;\, i \in S)$ 为不恒为零且非负的有限数列，对任意 $i \in S$ 有

$$u_i = \sum_{k \in S} u_k p_{k,i}. \tag{3.3}$$

当 $\sum_{i \in S} u_i = 1$，即 $(u_i\,;\, i \in S)$ 是概率分布时，称 $\mathbf{u}$ 为 X 的**平稳分布**，也可以叫做不变测度。

定理 3.1. 设 X 为不可约马氏链，X 存在不变测度 $(\pi_i,\, i \in S)$ 当且仅当 X 正常返，且此时 X 的平稳分布唯一。

定理 3.2. 假设我们有初始分布为 π_0 的不可约马氏链 $\{X_t,\, t \geq 0\}$，状态空间为 Ω，则有

$$\pi_t \to \pi, \quad t \to \infty$$

这里 π_t 为马尔科夫链在 t 时间点的边际分布，π 为平稳分布。此定理说明了当马氏链运行相当长时间后 (t 极大)，X_n 的分布为 π，与初始分布无关。

因此, 我们可以通过算法生成马氏链，进行足够多次数的迭代，使其达到平稳状态，且平稳分布为 π。得到的整条马氏链相当于分布 π 的样本，也就是待积后验分布的经验样本，从而可以对后验积分进行估计。

§3.3 Metropolis-Hastings 算法

上节提供了思路，使用马尔科夫链估计 $E(\theta|\boldsymbol{y})$，然而问题在于如何构造一条马尔科夫链使其平稳分布精确地对应后验分布。Metropolis 算法是以其发明者，美国物理学家和计算机科学家 Nicholas C. Metropolis 的名字命名的。该算法简单实用，可用于从任意维数的复杂目标分布中获取随机样本。

[1]本文涉及且介绍的仅具有离散参数和离散状态的情形，即**离散时间马氏链**

假设我们想从概率密度函数为 $f(\theta|\boldsymbol{y})$ 的分布中抽取样本 T，假设 θ^t 是从 f 中抽取的第 t 个样本，初始值为 θ^0。对于第 $t+1$ 次迭代，算法会基于当前样本 θ^t，从**提议分布 (Proposal distribution)** $q(\cdot|\theta^t)$ 中抽取样本，并且决定是否拒绝这个新样本。若新样本被接受，算法基于此新样本继续运行迭代，若新样本被拒绝，算法会选择当前样本重复取样过程。由上一节可知，产生的马氏链需要满足不可约、正常返和非周期的条件，使得平稳分布为待积后验分布。一般情况下，具有与后验分布相同支撑集的提议分布可以使产生的马氏链满足这些条件[33]。

Metropolis 算法假设提议分布 $q\left(\theta_{\text{new}}|\theta^t\right)$ 是对称分布，有 $q\left(\theta_{\text{new}}|\theta^t\right) = q\left(\theta^t|\theta_{\text{new}}\right)$，意味着由 θ_{new} 到 θ^t 的概率与 θ^t 到 θ_{new} 的概率相同。最常见的建议分布是正态分布 $N\left(\theta^t, \sigma\right)$，其中 σ 为固定值。Hastings(1970)[2] 结合了 Metropolis 算法，提出了 Metropolis-Hastings 算法，使用非对称的提议分布 (即 $q\left(\theta_{\text{new}}|\theta^t\right) \neq q\left(\theta^t|\theta_{\text{new}}\right)$)，算法如下：

算法 1 Metropolis-Hastings algorithm

输入: 目标分布 $f\left(\theta|\mathbf{y}\right)$；提议分布 $q(\cdot|\cdot)$
输出: 模拟样本 $\{\theta^0, \theta^1, \theta^2, \cdots, \theta^t\}$

1: 设置 $t = 0$。选择起始点 θ^0，可以是满足 $f\left(\theta^0|\mathbf{y}\right) > 0$ 的任意点
2: **REPEAT**
3: 　使用提议分布 $q\left(\cdot|\theta^t\right)$ 生成一个新样本 θ_{new}
4: 　计算接受概率 $r = \min\left\{\frac{f(\theta_{\text{new}}|\mathbf{y})q\left(\theta^t|\theta_{\text{new}}\right)}{f(\theta^t|\mathbf{y})q(\theta_{\text{new}}|\theta^t)}, 1\right\}$
5: 　从均匀分布 $U(0,1)$ 中任取样本 u
6: 　若 $u < r$ 令 $\theta^{t+1} = \theta_{\text{new}}$；否则令 $\theta^{t+1} = \theta^t$
7: 　令 $t = t + 1$
8: **UNTIL** $t \geq T$

本文选择使用以正态分布为提议分布的随机游走 Metropolis 算法，即下一次选样 X_{t+1} 的对称提议分布以上一次选样的 X_t 为对称中心进行取样，例如提议分布为标准正态时，取样 X_{t+1} 的提议分布为 $N(X_t, \sigma^2)$。

第四章　集成嵌套拉普拉斯近似 (INLA) 方法

直到现在，贝叶斯推断也依赖于马尔科夫链蒙特卡洛方法来模拟参数的后验分布，在多参数情形下会产生极大的时间成本，而实证中需要推断的后验分布往往都基于高维空间。Havard Rue、Martino 和 Chopin(2009)[7] 提出了一种新方法，可以更高效地进行贝叶斯推理。首先，这种方法的目的不是估计模型参数的联合后验分布，而是关注单个参数的边际后验。在多数情况下，使用边际后验足以对模型参数和潜在效应进行推断，不需要处理难以计算的多元后验分布[34]。其次，他们重点针对了属于隐高斯马尔可夫随机场 (Gaussian Markov random fields, GMRF) 的模型。于是他们基于拉普拉斯近似，提出了有关模型参数边际后验分布近似的新方法，也就是集成嵌套的拉普拉斯近似 (INLA) 方法。在这一章，我们规定 $\boldsymbol{x}_{-i}$ 为向量 $\boldsymbol{x}$ 去掉第 i 个元素的部分，$\Gamma(\tau; a, b)$ 为 τ 的分布。

§4.1　拉普拉斯近似

拉普拉斯近似法是很久以前用来做积分近似的方法，具体可参考 Barndorff-Nielsen & Cox(1989, chapter 3.3)[35]。拉普拉斯近似利用二阶泰勒展开，并结合高斯近似来对函数积分进行推断。对函数 $f(x)$ 计算积分：

$$I = \int_x f(x)dx,$$

其中 $f(x)$ 有极大值点 x^*，若 f 在极值点 x^* 有唯一最大值，则积分 I 的大部分贡献可认为来自于 x^* 邻域的积分。做变换 $f(x) = exp(log(f(x)))$，并对 $\log(f(x))$ 在 x^* 处做二阶泰勒展开，可得

$$I = \int_x \exp\left[\log f(x^*) + \left.\frac{\partial \log f(x)}{\partial x}\right|_{x=x^*}(x-x^*) + \frac{1}{2}\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}(x-x^*)^2\right] \mathrm{d}x, \tag{4.1}$$

其中因 x^* 为极值点，有 $\left.\frac{\partial \log f(x^*)}{\partial x}\right|_{x=x^*} = 0$，则

$$I = \exp\left[\log f(x^*)\right] \int_x \exp\left[\frac{1}{2}\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}(x-x^*)^2\right] dx \tag{4.2}$$

$$= f(x^*) \int_x \exp\left[\frac{D}{2}(x-x^*)^2\right] dx \tag{4.3}$$

$$= f(x^*) \int_x \exp\left[-\frac{(x-x^*)^2}{2\hat{\sigma}^2}\right] dx, \tag{4.4}$$

其中 $D = \left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}$，$\hat{\sigma}^2 = -\frac{1}{D} = -1/\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}$

可以看到式 (4.4) 被积函数满足高斯分布 $N(x^*, \hat{\sigma}^2)$ 核，因此可以得到函数的积分近似结果。若

x 为向量形式 $\boldsymbol{x}=\{x_1,\cdots,x_p\}$，则上式可表示为

$$I=f(\boldsymbol{x}^*)\int_{\boldsymbol{x}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^*)^{\mathrm{T}}H(\boldsymbol{x}^*)(\boldsymbol{x}-\boldsymbol{x}^*)\right]d\boldsymbol{x}, \tag{4.5}$$

其中 $\boldsymbol{x}^*$ 为峰值点，$H(\boldsymbol{x}^*)_{p\times p}$ 为 $\log(f(\boldsymbol{x}))$ 在 $\boldsymbol{x}^*$ 处的黑塞矩阵。

§4.2 隐高斯模型

INLA 方法的内容离不开隐高斯模型这个概念，这是一个非常实用的抽象概念，包含了绝大部分统计模型，在此基础上的统计推断也可归为一整类。“隐” 意指随机，贝叶斯理论中将未知参数视为随机变量，也可称为隐变量，此时含未知参数的模型为隐模型，隐变量构成的空间称为隐场。[36] 这个代表一类模型的抽象概念是由三阶层次模型表达式而得，假设观测 $\boldsymbol{y}$ 条件独立，给定隐高斯随机场 $\boldsymbol{x}$ 和参数 $\boldsymbol{\theta_1}$，有第一阶

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta_1}\sim\prod_{i\in\mathcal{I}}\pi(y_i|x_i,\boldsymbol{\theta_1}).$$

模型的多功能性体现在隐高斯场上，即第二阶

$$\boldsymbol{x}|\boldsymbol{\theta_2}\sim\mathcal{N}\left(\mu\left(\boldsymbol{\theta_2}\right),Q^{-1}\left(\boldsymbol{\theta_2}\right)\right).$$

它包含了统计模型的所有随机项。参数 $\boldsymbol{\theta}=(\boldsymbol{\theta_1},\boldsymbol{\theta_2})$ 有合理的先验分布，控制隐高斯场或数据的似然，即第三阶 $\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta})$。于是后验表示为

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})\propto\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i\in\mathcal{I}}\pi(y_i|x_i,\boldsymbol{\theta}). \tag{4.6}$$

同时需要做出以下假设：

1. 参数 $\boldsymbol{\theta}$ 的个数不应过大，理想情况下在 2 到 5 之间，但不能超过 20。
2. 隐场 $\boldsymbol{x}|\boldsymbol{\theta}$ 的分布为高斯分布，且当维数 n 很大时应属于高斯马尔科夫随机场。
3. 数据 $\boldsymbol{y}$ 关于 $\boldsymbol{x},\boldsymbol{\theta}$ 条件独立，意味着 y_i 仅依赖于隐场部分，例如 x_i。$\boldsymbol{x}$ 的大部分成分是未知的。

将隐高斯模型关联到更为人所熟知的模型，把 $\{x_i,\ i\in\mathcal{I}\}$ 解释为 η_i，后者为关于其他效应的可加模型：

$$\eta_i=\mu+\sum_j\beta_j z_{ji}+\sum_k f^{(k)}(u_{ji}). \tag{4.7}$$

其中 μ 为总截距，z 为具有效应 $\{\beta_j\}$ 的协变量。与通常的广义线性模型不同的地方在于其增加了 $\left\{f^{(k)}(\cdot)\right\}$ 项，是关于变量 u 的未知函数，当 $\left\{f^{(k)}\right\}$ 取不同函数时可得到不同实际应用的模型。在回归模型中，$f^{(k)}(\cdot)$ 可用来缓解协变量的共线性，具体查阅 Fahrmeir and Tutz(2001)[37]。或引入随机影响。常见模型有惩罚样条模型 (Lang and Brezger, 2004)[38] 和随机游走模型 (Fahrmeir and Tutz, 2001; Rue and Held, 2005)[8, 37] 等。在动态模型里，可将 (4.7) 中的 i 替换为时间 t，使得 $f(u_t)=f_t$。这样 $\{f_t\}$ 可表示为离散或连续时间的自回归模型或具有季节效应的时间序列模型 (Kitagawa and Gersch, 1996;West and Harrison, 1997)[39, 40] 等。

通过上述介绍，式 (4.6) 可表示为

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i\in\mathcal{I}} \pi\left(y_i|x_i,\boldsymbol{\theta}\right) \tag{4.8}$$

$$\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i\in\mathcal{I}} \log\left\{\pi\left(y_i|x_i,\boldsymbol{\theta}\right)\right\}\right] \tag{4.9}$$

若对 $\boldsymbol{x}$ 有线性约束的话则表示为 $\boldsymbol{Ax}=\boldsymbol{e}$，其中矩阵 $\boldsymbol{A_{k\times n}}$ 秩为 k。

§4.3 高斯马尔科夫随机场 (GMRF)

4.3.1 有向图和无向图

边没有方向的图称为**无向图**，相对地，边有方向的图叫做**有向图**。可由图4.1和4.2显示二者区别。

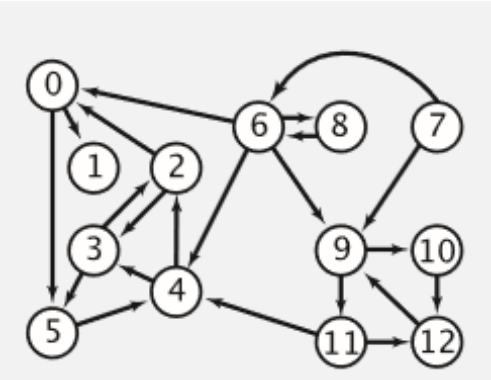

图 4.1　有向图示例.
Figure 4.1　Example of directed graph.

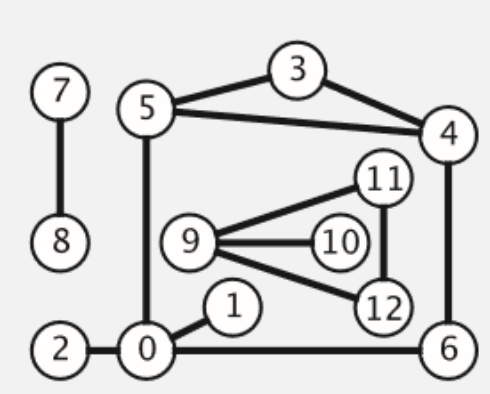

图 4.2　无向图示例.
Figure 4.2　Example of undirected graph.

无向图 $\mathcal{D}$ 可表示为 $\mathcal{D}=(\mathcal{P},\omega)$，其中 $\mathcal{P}$ 表示节点集，若两节点相邻则称其邻接，任意两节点间的边界为 ω，且边界是无方向的。无向图最多有 $n(n-1)/2$ 条边，此时称为无向完全图。

4.3.2 高斯马尔科夫随机场

Rue & Held(2005)[8] 和 Held & Rue (2010)[41] 中对高斯马尔科夫随机场 (GMRF) 有详细定义和介绍，GMRF 中的高斯随机变量 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 具备马尔科夫特性，即3.2节介绍的条件独立。马尔科夫特性使得精度矩阵 (即协方差矩阵的逆) $\boldsymbol{Q}$ 有如下性质：$Q_{ij}=0$ 当且仅当 x_i 与 x_j 在给定其他元素时条件独立。以一阶自回归模型为例：$a_t=\phi a_{t-1}+\epsilon_t,\ t=1,2,\cdots,m$。对于此模型，$a_t$ 与 a_s 的相关性为 $\phi^{|s-t|}$，协方差矩阵为 $m\times m$ 的稠密矩阵[1]。对于 $\forall|s-t|>1$, a_s 与 a_t 在给定其他元素 $a_k,\ k\neq i,j$ 时有条件独立。因此对于一阶自回归模型的精度矩阵，有 $Q_{st}=0,\ \forall|s-t|>1$。此时 $\boldsymbol{Q}$ 为稀疏矩阵，这会使计算更加方便。

[1]稠密矩阵：矩阵中非 0 元素个数占所有元素个数的比例较大，比例较小时称为稀疏矩阵

定义 4.1. 令 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 的条件独立特性可由无向图 $\mathcal{G}$ 表示，$\boldsymbol{x}$ 的密度可表示为

$$\pi(\mathbf{x})=(2\pi)^{-n/2}|\mathbf{Q}|^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{Q}(\mathbf{x}-\boldsymbol{\mu})\right\}, \tag{4.10}$$

其中 $\boldsymbol{\mu}$ 为 $\boldsymbol{x}$ 期望，且 $\boldsymbol{x}$ 具备马尔科夫特性，则称 $\boldsymbol{x}$ 为关于 $\mathcal{G}$ 的高斯马尔科夫随机场 $(GMRF)$[8]。

§4.4 集成嵌套拉普拉斯近似

我们的主要目标是估计后验边际 $\pi(x_i|\boldsymbol{y}),\pi(\boldsymbol{\theta}|\boldsymbol{y})$ 和 $\pi(\theta_i|\boldsymbol{y})$。后验边际可写为

$$\pi\left(x_i|\boldsymbol{y}\right)=\int\pi\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)\pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}, \tag{4.11}$$

$$\pi\left(\theta_j|\boldsymbol{y}\right)=\int\pi(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}_{-j}. \tag{4.12}$$

INLA 方法的关键特点在于使用下式做积分近似

$$\tilde{\pi}\left(x_i|\boldsymbol{y}\right)=\int\tilde{\pi}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)\tilde{\pi}(\boldsymbol{\theta}|vy)\mathrm{d}\boldsymbol{\theta}, \tag{4.13}$$

$$\tilde{\pi}\left(\theta_j|\boldsymbol{y}\right)=\int\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}_{-j}, \tag{4.14}$$

其中 $\tilde{\pi}(\cdot|\cdot)$ 为参数密度的近似。使用 $\pi\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ 和 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ 的近似来得到 $\pi\left(x_i|\boldsymbol{y}\right)$ 的近似。用 $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 对 $\boldsymbol{\theta}_{-j}$ 积分得到 $\pi\left(\theta_j|\boldsymbol{y}\right)$ 的近似。

为实现这一目的，首先做 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ 的拉普拉斯近似，即

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})=\frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\propto\left.\frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\tilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})}, \tag{4.15}$$

其中分母 $\tilde{\pi}_{\mathrm{G}}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ 为高斯近似

$$\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\propto\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}+\sum_{i\in\mathcal{I}}\log\pi\left(y_i|x_i,\boldsymbol{\theta}\right)\right) \tag{4.16}$$

$$=(2\pi)^{-n/2}|\boldsymbol{P}(\boldsymbol{\theta})|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))^T\boldsymbol{P}(\boldsymbol{\theta})(\boldsymbol{x}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right), \tag{4.17}$$

其中 $\boldsymbol{P}(\boldsymbol{\theta})=\boldsymbol{Q}(\boldsymbol{\theta})+diag(\boldsymbol{c}(\boldsymbol{\theta}))$，$\boldsymbol{c}(\boldsymbol{\theta})$ 为 $\sum_{i\in\mathcal{I}}\log\pi\left(y_i|x_i,\boldsymbol{\theta}\right)$ 在点 $\boldsymbol{\mu}(\boldsymbol{\theta})$ 的二阶泰勒展开的负值，$\boldsymbol{\mu}(\boldsymbol{\theta})$ 同式 (4.5) 的 $\boldsymbol{x}^*$ 为峰值点，可用牛顿迭代法推得，具体可参考 (Fahrmeir and Tutz,2001)[37]，下面进行简单介绍. 令 $g_i(x_i)$ 为 $\log\pi(y_i|x_i,\boldsymbol{\theta})$，$\boldsymbol{\mu}^{(0)}$ 为迭代的起始点，是对峰值点的任意猜测。对 $g_i(x_i)$ 在点 $\mu_i^{(0)}$ 做二阶泰勒展开

$$g_i\left(x_i\right)\approx g_i\left(\mu_i^{(0)}\right)+b_ix_i-\frac{1}{2}c_ix_i^2,$$

其中 $\boldsymbol{b}=\{b_i\},\boldsymbol{c}=\{c_i\}$ 均与 $\boldsymbol{\mu}^{(0)}$ 有关。对 $\{\boldsymbol{Q}+diag(\boldsymbol{c})\}\boldsymbol{\mu}=\boldsymbol{b}$ 求解得到 $\boldsymbol{\mu}^{(1)}$，再做二阶泰勒展开得到 $\boldsymbol{b^{(1)}},\boldsymbol{c^{(1)}}$。反复进行此过程直到收敛，即可得到峰值点 $\boldsymbol{\mu^{(*)}}$。

按照上述方法可得到 $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 的近似，下面讨论求 $\pi(\theta_j|\boldsymbol{y})$ 的近似。设 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_m)\in\mathbb{R}^m$。

(a) 通过最优化 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 求得 $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 的峰值点。这一步使用拟牛顿法 (quasi-Newton method) 进行近似，用有限差分法近似梯度，进而通过连续梯度向量的不同对 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 的二阶导进行近似。令 $\boldsymbol{\theta^*}$ 表示峰值点。

(b) 在峰值点 $\boldsymbol{\theta^*}$ 利用有限差分计算黑塞矩阵 $\mathbf{H}>0$。令 $\Sigma=\mathbf{H}^{-1}$ 作为 $\boldsymbol{\theta}$ 的协方差矩阵。为了方便，使用标准化的 $\boldsymbol{z}$ 代替 $\boldsymbol{\theta}$。对 Σ 进行特征分解 $\Sigma=\mathbf{V}\Lambda\mathbf{V}^T$，于是有 $\boldsymbol{\theta}=\boldsymbol{\theta^*}+\mathbf{V}\Lambda^{1/2}\boldsymbol{z}$。若

$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 为正态分布，则 $\boldsymbol{z}$ 服从 $\mathcal{N}(\mathbf{0}, \mathbf{I})$。

(c) 对 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 进行进一步分析。设 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 服从单峰分布，并假设 $m = 2$。图4.3为 $m = 2$ 时模拟的 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 得等高线图。其中新坐标轴对应 $\boldsymbol{z}$，新坐标原点对应峰值点。我们需要定位 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 概率密度的集中区域，具体做法为从峰值点以步长 δ_z 沿 z_1 轴方向进行条件为

$$\log\{\tilde{\pi}(\mathbf{z}|\mathbf{y})\} - \log\{\tilde{\pi}(\mathbf{z}'|\mathbf{y})\} < \delta_\pi \tag{4.18}$$

的搜索。δ_π 视不同情况设定。沿 z_2 轴方向进行相同的操作，结果可得到一组在坐标轴上的 $\{\mathbf{z}'\}$ 点，即下图 (b) 的黑点。使用横纵坐标轴上的不同横纵坐标对应处坐标平面内的不同点，即下图 (b) 的灰点，注意这些灰点同样需要满足式 (4.18) 的条件。得到的点在后面计算 $\tilde{\pi}(x_i|\boldsymbol{y})$ 时同样会用到。

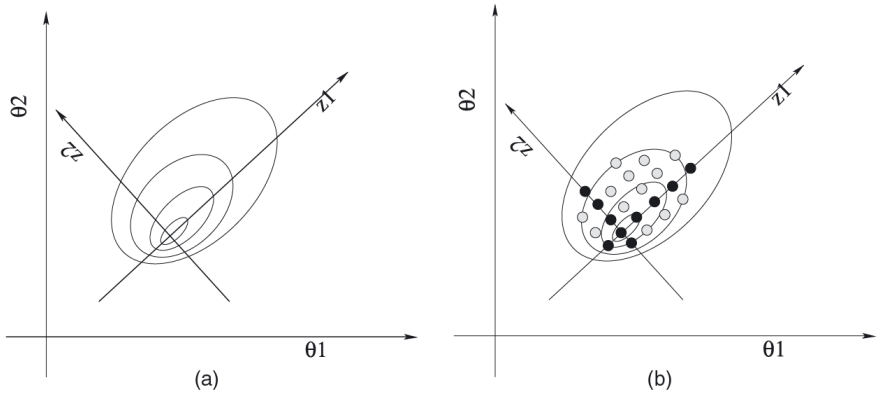

图 4.3　对 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 的探索.
Figure 4.3　Exploration of $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$.

(d) 使用前三步获得的点构造关于 $\log\{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})\}$ 的插值，然后基于插值使用数值积分来计算 $\pi(\theta_j|\boldsymbol{y})$。若对精度进行进一步要求，可调整步骤 (c) 中的搜索步长 δ_z(比如 $\frac{1}{2}$ 或 $\frac{1}{4}$)

接下来对 $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ 进行估计。由式 (4.18) 我们得到了一组点 $\{\boldsymbol{\theta}_k\}$，下面会使用到。这里给出三种近似 $\tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y})$，分别是高斯近似、拉普拉斯近似和简易拉普拉斯近似。尽管简易拉普拉斯近似在精度上会略低于拉普拉斯近似，但可以减少更多的成本。

1. 使用高斯近似

对 $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ 最简单的近似就是高斯近似，在求 $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ 时已经得到 $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$，因此只需要对边际做出推断即可。但 Rue and Martino(2007)[42] 提出，尽管高斯近似是合理的结果，但由于缺少偏态性而易造成误差。

2. 使用拉普拉斯近似

拉普拉斯近似形式如下式

$$\tilde{\pi}_{\mathrm{LA}}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right) \propto \left.\frac{\pi(\mathbf{x},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_{\mathrm{GG}}\left(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\boldsymbol{y}\right)}\right|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})}, \tag{4.19}$$

其中 $\tilde{\pi}_{\mathrm{GG}}$ 是对 $\pi(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\boldsymbol{y})$ 的高斯近似，且 $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ 峰值点，这里的 $\tilde{\pi}_{\mathrm{GG}}$ 与 $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ 无关。

然而，上式意味着计算量的增大，因为精度矩阵依赖于 x_i 和 $\boldsymbol{\theta}$。于是引入两种修正方法。

第一种修正方法是对峰值点进行估计：

$$\mathbf{x}^*_{-i}\left(x_i,\boldsymbol{\theta}\right) \approx \boldsymbol{E}_{\tilde{\pi}_{\mathrm{G}}}\left(\mathbf{x}_{-i}|\boldsymbol{x}_i\right). \tag{4.20}$$

等式右边部分来自高斯近似 $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$。

第二种修正方法基于直觉：只有与 x_i 相邻的 x_j 才会对 x_i 的边际存在影响。假设 x_i 与 x_j 的相关程度随 i 与 j 间的距离增大而减小，那么只有 x_i 周围的一组 $\{x_j\}$，称为 $\boldsymbol{R}_i(\boldsymbol{\theta})$，才能决定 x_i 的边际。利用 (4.20)，当 $i \neq j$ 时对一些 $a_{ij}(\boldsymbol{\theta})$ 有

$$\frac{E_{\tilde{\pi}_G}\left(x_j|x_i\right)-\mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta})\frac{x_i-\mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})}$$

因此可构造 $\boldsymbol{R}_i(\boldsymbol{\theta})$

$$\boldsymbol{R}_i(\boldsymbol{\theta}) = \left\{j : |a_{ij}(\boldsymbol{\theta})| > 0.001\right\},$$

这样就减少了计算 (4.19) 分母项时的阶数，只需要对 $|\boldsymbol{R}_i(\boldsymbol{\theta})| \times |\boldsymbol{R}_i(\boldsymbol{\theta})|$ 阶的稀疏矩阵进行因子分解即可。

通过上面方法简化修正式 (4.19) 后，我们需要计算 (4.19) 在不同 x_i 时的值以获得分布。为选择不同的 x_i 值，我们使用 $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ 的高斯边际，即 $\tilde{\pi}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right) = \mathcal{N}\left\{x_i;\mu_i(\boldsymbol{\theta}),\sigma_i^2(\boldsymbol{\theta})\right\}$ 中的均值和方差，基于由高斯-埃尔米特求积公式给出的相关横坐标选择，为标准化的

$$x_i^{(s)} = \frac{x_i-\mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})}$$

选择不同的值。最终拉普拉斯近似表示为

$$\tilde{\pi}_{\mathrm{LA}}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right) \propto \mathcal{N}\left\{x_i;\mu_i(\boldsymbol{\theta}),\sigma_i^2(\boldsymbol{\theta})\right\}\exp\left\{\text{cubic spline}\left(x_i\right)\right\}.$$

在选择的横坐标点处使用三次样条来拟合 $\log\tilde{\pi}_{\mathrm{LA}}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ 与 $\log\tilde{\pi}(x_i|\boldsymbol{\theta},\boldsymbol{y})$ 的差，然后用正交积分将密度归一化。

3. 使用简易拉普拉斯近似

简易拉普拉斯近似 $\tilde{\pi}_{\mathrm{SLA}}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ 是基于 $\tilde{\pi}_{\mathrm{LA}}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ 的分子分母在 $x_i = \mu_i(\boldsymbol{\theta})$ 附近进行三阶泰勒展开，

$$\log\left\{\tilde{\pi}_{\mathrm{SLA}}\left(x_i^s|\boldsymbol{\theta},\boldsymbol{y}\right)\right\} = C - \frac{1}{2}\left(x_i^{(s)}\right)^2 + \gamma_i^{(1)}(\boldsymbol{\theta})x_i^{(s)} + \frac{1}{6}\left(x_i^{(s)}\right)^3\gamma_i^{(3)}(\boldsymbol{\theta}) + \ldots, \tag{4.21}$$

其中 C 为常数。但因为式 (4.21) 的三阶项为无界，故不能化为密度函数。因此在高斯分布中引入偏度，得到偏态高斯分布 (Azzalini and Capitanio, 1999)[43]

$$\pi_{\mathrm{SN}}(z) = \frac{2}{\omega}\phi\left(\frac{z-\xi}{\omega}\right)\Phi\left(a\frac{z-\xi}{\omega}\right),$$

其中 $\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别为标准正态的密度函数和分布函数，$\xi,\omega>0,a$ 分别为位置、尺度和偏度参数。通过引入偏态高斯分布，使得式 (4.21) 在峰值点的三阶导为 $\gamma_i^{(3)}$。这样 $\gamma_i^{(3)}$ 只影响偏度而 $\gamma_i^{(1)}$ 影响均值。具体内容参考 Rue, H.(2009)[7]。

综合上文，我们得到 $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ 和 $\pi\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ 的近似，也讨论了 $\pi(\theta_j|\boldsymbol{y})$ 的近似，只剩下求 $\pi(x_i|\boldsymbol{y})$ 的近似。考虑到 (4.13) 会产生大量计算，我们选择使用

$$\tilde{\pi}\left(x_i|\boldsymbol{y}\right)=\sum_k \tilde{\pi}\left(x_i|\boldsymbol{\theta}_k,\boldsymbol{y}\right)\tilde{\pi}\left(\boldsymbol{\theta}_k|\boldsymbol{y}\right)\Delta_k,$$

即数值积分的方法进行计算，其中 $\boldsymbol{\theta}_k$ 为满足式 (4.18) 条件而选择的一组点，Δ_k 为权重。当 $\boldsymbol{\theta}$ 的维数小于等于 2 时，使用网格搜索积分节点；当维数大于 2 时，使用中心合成设计 (Central Composite Design,CCD) 定位积分节点。

基于上述内容，简要总结 INLA 算法 (对应 R 软件的 INLA 包)：

1. 求得 $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$，找到其峰值点并定位一组点 $\left\{\boldsymbol{\theta}^1,\cdots,\boldsymbol{\theta}^K\right\}$
2. 代入计算 $\tilde{\pi}(\boldsymbol{\theta}^1|\boldsymbol{y}),\cdots,\tilde{\pi}(\boldsymbol{\theta}^K|\boldsymbol{y})$
3. 使用简易拉普拉斯近似求 $\tilde{\pi}(x_i|\boldsymbol{\theta}^k,\boldsymbol{y}),\ k=1,\cdots,K$，也可指定使用高斯近似或拉普拉斯近似。
4. 计算 $\tilde{\pi}\left(x_i|\boldsymbol{y}\right)=\sum_{k=1}^K \tilde{\pi}\left(x_i|\boldsymbol{\theta}^k,\boldsymbol{y}\right)\tilde{\pi}\left(\boldsymbol{\theta}^k|\boldsymbol{y}\right)\Delta_k$ 其中 Δ_k 为权重。(4.12) 的积分也用类似方法求解。

第五章　实证分析与预测

§5.1　Logistic 回归建模

本节的实证分析使用前章介绍的研究生入学预测数据，响应变量为 COA，重命名为 y，代表学生能否成功被录取，1 表示成功录取，0 表示不会被录取，是二分类的分类变量。解释变量分别为 GRE、TOEFL、UR、SOP、LOR、CGPA、RE，分别表示 GRE 分数、TOEFL 分数、申请学校的评级、目的陈述评价分、推荐信评价分、本科 GPA 和是否有科研经历。其中 RE 变量为二分类变量。

建立 Logistic 回归，响应变量服从伯努利分布，即

$$y_i \sim Bern(p), \quad i = 1, \cdots, n$$

通常 p 通过 logit 链接函数

$$\text{logit}\,(p) = \log(\frac{p}{1-p})$$

与回归协变量联系，此时回归模型可表示为

$$\begin{aligned}\text{logit}\,(p) = \log(\frac{p}{1-p}) = \eta = \beta_0 \ &+ \beta_1 GRE + \beta_2 TOEFL + \beta_3 UR + \beta_4 SOP \\ &+ \beta_5 LOR + \beta_6 CGPA + \beta_7 RE. \end{aligned} \tag{5.1}$$

从式 (5.1) 中可推得

$$p = \frac{1}{1+\exp(-\eta)} = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

使用训练集的数据，$n = 351$，下面分别使用 MCMC 方法和 INLA 方法对参数 $\beta_0 - \beta_7$ 进行估计。

§5.2　基于 MCMC 方法的贝叶斯推断

本节将使用 MCMC 方法对式 (5.1) 进行贝叶斯方式的参数估计。考虑到不同软件的运行时间不同以及操作难易度不同，本文选取 SAS 进行 MCMC 的实现，SAS 软件带有 MCMC 过程模块，专为实现 MCMC 方法而设计。使用 MCMC 过程模块需要设定先验以及似然，因没有经验数据可以借鉴，本文选取非正常先验作为参数 $\beta_0 - \beta_7$ 的先验分布，即实数上的均匀分布，在 MCMC 过程模块中用 $N(0, \text{var} = 10000)$ 进行表示。在 MCMC 抽样中，需要先满足一定次数的迭代才会收敛至平稳，因此需要丢弃马尔科夫链样本的初始部分，使初始的迭代部分对后验分布推断的影响最小化。在 MCMC 过程模块中，参数 `NBI` 的作用就是设置舍弃链初始迭代的次数，设置 `NBI`= 40000；参数

NMC 为 MCMC 总迭代次数，包含需要舍弃的次数，设置 NMC= 100000；参数 SEED 为随机种子，设置 SEED= 2020。设置参数 $\beta_0 - \beta_7$ 初始值为系统随机。

SAS 运行的参数估计结果如表5.1：

表 5.1　MCMC 方法-运行结果.
Table 5.1　Result of MCMC method.

变量名	MEAN	Sd	中位数	25th percentile	75th percentile
Intercept	-4.922	0.813	-4.883	-5.440	-4.348
GRE	0.222	0.681	0.210	-0.231	0.677
TOEFL	0.968	0.572	0.952	0.580	1.354
UR	1.456	0.493	1.447	1.120	1.782
SOP	-0.195	0.558	-0.195	-0.567	0.175
LOR	-0.197	0.409	-0.196	-0.467	0.084
CGPA	4.894	1.049	4.846	4.161	5.567
RE	0.655	0.693	0.646	0.186	1.124

置信度 0.05 时参数的区间估计如表5.2：

表 5.2　MCMC 方法-区间估计.
Table 5.2　Interval estimation using MCMC method.

变量名	Alpha	等尾区间		最大后验密度区间	
Intercept	0.05	-6.623	-3.457	-6.556	-3.416
GRE	0.05	-1.103	1.585	-1.087	1.596
TOEFL	0.05	-0.114	2.117	-0.140	2.087
UR	0.05	0.514	2.443	0.506	2.424
SOP	0.05	-1.298	0.899	-1.335	0.853
LOR	0.05	-1.010	0.594	-1.005	0.596
CGPA	0.05	2.969	7.095	2.916	7.011
RE	0.05	-0.693	2.016	-0.685	2.022

运行结果中，Mean 代表马尔科夫链模拟后验分布样本的均值，即参数的估计，同样还给出了参数的分位数估计，表5.2给出各参数估计的置信区间，置信度 α 为 0.05。SAS 同时给出了各参数后验样本的蒙特卡洛标准误，并与标准差做比，结果如表5.3，比值越小，说明参数估计越可靠。

表 5.3　MCMC 方法-参数估计标准误.
Table 5.3　Estimation standard error using MCMC method.

变量名	MCSE	sd	MCSE/sd
Intercept	0.0136	0.8134	0.0167
GRE	0.0120	0.6808	0.0176
TOEFL	0.0090	0.5718	0.0158
UR	0.0077	0.4927	0.0157
SOP	0.0105	0.5581	0.0188
LOR	0.0068	0.4085	0.0165
CGPA	0.0162	1.0489	0.0155
RE	0.0117	0.6925	0.0169

图5.1为模拟出的马尔科夫链轨迹图，自相关图以及模拟后验分布概率密度图

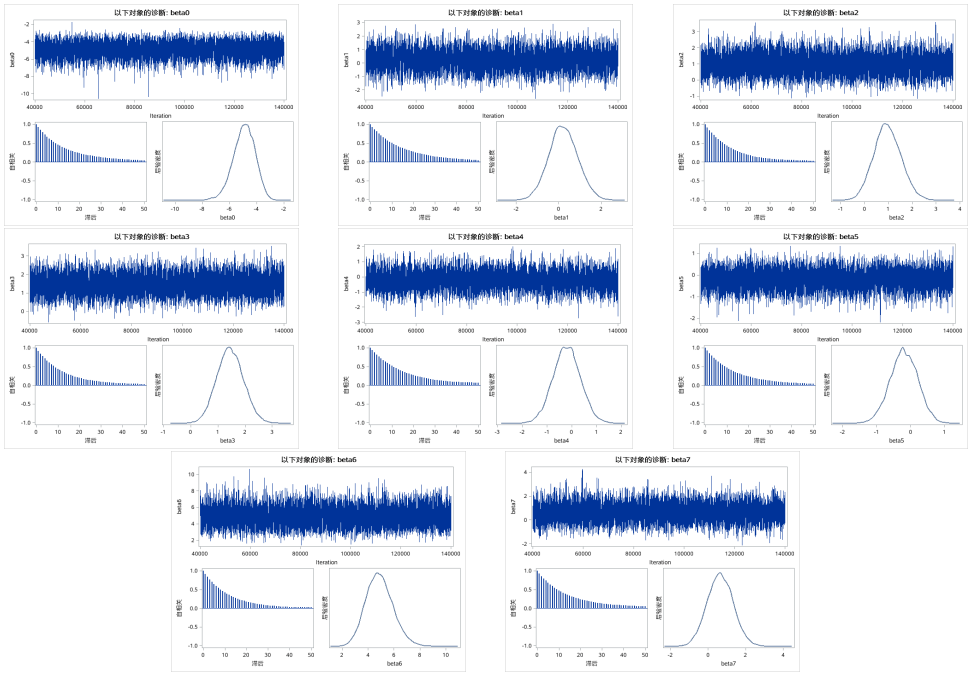

图 5.1　MCMC 模拟结果图.
Figure 5.1　MCMC simulation result.

从图中可以看到，各参数的轨迹均较为稳定，由于选择无信息先验，因此参数自相关降低较缓，但最终均显示平稳。

§5.3　基于 INLA 方法的贝叶斯推断

本节利用 R 软件的 INLA 包实现 INLA 方法，对式 (5.1) 参数进行估计。使用训练集数据建模。在 `inla` 函数中，参数 `family` 对应似然 `likehood`，本例中响应变量服从伯努利分布，也可看做实

验次数为 1 的二项分布，故设置 `control.family = 'binomial'`，并设置 logit 链接。运行结果如表5.4

表 5.4　INLA 运行结果.
Table 5.4　Result of INLA method.

变量名	均值	标准差	2.5% 分位数	中位数	97.5% 分位数
Intercept	-4.780	0.750	-6.376	-4.736	-3.427
GRE	0.214	0.647	-1.022	0.202	1.519
TOEFL	0.918	0.540	-0.098	0.902	2.025
UR	1.416	0.468	0.531	1.404	2.368
SOP	-0.187	0.540	-1.256	-0.184	0.866
LOR	-0.190	0.391	-0.968	-0.186	0.567
CGPA	4.724	1.002	2.896	4.674	6.831
RE	0.636	0.659	-0.625	0.624	1.963

表5.4中 Intercept 为截距项，即 β_0。变量 SOP 与 LOR 参数均为负数，意味着对响应变量具有负效应；变量 CGPA 参数估计值最大，说明本科时期的 GPA 对于判断该生能否被录取的影响最大，远高于其余变量；变量 GRE、SOP 与 LOR 的参数估计值相对较小，意味着 GRE 分数、自我陈述与推荐信的评价分数对于判断该生能否录取的影响程度相对其余变量较小。同样地，得到的后验概率密度曲线图如图5.2

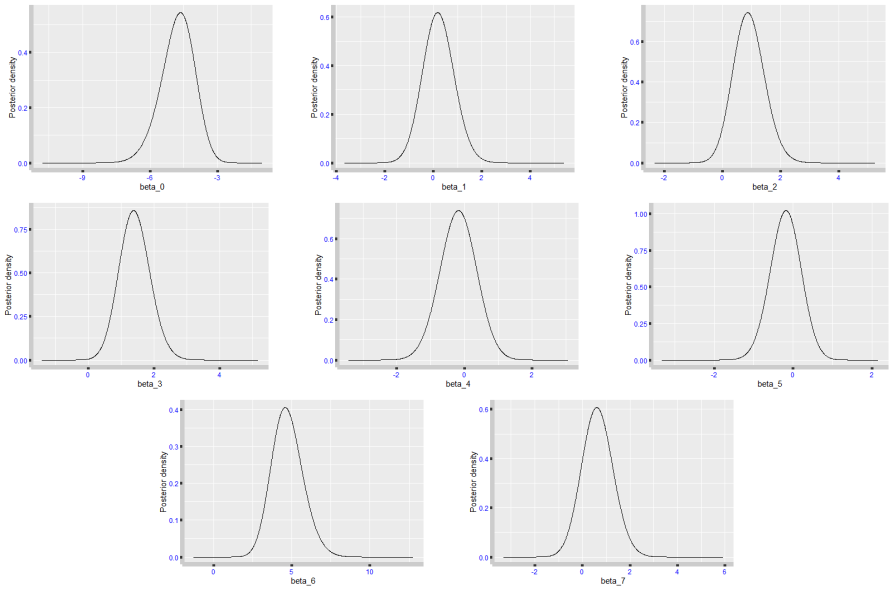

图 5.2　INLA 方法得到的参数后验分布.
Figure 5.2　Posterior distribution using INLA method.

第六章　INLA 方法与 MCMC 方法对比

§6.1　INLA 方法与 MCMC 方法对比

本节整理5.2节和5.3节中基于 INLA 方法与 MCMC 方法的模型参数估计结果，从参数估计结果、DIC 值以及运行时间三个方面对比两种方法，考察 INLA 方法能否准确快速的给出参数的贝叶斯估计。

6.1.1　参数估计结果对比

两种方法对式 (5.1) 中的参数的估计结果如表6.1，表中给出了每个参数的后验均值、中位数和区间估计等。

表 6.1　INLA 方法与 MCMC 方法运行结果对比.
Table 6.1　Comparison of INLA method and MCMC method results.

变量名	估计方法	均值	标准差	中位数	等尾区间		最大后验密度区间	
Intercept	INLA	-4.780	0.750	-4.736	-6.376	-3.427	-6.289	-3.363
	MCMC	-4.922	0.813	-4.883	-6.623	-3.457	-6.556	-3.416
GRE	INLA	0.214	0.647	0.202	-1.022	1.519	-1.046	1.490
	MCMC	0.222	0.681	0.210	-1.103	1.585	-1.087	1.596
TOEFL	INLA	0.918	0.540	0.902	-0.098	2.025	-0.127	1.988
	MCMC	0.968	0.572	0.952	-0.114	2.117	-0.140	2.087
UR	INLA	1.416	0.468	1.404	0.531	2.368	0.509	2.340
	MCMC	1.456	0.493	1.447	0.514	2.443	0.506	2.424
SOP	INLA	-0.187	0.540	-0.184	-1.256	0.866	-1.249	0.868
	MCMC	-0.195	0.558	-0.195	-1.298	0.899	-1.335	0.852
LOR	INLA	-0.190	0.391	-0.186	-0.968	0.567	-0.960	0.571
	MCMC	-0.197	0.409	-0.196	-1.010	0.594	-1.005	0.596
CGPA	INLA	4.724	1.002	4.674	2.896	6.831	2.813	6.727
	MCMC	4.894	1.049	4.846	2.969	7.095	2.916	7.011
RE	INLA	0.636	0.659	0.624	-0.625	1.963	-0.650	1.935
	MCMC	0.655	0.693	0.646	-0.693	2.016	-0.685	2.022

通过上表得出结论：

- 无论从均值还是中位数方面看，INLA 与 MCMC 对各变量的正负效应判断是一致的，且对于正效应变量，基于 INLA 的参数估计均小于基于 MCMC 的参数估计，对于负效应变量，基于 INLA 的参数估计均大于基于 MCMC 的参数估计。
- INLA 方法得到各参数估计标准差均小于对应 MCMC 方法得到的参数估计标准差，说明 INLA 方法得到的参数估计相对 MCMC 方法的参数估计更为可靠。
- 无论是等尾区间还是最大后验密度区间的估计，INLA 方法得到的区间长度均小于 MCMC 方法的结果，说明 INLA 对参数的定位较 MCMC 方法要更为精准。

6.1.2 DIC 值对比

偏差信息准则 (Deviance information criterion)(Spiegelhalter 等, 2002)[44] 是一种模型评价指标，在此之前的还有 AIC 准则和 BIC 准则。DIC 准则是对这两个准则的替代。DIC 值计算时使用后验密度，这意味着它考虑了先验信息。该标准可以应用于非嵌套模型和具有非独立同分布数据的模型。DIC 的计算并不繁琐，不需要像 AIC 和 BIC 那样在参数空间上最大化。DIC 越小表示拟合情况越好。

令 θ 为模型参数，DIC 定义为

$$DIC = \overline{D(\boldsymbol{\theta})} + p_D = D(\bar{\boldsymbol{\theta}}) + 2p_D \tag{6.1}$$

其中 $D(\boldsymbol{\theta}) = 2(\log(f(\boldsymbol{y})) - \log(p(\boldsymbol{y}|\boldsymbol{\theta})))$，称为贝叶斯偏差；$p(\boldsymbol{y}|\boldsymbol{\theta})$ 为似然函数；$f(\boldsymbol{y})$ 定义为与数据有关的函数，是一个常数，在比较具有相同似然函数的不同模型时是没有影响的，因此在比较 DIC 值时，这一部分的计算通常被忽略；$\bar{\boldsymbol{\theta}}$ 是后验均值；$\overline{D(\boldsymbol{\theta})}$ 是偏差的均值，度量模型对数据拟合的程度；$D(\bar{\boldsymbol{\theta}})$ 是后验均值 $\bar{\boldsymbol{\theta}}$ 的偏差，等价于 $-2log(p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$，意味着是得到的最好后验估计的偏差；$p_D$ 是有效参数个数，等于 $\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$，作为惩罚项纠正参数较多时对偏差的影响，其值越大模型相对较复杂。

表6.2给出了基于 INLA 与 MCMC 方法计算的模型 DIC 值

表 6.2　INLA 方法与 MCMC 方法的 DIC 值.
Table 6.2　DIC value of INLA method and MCMC method.

方法	$D(\bar{\boldsymbol{\theta}})$	$\overline{D(\boldsymbol{\theta})}$	p_D	DIC
INLA	109.65	117.03	7.38	124.41
MCMC	109.79	117.60	7.81	125.41

观察上表结果发现，基于 MCMC 方法的模型 DIC 值略大于基于 INLA 的模型 DIC 值，但二者差别极小，可以认为两种方法的估计结果几乎一样好，也可以说明 INLA 方法能较好的做出参数的贝叶斯推断。

6.1.3 运行时间对比

使用 INLA 方法和 MCMC 方法对 Logistic 回归模型进行参数估计，记录程序运行时间，MCMC 方法通过 SAS(9.4 版本) 实现，INLA 方法通过 R(3.6.3 版本) 实现。运行时间如表6.3

表 6.3　INLA 方法与 MCMC 方法的运行时间.
Table 6.3　Runtime of INLA method and MCMC method.

方法	时间 (s)
INLA	5.03
MCMC	22.19

可以看出 INLA 方法的程序运行明显快于 MCMC 方法。综合前两节，INLA 方法能够同 MCMC 方法一样做出准确的贝叶斯推断，且 INLA 方法的结果在准确度和运行时间上优于 MCMC 方法。

§6.2　预测结果对比

使用后验均值作为参数的贝叶斯估计，提取基于 INLA 方法得到的参数估计，带入 Logistic 回归方程。将测试集的观测数据代入回归方程得到测试集每条观测的入学概率预测 $\hat{p}_i$。当 $\hat{p}_i > 0.8$ 时，认定该学生被录取，即可得到测试集上的响应变量预测结果，并计算混淆矩阵。同时使用 KNN、RandomForest 和 XGboost 三种机器学习方法对训练集进行建模，在测试集上进行预测，并计算混淆矩阵。混淆矩阵也叫误差矩阵，可显示模型的预测精度，行表示预测的类别，列表示真实的类别。以 2×2 形式的混淆矩阵为例，如表6.4：

表 6.4　混淆矩阵组成.
Table 6.4　Composition of confusion matrix.

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP
	Negative	FN	TN

注：字母 T 和 F 分别对应 TRUE 和 FALSE，P 和 N 分别对应 Positive 和 Negative

我们希望模型对分类的预测越准越好，因此对应到混淆矩阵中我们希望 FP 与 FN 的数量越小越好。因此针对混淆矩阵延伸了多个指标，定义如下：

- 准确率 (Accuracy)= $\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$
- 灵敏度 (Sensitivity)= 召回率 (Recall)= $\frac{TP}{TP+FN} \times 100\%$
- 特异度 (Specificity)= $\frac{TN}{TN+FP} \times 100\%$
- 精确率 (Precision)= $\frac{TP}{TP+FP} \times 100\%$
- F1= $\frac{2Precision \cdot Recall}{Precision+Recall}$

- Kappa= $\frac{p_0-p_c}{1-p_c}$，其中 $p_0=\frac{TP+TN}{TP+TN+FP+FN}$，$p_c=\frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+TN+FP+FN)^2}$

其中灵敏度代表真实值为 Positive 的数据中模型预测正确的比例；特异度代表真实值为 Negative 的数据中模型预测正确的比例；精确率为模型预测为 Positive 的数据中预测正确的比例；F1 值与 Kappa 值都是模型评价指标，取值 0 到 1 间，越接近 1 代表模型越好。

6.2.1 各指标值对比

整理基于 INLA 方法估计参数的 Logistic 回归模型、KNN、RandomForest 和 XGboost 四种模型得到的混淆矩阵，整理结果如表6.5：

表 6.5 各方法混淆矩阵.

Table 6.5 Confusion matrices of each method.

		真实值				
预测值	**INLA**	0	1	**KNN**	0	1
	0	105	11	0	107	14
	1	2	31	1	0	28
	RF	0	1	**XG**	0	1
	0	106	13	0	105	7
	1	1	29	1	2	35

可以看到，四种模型均会偏向于将实际上被录取的情况预测为不被录取，即表6.4中 FP 的值较大。其中，基于 INLA 方法的 Logistic 回归模型的 FP 值小于 KNN 和 RandomForest 的 FP 值，说明前者在此处的误判情况优于后两者。对各方法计算延伸的指标值，结果如表6.6：

表 6.6 模型评价指标比较.

Table 6.6 Comparison of model evaluation indices.

指标 / 方法	准确率	灵敏度	特异度	精确率	F1 值	Kappa 值
Log-INLA	0.913	0.738	0.981	0.939	0.827	0.770
KNN	0.906	0.667	1.000	1.000	0.800	0.742
随机森林	0.906	0.691	0.991	0.967	0.806	0.746
XGboost	0.940	0.833	0.981	0.946	0.886	0.845

观察上表结果，从准确率方面来看，基于 INLA 的 Logistic 回归模型和 XGboost 的预测效果优于 KNN 和随机森林；从灵敏度与特异度方面来看，KNN 与随机森林的预测明显不平衡，较容易将真实值为 1 的观测预测为 0，即倾向判定学生不被录取，而基于 INLA 的 Logistic 回归模型和 XGboost 的误判情况要优于 KNN 和随机森林；从 F1 与 Kappa 值来看，XGboost 模型最好，基于 INLA 的 Logistic 回归模型其次，KNN 模型和随机森林模型相对最差。综合上述内容，可以看到基于 INLA 方法估计参数的 Logistic 回归模型的预测效果较为理想，并不逊色于机器学习算法。

6.2.2 各曲线图对比

除上一节介绍的各指标外，由指标构成的曲线也可以反映模型好坏，起到评价模型的作用。常见的评价曲线有特异度-灵敏度曲线、召回率-精度率曲线、ROC 曲线和提升图。曲线的组成原理为：设定 0 到 1 间的一个阈值 p_i，对阈值用预测概率进行类别判定 ($\hat{p} > p_i$ 取 1，否则取 0)，得到的判定结果可生成混淆矩阵并计算相应指标值，即可得到一组曲线坐标。设定不同阈值可获得多组坐标，最终得到相应曲线。上述曲线中前两种的横纵坐标由曲线名称可知，对于 ROC 曲线，横坐标为 1− 特异度，纵坐标为灵敏度。对于提升图，纵坐标为提升值，定义为 $\frac{TP}{TP+FP}/\frac{TP+FN}{TP+TN+FP+FN}$；横坐标为预测成 Positive 的比例，即 $\frac{TP+FP}{TP+TN+FP+FN} \times 100\%$。另将 ROC 曲线下方与坐标轴所围成的面积定义为 AUC 值，也是模型评价指标的一种，取值在 0.5 到 1 之间，且越接近 1 意味着评价越好。

图6.1到6.4为整理的各方法的曲线图，其中 ROC 图的图例带有计算的 AUC 值。

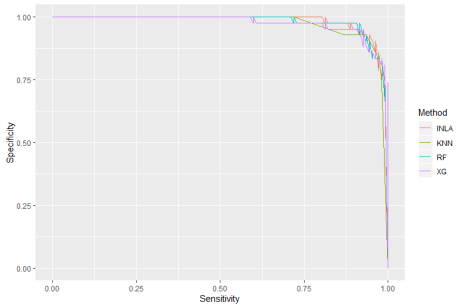

图 6.1　敏感度-特异度曲线.
Figure 6.1　Sensitivity-specificity curve.

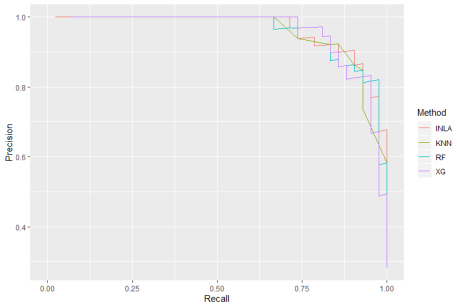

图 6.2　召回率-精确率曲线.
Figure 6.2　Recall-precision curve.

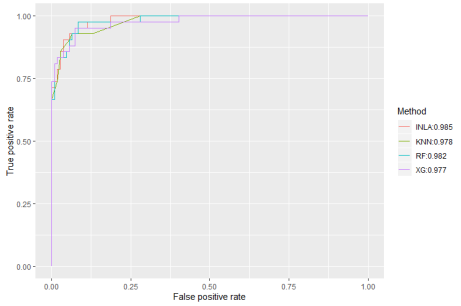

图 6.3　ROC 曲线图.
Figure 6.3　ROC curve.

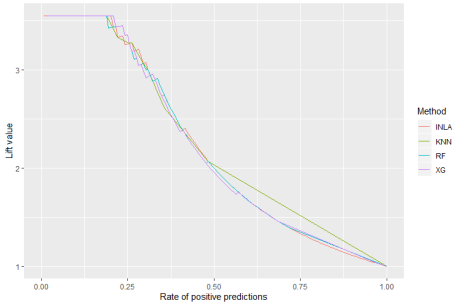

图 6.4　提升图.
Figure 6.4　Lift chart.

对于敏感度-特异度曲线图和召回率-精确率曲线图，曲线越靠近右上角说明模型越好；对于 ROC 图，曲线越靠近左上角说明模型越好。由图6.1,6.2可得，各模型的效果都较为接近，基于 INLA 的 Logistic 回归模型表现略优于其他方法；从 ROC 曲线图6.3可知，基于 INLA 的 Logistic 回归模型效果最好，AUC 值最高；针对提升图，本文所用实证例中判断学生能否被录取的概率阈值理论上应取较高的数值，意味着预测成 Positive 的比例并不高，因此在提升图中，应偏向于观察横坐标较小时图线的表现。则在图6.4中可以看到基于 INLA 的 Logistic 回归模型和 XGboost 模型的表现优于 KNN 与随机森林模型。综合来讲，基于 INLA 的 Logistic 回归模型在预测方面效果十分理想，各指标值与曲线的表现也优于大部分机器学习算法。

第七章　总结与展望

本文介绍了一种对参数做贝叶斯估计的新方法——INLA 方法，其将拉普拉斯方法与数值积分方法结合，可对参数后验分布做出较为精确的推断，且运算时间远小于目前泛用性最广的马尔科夫链蒙特卡洛模拟（MCMC）方法。文中使用研究生入学预测数据集，建立 Logistic 回归模型并对其参数进行了基于 MCMC 和 INLA 方法的贝叶斯估计：使用 SAS 软件进行 MCMC 方法的实现，使用 R 软件进行 INLA 方法的实现，并记录估计结果和运行时间，同时计算各自的 DIC 值，以此对两种方法进行比较。结果显示通过 INLA 方法可以准确得到与 MCMC 相近的参数估计，且前者程序的运行时间远小于后者，说明 INLA 方法在进行参数贝叶斯估计上有很好的的效果，对 MCMC 方法也有很好的的替代性。本文最后使用参数估计结果带入模型对数据进行了预测，并与 KNN、随机森林和 XGboost 三种机器学习算法对比了模型预测效果，结果显示使用贝叶斯方法估计参数的 Logistic 回归模型也具备较好的预测效果，并不明显逊于这三种机器学习算法。

在探究过程中，本文仍可从以下几点进行改进：首先针对实例分析中的数据集，整个数据集的样本量只有 500 条，相对较小，没有很好地体现在大数据背景下 INLA 方法所需时间成本极小的优势；其次对于模型参数的贝叶斯估计，因为没有找到合适的先验信息，本文使用了无信息先验作为先验分布。若能选取到更大的数据集以及合理的先验分布，可以更好、更严谨的体现 INLA 方法的优势。随着贝叶斯理论体系的不断完善以及实证中数据集的不断扩大，将贝叶斯方法与机器学习、深度学习相结合必将是未来研究的热点，需要做更多的探索和尝试。

参考文献

[1] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. Equation of State Calculations by Fast Computing Machines[J/OL]. The Journal of Chemical Physics, 1953, 21(6): 1087-1092. eprint: https://doi.org/10.1063/1.1699114. https://doi.org/10.1063/1.1699114. DOI: 10.1063/1.1699114.

[2] Hastings, W. Monte Carlo Sampling Methods Using Markov Chains and Their Applications[J]. Biometrika, 1970, 57: 97-109.

[3] Geman, S., Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, PAMI-6(6): 721-741. DOI: 10.1109/TPAMI.1984.4767596.

[4] Gelfand, A., Hills, S., Racine-Poon, A., Smith, A. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling[J]. Journal of The American Statistical Association - J AMER STATIST ASSN, 1990, 85: 972-985. DOI: 10.1080/01621459.1990.10474968.

[5] Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D. WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility[J/OL]. Statistics and Computing, 2000, 10(4): 325-337. https://doi.org/10.1023/A:1008929526011. DOI: 10.1023/A:1008929526011.

[6] Lunn, D., Spiegelhalter, D., Thomas, A., Best, N. The BUGS project: Evolution, critique and future directions[J]. Statistics in Medicine, 2009, 28(25): 3049-3067.

[7] Rue, H., Martino, S., Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations[J]. Journal of the Royal Statistical Society:Series B (Statistical Methodology), 2009, 71(2): 319-392.

[8] Rue, H., Held, L. Gaussian Markov Random Fields: Theory and Applications[M]. Chapman: Hall/CRC Press, 2005.

[9] Martino, S., Aas, K., Lindqvist, O., Neef, L., Rue, H. Estimating stochastic volatility models using integrated nested Laplace approximations[J]. The European Journal of Finance, 2011, 17(7): 487-503.

[10] Ugarte, M. D., Adin, A., Goicoa, T., Militino, A. F. On fitting spatio-temporal disease mapping models using approximate Bayesian inference[J/OL]. Statistical Methods in Medical Research, 2014, 23(6): 507-530. eprint: https://doi.org/10.1177/0962280214527528. https://doi.org/10.1177/0962280214527528. DOI: 10.1177/0962280214527528.

[11] Fong, Y., Rue, H., Wakefield, J. Bayesian inference for generalized linear mixed models[J/OL]. Biostatistics (Oxford, England), 2010, 11(3): 397-412. https://europepmc.org/articles/PMC2883299. DOI: 10.1093/biostatistics/kxp053.

[12] Riebler, A., Held, L. Projecting the future burden of cancer: Bayesian age-period-cohort analysis with integrated nested Laplace approximations[J]. Biom. J., 2017, 59: 531-549.

[13] Natário, I., Oliveira, M., Marques, S. Using INLA to Estimate a Highly Dimensional Spatial Model for Forest Fires in Portugal[J]. ICE Selected Engineering Papers, 2013, 10. DOI: 10.1007/978-3-319-05323-3_23.

[14] Kandt, J., Chang, S S., Yip, P., Burdett, R. The spatial pattern of premature mortality in Hong Kong: How does it relate to public housing?[J/OL]. Urban Studies, 2017, 54(5): 1211-1234. eprint: https://doi.org/10.1177/0042098015620341. https://doi.org/10.1177/0042098015620341. DOI: 10.1177/0042098015620341.

[15] Gomez-Rubio, V., Bivand, R. S., Rue, H. Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximation[J]. ArXiv e-prints, 2017, arXiv:1703.01273: arXiv:1703.01273. arXiv: 1703.01273 [stat.CO].

[16] Santermans, E., Robesyn, E., Ganyani, T., Sudre, B., Faes, C., Quinten, C., Van Bortel, W., Haber, T., Kovac, T., Van Reeth, F. Spatiotemporal Evolution of Ebola Virus Disease at Sub-National Level during the 2014 West Africa Epidemic: Model Scrutiny and Data Meagreness[J]. PLoS ONE, 2016.

[17] Niemi, J., Mittman, E., Landau, W., Nettleton, D. Empirical Bayes analysis of RNA-seq data for detection of gene expression heterosis[J]. J. Agric. Biol. Environ. Stat, 2015, 20(4): 614-628.

[18] Martino, S., AKERKAR, R., Rue, H. Approximate Bayesian Inference for Survival Models[J]. Scandinavian Journal of Statistics, 2011, 38: 514-528. DOI: 10.1111/j.1467-9469.2010.00715.x.

[19] Wang, X. F. Bayesian Nonparametric Regression and Density Estimation Using Integrated Nested Laplace Approximations[J]. Journal of Biometrics and Bio-statistics, 2013, 4(4).

[20] Yue, Y., Rue, H. Bayesian inference for additive mixed quantile regression models[J]. Computational Statistics & Data Analysis, 2011, 55: 84-96. DOI: 10.1016/j.csda.2010.05.006.

[21] Cameletti, M., Rubio, V. G., Blangiardo, M. Bayesian modeling for spatially misaligned health and air pollution data through the INLA-SPDE approach[J]. Spatial Statistics, 2019.

[22] Poggio, L., Gimona, A., Spezia, L., Brewer, M. Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA[J]. Geoderma, 2016, 277: 69-82. DOI: 10.1016/j.geoderma.2016.04.026.

[23] Selle, M., Steinsland, I., Hickey, J. M., Gorjanc, G. Flexible modelling of spatial variation in agricultural field trials with the R package INLA[J]. Theoretical and Applied Genetics, 2019, 132(6): 3277-3293. DOI: 10.1007/s00122-019-03424-y.

[24] 周翔, 姜婷婷, 徐丹, 杨榛, 梁剑平, 王亮. 基于 Logistic 回归建模和马尔可夫链蒙特卡罗方法计算后验描述丁酸梭菌株对于给定辐照剂量区的应答趋势[J]. 原子核物理评论, 2016, 33(4): 500-505.

[25] 付志慧, 武健, 马明玥. Rstan 包在四参数 Logistic 模型参数估计中的应用[J]. 沈阳师范大学学报 (自然科学版), 2019, 37(4): 309-314.

[26] Zucknick, M., Richardson, S. MCMC algorithms for Bayesian variable selection in the logistic regression model for large-scale genomic applications[J]., 2014.

[27] Nelder, J. A., Wedderburn, R. W. M. Generalized Linear Models[J]. Journal of the Royal Statistical Society, 1972, 135(3): 370-384.

[28] WEDDERBURN, R. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method[J]. Biometrika, 1974, 61: 439-447. DOI: 10.1093/biomet/61.3.439.

[29] McCullagh, P., Nelder, J. A. Generalized Linear Models 2nd ed[M]. London: Chapman, 1989.

[30] Metropolis, N., Ulam, S. The Monte Carlo Method[J]. Journal of the American Statistical Association, 1949, 44: 335-341.

[31] Tanner, M. A., Wong, W. H. The Calculation of Posterior Distributions by Data Augmentation[J]. Journal of the American Statistical Association, 1987, 82: 528-540.

[32] 钱敏平, 龚光鲁. 随机过程论 (第二版)[M]. 北京: 北京大学出版社, 1997.

[33] 欧卫星. Copula-MCMC 方法在证券投资组合中的应用研究[D]. 湖南: 湖南大学, 2011.

[34] Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., Lindgren, F. K. Bayesian Computing with INLA: A Review[J/OL]. Annual Review of Statistics and Its Application, 2017, 4(1): 395-421. eprint: https://doi.org/10.1146/annurev-statistics-060116-054045. https://doi.org/10.1146/annurev-statistics-060116-054045. DOI: 10.1146/annurev-statistics-060116-054045.

[35] Barndorff-Nielsen, O., Cox, D. Asymptotic Techniques for Use in Statistics[M]. Boca Raton: Chapman, 1989.

[36] 丁书珍. 变系数模型的 Bayesian-INLA 估计[D]. 新疆: 新疆大学, 2019: 17-18.

[37] Fahrmeir, L., Tutz, G. Multivariate Statistical Modelling Based on Generalised Linear Models (2nd edn.)[M]. New York: Springer-Verlag, 2001.

[38] Lang, S., Brezger, A. Bayesian P-Splines[J]. Journal of Computational & Graphical Statistics, 2004, 13(1): 183-212.

[39] Kitagawa, G., Gersch, W. Smoothness priors analysis of time series[M]. New York: Springer-Verlag, 1996.

[40] West, M., Harrison, J. Bayesian Forecasting and Dynamic Models, 2nd edn[M]. New York: Springer, 1997.

[41] Held, L., Rue, H. Conditional and intrinsic autoregressions[C]//AGelfand, P., Diggle, M., Fuentes, P., Guttorp, p. Handbook of Spatial Statistics. [S.l.]: CRC/Chapman, 2010.

[42] Rue, H., Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models[J]. Journal of Statistical Planning and Inference, 2007, 137(10): 3177-3192.

[43] Azzalini, A., Capitanio, A. Statistical Application of the Multivariate Skew Normal Distribution[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999, 61: 579-602. DOI: 10.1111/1467-9868.00194.

[44] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A. Bayesian measures of model complexity and fit[J/OL]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2002, 64(4): 583-639. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00353. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00353. DOI: 10.1111/1467-9868.00353.

附录 A　附录

§A.1　R 代码

```
setwd('C:\\Users\\lenovo\\Desktop\\毕业论文')
data0 <- read.csv("./data/graduate-admissions/Admission_Predict_Ver1.1.csv",
                  header = T) # 载入数据
plot(density(data0$Chance.of.Admit)) #查看可能性分布曲线
#hist(data0$Chance.of.Admit) #查看毕业可能性直方图
sum(is.na(data0)) # 检查缺失值
data1 <- data0[,-1] # 去掉id列
# 描述性统计
library(psych)
describe(data1,quant = c(.25,.75),type = 1)
# 假设可能性大于0.6的学生可以成功录取
# 可以看到成功录取的学生比例为80%
sum(data1$Chance.of.Admit>0.8)/nrow(data0)
# 将我们关注的变量命名为y
colnames(data1) <- c('GRE','TOEFL','UR','SOP','LOR','CGPA','RE','y')
# 数据可视化
library(GGally)
ggpairs(data1) # 绘制散点图，各变量频率曲线，相关系数

data1$y0 <- ifelse(data1$y>0.8,1,0) # 0.8以上为1, 0.8以下定为0
data1$y0 <- factor(data1$y0)

data <- data1[,-8] # data数据集为去掉y0即概率列
data2 <- data1[,-9] # data2数据集为去掉y即判断列

#### 预处理
```

```
library(caret)
# 进行划分
set.seed(2020)
# 70%为训练集，30%为测试集
Index <- createDataPartition(data$y0,p=0.7,list=F,times=1)
train <- data[Index,]
test <- data[-Index,]
dim(train);str(train)

###    INLA-logistic
library(INLA)
# 使用data数据集进行分析
# 标准化
trainscale <- preProcess(train[,-7],method = c('center','scale'))
traininla <- predict(trainscale,train) # 对训练集标准化
testinla <- predict(trainscale,test) # 对测试集标准化(基于训练集)
####  构造sas使用的数据集
trainmcmc <- traininla
colnames(trainmcmc)[8] <- 'COA'
write.csv(trainmcmc,file = './data/graduate-admissions/MCMC_data3.csv')
###   INLA建模
Ntrial <- rep(1,nrow(traininla)) # 这里是认为每次做一次实验，即trial为1
GAinla <- inla(y0 ~ GRE+TOEFL+UR+SOP+LOR+CGPA+RE, data = traininla,
            Ntrials = Ntrial,control.compute = list(dic = TRUE),
            family = "binomial",
                  control.family = list(
                    control.link = list(model = "logit"))
                )
summary(GAinla)
# 对参数各统计量总结 包括均值标准差和各分位数
beta0.s = inla.zmarginal(marginal=GAinla$marginals.fixed$`(Intercept)`)
# 计算HPD区间
(beta0.hpd = inla.hpdmarginal(0.95,marginal=GAinla$marginals.fixed$RE))
#  画出后验分布
beta0 = data.frame(inla.smarginal(marginal =
                          GAinla$marginals.fixed$`(Intercept)`))
```

```
ggplot(data = beta0,aes(x=x,y=y)) + geom_line() +
  theme(axis.ticks = element_line(size = 2),
        axis.text = element_text(colour = "blue"),
        axis.line = element_line(size = 3, colour = "grey80")) +
  labs(x = expression(beta_7),y = "Posterior density")

# 提取参数进行预测
(cof <- GAinla$summary.fixed$mean)
# 提取训练集数据
traindata <- data.frame(cbind(rep(1,nrow(traininla)),traininla[,-8]))
colnames(traindata)[1] <- 'intercept'
# 对每一行数据乘以得到的参数
result <- apply(t(apply(traindata,1,function(x) cof*x)),1,sum)
# 计算预测概率
prob <- exp(result)/(1+exp(result))
res <- ifelse(prob>0.8,1,0)
(table <- table(res,traininla$y0))
# 得到混淆矩阵
pre <- factor(res);ref <- factor(traininla$y0)
conM.inla <- confusionMatrix(pre,ref,positive = '1')
conM.inla
# 用得到的参数对测试集预测
testdata <- data.frame(cbind(rep(1,nrow(testinla)),testinla[,-8]))
colnames(testdata)[1] <- 'intercept'
# 对每一行数据乘以得到的参数
testresult <- apply(t(apply(testdata,1,function(x) cof*x)),1,sum)
# 计算预测概率
testprob <- exp(testresult)/(1+exp(testresult))
testres <- ifelse(testprob>0.8,1,0)
testpre <- factor(testres);testref <- factor(testinla$y0)
conM.inla <- confusionMatrix(testpre,testref,positive = '1')
conM.inla


### 使用SAS得到MCMC结果进行预测
cof_mcmc <- c(-4.922,0.222,0.968,1.456,-0.195,-0.197,4.894,0.655)
# 提取训练集数据
```

```r
test_mcmc <- data.frame(cbind(rep(1,nrow(testinla)),testinla[,-8]))
colnames(test_mcmc)[1] <- 'intercept'
result_mcmc <- apply(t(apply(test_mcmc,1,function(x) cof_mcmc*x)),1,sum)
# 计算预测概率
prob_mcmc <- exp(result_mcmc)/(1+exp(result_mcmc))
# 预测二分类结果
pre_mcmc <- factor(ifelse(prob_mcmc>0.8,1,0))
ref_mcmc <- factor(testinla$y0)
conM.mcmc <- confusionMatrix(pre_mcmc,ref_mcmc,positive = '1')
conM.mcmc
```

§A.2 SAS 代码

```
libname MCMC "C:\Users\lenovo\Desktop\毕业论文\data";
/* import data */
proc import out = MCMC.data
Datafile = "C:\Users\lenovo\Desktop\毕业论文\data\graduate-admissions
            \MCMC_data5.xls" dbms = excel
replace;
Getnames = YES;
run;

/* View data */
proc print data = MCMC.data;run;quit;
proc contents varnum data = MCMC.data;
    ods select position;
run;quit;

ods graphics on;
/* MCMC */
    /* NBI 是计算参数结果时去掉开头的迭代数 PLOTS是画出所有图像 */
  /* STATISTICS是计算参数所有统计量值 SEED是随机种子 NMC 是迭代次数*/
    /* 模型 logit(p)=beta0 + beta1x1 +...+ beta7x7 */
proc MCMC data = MCMC.data  nbi = 40000 nmc = 100000 seed = 2020
          DIAG = ALL DIC PLOTS = ALL STATISTICS = ALL
          propcov = quanew outpost = MCMC.out;
    parms beta0-beta7;
```

```
    prior beta: ~ normal(0,var=10000);
    mu = beta0 + beta1*GRE + beta2*TOEFL + beta3*UR + beta4*SOP +
              beta5*LOR + beta6*CGPA + beta7*RE;
  p = exp(mu)/(1+exp(mu));
  model COA ~ bern(p);
run;quit;
ods graphics off;
```

致　谢

自 2016 年入学，四年本科时间弹指即过，一路走来，收获颇丰，感叹反思自己并没有虚度光阴。论文选题到完成的过程中始终都得到导师汤银才教授的鼎力相助，解答了我论文撰写和排版中的许多问题，培养我的科研探索意识，使我在接下来的人生中有了新的目标与方向。汤老师一直都强调要培养学生的自主科研能力，善于发现问题，大胆进行设想，广泛查阅资料，增强语言表达。在此由衷的向汤老师表示感激，并祝愿老师身体健康，生活幸福。

其次，我还要感谢本科期间认识的所有同学与老师，是你们让我四年本科生活有滋有味，也给予了我学业、工作、科研比赛和情感生活上的的鼓励和帮助。尤其对我本科期间的室友徐通、夏亦扬和朱笑延，与他们相识我感到很幸运，一生当中难得结友如斯，祝福他们都能学业、事业有成，未来一帆风顺。最后感谢我的家人能一直支持和鼓励我，愿家人们身体健康，心想事成。