第一篇

分 层 模 拟

第 1 章　分层线性模型

1.1　概述

1.1.1　背景介绍

1. *分层数据结构*

数据存在于特定的时间和空间中，其表现形式通常是复杂的分层结构，这是一种非常普遍的现象. 比如，公司在制定决策以便提高劳动生产力方面，显然工人和公司都是分析的对象，对这两个层次的变化都必须进行考量. 其实这样的数据就有着一种分层结构：工人嵌套在公司里. 又比如在研究国家经济的发展与影响生育率的教育时，家庭和国家都是研究的对象，前者嵌套在后者之中，这基本的数据结构也是分层的. 再给出一个例子：关于教育方面的数据，学生被分成班级，班级嵌套在学校里，学校上面有社区，社区上面还有省、国家等.

具有分层结构的数据是一种普遍现象. 随着科学技术的飞速发展，当今世界许多科学研究领域主要面临的挑战是急剧增长的高维多元复杂分层数据，这种类型的数据普遍存在. 这些数据的有效分析无论是在理论研究方面还是在经验研究方面都引起了广泛的关注. 本书着力研究如下 6 大类型的高维空间里的分层结构数据：(1) 空间分层数据 (hierarchical data)；(2) 时间纵向数据 (longitudinal data); (3) 重复测量数据 (repeated measurement data); (4) 广义聚类数据 (generalized clustered data); (5) 名义分类数据 (nominal categorical data); (6) 有序分类数据 (ordinal categorical data). 如何从中挖掘出有用的信息，找出数据掩盖下的事物存在与发展的基本规律，促进统计学学科的发展，推动若干重要的相关领域及某些科学前沿取得突破，这些正是本书研究的目的.

2. *分层结构数据分析的历程*

由于分层数据分析在各学科领域中越来越受到重视，所以相关的研究显得异常活跃，文献中出现了各种各样的称谓. 在社会学研究方面，这些模型指分层线性模型 (Goldstein, 1995; Mason, 等, 1983); 在生物测定学方面，这些术语混合型效应模型和随机效应模型很普遍 (Elston, Grizzle, 1962; Laird, Ware, 1982; Singer, 1998); 在计量经济学文献中，也称为随机系数模型 (Fosenberg, 1973; Longford, 1993); 在统计文献中，称为协方差成分模型 (Dempster, Rubin, Tsutakawa, 1981;

Longford, 1987). 本书之所以采用分层线性模型这个术语，是因为它表达了数据的一个重要结构特点，这种数据使用范围广，常见于增长性研究、机构效应和综合研究. 这个术语是由 Lindley 和 Smith(1972),Smith(1973) 引进的. 在这种前提下，Lindley 和 Smith 详细阐述了复杂的误差结构嵌入数据的普遍框架. 不过后续的研究曾经一度衰落，因为这些模型的使用需要对不平衡的数据进行协方差成分的估计. 除了一些非常简单的问题之外，在 20 世纪 70 年代早期没有一种全面的估计方法行得通. Dempster, Laird 和 Rubin (1977) 研究了 EM 的算法，结果使之有了必要的突破：一种概念上的可行和协方差成分估计的广泛应用方法.Dempster, 等 (1981) 证明了这种分层数据结构方法的合理性.Laird 和 Ware(1982), Strenio, Weisberg 和 Bryk (1983) 把这种方法应用到增长性研究方面，而 Mason, 等 (1987) 则把它应用到多层结构横截面数据方面. 后来，通过对最小二层重新反复地广泛使用和一种 Fisher 得分算法，其他的多种协方差成分估计方法也就应运而生了 (Goldstein, 1986).

贝叶斯方法在这种情况下提供了一种有意义的可选择方法. 标准误差比在 ML 下将更趋向于实际. 而且，通过提供感兴趣的每个参数的后验分布，贝叶斯方法提供了有关研究问题的多种有兴趣的图表与数量证据的总结. 贝叶斯方法的出现并不新鲜. 比较新鲜的是较为方便的计算方法的出现，尤其是在分层数据和模型的背景下. 这些新方法的进展是关于使用蒙特卡罗方法来估计在先前被当作难处理的背景下后验分布的规则系统族. 这种方法包括数据增广 (Tanner, wong，1987) 和 Gibbs 取样 (Gelfand, 等, 1990; Gelfand, Smith, 1990).

3. *分位回归*

关于传统的分层模型所用的统计分析方法主要是均值回归. 该方法有许多不足之处. Koenker 和 Bassett (1978) 提出了分位回归，它可以看做是将经典的最小二乘方法从估计条件均值模型扩展到估计条件分位函数组合的模型. 一个重要特殊的情况就是中位数回归估计量，它是最小化绝对误差的和. 其他的条件分位函数的估计方法是通过最小化绝对误差的非对称的加权和. 简单地讲，均值回归研究的是给定解释变量后响应变量的平均变化趋势，而分位回归则试图全面刻画条件随机变量的各分位点随解释变量的变化情况，另外，它能估计出来的系数向量，即边际效应，对于响应变量的离群观测值来说，是稳健的; 给出在不同分位点上潜在的不同解，这具有很有用的解释意义.

1.1.2 复杂数据界定

众所周知，随着现代科学技术的飞速发展，许多科学研究领域产生了多种多的复杂数据. 当然复杂数据的统计建模涵盖了许多当代统计分支，推动了当代统

计学理论方法的进步与发展，并且其应用层面涉及生物信息学、流行病学和金融风险等，意义十分重大深远.

本书所研究的复杂数据的明显特征之一是：高维、高频、多元、复杂的“时空” 分层等性质. 数据分析方面的主要挑战来自：(1) 在高维空间中直接进行系统搜索变得非常困难；(2) 一般高维函数的精确逼近很棘手；(3) 高维函数积分的实现变得不可能；(4) 对感兴趣的高维多元条件随机变量分布的全面刻画尚无先例可循；(5) 如果没有考虑普遍存在的复杂“时空”分层数据的特征，常常使得传统的统计方法表现不佳，甚至失效，等等.

高维多元复杂数据的统计分析是目前全世界统计学界面临的最大挑战，这无疑是当前统计学中的研究热点问题. 本书以此选题，针对复杂数据相关问题开展研究.

1.1.3　经典模型

1. 线性分层分位回归模型

分层数据可以有很多层. 为了说明方便而且又不失一般性，我们在本节里只考虑具有两层的数据，有的书上又称为两水平数据. 其实，所得到的基本结果是很容易推广到多层数据上去. 假设我们有 $(\boldsymbol{X},\boldsymbol{W},Y)$ 的独立同分布 (i.i.d.) 观测值 $\{(\boldsymbol{X}_1,\boldsymbol{W}_1,Y_1),\cdots,(\boldsymbol{X}_n,\boldsymbol{W}_n,Y_n)\}$, 其中 Y_i $(i=1,2,\cdots,n)$ 是实值响应变量, $\boldsymbol{X}_i$ $(i=1,2,\cdots,n)$ 是已知的 $1\times d$ 维第一层预测值向量，$\boldsymbol{W}_i$ $(i=1,2,\cdots,n)$ 是已知的 $d\times f$ 第二层预测矩阵, 满足第一层模型

$$Y_i=\boldsymbol{X}_i\boldsymbol{\beta}_i+\epsilon_i,\qquad \epsilon_i\sim N(0,\sigma^2),\quad i=1,2,\cdots,n$$

其中 $\boldsymbol{\beta}_i$ $(i=1,2,\cdots,n)$ 是未知的 $d\times 1$ 维第一层系数向量, ϵ_i $(i=1,2,\cdots,n)$ 是 i.i.d. 不可观测随机效应变量, 假定它们与 $\boldsymbol{X}_i$ $(i=1,2,\cdots,n)$ 独立并且服从均值为 0 方差为 σ^2 的正态分布.

在第二层模型上, 第一层模型中的系数成了输出结果:

$$\boldsymbol{\beta}_i=\boldsymbol{W}_i\boldsymbol{\gamma}+\boldsymbol{u}_i,\qquad \boldsymbol{u}_i\sim N(\boldsymbol{0},\boldsymbol{T}),\quad i=1,2,\cdots,n$$

其中 $\boldsymbol{\gamma}$ 是 $f\times 1$ 固定效应向量, $\boldsymbol{u}_i(i=1,2,\cdots,n)$ 是 $d\times 1$ 维第二层随机效应向量，我们假定它们与 $\boldsymbol{W}_i(i=1,2,\cdots,n)$ 和 $\epsilon_i(i=1,2,\cdots,n)$ 独立并且服从均值向量为 $\boldsymbol{0}$ 向量协方差阵为方阵 $\boldsymbol{T}_{d\times d}$ 的多元分布.

将第二层模型代入到第一层模型，产生如下组合模型:

$$Y_i=\boldsymbol{X}_i\boldsymbol{W}_i\boldsymbol{\gamma}+\boldsymbol{X}_i\boldsymbol{u}_i+\epsilon_i,\ \epsilon_i\sim N(0,\sigma^2),\ \boldsymbol{u}_i\sim N(\boldsymbol{0},\boldsymbol{T}),\quad i=1,2,\cdots,n\quad(1.1)$$

为了全面刻画给定预测变量 $(\boldsymbol{X},\boldsymbol{W})=(\boldsymbol{x},\boldsymbol{w})$ 的条件下响应变量 Y 的条件分布 $F(y|\boldsymbol{x},\boldsymbol{w})$, 我们来考虑 Y 的分位函数. 假定 $F(y\,|\,\boldsymbol{x},\boldsymbol{w})$ 是 y 的增函数并且在 $\boldsymbol{x}$ 和 $\boldsymbol{w}$ 处连续, 那么在给定 $\boldsymbol{X}=\boldsymbol{x}$ 和 $\boldsymbol{W}=\boldsymbol{w}$ 的条件下，Y 的 τ 阶分位可定义为 $q_\tau(\boldsymbol{x},\boldsymbol{w})$, 它满足：

$$q_\tau(\boldsymbol{x},\boldsymbol{w})=\inf\{t\in R: F(t\,|\,\boldsymbol{x},\boldsymbol{w})\geqslant\tau\},\ 0<\tau<1.$$

可以直接证明在模型 (1.1) 之下，有

$$q_\tau(\boldsymbol{x},\boldsymbol{w})=\boldsymbol{x}\boldsymbol{w}\boldsymbol{\gamma}+(\boldsymbol{x}\boldsymbol{T}\boldsymbol{x}'+\sigma^2)^{1/2}\Phi^{-1}(\tau),$$

其中 $\Phi(\cdot)$ 是标准正态分布函数.

2. 半参数分层分位回归模型

为了简化陈述，我们只考虑两层的分层模型. 结果很容易推广到三层或者更多层. 假设有 J 个单元，每个单元 i 有 n_j 个元素，我们定义 Y_{ij} 是第 j 个单元的第 i 个个体的观测值. Y_{ij} 是一个实值随机变量，$\boldsymbol{X}_{ij}$ 是第一层 $d_1\times 1$ 的非参数部分的解释向量，$\boldsymbol{Z}_{ij}$ 是 $d_2\times 1$ 参数部分解释向量.

第一层单元内部模型:

$$Y_{ij}=m(\boldsymbol{X}_{ij}^{\mathrm{T}})+\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j+\varepsilon_{ij},\quad j=1,\cdots,J,\quad i=1,\cdots,n_j;$$

其中 $m(\cdot)$ 是一个未知的函数. ε_{ij} 是不可观测的随机误差项，其分布未知，$\boldsymbol{\beta}_j=(\boldsymbol{\beta}_{j1},\boldsymbol{\beta}_{j2},\cdots,\boldsymbol{\beta}_{jd_2})^{\mathrm{T}}$ 是一个 d_2 维的系数向量. 对于 $\boldsymbol{x}$ 点周围的 $\boldsymbol{X}_{ij}$，我们有下面的局部线性展开：

$$m(\boldsymbol{X}_{ij})\approx m(\boldsymbol{x})+(\boldsymbol{X}_{ij}-\boldsymbol{x})^{\mathrm{T}}\bigtriangledown m(\boldsymbol{x}).$$

从而

$$Y_{ij}\approx m(\boldsymbol{x})+(\boldsymbol{X}_{ij}-\boldsymbol{x})^{\mathrm{T}}\bigtriangledown m(\boldsymbol{x})+\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j+\varepsilon_{ij}\stackrel{\mathrm{def}}{=}\tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_j(x)+\varepsilon_{ij},$$

其中 $\tilde{\boldsymbol{X}}_{ij}=(1,(\boldsymbol{X}_{ij}-\boldsymbol{x})^{\mathrm{T}},\boldsymbol{Z}_{ij}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_j(\boldsymbol{x})=(m(\boldsymbol{x}),\bigtriangledown m(\boldsymbol{x})^{\mathrm{T}},\boldsymbol{\beta}_j^{\mathrm{T}})^{\mathrm{T}}$.

第二层单元之间模型:

$$\boldsymbol{\theta}_j(\boldsymbol{x})=\boldsymbol{W}_j\boldsymbol{\gamma}(\boldsymbol{x})+\boldsymbol{U}_j,$$

其中 $\boldsymbol{W}_j$ 是 $(1+d_1+d_2)\times f$ 的第二层的解释变量矩阵，$\boldsymbol{\gamma}(x)$ 是 $f\times 1$ 的固定效应向量，$\boldsymbol{U}_j$ 是 $(1+d_1+d_2)\times 1$ 的第二层的随机误差向量，和 ε_{ij} 相互独立，有均值向量 $E(\boldsymbol{U}_j)=\boldsymbol{0}$，协方差阵 $\mathrm{cov}(\boldsymbol{U}_j)=\boldsymbol{T}$.

合并两层，我们得到

$$Y_{ij} = \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}} \boldsymbol{W}_j \boldsymbol{\gamma}(\boldsymbol{x}) + \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}} \boldsymbol{U}_j + \varepsilon_{ij}. \tag{1.2}$$

令 $\xi_{ij} = \tilde{\boldsymbol{X}}_{ij}^{\mathrm{T}} \boldsymbol{U}_j + \varepsilon_{ij}$，其中 $\xi_{ij} \sim G_{ij}$. 计 $F(y)$ 为 Y 的分布函数. 为了得到给定条件 $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 下响应变量 Y 的分布 $F(y|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$，我们考虑 Y 的分位函数. 假定 $F(y|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 是 y 的增函数，且在 $\boldsymbol{x}$ 和 $\boldsymbol{w}$ 处连续，于是在条件 $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$ 下 Y 的 τ 阶分位数定义为 $q_\tau(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w})$. 可以直接证明在模型 (1.2) 之下，有

$$q_\tau(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) = \inf\{t \in \mathrm{R} : F(t|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}) \geqslant \tau\}, \quad 0 < \tau < 1.$$

根据上式，很容易证明有

$$F^{-1}_{Y|\boldsymbol{x},\boldsymbol{z},\boldsymbol{w}}(\tau) = \tilde{\boldsymbol{X}}^{\mathrm{T}} \boldsymbol{W}_j \boldsymbol{\gamma}(\boldsymbol{x}) + G^{-1}(\tau).$$

1.1.4　主要参考文献

分层结构模型从 1970 年开始被研究，参见 (Goldstein, 1995), (Mason, 等, 1983)，(Elston 和 Grizzle, 1962)，(Laird 和 Ware, 1982)，(Singer, 1998)，(Rosenverg, 1973)，(Longford, 1993)，(Dempster, 等, 1981)，(Longford, 1987)，(Bryk 和 Raudenbush, 1992), (Chen, Tang 和 Tian, 2013)，(Tian, 等, 2008)，(Lindley 和 Smith, 1972)，(Smith, 1973)，(Mason, Wong 和 Entwistle, 1983), (Goldstein, 1995), (Elston, 1962)，(Laird, 1982)，(Longford, 1987)，(Singer, 1998)，(Rosenberg, 1973)，(Longford, 1993)，(Kass 和 Steffey, 1989)，(Dempster, Rubin 和 Tsutakawa, 1981) 以及 (Hobert, 2000). 本节主要参考 (Tian, 2006)，引入分层分位回归模型.

1.2　贝叶斯估计法

1.2.1　引言

本节关注的将仅限于线性模型，$E(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{\theta}$，其中 $\boldsymbol{y}$ 是观测值向量, $\boldsymbol{A}$ 为已知的设计阵，$\boldsymbol{\theta}$ 为未知的参数向量. 在这种情形下，$\boldsymbol{\theta}$ 的估计通常由最小二乘导出. 我们证明过如果能得到参数的先验信息，那么这种导出通常是对的. 并且进一步探索可以找到改进的甚至有时会有很大改进的估计. 在本节中基于 Finetti (1964) 重要的可交换性概念，我们探索一种特殊形式的先验信息.

讨论完全限定在贝叶斯框架内. 最近有关于统计中贝叶斯和非贝叶斯方法长处的诸多讨论：例如 Cornfield (1969) 的文章以及接踵而来的讨论. 没必要也不希望将这类文献加到这篇文章中来，因为我们知道没有理由反对在这里采用

贝叶斯方法. 然而不致力于这种方法的读者应该记得许多抽样理论学派的方法基本上都是不牢靠的；参见 Lindley (1971b) 的综述. 特别地，典型的最小二乘估计令人不满意；或者在这种学派下，在维数超过两维时所得估计是不容许的. 由著名的最小二乘理论（例如参见（Plackett,1960）[59] 知道，通过变换可以将线性模型写为这种形式：当 $i \leqslant p$ 时，$E(z_i) = \xi_i$；当 $i > p$ 时，$E(z_i) = 0$，这里 z_i 为数据的变换，ξ_i 为参数. 加进正态性假设，我们可以采用 Brown (1966) 的结果，他推广了 Stein (1956) 的结论，该结论表明对于更广泛的损失函数而言当 $i \leqslant p$ 时，由 z_i 作为 ξ_i 的估计是不容许的. 在本文第一部分我们评价了贝叶斯估计的可容许性，并且试图向传统理论的信服者证明至少在某些情况下贝叶斯估计是优于最小二乘估计的.

1. 可交换性

我们以一个简单的例子开始. 假设在一般的线性模型中，设计阵为单位矩阵使得 $E(y_i) = \theta_i$，$i = 1, \cdots, n$，并且 $y_1, \cdots, y_n$ 为独立的正态随机变量，方差已知为 σ^2. 如果 y_i 是平均产量为 θ_i 的某块地上第 i 个品种的产量观测值，就会产生这种简单的模型. 考虑 θ_i 的先验信息时，假设它们分布的可交换性通常是合理的，也就是说通过下标的置换它是不变的：特别地，θ_7 的先验信息与 θ_4 或者其他的 θ_i 的先验信息相同. 对于一对、三个一组或更多的也同样成立. 现在获得可交换分布 $p(\boldsymbol{\theta})$ 的一种途径就是假设

$$p(\boldsymbol{\theta}) = \int \prod_{i=1}^{n} p(\theta_i|\mu) \mathrm{d}Q(\mu),$$

其中对于每个 μ，$p(\theta_i|\mu)$ 和 $Q(\mu)$ 描述了任意的概率分布. 换而言之，在给定 μ 下，$p(\boldsymbol{\theta})$ 通过 $Q(\mu)$ 为独立同分布的一个混合. 事实上，Hewitt 和 Savage (1955) 在 Finetti 原创结果的推广下证明了如果对于每个 n 都假设有可交换性，那么混合体将成为产生一个可交换分布的唯一方式.

在本节中，我们研究了具有可交换先验信息和假设可交换性的情形. 例子中暗含了 $E(\theta_i) = \mu$，即对于每个 i 都有一个共同的值. 换而言之参数的线性结构与观测值 $\boldsymbol{y}$ 的线性结构相似. 如果我们加入 θ_i 作为一个随机抽样来自于正态的前提，那么两阶段 $\boldsymbol{y}$ 和 $\boldsymbol{\theta}$ 的平行性将更接近. 在本节中我们研究如下情形，一般线性模型参数本身就其他量 (我们称这些其他量为超参数) 而言具有一般的线性结构. 在这个简单的例子中只有一个超参数 μ.

确实我们将会发现有必要进一步令超参数也具有线性结构，称为三阶段模型并在下一节中分析. 推广到任何阶段是非常直接的.

返回到简单的例子 $E(y_i) = \theta_i$, $E(\theta_i) = \mu$，各自的方差分别为 σ^2 和 τ^2. 一

旦给定了 μ 的先验分布，情况就完全确定了（实际上这就是刚提到的第三阶段）. 设 μ 在整个实数轴上服从均匀分布 —— 这种情形通常代表 μ 的先验信息不明确. Lindley (1971a) 得出了 θ_i 的后验分布，而且发现它的均值为

$$E(\theta_i|y) = \frac{y_i/\sigma^2 + y/\tau^2}{1/\sigma^2 + 1/\tau^2}, \tag{1.3}$$

其中 $y_\cdot = \sum y_i/n$. 在刚引用的参考文献中有详细的分析，所以我们很满意地将以简洁的讨论作为下面部分的一般性理论介绍.

估计式 (1.3)，被称为贝叶斯估计，正是它将成为通常的最小二乘估计的替代量. 我们将他们记为 θ_i^*，保留通常的记号 $\hat{\theta}_i$ 做普通的估计. 注意到 θ_i^* 是 $y_i = \hat{\theta}_i$ 和总体均值 $y_\cdot$ 的加权平均，它们的权与 y_i 和 θ_i 的方差成反比. 因此自然的估计将会被拉向中心值 y，极端值将会做最大移动. 我们将会发现即使是在一般的模型中加权平均的现象也会出现. 当然估计 (1.3) 依赖于 σ^2 和 τ^2，通常它们未知但是容易得到它们的估计值. 如果对于每个 i, 有 y_i 的重复值，那么 σ^2 可以用通常的组内方差来估计. 由于对于 θ_i 可重复 (在可交换性假定下来自正态分布 $N(\mu, \tau^2)$)，τ^2 可以估计. 例如 $\sum(\theta_i^* - \theta_\cdot^*)^2/(n-1)$ 可以作为 τ^2 的一个合理估计，尽管刚引用的参考文献中显示了该估计可以进一步改进.σ^2 和 τ^2 的估计可以替换 (1.3) 中已知值，然后重复以上循环.

让我们岔开贝叶斯观点，尝试着去说服一个传统的统计学家说 (1.3) 式也是他们可以考虑的合理估计，并且确实比最小二乘估计好. 当然，θ_i^* 是 θ_i 的有偏估计，所以它的长处不可以通过它的方差来判定. 作为替代我们使用均方误差 $E(\theta_i^* - \theta_\cdot^*)^2$ 来评判. 这仅是 n 个估计中的一个估计量的优点判定准则，于是我们考察 n 个值上的平均均方误差. 与 $\hat{\theta}_i$ 相比，简单但冗长的计算使得 θ_i^* 的均方误差可以找到并且可以和 $\hat{\theta}_i$ 的相应的量做比较.θ_i^* 的平均（MSE）比 $\hat{\theta}_i$ 的小的条件是

$$\sum(\theta_i - \theta)^2/(n-1) < 2\tau^2 + \sigma^2. \tag{1.4}$$

θ_i^* 的 MSE 依赖于 θ_i，因此该条件也依赖于 θ_i. 因此贝叶斯估计也并不总是优于最小二乘估计. 但考虑得到 (1.4) 式时，由假定在给定 μ 和 τ^2 后，θ_i 是来自 $N(\mu, \tau^2)$ 的随机样本，所以 θ_i 已知时 (1.4) 式的左边是 τ^2 的估计. 因此条件为 τ^2 的估计小于 $2\tau^2 + \sigma^2$. 估计量的分布为多元卡方分布，简单的计算表明按照正态分布 $N(\mu, \tau^2)$，(1.4) 式满足的几率在 n 小至 4 时就很高了，且当 n 增大时几率迅速地趋于 1. 但是正如我们所看到的 τ^2 自身可以估计，所以有了这一点，我们几乎可以肯定 (1.3) 式中 θ_i^* 的 MSE 比 $\hat{\theta}_i$ 的要小. 特别地，期望（基于 θ 分布）往往支持贝叶斯估计.

此论证具有启发意义. 我们的估计与 Stein (1956) 的结果类似，他严密地证明了在平均 MSE 意义下是优于最小二乘估计的.Brown (私下交谈) 指出已知 σ^2, τ^2 时，(1.3) 式是可容许估计量. 本质上这是因为先验分布的不恰当被限定在一维的 μ 中. 我们离题似地详述了以上观点.

如果使用恰当地的先验分布（就是说它在整个空间中的积分为 1）以及有界的功效函数，那么通过在整个参数分布上最大化期望功效函数得到的估计往往是可容许的. 这一点很容易说明，这是因为在所述的两个条件下，所有一般的数学运算都是有效的，例如积分顺序的交换. 如果所述的条件中任何一个遭到破坏就会出现困难. 导致 MSE 的二次损失是无界的，但为了估计 $\boldsymbol{e}$ 可以很方便地替换为

$$1-\exp\{-(\boldsymbol{\theta}-\boldsymbol{e})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{\theta}-\boldsymbol{e})\} \tag{1.5}$$

其中 $\boldsymbol{\Lambda}$ 是半正定的矩阵，特别地为单位矩阵. 使用模糊的先验信息，如均匀分布，就会像 Stein 所证明的那样会造成困难，正是由于这一特证，至少在维数高于两维时，会产生不容许估计. 在下节的一般性理论中，就有界损失函数 (1.5) 式而论，只要先验信息提供得恰当，那么我们的估计将是可容许的. 如果不恰当性被限定在至多两维之内，我们将推测可容许性.

于是返回到不等式 (1.4) 式，我们看到了在传统的框架下有很好的理由去偏爱新估计量. 在 Hoerl 和 Kennard (1970a, b) 的文章中可找到更进一步的论证，他们讨论了估计的一种特殊情况，我们将要在 1.2.5 节中予以说明. 我们并不严肃地对待这些论证，感觉到由于贝叶斯观点被如此多的普通考虑所支持，其中的准则例如 MSE, 扮演很轻的角色甚至不起作用，同时我们还感觉到他们提供的附加的证实也显得并不重要.

在继续一般性的讨论前，有一点必须强调：在例子中我们已经假设了一个可交换的先验分布. 因此只有当假设真正成立时，才建议采用估计 (1.3) 式. 这正是贝叶斯的强大力量所在，贝叶斯观点提供了一种能描述任何推断和结论的正式体系. 从现实问题到数学公式，制定和展示假设是必要的（这适用于任何形式体系，例如欧几里得几何，而不仅仅是贝叶斯统计）. 这里可交换性是一种假设，在使用基于这些假设的估计量之前，必须评估可交换性与实际的相关性. 例如，如果正如以上所建议的，我们的模型描述了在农田试验时观测到的 n 个品种的产量. 如果一个或多个品种是控制组，其余的是实验组，那么可交换性的假设将会不合适. 然而，假设可修正为控制组内的可交换性和实验组内的可交换性. 与两因素的分类分为行和列类似，分别假设行和列间的可交换性也是合理的. 在任何应用中，先验分布的特殊形式必须仔细考虑.

应该指出在为以上形式的 θ_i 指定一个先验分布的同时，我们又将其作为来自 $N(\mu,\tau^2)$ 的一个随机样本，所以我们没有转到模型 II，随机效应的情形，Fisk (1967) 和 Nelder (1968) 已讨论过. 我们对混合效应估计感兴趣. 我们已经研究过真实模型 II 并且得到了 μ 的估计（在后面的一般模型中的 $\boldsymbol{\theta}_2$），但是这一点将会另有说明.

现在我们转而考虑一般性的理论. 对于熟悉矩阵代数的人来说数学部分并不难，主要的结果将作为定理和推论给出. 第二部分的结果均假设方差已知. 关于扩展到未知方差的情形将在后面描述.

2. 一般贝叶斯线性模型

$\boldsymbol{y}\sim N(\boldsymbol{\mu},\boldsymbol{D})$ 表示随机向量 $\boldsymbol{y}$ 服从多元正态分布，均值为 $\boldsymbol{\mu}$，方差 $\boldsymbol{D}$ 为半正定矩阵.

下面我们进入到主要结果. 我们考虑线性模型，形式为 $E(\boldsymbol{y})=\boldsymbol{A}_1\boldsymbol{\theta}_1$，下标表明了这是模型的第一层. 我们推广到一个任意的 $\boldsymbol{y}$ 的方差矩阵 $\boldsymbol{C}_1.\boldsymbol{\theta}_1$ 的先验分布表示为超参数 $\boldsymbol{\theta}_2$ 的线性模型形式，$E(\boldsymbol{\theta}_1)=\boldsymbol{A}_2\boldsymbol{\theta}_2$，协方差矩阵为 $\boldsymbol{C}_2$. 可以写出任意多层的模型：对于我们来说三层就已经足够了，假设最后一层的均值和方差都是已知的. 为了推断特别是估计，我们需要 $\boldsymbol{\theta}_1$ 的后验分布. 这由以下结果给出.

定理 1.2.1 假设给定 $\boldsymbol{\theta}_1$,

$$\boldsymbol{y}\sim N(\boldsymbol{A}_1\boldsymbol{\theta}_1,\boldsymbol{C}_1), \tag{1.6}$$

给定 $\boldsymbol{\theta}_2$,

$$\boldsymbol{\theta}_1\sim N(\boldsymbol{A}_2\boldsymbol{\theta}_2,\boldsymbol{C}_2), \tag{1.7}$$

给定 $\boldsymbol{\theta}_3$,

$$\boldsymbol{\theta}_2\sim N(\boldsymbol{A}_3\boldsymbol{\theta}_3,\boldsymbol{C}_3). \tag{1.8}$$

于是，给定 $\{\boldsymbol{A}_i\}$, $\{\boldsymbol{C}_i\}$, $\boldsymbol{\theta}_3$ 和 $\boldsymbol{y}$ 时，$\boldsymbol{\theta}_1$ 的后验分布服从 $N(\boldsymbol{Dd},\boldsymbol{D})$, 其中

$$\boldsymbol{D}^{-1}=\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{A}_1+\{\boldsymbol{C}_2+\boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2^{\mathrm{T}}\}^{-1}, \tag{1.9}$$

$$\boldsymbol{d}=\boldsymbol{A}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{y}+\{\boldsymbol{C}_2+\boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2^{\mathrm{T}}\}^{-1}\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\theta}_3. \tag{1.10}$$

(这里 $\boldsymbol{\theta}_i$ 为 p_i 维向量，方差矩阵 $\boldsymbol{C}_i$ 为满秩的.)

第一部分中所描述的情形已经由 Lindley (1971a) 详细讨论过，尽管超出了 Lindley (1969) 所描述的一般框架. 感兴趣的读者可以很容易地将这个例子融入到本节内容的讨论中来.