含组结构和层次结构模型的规则化路径估计

马景义 著

中央财经大学统计学院学术文库

中国统计出版社
China Statistics Press

含组结构和层次结构模型的规则化路径估计

马景义 著

中央财经大学统计学院学术文库

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

含组结构和层次结构模型的规则化路径估计/马景义著.
—北京:中国统计出版社,2008.5
(中央财经大学统计学院学术文库)
ISBN 978-7-5037-5372-5

Ⅰ.含…
Ⅱ.马…
Ⅲ.数理统计—数学模型
Ⅳ.O212

中国版本图书馆 CIP 数据核字(2008)第 026648 号

含组结构和层次结构模型的规则化路径估计

作　　者/马景义
责任编辑/吕　军
装帧设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/880×1230mm　1/32
字　　数/100 千字
印　　张/3.375
版　　别/2008 年 12 月第 1 版
版　　次/2008 年 12 月第 1 次印刷
书　　号/ISBN 978-7-5037-5372-5/O·64
定　　价/11.00 元

序

长期以来，统计学在自然科学和社会科学中都有着极其重要而广泛的应用。统计方法作为数量分析的一种有效方法已经成为理、工、农、医、人文、社会、经济、管理、军事、法律等诸多学科领域的基本方法。特别是近几年来，随着社会经济、科学技术的飞速发展，统计方法与具体的实际问题紧密结合，并辅之以高效的计算机技术，更从广度和深度上大大拓展了统计学在理论、方法和应用等方面的内容体系。可以说，在当今信息爆炸的时代，离开数据、离开数据分析、离开统计应用，就意味着远离科学、远离进步、远离财富。

为繁荣统计与数据分析领域的科学研究事业，鼓励广大教师积极开展学术研究，并使研究成果能够尽快地服务于经济和社会的发展，在中央财经大学学术著作出版资助项目的支持下，中央财经大学统计学院组织出版了“中央财经大学统计学院学术文库”（以下简称文库）。希望它能够为从事统计理论、方法和应用的研究人员以及实际工作者们提供及时而有效的帮助，也能够对广大学生读者深入理解统计、数据分析的内涵有所帮助和启迪。

文库的选题主要源于统计理论、方法和应用领域的相关博士论文，一旦有好的选题，即可列入出版计划。此次出版的书目包括下列四部：

《经济周期波动：测度方法与中国经验分析》，吕光明　著

《中国各地区金融发展与固定资产投资实证研究》，赵楠　著

《成分数据多元分析方法研究》，孟洁　著

《含组结构和层次结构模型的规则化路径估计》，马景义　著

文库的选题注重引进当前国内外在统计理论、方法和应用方面的前沿研究内容，注重与计算机技术相结合，着力突出以下特点：

1. 创新性：文库所收录的专著，在经典统计理论的基础上，进行方法或应用方面的创新，并付之于实践应用，用以解决实际问题。

2. 广泛性：文库包括统计方法在经济、金融领域的实证研究、特殊数据类型的统计建模研究、数据挖掘算法研究等多方面的内容，对自然科学和社会科学领域的广大读者具有一定的参考和借鉴价值。

3. 可操作性：文库中涉及到的统计方法、应用案例等，读者都可以按照其中具体介绍的实施步骤进行演练或用于解决自己工作生活中的数据处理问题；同时，充分考虑读者需求，文库中部分涉及并介绍了实现相关模型方法的应用软件或可编程软件。

4. 通俗性：行文按照“统计原理→模型算法→案例应用”的组织形式，努力体现深入浅出的结构安排和文字风格，便于读者的理解，不同读者群可以有所取舍地阅读和学习。

总之，统计学作为数据处理的方法论，具有广泛的应用领域。本文库的初衷正是希望为相关读者奉献一系列具有一定理论高度，且具备一定指导性和实战性的统计方法和应用书籍，可以让众多领域的科研人员、管理人员、高校学生从中获益。本文库的出版感谢中国统计出版社的大力协助和支持，希望我们的共同努力能够筑就统计学未来的辉煌。同时，欢迎广大读者提出批评意见，以利于我们不断提高。

邱　东

2008 年 4 月

目　录

第1章 引言

本书的核心内容是: 在广义线性模型模型的框架下, 给出在有组结构和层次结构约束情况下, 如何选择罚函数, 及模型在选择相应罚函数后规则化估计的算法. 本文的研究思路为: 首先, 讨论有组结构模型的情况; 再次, 把问题更一般化, 讨论同时有组结构和层次结构模型的情况.

本章中将依次给出研究的理论背景和现实意义, 工作概述.

1.1 论文理论背景和现实意义

伴随着计算机技术和信息技术的高速发展, 不同领域内, 人们不得不分析在大小和复杂度与以往有着天壤之别的数据. 于是机器学习技术应用而生. 伴随着机器学习和其它学科的交叉应用, 模型选择成为机器学习理论研究中的关键问题.

传统的模型选择方法, 如逐步回归, 最优子集回归等的关键环节有这样两步: 第一步, 估计出若干模型, 第二步, 用AIC, BIC, F统计量等指标来选择最优模型.

传统方法中会涉及矩阵求逆操作, 而在处理海量数据的统计建模中自变量往往高维而且具有高度共线性, 这导致这传统算法很不稳定, 见Breiman (1996). 另外, 因为自变量高维, 我们常常希望模型保持稀疏性(sparsity) [1], 从而能更好的解释模型, 通常传统方法不

[1] 稀疏性指预测模型有较多的模型系数估计为0.

能满足.

Tibshirani(1996)提出的Lasso, 通过规范化(Regularization)方法估计模型参数, 集模型参数估计和高维变量选择为一体, 而且有助于我们得到稀疏的预测模型. 这种想法一经提出，很快成为模型选择研究的新概念. Hastie, et, al.(2001)用基于统计框架下的机器学习概括了这些新的领域. 这些方法不仅已广泛地运用到各种统计模型中, 特别是围绕Lasso的模型选择研究表现活跃, 取得了显著的进展.

Hastie, et, al.(2001)还展示了模型选择方法的诸多运用领域，如基因、生物医学、大气，地质，水文，地理学、生态学、环境科学、流行病学，地震，遥感，天文及影像处理等诸多领域的研究中，在计算机和工程学中的应用主要有金融、网络、电信等领域.

本书的探讨的内容是模型有组结构和层次结构约束情况下, 如何拓展Lasso, 得到模型的相应规则化路径估计.

1.2 规则化方法(Regularization Methods)的定义

目前, 国内的文献中, "Regularization"这个词常见的有两种翻译. 第一种是"正则化", 第二种是"规范化". 为什么, 本书中要翻译成"规则化"呢? 本节后面的部分会讨论, Regularization Methods 其实就是在损失函数中, 通过某个罚函数来体现我们对预测模型的某种性质的偏好(罚函数"罚"的方面). 也就是, 罚函数体现的是选择模型的规则.

规则化方法的作用原理如下.

在规则化方法中, 首先会定义下面的实值损失函数

$$\Gamma(\mathbf{b};\lambda) = L(\mathbf{b};Y,Z) + \lambda T(\mathbf{b}). \tag{1.1}$$

(1.1)中

- $\mathbf{b} \in \mathcal{B}$是模型参数, Y和Z分别是因变量的观测值和设计矩阵(Design Matrix) [2].

- $L(\mathbf{b}; Y, Z)$是经验损失函数 (Empirical Loss Function) , 它用来衡量模型的拟合好坏. 如果模型拟合程度越高, 那么 $L(\mathbf{b}; Y, Z)$的值越小;

- $T(\mathbf{b})$是罚函数(Penalty Function), 它用来度量模型的复杂程度. 如果模型越简单, 那么$T(\mathbf{b})$的值越小.

- $\lambda(\geq 0)$是调整参数(Tuning Parameter), 或规则化参数 (Regularization Parameter) . 它用来控制对模型的规则化程度的大小.

下一步, 假定我们可以, 根据(1.1)可以得到$\mathbb{R}^+ \to \mathcal{B}$的映射

$$\hat{\mathbf{b}}_\lambda = \arg\min_{\mathbf{b}} \Gamma(\mathbf{b}; \lambda). \tag{1.2}$$

记(1.2)定义的映射的像集为$\mathcal{B}_\lambda(\subseteq \mathcal{B})$. 那么我们就可以根据AIC, BIC, 交叉验证等准则[3], 选择$\mathcal{B}_\lambda$的某个元素作为预测模型的系数估计.

一般来说, 传统统计中, 统计模型的损失函数就是经验损失函数. 然而, 经验损失函数意在控制预测模型对样本拟合程度. 所以, 规则化方法被提出, 罚函数$T(\mathbf{b})$被引入.

为什么我们希望模型越简单呢?

[2]这里只是一般性描述, 所以不限定$\mathbf{b}$, Z和Y的维数.

[3]AIC, BIC, 交叉验证等准则用来度量模型的预测能力.

这里引入奥卡姆剃刀原理(Ockham's Razor)[4]. 这个原理即"切勿浪费较多东西去做用较少的东西同样可以做好的事情". 后来以一种更为广泛的形式为人们所知, 即"如无必要, 勿增实体". 也就是, 如果有两个类似的解决方案, 选择最简单的。

我们再来看(1.1)和(1.2)的等价形式

$$\begin{aligned} \hat{\mathbf{b}} &= \arg\min_{\mathbf{b}} L(\mathbf{b}; Z, Y), \\ & s.t. \quad T(\mathbf{b}) \leq s, s \geq 0. \end{aligned} \tag{1.3}$$

(1.3)中$s.t.$是英文"subject to"的缩写.

也就是说, (1.1)和(1.2)定义的规则化方法是, 首先, 控制模型的复杂程度在一定的范围内, 然后, 在这个范围内, 寻找使模型的拟合程度最好的模型. 如此, 可以得到一个预测模型估计的集合, 这个集合中的任意一个元素对应固定的复杂程度下, 拟合效果最好的模型. 最后, 我们再用评价模型预测能力的原则从这个预测模型的集合中寻找最优模型.

注 1.1. 在上文中, 假定我们可以, 根据(1.1)可以得到$\mathbb{R}^+ \to \mathcal{B}$的映射(1.2). 换而言之, 为了上述假定成立, (1.1)中$\Gamma(\mathbf{b}; \lambda)$必须满足一定的条件.目前, 统计学习(Statistical Learning)领域, $L(\mathbf{b}; Z, Y)$ 和 $T(\mathbf{b})$ 的选择都有一定的规则[5], 可以保证上述假定成立, 这在后面就不再交代了.

1.3 经验损失函数, 组结构约束模型, 层次结构约束模型

假定我们是有因变量因变量y和自变量$x_1, \ldots, x_{K_1}$, 以及因变量

[4]奥卡姆剃刀原理是由14的世纪哲学家、圣方济各会修士奥卡姆郡的威廉(William of Occam, 约1285年至1349年)提出的一个原理.

[5]$L(\mathbf{b}; Z, Y)$和$T(\mathbf{b})$一般都是凸函数.

的n次实现所得到得向量$Y=(y_1,\ldots,y_n)^{\mathrm{T}}$, 和对应的自变量的$n\times K_1$的数据阵$X$. 记$X=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathrm{T}}$, 即$\mathbf{x}_i(i=1,\ldots,n)$是$K_1$维向量, 进一步, 记$\mathbf{x}_i=(x_{1i},\ldots,x_{K_1i})^{\mathrm{T}}$.

下面将给出本书中经验损失函数, 组结构和层次结构模型的定义.

1.3.1 经验损失函数

根据因变量的类型, 我们可以适当的选择经验损失函数, 记经验损失函数为

$$L(\mathbf{b};Y,Z)=-\sum_{i=1}^{n}\log f(y_i|\theta_i), \tag{1.4}$$

如果y是Binary型, 我们可以通过一定的编码规则, 使得 $y\in\{0,1\}$. 进一步, 可以取$f(\cdot)$是Bernoulli分布密度函数.

Z是$n\times(J+1)$的矩阵[6], 是模型的设计矩阵(Design Matrix), 我们可以表示这样表示Z,

$$Z=(\mathbf{z}_1,\ldots,\mathbf{z}_n)^{\mathrm{T}}.$$

那么$\mathbf{z}_i$, $i=1,\ldots,n$是$J+1$ 维向量. 而模型参数θ_i是$\mathbf{z}_i$的线性函数.

我们首先考虑θ_i的“ANOVA”表示

$$\theta_i=\sum_{p=0}^{P}\theta_{ip}\quad i=1,\ldots,n. \tag{1.5}$$

[6] Z更详细的定义将在本节后面部分给出.

其中

$$\theta_{i0} = b_0, \tag{1.6}$$

$$\theta_{i1} = \sum_{k}^{K_1} f_k(x_{ki}), \tag{1.7}$$

$$\theta_{ip} = \sum_{1\le k_1<\cdots<k_p<K} f_{k_1\ldots k_p}(x_{k_1 i},\ldots,x_{k_p i}), p=2,\ldots,P. \tag{1.8}$$

这里, θ_{i0}是常数项; θ_{i1}称作自变量的主效应; $\theta_{ip}(2\le p\le P)$称作自变量的p阶交互效应[7].

首先, 考虑主效应θ_{i1}的形式.

记x_k的定义域为$\mathbb{U}_k(\subseteq \mathbb{R})$, $k=1,\ldots,K_1$. 假定$f_k(x_k)$是定义在$\mathbb{U}_k$上的J_{1k}维线性函数空间$\mathcal{S}_k$的元素, 并假定$\mathcal{S}_k$的一组基为

$$\{B_{k1}(x_k),\ldots,B_{kJ_{1k}}(x_k)\}.$$

于是, $f_k(x_{ki})$, 可以唯一的表示

$$f_k(x_{ki}) = \sum_{j=1}^{J_{1k}} \kappa_{kj} B_{kj}(x_{ki}). \tag{1.9}$$

记

$$b_{1kj} = \kappa_{kj}, \tag{1.10}$$

$$z_{1kij} = B_{kj}(x_{ki}),$$

$$k=1,\ldots,K_1, j=1,\ldots,J_{1k}.$$

也就是说, 如(1.5), θ_{i1}表示为K_1项和后, 第k项为

$$\sum_{j=1}^{J_{1k}} b_{1kj} z_{1kij}, \quad k=1,\ldots,K_1. \tag{1.11}$$

[7]这里$1\le P\le K_1$, 也就是我们允许假定$P=1$. 如果假定$P=1$, 那么模型中只有主效应, 没有交互效应.

其次, 考虑p阶交互效应θ_{ip}, $p=2,\ldots,P$. 为了方便, 记

$$K_p=\binom{K_1}{p},\quad p=2,\ldots,P.$$

由(1.8), 我们可知p阶交互效应θ_{ip}是K_p项求和, 可以把这K_p项求和按字典顺序排列. 把θ_{ip}的K_p项按字典顺序后, 记排列在第$k(1\le k\le K_p)$个位置的那项为

$$f_{k_1\ldots k_p}(x_{k_1i},\ldots,x_{k_pi}).$$

也就是说, 如果我们知道θ_{ip}的K_p项中某项, 按字典顺序排列后, 排在第k项, 那么我们可以确定$k_1,\ldots,k_p$, 也就是可以确定θ_{ip}的K_p项中第k项为

$$f_{k_1\ldots k_p}(x_{k_1i},\ldots,x_{k_pi}).$$

记$J_{pk}=J_{1k_1}\ldots J_{1k_p}$. 我们进一步假定$f_{k_1\ldots k_p}(x_{k_1i},\ldots,x_{k_pi})$ 是定义在

$$\mathbb{U}_{k_1}\times\cdots\times\mathbb{U}_{k_p}$$

上的J_{pk}维线性函数空间

$$\mathcal{S}_{k_1}\oplus\cdots\oplus\mathcal{S}_{k_p}$$

中的元素, 即

$$\begin{aligned}&f_{k_1\ldots k_p}(x_{k_1i},\ldots,x_{k_pi})\\=&\sum_{j_1=1}^{J_{1k_1}}\cdots\sum_{j_p=1}^{J_{1k_p}}\alpha_{j_1\ldots j_p}B_{k_1j_1}(x_{k_1i})\ldots B_{k_pj_p}(x_{k_pi}).\end{aligned}\tag{1.12}$$

然后, 我们把(1.7)中的J_{pk}项按字典顺序排列后, 假定第$j(1\le j\le J_{pk})$项为

$$\alpha_{j_1\ldots j_p}B_{k_1j_1}(x_{k_1i})\ldots B_{k_pj_p}(x_{k_pi}).\tag{1.13}$$

也就是说, 按(1.7)的假定, 如果$f_{k_1\ldots k_p}(x_{k_1 i},\ldots,x_{k_p i})$表示为$J_{pk}$项后, 如果我们知道这$J_{pk}$项和中某项按字典顺序, 排列后所处的位置是$j$, 那么我们可以确定相应的$j_1,\ldots,j_p$, 也即可以确定这项为

$$\alpha_{j_1\ldots j_p}B_{k_1 j_1}(x_{k_1 i})\ldots B_{k_p j_p}(x_{k_p i}). \tag{1.14}$$

为了方便[8],记

$$\begin{aligned} b_{pkj} &= \alpha_{j_1\ldots j_p}, \qquad (1.15)\\ z_{pkij} &= B_{k_1 j_1}(x_{k_1 i})\ldots B_{k_p j_p}(x_{k_p i}),\\ &\quad p=2,\ldots,P, k=1,\ldots,K_p, i=1,\ldots,n, j=1,\ldots,J_{pk};\\ Z_{pkj} &= (z_{pk1j},\ldots,z_{pknj})^{\mathrm{T}},\\ &\quad p=2,\ldots,P, k=1,\ldots,K_p, j=1,\ldots,J_{pk}. \end{aligned}$$

那么$f_k(x_{ki})=\mathbf{z}_{1ki}^{\mathrm{T}}\mathbf{b}_{1k}$. 所以

$$\theta_{i1}=\sum_{k=1}^{K}\mathbf{z}_{1ki}^{\mathrm{T}}\mathbf{b}_{1k}.$$

所以, 如(1.5), θ_{ip}是K_p项和, 这K_p项按字典顺序排列后, 第k项, 也就是(1.12), 可以表示为

$$\sum_{j=1}^{J_{pk}}b_{pkj}z_{pkij} \tag{1.16}$$

到这里, 我们可以统一记号, 将(1.11)和(1.16)统一. 也就是θ_{ip}表示为K_p项, 这K_p项按字典顺序排列后, 第k项, 可以表示为

$$\sum_{j=1}^{J_{pk}}b_{pkj}z_{pkij}, k=1,\ldots,K_p, p=1,\ldots,P. \tag{1.17}$$

[8] z_{pkij}共有4个下标: 第1个下标"p"表示z_{pkij}与p阶效应, 也就是θ_{pi}有关; 第2个下标"k"表示z_{pkij}相应的p阶效应θ_{pi}的第k项; 第3个下标"i"表示z_{pkij}相应的p阶效应θ_{i1}和第i组样本y_1有关; 第4个下标"j"表示z_{pkij}和p阶效应θ_{pi}第k项(表示为J_{1k}项和)的第j项有关系.

然后, 定义J_{pk}维向量$\mathbf{b}_{pk}$和$\mathbf{z}_{pki}$, $\mathbf{b}_{pk}$和$\mathbf{z}_{pki}$具体形式为

$$\begin{aligned}
\mathbf{b}_{pk} &= (b_{pk1},\ldots,b_{pkJ_{pk}})^{\mathrm{T}}, \\
&\quad k=1,\ldots,K_p, p=1,\ldots,P; \\
\mathbf{z}_{pki} &= (z_{pki1},\ldots,z_{pkiJ_{pk}})^{\mathrm{T}}, \\
&\quad i=1,\ldots,n, k=1,\ldots,K_p, p=1,\ldots,P; \\
Z_{pk} &= (\mathbf{z}_{pki},\ldots,\mathbf{z}_{pkn})^{\mathrm{T}}, \\
&\quad k=1,\ldots,K_p, p=1,\ldots,P.
\end{aligned} \tag{1.18}$$

所以(1.17), 也又可以表示为

$$\mathbf{z}_{pki}^{\mathrm{T}}\mathbf{b}_{pk}, \quad i=1,\ldots,n, k=1,\ldots,K_p, p=1,\ldots,P. \tag{1.19}$$

进一步, 记$J_p=\sum_{k=1}^{K_p}J_{pk}$, 定义$J_p$维向量$\mathbf{b}_p$和$\mathbf{z}_{pi}$. $\mathbf{b}_p$和$\mathbf{z}_{pi}$具体形式为

$$\begin{aligned}
\mathbf{b}_p &= (\mathbf{b}_{p1}^{\mathrm{T}},\ldots,\mathbf{b}_{pK_p}^{\mathrm{T}})^{\mathrm{T}}, p=1,\ldots,P; \\
\mathbf{z}_{pi} &= (\mathbf{z}_{p1i},\ldots,\mathbf{z}_{pK_pi}^{\mathrm{T}})^{\mathrm{T}}, i=1,\ldots,n, p=1,\ldots,P; \\
Z_p &= (\mathbf{z}_{pi},\ldots,\mathbf{z}_{pn})^{\mathrm{T}}, p=1,\ldots,P.
\end{aligned} \tag{1.20}$$

所以, (1.7)和(1.8)都可以表示为

$$\theta_{pi}=\mathbf{z}_{pi}^{\mathrm{T}}\mathbf{b}_p, \quad p=1,\ldots,P, i=1,\ldots,n.$$

最后, 记$J=\sum_{p=1}^{P}J_p$, 定义$J+1$维向量$\mathbf{b}$和$\mathbf{z}$; $\mathbf{b}$和$\mathbf{z}$的具体形式为

$$\begin{aligned}
\mathbf{b} &= (b_0,\mathbf{b}_1^{\mathrm{T}},\ldots,\mathbf{b}_P^{\mathrm{T}})^{\mathrm{T}}, \\
\mathbf{z}_i &= (1,\mathbf{z}_{1i}^{\mathrm{T}},\ldots,\mathbf{z}_{Pi}^{\mathrm{T}})^{\mathrm{T}}, i=1,\ldots,n; \\
Z &= (\mathbf{z}_1,\ldots,\mathbf{z}_n)^{\mathrm{T}}
\end{aligned} \tag{1.21}$$

于是可以表示(1.5)为

$$\theta_i = \mathbf{z}_i^{\mathrm{T}}\mathbf{b}, \quad i = 1, \ldots, n \tag{1.22}$$

我们还可以把(1.22)表示成向量的形式. 记$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$, 则(1.22)可以表示为

$$\boldsymbol{\theta} = Z\mathbf{b} \tag{1.23}$$

在下面的例子中会具体说明.

例 1.1. (对数线性模型.)

表 1.1 3×3的列联表:

	$x_2 = 0$	$x_2 = 1$	$x_2 = 3$
$x_1 = 0$	n_{00}	n_{01}	n_{02}
$x_1 = 1$	n_{10}	n_{11}	n_{12}
$x_1 = 2$	n_{20}	n_{21}	n_{22}

将$n_{rc}(r = 0, 1, 2, c = 0, 1, 2)$按字典顺序排列, 并按字典顺序记做$y_1, \ldots, y_9$; 对于$y_i(i = 1, \ldots, 9)$, 相应的列联表的两个变量的值记做$x_{1i}, x_{2i}$.

定义

$$B_{kj}(x_{ki}) = \begin{cases} 1, & x_{ki} = j; \\ 0, & x_{ki} \neq j. \end{cases} \tag{1.24}$$

其中$j = 1, 2$; $k = 1, 2$; $i = 1, \ldots, 9$.

所以, $K = 2$, $J_{11} = 2$, $J_{12} = 2$. 对应的Z_{11}, Z_{12}都是9×2的矩阵. 如果对应的(1.5)中$P = 2$, 那么$K_2 = 1$. 对应的Z_{21}是9×4的矩