现代非参数统计中的窗宽选择及应用

田茂再 著

科学出版社

现代非参数统计中的窗宽选择及应用

田茂再 著

科学出版社
北京

内 容 简 介

非参数回归是统计学的一个重要分支，其极具挑战性的重要环节就是“窗宽选择”．本书全面、系统、严格地阐明了非参半参数建模理论与方法，尽力反映复杂高维多元大数据非参数统计研究成果，利用统计学基础理论和方法来解决当代非参半参数统计窗宽选择的一些国际前沿问题．

本书可作为统计学及其相关领域的大学生、研究生的教材，也可供教师和科技人员参考．

图书在版编目(CIP)数据

现代非参数统计中的窗宽选择及应用/田茂再著．—北京：科学出版社，2019.9

ISBN 978-7-03-060967-0

Ⅰ. ①现… Ⅱ. ①田… Ⅲ. ①非参数统计-研究 Ⅳ. ①O212.7

中国版本图书馆 CIP 数据核字(2019) 第 060898 号

责任编辑：王丽平／责任校对：彭珍珍
责任印制：吴兆东／封面设计：陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京虎彩文化传播有限公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 9 月第 一 版 开本：720 × 1000 1/16
2019 年 9 月第一次印刷 印张：32 1/2
字数：600 000

定价：198.00 元

(如有印装质量问题，我社负责调换)

前　言

随着现代科学技术的迅猛发展, 统计学产生了一个非常重要的分支 —— 非参数统计, 它比较灵活, 不必事先对数据的分布形式做出任何假设, 这使得它真正做到了 "让数据自身说话", 所以该分支在实践中有着广泛的应用, 近年来有关该领域的研究得到了长足的发展, 取得了显著的成果. 目前, 有关大数据非参数统计的研究富有挑战性, 是统计学的前沿研究领域之一. 众所周知, 非参数估计中最重要的一步就是 "窗宽选择", 然而系统地研究这方面问题的专著尚未出现 (就笔者所知), 所以本书视角独特, 别具匠心, 它的出版可以弥补该领域的空白. 本书它利用最前沿的统计理论和方法来解决四大当代非参数统计学中窗宽选择的前沿问题:

(1) 均值回归模型下的 "插入法" 是自动窗宽选择法中常用的方法之一. 插入方法基于给定偏差方差函数的最优窗宽的渐近公式. 这种方法的主要缺陷就是应用潜在的未知函数的高阶导数来计算偏差, 这显然比最初估计函数本身更困难, 再者算法非常复杂并且需要一些附加参数.

(2) 均值回归模型下的 "质量拟合法" 也是自动窗宽选择法中常用的方法之一. 这个方法不用处理偏差的估计. 这组方法应用所谓的拟合质量统计量, 如交叉核实 (CV), 广义交叉核实 (GCV), C_p, Akaike 准则等. 这些统计量可以表达成形式如 $Q = \log\hat{\sigma}^2 + \psi(h)$, 其中 $\hat{\sigma}^2 = 1/N\Sigma_s(z_s - \hat{y}_0(x_s, h))^2$, 并且 $\psi(h)$ 是随着窗宽 h 变小而增加的惩罚函数. 通过最小化 Q 选择窗宽 h, 并且惩罚项 $\psi(h)$ 保证了所选取的窗宽 h 的值不会太小. 然而, 最小化 Q 不会直接得到准确的最优值. 关注这种方法的大多数文章与基于数据的全局窗宽 (不变) 有关, 即选择固定的窗宽.

(3) 分位回归模型下的 "置信区间交集法"(ICI), 该方法不需要估计偏差, 并且与前面通过准确最优化目标提到的拟合质量估计量不同.

(4) 分位回归模型下的 "适应性法", 该方法的优点在于完全的适应性, 即不需要假定模型的先验信息, 不存在高维灾难问题, 而且保留某些需要的特征, 例如跳跃点, 瞬间跳出变化.

(5) 以上方法的实际应用.

本书的部分研究成果曾经荣获 "北京市第十三届哲学社会科学优秀成果奖二等奖" 和 "第十二届北京市统计科研优秀成果奖一等奖".

在本书写作过程中, 自始至终有硕士生、博士生参加翻译、校正等工作: 戴成、范洁瑜、钱政超、张宁、安姝静、陈博钰、范燕、姜春波、马维华、苏宇楠、张圆圆、

陈彦靓、郭洁、荣耀华、李兆媛、司世景、熊巍、胡亚南、田玉柱、王榛、杨亚琦、李二倩、罗静、史普欣、王晓荷、袁梦、晏振、梁晓琳等. 在此对他们表示衷心的感谢!

本书获得以下基金资助: 中国人民大学科学研究基金 (中央高校基本科研业务费专项资金资助) 项目 (No.18XNL012)——“大数据分析的稳健统计理论与应用研究”.

本书的目的就是给读者一些非参数统计中窗宽选择的知识, 侧重于理论与方法的系统性论述.

由于作者水平有限, 疏漏和不妥之处在所难免, 甚望批评指正!

田茂再

2019 年 4 月 22 日

目　　录

第 1 章　均值回归模型下的“插入法”

1.1　基于锥形 Fourier 级数估计的窗宽选择

本节主要考虑了非参数回归中窗宽参数的选择问题, 研究了一个锥形Fourier级数估计, 并讨论了它和核估计的关系. 首先考虑一个基于均方误差意义下的无偏估计方法, 并且在此情况下选择的窗宽是渐近最优的, 同时证明了其他方法与之是渐近等价的. 但是, 几个模拟研究表明在小样本和中等样本容量下, 这些方法却表现出很大的不同.

1.1.1　引言

近年来, 非参数概率密度估计和回归估计在理论研究和实际应用中都已变得相当普遍. 这些方法的应用都要求选择一个好的光滑参数, 并且许多基于数据驱动的选择方法已经被提出. 各种交叉验证方法尤其引人注目. 尽管这方面理论上已经有了很大的发展, 如 Craven 和 Wahba(1979), Chow 等 (1983) 以及 Speckman(1994) 等的工作, 但是仍然有许多问题值得进一步探讨.

考虑模型

$$y_i = f(x_i) + \varepsilon_i,$$

其中 $x_i = i/n, i = 0, 1, \cdots, n$, ε_i 是均值为 0 方差为 σ^2 的独立随机变量. 在上述假定下, f 的一个核估计为

$$f_n(x) = \sum_{i=0}^{n-1} \frac{\lambda}{n} \cdot \omega(\lambda(x_i - x)) \cdot y_i, \tag{1.1.1}$$

这里$\omega(\cdot)$是一个关于 0 对称的核函数, λ 是窗宽的倒数, 它可以基于数据选择. Gasser 和 Müller(1984) 曾指出, 如果 f 不是光滑的周期函数, 则 (1.1.1) 所给估计在边界点附近应该被修正.

本节将假定 f 是光滑周期的, 并将考虑 (1.1.1) 的一个循环格式. 这使得利用 Fourier 分析方法 (注意到 (1.1.1) 是一个卷积形式) 成为可能. 为了简单起见, 将对 (1.1.1) 式进行修正. 总体来说, 这个修正就是通过舍掉 ω 中 “混淆” 的 Fourier 系数, 并且用一个锥形 Fourier 级数估计替换原有核估计后所得到的. 这种修正不仅简化了有关技术证明, 而且使得必要的证明更容易实现.

考虑通过最小化下面的风险函数来选取 λ:

$$R_n(\lambda) = E\left(\frac{1}{n}\sum_{i=0}^{n-1}(f(x_i) - f_n(x_i))^2\right),$$

这里 $R_n(\lambda)$ 是积分均方误差的离散近似.

如果令

$$\boldsymbol{Y} = (y_0, \cdots, y_{n-1})^{\mathrm{T}}, \quad \boldsymbol{f} = (f(x_0), \cdots, f(x_{n-1}))^{\mathrm{T}},$$

$$\boldsymbol{W}(\lambda) = \left[\frac{\lambda}{n} \cdot \omega(\lambda(x_i - x_j))\right],$$

那么 $R_n(\lambda)$ 可以表达为

$$R_n(\lambda) = E\left(\frac{1}{n}\|\boldsymbol{f} - \boldsymbol{W}(\lambda)\boldsymbol{Y}\|^2\right),$$

残差平方和是

$$\mathrm{RSS}_n(\lambda) = \|\boldsymbol{Y} - \boldsymbol{W}(\lambda)\boldsymbol{Y}\|^2,$$

简单计算后得

$$\begin{aligned}\frac{1}{n}E\mathrm{RSS}_n(\lambda) &= R_n(\lambda) + \sigma^2 - \frac{2\sigma^2}{n}\mathrm{trace}(\boldsymbol{W}(\lambda)) \\ &= R_n(\lambda) + \sigma^2 - \frac{2}{n}\sigma^2\lambda\omega(0).\end{aligned} \tag{1.1.2}$$

进而, 如果 σ^2 已知, 可以构造风险函数的一个无偏估计

$$\hat{R}_n(\lambda) = \frac{1}{n}\mathrm{RSS}_n(\lambda) - \sigma^2 + \frac{2}{n}\sigma^2\lambda\omega(0), \tag{1.1.3}$$

然后最小化 $\hat{R}_n(\lambda)$ 来选择 λ. 这种方法被 Mallows(1973) 建议用来做回归变量选择, 也被 Craven 和 Wahba(1979) 建议用来在样条光滑方法中选择光滑化参数. 然而一般情况下, σ^2 是未知的, 我们可以通过逐次差分 $y_{k+1} - y_k$ 或把连接三个点所拟合直线残差的单一估计组合在一起来构造 σ^2 的估计 $\hat{\sigma}^2$, 再最小化下式选择 λ,

$$\tilde{R}_n(\lambda) = \frac{1}{n}\mathrm{RSS}_n(\lambda) - \hat{\sigma}^2 + \frac{2}{n}\hat{\sigma}^2\lambda\omega(0). \tag{1.1.4}$$

交叉验证法可以达到同样的目的. 用 $f_n^{(k)}(x)$ 表示删去第 k 个数据点后回归函数的估计, λ 可以通过最小化下式选择

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{k=0}^{n-1}\left(y_k - f_n^{(k)}(x_k)\right)^2. \tag{1.1.5}$$

Hall (1982) 证明了密度函数的一个交叉验证估计的渐近有效性. Wong (1982) 证明了等距数据交叉验证核估计的相合性, 其所用方法有别于本节中所用方法. Li (1983a) 证明了基于交叉验证的近邻回归估计的相合性. Li (1983b) 的另一篇有趣的文章也涉及了交叉验证、风险估计和非参数回归的 Stein 估计问题. Breiman 和 Freedman (1983) 讨论了回归变量选择的有关问题. Rice (1984a) 研究了非参数回归中的窗宽选择问题, 分析了锥形傅里叶序列估计, 并讨论了这种估计与核估计之间的关系. 然而, 本节给出了窗宽估计 (修正估计的窗宽) 的收敛速度.

1.1.2 无偏风险估计

考虑最小化 $\hat{R}_n(\lambda)$ 的估计 λ_n, 定理 1.1.1 将证明当 $n \to \infty$ 时 $\hat{R}_n(\lambda_n) \to_p 0$. 定理 1.1.2 将证明 $\dfrac{\lambda_n - \lambda_n^*}{\lambda_n^*} \to_p 0$. 定理 1.1.3 给出了 λ_n 的渐近正态分布, 证明了 $\dfrac{\lambda_n - \lambda_n^*}{\lambda_n^*}$ 的标准差与 $\dfrac{1}{\sqrt{\lambda_n^*}}$ 同阶.

假定 ω 是一个对称的概率密度函数且 $\int x\omega(x)dx = 0$. 为了得到有关结果, 进一步假设 ω 是光滑的.

由于 (1.1.1) 式的估计是一个卷积形式, 所以可以方便地借助 Fourier 分析方法来研究它及 $\arg\min \hat{R}_n(\lambda)$ 的性质. 假定模型是循环的, 即 f 是一个周期函数. 这里周期性的假定条件有点苛刻, 但是如果 f 不是周期的, 那么 (1.1.1) 关于 x 的估计在边界附近将会表现得不尽如人意.

首先引入一些记号. 令

$$y_{jn} = f\left(\frac{j}{n}\right) + \varepsilon_{jn}, \quad j = 0, 1, \cdots, n-1,$$

其中 ε_{jn} 是均值为 0 方差为 σ^2 的独立随机变量. 再让

$$\hat{y}_{kn} = \frac{j}{\sqrt{n}} \sum_{j=0}^{n-1} y_{jn} e^{-2\pi ijk/n} \tag{1.1.6}$$

表示序列 $y_{jn}/\sqrt{n}$ 的有限 Fourier 变换 (Cooley et al., 1977). 于是, 有

$$E\hat{y}_{kn} = \left(\frac{1}{\sqrt{n}}\right) \sum_{j=0}^{n-1} f(j/n) e^{-2\pi ijk/n} = \sqrt{n} f_{kn}, \tag{1.1.7}$$

其中

$$f_{kn} = \left(\frac{1}{n}\right) \sum_{j=0}^{n-1} f(j/n) e^{-2\pi ijk/n} \tag{1.1.8}$$

是序列 $(1/n)f(j/n)$ 的第 k 个有限 Fourier 系数. 注意到如果

$$f_k = \int_0^1 f(x)e^{-2\pi ikx}dx,$$

那么就有

$$f_{kn} = f_k + \sum_{s\neq 0} f_{k+sn}. \tag{1.1.9}$$

我们将假定 ω 是非负的且支撑在 $[-1/2, 1/2]$ 上. 为了卷积 (1.1.1) 被更好地定义, 对于每个 λ, 函数 $\lambda\omega(\lambda x)$ 被扩展为周期函数, 则序列 $(\lambda/n)\omega(\lambda j/n)$ 的有限 Fourier 系数可表示为 $\omega_{kn}(\lambda)$. 和 f 类似, 有

$$\omega_{kn}(\lambda) = \omega_k(\lambda) + \sum_{s\neq 0} \omega_{k+sn}(\lambda). \tag{1.1.10}$$

如果 ω 的支撑在 $[-1/2, 1/2]$ 上, 则

$$\begin{aligned}\omega_k(\lambda) &= \lambda\int_{-1/2}^{1/2} \omega(\lambda x)e^{-2\pi ikx}dx \\ &= \int \omega(x)e^{-2\pi ikx/\lambda}dx = \tilde{\omega}(k/\lambda), \quad \lambda > 1,\end{aligned} \tag{1.1.11}$$

其中 $\tilde{\omega}$ 表示 ω 的 Fourier 变换.

下面我们开始讨论修正估计. 因为 (1.1.1) 在 j/n 处的估计是一个卷积, 所以它的有限 Fourier 系数是 $\sqrt{n}\hat{y}_{kn}\omega_{kn}(\lambda)$. 众所周知, 在这种循环情形下, 核光滑化的效应就是使得高阶 Fourier 系数逐步消失. 因此下面我们将研究的修正估计就被定义为 Fourier 系数为 $\sqrt{n}\hat{y}_{kn}\tilde{\omega}(k/\lambda)$ 的三角多项式. 这里 $a_n \leqslant k \leqslant b_n$, 其中

$$a_n = \begin{cases} -n/2, & n \text{ 是偶数}, \\ -(n-1)/2, & n \text{ 是奇数}, \end{cases}$$

$$b_n = \begin{cases} n/2 - 1, & n \text{ 是偶数}, \\ (n-1)/2, & n \text{ 是奇数}. \end{cases}$$

因此估计就是一个锥形 Fourier 级数. 通过删去“混淆”的 Fourier 系数可以减少很多繁杂的证明细节. 尽管我们没有去做, 但是可以推测, 没有修正的估计也可以以类似的方式得到.

利用 Parseval 定理

$$\begin{aligned}\left(\frac{1}{n}\right)\mathrm{RSS}(\lambda) &= \left(\frac{1}{n}\right)\sum_{k=a_n}^{b_n} |\hat{y}_{kn} - \hat{y}_{kn}\tilde{\omega}(k/\lambda)|^2 \\ &= \left(\frac{1}{n}\right)\sum_{k=a_n}^{b_n} |\hat{y}_{kn}|^2|1 - \tilde{\omega}(k/\lambda)|^2\end{aligned} \tag{1.1.12}$$