文献汇报

刘开元

北京工业大学

2023 年 9 月 25 日

Bandwidth Selection in Nonparametric Kernel Testing:

- **期刊：** Journal of the American Statistical Association, JASA. (STATISTICS & PROBABILITY, 15/125, Q1, 3.7)
- **一作：** Jiti Gao, University of Adelaide(莫纳什大学杰出教授). 计量经济学，统计学。

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Publisher name: TAYLOR & FRANCIS INC

Journal Impact Factor™

3.7	5.2
2022	Five Year

JCR Category	Category Rank	Category Quartile
STATISTICS & PROBABILITY *in SCIE edition*	15/125	Q1

高集体教授

在线学术报告｜高集体教授：计量经济学与统计学中的非线性：一段个人的探索旅程

Nonlinearity in Econometrics and Statistics: A Personal Journey

主讲嘉宾：高集体 教授

分享时间：11月28日星期六18:00

- 高集体，澳大利亚莫纳什大学计量经济学与商务统计系杰出教授，Journal of Econometrics等学术期刊副主编。高集体教授于2012年被选为澳大利亚社会科学院院士。
- 高集体教授在计量经济学、金融计量经济学、非参数与半参数计量经济学、面板数据与时间序列计量经济等研究领域均有创造性的贡献，已在计量、金融、统计、经济学的顶级期刊发表百余篇论文，并著有两本专著。高集体教授主持多项澳大利亚国家教授级研究员项目、澳大利亚研究委员会基金项目，谷歌学术统计显示高教授的研究工作被引用次数已超过4000次。

方法提出

参数回归模型 VS 非参数模型

1. 基于参数估计和非参数估计 (NW、GM、局部线性/多项式) 的距离;
2. 关注于条件方差函数的积分，基于参数估计和该方差积分的 NW 估计的距离;
   - Fixed alternatives OR sequences of local alternatives
3. 依赖于带宽参数的选择;
   - CV→Estimation-based(检验效果可能不是最优);
   - Suitable bandwidth→ 实操可能有问题 (Horowitz and Spokoiny, 2001);
4. **本文**提出最优带宽的选择方法。

非参数模型设定检验中的最优带宽选择

- 非参数回归模型：

$$Y_i = m(X_i) + e_i, \qquad i = 1, 2, \cdots, n.$$

- 检验问题：

$$\begin{aligned} &\mathcal{H}_0 : m(x) = m_{\theta_0}(x), \\ &\mathcal{H}_1 : m(x) = m_{\theta_1}(x) + \Delta_n(x), \qquad x \in \mathcal{R}^d, \end{aligned}$$

where $1 \leq d \leq 3$ to avoid curese of dimensionality.

- 检验统计量，Eqs (8)：

$$T_n(h) = \frac{1}{n\sqrt{h^d}\sigma_n} \sum_{i=1}^{n} \sum_{j=1, \neq i}^{n} e_i K\left(\frac{X_i - X_j}{h}\right) e_j.$$

一些符号

- X_i : 平稳时间序列，e_i : 扰动项，期望为 0 方差存在，$m(\cdot)$:$\mathcal{R}^d$ 上的未知函数;
- $\Delta_n(x) \neq 0$: 未知的非参数函数，$K(\cdot)$: 核函数，$\pi(X)$:X 的边际密度函数;
- $\sigma_n^2 = 2\mu_2^2 v_2 \int K^2(u)du$，其中 $\mu_k = E\left[e_1^k\right], v_l = E\left[\pi^l(X_1)\right]$。

易知，$T_n(h)$ 是以下二次形式的主项:

$$Q_n(h) = \frac{1}{n\sqrt{h^d}\sigma_n} \sum_{i=1}^{n}\sum_{j=1}^{n} e_i K\left(\frac{X_i - X_j}{h}\right) e_j \quad (T_n(h) + i = j).$$

$T_n(h)$ 和 $Q_n(h)$ 均可写成以下形式:

$$R_n(h) = \sum_{i=1}^{n}\sum_{j=1}^{n} e_i \phi_n\left(X_i, X_j\right) e_j,$$

其中 $\phi_n(\cdot,\cdot)$ 依赖于 n, h, K.

$T_n(h)$ 的估计版本

$T_n(h)$ 中有很多未知量，给出估计版本：

$$\widehat{T}_n(h) = \frac{\sum_{i=1}^n \sum_{j=1,\neq i}^n \widehat{e}_i K\left(\frac{X_i - X_j}{h}\right) \widehat{e}_j}{n\sqrt{h^d}\widehat{\sigma}_n},$$

其中，

$$\widehat{e}_i = Y_i - m_{\hat{\theta}}(X_i), \quad \widehat{\sigma}_n^2 = 2\widehat{\mu}_2^2\widehat{\nu}_2 \int K^2(u)\,du, \quad \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2, \quad \widehat{\upsilon}_2 = \frac{1}{n}\sum_{i=1}^n \widehat{\pi}^2(X_i),$$

其中 $\widehat{\theta}$ 是 H_0 下 $\sqrt{n}$ 相合估计。$\widehat{\pi}(x) = \frac{1}{n\hat{b}_{\text{cv}}^d} \times \sum_{i=1}^n K\left(\frac{x - X_i}{\hat{b}_{cv}}\right)$ 是核密度估计，$\widehat{b}_{\text{cv}}$ 是 CV 方法选的带宽。

对于给定的 h，满足：$\widehat{T}_n(h) = T_n(h) + o_p\left(\sqrt{h^d}\right)$.

$T_n(h)$ 的 Bootstrap 版本

实际操作时临界值无法计算，给出 Bootstrap 版本：

首先生成 $Y_i^* = m_{\hat{\theta}}(X_i) + \sqrt{\hat{\mu}_2} e_i^*$，其中 $e_i^* \sim N(0,1)$。使用新生成的数据集 $\{(X_i, Y_i^*)\}$ 进行计算：

$$\widehat{T}_n^*(h) = \frac{\sum_{i=1}^n \sum_{j=1.\neq i}^n \widehat{e}_i^* K\left(\frac{X_i - X_j}{h}\right) \widehat{e}_j^*}{n\sqrt{h^d}\widehat{\sigma}_n^*},$$

其中 $\widehat{e}_i^* = Y_i^* - m_{\hat{\theta}^*}(X_i)$，$\widehat{\sigma}_n^{*2} = 2\widehat{\mu}_2^{*2}\widehat{v}_2 \int K^2(u)du$，$\widehat{\mu}_2^* = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^{*2}$。

笔误？

$\widehat{\sigma}_n^{*2} = 2\widehat{\mu}_2^{*2}\widehat{v}_2 \int K^2(u)du \longrightarrow \widehat{\sigma}_n^{*2} = 2\widehat{\mu}_2^{*2}\widehat{v}_2^* \int K^2(u)du$?

主要定理和结论

定理 1

假设 A.1 和 A.2 成立，则在 $\mathcal{H}_0$ 下：

- $$\sup_{x\in R^1} |P^*(\widehat{T}_n^*(h) \le x) - P(\widehat{T}_n(h) \le x)| = O(\sqrt{h^d}),$$

  说明可以用 l_α^* 代替 l_α。

- $$p(\widehat{T}_n(h) > l_a^*) = \alpha + O(\sqrt{h^d}),$$

  说明 Bootstrap 版本得到的size function和 α 接近。

size function 和 power function

定义 size function 和 power function 分别为：

$$\alpha_n(h) = P(\widehat{T}_n(h) > l_\alpha \mid H_0), \quad \beta_n(h) = P(\widehat{T}_n(h) > l_\alpha \mid \mathcal{H}_1).$$

把 l_α 换成 l_α^* 得到 $\alpha_n^*(h)$ 和 $\beta_n(h)$。

定义：矩母函数和累积量母函数

- X 的矩母函数 (m.g.f.) 定义为 $\psi_X(t) = \mathrm{E}\left\{\exp\left(t^\top X\right)\right\}$
- 当 $0 < \psi_X(t) < \infty$ 时，X 的累积量母函数 (c.g.f.) 定义为 $\kappa_X(t) = \log \psi_X(t)$.

定理：One-Term Edgeworth Expansion

设 $X_1, X_2, \cdots \overset{\text{i.i.d}}{\sim} X, X$ 分布绝对连续，均值 μ 、方差 σ^2 和四阶矩 $\mathrm{E}(\mathrm{X}^4) < \infty$. 则 $Z_n = \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}$ 的分布函数可写为:

$$F_{Z_n}(x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + O(n^{-1})$$

对 x 一致成立, 其中 $p_1(\mathrm{x}) = -\frac{1}{6}\kappa_3\left(x^2 - 1\right)$。

Edgeworth 展开可理解为中心极限定理的推广，独立和的高阶展开; 类似于对于非随机函数的 Taylor 展开。

定理 2a

假设 A.1 和 A.2 成立时，size function 的 Edgeworth 展开为：

$$\alpha_0(h) = 1 - \Phi(l_k - s_n) - \kappa_n(1 - (l_\alpha - s_n)^2)\phi(l_\alpha - s_n) + o(\sqrt{h^d}),$$

其中 $\Phi(\cdot)$ 和 $\phi(\cdot)$ 分别为标准正态分布的分布函数和密度函数，$s_n = C_0(m)\sqrt{h^d}$，其中

$$C_0(m) = \frac{\int (\frac{\partial m_{\theta_0}(x)}{\partial \theta})^\tau (E[(\frac{m_{\theta_0}(X_1)}{\partial \theta})(\frac{m_{\theta_0}(X_1)}{\partial \theta})^\tau])^{-1} (\frac{m_{\theta_0}(x)}{\partial \theta})\pi^2(x)\,dx}{\sqrt{2v_2 \int K^2(v)\,dv}}.$$

定理 2b

假设 A.1-A.3 成立时，power function 的 Edgeworth 展开为：

$$\beta_n(h) = 1 - \Phi(l_\alpha - r_n) - \kappa_n(1 - (l_\alpha - r_n)^2)\phi(l_\alpha - r_n) + o(\sqrt{h^d}),$$

其中 $r_n = nC_n^2\sqrt{h^d}$，

$$C_n^2 = \frac{\int \Delta_n^2(x)\pi^2(x)\,dx}{\sigma^2\sqrt{2v_2 \int K^2(v)\,dv}}.$$

符号简化

令 z_n 是标准正态分布的 $1-\alpha$ 分位数，$d_j = (z_a^2 - 1)c_j, j = 1, 2$，其中

$$c_1 = \frac{4K^{(3)}(0)\mu_2^3 v_3}{3\sigma_n^3}, \quad c_2 = \frac{\mu_3^2 K^2(0)}{\sigma_n^3}.$$

定理 3

令 $d_0 = d_1 - C_0(m)$，假设定理 2a 的条件满足，在 $\mathcal{H}_0$ 下

$$l_\alpha \approx z_\alpha + d_0\sqrt{h^d} + d_2\frac{1}{n\sqrt{h^d}} \quad \text{in probability}$$

最优带宽的选择

- **选择目的：** 寻找最优带宽 h_{ew}，使得：

$$h_{\mathrm{ew}} = \arg \max_{h \in H_n(\alpha)} \beta_n(h),$$

其中

$$H_n(\alpha) = \{h : \alpha - c_{\min} < \alpha_n(h) < \alpha + c_{\min}\}.$$

$c_{\min}$ 是一个很小的数。

- **推导思路：** 换元$\longrightarrow x_{\mathrm{ew}} = \sqrt{h_{\mathrm{ew}}^d}$。
- **具体形式：**

$$\widehat{h}_{\mathrm{ew}} = \widehat{a}_1^{-1/(2d)} \widehat{t}_n^{-3/(2d)},$$

文章Eqs (36)部分。

模拟以及实例分析

- **Example 1(模拟):** 线性模型 ($\mathcal{H}_0$) 和非线性模型 ($\mathcal{H}_1$)，扰动项正态分布，和 CV 带宽比较；$\star$
- **Example 2(模拟):** 线性模型 ($\mathcal{H}_0$) 和非线性模型 ($\mathcal{H}_1$)，扰动项不同分布，和现有方法比较；
- **Example 3(实例):** 高频 7 天欧元存款利率，$n = 5505$，线性模型 ($\mathcal{H}_0$) 和非线性模型 ($\mathcal{H}_1$)，和 Plug-in 带宽比较。

Example 1

模拟设置

考虑 2 个预测变量

$$X_{i1} = 0.5X_{i-1,1} + u_i,$$
$$X_{i2} = 0.5X_{i-1,2} + v_i$$

其中 $X_{01} = X_{02} = 0$，u_i 和 v_i 服从标准正态分布。
模型：

$$\mathcal{H}_0 : Y_i = X_{i1} + X_{i2} + e_i,$$
$$\mathcal{H}_1 : Y_i = X_{i1} + X_{i2} + c_n(X_{i1}^2 + X_{i2}^2) + e_i,$$

其中扰动项 $e_i \sim N(0,1)$，$c_n = c_{jn}(j = 1, 2)$，$c_{1n} = n^{-1/2}\sqrt{\log\log(n)}$ 或 $c_{2n} = n^{-7/18}$。

正态核；试验次数 $M = 1000$；Bootstrap 次数 $B = 250$；样本量 n =250、500、750；显著性水平 1%、5%、10%。

样本量	置信水平	H0	H1, j=1	H1, j=2
50	0.01	0.012	0.226	0.348
150		0.012	0.338	0.628
250		0.014	0.364	0.71
750		0.008	0.384	0.888
50	0.05	0.062	0.442	0.59
150		0.054	0.574	0.83
250		0.046	0.632	0.914
750		0.05	0.7	0.978
50	0.1	0.118	0.556	0.732
150		0.11	0.706	0.934
250		0.118	0.78	0.972
750		0.102	0.814	0.996

Table 1. Simulated size and power values at the 1% significance level

Sample size	Null hypothesis is true		Null hypothesis is false			
n	α_{01}	α_{02}	β_{11}	β_{21}	β_{12}	β_{22}
250	.012	.016	.212	.239	.294	.272
500	.018	.014	.270	.303	.318	.334
750	.014	.008	.310	.367	.408	.422

Example 2

模拟设置

考虑 1 个预测变量 $X_i \sim N(0, 25)$ 并且在 5% 和 95% 分位数处截断；考虑不同的扰动项：

1. $e_i \sim N(0, 4)$；
2. 90% 的 $e_i \sim N(0, 1.56)$，10% 的 $e_i \sim N(0, 25)$；
3. 方差为 4 的一型极值分布 (非对称)；

模型：

$$\mathcal{H}_0 : Y_i = 1 + X_i + e_i,$$
$$\mathcal{H}_1 : Y_i = 1 + X_i + \frac{5}{\tau}\phi(\frac{X_i}{\tau}) + e_i,$$

其中 $\tau = 1$ 或 0.25。

$K(x) = \frac{15}{16}(1 - x^2)^2 I(|x| \leq 1)$；**试验次数** $M = 1000$；**Bootstrap 次数** $B_0 = 99$，$B_1 = 250$；**样本量** $n = 250$；**显著性水平 5%**。(符号混淆：M or B)

Table 4. Simulated size and power values at the 5% significance level

Distribution	τ	Probability of rejecting null hypothesis: Andrews test	HM test	HS test	EL test	$\widehat{T}_n(\widehat{h}_{\text{new}})$ test
Null hypothesis is true						
Normal		.057	.060	.066	.053	.049
Mixture		.053	.053	.048	.055	.052
Extreme		.063	.057	.055	.057	.052
Null hypothesis is false						
Normal	1.0	.680	.752	.792	.900	.907
Mixture	1.0	.692	.736	.835	.905	1.000
Extreme	1.0	.600	.760	.820	.924	.935
Normal	.25	.536	.770	.924	.929	.993
Mixture	.25	.592	.704	.922	.986	.999
Extreme	.25	.604	.696	.968	.989	.989

Example 3

- **计量经济学数据：** 欧元存款利率；
- **样本量大：** $n = 5505$；
- 基于现有文献探索使用何种模型；
- 正态核；Bootstrap 次数 $M(B) = 1000$。

结论

- **本文所提出方法：** $\widehat{p}_1 = 0.102$
- **Plug-in 带宽：** $\widehat{p}_2 = 0.072$

均得到了接受原假设的结论，用线性模型建模。

谢谢，欢迎批评指正！