文章编号: 1000- 2243(2006)06- 0794- 04

多元条件密度函数的投影追踪估计方法

叶阿忠[1], Rob J Hyndman[2]

(1. 福州大学管理学院, 福建 福州 350002; 2. 莫纳什大学计量经济和商务统计系, 墨尔本 3800 澳大利亚)

摘要: 由多元条件密度函数 $f_{Y|X}(y|x)(x\in\Re^p)$ 可以知道许多被解释变量 Y 对解释变量 X 的回归关系的信息, 其条件期望就是回归函数, 条件方差就是回归误差项的条件方差. 为了克服高维空间数据稀松性带来的估计上的困难, 提出多元条件密度函数的投影追踪估计方法, 通过最小化 Kullback- Leibler距离, 得到了最优初始条件密度函数和每一步的增量函数和方向向量, 还给出了估计步骤及其终止法则.

关键词: 多元条件密度函数; 投影追踪; Kullback- Leibler距离; 终止法则

中图分类号: O212; F224 文献标识码: A

Projection pursuit estimator for multivariate conditional densities

YE A- zhong[1], Rob J Hyndman[2]

(1. College of Management, Fuzhou University, Fuzhou, Fujian 350002, China; 2. Department of Econometrics and Business Statistics, Monash University, Melbourne 3800, Australia)

Abstract As we know, the conditional density function of a random variable given a dependent random vector shows us a lot of information about their related regression. Its mean is regression function and its variance is the conditional variance of regression error. To overcome the curse of dimensionality, we present the projection pursuit estimator for multivariate conditional densities in this paper. The optimal initial conditional density function, augmenting function and direction vector are obtained by minimizing the Kullback- Leibler distance. We also give the estimation procedure and its termination criteria.

Keywords multivariate conditional densities; projection pursuit; Kullback- Leibler distance; termination criteria

自从 Huber[1]和 Friedman, Stuetzle, Schroeder[2]建立多元密度函数的投影追踪估计以来, 还没有人应用该方法估计多元条件密度函数. 多元条件密度函数 $f_{Y|X}(y|x)(x\in\Re^p)$ 可以告诉我们许多被解释变量 Y 对解释变量 X 的回归关系的信息, 其条件期望就是回归函数, 条件方差就是回归误差项的条件方差. 但目前仅有少量的条件密度函数估计的论文发表, 而且, 这方面的研究几乎都是有关一元条件密度函数的[3- 5]. 因为高维空间数据的稀松性, 对多元条件密度函数进行非参数估计是困难的, 除非数据的观察个数很大, 这也就是著名的维数诅咒问题[6]. 为了克服高维空间数据稀松性带来的估计上的困难, 投影追踪的降维估计方法已经成功地应用于多元密度函数和多元非参数回归的估计[2, 7- 8]. 本文提出多元条件密度函数的投影追踪估计方法, 通过最小化 Kullback- Leibler距离, 得到了最优初始条件密度函数和每一步的增量函数和方向向量, 还给出了估计步骤及其终止法则.

1 投影追踪估计

多元条件密度函数 $f_{Y|X}(y|x)(x\in\Re^p)$ 的投影追踪估计具有如下形式:

收稿日期: 2006- 03- 23

作者简介: 叶阿忠(1963-), 男, 博士, 教授.

基金项目: 国家自然科学基金资助项目(70371025); 教育部人文社会科学研究资助项目(02JA790014)

$$g_M(y|\boldsymbol{x}) = g_0(y|\boldsymbol{x}) \prod_{m=1}^{M} h_m(y|\theta_m \boldsymbol{x}) \tag{1}$$

其中: g_0 是初始条件密度函数; h_m 是第 m 步的待定的函数; θ_m 是第 m 步的待定的方向向量; M 是待定的正整数; $\|\theta_m\| = 1$, $\int_{-\infty}^{+\infty} g_M(y|\boldsymbol{x})\,dy = 1$.

由式(1)得到如下递推的关系:

$$g_m(y|\boldsymbol{x}) = g_{m-1}(y|\boldsymbol{x}) h_m(y|\theta_m \boldsymbol{x}), \quad \int_{-\infty}^{+\infty} g_m(y|\boldsymbol{x})\,dy = 1 \qquad (m = 1, 2, \ldots, M) \tag{2}$$

称 h_m 为增量函数. 假定 $g_{m-1}(y|\boldsymbol{x})$给定, 目标是寻求 $f_{Y|X}(y|\boldsymbol{x})$的最佳估计 $g_m(y|\boldsymbol{x})$. 用 $g_m(y|\boldsymbol{x})$与 $f_{Y|X}$的 Kullback-Leibler距离来度量 $g_m(y|\boldsymbol{x})$的拟合优度. 通过最大化 $g_m(y|\boldsymbol{x})$的拟合优度来选择方向向量 θ_m 和它对应的增量函数 $h_m(y|\theta_m \boldsymbol{x})$.

2 Kullback-Leibler距离

定义 **1** 条件密度函数 $g(y|\boldsymbol{x})$与 $f_{Y|X}(y|\boldsymbol{x})$的 Kullback-Leibler距离为

$$d_{KL}(f_{Y|X}, g) = \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{f_{Y|X}(y|\boldsymbol{x})}{g(y|\boldsymbol{x})}\,dy\,d\boldsymbol{x} \tag{3}$$

引理 **1** Kullback-Leibler距离满足

$$d_{KL}(f_{Y|X}, g) \geqslant 0.5 \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) p\left(\frac{f_{Y|X}(y|\boldsymbol{x})}{g(y|\boldsymbol{x})} - 1\right) dy\,d\boldsymbol{x} \geqslant 0 \tag{4}$$

其中当 $|z| \leqslant 1$时, $p(z) = 0.5z^2$, 否则, $p(z) = |z| - 0.5$

引理 **2**

$$\int |f_{Y|X}(y|\boldsymbol{x}) - g(y|\boldsymbol{x})| f_X(\boldsymbol{x})\,dy\,d\boldsymbol{x} \leqslant [2 d_{KL}(f_{Y|X}, g)]^{0.5} \tag{5}$$

可见, $g_m(y|\boldsymbol{x})$与 $f_{Y|X}$的 Kullback-Leibler距离 $d_{KL}(f_{Y|X}, g_m)$越小, $g_m(y|\boldsymbol{x})$拟合 $f_{Y|X}(y|\boldsymbol{x})$的程度越好.

3 定理及其证明

定理 **1** 假定条件密度函数 $f_{Y|X}(y|\boldsymbol{x})$的条件期望为 $\mu_f(\boldsymbol{x})$, 条件方差为 $\sigma_f^2(\boldsymbol{x})$, 则最小化 $d_{KL}(f_{Y|X}, g_0)$的最优的高斯条件密度函数 $g_0^{\text{opt}}(y|\boldsymbol{x})$的条件期望为 $\mu_f(\boldsymbol{x})$, 条件方差为 $\sigma_f^2(\boldsymbol{x})$.

证明　让 $g_0^{\text{opt}}(y|\boldsymbol{x})$是条件期望为 $\mu_f(\boldsymbol{x})$, 条件方差为 $\sigma_f^2(\boldsymbol{x})$的正态条件密度函数. 设 $g(y|\boldsymbol{x})$是条件期望为 $\mu(\boldsymbol{x})$, 条件方差为 $\sigma^2(\boldsymbol{x})$的条件正态密度函数, 则

$$d_{KL}(f_{Y|X}, g) - d_{KL}(f_{Y|X}, g_0^{\text{opt}}) = \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{g_0^{\text{opt}}(y|\boldsymbol{x})}{g(y|\boldsymbol{x})}\,dy\,d\boldsymbol{x}$$

$$= 0.5 \int \left[\log \frac{\sigma^2(\boldsymbol{x})}{\sigma_f^2(\boldsymbol{x})} + \frac{(y - \mu(\boldsymbol{x}))^2}{\sigma^2(\boldsymbol{x})} - \frac{(y - \mu_f(\boldsymbol{x}))^2}{\sigma_f^2(\boldsymbol{x})}\right] f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x})\,dy\,d\boldsymbol{x}$$

$$= 0.5 \int \left[\log \frac{\sigma^2(\boldsymbol{x})}{\sigma_f^2(\boldsymbol{x})} + \frac{\sigma_f^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} + \frac{(\mu(\boldsymbol{x}) - \mu_f(\boldsymbol{x}))^2}{\sigma^2(\boldsymbol{x})} - 1\right] f_X(\boldsymbol{x})\,d\boldsymbol{x}$$

$$\geqslant 0.5 \int \frac{(\mu(\boldsymbol{x}) - \mu_f(\boldsymbol{x}))^2}{\sigma^2(\boldsymbol{x})} f_X(\boldsymbol{x})\,d\boldsymbol{x} \geqslant 0$$

可见, 仅当 $g = g_0^{\text{opt}}$时, $d_{KL}(f_{Y|X}, g)$达最小.

定理 **2** 假定方向向量 θ_m 给定, 则最小化 $d_{KL}(f_{Y|X}, g_m)$的最优增量函数为:

$$h_m^{\text{opt}}(y|\theta_m \boldsymbol{x}) = \frac{f_{Y|\theta_m X}(y|\theta_m \boldsymbol{x})}{g_{m-1}^{(\theta_m)}(y|\theta_m \boldsymbol{x})} \tag{6}$$

其中 $g_{m-1}^{(\theta_m)}(y|\theta_m\boldsymbol{x})$是 $g_{m-1}(y|\boldsymbol{x})$在 $\theta_m\boldsymbol{x}$上的一维条件密度函数. 此时, Kullback- Leibler距离减少

$$d_{KL}(f_{Y|X}, g_{m-1}) - d_{KL}(f_{Y|X}, g_m) = d_{KL}(f_{Y|\theta_m X}, g_{m-1}^{(\theta_m)}) \tag{7}$$

证明 不失一般性, 不妨设 $\theta_n = (1, 0, \ldots, 0)$, 则 $\theta_n\boldsymbol{x} = x_1$. 最小化

$$d_{KL}(f_{Y|X}, g_m) = \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{f_{Y|X}(y|\boldsymbol{x})}{g_{m-1}(y|\boldsymbol{x}) h_m(y|x_1)} \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{\dfrac{f_{YX}(y, \boldsymbol{x})}{f_X(\boldsymbol{x})}}{\dfrac{g_{m-1}(y, \boldsymbol{x})}{g_{m-1}(\boldsymbol{x})} h_m(y|x_1)} \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{\dfrac{f_{YX}(y, \boldsymbol{x}) f_{YX_1}(y, x_1) f_{X_1}(x_1)}{f_{YX_1}(y, x_1) f_{X_1}(x_1) f_X(\boldsymbol{x})}}{\dfrac{g_{m-1}(y, \boldsymbol{x}) g_{m-1}(y, x_1) g_{m-1}(x_1)}{g_{m-1}(y, x_1) g_{m-1}(x_1) g_{m-1}(\boldsymbol{x})} h_m(y|x_1)} \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log \frac{\dfrac{f_{X|YX_1}(y, \boldsymbol{x}|y, x_1) f_{Y|X_1}(y|x_1)}{f_{X|X_1}(\boldsymbol{x}|x_1)}}{\dfrac{g_{m-1}(y, \boldsymbol{x}|y, x_1) g_{m-1}^{(\theta_m)}(y|x_1)}{g_{m-1}(\boldsymbol{x}|x_1)} h_m(y|x_1|)} \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \left[\log \frac{f_{X|YX_1}(y, \boldsymbol{x}|y, x_1)}{f_{X|X_1}(\boldsymbol{x}|x_1)} - \log \frac{g_{m-1}(y, \boldsymbol{x}|y, x_1)}{g_{m-1}(\boldsymbol{x}|x_1)} \right] \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$+ \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) [\log f_{Y|X_1}(y|x_1) - \log g_{m-1}^{(\theta_m)}(y|x_1) h_m(y|x_1)] \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

等价于最小化

$$\int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) [\log f_{Y|X_1}(y|x_1) - \log g_{m-1}^{(\theta_m)}(y|x_1) h_m(y|x_1)] \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X_1}(y|x_1) f_{X_1}(x_1) \log \frac{f_{Y|X_1}(y|x_1)}{g_{m-1}^{(\theta_1)}(y|x_1) h_m(y|x_1)} \mathrm{d}y\,\mathrm{d}x_1 = d_{KL}(f_{Y|X_1}, g_{m-1}^{(\theta_m)} h_m)$$

也就是最优的增量函数由式(6)给出. 此时, Kullback- Leibler距离减少

$$d_{KL}(f_{Y|X}, g_{m-1}) - d_{KL}(f_{Y|X}, g_m) = \int f_{Y|X}(y|\boldsymbol{x}) f_X(\boldsymbol{x}) \log h_m^{\mathrm{opt}}(y|x_1) \mathrm{d}y\,\mathrm{d}\boldsymbol{x}$$

$$= \int f_{Y|X_1}(y|x_1) f_{X_1}(x_1) \log h_m^{\mathrm{opt}}(y|x_1) \mathrm{d}y\,\mathrm{d}x_1 = \int f_{Y|X_1}(y|x_1) f_{X_1}(x_1) \log \frac{f_{Y|X_1}(y|x_1)}{g_{m-1}^{(\theta_m)}(y|x_1)} \mathrm{d}y\,\mathrm{d}x_1$$

$$= d_{KL}(f_{Y|\theta_m X}, g_{m-1}^{(\theta_m)})$$

定理 **3** 最小化 $d_{KL}(f_{Y|X}, g_{m-1}h_m^{\mathrm{opt}})$的最优方向向量 θ_n^{opt}最大化

$$W(\theta_n) = \int f_{Y|X}(y|x) f_X(x) \log h_m^{\mathrm{opt}}(y|\theta_n\boldsymbol{x}) \mathrm{d}y\,\mathrm{d}\boldsymbol{x} \tag{8}$$

证明 由定理 2容易推得.

4 终止法则

从定理 2知, m 越大, g_m 就越趋近于 $f_{Y|X}$. 显然, M 过大, 将增加计算量. 所以, 在实践中, 必须确定式(1)中的 M. 易见, 如果 g_{m-1}接近于 $f_{Y|X}$, 则 h_m^{opt}接近于 1. 因而, 当 h_m^{opt}接近于 1时, 可以确定 $M = m - 1$, 否则, $M \geqslant m$.

5 估计步骤

给定样本数据 $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$, 估计条件密度函数 $f_{Y|X}(y|\boldsymbol{x})$的过程如下:

1) 由定理 1, 设

$$g_0(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-a-\theta_0\boldsymbol{x})^2}{2\sigma^2}\right\} \tag{9}$$

通过最大化对数似然函数

$$ll(a,\ \theta_0,\ \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - a - \theta_0\boldsymbol{X}_i)^2 \tag{10}$$

获得未知参数 a, θ_0, σ^2 的估计.

2) 假定 $g_{m-1}(y|\boldsymbol{x})$和方向向量 θ_n 已知, 采用核估计方法估计一元条件密度函数 $f_{Y|\theta_m X}(y|\theta_n\boldsymbol{x})$[3, 9]. 采用 Monte Carlo抽样方法估计 $g_{m-1}^{(\theta_m)}(y|\theta_m\boldsymbol{x})$[2, 10], 即对于每个 $\boldsymbol{X}_i$ ($1\leqslant i\leqslant n$), 由密度函数 $g_{m-1}(y|X_i)$产生 Monte Carlo随机数据 Y_{i1}, Y_{i2}, ..., Y_{is} (s是一固定的正整数, 为重抽样的数据个数, 可取 $s=30$), 然后, 利用数据 $\{(Y_{ij},\ \boldsymbol{X}_i)_{j=1}^{s}\}_{j=1}^{n}$得到 $g_{m-1}^{(\theta_m)}(y|\theta_m\boldsymbol{x})$的 Hyndman核估计. 进而得到增量函数 $h_m^{\mathrm{opt}}(y|\theta_n\boldsymbol{x}) = \dfrac{f_{Y|\theta_m X}(y|\theta_n\boldsymbol{x})}{g_{m-1}^{(\theta_m)}(y|\theta_n\boldsymbol{x})}$.

3) 最大化

$$w(\theta_m) = \frac{1}{n}\sum_{i=1}^{n}\log h_m^{\mathrm{opt}}(Y_i|\theta_m\boldsymbol{X}_i) \tag{11}$$

得到方向向量 θ_n 的最优估计 θ_n^{opt}. 如果 $h_m^{\mathrm{opt}}(y|\theta_n\boldsymbol{x})$接近于 1, 则取$M=m-1$, 否则, 让 $g_m(y|\boldsymbol{x}) = g_{m-1}(y|\boldsymbol{x})h_m^{\mathrm{opt}}(y|\theta_m^{\mathrm{opt}}\boldsymbol{x})$, 重复 2)和 3), 如此反复直到确定$M$.

6 结语

采用投影追踪的降维技术, 提出多元条件密度函数的投影追踪估计方法, 通过最小化 Kullback-Leibler距离, 得到了最优初始条件密度函数和每一步的增量函数和方向向量, 还给出了估计步骤及其终止法则, 从而有效地解决了因高维空间数据的稀松性对多元条件密度函数进行非参数估计所带来的维数诅咒问题.

参考文献:

[1] Huber P J. Projection pursuit[J]. The Annals of Statistics, 1985, 13(2): 435-475.

[2] Friedman J H, Stuetzle W, Schroeder A. Projection pursuit density estimation[J]. J Amer Statist Assoc, 1984, 79: 599-608.

[3] Bashtannyk D M, Hyndman R J. Bandwidth selection for kernel conditional density estimation[J]. Computational Statistics and Data Analysis, 2001, 36(3): 279-298.

[4] Fan J, Yao Q, Tong H. Estimation of conditional densities and sensitivity measures in nonlinear dynamical system[J]. Biometrika, 1996, 83(1): 189-206.

[5] Hyndman R J, Yao Q. Nonparametric estimation and symmetry tests for conditional density functions[J]. Journal of Nonparametric Statistics, 2002, 14(3): 259-278.

[6] Wand M P, Jone M C. Kernel smoothing[M]. London: Chapman and Hall, 1995.

[7] Hardle W. Applied nonparametric regression[M]. Cambridge: Cambridge University Press, 1990.

[8] Xia Y, Tong H, Li W K, *et al*. An adaptive estimation of dimension reduction space[J]. JR Statist Soc: B, 2002, 64: 363-410.

[9] Hyndman R J, Bashtannyk D M, Grunwald G K. Estimating and visualizing conditional densities[J]. J Comput Graph Stat, 1996, 5: 315-336.

[10] Zhu M. On the forward and backward algorithms of projection pursuit[J]. The Annals of Statistics, 2004, 32(1): 233-244.