相互近邻域的函数型回归估计

黄收友

(北京航空航天大学 数学与系统科学学院，北京 100083)

摘　要： 令 H 为可分的希尔伯特空间，(X,Y) 是在 $H\times\mathbb{R}$ 上取值的随机对。本文通过对每个 X_i 进行 d 项傅里叶级数展开，从而将无穷维降到有限维，进而证明了 MNN 估计的相容性。其中维数和邻域数都是从观察样本中自动选取的。

关键词： 希尔伯特空间；相互最近邻域；回归函数；相容性

中图分类号： O212.4　　**文献标识码：** A　　**文章编号：** 2095－414X(2014)06－0082－04

1　引言

回归函数估计在经济、医学和模式识别等众多领域有广泛的应用[3]. 除了分类以外[2]，k－近邻域(kNN)方法已被广泛用于回归函数估计，在过去的几十年，这方面已有大量的研究，上世纪五十年代，Fix 和 Hodges [5] 首次提出近邻域的方法，随后，Stone 在文献[7]中用 kNN 的方法研究了一类非参回归函数的估计，并证明了一个结论. 即：kNN 估计是普遍相容的. 本文研究了相互近邻域(MNN)的相容性，关于 MNN 的研究可以追溯到上世纪七十年代[4]，由 Chidananda 等人首次提出 MNN 的概念,已应用于决策分析、数据挖掘和模式识别等方面. 与标准的 kNN 方法相比，MNN 方法会增加一定的计算复杂度，但是它能得到比 kNN 更好的估计效果，文献[1]阐述了 MNN 方法的相关理论性质，并证明了其在欧氏空间的相容性及最优收敛率.本文证明了 MNN 在可分希尔伯特空间中的相容性.

2　相互近邻域的相容性

下面先介绍一些基本定义. 令 H 是可分的希尔伯特空间，(X,Y) 为在 $\mathbb{Z}=H\times\mathbb{R}$ 上取值的随机对. $\{(\mathrm{X}_i,\mathrm{Y}_i)\}_{i=1}^n$ 为 n 个独立的样本点，且与随机对 (X,Y) 同分布. 假设 $\{\phi_j\}_{j\ge1}$ 是 H 上的一组标准正交基，对任意的 X_i，则有 $X_i=\sum_{j=1}^{\infty}X_{ij}\phi_j$，其中 $X_{ij}=\langle X_i,\phi_j\rangle$，$\langle\cdot,\cdot\rangle$ 表示内积.

众所周知，函数型数据研究中经常遇到维数灾难的情形，为了克服这一问题，通常不得不使用一些降维技巧来处理高维甚至无数维的数据. 通过考虑 X 在希尔伯特空间中一组正交基下的前 d 项展开，我们把无穷维降至有限维. 具体说来，就是选择一个合适的维数 d，用 $\sum_{j=1}^{d}X_{ij}\phi_j$ 去逼近样本点 X_i. 另一方面，$H^{(\mathrm{d})}$ 是由 $\{\phi_j\}_{j=1}^{d}$ 线性扩张而成的，对每个向量 $X_i^{(\mathrm{d})}\in H^{(\mathrm{d})}$，则有

$$X_i^{(\mathrm{d})}=\sum_{j=1}^{d}X_{ij}\phi_j\,.$$

在本文中，我们假设 Y 是有界的.回归函数 $f_\rho(x)=\mathrm{E}[\mathrm{Y}\mid\mathrm{X}=x]$,其相应的误差为 $\sigma_\rho^2=\int_{\mathbb{Z}}(\mathrm{y}-f_\rho(x))^2d\rho(x,\mathrm{y})$. 类似的，我们分别定义 $f_{\rho,d}$ 和 $\sigma_{\rho,d}^2$ 为 $H^{(\mathrm{d})}$ 上的回归函数和相应误差. $D_n^{(\mathrm{d})}=\left((\mathrm{X}_{(1)}^{(\mathrm{d})}(x),\mathrm{Y}_{(1)}(x)),\cdots,(\mathrm{X}_{(\mathrm{n})}^{(\mathrm{d})}(x),\mathrm{Y}_{(n)}(x))\right)$ 表示按照 $\left\|\mathrm{X}_{(\mathrm{i})}^{(\mathrm{d})}(x)-x\right\|$ 大小单调递增的顺序重排的数据,定义 $\mathbb{N}_k(x)$ 是 x 在集合 $D_n^{(\mathrm{d})}$ 上的 k 邻域，$\mathrm{N}_k(\mathrm{X}_i^{(\mathrm{d})})$ 表示 X_i 在集合 $(D_n^{(\mathrm{d})}\setminus\{\mathrm{X}_i^{(\mathrm{d})}\})\cup\{x\}$ 上最近 k 邻域. 令 x 的相互最近邻域为

$$\Omega_k(x)=\{\mathrm{X}_i^{(\mathrm{d})}\in\mathbb{N}_k(x):x\in\mathrm{N}_k(\mathrm{X}_i^{(\mathrm{d})})\}.$$

作者简介： 黄收友（1982−），男，博士研究生，研究方向：统计学习理论.

基金项目： 国家自然科学基金（11171014 和 91130009）；国家重点基础研究发展计划项目（973−2010CB731900）.

令 $M=|\Omega_k(x)|$ 表示 x 的相互最近邻域中所含点的个数，显然，M 取值于 0 和 k 之间，利用 MNN 的定义，我们给出回归函数的如下估计

$$f_{n,M,\mathrm{d}}(x)=\frac{1}{M}\sum_{i:X_i\in\Omega_k(x)}Y_{(\mathrm{i})}(x)\cdot$$

接下来，我们将数据分成两部分，其中一部分是包含 l 个样本点的训练集 $\{(\mathrm{X}_i,\mathrm{Y}_i),\mathrm{i}\in\mathrm{P}_l\}$，另一部分是包含 m 个样本点的验证集 $\{(\mathrm{X}_j,\mathrm{Y}_j),\mathrm{j}\in\Theta_m\}$，且 $m+l=n$．我们构建基于训练集的 MNN 估计 $f_{l,M,d}$，如此同时，利用验证集按下列方式选取 $\hat{d}$ 和 $\hat{M}$．

$$(\hat{\mathrm{d}},\hat{\mathrm{M}})\in\arg\min_{d>1}\min_{1\le M\le l}\left\{\frac{1}{m}\sum_{j\in\Theta_m}(\mathrm{Y}_j-f_{l,M,d}(\mathrm{X}_j^{(\mathrm{d})}))^2+\frac{\lambda_d}{\sqrt{m}}\right\},\quad (\frac{\lambda_d}{\sqrt{m}}\text{是惩罚项})$$

根据这种方法得到相应的估计函数和误差分别为

$$\hat{f}_n(x)=f_{l,\hat{M},\hat{d}}(x^{(\hat{\mathrm{d}})}),\ \varepsilon(\hat{f}_n)=\int_{\mathbb{Z}}(\mathrm{y}-\hat{f}_n(x))^2d\rho(x,y)=\int_H(\hat{f}_n(x)-f_\rho(x))^2d\rho_X(x)+\sigma_\rho^2$$

下面给出本文的主要结论:

定理 1　假设 Λ 有界，当 $l\to\infty$ 时，$M/l\to 0$ 且 $\frac{\ln l}{m}\to 0$，则有

$$E\int_H(\hat{f}_n(x)-f_\rho(x))^2\to 0,\qquad n\to\infty.$$

3　证明

在本节中，我们将阐述相关定理和命题，另外，给出了两个相关的引理命题的证明.

命题 1　令 $(\mathrm{Y}-f_{l,M,d}(x))^2\le T(\mathrm{T}\ge 0)$ 几乎处处成立，$\Lambda\le\infty$，对整数 $l>1/\Lambda$ 和 $m=n-l$，则存在一个常数 $C_{T,\Lambda}\ge 0$ 满足

$$E\int_H(\hat{f}_n(x)-f_\rho(x))^2\le \mathrm{C}_{T,\Lambda}\sqrt{\frac{\ln l}{m}}+\inf_{d\ge 1}\left\{(\sigma_{\rho,d}^2-\sigma_\rho^2)+\inf_{1\le M\le l}\left(E\int_{H^{(\mathrm{d})}}(f_{l,M,\mathrm{d}}(x)-f_{\rho,\mathrm{d}}(x))^2\right)+\lambda_d/\sqrt{m}\right\},\tag{1}$$

其中 $\Lambda=\sum_{d=1}^{\infty}e^{-2(\lambda_d/T)^2}$．

注: 不等式（1）右边的第一项可以看作是逼近误差，是因为用有限维空间去逼近样本点空间产生的．由下面引理 1 和引理 2，我们可知上述不等式的第一项和第二项都收敛到零.

命题 1 的证明．令 $L(\mathrm{M},\mathrm{d})=\mathrm{E}[(\mathrm{Y}-f_{l,M,d}(\mathrm{X}^{(\mathrm{d})}))^2\,|\,(\mathrm{X}_i,\mathrm{Y}_i),1\le\mathrm{i}\le\mathrm{n}]$，且 $\hat{L}(\mathrm{M},\mathrm{d})=\frac{1}{m}\sum_{j\in\Theta_m}(\mathrm{Y}_j-f_{l,M,d}(\mathrm{X}_j{}^{(\mathrm{d})}))^2$．我们的目标是最小化 $\hat{L}(\mathrm{M},\mathrm{d})+\lambda_d/\sqrt{m}$．取 $\varepsilon>0$，对所有 $d>1$ 和满足 $1<M<l$ 的任意数 M，根据 $(\hat{M},\hat{d})$ 的定义可得

$$\hat{L}(\hat{M},\hat{d})+\lambda_{\hat{d}}/\sqrt{m}\le\hat{L}(M,d)+\lambda_d/\sqrt{m}.$$

从而可得

$$\mathrm{P}\{L(\hat{M},\hat{d})-\hat{L}(M,d)>\lambda_d/\sqrt{m}+\varepsilon\}\le\mathrm{P}\{L(\hat{M},\hat{d})-\hat{L}(\hat{M},\hat{d})>\lambda_{\hat{d}}/\sqrt{m}+\varepsilon\}.$$

进一步，根据可数个随机事件之并的概率不超过其概率之和，以及 Hoeffding 不等式可得

$$\mathrm{P}\{L(\hat{M},\hat{d})-\hat{L}(M,d)>\lambda_d/\sqrt{m}+\varepsilon\}\le\sum_{d=1}^{\infty}\sum_{M=1}^{l}\mathrm{P}\{L(M,d)-\hat{L}(M,d)>\lambda_d/\sqrt{m}+\varepsilon\}$$

$$=\sum_{d=1}^{\infty}\sum_{M=1}^{l}E\mathrm{P}\{L(M,d)-\hat{L}(M,d)>\lambda_d/\sqrt{m}+\varepsilon\,|\,(\mathrm{X}_i,\mathrm{Y}_i),\mathrm{i}\in\mathrm{P}_l\}$$

$$\le\sum_{d=1}^{\infty}l\exp\{-2[\lambda_d/\sqrt{m}+\varepsilon]^2\times(m/M^2)\}\le l\exp\{-2m\varepsilon^2/T^2\}\sum_{d=1}^{\infty}e^{-2(\lambda_d/T)^2}$$

$$=l\Lambda\exp\{-2m\varepsilon^2/T^2\},(\Lambda=\sum_{d=1}^{\infty}e^{-2(\lambda_d/T)^2})$$

另一方面，对所有的 $1\le M\le l$ 和 $d\ge 1$，则有

$$EL(\hat{M},\hat{d}) \le E\hat{L}(M,d) + \lambda_d \big/ \sqrt{m} + \int_0^\infty \mathrm{P}\{L(\hat{M},\hat{d}) - \hat{L}(M,d) > \lambda_d \big/ \sqrt{m} + \varepsilon\} d\varepsilon ,$$

对所有的 $\mu > 0$，则有

$$EL(\hat{M},\hat{d}) \le E\hat{L}(M,d) + \lambda_d \big/ \sqrt{m} + \mu + l\Lambda \int_\mu^\infty \exp\{-2m\varepsilon^2 / T^2\} d\varepsilon .$$

注意到

$$\int_\mu^\infty \exp\{-2m\varepsilon^2 / T^2\} d\varepsilon \le \frac{1}{2}\int_\mu^\infty (2 + \frac{T^2}{2m\varepsilon}) \exp\{-2m\varepsilon^2 / T^2\} \mathrm{d}\varepsilon$$

$$= -\frac{1}{2}\left[\frac{T^2}{2m\varepsilon}\exp\{-2m\varepsilon^2 / T^2\}\right]_\mu^\infty = \frac{T^2}{4m\mu}\exp\{^{-2m\mu^2}\!/_{T^2}\} .$$

若 $\mu = T\sqrt{\ln(l\Lambda)/2m}$，则有

$$EL(\hat{M},\hat{d}) \le E\hat{L}(M,d) + \lambda_d \big/ \sqrt{m} + T\sqrt{\ln(l\Lambda)/2m} + T\big/\sqrt{8m\ln(l\Lambda)} .$$

然而对每个固定的 M 和 d，都有 $E\hat{L}(M,d) = EL(M,d)$，从而证得不等式(1).

在证明定理 1 之前，我们先给出两个引理.

引理 1 当 $d \to \infty$ 时，$\sigma_{\rho,d}^2 - \sigma_\rho^2 \to 0$.

证明：$\sigma_{\rho,d}^2 - \sigma_\rho^2 = E\left[Y - E[\mathrm{Y} \mid \mathrm{X}^{(\mathrm{d})}]\right]^2 - E\left[Y - E[\mathrm{Y} \mid \mathrm{X}]\right]^2$

$$= E\left[Y - E[\mathrm{Y} \mid \mathrm{X}]\right]^2 + E\left[E[\mathrm{Y} \mid \mathrm{X}] - \mathrm{E}[\mathrm{Y} \mid \mathrm{X}^{(\mathrm{d})}]\right]^2 - E\left[Y - E[\mathrm{Y} \mid \mathrm{X}]\right]^2$$

$$= E\left[E[\mathrm{Y} \mid \mathrm{X}] - E[\mathrm{Y} \mid \mathrm{X}^{(\mathrm{d})}]\right]^2 .$$

由于 $E[\mathrm{Y}^2] < \infty$，$\left(E[\mathrm{Y} \mid \mathrm{X}^{(\mathrm{d})}]\right)_{d \ge 1}$ 是有界鞅，从而证得引理 1.

引理 2 假设 $M \to \infty$ 且当 $l \to \infty$ 时 $M/l \to \infty$，对任意固定的 d，则有

$$E\int_{H^{(\mathrm{d})}} \left(f_{l,M,\mathrm{d}}(x) - f_{\rho,d}(x)\right)^2 \to 0 .$$

具体证明可以参照文献[6]中的定理 6.1.

定理 1 的证明. 取定 $\varepsilon > 0$，根据引理 1，则存在 d_0，对所有的 $d \ge d_0$ 满足 $\sigma_{\rho,d}^2 - \sigma_\rho^2 \le \varepsilon$，另一方面，结合引理 2，则有

$$E\int_{H^{(\mathrm{d}_0)}} \left(f_{l,M,\mathrm{d}_0}(x) - f_{\rho,d_0}(x)\right)^2 \to 0, \quad l \to \infty .$$

利用命题 1，则有

$$E\int_H (\hat{f}_n(x) - f_\rho(x))^2$$

$$\le \inf_{d \ge 1}\left[(\sigma_{\rho,d}^2 - \sigma_\rho^2) + \inf_{1 \le M \le l}\left(\mathrm{E}\int_{H^{(\mathrm{d})}}\left(f_{l,M,\mathrm{d}}(x) - f_{\rho,d}(x)\right)^2\right) + \frac{\lambda_{d_0}}{\sqrt{m}}\right] + C_{T\Lambda}\sqrt{\ln l/m}$$

$$\le (\sigma_{\rho,d_0}^2 - \sigma_\rho^2) + \inf_{1 \le M \le l}\left(\mathrm{E}\int_{H^{(\mathrm{d}_0)}}\left(f_{l,M,\mathrm{d}_0}(x) - f_{\rho,d_0}(x)\right)^2\right) + \frac{\lambda_{d_0}}{\sqrt{m}} + C_{T\Lambda}\sqrt{\ln l/m} \le \varepsilon + o(1), \quad n \to \infty .$$

由于 ε 是任意的，从而得证.

4 结论

尽管 *MNN* 比传统的 *kNN* 计算复杂度大，但它提高了估计效果，减少了离异值点的干扰，对于函数型数据来说，同样可以证得它也是弱相容的.

参考文献：

[1] Arnaud G., Nick H. On the mutual nearest neighbors estimate in regression[J]. Journal of Machine Learning Research，2013,(14)：2361-2376.

[2] Biau G., Buena F., Wegkamp M. Functional classification in Hilbert Spaces[J]. IEEE Transactions on Information Theory，2005，(51)：2163-2172.

[3] Burba F., Ferraty F., Vier P. k-NearestNeighbour method in functional nonparametric regression[J]. Journal of Nonparametric

Statistics，2009，(21)：453–469.

[4] Chidananda K., Krishna G. Agglomerative clustering using the concept of mutual nearest neighbourhood[J]. Pattern Recognition，1978，(10)：105–112.

[5] FIX E., Hodges J. Discriminatory Analysis. Nonparametric Discrimination: Consistency properties[J]. International Statistical Review，1989，57(3)：238–247.

[6] Kohler M., Krzyzak A., Walk H. A distribution–free theory of nonparametric regression[J]. Springer Series in Statistics, 2002, XVI: 650.

[7] Stone, C. Consistent nonparametric regression[J]. The Annals of Statistics，1977，(5)：595–645.

A Mutual Nearest Neighbor Estimate for Functional Regression

HUANG Shou-you

(School of Mathematics and System Science, Beihang University, Beijing 100083, China)

Abstract：Let H be a separable Hilbert space, (X,Y) be a random pair taking values in $H \times \mathbb{R}$. By considering only the first d coefficients of a Fourier series expansion of each X_i, we reduce the dimension of space H, and then prove the consistency of mutual nearest neighbor estimation of the regression function. Both the dimension and the number of neighbors are automatically selected from the observation which uses data dependent splitting devices.

Key words：Hilbert Space; Mutual Nearest Neighbor; Regression Function; Consistency