

基于连接函数的投影寻踪主成分分析

王晓光, 宋立新*

(大连理工大学 数学科学学院, 辽宁 大连 116024)

摘要: 基于连接函数的相依度量, 提出了一种探索连续型随机向量之间相关关系的方法. 通过连接函数构造的投影指标函数可以给出高维随机向量的投影寻踪主成分分析, 并且可以证明这样得到的最优样本投影方向具有强相合性. 此外, 又给出了两个以上随机向量之间的典型相关分析. 结果表明所提出的方法具有优良的理论性和实用前景.

关键词: 投影寻踪; 连接函数; 经验连接函数; 主成分分析; 典型相关分析

中图分类号: O212.4 **文献标志码:** A

0 引言

随着现代科技的发展越来越多的数据产生, 其中高维数据的统计分析至关重要, 许多在低维空间中表现优良的传统统计分析方法在高维空间表现糟糕或根本无法使用, 学者们开始考虑降维. 统计中降维的思想有很多, 作为一种投影降维方法, 投影寻踪通过将高维数据投影到低维(1~3维)空间, 如2~3维的彩色图像以及计算机屏幕上显示的点云等可视图形, 以此来揭示高维数据中的隐含规律和结构^[1~3]. 目前, 投影寻踪方法已经被广泛地应用于多元分析领域, 如多元回归分析、聚类分析、判别分析、主成分分析和典型相关分析等统计方法.

投影寻踪方法的主要步骤是构造一个投影指标函数, 用来选取投影方向. 即在已知 p 维随机向量 $\mathbf{X}$ 的 n 个观察值 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ 时, 通过投影指标函数计算得出单位投影方向阵 $\mathbf{A} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_k)^T, 1 \leq k \leq p, \|\alpha_i\| = 1, i = 1, 2, \dots, k$, 从而将 $\mathbf{X}$ 投影到 k 维空间, $k = 1, 2, 3$. 文献[1~4]分别提出了几种投影指标函数, 这些指标主要用来处理抛物型分布的随机向量, 如正

态分布或者 t -分布, 而且这些指标都是基于线性相关关系提出的.

在研究相关关系时, 当前较多采用的是连接函数(Copula), 它作为联合分布与其一元边际之间的函数, 能够描述非线性相关关系, 因而被广泛应用于经济、金融、保险等领域. 通过加工连接函数, 可以获得多种非线性相关的度量指标. 本文提出的投影指标函数正是基于这类相依度量来构造的, 进而对连续型的高维随机向量进行降维, 如 L_2 度量^[5]:

$$\|C\|_{L_2} = (90 \int_{\mathbb{R}^k} |C_{a_1, \dots, a_k}(u_1, \dots, u_k) - u_1 \cdots u_k|^2 du_1 \cdots du_k)^{1/2}$$

其中 k 元函数 $C_{a_1, \dots, a_k}(u_1, \dots, u_k)$ 是投影后随机变量 $a_1^T \mathbf{X}, \dots, a_k^T \mathbf{X}$ 间的连接函数. 所关心的投影方向矩阵 $\mathbf{A}$ 使 $\|C\|_{L_2}$ 达到最小值. 对于样本形式, 当给定 n 个向量观测值 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ 时, 经验连接函数^[6]被用来作为 $\|C\|_{L_2}$ 中连接函数的样本估计. $\mathbf{A}\mathbf{X}$ 是 $\mathbf{X}$ 的低维线性投影, 但此处 $\mathbf{X}$ 不必具有线性相依结构. 这样通过改变 k 的值就可以给出投影寻踪主成分分析方法. 这种方法也可看作

是通过连接函数来做独立成分分析, Ma 等^[7]提出了参数连接函数的独立成分分析方法, 然而他们的方法需要提前准确了解参数连接函数的具体形式, 这往往是很难办到的, 因此本文选用非参数的经验连接函数.

此外, 两个连续随机向量之间的典型相关分析也可以通过本文以下思路得到. 将 $\mathbf{a}^T \mathbf{X}$ 和 $\mathbf{b}^T \mathbf{Y}$ 之间的连接函数作为 $\|C\|_{L_2}$ 中的函数 C , 然后取极大值点 $(\mathbf{a}, \mathbf{b})$, 那么 $\mathbf{a}^T \mathbf{X}$ 和 $\mathbf{b}^T \mathbf{Y}$ 就是最大相关的, 进一步可以给出第二、三相关方向, 而且在以上两种分析方法中所选用的 L_2 度量可以换作由连接函数构造的肯德尔的 τ 和斯皮尔曼的 ρ 等秩相关系数, 结论依然成立.

1 投影寻踪分析

为了直观地了解投影寻踪, 首先给出一个可用于投影寻踪分析的连续随机向量的相关结构, 体会如何将高维随机向量 $\mathbf{X}$ 通过 $\mathbf{AX}$ 投影到低维空间.

例 1 若 U_1, U_2 为两个独立的随机变量, 希

望给出随机向量 $\mathbf{X} = \begin{pmatrix} U_1 + U_2 \\ 2U_1 + U_2 \\ U_1 + 2U_2 \end{pmatrix}$ 的最佳投影方

向. 假设已知 $\mathbf{X}$ 的结构, 利用上面投影寻踪的方法, 可作变换

$$\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \mathbf{X} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}.$$
 所得最优投

影方向为 $\mathbf{a}_1 = \left(-\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2} \quad 0\right)^T, \mathbf{a}_2 = \left(-\frac{\sqrt{2}}{2} \quad 0 \quad \frac{\sqrt{2}}{2}\right)^T.$

正如例 1, 投影寻踪的主要目标是寻找最佳投影方向, 而这是通过投影指标函数来实现的. 投影指标往往是建立在相依度量的基础上, 而连接函数可以用来构造描述非线性相关的相依度量. 本文选取建立在随机变量 $X_1, X_2, \dots, X_k$ 的连接函数和独立连接函数 $\Pi = u_1 \cdots u_k$ 的 L_q 距离上的相依度量. 对任意的 $q, 1 \leq q \leq +\infty, L_q$ 距离形式

如下^[5]:

$$\|C\|_{L_q} = \left(t_{kq} \int \cdots \int_{I^k} |C(u_1, \dots, u_k) - u_1 \cdots u_k|^q du_1 \cdots du_k\right)^{1/q} \tag{1}$$

那么投影指标函数可取为 $q = 2$ 时的简化形式:

$$h(\mathbf{a}_1, \dots, \mathbf{a}_k) = \int \cdots \int_{I^k} (C_{\mathbf{a}_1, \dots, \mathbf{a}_k}(u_1, \dots, u_k) - u_1 \cdots u_k)^2 du_1 \cdots du_k \tag{2}$$

若记 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ 为 $\mathbf{X}$ 的 n 个观测值, 选用经验连接函数估计式(2)中的连接函数, 那么式(2)的估计形式为

$$h_n(\mathbf{a}_1, \dots, \mathbf{a}_k) = \int \cdots \int_{I^k} \left(\hat{C}_{\mathbf{a}_1, \dots, \mathbf{a}_k}(u_1, \dots, u_k) - u_1 \cdots u_k\right)^2 du_1 \cdots du_k \tag{3}$$

其中 $\hat{C}_{\mathbf{a}_1, \dots, \mathbf{a}_k}(u_1, \dots, u_k)$ 是基于样本 $(\mathbf{a}_1^T \mathbf{X}_1, \dots, \mathbf{a}_k^T \mathbf{X}_1), \dots, (\mathbf{a}_1^T \mathbf{X}_n, \dots, \mathbf{a}_k^T \mathbf{X}_n)$ 的经验连接函数:

$$\begin{aligned} \hat{C}_{\mathbf{a}_1, \dots, \mathbf{a}_k}(u_1, \dots, u_k) = \\ \frac{1}{n} \sum_{i=1}^n I_{\{F_{1n}(\mathbf{a}_1^T \mathbf{X}_i) \leq u_1, \dots, F_{kn}(\mathbf{a}_k^T \mathbf{X}_i) \leq u_k\}}; \\ u_1, \dots, u_k \in [0, 1] \end{aligned} \tag{4}$$

其中 $F_{in}(x) = \frac{1}{n} \sum_{j=1}^n I_{\{\mathbf{a}_i^T \mathbf{X}_j \leq x\}}$ 是 $(\mathbf{a}_i^T \mathbf{X}_1, \dots, \mathbf{a}_i^T \mathbf{X}_n)$ 的经验分布函数, $u_i \in \mathbf{R}, i = 1, 2, \dots, k$. 若 $(\mathbf{a}_{1n}, \dots, \mathbf{a}_{kn})$ 是 $h_n(\mathbf{a}_1, \dots, \mathbf{a}_k)$ 的极小值点, 则 $(\mathbf{a}_{1n} \cdots \mathbf{a}_{kn})$ 就是最佳投影方向.

进一步, 给出投影寻踪主成分分析方法. 投影寻踪方法即寻找使得投影指标函数达极值的单位向量, 若依次给出相互正交的单位向量, 则给出了所有的主成分. 传统的主成分分析可以看作是投影寻踪的一个特例, 即取投影指标函数为样本方差的情况. 不同的投影指标函数对应着不同的投影寻踪方法. 本文的投影指标函数基于连接函数的相依度量, 因而具有刻画非线性相关的特点, 其步骤如下.

Step 1 令

$$h_n(\mathbf{a}_{1n}, \mathbf{a}_{2n}) = \max_{|\mathbf{a}_1|=1, |\mathbf{a}_2|=1} h_n(\mathbf{a}_1, \mathbf{a}_2)$$

从而给出前两个主成分 $\mathbf{a}_{1n}^T \mathbf{X}, \mathbf{a}_{2n}^T \mathbf{X}$.

Step 2 固定 a_{1n}, a_{2n} , 求 a_{3n} 使得

$$h_n(a_{1n}, a_{2n}, a_{3n}) = \max_{|a_3|=1, a_3 \perp a_1, a_3 \perp a_2} h_n(a_1, a_2, a_3)$$

从而给出第三个主成分 $a_{3n}^T X$.

Step 3 如此重复, 直至某个主成分使得

$h_n(a_{1n}, \dots, a_{kn})$ 取值接近于 0 为止. 这样即依次给出了各个主成分.

下面给出样本投影方向的相合性结论. 不失一般性, 令 $k = 2$, 即 $A = (a_1 \ a_2)^T$, 样本方向 $(a_{1n} \ a_{2n})$. 记 $H_{a_1, a_2}(x, y)$ 表示一维变量 $a_1^T X$ 和 $a_2^T X$ 的二维联合分布, 其一维边际分布为 $F_{a_1^T X}(x) = H(x, +\infty)$, $G_{a_2^T X}(y) = H(+\infty, y)$, 对应边际密度为 $f_{a_1^T X}, g_{a_2^T X}$. 对于样本投影方向, $H_{a_{1n}, a_{2n}}$ 为 $a_{1n}^T X$ 和 $a_{2n}^T X$ 的联合分布. 记 x_u^a, x_v^a 分别为 $a_1^T X$ 和 $a_2^T X$ 的 u, v 分位数, 其中 $0 < u, v < 1$. 在给出本节主要结论之前, 首先列出用到的 4 个引理.

引理 1 若当 $n \rightarrow +\infty$ 时, 满足 $a_{1n} \rightarrow a_1, a_{2n} \rightarrow a_2$, 则可得如下两个结论:

- (1) 对 $\forall u, v, C_{a_{1n}, a_{2n}}(u, v) \rightarrow C_{a_1, a_2}(u, v)$;
- (2) 对 $\forall u_1, u_2, v_1, v_2$, 满足 $0 < u_1 < u_2 < 1, 0 < v_1 < v_2 < 1$, 令 $B = [u_1, u_2] \times [v_1, v_2]$,

$$\sup_{(u, v) \in B} |C_{a_{1n}, a_{2n}}(u, v) - C_{a_1, a_2}(u, v)| \rightarrow 0$$

其中 $C_{a_{1n}, a_{2n}}(u, v)$ 为变量 $a_{1n}^T X$ 和 $a_{2n}^T X$ 之间的连接函数, $C_{a_1, a_2}(u, v)$ 为变量 $a_1^T X$ 和 $a_2^T X$ 之间的连接函数.

证明 若 $a_{1n} \rightarrow a_1, a_{2n} \rightarrow a_2$, 则有 $a_{1n}^T X \xrightarrow{d} a_1^T X, a_{2n}^T X \xrightarrow{d} a_2^T X$, 记号“ $\xrightarrow{d}$ ”代表随机变量列的弱收敛. 根据 Sklar 定理, 有

$$C_{a_{1n}, a_{2n}}(u, v) = H_{a_{1n}, a_{2n}}(F_{a_{1n}^T X}^{-1}(u), G_{a_{2n}^T X}^{-1}(v))$$

$$C_{a_1, a_2}(u, v) = H_{a_1, a_2}(F_{a_1^T X}^{-1}(u), G_{a_2^T X}^{-1}(v))$$

证毕.

其实引理 1 的证明可通过 $a_{1n}^T X, a_{2n}^T X$ 的联合分布弱收敛结果得到.

根据引理 1 中结论(1) 可以得出对任意 u, v , 连接函数 $C_{a_1, a_2}(u, v)$ 相对 a_1, a_2 是连续函数. 著名的非参数秩相关系数肯德尔的 τ 和斯皮尔曼的

ρ 均可由连接函数表示:

$$\tau = 4 \iint_{I^2} C(u, v) dC(u, v) - 1$$

$$\rho = 12 \iint_{I^2} (C(u, v) - uv) dudv$$

用 $C_{a_{1n}, a_{2n}}(u, v)$ 估计其中的连接函数, 可得到样本形式的 τ_n, ρ_n , 此时有以下引理.

引理 2 在引理 1 的假设成立时, 当 $n \rightarrow +\infty$, 有 $\tau_n \rightarrow \tau, \rho_n \rightarrow \rho$.

引理 2 的证明方法可参见文献[8]中定理 5.1.9 的证明.

引理 3 取 $0 < u_1 < u_2 < 1$, 对 $\forall u \in B_u = [u_1, u_2]$, 记 $F_{a^T X}$ 的 u 分位数为 x_u^a , 假设

$$(1) \exists \delta_1 > 0, \inf_a f_{a^T X}(x_u^a) > \delta_1 > 0;$$

$$(2) f_{a^T X}(x) \text{ 在 } x_u^a \text{ 处等度连续 (对 } \forall \epsilon > 0, \exists \delta > 0, \text{ 当 } |y - x_u^a| < \delta \text{ 时, } \sup_a |f_{a^T X}(y) - f_{a^T X}(x_u^a)| < \epsilon);$$

则

$$\sup_a |(F_{a^T X}^{-1})_n^{-1}(u) - F_{a^T X}^{-1}(u)| \rightarrow 0, \text{ a. s. ; } n \rightarrow +\infty$$

证明 只给出如下情形的证明过程, 当 $u = \frac{1}{2}$ 时, $\sup_a |m_n^a - m^a| \rightarrow 0, \text{ a. s. }$, 其中 m_n^a, m^a 分别为 $(F_{a^T X})_n, F_{a^T X}$ 的 $u = \frac{1}{2}$ 分位数. 而 $(F_{a^T X})_n$ 是对应 $F_{a^T X}$ 的经验分布.

采取反证法来证明. 假设 $\sup_a |m_n^a - m^a| \not\rightarrow 0, \text{ a. s. }$ 不成立, 则存在集合 $C, P(C) > 0, \forall \omega \in C$, 使得 $\sup_a |m_n^a(\omega) - m^a| \not\rightarrow 0$. 即 $\exists \epsilon_0 > 0$, 对 $\forall N, \exists n \geq N$, 对某一个 a' , $|m_n^{a'}(\omega) - m^{a'}| > \epsilon_0$. 若假设(2) 成立, 取 $\delta_2 = \delta \wedge \epsilon_0$, 于是 $m_n^{a'}(\omega) > \epsilon_0 + m^{a'} > \delta_2 + m^{a'}, m_n^{a'}(\omega) < m^{a'} - \epsilon_0 < m^{a'} - \delta_2$. 若假设(1) 也成立, 同时取 $\epsilon = \frac{1}{3}\delta_1$, 那么只要 $y \in (m^{a'} - \delta, m^{a'} + \delta)$, 则

$$f_{a^T X}(y) > -\epsilon + f_{a^T X}(m^{a'}) > -\frac{1}{3}\delta_1 + \delta_1 = \frac{2}{3}\delta_1 \quad (5)$$

由于只要 $y \in (m^{a'} - \delta_2, m^{a'} + \delta_2), (m^{a'} - \delta_2, m^{a'} + \delta_2) \subset (m^{a'} - \delta, m^{a'} + \delta)$, 故式(5)也成立. 于是

$$|F_{a^T X}(m_n^{a'}(\omega)) - F_{a^T X}(m^{a'})| =$$

$$\begin{cases} \int_{m^{a'}}^{m_n^{a'}(\omega)} f_{a^T X}(t) dt \geq \int_{m^{a'}}^{m^{a'} + \delta_2} f_{a^T X}(t) dt > \frac{2}{3} \delta_1 \delta_2 > 0; \\ m_n^{a'}(\omega) \geq m^{a'} \\ \int_{m_n^{a'}(\omega)}^{m^{a'}} f_{a^T X}(t) dt \geq \int_{m^{a'} - \delta_2}^{m^{a'}} f_{a^T X}(t) dt > \frac{2}{3} \delta_1 \delta_2 > 0; \\ m_n^{a'}(\omega) < m^{a'} \end{cases}$$

但是

$$\sup_a |F_{a^T X}(m_n^a) - F_{a^T X}(m^a)| \leq S_n + T_n$$

其中 $S_n = \sup_a |F_{a^T X}(m_n^a) - (F_{a^T X})_n(m_n^a)|$, $T_n =$

$$\sup_a |(F_{a^T X})_n(m_n^a) - F_{a^T X}(m^a)|.$$

由于

$$\begin{aligned} \sup_a |(F_{a^T X})_n(x) - F_{a^T X}(x)| &= \\ \sup_a |P_n I_{\{a^T X \leq x\}} - P I_{\{a^T X \leq x\}}| &= \\ \sup_a |P_n I_{\{a^T X - x \leq 0\}} - P I_{\{a^T X - x \leq 0\}}| &= \\ \sup_{g \in \mathcal{G}} |P_n I_{\{g \leq 0\}} - P I_{\{g \leq 0\}}| &\rightarrow 0, \text{ a. s. } \end{aligned}$$

其中 g 为一个有限维映射, 根据文献[9]中引理18和定理21, 收敛成立. 故 $S_n \rightarrow 0$, a. s.. 由于

$$F_{a^T X}(m^a) = \frac{1}{2}, \text{ 且 } \frac{[n/2]}{n} \leq (F_{a^T X})_n(m_n^a) \leq$$

$$\frac{[n/2] + 1}{n}, \text{ 有 } T_n \leq \frac{1}{n} \rightarrow 0. \text{ 于是, } \sup_a |F_{a^T X}(m_n^a) -$$

$F_{a^T X}(m^a)| \rightarrow 0$, a. s.. 这与开始的假设矛盾, 因此

$$\sup_a |m_n^a - m^a| \rightarrow 0, \text{ a. s. .}$$

引理4 除了引理3的假设(1)、(2), 若更有假设:

(3) $F_{a^T X}$ 等度连续 (对 $\forall \epsilon > 0, \exists \delta' > 0$, 当 $|x_1 - x_2| < \delta', \sup_a |F_{a^T X}(x_1) - F_{a^T X}(x_2)| < \epsilon$);

则

$$\begin{aligned} \sup_{u \in B_u} \sup_a |(F_{a^T X})_n^{-1}(u) - F_{a^T X}^{-1}(u)| &\rightarrow 0, \\ \text{a. s. ; } n \rightarrow +\infty \end{aligned}$$

证明 在假设(3)条件之下, 由引理1的证明方法, 令 $K_1^a = F_{a^T X}^{-1}(u_1), K_2^a = F_{a^T X}^{-1}(u_2)$, 用 $\delta_3 = \delta' \wedge \epsilon/2$ 来分割区间 $(K_1^a - \delta_3, K_2^a + \delta_3)$. 分点为

$$M_1^a = K_1^a - \delta_3, M_2^a = K_1^a, \dots, M_{[(K_2^a - K_1^a)/\delta_3] + 3}^a = K_2^a + \delta_3. \text{ 当 } M_i^a \leq F_{a^T X}^{-1}(u) \leq M_{i+1}^a \text{ 时, 令 } t = F_{a^T X}^{-1}(u). \text{ 由引理3, 对 } \forall \epsilon > 0, \exists N, \text{ 使得}$$

$$\begin{aligned} (F_{a^T X})_n^{-1}(u) - F_{a^T X}^{-1}(u) &\leq \\ (F_{a^T X})_n^{-1} F_{a^T X} M_{i+1}^a - t &= \\ (F_{a^T X})_n^{-1} F_{a^T X} M_{i+1}^a - M_{i+1}^a + M_{i+1}^a - t &\leq \\ (F_{a^T X})_n^{-1} F_{a^T X} M_{i+1}^a - \\ (F_{a^T X})^{-1} F_{a^T X} M_{i+1}^a + M_{i+1}^a - t &\leq \\ \sup_a |(F_{a^T X})_n^{-1} F_{a^T X} M_{i+1}^a - \\ (F_{a^T X})^{-1} F_{a^T X} M_{i+1}^a| + \delta_3 &< \\ \epsilon/2 + \delta_3 \leq \epsilon; n > N \end{aligned}$$

另一方面, $\geq$ 同时成立. 于是 $|(F_{a^T X})_n^{-1}(u) - F_{a^T X}^{-1}(u)| < \epsilon$. 进一步,

$$\begin{aligned} \sup_a \sup_{u \in B_u} |(F_{a^T X})_n^{-1}(u) - F_{a^T X}^{-1}(u)| &= \\ \sup_a \sup_{K_1^a \leq F_{a^T X}^{-1}(u) \leq K_2^a} |(F_{a^T X})_n^{-1}(u) - F_{a^T X}^{-1}(u)| &= \\ \sup_a \left[\max_{1 \leq i \leq \left[\frac{K_2^a - K_1^a}{\delta_3} \right] + 3} \sup_{M_i^a \leq F_{a^T X}^{-1}(u) \leq M_{i+1}^a} |(F_{a^T X})_n^{-1}(u) - \right. \\ \left. F_{a^T X}^{-1}(u) \right] &\leq \epsilon \end{aligned}$$

在以上基础之上, 最优投影方向 (a_{1n}, a_{2n}) 的强相合性可以给出.

定理1 假设作为 $h(a_1, a_2)$ 的极小值点 (a_{10}, a_{20}) 具有唯一性, 且假设:

(1) 对任意 u, v , 存在 $\delta_1 > 0, \delta_2 > 0$, 使得 $\inf_{a_1} f_{a_1^T X}(x_{u_1}^{a_1}) > \delta_1 > 0$ 且 $\inf_{a_2} g_{a_2^T X}(x_{v_2}^{a_2}) > \delta_2 > 0$;

(2) 对任意 $u, v, f_{a_1^T X}(x)$ 在点 $x_{u_1}^{a_1}$ 处等度连续, $g_{a_2^T X}(x)$ 在点 $x_{v_2}^{a_2}$ 处等度连续;

(3) $F_{a_1^T X}$ 和 $G_{a_2^T X}$ 等度连续;

则

$$(a_{1n}, a_{2n}) \rightarrow (a_{10}, a_{20}), \text{ a. s. ; } n \rightarrow +\infty$$

证明 利用文献[9]第2章中问题4的证明方法, 只需验证 $\sup_{(a_1, a_2)} |h_n(a_1, a_2) - h(a_1, a_2)| \rightarrow 0$, a. s.. 若令 $\tilde{C}_{a_1, a_2}(u, v) = H((F_{a_1^T X})_n^{-1}(u), (G_{a_2^T X})_n^{-1}(v))$, 首先考虑

$$\sup_{(u, v) \in B} \sup_{(a_1, a_2)} |\tilde{C}_{a_1, a_2}(u, v) - C_{a_1, a_2}(u, v)| \leq$$

$$S_n + T_n$$

其中 $S_n = \sup_{(u,v) \in B} \sup_{(a_1, a_2)} | \hat{C}_{a_1, a_2}(u, v) - \tilde{C}_{a_1, a_2}(u, v) |, T_n = \sup_{(u,v) \in B} \sup_{(a_1, a_2)} | \tilde{C}_{a_1, a_2}(u, v) - C_{a_1, a_2}(u, v) |$.

对于第一项 S_n , 根据定义

$$\hat{C}_{a_1, a_2}(u, v) = \frac{1}{n} \sum_{i=1}^n I_{\{(F_{a_1}^T X)_n (a_1^T X_i) \leq u, (G_{a_2}^T X)_n (a_2^T X_i) \leq v\}} = \frac{1}{n} \sum_{i=1}^n I_{\{a_1^T X_i \leq (F_{a_1}^T X)_n^{-1}(u), a_2^T X_i \leq (G_{a_2}^T X)_n^{-1}(v)\}}$$

利用经验过程方法, 记

$$P_n f = \frac{1}{n} \sum_{i=1}^n I_{\{a_1^T X_i \leq x, a_2^T X_i \leq y\}}$$

根据经验过程的一致强大数定律,

$$\sup_f | P_n f - P f | \rightarrow 0, \text{ a. s.}$$

则 $S_n = \sup | \hat{C}_{a_1, a_2}(u, v) - \tilde{C}_{a_1, a_2}(u, v) | \rightarrow 0, \text{ a. s.}$ 另一方面根据引理4,

$$T_n = \sup_{(u,v) \in B} \sup_{(a_1, a_2)} | \tilde{C}_{a_1, a_2}(u, v) - C_{a_1, a_2}(u, v) | = \sup_{(u,v) \in B} \sup_{(a_1, a_2)} | H((F_{a_1}^T X)_n^{-1}(u), (G_{a_2}^T X)_n^{-1}(v)) - H(F_{a_1}^{-1}(u), G_{a_2}^{-1}(v)) | \rightarrow 0, \text{ a. s.}$$

于是 $\sup_{(u,v) \in B} \sup_{(a_1, a_2)} | \hat{C}_{a_1, a_2}(u, v) - C_{a_1, a_2}(u, v) | \rightarrow 0$. 那么

$$\begin{aligned} & \sup_{(a_1, a_2)} | h_n(a_1, a_2) - h(a_1, a_2) | = \\ & \sup_{(a_1, a_2)} | \iint_{I^2} [(\hat{C}_{a_1, a_2}(u, v) - uv)^2 - \\ & (C_{a_1, a_2}(u, v) - uv)^2] dudv | = \\ & \sup_{(a_1, a_2)} | \iint_{I^2} (\hat{C}_{a_1, a_2}(u, v) - C_{a_1, a_2}(u, v)) \times \\ & (\hat{C}_{a_1, a_2}(u, v) + C_{a_1, a_2}(u, v) - 2uv) dudv | \leq \\ & 4 \iint_{I^2} \sup_{(u,v) \in B} \sup_{(a_1, a_2)} | \hat{C}_{a_1, a_2}(u, v) - \\ & C_{a_1, a_2}(u, v) | dudv \rightarrow 0, \text{ a. s.} \end{aligned}$$

注1 定理1对 $h(a_1, a_2)$ 具有唯一极小值点的条件并不一定是一个点, 可取为一个集合. 利用文献[9]第2章中问题1的讨论, 记 D 为 $h(a_1, a_2)$ 的最小值点集合, 则在 D 上可以建立如定理1的结论, 只要满足 (a_{1n}, a_{2n}) 收敛到集合 D 中的某个

点以概率1成立.

2 典型相关分析

本章考虑另外一种多元分析方法, 随机向量之间的典型相关分析, 这是一种研究向量 $\mathbf{X}$ 和 $\mathbf{Y}$ 之间多元相关关系的方法. 受主成分分析思想的启发, 把两个随机向量之间的相关性转化为两个变量的相关性来考虑, 即考虑它们的线性组合之间的相关性. 做法如下: 首先寻找单位相关方向向量 $\mathbf{a}, \mathbf{b}$, 从而得到线性组合 $\mathbf{a}^T \mathbf{X}$ 和 $\mathbf{b}^T \mathbf{Y}$, 此时随机向量 $\mathbf{X}$ 和 $\mathbf{Y}$ 满足最大相关, 称 $\mathbf{a}, \mathbf{b}$ 为第一对典型变量. 然后可以给出第二对典型变量、第三对典型变量... 这样通过研究相关性较大的几对典型变量来了解两个高维随机向量之间的相关关系, 问题得以简化, 从而易于抓住本质.

典型相关分析要借助于一个相依度量指标函数来给出单位相关方向向量, 这里的指标函数是完成分析的关键. 指标函数(2)在本章同样适用, 不过由于是考虑最大相关, 因此要找极大值点来确定单位相关方向向量. 这里完成全部典型相关分析的步骤与上章类似, 因此不再重复. 下面主要给出样本形式的渐近性质, 若 $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_m$ 是 $n+m$ 个分别来自 $\mathbf{X}$ 和 $\mathbf{Y}$ 的观测值向量, 最优方向 $(\mathbf{a}_n, \mathbf{b}_n)$ 可以取为式(3)的极大值点. 仿照上章可以给出 $(\mathbf{a}_n, \mathbf{b}_n)$ 的强相合性结果.

定理2 若作为指标函数 $h(\mathbf{a}, \mathbf{b})$ 的极大值点 $(\mathbf{a}_0, \mathbf{b}_0)$ 具有唯一性, 且满足如下假设:

- (1) 对任意 u, v , 存在 $\delta_1 > 0, \delta_2 > 0$, 使得 $\inf_a f_{a^T \mathbf{X}}(x_v^a) > \delta_1 > 0$, 且 $\inf_b g_{b^T \mathbf{Y}}(y_v^b) > \delta_2 > 0$;
- (2) 对任意 $u, v, f_{a^T \mathbf{X}}(x)$ 在点 x_v^a 处等度连续, $g_{b^T \mathbf{Y}}(y)$ 在点 y_v^b 处等度连续;
- (3) $F_{a^T \mathbf{X}}$ 和 $G_{b^T \mathbf{Y}}$ 等度连续;

则

$$(\mathbf{a}_n, \mathbf{b}_n) \rightarrow (\mathbf{a}_0, \mathbf{b}_0), \text{ a. s. ; } n \rightarrow +\infty$$

此定理的证明过程与定理1的证明一样.

实际上本文的指标函数也可以换做其他的由连接函数构造的秩相关系数, 如肯德尔的 τ 和斯

皮尔曼的 ρ 等, 它们的 k 维形式为^[10]

$$\tau = \frac{1}{2^{k-1} - 1} \left[2^k \int_{I^k} C(u_1, \dots, u_k) dC(u_1, \dots, u_k) - 1 \right]$$

$$\rho = \frac{k+1}{2^k - (k+1)} \left[2^{k-1} \times \left(\int_{I^k} C(u_1, \dots, u_k) d\Pi(u_1, \dots, u_k) + \int_{I^k} \Pi(u_1, \dots, u_k) dC(u_1, \dots, u_k) \right) - 1 \right]$$

3 结 语

连接函数刻画了随机变量之间的相关性, 本文用其构造了两种多元分析方法: 投影寻踪主成分分析和典型相关分析. 同时证明了其强相合性, 因此这种构造方法具有较好的理论基础. 据此可以看出, 这两种方法具有丰富的应用前景.

参考文献:

- [1] FRIEDMAN J H, TUKEY J W. A projection pursuit algorithm for exploratory data analysis [J]. **IEEE Transactions on Computers**, 1974, **23**(9): 881-890
- [2] HUBER P J. Projection pursuit (with discussion)

- [J]. **Annals of Statistics**, 1985, **13**(2): 435-525
- [3] HALL P. On polynomial-based projection indices for exploratory projection pursuit [J]. **Annals of Statistics**, 1989, **17**(2): 589-605
- [4] NASON G P. Design and choice of projection indices [J]. **Journal of the Royal Statistical Society**, 2001, **63**(Part 3): 551-567
- [5] SCHWEIZER B, WOLFF E F. On nonparametric measures of dependence for random variables [J]. **Annals of Statistics**, 1981, **9**(4): 879-885
- [6] DEHEUVELS P. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance [J]. **Bulletin de la Classe des Sciences, 5e Serie**, 1979, **65**: 274-292
- [7] MA Jian, SUN Zeng-qi. Copula component analysis [J]. **Lecture Notes in Computer Science**, 2007, **4666**: 73-80
- [8] NELSEN R B. **An Introduction to Copulas** [M]. New York: Springer-Verlag, 1999
- [9] POLLARD D. **Convergence of Stochastic Processes** [M]. New York: Springer-Verlag, 1984
- [10] JOE H. **Multivariate Models and Dependence Concepts** [M]. New York: Chapman & Hall, 1997

Analyses of projection pursuit principal components based on Copulas

WANG Xiao-guang, SONG Li-xin*

(School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China)

Abstract: An approach to exploring the dependence relation of continuous random vectors is presented based on the dependence measure constructed by Copula. Copula is used to construct the index function in the analyses of the projection pursuit principal components with high-dimensional random vector, and the obtained optimal projection directions are strong consistency. In addition, the canonical correlation analyses of two or more random vectors are discussed. The conclusions show that the proposed method has excellent theoretical properties and practical prospects.

Key words: projection pursuit (PP); Copula; empirical Copula; analyses of principal components; canonical correlation analysis