系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
8 (3) (1988), 234—242

PP 型拟合优度检验*

张 航
(中国科学院系统科学研究所)

I. 引 言

摄影寻踪 (Projection Pursuit,简称 PP) 是一种新兴的用来处理高维数据的统计方法,其主要思想是通过极大化某个投影指标(通常是分布函数的泛函)来寻找低维投影,通过对其低维投影数据的研究来发现高维数据的性质. PP 方法自首次提出,已用于处理一些非正态多维数据分析问题,如回归、判别、聚类、密度估计等[2,4].

PP 的一个特点就是能把许多一维方法用于多维问题. 本文就是用 PP 方法,以经典的 χ^2 检验统计量作为投影指标,得到一种高维分布的拟合优度检验,并对其极限分布进行研究.

II. 主 要 结 果

$(R^k, \mathscr{B}^k)$ 是 k 维 Lebesque 可测空间, $S_1, \cdots, S_m$ 是 R^1 上的 m 个不相交的区间,具 $\bigcup_{i=1}^{m} S_i = R^1$. 本文中的 a, b 均代表 k 维方向. 令 $S_i^a = \{x: a^T x \in S_i, x \in R^k\}$; $I_i^a = Is_i^a$, Is_i^a 是 S_i^a 的示性函数.

P, G 为 $(R^k, \mathscr{B}^k)$ 上的概率分布, P 连续,已知或仅含有未知参数. 我们要检验

$$H_0: G = P \longleftrightarrow k: G \neq P.$$

有样本 $X_1, \cdots, X_n$ i.i.d $\sim G$, 对任意方向 a, $a^T X_1, \cdots a^T X_n$ i.i.d $\sim G_a$, 其中 G_a 是 G 在方向 a 上的边际分布.

传统的 χ^2 统计量 $z_n^a = \sum_{i=1}^{m} n(P_{ni}^a - \pi_i^a)^2/\pi_i^a$ 可以用来检验 G_a 是否等于 P_a, 这里 $\pi_i^a = P(S_i^a)$, $P_{ni}^a = \frac{1}{n}\sum_{j=1}^{n} I_i^a(X_j)$.

G_a 与 P_a 差别越大,则对于较大的样本, z_n^a 的值也较大,用 z_n^a 作检验应该越显著;反之,若 $\sup_a z_n^a$ 的值较小,则任何方向上, G_a 与 P_a 都相差不大, G 与 P 就没有明显差别. 所以可以考虑用 $z_n = \sup_a z_n^a$ 来作检验统计量.

下面我们分别对 F 已知和 F 含有未知参数两种情形给出 z_n 的极限分布. 为方便起

* 国家科学基金资助课题.
1986 年 11 月 17 日收到, 1987 年 9 月 20 日收到修改稿.

见，本文中实数的绝对值和向量的欧氏模都用$|\cdot|$表示，而把$\|\cdot\|$留作它用.

定理 1. 设 π_i^a 满足 $\inf_a \pi_i^a > 0$, $1 \leqslant i \leqslant m$. 则当 H_0 成立时，

$$\{z_n^a : |a| = 1\} \xrightarrow{\mathscr{L}} \{Y^a : |a| = 1\},$$

$$z_n \xrightarrow{\mathscr{L}} z = \sup_a (Y^a)^\tau Y^a,$$

其中 "$\xrightarrow{\mathscr{L}}$" 表示依分布收敛，$\{Y^a : |a| = 1\}$ 是一 m 维 Gauss 过程：

$$\forall a,\ Y^a \sim N(0,\ \Sigma^a),\ \Sigma^a = (\sigma_{ij}^a)_{m\times m},$$

$$\sigma_{ij}^a = \begin{cases} 1 - \pi_i^a, & i = j, \\ -\sqrt{\pi_i^a \pi_j^a}, & i \neq j; \end{cases}$$

$$\forall a, b,\ \begin{pmatrix} Y^a \\ Y^b \end{pmatrix} \sim N(0,\ A^{ab}),$$

$$A^{ab} = \begin{pmatrix} \Sigma^a & \Sigma^{ba} \\ \Sigma^{ab} & \Sigma^b \end{pmatrix},\ \Sigma^{ab} = \Sigma^{ba} = (\sigma_{ij}^{ab})_{m\times m},$$

$$\sigma_{ij}^{ab} = [P(S_i^a \cap S_j^b) - \pi_i^a \pi_j^b] / \sqrt{\pi_i^a \pi_j^b}.$$

定理 2. 设随机变量 X 的分布族为 $\{P_\theta : P_\theta \ll P,\ \theta \in \Theta\}$，其中 P 是 Lebesque 可测空间 $(R^k,\ \mathscr{B}^k)$ 中的连续概率测度，Θ 为 R^k 中一开集. 又 $X_1, \cdots, X_n, \cdots$ i.i.d$\sim P_\theta$, $\theta \in \Theta$ 为未知参数，记 $P_{ni}^a = \frac{1}{n}\sum_{j=1}^{n} I_i^a(X_j)$; $\pi_i^a(\varphi) = P_\varphi(S_i^a)$, $\varphi \in \Theta$. 设 $\pi_i^a(\varphi)$ 满足

1) $\pi_i^a(\varphi)$ 对 $(a,\ \varphi)$ 连续;

2) $\forall \varphi \in \Theta$, $\inf \pi_i^a(\varphi) > 0$, $i = 1, \cdots, m$;

3) 若 $\theta_1 \neq \theta_2$, 则 $\sum_{i=1}^{m} |\pi_i^a(\theta_1) - \pi_i^a(\theta_2)| > 0$, $\forall a$;

4) $\forall i = 1, \cdots, m$, $\partial \pi_i^a(\varphi)/\partial \varphi$ 在 Θ 内存在，对 (a, φ) 连续，且作为 φ 的函数，对 a 等度连续;

5) 方阵 $I^a(\varphi) = (I_{rs}^a(\varphi))$ 非异，其中

$$I_{rs}^a(\varphi) = \sum_{i=1}^{m} \frac{1}{\pi_i^a(\varphi)} \frac{\partial \pi_i^a(\varphi)}{\partial \varphi_r} \frac{\partial \pi_i^a(\varphi)}{\partial \varphi_s},\quad r, s = 1, \cdots, K,\ \forall \varphi \in \Theta.$$

则存在 $\hat{\theta}_n^a$ 为似然方程

$$\sum_{i=1}^{m} \frac{P_{ni}^a}{\pi_i^a(\varphi)} \frac{\partial \pi_i^a(\varphi)}{\partial \varphi_j} = 0,\ 1 \leqslant j \leqslant K$$

的一致相合解，即 $P(\sup_a |\hat{\theta}_n^a - \theta| \xrightarrow[n\to\infty]{} 0) = 1$.

若记 $\hat{K}_n^a = \sum_{i=1}^{m} n(P_{ni}^a - \pi_i^a(\hat{\theta}_n^a))^2 / \pi_i^a(\hat{\theta}_n^a)$, $K_n = \sup_a \hat{K}_n^a$，则有

$$K_n \xrightarrow{\mathscr{L}} K = \sup_a (Y^a)^\tau C^a Y^a,$$

其中 $\{Y^a : |a| = 1\}$ 与定理 1 中相似 $(P = P_\theta)$, $C^a = I_m - (B^a)(I^a(\theta))^{-1}(B^a)^\tau$, $B^a = (B_{ij})^a$, $B_{ij}^a = \dfrac{1}{\sqrt{\pi_i^a(\varphi)}} \dfrac{\partial \pi_i^a(\varphi)}{\partial \varphi_j}\Big|_{\varphi=\theta}$.

III. 预 备 知 识

$(S, \sigma(S), P)$ 为一概率空间，$\mathscr{F}$ 是 S 的子集类，称 $\mathscr{F}$ 是一多项式集类，如果存在一个多项式 $F(x)$，使 $\forall A \subset S$，$\#A = n$（$\#A$ 表示 A 中元素个数）有

$$\#\{A \cap D : D \in \mathscr{F}\} \leqslant F(n), n \text{ 为任何自然数.}$$

对于多项式集类有

引理 3.1 ([5], p.20). 设 $\mathscr{D}$ 为 S 上的实函数构成的一个有限维集合，则 $\{D : D = [s : d(s) \geqslant 0], d \in \mathscr{D}\}$ 是一多项式集类.

设 $X_1, X_2, \cdots, X_n$ 为 $(S, \sigma(S))$ 上的一组 i.i.d 样本，$X_1 \sim P$，记 P_n 为经验概率测度.

设 $\mathscr{F}$ 是 $(S, \sigma(S))$ 上的可测实函数集，$\mathscr{F}$ 有指标集 T，$\mathscr{F} = \{f(\cdot, t) : t \in T\}$. 设 T 为一可分度量空间，$\mathscr{B}(T)$ 是 T 上的 Borel 域. 我们称 $\mathscr{F}$ 是一可容许集，如果 $\mathscr{F}$ 的指标集 T 满足：

i) 函数 $f(\cdot, \cdot)$ 是 $\sigma(S) \times \mathscr{B}(T)$ 可测；

ii) T 是紧致度量空间的一个解析子集.

称 $\mathscr{D}$ 是可容许集，如果 $\{I_D : D \in \mathscr{D}\}$ 是一可容许函数集.

引理 3.2 ([5], p. 22). 若 $\mathscr{D}$ 为一可容许多项式集类，则

$$\sup_{\mathscr{D}} |P_n(D) - P(D)| \xrightarrow[n \to \infty]{} 0, \quad \text{a. s.}$$

若 f 为 S 上的实函数，f 的图集定义为 $G_f = \{(s, t) : 0 \leqslant t \leqslant f(s) \text{ 或 } 0 \geqslant t \geqslant f(s), s \in S\}$

$\mathscr{C}$ 为 S 上的实函数构成的空间，在上面定义了某种距离 d，$\mathscr{F} \subset \mathscr{C}$. $\forall \varepsilon > 0$，$\mathscr{F}$ 的 ε-覆盖数 $N(\varepsilon, d, \mathscr{F})$ 定义为 $\min\{i : \exists\, i \text{ 个 } f_1 \cdots f_i \in \mathscr{C} \text{ 使得 } \min_{1 \leqslant j \leqslant i} d(f_j, f) < \varepsilon, \forall f \in \mathscr{F}\}$.

记 $L^1(P) = \{f : \int |f| dP < \infty\}$，$\mathscr{L}^2(P) = \{f : \int f^2 dP < \infty\}$，当 $\mathscr{C}$ 为 $L^1(P)$，$L^2(P)$ 距离 d 为 L_P^1，L_P^2 即 $\int |f - g| dP$，$\left(\int |f - g|^2 dP\right)^{\frac{1}{2}}$ 时，相应的覆盖数分别记作 $N_1(\varepsilon, P, \mathscr{F})$，$N_2(\varepsilon, P, \mathscr{F})$.

若 $F \in \mathscr{C}$，且满足：$\forall f \in \mathscr{F}$，$|f(s)| \leqslant F(s)$，$\forall s \in S$，称 F 是 $\mathscr{F}$ 的封套.

引理 3.3 ([5], p. 27). 设 $\mathscr{F}$ 是 S 上的一个可测函数集，有封套 F，Q 为 S 上概率测度，$0 < \int F dQ < \infty$，如果 $\mathscr{F}$ 的图象集构成一个多项式集类，则 $N_1(\varepsilon QF, Q, \mathscr{F}) \leqslant A\varepsilon^{-w}$. 这里非零常数 A，w 只与 $\mathscr{F}$ 的图集有关，同 Q 无关，$QF = \int F dQ$.

记 $E_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Ef(X_i))$，$f$ 为一可积函数. $\forall \delta > 0$，定义随机覆盖积分

$$J_2(\delta, P_n, \mathscr{F}) = \int_0^{\delta} [2 \log (N_2(u, P_n, \mathscr{F})^2/u)]^{\frac{1}{2}} du. \tag{1}$$

引理 3.4 ([5], p. 150). 设 $\mathscr{F}$ 是 $\mathscr{L}^2(P)$ 的可容许函数集，有封套 $F\in\mathscr{L}^2(P)$，假设 $\forall\eta>0$，$\varepsilon>0$，$\exists\nu>0$，使得 $\lim\limits_{n\to\infty}\sup \mathrm{P}[J_2(\nu, P_n, \mathscr{F})>\eta]<\varepsilon$，则 $\exists\delta>0$，使 $\lim\limits_{n\to\infty}\sup \mathrm{P}[\sup\limits_{[\delta]}|E_n(f-g)|>\eta]<\varepsilon$. 这里 $[\delta]=\{(f,g): f,g\in\mathscr{F}, \rho_P(f-g)\leqslant\delta\}$，$\rho_P$ 为 $\mathscr{L}^2_P$ 距离.

设 $\mathscr{F}$ 为一逐点有界的可容许函数集，且 $\sup_{\mathscr{F}}|\int f dP|<\infty$. 记 $\mathscr{X}$ 为 $\mathscr{F}$ 上所有有界函数组成的空间，随机过程 $E_n=\{E_nf: f\in\mathscr{F}\}$ 可看成是 $\mathscr{F}$ 上的随机有界函数，其样本轨道 (Sample path) 属于 $\mathscr{X}$，在 $\mathscr{X}$ 上定义一致距离：$x, y\in\mathscr{X}$，$d(x,y)=\|x-y\|=\sup_{\mathscr{F}}|x(f)-y(f)|$. 又令 $C(\mathscr{F}, P)=\{x: x\in\mathscr{X}, x$ 对 $\mathscr{F}$ 中的距离 ρ_P 而言一致连续$\}$.

引理 3.5 ([5], p. 157). 设 $\mathscr{F}$ 是一个逐点有界、全有界、可容许的 $L^2(P)$ 子集，如果 $\forall\eta>0$，$\varepsilon>0$，$\exists\delta>0$，使得

$$\lim_{n\to\infty}\sup \mathrm{P}\{\sup_{[\delta]}|E_n(f-g)|>\eta\}<\varepsilon,$$

则作为 $\mathscr{X}$ 中的随机元序列，$\{E_n\}$ 依律收敛到 $\mathscr{X}$ 中的随机元 E_P，$E_P=\{E_Pf: f\in\mathscr{F}\}$ 是一个紧 (tight) 的高斯过程，其样本轨道在 $C(\mathscr{F}, P)$ 中，E_Pf 期望为零，E_Pf 和 E_Pg 的协方差为 $\int fgdP-\left(\int fdP\right)\left(\int gdP\right)$，$\forall f, g\in\mathscr{F}$.

由引理 3.4 和引理 3.5 立即得到

推论 3.1. 设 $\mathscr{F}$ 是 $L^2(P)$ 中逐点有界，全有界且可容许的子集，有封套 $F\in L^2(P)$ 若对 $\forall\eta>0, \varepsilon>0, \exists\nu>0$，使 $\limsup \mathrm{P}[J_2(\nu, P_n, \mathscr{F})>\eta]<\varepsilon$，则 E_n 依律收敛到引理 3.5 中的高斯过程 E_P.

IV. 定 理 证 明

现在回到第 II 节的问题，沿用前面的记号，令 $\mathscr{F}_i=\{I_i^a: |a|=1\}$，$i=1, 2, \cdots, m$. 由于 $\{a^Tx-t: |a|=1, t\in R^1\}$ 是 R^k 中的有限维实函数集，应用引理 3.1 易得 $\mathscr{D}_i=\{S_i^a: |a|=1\}$，$i=1, \cdots, m$ 是多项式集类，再由 $I_i^a=I_{S_i^a}$ 只取 0，1 两个值的性质，可以证明 $\mathscr{F}_i$ 的图集也是多项式集类.

故由引理 3.3 知，$\exists A, w$，使对任何概率测度 Q 有 $N_1(\varepsilon, Q, \mathscr{F}_i)\leqslant A\varepsilon^{-w}(\forall\varepsilon>0)$，由此即知，$\mathscr{F}_i$ 是全有界的.

因为在 $\mathscr{F}_i$ 中 L_P^1 距离等于 L_P^2 距离的平方，故对 $\forall\, 0<\delta<\frac{1}{e}$，随机覆盖积分的定义 (1) 给出

$$\begin{aligned}J_2(\delta, P_n, \mathscr{F}_i)&=\int_0^\delta[2\log(N_2(u, P_n, \mathscr{F}_i)^2/u)]^{\frac{1}{2}}du\\&=\int_0^\delta[2\log(N_1(u^2, P_n, \mathscr{F}_i)^2/u)]^{\frac{1}{2}}du\\&\leqslant\int_0^\delta[2\log(A^2u^{-4w-1})]^{\frac{1}{2}}du\end{aligned}$$

$$\leqslant C\int_0^{\delta} -\log u du = C(\delta - \delta\log\delta) \xrightarrow[\delta\to 0]{} 0,$$

其中 C 为只与 A，w 有关的正常数.

由于 $\{a:|a|=1\}$ 是紧致度量空间，容易验证 $\mathscr{F}_i(1\leqslant i\leqslant m)$ 是可容许集；$\mathscr{F}_i$ 逐点有界，有封套 $F=1$ 都是显然的，从而推论 3.1 的一切条件满足. 记 $\mathscr{X}_i$ 是定义在 $\mathscr{F}_i$ 上的有界实函数全体，由推论 3.1 可知

$$\{E_n f : f\in\mathscr{F}_i\} \xrightarrow{\mathscr{L}} \{E_P f : f\in\mathscr{F}_i\},\ i=1,2,\cdots,m.$$

回到第 II 节的记号，并注意 E_n 的定义(见第 III 节)，上式即是

$$\{\sqrt{n}\,(P_{ni}^a - \pi_i^a) : |a|=1\} \to \{E_P(I_i^a) : |a|=1\}. \tag{2}$$

为完成定理，还需要下面的引理.

引理 4.1. 设 $\mathscr{X}$ 是 $\mathscr{F}$ 上有界实值泛函全体，$X_n(n=1,2,\cdots)$ 和 X 都是 $\mathscr{X}$ 中的随机元，$X_n \xrightarrow{\mathscr{L}} X$ 且 X 的样本轨道在 $C(\mathscr{F},P)$ 中，又 $\alpha=\{\alpha(f):f\in\mathscr{F}\}$ 满足 $\|\alpha\| = \sup_{\mathscr{F}}|\alpha(f)| < \infty$，则

$$\{X_n(f)\alpha(f) : f\in\mathscr{F}\} \xrightarrow{\mathscr{L}} \{X(f)\alpha(f) : f\in\mathscr{F}\}.$$

证. 定义 H 为从 $\mathscr{X}$ 到 $\mathscr{X}$ 的映象：$\forall x\in\mathscr{X}$，$(Hx)(f) = x(f)\alpha(f)\ (f\in\mathscr{F})$. 由于 $\|Hx-Hy\| = \sup_{\mathscr{F}}|x(f)\alpha(f) - y(f)\alpha(f)| \leqslant \sup_{\mathscr{F}}|\alpha(f)|\sup_{\mathscr{F}}|x(f)-y(f)| = \|\alpha\|\cdot\|x-y\|$，故 H 是连续映象；又 X 的样本轨道都在 $C(\mathscr{F},P)$ 中，由连续映象定理 ([5], p. 70) 知 $HX_n \xrightarrow{\mathscr{L}} HX$. 这就是本引理的结论.

定理 1 的证明

先固定 $\mathscr{F}_i$ 和 $\mathscr{X}_i\ (i=1\cdots m)$. 对每个 $I_i^a\in\mathscr{F}_i$，令 $\alpha(I_i^a) = (P(S_i^a))^{-\frac{1}{2}} = (\pi_i^a)^{-\frac{1}{2}}$. 由于 $\inf_a \pi_i^a > 0$，依引理 4.1 和 (2) 可知

$$\{\sqrt{n}\,(P_{ni}^a - \pi_i^a)/\sqrt{\pi_i^a} : |a|=1\} \xrightarrow{\mathscr{L}} \{B_P(I_i^a)/\sqrt{\pi_i^a} : |a|=1\}. \tag{3}$$

记 $\mathscr{X} = \mathscr{X}_1\times\cdots\times\mathscr{X}_m$. 注意到 $\mathscr{X}_i$ 的元素 $x_i = \{x_i(f) : f\in\mathscr{F}_i\}$ 的指标集实际上是 $\{a:|a|=1\}$. 为方便起见，可以把 $x\in\mathscr{X}$ 表示成 $x = \{(x_1(a),\cdots,x_m(a))^{\mathrm{T}} : |a|=1\}$. 定义 $\mathscr{X}$ 上距离为

$$\|x-y\| = \sup\{|x_i(a)-y_i(a)| : |a|=1, i=1,2,\cdots,m\}.$$

令

$$Y_n^a = (Y_{n1}^a,\cdots,Y_{nm}^a)^{\mathrm{T}} = (\sqrt{n}\,(P_{n1}^a-\pi_1^a)/\sqrt{\pi_1^a},\cdots,\sqrt{n}\,(P_{nm}^a-\pi_m^a)/\sqrt{\pi_m^a})^{\mathrm{T}},$$

$$Y^a = (Y_{(1)}^a,\cdots,Y_{(m)}^a)^{\mathrm{T}} = (B_P(I_1^a)/\sqrt{\pi_1^a},\cdots,B_P(I_m^a)/\sqrt{\pi_m^a})^{\mathrm{T}},$$

$$Y_n = \{Y_n^a : |a|=1\},\quad Y = \{Y^a : |a|=1\}.$$

显然 $Y_n(n=1,2,\cdots)$ 和 Y 都是 $\mathscr{X}$ 上的随机元，由 (3) 不难证明

$$\{Y_n^a : |a|=1\} \xrightarrow{\mathscr{L}} \{Y^a : |a|=1\}.$$

由高斯过程 B_P 的性质 (参见引理 3.5) 即知 $Y_{(i)}^a$ 的期望为 0 $(i=1,\cdots m;\ |a|=1)$.

$$\mathrm{cov}(Y_{(i)}^a, Y_{(j)}^b) = \int (I_i^a/\sqrt{\pi_i^a})(I_j^b/\sqrt{\pi_j^b})dP - \int I_i^a/\sqrt{\pi_i^a}\,dP\int I_j^b/\sqrt{\pi_j^b}\,dP$$

$$= [P(S_i^a \cap S_j^b) - \pi_i^a \pi_j^b] / \sqrt{\pi_i^a \pi_j^b}.$$

最后，注意到 $Mx \triangleq \sup_a \sum_{i=1}^{m} x_i^2(a)$ 是 $\mathscr{R}$ 上连续映象，由连续映象定理（[5], p. 70）即得

$$z_n = MY_n \xrightarrow{\mathscr{L}} MY = z.$$

定理 2 的证明

首先，由于 $(P_{n1}^a, \cdots, P_{nm}^a)$ 服从多项分布 $M(n; \pi_1^a(\theta), \cdots, \pi_m^a(\theta))$，不难看出似然方程为

$$\sum_{i=1}^{m} \frac{P_{ni}^a}{\pi_i^a(\varphi)} \frac{\partial \pi_i^a(\varphi)}{\partial \varphi_j} = 0, \quad j = 1, 2, \cdots K. \tag{4}$$

记球面 $C(\delta) = \{\varphi : |\varphi - \theta| = \delta, \ \varphi \in \Theta\}$，由题设知：向量 $(\pi_i^a(\theta), \ i = 1 \cdots m)$ 及 $(\pi_i(\varphi), \ i = 1 \cdots m)$ 是不同的，从而它们的 Kullback-Leibler 信息数

$$A^a(\varphi) \triangleq \sum_{i=1}^{m} \pi_i^a(\theta) \log \frac{\pi_i^a(\theta)}{\pi_i^a(\varphi)}$$

在球面 $C(\delta)$ 上为正（参见 [3]）又 $\pi_i^a(\varphi)$ 关于 (a, φ) 连续，故 $\inf\{A^a(\varphi) : |a| = 1, \varphi \in C(\delta)\} > 0$. 由于 $\{S_i^a : |a| = 1\}$ 是多项式集类，依引理 3.2，有

$$P_\theta\{\sup_a |P_{ni}^a - \pi_i^a(\theta)| \to 0, (n \to \infty)\} = 1, \ i = 1, \cdots, m.$$

于是概率为 1 地有

$$\inf\left\{\sum_{i=1}^{m} P_{ni}^a \log \frac{\pi_i^a(\theta)}{\pi_i^a(\varphi)} : |a| = 1, \ \varphi \in C(\delta)\right\} > 0, \ n \text{ 足够大时.}$$

即对 $|a| = 1$ 和 $\varphi \in C(\delta)$ 一致地有

$$\sum_{i=1}^{m} P_{ni}^a \log \pi_i^a(\theta) > \sum_{i=1}^{m} P_{ni}^a \log \pi_i^a(\varphi), \ n \text{ 足够大时.}$$

这说明除了一个零测集外，对数似然函数 $l_n^a(\varphi) \triangleq \sum_{1}^{m} P_{ni}^a \log \pi_i^a(\varphi)$ 在闭球 $|\theta - \varphi| \leqslant \delta$ 中心 θ 处的值，大于在球面 $C(\delta)$ 上任一点的值. 由 $\pi_i^a(\varphi)$ $(1 \leqslant i \leqslant m)$ 的连续性知 $l_n^a(\varphi)$ 在 $|\theta - \varphi| \leqslant \delta$ 内至少有一个局部极大点，该极大点显然是似然方程 (4) 的一个根. 取 $\delta_n \downarrow 0$. 记 $\hat{\theta}_n^a = \hat{\theta}_n^a(X_1, \cdots, X_n)$ 为 $l_n^a(\varphi)$ 在 $C(\delta_n)$ 中的一个局部极大点，则 $\hat{\theta}_n^a$ 满足方程 (4)，且显然有

$$P_\theta\{\sup_{|a|=1} |\hat{\theta}_n^a(X_1, \cdots, X_n) - \theta| \to 0 (n \to \infty)\} = 1. \tag{5}$$

这就是本定理的第一个结论.

对于偏导列向量 $\frac{\partial \pi_i^a(\varphi)}{\partial \varphi}$，记

$$\frac{\partial \pi_i^a(\varphi)}{\partial \varphi}\Big|_{\varphi=\mu} \triangleq \frac{\partial \pi_i^a}{\partial \mu},$$

则由 (5) 和 $\pi_i^a(\varphi)$ 及 $\frac{\partial \pi_i^a(\varphi)}{\partial \varphi}$ 的连续性知

$$P_\theta\{\sup_{|a|=1}|\pi_i^a(\hat\theta_n^a)-\pi_i^a(\theta)|\to 0\ (n\to\infty)\}=1, \tag{6}$$

$$P_\theta\left\{\sup_{|a|=1}\left|\frac{\partial\pi_i^a}{\partial\hat\theta_n^a}-\frac{\partial\pi_i^a}{\partial\theta}\right|\to 0\ (n\to\infty)\right\}=1. \tag{7}$$

记

$$\eta_{ni}^a=\sqrt{n}\ (P_{ni}^a-\pi_i^a(\hat\theta_n^a))/\sqrt{\pi_i^a(\hat\theta_n^a)},\ \eta_n^a=(\eta_{n1}^a,\cdots,\eta_{nm}^a);$$

$$Y_{ni}^a=\sqrt{n}\ (P_{ni}^a-\pi_i^a(\theta))/\sqrt{\pi_i^a(\theta)},\ Y_n^a=(Y_{n1}^a,\cdots,Y_{nm}^a);$$

$$u_{ni}^a=\sqrt{n}\ (\pi_i^a(\hat\theta_n^a)-\pi_i^a(\theta))/\sqrt{\pi_i^a(\theta)},\ u_n^a=(u_{n1}^a,\cdots,u_{nm}^a).$$

显然 $K_n^a=(\eta_n^a)^r(\eta_n^a)$,而且由(6)知

$$P_\theta\{\sup_{|a|=1}|\eta_n^a-(Y_n^a-u_n^a)|\to 0\quad(n\to\infty)\}=1. \tag{8}$$

因此只须求出 $\{Y_n^a-u_n^a:|a|=1\}$ 的极限分布就容易得到 $\{K_n^a:|a|=1\}$ 和 K_n 的极限分布.

依定理 1,并以 P_θ 代替定理 1 中的 P,得

$$Y_n\triangleq\{Y_n^a:|a|=1\}\xrightarrow{\mathscr{L}}\{Y^a:|a|=1\}\triangleq Y, \tag{9}$$

其中 Y 是一个为定理 1 所述的高斯过程,不过 $P=P_\theta$.

下面看 u_n^a. 利用中值公式有

$$u_{ni}^a=\sqrt{n}\cdot\frac{1}{\sqrt{\pi_i^a(\theta)}}\left(\frac{\partial\pi_i^a}{\partial\bar\theta}\right)^r(\hat\theta_n^a-\theta),$$

其中 $\bar\theta$ 的各分量 $\bar\theta_j$ 在 $\hat\theta_n^a$ 和 θ 的相应分量 $\hat\theta_{nj}^a$ 和 θ_j 之间. 记

$$B^a=\left(\frac{\partial\pi_1^a}{\partial\theta}\Big/\sqrt{\pi_1^a(\theta)},\cdots,\frac{\partial\pi_m^a}{\partial\theta}\Big/\sqrt{\pi_m^a(\theta)}\right)$$

为 $K\times m$ 矩阵,利用(7)有

$$P_\theta\{\sup_{|a|=1}|u_n^a-B^a[\sqrt{n}\,(\hat\theta_n^a-\theta)]|\to 0\ (n\to\infty)\}=1. \tag{10}$$

由于 $\hat\theta_n^a$ 满足(4)又 $\sum_{i=1}^a\pi_i^a(\varphi)=1\ (\varphi\in\Theta)$ 给出了 $\sum_{i=1}^m\frac{\partial\pi_i^a}{\partial\hat\theta_n^a}=0$. 再利用中值公式可得

$$\begin{aligned}z_{nr}^a&\triangleq\sum_{i=1}^m\frac{\sqrt{n}\ (P_{ni}^a-\pi_i^a(\theta))}{\pi_i^a(\hat\theta_n^a)}\frac{\partial\pi_i^a}{\partial\hat\theta_{nr}^a}\\&=\sum_{i=1}^m\frac{\sqrt{n}\ (\pi_i^a(\hat\theta_n^a)-\pi_i^a(\theta))}{\pi_i^a(\hat\theta_n^a)}\frac{\partial\pi_i^a}{\partial\hat\theta_{nr}^a}\\&=\sum_{i=1}^m\frac{\sqrt{n}}{\pi_i^a(\hat\theta_n^a)}\sum_{s=1}^K(\hat\theta_{ns}^a-\theta_s)\frac{\partial\pi_i^a}{\partial\bar\theta_s}\frac{\partial\pi_i^a}{\partial\hat\theta_{nr}^a}\\&=\sum_{s=1}^K\sqrt{n}\ (\hat\theta_{ns}^a-\theta_s)\sum_{i=1}^m\frac{1}{\pi_i^a(\hat\theta_n^a)}\frac{\partial\pi_i^a}{\partial\bar\theta_s}\frac{\partial\pi_i^a}{\partial\hat\theta_{nr}^a},\end{aligned}$$

其中 $\bar\theta_s$ 在 $\hat\theta_{ns}^a$ 和 θ_s 之间. 利用(6)和(7)可知

$$P_\theta\left\{\sup_{|a|=1}\left|\sum_{i=1}^m\frac{1}{\pi_i^a(\hat\theta_n^a)}\frac{\partial\pi_i^a}{\partial\bar\theta_s}\frac{\partial\pi_i^a}{\partial\hat\theta_{nr}^a}-I_{rs}^a(\theta)\right|\to 0\ (n\to\infty)\right\}=1.$$

记 $z_n^a=(z_{n1}^a,\cdots,z_{nK}^a)$ 即得

$$P_\theta\{\sup_{|a|=1}|z_n^a - I^a(\theta)\sqrt{n}(\hat{\theta}_n^a - \theta)| \to 0 \ (n\to\infty)\} = 1. \tag{11}$$

另一方面由 z_n^a 的定义和(6),(7)两式,我们有

$$P_\theta\{\sup_{|a|=1}|z_n^a - (B^a)^\tau Y_n^a| \to 0 \ (n\to\infty)\} = 1.$$

再利用(10)可知

$$P_\theta\{\sup_{|a|=1}|u_n^a - B^a(I^a(\theta))^{-1}(B^a)^\tau Y_n^a| \to 0 \ (n\to\infty)\} = 1.$$

注意到 $C^a = I_m - B^a(I^a(\theta))^{-1}(B^a)^\tau$,我们有

$$P_\theta\{\sup_{|a|=1}|(Y_n^a - u_n^a) - C^a Y_n^a| \to 0 (n\to\infty)\} = 1.$$

由(9),(8)和(7),利用连续映象定理和类似定理1的推理便得

$$K_n = \sup_{|a|=1} K_n^a \xrightarrow{\mathscr{L}} K = \sup_{|a|=1}(Y^a)^\tau C^a Y^a$$

这就完成了定理证明.

V. Bootstrap 逼近

由于定理1,定理2中最后的极限分布不能精确算出来,所以检验的渐近否定域就不好构造出来. 利用 Bootstrap 方法,可以解决这一困难.

设 P 如定理1中所述, $X_1,\cdots,X_n$ i.i.d $\sim P$. P_n 为 $X_1,\cdots,X_n$ 的经验分布、取样本 $\bar{X}_1,\cdots,\bar{X}_n$ i.i.d $\sim P_n$. 设 $a_1, a_2,\cdots, a_{K_n}$ i.i.d $\sim \mu$, 这里 μ 是 K 维单位球面上的均匀分布,令

$$\bar{P}_{ni}^a = \frac{1}{n}\sum_{j=1}^n I_i^n(\bar{X}_j),$$

$$\bar{z}_n^a = \sum_{i=1}^m \frac{n(\bar{P}_{ni}^a - P_{ni}^a)^2}{P_{ni}^a},$$

$$\bar{z}_n = \max_{1\le k\le K_n} \bar{z}_n^a, \quad \bar{\bar{z}}_n = \sup_a \bar{z}_n^a.$$

则我们可以得到

定理 1′. 设定理1的条件满足. 设 $t_n(\alpha, P_n)$ 为 $\bar{z}_n$ 的 $1-\alpha$ 分位点, $0<\alpha<1$. $\bar{t}_n(\alpha, P_n)$ 是 $\bar{\bar{z}}_n$ 的 $1-\alpha$ 分位点, r 为 $\sup_a(Y^a)^\tau Y^a$ (Y^a 如定理1)的 $1-\alpha$ 分位点,则以概率为1地有

$$t_n(\alpha, P_n) \to r \text{ 以概率 } \mu^{K_n}\times P_n^n,$$

$$\bar{t}_n(\alpha, P_n) \to r \text{ 以概率 } P_n^n,$$

该结论的证明,作者将在另一文中讨论.

本文在成平、李国英教授的指导下完成的,在此表示感谢.

参　考　文　献

[1] Billingsley, P., Convergence of Probability Measures, Wiley, New York, 1968.
[2] 成平,李国英等,投影寻踪讲义,中国科学院系统所统计室, 1985.

[3] 陈希孺,数理统计引论,科学出版社,北京 1981 年.
[4] Huber, P.J., Projection pursuit, *Ann. Statist.*, **13**(1985), 435—475.
[5] Pollard, D., Convergence of stochastic Processes, New York, Springer-Verlag, 1984.

A PP GOODNESS-OF-FIT TEST AND ITS ASYMPTOTIC PROPERTIES

ZHANG HANG

(Institute of Systems Science, Academia Sinica)

ABSTRACT

By the use of the Projection Pursuit method, we get a multivariate goodness-of-fit test with the classical χ^2-test statistic as an index. The asymptotic propeerties of the test are discussed, and the asymptotic rejection region is obtained by bootstraping the test statistic.