系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
9 (4) (1989), 370—382

PP 检验的渐近功效[1)]

张 健　　成 平
（中国科学院系统科学研究所）

引　言

投影寻踪 (Projection Pursnit，简称 PP) 是一种处理高维数据的统计方法，近年来借助于计算机的发展，它的理论得到了迅速的发展. 假如 Q 为一能反映统计性质的指标，对每个 p 维方向 a，计算出 $Q(a^\tau X_1,\cdots,a^\tau X_n)$，从中找出使 $Q(a^\tau X_1,\cdots,a^\tau X_n)$ 最大的方向 a_0，通过研究数据 $a_0^\tau X_1,\cdots,a_0^\tau X_n$ 的性质，来了解原数据 $X_1,\cdots,X_n$ 的性质，这就是所谓数值 PP 的基本想法. 如果用一维检验或估计统计量 Q 作指标，则可得到多维检验或估计 $\sup_{\|a\|=1} Q(a^\tau X_1,\cdots,a^\tau X_n)$. 见文献 [2,3,5—7]. 在上述文章中这些检验和估计的渐近性质得到了部分研究，但当对立假设成立时，有关这些检验统计量的渐近分布问题所得结果尚不多，本文系统地研究了这类问题. 我们部分地采用[4]中的记号：$Pf(X)$ 记 $f(X)$ 的期望 ($X\sim P$)，P_n 为经验分布，"$\rightsquigarrow$"记依分布收敛.

§1. 主要结果与证明

定理. 设有概率分布 P，$X\sim P$，分布泛函 $V(t)\triangleq V(t,P)$ 及样本泛函 $V_n(t)\triangleq V_n(t,P_n)$，$t\in\pi$，其中 P_n 是经验分布，$\pi\subseteq\mathscr{G}$，$(\mathscr{G},d)$ 是一距离空间，$V(t)$ 关于距离 d 在 π 上连续，$B=\{t\in\pi: V(t)=\sup_{t\in\pi}V(t)\}$ 非空. 令

$$S_n(t)\triangleq\sqrt{n}\,(V_n(t)-V(t)),\ n\geqslant 1,\ t\in\pi.$$

若存在过程 $\{S(t): t\in\pi\}$ 满足

$$\sup_{t\in\pi}|S_n(t)-S(t)|\to 0\ \text{a.s}(P) \qquad (*)$$

且 $\{S(t): t\in\pi\}$ 的样本轨迹在 π 上一致连续，一致有界. 则

$$\sqrt{n}\,(\sup_{t\in\pi}V_n(t)-\sup_{t\in\pi}V(t))\to\sup_{t\in B}S(t),\ \text{a.s.}(P).$$

注. 1) 对 inf 也有同样的结论.

2) 条件 $S(t)$ 在 π 上一致连续，可改成存在 $\delta>0$，使 $S(t)$ 在 $B^\delta=\{t\in\pi: d(t,B)<\delta\}$ 内一致连续，则定理仍然成立.

证. 因为 $V(t)$ 连续，所以 B 是闭集，令 $B^c\triangleq\pi-B$. 现令 $h=\sup_{t\in\pi}V(t)$，则

1989 年 7 月 5 日收到.
1) 献给系统科学研究所建所十周年 (1979—1989).

$$\begin{aligned}
&\sqrt{n}\,(\sup_{t\in\pi} V_n(t) - \sup_{t\in\pi} V(t)) \\
&= \sqrt{n}\,\sup_{t\in\pi}(V_n(t) - h) \\
&= \sqrt{n}\max\{\sup_{t\in B}(V_n(t) - h),\ \sup_{t\in B^c}(V_n(t) - h)\} \\
&= \sqrt{n}\max\{\sup_{t\in B}(V_n(t) - V(t)),\ \sup_{t\in B^c}(V_n(t) - h))\} \\
&= \max\{\sup_{t\in B} S_n(t),\ \sup_{t\in B^c}(S_n(t) + \sqrt{n}\,(V(t) - h)\},
\end{aligned}$$

现任意固定使(*)式收敛性成立的样本点。

由 $S(t)$ 在 π 上一致连续性知

$\forall \varepsilon > 0$，$\exists \delta > 0$，使当 $t_1, t_2 \in \pi$，$d(t_1, t_2) < \delta$ 时，有 $|S(t_1) - S(t_2)| < \varepsilon$。

因为 B 是闭集，所以任给 $t \in B^c \cap \{t: d(t, B) < \delta\}$，存在 $b_t \in B$，使 $d(b_t, t) = d(t, B) < \delta$，从而

$$S(t) < \varepsilon + S(b_t) \leqslant \varepsilon + \sup_{t\in B} S(t)，\text{ 也即}$$

$$\begin{aligned}
&\sup_{t\in B^c\cap\{t: d(t,B)<\delta\}} (S_n(t) - S(t) + S(t) + \sqrt{n}\,(V(t) - h)) \\
&\leqslant \sup_{t\in\pi}|S_n(t) - S(t)| + \varepsilon + \sup_{t\in B} S(t) \to \varepsilon + \sup_{t\in B} S(t).
\end{aligned}$$

而

$$\begin{aligned}
&\sup_{t\in B^c\cap\{t: d(t,B)\geqslant\delta\}} (S_n(t) - S(t) + S(t) + \sqrt{n}\,(V(t) - h)) \\
&\leqslant \sup\{\sup_{t\in\pi}|S_n - S(t)| + \sup_{t\in\pi} S(t) + \sqrt{n}\,(V(t) - h):\\
&\qquad t \in B^c \cap \{t: d(t, B) \geqslant \delta\}\} \xrightarrow[n\to+\infty]{} -\infty.
\end{aligned}$$

由 ε 的任意性知：

$$\overline{\lim_{n\to\infty}}\ \sup_{t\in B^c}(S_n(t) + \sqrt{n}\,(V(t) - h)) \leqslant \sup_{t\in B} S(t),$$

显然，$\sup_{t\in B} S_n(t) \to \sup_{t\in B} S(t)$。所以，$\sqrt{n}\,(\sup_{t\in\pi} V_n(t) - \sup_{t\in\pi} V(t)) \to \sup_{t\in B} S(t)$。

推论 1. M-型 PP 指标其最大特征根的极限分布。

设 $\phi(t_1, t_2)$ 关于 t_2 单调减(或是 t_2 的偶函数且在 $(0,\infty)$ 上单调减)。T^a 是 $\int \phi(a^\tau X, t)\,dp = 0$ 的最小解，T_n^a 是 $\int \phi(a^\tau X, t)\,dP_n = 0$ 的最小解，P_n 是经验分布。

设 $N_{\varepsilon_0} = \{(b, t): \exists a \in S_p,\ \|b - a\| \leqslant \varepsilon_0 | t - T^a| \leqslant \varepsilon_0\}$，$S_p = \{a \in \mathbf{R}^p;\ \|a\| = 1\}$，$\varepsilon_0 \geqslant 0$。

假定 $\dfrac{\partial\phi(t_1, t_2)}{\partial t_2}$ 二元连续，$P\,\dfrac{\partial\phi(a^\tau X, T^a)}{\partial t_2} \neq 0$，且存在 $M_i(X), 0 \leqslant i \leqslant 4$，满足：

$$\sup_{N_{t_0}}|\phi(a^\tau X,t^\tau)|\leqslant M_0(X),\quad \sup_{N_{t_0}}\left|\frac{\partial\phi(a^\tau X,t^\tau)}{\partial t_1}\right|\leqslant M_1(X),$$

$$\sup_{N_{t_0}}\left|\frac{\partial\phi(a^\tau X,t^\tau)}{\partial t_2}\right|\leqslant M_2(X),\quad \sup_{N_{t_0}}\left|\frac{\partial^2\phi(a^\tau X,t^\tau)}{\partial t_1\partial t_2}\right|\leqslant M_3(X),$$

$$\sup_{N_{t_0}}\left|\frac{\partial^2\phi(a^\tau X,t^\tau)}{\partial t_2^2}\right|\leqslant M_4(X),\ \text{且}$$

$$PM_0^2(X)<+\infty,\ PM_0^2(X)\|X\|^\delta<+\infty\ (\text{某}\ \delta>0),\ PM_1(X)\|X\|<+\infty,$$

$$PM_1^2(X)<+\infty,\ PM_2^2(X)<+\infty,\ PM_3(X)\|X\|<+\infty,PM_4(X)<+\infty.$$

则 $\sqrt{n}(\sup_{\|a\|=1}T_n^a-\sup_{\|a\|=1}T^a)\rightsquigarrow\sup_{a\in B}W(h(a))$，其中

$$h(a)\triangleq h(a,X)=\frac{\phi(a^\tau X,T^a)}{-P\dfrac{\partial\phi(a^\tau X,T^a)}{\partial t_1}},\quad \mathscr{F}=\{h(a):\|a\|=1\},$$

W 是 $\mathscr{F}$ 上的 P-桥，$B=\{b:\|b\|=1,\ T^b=\sup_{\|a\|=1}T^a\}$. P-桥的定义见 Pollard[4].

注记. 如果 $\phi(t_1,t_2)=\dfrac{t_1^2-t_2^2}{t_1^2+t_2^2}$，$X\sim$ 椭球等高分布，则只要 $P\|X\|^\delta<+\infty$（某 $\delta>0$），本定理的结论就成立.

证. 由[7]知

$$\sup_{\|a\|=1}|\sqrt{n}\ (T_n^a-T^a)-\sqrt{n}\,(P_n-P)h(a,X)|=o_p(1),$$

注意到

$$\sqrt{n}(\sup_{\|a\|=1}T_n^a-\sup_{\|a\|=1}T^a)$$

$$-\sqrt{n}\,\{\sup_{\|a\|=1}V_n(a)-\sup_{\|a\|=1}V(a)\}+o_p(1),$$

其中 $V_n(a)\triangleq P_n(h(a,X)+T^a)$，$V(a)\triangleq T^a$，

$$S_n(a)\triangleq\sqrt{n}\,(V_n(a)-V(a))=\sqrt{n}\,(P_n-P)h(a,X),\ \|a\|=1.$$

由 Dudley-Philipp[1] 或 Pollard [6]表现定理知，存在过程 $\{\bar{S}_n(a),S(a):\|a\|=1\}$，满足

$\{\bar{S}_n(a):\|a\|=1\}$ 与 $\{S_n(a):\|a\|=1\}$同分布，$\{S(a):\|a\|=1\}=\{S(h(a)):\|a\|=1\}$，后者是 $\mathscr{F}$ 上 P-桥且

$$\sup_{\|a\|=1}|\bar{S}_n(a)-S(a)|\to 0\ \text{a.s.}P.$$

因为 $S(h(a))$ 的样本轨迹在 $\mathscr{F}$ 上一致连续，其中指标集 $\mathscr{F}$ 的本身度量为 $L_2(P)$ 半模，而由 $PM_0^2(X)<+\infty$ 知，当

$$\|a-b\|\to 0\ \text{时},\ P|h(a)-h(b)|^2\to 0,$$

注意到 $\{a\in\mathbf{R}^p:\|a\|=1\}$ 是闭集，所以

$S(a)$ 的样本轨迹在 $\{a\in\mathbf{R}^p:\|a\|=1\}$ 上关于欧氏模一致连续，且一致有界.

综上定理的条件满足，因此

$$\sqrt{n}(\sup_{\|a\|=1}T_n^a-\sup_{\|a\|=1}T^a)\rightsquigarrow\sup_{a\in B}W(h(a)).$$

推论 2. Friedman 判别标准的功效.

设 $L_j \triangleq L_j(a,Z) \triangleq Q_j(2\Phi(a^\tau Z)-1)$, $1 \leqslant j \leqslant J$.

Q_j 是[-1,1]上的勒让德多项式, $1 \leqslant j \leqslant J$.

$$\Phi \sim N(0, I_p),\quad Z \sim P.$$

一维投影指标:

$$I_p(a) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)P_R^2(Q_j(R)),\ R = (2\Phi(a^\tau Z)-1), P_R(Q_j(R)) \text{ 是 } Q_j(R) \text{ 的期望}$$

$$I_n(a) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)[P_n L_j(a,Z)]^2;$$

二维投影指标:

$$I_p(a,b) = \frac{1}{4}\sum_{j=1}^{J}(2j+1)E^2[Q_j(R_1)] + \frac{1}{4}\sum_{k=1}^{J}(2k+1)E^2[Q_k(R_2)]$$
$$+ \frac{1}{4}\sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)[E(L_j(a,Z)L_k(a,Z))]^2.$$

其中 $R_1 = 2\Phi(X_1)-1$, $X_1 = a^\tau Z$, $R_2 = 2\Phi(X_2)-1$, $X_2 = b^\tau Z$, $\|a\| = 1 = \|b\|$, $a^\tau b = 0$, E 表示取期望.

$$I_n(a,b) = \frac{1}{4}\sum_{j=1}^{J}(2j+1)[P_n L_j(a,Z)]^2 + \frac{1}{4}\sum_{k=1}^{J}(2k+1)[P_n L_k(b,Z)]^2$$
$$+ \frac{1}{4}\sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)[P_n L_j(a,Z)L_k(b,Z)]^2.$$

当 $P \neq \Phi$ 时,

$$\sqrt{n}\left(\sup_{\|a\|=1} I_n(a) - \sup_{\|a\|=1} I_p(a)\right) \sim \sup_{a\in B_1}\sum_{j=1}^{J}(2j+1)(E_p L_j(a,Z))\cdot PL_j(a,Z),$$

其中 $B_1 = \{b: I_p(b) = \sup_{\|a\|=1} I_p(a)\}$;

$$\sqrt{n}\left(\sup\{I_n(a,b): \|a\| = 1 - \|b\|, a^\tau b = 0\} - \sup\{I(a,b): \|a\| = 1 - \|b\|,\right.$$
$$\left. a^\tau b = 0\}\right) \rightsquigarrow 2\sup_{(a,b)\in B_{11}}\left\{\frac{1}{4}\sum_{j=1}^{J}(2j+1)(E_p L_j(a,Z))PL_j(a,Z)\right.$$
$$+ \frac{1}{4}\sum_{k=1}^{J}(2k+1)(E_p L_k(b,Z))PL_k(k,Z) + \frac{1}{4}\sum_{j+k<J}(2j+1)$$
$$\left.\cdot(2k+1)(E_p L_j(a,Z)L_k(b,Z))P(L_j(a,Z)L_k(b,Z))\right\}.$$

其中 $B_{11} = \{(a,b): a^\tau b = 0, \|a\| = 1 = \|b\|, I_p(a,b) = \sup\{I_p(c,d): c^\tau d = 0, \|c\| = \|d\| = 1\}\}$, E_p 是 $\mathscr{F} = \{L_j(a,Z), L_j(c,Z)L_k(d,Z): i+k \leqslant J, 1 \leqslant j \leqslant J, c^\tau d = 0, \|a\| = \|c\| = \|d\| = 1\}$ 上的 P-桥.

证. 取 $V_n(a) = I_n(a)$, $V(a) = I_p(a)$.

因为 $S_n(a) \triangleq \sqrt{n}(I_n(a) - I_p(a))$

$$-\sum_{j=1}^{J}(2j+1)[\sqrt{n}(P_n-P)L_j(a,Z)]PL_j(a,Z)+o_p(1).$$

由 Dudley-Philipp[3] 及 Pollard[4] 表现定理知，

存在 $\left\{\bar{S}_n(a),\sum_{j=1}^{J}(2j+1)(E_pL_j(a,Z))PL_j(a,Z):\|a\|=1\right\}$ 满足 $\bar{S}_n(a)$ 与 $S_n(a)-o_p(1)-\sum_{j=1}^{J}(2j+1)[\sqrt{n}(P_n-P)L_j(a,Z)]PL_j(a,Z)$ 同分布且

$$\sup_{\|a\|=1}\left|\bar{S}_n(a)-\sum_{j=1}^{J}(2j+1)E_pL_j(a,Z)PL_j(a,Z)\right|\to 0,\ \text{a.s.}P.$$

令

$$S(a)=\sum_{j=1}^{J}(2j+1)E_pL_j(a,Z)\cdot PL_j(a,Z),$$

由于 $\|a-b\|\to 0$ 时，$P|L_j(a,Z)-L_j(b,Z)|^2\to 0$，所以，$S(a)$ 是 $\{a\in R^p:\|a\|=1\}$上的一致连续函数.

综上定理条件满足，从而

$$\sqrt{n}\left(\sup_{\|a\|=1}I_n(a)-\sup_{\|a\|=1}I_p(a)\right)\rightsquigarrow\sup_{a\in B_1}\sum_{j=1}^{J}(2j+1)(E_pL_j(a,Z))PL_j(a,Z).$$

因为

$$\begin{aligned}&\sqrt{n}(I_n(a,b)-I(a,b))\\&\quad-\frac{1}{2}\sum_{j=1}^{J}(2j+1)[\sqrt{n}(P_n-P)L_j(a,Z)]PL_j(a,Z)\\&\quad+\frac{1}{2}\sum_{k=1}^{J}(2k+1)[\sqrt{n}(P_n-P)L_k(b,Z)]PL_k(b,Z)\\&\quad+\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J-i}(2j+1)(2k+1)\sqrt{n}(P_n-P)(L_j(a,Z)L_k(b,Z))\\&\quad\cdot P(L_j(a,Z)L_k(b,Z))+o_p(1).\end{aligned}$$

用类似的强逼近办法得到

$$\begin{aligned}&\sqrt{n}(\sup\{I_n(a,b):a^Tb=0,\|a\|=\|b\|=1\}-\sup\{I(a,b):a^Tb=0,\|a\|=\|b\|=1\})\\&\rightsquigarrow\sup_{a\in B_{11}}\left\{\frac{1}{2}\sum_{j=1}^{J}(2j+1)(E_pL_j(a,Z))PL_j(a,Z)\right.\\&+\frac{1}{2}\sum_{k=1}^{J}(2k+1)(E_pL_k(a,Z))PL_k(b,Z)\\&\left.+\frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{J-i}(2j+1)(2k+1)E_p(L_j(a,Z)L_k(b,Z))P(L_j(a,Z)L_k(b,Z))\right\}.\end{aligned}$$

推论 3. PP-Neyman 检验的功效

设 $H_0: G = P \longleftrightarrow H_1: G \neq P$, P 的 a 方向分布函数记为 F^a. G, P 连续分布.

$$K_n(a) = \sum_{i=1}^{J} [P_n \pi_i(F^a(a^\tau X))]^2$$

$$K_G(a) = \sum_{i=1}^{J} [G\pi_i(F^a a^\tau X))]^2,$$

$$\sqrt{n}\left(\sup_{\|a\|=1} K_n(a) - \sup_{\|a\|=1} K_G(a)\right)$$

$$\rightsquigarrow \sup_{a \in B} 2\left\{\sum_{i=1}^{J} E_G \pi_i(F^a(a^\tau X)) \cdot G\pi_i(F^a(a^\tau X))\right\}.$$

其中 $B = \{a \in \mathbf{R}^p: \|a\| = 1, K_G(a) = \sup_{\|b\|=1} K_G(b)\}$, E_G 是 $\{\pi_i(F^a(a^\tau X)): 1 \leqslant i \leqslant J, \|a\| = 1\}$ 上的 P-桥, 而 $\{\pi_i: 1 \leqslant i \leqslant J, \pi_0 \triangleq 1\}$ 是 $[0,1]$ 上的正交多项式.

证. 完全类似推论 2, 故略.

推论 4. PP-型 χ^2 拟合优度检验的功效.

假定 G, P 都是连续分布,

$$H_0: G = P \longleftrightarrow H_1: G \neq P, X_i \sim G, 1 \leqslant i \leqslant n.$$

取 $S_i \triangleq (t_i, t_{i+1}]$, $t_i \uparrow$, $t_{m+1} = +\infty$, $t_1 = -\infty$,

$$Z_n^a = \sum_{i=1}^{m} \frac{(P_n[a^\tau X \in S_i] - P[a^\tau X \in S_i])^2}{P[a^\tau X \in S_i]},$$

若 $\min_{1 \leqslant j \leqslant m} \inf_{\|a\|=1} P[a^\tau X \in S_i] > 0$, 则

$$\sqrt{n}\left(\sup_{\|a\|=1} Z_n^a - \sup_{\|a\|=1} Z^a\right) \rightsquigarrow 2 \sup_{a \in B} \sum_{i=1}^{m} \frac{E_G I_{[a^\tau X \in S_i]}}{P[a^\tau X \in S_i]} (G - P) I_{[a^\tau X \in S_i]},$$

其中 $B = \{a: Z^a = \sup_{\|b\|=1} Z^b, \|a\| = 1\}$, E_G 是 $\{I_{[a^\tau X \in S_i]}: 1 \leqslant i \leqslant m, \|a\| = 1\}$ 上的 G-桥.

证. 因为

$$S_n(a) \triangleq \sqrt{n}(Z_n^a - Z^a) = 2 \sum_{i=1}^{m} \frac{\sqrt{n}(P_n - G) I_{[a^\tau X \in S_i]}}{P(a^\tau X \in S_i)}$$
$$\cdot (G - P)[a^\tau X \in S_i] + o_p(1).$$

令

$$S(a) = 2 \sum_{i=1}^{m} \frac{E_G(I_{[a^\tau X \in S_i]})[(G - P) I_{[a^\tau X \in S_i]}]}{P[a^\tau X \in S_i]},$$

$E_G I_{[a^\tau X \in S_i]}$ 关于 $L_2(G)$ 半模在 $\{I_{[a^\tau X \in S_i]}: \|a\| = 1\}$ 上一致连续, 又因为

当 $\|a - b\| \to 0$ 时,

$$G(I_{[a^\tau X \in S_i]} - I_{[b^\tau X \in S_i]})^2 \to 0, \quad \inf_{\|a\|=1} P(a^\tau X \in S_i] > 0, \quad 1 \leqslant i \leqslant m.$$

从而 $S(a)$ 的样本轨迹 关于 a 在 $\{a \in \mathbf{R}^p: \|a\| = 1\}$ 上一致连续, 一致有界.

用 Dudley-Philipp[3] 及 Pollard[4] 的表现定理知, 定理的条件都成立, 从而即证.

推论 5. PP Kolmogrov-Smirnov 统计量的功效.

假定 G, P 是连续分布,

$$H_0: G = P \longleftrightarrow H_1: G \neq P,\ X_i \sim G,\ 1 \leqslant i \leqslant n.$$

$$\beta_n = \sup_{\|a\|=1} \sup_{|t|<+\infty} |(P_n - P) I_{[a^\tau x \leqslant t]}|$$

$$\beta = \sup_{\|a\|=1} \sup_{|t|<+\infty} |(G - P) I_{[a^\tau x \leqslant t]}|,$$

$$V_n(a,t) = |(P_n - P) I_{[a^\tau x \leqslant t]}|$$

$$V(a,t) = |(G - P) I_{[a^\tau x \leqslant t]}|,$$

显然 $V(a,t)$ 是 (a,t) 的连续函数.

假设 H_1 成立,则

$$\sqrt{n}(\beta_n - \beta) \rightsquigarrow \sup_{(a,t)\in B} E_G I_{[a^\tau x \leqslant t]} \operatorname{sgn}((G - P) I_{[a^\tau x \leqslant t]}),$$

其中

$$B = \{(a,t): a \in \mathbf{R}^p, \|a\| = 1, t \in \mathbf{R}^1, |(G - P) I_{[a^\tau x \leqslant t]}| = \sup_{\|b\|=1} \sup_{|s|<+\infty} |(G - P) I_{[b^\tau x \leqslant s]}|\},$$

E_G 是 $\{I_{[a^\tau x \leqslant t]}: \|a\| = 1, |t| < +\infty\}$ 上的 G-桥.

注记. PP Anderson-Darling 统计量:

$$\sup_{\|a\|=1} \sup_{|t|<+\infty} \frac{|(P_n - P) I_{[a^\tau x \leqslant t]}|}{[1 - F^a(t)] F^a(t)}$$ 也有类似的结论,

其中 F^a 是 P 在 a 方向上的分布函数,

证. 当 $(G - P) I_{[a^\tau x \leqslant t]} \neq 0$ 时,

$$S_n(a,t) \triangleq \sqrt{n}(V_n(a,t) - V(a,t))$$

$$= \sqrt{n}(P_n - G) I_{[a^\tau x \leqslant t]} \operatorname{sgn}((G - P) I_{[a^\tau x \leqslant t]} + o_p(1).$$

$o_p(1)$ 与 a 无关;

当 $(G - P) I_{[a^\tau x \leqslant t]} = 0$ 时,

$$S_n(a,t) \triangleq \sqrt{n}(V_n(a,t) - V(a,t))$$

$$= \sqrt{n} |(P_n - G) I_{[a^\tau x \leqslant t]}|$$

$$S(a,t) = \begin{cases} E_G(I_{[a^\tau x \leqslant t]}) \operatorname{sgn}((G - P) I_{[a^\tau x \leqslant t]}), & (G - P) I_{[a^\tau x \leqslant t]} \neq 0, \\ |E_G I_{[a^\tau x \leqslant t]}| \operatorname{sgn}((G - P) I_{[a^\tau x \leqslant t]}), & \text{否则}. \end{cases}$$

因为 H_1 成立,所以,

存在 $\delta > 0$, 使在 B 的 δ 闭邻域内 $(G - P) I_{[a^\tau x \leqslant t]} \neq 0$. 由于 $S(a,t)$ 在 $(G - P) I_{[a^\tau x \leqslant t]} \neq 0$ 上是 $G(a^\tau X \leqslant t)$ 的一致连续函数.

又因为 G, P 是连续分布,即 $G[a^\tau X \leqslant t]$, $P[a^\tau X \leqslant t]$ 是 (a,t) 的一致连续函数,从而 $S(a,t)$ 在 B 的 δ 闭邻域内关于 (a,t) 一致连续.

从而定理注 2)的条件满足,即证.

推论 6. PP-VonMise-Smirnov 统计量.

设 $H_0: G = P \longleftrightarrow H_1: G \neq P; G, P$ 连续;

$$w_n = \sup_{\|a\|=1} \int ((P_n - P) I_{[a^T x < t]})^2 dF^a(t)$$

$$= \sup_{\|a\|=1} \int ((P_n - P) I_{[a^T x < a^T Y]})^2 dF(Y);$$

F^a 是 P 在 a 方向上的边缘分布；

$$w = \sup_{\|a\|=1} \int ((G - P) I_{[a^T x < t]})^2 dF^a(t);$$

$$V_n(a) = \int ((P_n - P) I_{[a^T x < t]})^2 dF^a(t);$$

$$V(a) = \int ((G - P) I_{[a^T x < t]})^2 dF^a(t);\ \text{则}$$

$$\sqrt{n}(w_n - w) \rightsquigarrow 2 \sup_{a \in B} \int E_G(I_{[a^T x < t]}) \cdot (G - P) I_{[a^T x < t]} dF^a(t).$$

其中 $B = \{a: \|a\| = 1,\ V(a) = \sup_{\|b\|=1} V(b)\}$，$E_G$ 是 $\{I_{[a^T x < t]}: \|a\| = 1, |t| < +\infty\}$ 上的 G-桥.

证. $S_n(a) \triangleq \sqrt{n}(V_n(a) - V(a))$

$$= \int \sqrt{n}\ (P_n - G) I_{[a^T x < t]} (P_n + G - 2P) I_{[a^T x < t]} dF^a(t)$$

$$= 2 \int \sqrt{n}\ (P_n - G) I_{[a^T x < t]} (G - P) I_{[a^T x < t]} dF^a(t) + o_p(1).$$

取强逼近过程.

$$S(a) = 2 \int E_G I_{[a^T x < t]} \cdot (G - P) I_{[a^T x < t]} dF^a(t).$$

由 P, G 连续及 E_G 在 $\{I_{[a^T x < t]}: \|a\| = 1, |t| < +\infty\}$ 上关于 $L_2(G)$ 一致连续知，

$S(a)$ 关于 a 连续，从而 $S(a)$ 在 $\{a \in R^p: \|a\| = 1\}$ 上一致连续且一致有界. 由 Pollard[6] 表现定理知所需要的条件成立，即证.

推论 7. PP-L 统计量(作为位置和刻度的估计).

设 $X_i \sim G$, $1 \leqslant i \leqslant n$, G 连续分布函数，G_n 是 G 的经验分布函数，G^a 是 G 在 a 方向上的边缘分布函数，同理 G_n^a 是 G_n 在 a 方向上的边缘分布函数. h 是偏对称 (Skew-Symmetric)，有界连续(或对称有界连续).

$$T(G^a) = \int x h(G^a(x)) dG^a(x),$$

$$T(G_n^a) = \int x h(G_n^a(x)) dG_n^a(x).$$

设 G 满足[10]定理 3.2 的条件，则

$$T(G_n^a) - T(G^a) = -\int \frac{G_n^a(x) - G^a(x)}{g_a(x)} h(G^a(x)) dG^a(x) + O(n^{-\frac{3}{4}} \log n),$$

$g_a(t)$ 是 G^a 的密度 (t 处的).

令 $V(a) = T(G^a)$, $V_n(a) = T(G_n^a)$，有

$$\sqrt{n}(\sup_{\|a\|=1} T(G_n^a) - \sup_{\|a\|=1} T(G^a)) \rightsquigarrow \sup_{a \in B} \left[-\int \frac{E_G I_{[a^T x < t]}}{g_a(t)} h(G^a(t)) dG^a(t) \right],$$

其中 $B=\{a:\|a\|=1,\ T(G^a)=\sup_{\|b\|=1}T(G^b)\}$, E_G 是$\{I_{[a^TX<t]}:\|a\|=1,|t|<+\infty\}$上的 G-桥.

证. 类同推论 6,故略.

注. 作为检验统计量 $\sup_{\|a\|=1}|T(G_n^a)-T(G^a)|$ 或 $\sup_{\|a\|=1}\left|\frac{T(G_n^a)}{T(G^a)}-1\right|$, 其中 $H_0:P=G\longleftrightarrow H_1:P\neq G$, $X_i\sim P$, $1\leqslant i\leqslant n$, P, G 是位置或刻度参数族中的元素. 此时,用本文的方法,可得到这些统计量的功效.

推论 8. PP-Mann-Whitney 检验.

设 $G(X)=P(X-r)$, r 为未知 p 维参数向量, P 是连续分布.

$$H_0:r=0\longleftrightarrow H_1:r\neq 0.$$

取独立样本: $X_1,\cdots,X_n$ i. i. d $\sim P$,

$Y_1,\cdots,Y_n$ i. i. d $\sim G$.

$$U_n^a=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[I_{[a^TX_i<a^TY_j]}-\frac{1}{2}\right]$$

$$=(P\times G)_n I_{[a^TX<a^TY]}-\frac{1}{2},$$

其中 $(P\times G)_n=\frac{1}{n^2}\sum_{1\leqslant i,j\leqslant n}\delta_{(X_i,Y_j)}$, $\delta_{(X_i,Y_j)}$ 为单点测度.

$$V_n(a)=(P\times G)_n I_{[a^TX<a^TY]}-\frac{1}{2},$$

$$V(a)=(P\times G)I_{[a^TX<a^TY]}-\frac{1}{2}.$$

当 H_1 成立时,

$\sqrt{6n}\,(\sup_{\|a\|=1}V_n(a)-\sup_{\|a\|=1}V(a))\rightsquigarrow\sup_{a\in B}Z^a$, Z^a 是

$\mathscr{F}=\{I_{[a^TX<a^TY]}:\|a\|=1\}$ 上的 $P\times G$-桥,

$B=\{a:\|a\|=1,\ V(a)=\sup_{\|b\|=1}V(b)\}$.

注. 类似,对 PP-U 统计量也有类似的结论.

证. 类似前面的证法,故略.

推论 9. PP 两样本检验

设 F 为 k 维分布函数, $G(X)=F(X-r)$, r 为 p 维未知参数.

$H_0:r=0,\ \longleftrightarrow H_1:r\neq 0$.

设有样本 $X_1,\cdots,X_n$ i. i. d F

$$Y_1,\cdots,Y_m \text{ i. i. d } G,\ \frac{m}{n}=\lambda,\ 0<\lambda<+\infty,$$

$\mathrm{med}(a^TX)\triangleq a^TX_1,\cdots,a^TX_n$ 的中位数,

$\mathrm{mad}(a^TX)=|a^TX_1-\mathrm{med}\,a^TX|,\cdots,|a^TX_n-\mathrm{med}\,a^TX|$的中位数,

$$T_{n,m}^a=\frac{\mathrm{med}\,a^TX-\mathrm{med}\,a^TY}{\mathrm{mad}(a^T(X\cup Y))},$$

$$T_{s,m}=\sup_{\|a\|=1}T^a_{n,m}.$$

设 F^a 为 F 在 a 方向上边缘分布函数，密度为 f^a. $\bar{F}^a\triangleq\frac{1}{1+\lambda}F^a+\frac{\lambda}{1+\lambda}G^a$, $\bar{d}_a$ 为 $\bar{F}^a$ 的中位数.

$|a^\tau X-\bar{d}_a|$ 的分布记作 F_a，密度 f_a,

$|a^\tau Y-\bar{d}_a|$ 的分布记作 G_a，密度 g_a,

$\bar{F}_a\triangleq\frac{1}{1+\lambda}F_a+\frac{\lambda}{1+\lambda}G_a$, $\bar{d}_a$ 为 $\bar{F}_a$ 的中位数. 则

1) $$\sqrt{\frac{mn}{m+n}}T^a_{n,m}=\sqrt{\frac{mn}{m+n}}\left(\frac{G_m I_{[a^\tau Y<0]}}{f^a(-a^\tau r)}-\frac{F_n I_{[a^\tau X<0]}}{f^a(0)}\right)\frac{1}{\bar{d}_a}+o_p(1).$$

其中，G_m, F_n 分别是 G 和 F 的经验分布，$\bar{F}^a$, $\bar{F}_a$ 满足[6]定理 3.2 条件.

2) 令 $V(a)=\left(\frac{F^a(-a^\tau r)}{f^a(-a^\tau r)}-\frac{F^a(0)}{f^a(0)}\right)\frac{1}{\bar{d}_a}$, $\|a\|=1$；那么

$$\sqrt{\frac{mn}{m+n}}\left(\sup_{\|a\|=1}T^a_{n,m}-\sup_{\|a\|=1}V(a)\right)$$

$$\rightsquigarrow\sup_{a\in B}\left(\sqrt{\frac{1}{1+\lambda}}E_G\frac{I_{[a^\tau Y<0]}}{f(-a^\tau r)}-\sqrt{\frac{\lambda}{1+\lambda}}E_F\frac{I_{[a^\tau X<0]}}{f^a(0)}\right)\frac{1}{\bar{d}_a}.$$

其中，$B=\{a:\|a\|=1,\ V(a)=\sup_{\|b\|=1}V(b)\}$，$E_G$, E_F 分别是 $\{I_{[a^\tau Y<0]}:\|a\|=1\}$, $\{I_{[a^\tau X<0]}:\|a\|=1\}$ 上的 G-桥与 F-桥.

证. 由[6]定理 3.2 知，欲证(1)，剩下只须证：

i) $\sup_{\|a\|=1}|\mathrm{med}(a^\tau(X\cup Y))-\bar{d}_a|\xrightarrow[n\to+\infty]{}0$,

ii) $\mathrm{med}\,(|a^\tau X-\bar{d}_a|\cup|a^\tau Y-\bar{d}_a|)\to\bar{d}_a$, 对 $\|a\|=1$ 一致，只须证 ii)，(i) 类似)首先令

$$\mathscr{F}_0=\{I_{[|a^\tau x|<t]}-I_{[|a^\tau x|<s]}:\|a\|=1,\ s,\ t\in\mathbf{R}^1\}$$

$$\mathscr{F}_n=\{I_{[|a^\tau x-\bar{d}_a|<t]}-I_{[|a^\tau x-\bar{d}_a|<\xi_a]}:\|a\|=1,\ t\in I^a_n\}\leqslant\mathscr{F}_0.$$

$$I^a_n=\left(\xi_a-\frac{\ln n}{n^{\frac{1}{4}}},\ \xi_a+\frac{\ln n}{n^{\frac{1}{4}}}\right),\ \text{则有}$$

$$\sup_Q N_1(\varepsilon,Q,\mathscr{F}_n)\leqslant A_0\varepsilon^{-m_0},\ \forall n\geqslant 1.$$

$$\sup_{f\in\mathscr{F}_n}Pf^2=\sup_{\|a\|=1}\sup_{x\in I^a_n}|F_a(x)-F_a(\xi_a)|$$

$$=\sup_{\|a\|=1}\sup_{x\in I^a_n}|F^a(x)-F^a(\xi_a)+F^a(-\xi_a)-F^a(-x)|$$

$$\leqslant 2\sup_{\|a\|=1}\sup_{x\in I^a_n}[f^a(x)Vf^a(-x)]\frac{\log n}{n^{\frac{1}{4}}}$$

$$\leqslant\sup_{\|a\|=1}M(\xi_a)\cdot\frac{\log n}{n^{\frac{1}{4}}},\ f^a(x)Vf^a(-x)\triangleq\max\{f^a(x),f^a(-x)\}.$$

f 在 ξ_a, $-\xi_a$ 附近有界时，$M(\xi_a)$ 有界.

令 $\delta_n^2 = \sup_{\|a\|=1} M(\xi_n) \cdot \frac{\log n}{n}$, $\alpha_n^2 = \frac{1}{\sup_{\|a\|=1} M(\xi_n)\, n^{\frac{1}{2}}}$, 则由 [4] 中 p 34, 定理 37 知, 存在 $L>0$ 使

$$\sup_{g_n} |F_{n,n}f - F_n f| \leqslant L\delta_n^2 \alpha_n = L\left(\sup_{\|a\|=1} M(\xi_n)\right)^{\frac{1}{2}} \log n / n^{\frac{1}{2}}, \text{ a. s.}$$

也即 $\sup\{[F_{n,n}(x) - F_{n,n}(\xi_n)] - [F_n(x) - F_n(\xi_n)]| : \|a\| = 1,\ x \in I_n^a\}$

$$= O(n^{-\frac{1}{2}} \log n) \text{ a. s.}$$

同理 $\sup\{|[G_{\lambda,m}(x) - G_{\lambda,m}(\xi_n)] - [G_n(x) - G_n(\xi_n)]| : \|a\| = 1,\ x \in Z_m^a\}$

$$= O(m^{-\frac{1}{2}} \log m) \text{ a. s.}$$

现设 $\bar{F}_{n,n,m}$ 为 $|a^\tau X_1 - \hat{d}_n|, \cdots, |a^\tau X_n - \hat{d}_n|$,

$|a^\tau Y_1 - \hat{d}_n|, \cdots, |a^\tau Y_m - \hat{d}_n|$ 的经验分布, 则

$$\bar{F}_{n,n,m} = \frac{1}{n+m}\left(\sum_{i=1}^{n} \delta_{[|a^\tau X_i - \hat{d}_n|]} + \sum_{j=1}^{m} \delta_{[|a^\tau Y_j - \hat{d}_n|]}\right)$$

$$= \frac{1}{1+\lambda} F_{n,n} + \frac{\lambda}{1+\lambda} G_{n,m} \to \bar{F}_n \text{ a. s. 对 } \|a\| = 1 \text{ 一致.}$$

由此, $\sup_{\|a\|=1} \sup_{x \in I_n^a} \left\{\left| \frac{1}{1+\lambda}(F_{n,n}(x) - F_{n,n}(\xi_n) - F_n(x) + F_n(\xi_n)) \right.\right.$

$$\left.\left. + \frac{\lambda}{1+\lambda}(G_{\lambda,m}(x) - G_{n,m}(\xi_n) - G_n(x) + G_n(\xi_n)) \right|\right\}$$

$$= \sup_{\|a\|=1} \sup_{x \in I_n^a} \{[\bar{F}_{n,n,m}(x) - \bar{F}_{n,n,m}(\xi_n)] - [\bar{F}_n(x) - \bar{F}_n(\xi_n)]\}$$

$$= O(n^{-\frac{1}{2}} \log n). \tag{3}$$

令 $k_n = (n+m)p + o(n^{\frac{1}{2}} \log n)$. 下证

若设 $V_{n,m}^a$ 为 $|a^\tau X_1 - \hat{d}_n|, \cdots, |a^\tau X_n - \hat{d}_n|$;

$|a^\tau Y_1 - \hat{d}_n|, \cdots, |a^\tau Y_m - \hat{d}_n|$ 的第 k_n 秩次统计量. 则 n 充分大时,

$$\forall \|a\| = 1,\ V_{n,m}^a \in I_n^a.$$

实际上, 令 $a_n = \frac{\ln n}{\sqrt{n}}$, ξ_n 是 $\bar{F}_n$ p 分位点.

$$\{\inf_a (V_{n,m}^a - \xi_n + a_n) \leqslant 0\} = \left\{\sup_a \bar{F}_{n,n,m}(\xi_n - a_n) \geqslant \frac{k_n}{n+m}\right\}$$

$$\leqslant \{\sup_{\|a\|=1} \{|\bar{F}_{n,n,m}(\xi_n - a_n) - \bar{F}_n(\xi_n - a_n)|\} \geqslant t_n\}$$

$$\leqslant \{\sup_{\|a\|=1} \sup_{\tau \in I_n^a} \{|\bar{F}_{n,n,m}(x) - \bar{F}_{n,n,m}(\xi_n) - \bar{F}_n(x) + \bar{F}_n(\xi_n)|\} \geqslant t_n\}$$

$$t_n = \inf_{\|a\|=1}\left(\frac{k_n}{n+m} - \bar{F}_n(\xi_n - a_n)\right) \geqslant \frac{1}{2} M \log n / n^{\frac{1}{2}},\ M \text{是常数.}$$

由(3)式知, $\inf_{\|a\|=1} (V_{n,m}^a - \xi_n + a_n) > 0$ a. s., n 充分大.

同理 $\sup_{\|a\|=1} (V_{n,m}^a - \xi_n + a_n) < 0$, a. s, n 充分大.

最后再证

$$V^a_{n,m} = \xi_a + [k_a - (n+m)\bar{F}_{a,n,m}(\xi_a)]/(n+m)\bar{f}_a(\xi_a) + O(n^{-\frac{1}{2}}\log n), \tag{4}$$

其中，$$\bar{f}_a(\xi_a) = \frac{1}{1+\lambda} f_a(\xi_a) + \frac{\lambda}{1+\lambda} g_a(\xi_a).$$

实际上，由 $\bar{F}_a(V^a_{n,m}) - \bar{F}_a(\xi_a) = \bar{f}_a(\xi_a)(V^a_{n,m} - \xi_a) + \bar{F}'_a(\theta^a_n)(V^a_{n,m} - \xi_a)^2$ 及 (3) 式和 $V^a_{n,m} \in I^a_n$，$\forall \|a\| = 1$，n 充分大，马上得到(4)式．特别有 ii) 成立．

推论 9 的第二部分的证明．与前面方法类同，故略．

推论 10. 两样本检验．

$H_0: F = G$，$H_1: F \neq G$，F，G 连续，$\frac{n_1}{n_1+n_2} = \lambda$，$0 < \lambda < 1$．

$$V_{n_1,n_2}(a) = \int (F^a_{n_1}(t) - G^a_{n_2}(t))^2 dF^a(t),$$

$$V(a) = \int (F^a(t) - G^a(t))^2 dF^a(t), \text{ 则}$$

当 H_1 成立时，

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\left[\sup_{\|a\|=1} V_{n_1,n_2}(a) - \sup_{\|a\|=1} V(a)\right]$$
$$\sim 2 \sup_{x \in B}\left[\sqrt{1-\lambda}(E_F I_{[a^T x < t]}) - \sqrt{\lambda}(E_G I_{[a^T Y < t]})\right]$$
$$\cdot (F^a(t) - G^a(t)) dF^a(t)$$

其中 E_F，E_G 分别是 $\{I_{[a^T x < t]}: \|a\| = 1\}$，$\{I_{[a^T Y < t]}: \|a\| = 1\}$ 上的 F-桥和 G-桥．

注记：1)正态性检验

$$\delta_n = \sup_{a,t} \left| F^a_n(t) - \Phi\left(\frac{t - a^T \mu_n}{\sqrt{a^T V_n a}}\right)\right|,$$

其中 μ_n 是均值 μ 的样本均值估计，V_n 是协方差 V 的样本协方差估计，Φ 是标准正态分布函数，$F^a_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{[a^T x_i < t]}$．$\delta_n$ 的功效也有与前面类同的结论，这里不再赘述．

2) M-型 PP 位置检验．

T^a_n 是 $\int \phi(a^T X, t) dP_n = 0$ 的最小解，

$T^a \triangleq T^a(P)$ 是 $\int \phi(a^T X, t) dP = 0$ 的最小解；

如果 P_n 是 G 的经验分布，$G \neq P$，$\sup_{\|a\|=1} |T^a(G) - T^a(P)| > 0$．则有：

$$\sqrt{n}\left\{\sup_{\|a\|=1} |T^a_n - T^a(P)| - \sup_{\|a\|=1} |T^a(G) - T^a(P)|\right.$$
$$\sim \sup_{a \in B} E_G h(a) \cdot \operatorname{sgn}(T^a(G) - T^a(p)),$$

其中　$B = \{a: \|a\| = 1, \sup_{\|b\|=} |T^b(G) - T^b(P)| = |T^a(G) - T^a(P)|\}$，

而 $h(a) \triangleq \dfrac{\phi(a^TX, T^*(G))}{-G\dfrac{\partial\phi(a^TX, T^*(G))}{\partial s_2}}$，$E_G$ 是 $\{h(a):\|a\|=1\}$ 上的 G-桥

证. 本推论的证明也类同前面的方法，故略.

参 考 文 献

[1] Dudley, R. M., Philipp, W., Invariance principles for sums of Banach space valued random elements and Empirical processes, *Z. W. verw. Geb.*, 62 (1983), 509—552.

[2] Gideon, R. A., Rentice, M. J., Pyke, R. (1989), The Limiting distribution of the rank correlation coefficient Rg. (to appear in a Festschrift for Ingram Olkin, Springer).

[3] Li, G., Chen, Z., Projection-Pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, *JASA*, 80(1985), 759—766.

[4] Pollard, D., Convergence of Stochatic Processes, Springer-Verlag, New York, 1984.

[5] 张 岚，PP 型拟合优度检验，系统科学与数学，8:3(1988)，234—242.

[6] 张 岚，几类 PP 型检验统计量的性质，应用数学学报，12:1(1989)，82—95.

[7] 张 健，朱力行，成平 M-型 PP 指标其特征值和特征向量的渐近理论及其应用(I)(II)，投《中国科学》，1988.

ASYMPTOTIC POWERS OF SOME PP TESTS

ZHANG JIAN CHENG PING

(Institute of Systems Science, Academia Sinica)

ABSTRACT

A general method of dealing with the asymptotic powers of the PP tests is given and by this method asymptotic powers of a number of useful PP tests and estimators are obtained. In particular, the asymptotic distribution of the largest eigenvalue of M-type PP index is derived.