Chin. Ann. of Math.
12B:4(1991), 445—453.

THE PROBABILITY INEQUALITIES FOR THE SUPREME OF THE EMPIRICAL PROCESSES AND APPLICATIONS TO THE TESTING OF DATA STRUCTURE

ZHU LIXING (朱力行)*

Abstract

This paper extends Huber's results and achieves the probability inequalities on the distribution of the Kolmogorov distance in the worst direction for some distribution class, including the elliptically symmetric distrtbution, uniform distribution on square. Combining these results with öhrvik's results the author argues the testing problem of data structure obtained by projection pursuit technique.

§ 1. Introduction

Suppose that P is a nonatomic, d-dimensional distribution ($d>1$), and $\underset{\sim}{X}_1\cdots$, $\underset{\sim}{X}_n$ are an i. i. d. sample from P. Let P_n be the empirical measure determined by this sample. Moreover, let $S_d=\{\boldsymbol{a}:\boldsymbol{a}\in R^d,\ ||\boldsymbol{a}||=1\}$ for each $\boldsymbol{a}\in S_d$ and $t\in R^1$. Write $H(\boldsymbol{a}, t)$ for $\{\boldsymbol{a}^T\underset{\sim}{X}<t\}$, the open half space in R^d.

Define

$$D_n(H(\boldsymbol{a}, t))=|P_n(H(\boldsymbol{a}, t))-P(H(\boldsymbol{a}, t))|, \tag{1.1}$$

$$D_n=\sup_{\boldsymbol{a}, t} D_n(H(\boldsymbol{a}, t))). \tag{1.2}$$

D_n, clearly, is the maximum Kolmogorov distance between P_n and P all possible one-dimensional projections. And we easily see that

$$D_n=\sup_{\boldsymbol{a}\in R^d}\sup_{t\in R^1} D_n(H(\boldsymbol{a}, t)), \tag{1.3}$$

that is, in definition of D, the constraint of $\boldsymbol{a}\in S_d$ is not necessary.

Let

$$\eta(n, d, \varepsilon)=P(D_n>\varepsilon). \tag{1.4}$$

For $d=1$, we have that well-known inequality $\eta(n, 1, \varepsilon)<2\exp(-2n\varepsilon^2)$. When $d>1$, as a special case of Devroye[1], who refined an earlier result of Vapnik and Červonenkis[2],

$$\eta(n, d, \varepsilon)\leqslant 4e^8(n^2/d)^d\exp(-2n\varepsilon^2). \tag{1.5}$$

Manuscript received April 6, 1989.

* Instituie of Systems Science, Academia Sinica, Beijing, China.

Huber[3], under the condition of the spherically symmetric distribution, obtained a more refined result

$$\eta(n,\ d,\ \varepsilon)\leqslant 2(en/d)^d\exp(-2n(\varepsilon-d/n)^2) \tag{1.6}$$

for $\varepsilon > d/n$.

In this paper, we achieve the similar fashion of (1.6) for a more general distribution class including the elliptically symmetric distribution, the uniform distribution on square. Furthermore, combining the simulated results of Ohrvik, we argue the testing problem for the data structure obtained by projection pursuit technique. The general inequalities is put in section 2. Section 3 contains two special cas es of the theorems in Section 2. Finally, we discuss the testing problem in Section 4.

§ 2. Probability Inequalities

For convenience, we first put some notations. For any fixed r points $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r$ in R^d, let $H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}, t)=\{\boldsymbol{a}^T\underset{\sim}{X}<t\}$, where $\boldsymbol{a}$ and t satisfy the constraint condition: $||\boldsymbol{a}||=1$, $t\in R^1$, and $\boldsymbol{a}^T\boldsymbol{h}_i=t$, $1<i<r$, namely, the boundary of $H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}, t)$ contains $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r$. And let the class of the set

$$\mathscr{F}(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r)=\{H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}, t):\ \boldsymbol{a}\in S_d,\ t\in R^1,\ \boldsymbol{a}^T\boldsymbol{h}_i=t,\ i=1, \cdots, r\}.$$

Theorem 1. *Suppose that any one-dimensional marginal distribution of P at direction $\boldsymbol{a}$, $P\{H(\boldsymbol{a}, t)\}$, continues and increases strictly with respect to t, and the following conditions are fulfilled:*

(i) *there exists at least $t\neq 0$ and an $\boldsymbol{a}\in S_d$ for which $\boldsymbol{a}^T\boldsymbol{h}_i=t$, $i=1, \cdots, r$;*

(ii) *there exists unique $H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}_0, t_0) \in\mathscr{F}(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r)$ such that*

$$P\{H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}_0, t_0)\}=\sup_{\mathscr{F}(h_1,\cdots,h_r)}\{H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}, t)\}, \tag{2.1}$$

and $H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{h}_0, t_0)$ is the unique extremum point for functional $P(H(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r, \boldsymbol{a}, t)$;

(iii) *for every $\boldsymbol{a}\in S_d$*

$$PH(\boldsymbol{a}, 0)\}=1/2.$$

Then for each $\varepsilon>d/n$

$$\eta(n,\ d,\ \varepsilon)\leqslant 4(en/d)^d(2\pi d)^{-1/2}\exp(-2n(\varepsilon-d/n)^2)). \tag{2.2}$$

Theorem 2. *If there exists a function $f(\boldsymbol{a})\neq 0$ for which $P\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}\leqslant t)$ depends only on t, and continues and increases strictly with respect to t, then (2.2) also holds.*

Proof of Pheorem 1 For the fixed sample $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$, we shall show that $\boldsymbol{D}_n$ is the maximum value over $\boldsymbol{D}_n$ $(H(\boldsymbol{a}_i, t_i)$, $i=1, \cdots, \sum_{r=1}^{d}\binom{n}{r})$.

It is easy to see that for some $\boldsymbol{a} \in S_d$, $t \in R^1$

$$D_n = P_n(H(\boldsymbol{a}_0, t_0) - P(H(\boldsymbol{a}_0, t_0) \tag{2.3}$$

or

$$D_n = P(H(\boldsymbol{a}_0, t_0) - P_n(H(\boldsymbol{a}_0, t_0). \tag{2.4}$$

We first do some preliminary work to simplify the proof. In the following we assume that the boundary of $H(\boldsymbol{a}_0, t_0)$, $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t_0\}$ contains at least one sample point and $H(\boldsymbol{a}_0, t_0)$ contains the origin point. Here we give some illustrations for the reason of the assumptions. Since $P(H(\boldsymbol{a}_0, t)$ is continuous and strictly increasing with respect to t, and $P(H(\boldsymbol{a}_0, t)$ takes constant values piecewise, if $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t\}$ would not contains any sample point $\underset{\sim}{X}_i$ $(1 \leqslant i \leqslant n)$, then the euclidean distances between $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t_0\}$ and $\underset{\sim}{X}_i$'s would be larger than zero due to the compactness of $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t_0\}$. In this case we can find $t_1 < t_0 < t_2$ for which

$$P\{H(\boldsymbol{a}_0, t_1)\} = P_n\{H(\boldsymbol{a}_0, t_0)\} = P_n\{H(\boldsymbol{a}_0, t_2)\}$$

and

$$P\{H(\boldsymbol{a}_0, t_1)\} < P\{H(\boldsymbol{a}_0, t_0)\} < P\{(\boldsymbol{a}_0, t_2)\}. \tag{2.5}$$

This is a contradiction with (2.3) or (2.4). On the other hand, since for any fixed open half space H, and the complement, H' say, the continuity of P implies $D_n(H) = D_n(H')$ in distribution, we can only investigate the case that $H(\boldsymbol{a}, t)$ contains the origin point.

Furthermore, by the continuity of $P\{H(\boldsymbol{a}, t)$ with respect to t for any $\boldsymbol{a} \in S_d$, it is easy to see that the probability value of two sample points being the same is zero. Hence in probability one any d sample point determines uniquely a superplane in R^d space. Meanwhile, we have in probability one $t > 0$ due to $P\{\boldsymbol{a}^T \quad = 0\} = 0$.

If $r(>d)$ sample points are contained in $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t_0\}$, then $\{\boldsymbol{a}_0^T \underset{\sim}{X} = t_0\}$ can be determined uniquely by any d sample points among r sample points; on the other hand, when $r < d$, we can show that the open half space $H(\boldsymbol{a}_0, t_0)$ satisfies the condition (2.1) in Theorem 1. Indeed, if (2.1) were not satisfied, for the same reason as above by the compacthess of the superplane and the condition (ii), there would exist other two open half spaces $H(\boldsymbol{a}_1, t_1)$ and $H(\boldsymbol{a}_2, t_2)$ for which

$$P_n\{H(\boldsymbol{a}_1, t_1)\} = P_n\{H(\boldsymbol{a}_0, t_0)\} = P_n\{H(\boldsymbol{a}_2, t_2)\}$$

and

$$P\{H(\boldsymbol{a}_1, t_1)\} = P\{H(\boldsymbol{a}_0, t_0)\} = P\{H(\boldsymbol{a}_2, t_2)\},$$

which raises a contradiction with (2.3) and (2.4). Hence $H(\boldsymbol{a}_0, t_0)$ must satisfy (2.1). Furthermore, in the case of $r < d$, with the different combinations $\{\underset{\sim}{X}_{i_1}, \cdots, \underset{\sim}{X}_{i_r}\}$, the number of H satisfying (2.1) is at most equal to $\binom{n}{r}$. Then with varying r $(<d)$, we see that $H(\boldsymbol{a}_0, t_0)$ must be one of $N\left(= \sum_{r=1}^{d} \binom{n}{r}\right)$ half spaces H.

We now investigate the probability value of $H(a, t)$ whose boundary $\{\boldsymbol{a}^T\underset{\sim}{X}=t\}$ contains r sample points $\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr}$. Denote by $D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})$ the conditional distribution given these r sample points, it is easy to see that $D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})$ can be expressed by a random variable Y with binomial distribution $B(n-r, p)$, where $p=P(H(\boldsymbol{a}, t))$. If (2.3) holds, then

$$D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})=Y/n+r/(n-p)=[Y-(n-r)]/n+r(1-p)/n;$$

if (2.4) holds, then

$$D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})=p-Y/n=rp/n-[Y-(n-r)p]/n.$$

We have in this two cases

$$D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})\leqslant|[Y-(n-r)p]/n|+r/n. \tag{2.6}$$

By Hoeffding inequality, for each $r\leqslant d$

$$\begin{aligned}P\{D(_nH|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})>\varepsilon+d/n\}&\leqslant P\{D_n(H|\underset{\sim}{X}_{j1}, \cdots, \underset{\sim}{X}_{jr})>\varepsilon+r/n\}\\ &\leqslant 2\exp(-2\varepsilon^2n^2/(n-r))\\ &\leqslant 2\exp(-2n\varepsilon^2).\end{aligned} \tag{2.7}$$

Then

$$P\{D_n(H)>\varepsilon+d/n\}\leqslant 2\exp(-2n\varepsilon^2). \tag{2.8}$$

Therefore, by [2] we have $\sum_{r=1}^{d}\binom{n}{r}\leqslant 2(en/d)^d(2\pi d)^{-1/2}$, completing the proof.

Proof of Theorem 2 By (1.3) and $D_n(H)=D_n(H')$,

$$\begin{aligned}D_n&=\sup_{\boldsymbol{a}\in S_d}\sup_{t\in R^1}|P_n(H(\boldsymbol{a}, t))-P(H(\boldsymbol{a}, t))|\\ &=\sup_{\boldsymbol{a}\in S_d}\sup_{t\in R^1}|P_n(H(f(\boldsymbol{a})\boldsymbol{a}, t))-P(H(f(\boldsymbol{a})\boldsymbol{a}, t))|\\ &\leqslant\sup_{\boldsymbol{a}\in R^d}\sup_{t\in R^1}|P_n(H(\boldsymbol{a}, t))-P(H(\boldsymbol{a}, t))|=D_n.\end{aligned}$$

The second equation invokes $D_n(H)=D_n(H')$ for the case of $f(\boldsymbol{a})<0$. Let $\mathscr{F}'(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r)=\{H(f(\boldsymbol{a})\boldsymbol{a}, t): \boldsymbol{a}\in S_d, t\in R^1, f(\boldsymbol{a})\ \boldsymbol{a}^T\boldsymbol{h}_i=t, i=1, \cdots, r\}$. For any $H(f(\boldsymbol{a})\boldsymbol{a}, t)\in\mathscr{F}'(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r)$, define $\boldsymbol{h}$ as the intersection point between $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$ and its vertical vector crossing the origin point $\underset{\sim}{O}$, and let $\underset{\sim}{M}$ be the affine manifold spanned by the r points $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_r$. Q is defined as the point on $\underset{\sim}{M}$ closest to the origin $\underset{\sim}{O}$ in the euclidean nor m. Since $\underset{\sim}{O}\boldsymbol{h}$ is perpendicular to $H(f(\boldsymbol{a})\boldsymbol{a}, t)$ and $\underset{\sim}{M}$ lies in $H(f(\boldsymbol{a})\boldsymbol{a}, t)$, $O\boldsymbol{h}$ and $Q\boldsymbol{h}$ are the orthogonal vectors. Hence

$$||\underset{\sim}{O}\boldsymbol{h}||=||\underset{\sim}{OQ}||-||\underset{\sim}{Q}\boldsymbol{h}||. \tag{2.9}$$

Moreover, $||\underset{\sim}{Q}\boldsymbol{h}||$ is the strictly increasing function with respect to the angle $\theta(0\leqslant\theta\leqslant\pi/2)$ between $\underset{\sim}{OQ}$ and $\underset{\sim}{O}\boldsymbol{h}$. $||\underset{\sim}{Q}\boldsymbol{h}||=0$, for $\theta=0$; $=||\underset{\sim}{OQ}||$, for $\theta=\pi/2$, and that $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$ perpendicular to $\underset{\sim}{OQ}$ is unique.

On the other hand, the condition of $P\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}<t\}$ not depending on $\boldsymbol{a}$ implies that $P\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}<t\}$ is nothing but the strictly increasing function with respect to the distance between the origin point $\underset{\sim}{O}$ and $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$. Therefore, the angle between $\underset{\sim}{OQ}$ and $\underset{\sim}{O}\boldsymbol{h}$ equals that between the superplanes perpendicular to,

respectively, $\underset{\sim}{OQ}$ and $O\boldsymbol{k}$. Consequently,

$$\cos\theta = |f(\boldsymbol{a}_0)f(\boldsymbol{a})\boldsymbol{a}_0\boldsymbol{a}|/|f(\boldsymbol{a}_0)f(\boldsymbol{a})|\cdot||\boldsymbol{a}_0||\cdot||\boldsymbol{a}|| = |\boldsymbol{a}_0^T\boldsymbol{a}|, \qquad (2.10)$$

where $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$ is the superplane perpendicular to $\underset{\sim}{OQ}$. Since each $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$ considered contains $\boldsymbol{k}_1, \cdots, \boldsymbol{k}_r$, the direction corresponding to $\{f(\boldsymbol{a})\boldsymbol{a}^T\underset{\sim}{X}=t\}$ possesses $d-r$ freedom variables. Write $(x_1, \cdots, x_d)$ for $\underset{\sim}{X}^T$. We assume without loss of generality that $\{f(\boldsymbol{a}_0)\boldsymbol{a}_0^T\underset{\sim}{X}=t\}$ is parallel to $\{x_1=0\}$, that is, $\boldsymbol{a}=(1, 0, \cdots, 0)$, $\{f(\boldsymbol{a}_0)\boldsymbol{a}_0^T\underset{\sim}{X}=t\}=\{x_1=||\underset{\sim}{OQ}||\}$. (Otherwise, we can rotate the coordinate system.) Therefore, the $\boldsymbol{a}$ can be expressed as $\left(\cos\theta_1, \sin\theta_1\cos\theta_2, \cdots, \prod_{j=1}^{d-r}\sin\theta_j, \theta_j\cdots, 0\right)$, where $-\pi/2\leqslant\theta_1\leqslant\pi/2$, $0\leqslant\theta_j\leqslant 2\pi$, $j=2, \cdots, d-r$. So

$$\cos\theta = |\cos\theta_1|. \qquad (2.11)$$

Then for each $H(f(\boldsymbol{a})\boldsymbol{a}, t)\in\mathscr{F}'(\boldsymbol{k}_1, \cdots, \boldsymbol{k}_r)$ such that $\theta>0$, we can find $0<\theta'<\theta<\theta''<\pi/2$ by changing θ_1. Consequently, it is easy to see that (ii) in Theorem 1 is fulfilled. The proof is completed.

§ 3. Two Examples

We here check two usual distributions satisfying the conditions of the theorems: the elliptically symmetric distribution and the uniform distribution on square.

Example 1. Suppose that P is a nonatomic, elliptically symmetric. We know that P is said to be an elliptically symmetric if there exists a $d\times 1$ vector $\underset{\sim}{\mu}$, a d-order nonnegative definite matrix $\underset{\sim}{\Sigma}$ and a real function Φ for which the characteristic function of P is of the form

$$\exp(i\underset{\sim}{t}^T\underset{\sim}{\mu})\cdot\Phi(\underset{\sim}{t}^T\underset{\sim}{\Sigma}\underset{\sim}{t}). \qquad (3.1)$$

In particular, if $\underset{\sim}{\mu}=0$ and $\underset{\sim}{\Sigma}=I^d$, we say that P is spherically symmetric. Without loss of generality, assume that $\underset{\sim}{\mu}=0$. By the nonatomicity of P we know that $\underset{\sim}{\Sigma}\triangleq A\,A^T$ is positive definite. So there exists a spherically symmetric random vector $\underset{\sim}{Y}$ such that $\underset{\sim}{X}\overset{d}{=}A\underset{\sim}{Y}$, where the notation "$\overset{d}{=}$" means that two sides have the same distribution. Moreover, by the spherical symmetry of $\underset{\sim}{Y}$, the distribution of $\boldsymbol{a}^T\underset{\sim}{Y}$ is the same for each $\boldsymbol{a}\in S_q$, so we have

$$\boldsymbol{a}^T\underset{\sim}{X}=\boldsymbol{a}^TA\underset{\sim}{Y}=||\boldsymbol{a}^TA||\,\boldsymbol{a}^TA\underset{\sim}{Y}/||\boldsymbol{a}^TA||=||\boldsymbol{a}^TA||\,Y_1, \qquad (3.2)$$

where Y_1 is the first component of $\underset{\sim}{Y}$. To check the conditions in Theorem 1, we first show the following facts:

1) For $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$ i. i. d. with P there exist $\underset{\sim}{Y}_1, \cdots, \underset{\sim}{Y}_n$ i. i. d. with the same distribution as that of $\underset{\sim}{Y}$ such that

$$\{\boldsymbol{a}^T\underset{\sim}{X}_1, \cdots, \boldsymbol{a}^T\underset{\sim}{X}_n): \boldsymbol{a}\in S_d\}\overset{d}{=}\{\boldsymbol{a}^TA\underset{\sim}{Y}_1, \cdots, \boldsymbol{a}^TA\underset{\sim}{Y}_n): \boldsymbol{a}\in S_d\}. \qquad (3.3)$$

To show (3.3), it is enough to show that any finite dimensional distributions in two sides are the same, that is, for any $a_1, \cdots, a_m \in S_d$

$$(a_1^T \underset{\sim}{X}_1, \cdots, a_1^T \underset{\sim}{X}_n, \cdots, a_m^T \underset{\sim}{X}_1, \cdots, a_m^T \underset{\sim}{X}_n) \overset{d}{=} (a_1^T A \underset{\sim}{Y}_1, \cdots, a_1^T A \underset{\sim}{Y}_n, \cdots, a_m^T A \underset{\sim}{Y}_1, \cdots, a_m^T A \underset{\sim}{Y}_n). \tag{3.4}$$

Since $\underset{\sim}{X} = A\underset{\sim}{Y}$, for any Borel set B in R^d, we have

$$P\{\underset{\sim}{X} \in B\} = P\{A\underset{\sim}{Y} \in B\}. \tag{3.5}$$

On the other hand

$$\begin{aligned} P\{a_1^T \underset{\sim}{X}_1 \leqslant t_1, \cdots, a_m^T \underset{\sim}{X}_1 \leqslant t_m\} &= P\{\cap \{a_i^T \underset{\sim}{X}_1 \leqslant t_i\}\} \\ &= P\{\underset{\sim}{X}_1 \in \cap H(a_i, t_i)\} = P\{A\underset{\sim}{Y}_1 \in H(a_i, t_i)\} \\ &= P\{\cap \{a_i^T A \underset{\sim}{Y}_1 \leqslant t_i\}\} = P\{a_1^T A \underset{\sim}{Y}_1 \leqslant t_1, \cdots, a_m^T A \underset{\sim}{Y}_1 \leqslant t_m\}, \end{aligned}$$

from which together with independence we derive (3.4).

2) $P\{a^T A \underset{\sim}{Y} / \| |a^T A| \leqslant t\} = P\{Y_1 \leqslant t\}$ depends only on t and is continuous and strictly monotone with respect to t. This is an easy assertion since P has no any atomic and Σ is a positvie definite matrix.

By (3.3) and 2), we have

$$\begin{aligned} D &= \sup_{a \in S_d} \sup_{t \in R^1} |P_n H(a, t) - PH(a, t)| \\ &= \sup_{a \in S_d} \sup_{t \in R^1} \left| \frac{1}{n} \sum_{j=1}^n I[\|a^T A\|^{-2} a^T A \underset{\sim}{Y}_j \leqslant \|a^T A\|^{-1} t] - P\{\|a^T\|^{-1} a^T A \underset{\sim}{Y} \leqslant \|a^T A\|^{-1} t\} \right| \\ &= \sup_{a \in S_d} \sup_{t \in R^1} \left| \frac{1}{n} \sum_{j=1}^n I[\|a^T A\|^{-1} a^T A \underset{\sim}{Y}_j \leqslant t] - P\{a^T A\|^{-1} a^T A \underset{\sim}{Y} \leqslant t\} \right| \end{aligned}$$

which together with Theorem 2 concludes the proof.

Example 2. The uniform distribution on square.

Assume, without loss of generality, that the distribution is on $[-1, 1]^2$. Now we check (i) and (ii) in Theorem 1. Consider the area of intersect set of $[-1, 1]^2$ and $H(a, t)$ $(t>0)$ whose boundary crosses some point in $[-1, 1]^2$. Write the crossed point as $(1-c_2, 1-c_1)$, $0<c_1<c_2<1$. Calculate the area of $[-1, 1]^2 \cap \{a^T \underset{\sim}{X} \geqslant t\}$, where the notation "$\cap$" means the intersection of two sets. We first rotate the straight line $a^T \underset{\sim}{X} = t$ counter clockwise. Divide the investigation into five cases:

1) The area S has the form $S = cb/2$, $0<c<2$, $0<b<2$, where c is the distance between the point (1, 1) and the intersect point of $\underset{\sim}{Y} = 1$ and $a^T \underset{\sim}{X} = t$, and b stands for that between the point (1, 1) and the intersect point of $\underset{\sim}{X} = 1$ and $a^T \underset{\sim}{X} = t$. It is easy to see that $c = c_2 + c_1/\operatorname{tg}\theta$, $b = c_1 + c_2 \operatorname{tg}\theta$, where θ is the angle between $a^T \underset{\sim}{X} = t$ and $\underset{\sim}{Y} = 1$. So

$$S_\theta = cb/2 = (2c_1c_2 + c_1^2/\operatorname{tg}\theta + c_2^2 \operatorname{tg}\theta)/2, \tag{3.6}$$

and $\operatorname{arctg} c_1/(2-c_2) \leqslant \theta \leqslant \operatorname{arctg}(2-c_1)/c_2$. Differentiating S_θ with respect to θ

$$(S_\theta)' = (-c_1^2/\sin^2\theta + c_2^2/\cos^2\theta)/2 \tag{3.7}$$

we have $(S_\theta)'>0$, for $\theta>\theta_0 \triangleq \operatorname{arctg} c_1/c_1$; <0, for $\theta<\theta_0$. Hence S_θ is a strictly convex function, and $S_{\theta_0} = \min$.

2) The vertical coordinate of the intersect point of $\underset{\sim}{X} = -1$ and $a^T \underset{\sim}{X} = t$ is

$(1-t_1)(1-c_1<1-t_1<1,\ t_1>0)$.

The area is

$$S_{t_1}^{(1)}=2t_1+[2(c_1-t_1)c_2+(c_1-t_1)(2-c_2)+c_2^2(c_1-t_1)]/2$$

and increases strictly with respect to t, and

$$S_{t_1}^{(1)}\geqslant[2c_1c_2+c_1(2-c_2)+c_2^2c_1(2-c_2)^{-1})/2=S_{\operatorname{arctg}c_1(2-c_2)^{-1}}. \tag{3.8}$$

3) The vertical coordinate of intersect point of $\underset{\sim}{X}=1$ and $\boldsymbol{a}^T\underset{\sim}{X}=t$ is $1-c_1<Y<1$. Let θ_1 be the angle between $y=1-c_1$ and $\boldsymbol{a}^T\underset{\sim}{X}=t$. We have

$$S_{\theta_1}^{(2)}=2c_1+[(2-c_2)^2\operatorname{tg}\theta_1-c_2^2\operatorname{tg}\theta_1)]/2. \tag{3.9}$$

So $S_{\theta_1}^{(2)}$ increases strictly with respect to θ_1, and

$$S_{\theta_1}^{(2)}\geqslant S_{\theta}^{(2)}=S_{c_1}^{(1)}=2c_1. \tag{3.10}$$

Let $t(\theta_1)$ mean the distance from the horizontal coordinate of the intersect point of $\boldsymbol{a}^T\underset{\sim}{X}=t$ and $\underset{\sim}{Y}=1$ to the point (1, 1). We easily see that $t(\theta_1)$ increases strictly with respect to θ_1, where θ_1 is the same as that in the case 3).

4) The vertical coordinate of the intersect point of $\boldsymbol{a}^T\underset{\sim}{X}=t$ and $\underset{\sim}{X}=1$, and that of the intersect point of $\boldsymbol{a}^T\underset{\sim}{X}=t$ and $\underset{\sim}{X}=-1$ are, respectively, larger than 1 and -1.

It is easy to see that

$$S_{\theta_1}^{(3)}=2c_1+[(2-c_2)^2\operatorname{tg}\theta_1-c_2^2\operatorname{tg}\theta_1+t^2(\theta_1)\operatorname{tg}\theta_1]/2 \tag{3.11}$$

and $S_{\theta_1}^{(3)}$ increases strictly with respect to θ_1.

$$S_{\theta_1}^{(3)}\geqslant S_{\operatorname{arctg}c_1/c_2}^{(3)}=S_{\operatorname{arctg}c_1/c_2}^{(2)}\geqslant S_{\theta}^{(2)}. \tag{3.12}$$

5) The horizontal coordinate of the intersect point of $\boldsymbol{a}^T\underset{\sim}{X}=t$ and $\underset{\sim}{Y}=-1$, t_2-1 say, is larger than -1.

Since the origin point (0, 0) lies in the half plane $\{\boldsymbol{a}^T\underset{\sim}{X}\leqslant t\}$, the angle between $\boldsymbol{a}^T\underset{\sim}{X}=t$ and $\underset{\sim}{Y}=1-c_1$ is less than or equal to $\operatorname{arctg}(1-c_1)/(1-c_2)$, and $-1<t_3-1<-(1-c_2)/(1-c_1)$. Then the area is

$$S_{\theta_1}^{(4)}=2c_1+\{[(2-c_2)^2-t_2^2-c_2^2+t^2(\theta_1)]\operatorname{tg}\theta_1\}/2. \tag{3.13}$$

Furthermore, it is easy to see that $t^2(\theta_1)\geqslant t_2^2$. Thus $S_{\theta_1}^{(4)}$ increases strictly with respect to θ_1, and

$$S_{\theta_1}^{(4)}\geqslant S_{\theta}^{(3)}. \tag{3.14}$$

So far, we have derived the fact that the area considered increases strictly with counter clockwise rotation for that line of $\{\boldsymbol{a}^T\underset{\sim}{X}=t\}$ crossing $(1-c_2, 1-c_1)$ from the case 2) to the case 5). Similar to the the cases 2) ~ 5), when rotating $\{\boldsymbol{a}^T\underset{\sim}{X}=t\}$ clockwise, the area of $\{\boldsymbol{a}^T\underset{\sim}{X}\geqslant t\}\cap[-1, 1]^2$ also increases strictly. Together with the case 1), we have showed that the area of $\{\boldsymbol{a}^T\underset{\sim}{X}=t\}\cap[-1, 1]^2$ achieves uniquely maximum at $\theta=\operatorname{arc\,tg} c_1/c_2$. So the conditions (i) and (ii) in Theorem 1 are checked. Moreover, the condition (iii) is clearly satisfied.

§ 4. The Discussion for the Testing of the Data Structure

Projection pursuit is one of the most promising approaches toward discovering and extracting unspecified structure from a high-dimensional data set: the concrete way is to search through all low-dimensional linear projections of the set and pick the interesting ones. This can be done either informally by visual search or more objectively by maximizing some numerical projection index. Clearly, the interpretation of the apparent structure must remain subjective, that is, whether found by visual search or by numerical optimization, the perceived structure may be spurious. So to make statistical test is necessary. In our case the null hypothesis to be tested corresponds to the assumption that there is no structure at all. The testing purpose is to decide whether it is worth to follow up on a promising, but possibly spurious lead. We ought to know whether we are working above or below the general noise level. If we dip below that level, that is, the random noise is high, our chance of picking non-spurious leads becomes dim. Hence it is an interesting problem how to test the degree of the random noise.

Utilizing the Kolmmogorov statistic as the testing statistic, öhrvik (1987, 1988) obtained two empirical formulas via simulating experiments

$$\eta(n,\ d,\ \varepsilon)\sim 2\exp\{-2n\varepsilon^2+(d-1)\log(2en/d)/\sqrt{\pi}\},\tag{4.1}$$

and

$$\eta(n,\ d,\ \varepsilon)\sim 2\exp\{-2n\varepsilon^2+2.464(d-1)\}.\tag{4.2}$$

(4.1) is close to our inequality (2.2).

If the empirical formulas above hold, and if we let $d,\ n\to\infty$ such that $\nu=n/d$ is kept constant, then letting

$$D_\infty=(\log(2e\nu)/2\nu\sqrt{\pi})^{1/2} \text{ or } D_\infty=(1.232/\nu)^{1/2},\tag{4.3}$$

$$|D_n-D_\infty|=O_p(1/d),\tag{4.4}$$

namely, D_∞ can be considerd as a convenient central value. Huber (1988) suggested using this value as a testing level. If the statistic observed is less than this value, we probably are gazing at more random noise and it hardly will be worth our time and effort to follow up on even a conspicuous structure seen in that projection. This testing procedure differs from a usual test since the central part of the distribution of the statistic is more important than the tail part of that. Of course, we must be careful in applying D_∞ obtained by simulating results since we only achieve in theory an inequality.

References

[1] Devroye, L., Bounds for the uniform derivation of empirical measure, *J. Multivariate Anal.*, 12

(1982), 72—79.

[2] Vapnik, V. N. & Cervonenkis, A. Ya, *On* the uniforom convergence of relative frequencies of events to their probabilities, *Theory ppobab. Appl.*, **16**, (1971), 264—280.

[3] Huber, P. J., Spurious structure? or improved bounds for the *Kolmogorov* index, Research Report, Harvard Univ., (1988).

[4] Ohrvik, J., Structure in noise, Research Report, Univ. of Stockholm, Dept. of Statistic, (1987).

[5] Ohrvik, J., On the distribution off the *Kolmogorov* distance in the worst direction, Research Report, Univ. of Stockholm, Dept. of Statistic, (1988).