Ann. Inst. Statist. Math.
Vol. 40, No. 2, 343-352 (1988)

CONSISTENT ESTIMATION OF LOCATION REGION

DAVID K. BLOUGH

Department of Agricultural Economics, University of Arizona, Tucson, AZ 85721, U.S.A.

(Received May 1, 1987; revised October 13, 1987)

Abstract. Asymmetric multivariate probability distributions can be difficult to characterize in terms of their location. The works of Doksum (1975, *Scand. J. Statist.*, **2**, 11–22) and Blough (1985, *Ann. Inst. Statist. Math.*, **37**, 545–555) provide the construction of a location region for a given distribution. Any point in this closed, convex region will serve as a location parameter. It is the purpose of this paper to obtain a consistent estimator of the location region. Consistency is defined in terms of an appropriate pseudometric.

Key words and phrases: Asymmetry, location region, consistency, pseudometric.

1. Introduction

Establishing the location of an asymmetric probability distribution is difficult due to the fact that there are many "reasonable" measures of location. This problem is compounded for multivariate distributions. Doksum (1975) addresses this issue in the univariate case, and Blough (1985) extends these results to the multivariate case. For an n-variate distribution function F, a closed, convex region in R^n is constructed, any point of which is a reasonable location parameter for F (Blough (1985)). Reasonable here refers to equivariance under certain transformations of R^n.

It is the purpose of this paper to develop a consistent estimator of the n-variate location region. In Section 2, a brief review of the construction of the location region is given. An estimator of this region will be given in Section 3, and its consistency will be established. Examples are provided in Section 4, using both computer generated and "real" data. The bivariate case will be discussed throughout for ease of representation, but generalizations to higher dimensions are straight forward.

2. Location region

Paralleling Doksum (1975), Blough (1985) used three methods to

construct a two-dimensional location region for a bivariate distribution function F. It was then established that these three methods yield essentially the same closed, convex region in the plane. Hence, this region contains all reasonable measures of location. In particular, it contains measures which satisfy order relations and are equivariant under translations, positive definite transformations and orthogonal transformations. These properties were used in the development of Method II of Blough's work. It should be noted that these axioms of location are common throughout the literature. For example, in the development of location measures by Oja (1983), he characterizes equivariance by way of non-singular affine transformations. But since every non-singular matrix A can be expressed as $A=SR$ where S is symmetric positive definite and R is orthogonal (see, for example, Birkhoff and MacLane (1971), pp. 259–260), Blough's axioms encompass those of Oja. Thus, the new measures of location presented by Oja can be found in the location region as constructed by Blough.

For purposes of estimation, Blough's Method III will be exploited. Let $\alpha \in (0, 2\pi]$, and let R_α be the transformation obtained by rotating the plane (counterclockwise) through an angle α. For $X=\begin{pmatrix} Y \\ Z \end{pmatrix}$, a bivariate random vector with distribution function F, let F_α denote the distribution function of $R_\alpha(X)$, with respective univariate marginals G_α and H_α. Method III uses the bivariate function of symmetry in direction α defined as

$$\theta_\alpha\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (1/2)[G_\alpha^{-1}(u) + G_\alpha^{-1}(1-u)] \\ (1/2)[H_\alpha^{-1}(v) + H_\alpha^{-1}(1-v)] \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} m_\alpha(u) \\ n_\alpha(v) \end{pmatrix} \quad \text{for} \quad u, v \in (0, 1/2] .$$

Let

$$\theta^*_{G_\alpha} = \inf\{m_\alpha(u) : 0 < u \leq 1/2\} ,$$

$$\theta^{**}_{G_\alpha} = \sup\{m_\alpha(u) : 0 < u \leq 1/2\} ,$$

$$\theta^*_{H_\alpha} = \inf\{n_\alpha(v) : 0 < v \leq 1/2\} ,$$

$$\theta^{**}_{H_\alpha} = \sup\{n_\alpha(v) : 0 < v \leq 1/2\} .$$

Then, using component-wise ordering of vectors, let

$$B_\alpha = \left\{ x : \begin{pmatrix} \theta^*_{G_\alpha} \\ \theta^*_{H_\alpha} \end{pmatrix} \leq x \leq \begin{pmatrix} \theta^{**}_{G_\alpha} \\ \theta^{**}_{H_\alpha} \end{pmatrix} \right\} .$$

The location region for F is

$$B_F = \bigcap_{\alpha \in (0, 2\pi]} R_{-\alpha}(B_\alpha) .$$

3. Estimation of location region

Given a random sample of size n from a bivariate distribution with distribution function F, we seek to construct from these n points an estimate of the location region. Since for each $\alpha \in (0, 2\pi]$, B_α is the Cartesian product of the ranges of the two marginal functions of symmetry (Blough (1985)), we can apply Doksum's univariate results marginally. So given a random sample

$$x_1 = \begin{pmatrix} y_1 \\ z_1 \end{pmatrix}, \; x_2 = \begin{pmatrix} y_2 \\ z_2 \end{pmatrix}, \ldots, \; x_n = \begin{pmatrix} y_n \\ z_n \end{pmatrix},$$

we choose a finite set of angles $\{\alpha_1, \alpha_2, \ldots, \alpha_L\}$ equally spaced in the interval $(0, 2\pi]$. Let

$$\begin{pmatrix} y_{ij} \\ z_{ij} \end{pmatrix} = R_{\alpha_j} x_i , \quad j = 1, 2, \ldots, L; \quad i = 1, 2, \ldots, n .$$

For each rotation α_j, the natural estimates of $\theta^*_{G_j}$, $\theta^{**}_{G_j}$, $\theta^*_{H_j}$ and $\theta^{**}_{H_j}$ are, respectively,

$$\hat{\theta}^*_{G_{nj}} = \min_{1 \le i \le n} [y_{(i)j} + y_{(n-i+1)j}] ,$$

$$\hat{\theta}^{**}_{G_{nj}} = \max_{1 \le i \le n} [y_{(i)j} + y_{(n-i+1)j}] ,$$

$$\hat{\theta}^*_{H_{nj}} = \min_{1 \le i \le n} [z_{(i)j} + z_{(n-i+1)j}] ,$$

and

$$\hat{\theta}^{**}_{H_{nj}} = \max_{1 \le i \le n} [z_{(i)j} + z_{(n-i+1)j}], \quad j = 1, 2, \ldots, L .$$

Here $y_{(i)j}$ and $z_{(i)j}$ are the i-th smallest observations after rotation R_{α_j} for the two respective coordinates. Hence,

$$\hat{B}_{n\alpha_j} = [\hat{\theta}^*_{G_{nj}}, \hat{\theta}^{**}_{G_{nj}}] \times [\hat{\theta}^*_{H_{nj}}, \hat{\theta}^{**}_{H_{nj}}] ,$$

and for the estimated location region we take

$$\hat{B}_{F_n} = \bigcap_{j=1}^{L} R_{-\alpha_j} (\hat{B}_{n\alpha_j}) .$$

Having thus constructed $\hat{B}_{F_n}$, it will now be shown that in some sense $\hat{B}_{F_n}$ is a consistent estimator of the location region B_F. To this end, let P denote the probability measure on R^2 induced by F. For any two P-measurable sets A and B, define the pseudometric ρ as

$$\rho(A, B) = P(A \Delta B) ,$$

where $A \Delta B$ denotes the symmetric difference of A and B.

DEFINITION 3.1. The empirical location region $\hat{B}_{F_n}$ will be called a consistent estimator of B_F if

$$\rho(\hat{B}_{F_n}, B_F) \to 0 \quad \text{as} \quad n \to \infty .$$

We then have the following:

THEOREM 3.1. *Let X be a bivariate random vector with distribution function F. Assume the following:*

(1) *F_α is strictly increasing in each argument for all $\alpha \in (0, 2\pi]$, and*

(2) *(smoothness) For each $\alpha_0 \in (0, 2\pi]$, if $\{\alpha_j\}_{j=1}^{\infty}$ is any sequence in $(0, 2\pi]$ such that $\lim_{j\to\infty} \alpha_j = \alpha_0$.*

Then $\lim_{j\to\infty} G_{\alpha_j}^{-1}(u) = G_{\alpha_0}^{-1}(u)$ for all $u \in (0, 1)$ and $\lim_{j\to\infty} H_{\alpha_j}^{-1}(u) = H_{\alpha_0}^{-1}(u)$ for all $u \in (0, 1)$. Then $\hat{B}_{F_n}$ is a consistent estimator of B_F (see Appendix for proof).

4. Examples

Figures 1–4 are plots of estimated location regions for six data sets. In each case, dispersion and correlation dependencies are removed by first "sphering" the data. The original orientation and location of the data will be preserved. That is, if S represents the observed variance-covariance matrix based on n observations, we can find an orthogonal matrix P such that $S = P'DP$, where D is diagonal. If $\bar{x}$ is the observed mean vector, we transform the data to

$$w_i = P'D^{-1/2}P(x_i - \bar{x}) + \bar{x} , \quad i = 1, 2, \ldots, n .$$

The estimated location region is then constructed using the techniques of Section 3 with $w_1, w_2, \ldots, w_n$.

Figure 1 represents the estimated location region for a bivariate normal distribution with mean vector $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and variance-covariance matrix $\Sigma = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$. A computer-generated sample of $n = 50$ observations was used.

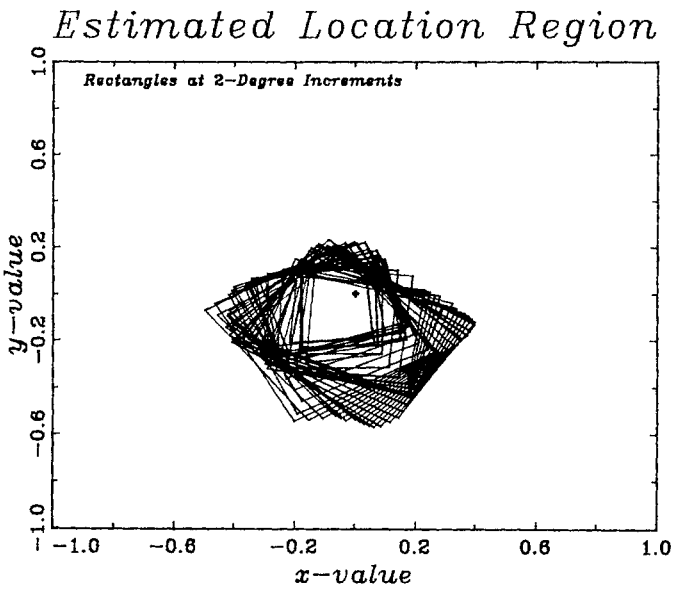

Fig. 1. Computer-generated bivariate normal data; $n=50$.

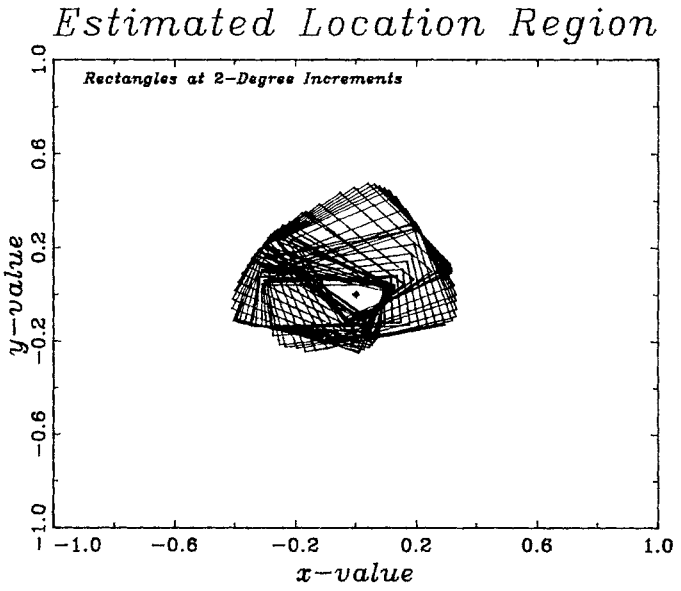

Fig. 2. Computer-generated bivariate normal data; $n=500$.

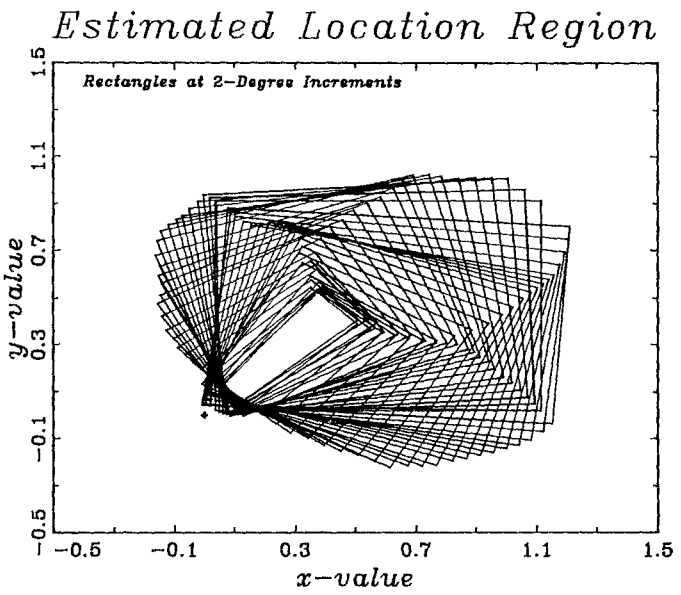

Fig. 3. Computer-generated bivariate i.i.d. beta data; $n=300$.

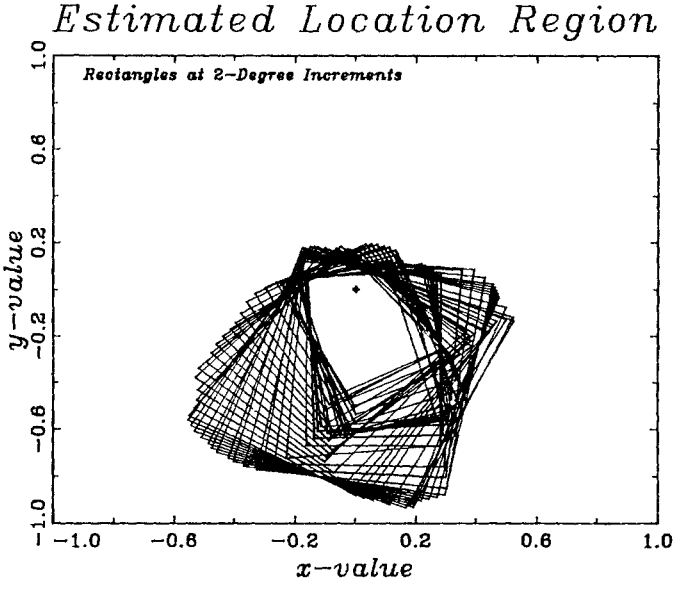

Fig. 4. Vapor pressure deficit residuals; $n=41$.

Figure 2 represents the estimated location region for $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 1/2\\1/2 & 1\end{pmatrix}\right)$, but now constructed from a computer-generated sample of $n=500$ points. The shrinkage, relative to Fig. 1 is apparent. This is evidence of the consistency of the estimator, and the fact that in this case $B_F=\left\{\begin{pmatrix}0\\0\end{pmatrix}\right\}$.

Table 1. Vapor pressure deficit data: time series residuals.

OBS	Ground level	Top of crop canopy
1	−0.174	−0.257
2	−0.330	−0.340
3	0.439	0.289
4	0.005	0.105
5	0.090	0.153
6	−0.147	−0.116
7	−0.932	−1.091
8	−0.092	−0.150
9	−1.144	−1.542
10	0.508	0.354
11	0.728	0.813
12	0.176	0.328
13	0.086	0.214
14	0.058	0.168
15	0.120	0.082
16	−0.124	−0.064
17	−0.004	−0.090
18	0.378	0.369
19	−0.433	−0.387
20	0.057	−0.016
21	−0.684	−0.988
22	0.062	0.017
23	0.243	0.463
24	0.639	0.800
25	−0.243	−0.213
26	0.146	0.133
27	−0.026	−0.019
28	−0.383	−0.433
29	0.149	0.102
30	0.172	0.190
31	−0.099	−0.134
32	−0.309	−0.464
33	−0.001	−0.063
34	0.215	0.279
35	0.071	0.084
36	−0.011	0.012
37	0.171	0.092
38	−0.509	−0.618
39	0.002	−0.032
40	−0.141	−0.091
41	−0.518	−0.864

Figure 3 represents the estimated location region for $X=\begin{pmatrix} Y \\ Z \end{pmatrix}$, where Y and Z are independent beta random variables, each with parameters $\alpha=3$, $\beta=15$. Since interchanging Y and Z leaves the distribution of X unchanged, the distribution is symmetric about the equi-angular line. Thus, B_F is some interval on the equi-angular line. This is reflected in the obvious elongation of $\hat{B}_{F_n}$ in the figure. The computer-generated sample was of size $n=300$.

Figure 4 represents the estimated location region for $n=41$ bivariate residuals obtained from a time series analysis. Irene Terry of the Department of Entomology at the University of Arizona collected data from a cotton field daily over a period of 41 days. Each day the vapor pressure deficit was measured at ground level and at the top of the crop canopy. A bivariate IMA (1, 1) time series model was fit to the data, and Table 1 presents the residuals from the fit. Figure 4 then is the estimated location region for these residuals. The plot strongly suggests that they come from a distribution symmetric about $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$. (This provides partial confirmation of the assumption that the underlying distribution is bivariate normal with mean $\mu=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$).

5. Conclusion

Consistent estimators of bivariate location region have been developed and implemented in some example data sets by way of a FORTRAN plotting program. Extending this results to three dimensions is straightforward. These techniques, when used in conjunction with interactive computer graphics could provide statisticians with two- and three-dimensional portrayals of location regions. Thus, bivariate and trivariate location could be quickly and consistently ascertained. These are directions for further implementation of these techniques.

Appendix

Proof of Theorem 3.1

LEMMA A.1. *For P-measurable sets* $\{A_j\}$ *and* $\{B_j\}$,

$$\left(\bigcap_j A_j\right) \Delta \left(\bigcap_j B_j\right) \subset \bigcup_j (A_j \Delta B_j) .$$

PROOF.

$$\left(\bigcap_j A_j\right) \Delta \left(\bigcap_j B_j\right) = \left[\left(\bigcap_j A_j\right) - \left(\bigcap_j B_j\right)\right] \cup \left[\left(\bigcap_j B_j\right) - \left(\bigcap_j A_j\right)\right]$$

But

$$\left[\left(\bigcap_j A_j\right) - \left(\bigcap_j B_j\right)\right] \subset \bigcup_j (B_j - A_j)$$

(see, for example, Chung (1974), p. 18). Similarly

$$\left[\left(\bigcap_j B_j\right) - \left(\bigcap_j A_j\right)\right] \subset \bigcup_j (A_j - B_j) .$$

Hence,

$$\begin{aligned}\left[\left(\bigcap_j A_j\right) \Delta \left(\bigcap_j B_j\right)\right] &\subset \left[\bigcup_j (A_j - B_j)\right] \cup \left[\bigcup_j (B_j - A_j)\right] \\ &= \bigcup_j [(A_j - B_j) \cup (B_j - A_j)] \\ &= \bigcup_j (A_j \Delta B_j) .\end{aligned}$$

PROOF OF THEOREM 3.1. By the smoothness assumption,

$$B_{\alpha_j} \to B_{\alpha_0} \quad \text{as} \quad j \to \infty .$$

This, together with the fact that $(0, 2\pi]$ is separable implies B_F can be constructed in a countable sequence of steps as follows:

(1) Construct B_{α_1} with $\alpha_1 = 0$. Let $\tilde{B}_{F,1} = B_{\alpha_1}$,

(2) Construct B_{α_2} with $\alpha_2 = \pi/2$ and let $\tilde{B}_{F,2} = B_{\alpha_1} \cap R_{-\alpha_2}(B_{\alpha_2})$,

(3) Construct B_{α_3} with $\alpha_3 = \pi/4$, and B_{α_4} with $\alpha_4 = 3\pi/2$ and let $\tilde{B}_{F,3} = \bigcap_{j=1}^{4} R_{-\alpha_j}(B_{\alpha_j})$.

Continuing in this way, constructing 2^{k-1} rectangles at stage k, all of them representing rotations equally spaced over the interval $(0, 2\pi]$, we have

$$\tilde{B}_{F,k} = \bigcap_{j=1}^{2^{k-1}} R_{-\alpha_j}(B_{\alpha_j}) .$$

Letting $\tilde{B}_F = \bigcap_{j=1}^{\infty} R_{-\alpha_j}(B_{\alpha_j})$, we clearly have

$$\rho(\tilde{B}_F, B_F) = 0 ,$$

and since

$$\tilde{B}_{F,k} \downarrow \tilde{B}_F \quad \text{as} \quad k \to \infty ,$$

$$\rho(\tilde{B}_{F,k},\ \tilde{B}_F) \to 0 \quad \text{as} \quad n \to \infty .$$

Now $\rho(\tilde{B}_{F,k},\ B_F) \leq \rho(\tilde{B}_{F,k},\ \tilde{B}_F) + \rho(\tilde{B}_F,\ B_F)$ implying

$$\rho(\tilde{B}_{F,k},\ B_F) \to 0 \quad \text{as} \quad k \to \infty .$$

Let $\varepsilon > 0$ be given. Then there exists a positive integer N such that

$$\rho(\tilde{B}_{F,N},\ B_F) < \frac{\varepsilon}{2} .$$

Also, there exist positive integers $M_1, M_2, \ldots, M_{2^{N-1}}$ such that for $n > M_j$

$$\rho(R_{-\alpha_j}(\hat{B}_{n\alpha_j}),\ R_{-\alpha_j}(B_{\alpha_j})) < \frac{\varepsilon}{2^N} \quad \text{a.s.} ,$$

for $j = 1, 2, \ldots, 2^{N-1}$. This follows immediately from the consistency of $\hat{\theta}^*_{G_{nj}}$, $\hat{\theta}^{**}_{G_{nj}}$, $\hat{\theta}^*_{H_{nj}}$ and $\hat{\theta}^{**}_{H_{nj}}$.

Let $M = \max\{M_1, M_2, \ldots, M_{2^{N-1}}\}$. Then for all $n > M$ we have

$$\begin{aligned}
\rho(\hat{B}_{F_n},\ B_F) &= \rho\left(\bigcap_{j=1}^{2^{N-1}} R_{-\alpha}(\hat{B}_{n\alpha_j}),\ B_F\right) \\
&\leq \rho\left(\bigcap_{j=1}^{2^{N-1}} R_{-\alpha_j}(\hat{B}_{n\alpha_j}),\ \tilde{B}_{F,N}\right) + \rho(\tilde{B}_{F,N},\ B_F) \\
&\leq P\left(\bigcap_{j=1}^{2^{N-1}} ((R_{-\alpha_j}(\hat{B}_{n\alpha_j}))\ \Delta\ (R_{-\alpha_j}(B_{\alpha_j}))\right) \\
&\quad + \rho(\tilde{B}_{F,N},\ B_F) \quad \text{(by the lemma)} \\
&\leq \sum_{j=1}^{2^{N-1}} \rho(R_{-\alpha_j}(\hat{B}_{n\alpha_j}),\ R_{-\alpha_j}(B_{\alpha_j})) + \rho(\tilde{B}_{F,N},\ B_F) \\
&< (2^{N-1}) \cdot \frac{\varepsilon}{2^N} + \frac{\varepsilon}{2} \\
&= \varepsilon .
\end{aligned}$$

REFERENCES

Birkhoff, G. and MacLane, S. (1971). *A Survey of Modern Algebra*, 3rd ed., MacMillan, New York.

Blough, David K. (1985). Measures of location in the plane, *Ann. Inst. Statist. Math.*, **37**, 545–555.

Chung, K. L. (1974). *A Course in Probability Theory*, Academic Press, New York.

Doksum, K. A. (1975). Measures of location and asymmetry, *Scand. J. Statist.*, **2**, 11–22.

Oja, H. (1983). Descriptive statistics for multivariate distributions, *Statist. Probab. Lett.*, **6**, 327–332.