L_1-Statistical Procedures and Related Topics
IMS Lecture Notes – Monograph Series (1997) Volume 31

Rank plots in the affine invariant case

B.M. Brown

University of Tasmania, Australia

T.P. Hettmansperger

Pennsylvania State University, USA

J. Möttönen and H. Oja

University of Oulu, Finland

Abstract: The bivariate ranks and quantiles based on the bivariate affine equivariant median are considered. Correspondences between two different plots for bivariate data, the direct diagram and the Oja rank plot, are described. Several illustrative examples are given.

Key words: Affine invariance, affine equivariance, bivariate quantile, bivariate rank, multivariate median.

AMS subject classification: 62G30, 62F07, 62H99.

1 Introduction

Rank methods occupy a central role among standard univariate statistical methods, and form the backbone of conventional nonparametrics. Consequently, it has been recently of some interest to explore concepts of rank for multivariate data, and in particular, for bivariate data. There are various alternatives, including ideas based on depth (Liu, 1990,1992; Liu and Singh, 1993). But another analytic definition of bivariate rank which leads to appealing bivariate analogues of univariate rank statistical methods is reached through the gradients of a convex objective function used to define a bivariate median; see Brown and Hettmansperger (1987a,b), Hettmansperger, Nyblom and Oja (1992), Hettmansperger, Möttönen and Oja (1997a,b) and Möttönen and Oja (1995). To show how this idea works, the notion of univariate rank is set up this way in Section 2.

There are several possible definitions of bivariate median which could be used to develop a notion of bivariate rank (Small, 1990; Niinimaa and Oja, 1997). Among these, the bivariate median of Oja (1983) is affine invariant. The resulting bivariate ranks are called Oja ranks, and defined in Section 3. They lead to the idea of bivariate quantile, which is a data item or region or chord between data items having prescribed rank.

The purpose of this paper is to examine the connections between and uses of two corresponding plots. The first, called the **direct diagram**, is just a plot of data in R^2 (the observation points with the lines going through pairs of observations). The second diagram, called an **Oja rank plot**, describes data items and regions having particular Oja rank values. From it, quantiles in the Oja sense can be read off. The Oja rank plot is developed in Section 4. In certain senses, the two plots have a duality relationship. Such connections, and other properties, are listed in Section 4. The correspondences between the two plots indicate considerable potential for higher dimensional versions to be useful in informal data analysis.

2 Univariate ranks

Given univariate data $x_1, \ldots, x_n$, the median m is defined as the choice $\theta = m$ to minimize the dispersion criterion

$$S(\theta) = \sum_{i=1}^{n} |x_i - \theta|.$$

Clearly $S'(\theta) = -\sum sgn(x_i - \theta)$, and this gradient function can be used to define univariate rank. In order to facilitate a clear analog with the coming bivariate case it is convenient to define $sgn(t) = +1$ if $t > 0$, -1 if $t < 0$, but $sgn(t)$ can take any value in $[-1, +1]$ when $t = 0$. If $x_{(j)}$ denotes the jth order statistic, then $\frac{1}{2}S'(x_{(j)})$ is any value in $[j - 1 - \frac{n}{2}, j - \frac{n}{2}]$ while in general, if $x_{(j-1)} < \theta < x_{(j)}$, then

$$\frac{1}{2}S'(\theta) = j - 1 - \frac{n}{2}.$$

The rank of the position θ among $\{x_i\}$ can therefore be defined as

$$R(\theta) = \frac{1}{2}S'(\theta),$$

and $-\frac{1}{2}n \leq R(\theta) \leq \frac{1}{2}n$, with $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$.

Inversion of the rank function leads to the notion of quantile. For $0 < p < 1$, the pth quantile ξ_p is the solution $\theta = \xi_p$ of $R(\theta) = (2p - 1)(\frac{n}{2})$. It is easy to verify that if $p = (j-1)/n$, then ξ_p is any value within $[x_{(j-1)}, x_{(j)}]$, while if $(j-1)/n < p < j/n$, then $\xi_p = x_{(j)}$.

The preceding definitions of univariate rank and quantile will now be extended to the bivariate case, using the Oja dispersion function for a bivariate median. Note now that later it is convenient to define bivariate quantiles in a way which is notationally different from the univariate case.

3 Oja bivariate ranks

Now suppose that $x_1, \ldots, x_n$ all $\in R^2$. The Oja bivariate median m is the choice $\theta = m$ to minimize

$$S(\theta) = \sum_{i<j} A(x_i, x_j, \theta)$$

where $A(a, b, c)$ is the area of the triangle having vertices a, b and c. The corresponding gradient function is

$$\nabla S(\theta) = \frac{1}{2} \sum_{i<j} u(x_i, x_j; \theta)$$

where u is a "repulsion vector", having magnitude $|x_i - x_j|$ and direction perpendicular to and away from the chord between x_i and x_j, towards θ. See Brown and Hettmansperger (1987a) for details. Correspondingly, the Oja rank of θ is defined as

$$\begin{aligned} R(\theta) &= \nabla S(\theta), \\ &= \frac{1}{2} \sum_{i<j} u(x_i, x_j; \theta) \end{aligned}$$

Note that ranks $R(\theta)$ are bivariate vectors, with direction as well as magnitude. The orientation of $R(\theta)$ among other $\{R(x_i)\}$ will be roughly similar to that of θ among $\{x_i\}$, but in a general sense the ranks display more regularity than the original data, resembling the situation for univariate ranks.

An important observation is that $R(\theta)$ remains constant as θ changes locally. Furthermore, R changes value only when θ crosses a line connecting some x_i, x_j. Then the increment to $R(\theta)$ is $\pm u(x_i, x_j; \theta)$. This observation establishes the fundamental basic relationship between the direct diagram, the plot of data $\{x_i\}$, and the Oja rank plot, of the rank values. It can be described as follows.

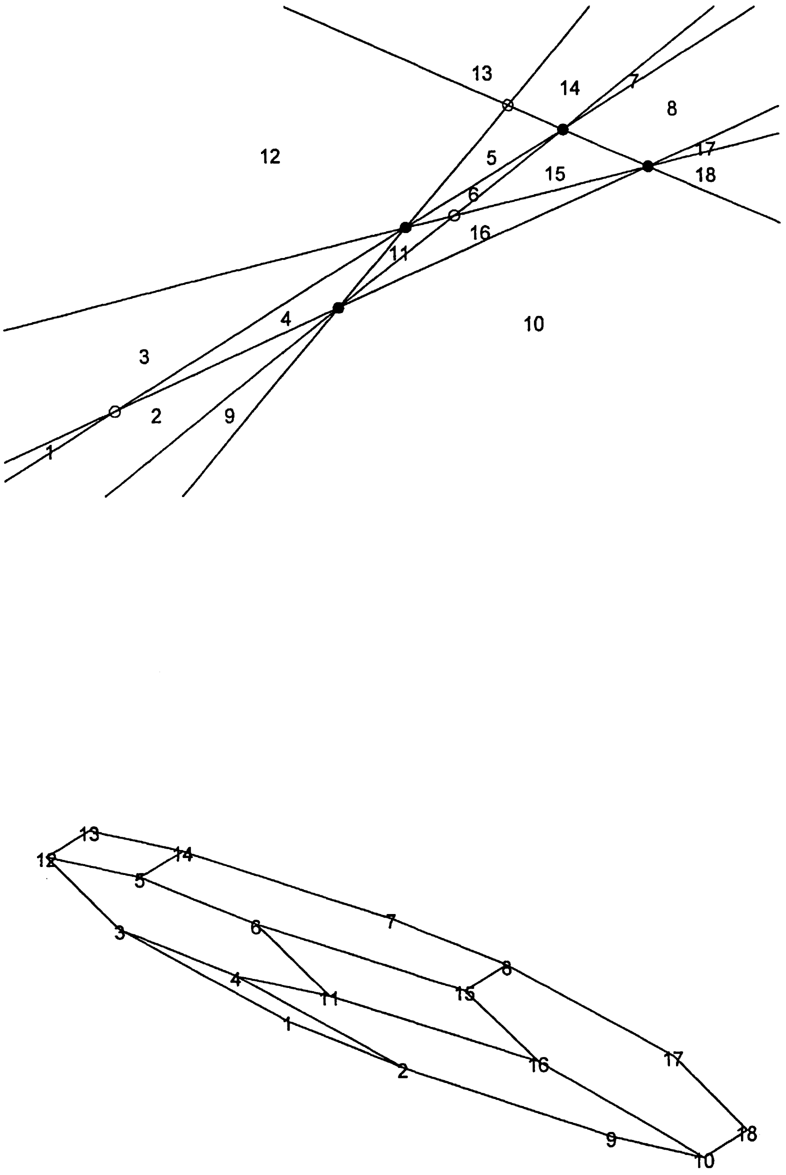

Figure 1: Direct diagram (upper case): Black points are original four data points; white points are secondary points. Tiles are numbered. Rank plot (lower case): Numbered vertices are rank plots of corresponding tiles.

Drawing in extended lines connecting all observation pairs x_i, x_j defines a **natural tiling** in the direct diagram. The tiles are polygons with vertices at observations, or at intersections of a line connecting x_i, x_j with a line connecting x_k, x_l, for some i, j, k, l. These intersections are called **secondary points**. For n original data points in the general position (no parallel lines), there are $3\binom{n}{4} = O(n^4/8)$ secondary points. For the points not in the general position, the number of secondary points (and consequently the number of tiles) is smaller.

The ranks for all θ within a tile of the direct diagram are constant vectors. That constant is a point of the Oja rank plot, so rank plot points correspond to tiles in the direct diagram. Neighbouring tiles have rank values differing by a repulsion vector u. Therefore the rank value of any point of a boundary between tiles can be any point in the rank plot on the chord between the rank value points of the two tiles.

Furthermore, in the direct diagram, n observation points and $3\binom{n}{4}$ secondary points lie at the junction of several tiles. Correspondingly, their rank values are not unique, but any value in the polygonal region of the rank plot whose vertices are the ranks of the abutting tiles. For illustration, see Section 4 and in particular Figure 1.

These observations lead to a number of further relationships between direct diagram and rank plot, described in the next section.

4 Direct diagram and rank plots

4.1 Relations between the tiles

The regions of the rank plot are more regular than the tiles in the direct diagram. Figure 1 show a case of $n = 4$ data points with just 3 secondary points. A small number of points minimizing the clutter in the figure was used to illustrate the duality relationship in a simple case. See Figure 3 for the rank plot of 10 points.

In general, every data point has $n - 1$ lines emanating from it, towards other data points, so in the rank plot the rank region for a data point is a $2(n - 1)$-sided polygon whose opposite sides are parallel and of equal length, i.e. an order-$(n - 1)$ parallelogram. By contrast, the rank regions of secondary points are conventional 4-sided parallelograms. Together, the two types of parallelogram form an orderly tiling of R^2 in the rank plot. All sides of regions are repulsion vectors as occurring in the definition of Oja rank. Figure 2 provide illustrations.

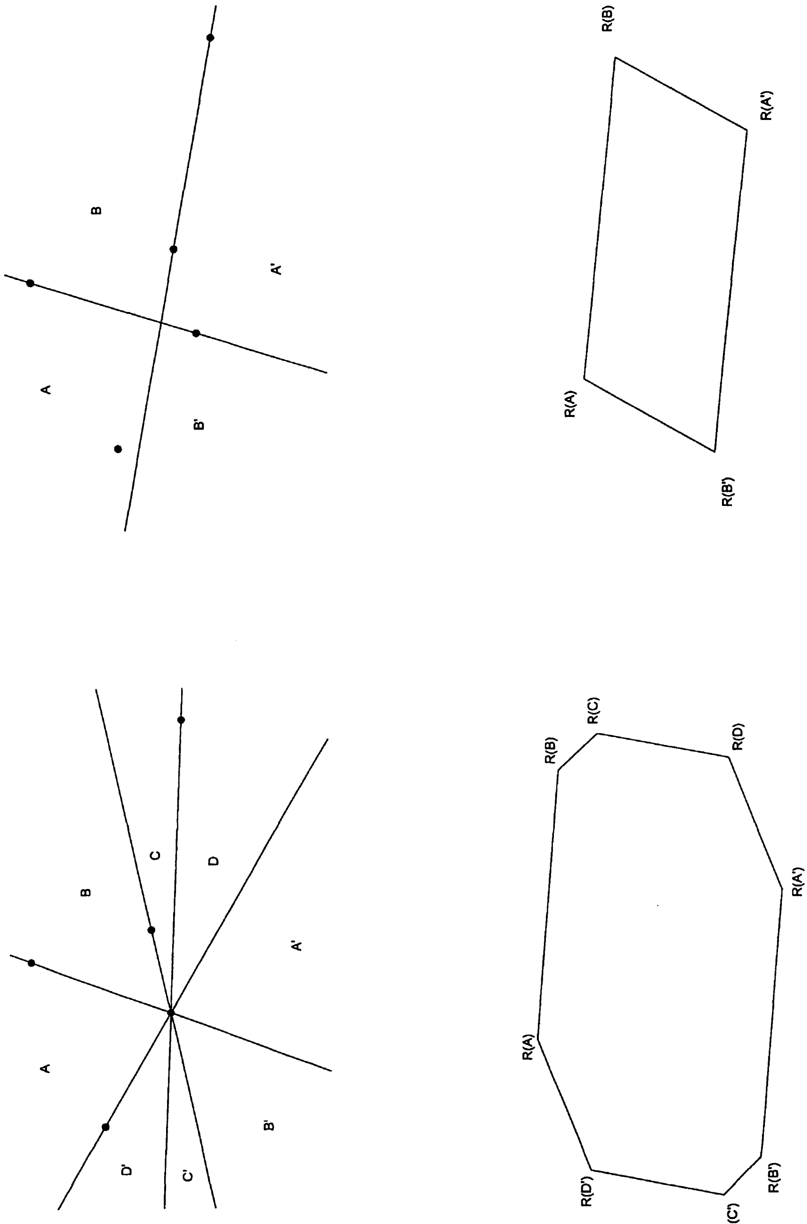

Figure 2: Direct diagrams and rank plots near a data point (lower case) and near a secondary data point (upper case).

There is complete projective duality between the two plots; that is, points in the direct diagram are associated with regions in the rank plot, chords in the direct diagram are associated with chords in the rank plot, and regions in the direct diagram are associated with points in the rank plot:

Direct plot	Rank plot
point	tile: $(n-1)$ parallelogram
secondary point	tile: 2 parallelogram
chord	chord
extended line	set of chords
tile	point

Only the location information is lost in the rank plot: The original data can be recaptured from the rank plot and the value of any data point (or Oja median). Let B be the sample covariance matrix computed on the Oja ranks. The **standardized rank plot** is then obtained if the rank plot items are multiplied (from the left) by $B^{-1/2}$. Both location and scale information is lost in the standardized rank plot.

Note that the Oja rank vectors are location invariant and affine equivariant in the sense that if the original observation vectors are multiplied by a full rank matrix A, the rank plot items will be multiplied by $A^* = abs(det(A))(A^{-1})^T$. If A is orthogonal then $A^* = A$ and if $A = diag(a_1, a_2)$ then $A^* = diag(a_2, a_1)$. For elliptical distributions, the eigenvectors from the Oja rank covariance matrix are then the eigenvectors of the conventional covariance matrix but the eigenvalues are reversed. The fact is connected to rank plot scale elongation occurring in orthogonal directions to scale elongations in the direct diagram; see Section 4.3.

4.2 The rank plot boundary

In using a rank value to assess the position of a point among points in a data cloud, it is useful to know the extremities of the rank plot. The rank plots are not standardized and the rank plot boundaries are determined by the data. The boundary tiles in the direct diagram each have an open face extending to ∞, and the rank values of these tiles form the vertices of the convex hull of the rank plot. Plotting these vertices will delineate the rank plot boundary, but there is a quicker more informal method of describing approximately where this boundary is.

This method, yielding an approximate boundary of the rank plot, is as follows.

Consider θ far away from the original data cloud, in direction α, i.e.

approximately

$$\theta = \hat{\theta} + r\begin{pmatrix}\cos\alpha\\ \sin\alpha\end{pmatrix}$$

with $r \to \infty$, where $\hat{\theta}$ is the Oja median. For r large, $R(\theta)$ does not depend on r. Then the contribution to $R(\theta)$ in direction α of the vector $(1/2)u(x_i, x_j; \theta)$ is approximately of magnitude

$$\frac{1}{2}r_{ij}|\sin(\alpha - \alpha_{ij})|, \tag{1}$$

where $r_{ij} = |x_i - x_j|$ and the line joining x_i, x_j has direction α_{ij} (or $\alpha_{ij} + \pi$). The sum of terms like (1) is awkward because of the absolute value, but a smooth approximation comes from using

$$|\sin x| \simeq a + \frac{1}{2}(1-a)(1-\cos 2x), \quad -\pi < x < \pi \tag{2}$$

Any value of a with $0 < a < 1$ may be used; the error of approximation varies between a at $x = 0$ and $(-1/2)(1-2a)^2/(1-a)$ when $|\sin x| = (1/2)(1-a)^{-1}$. The minimax error is $(2 - 2^{1/2})/4 = 0.1464$ at $a = (2 - 2^{1/2})/4 = 0.1464$ and this is a convenient choice for a. Then summing over terms in (1), using (2), gives

$$\begin{aligned} R(\theta) &= \frac{1}{4}(1+a)\sum_{i<j} r_{ij} - \frac{1}{4}(1-a)\gamma\cos(2\alpha - 2\omega), \\ &= \frac{1}{4}(1+a)R - \frac{1}{4}(1-a)\gamma[\cos(2\alpha)\cos(2\omega) - \sin(2\alpha)\sin(2\omega)], \end{aligned}$$

where

$$R = \sum_{i<j} r_{ij},$$

$$\gamma^2 = (\sum_{i<j} r_{ij}\cos(2\alpha_{ij}))^2 + (\sum_{i<j} r_{ij}\sin(2\alpha_{ij}))^2,$$

$$\cos(2\omega) = \gamma^{-1}\sum r_{ij}\cos(2\alpha_{ij})$$

and

$$\sin(2\omega) = \gamma^{-1}\sum r_{ij}\sin(2\alpha_{ij}).$$

The parameters $\cos(2\omega)$, $\sin(2\omega)$, ω, γ and R are easy to calculate and along with a, describe $R(\theta)$ as a simple cosine function of α, minimum at $\alpha = \omega \pm \pi$ and maximum at $\alpha = \omega$. The corresponding shape of the approximate rank plot boundary is approximately an ellipse whose major and minor axes give a rough indication of the principal components of the cloud of Oja ranks. See Figure 3 for the rank plot of a data set of 10 observations.

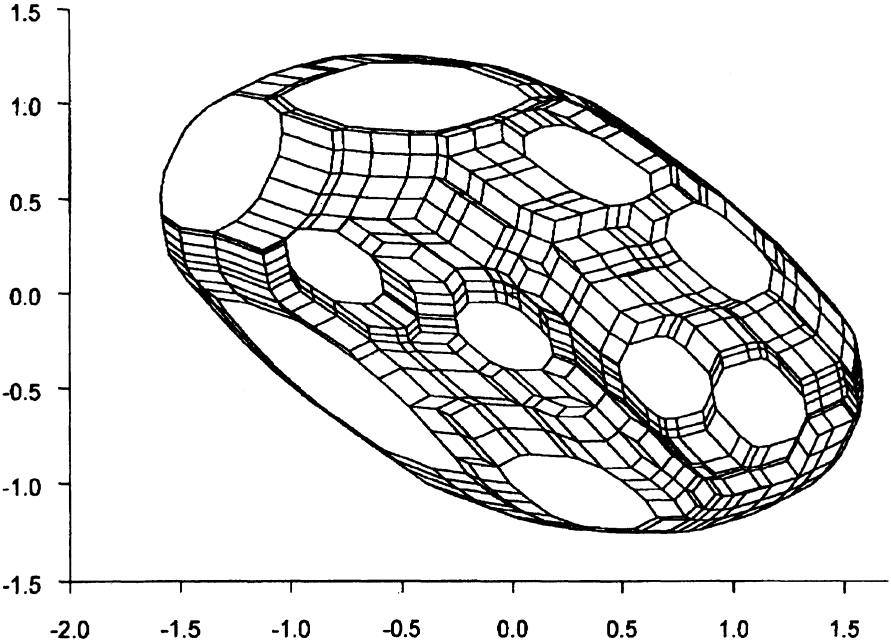

Figure 3: Rank plot for a data set of ten bivariate observations.

Note that

$$\lim_{\theta_1 \to \infty} R(\theta) = \frac{1}{2} \begin{pmatrix} \sum_{i<j} |x_{j2} - x_{i2}| \\ -\sum_{i<j} sign(x_{j2} - x_{i2})(x_{j1} - x_{i1}) \end{pmatrix}.$$

If $x_1, ..., x_n$ are i.i.d. from a spherical bivariate distribution with marginal Gini mean differences $\tau = E(|x_{11} - x_{21}|) = E(|x_{12} - x_{22}|)$ then clearly

$$\lim_{\theta_1 \to \infty} \binom{n}{2}^{-1} R(\theta) \to_P \frac{1}{2} \binom{\tau}{0}$$

and, for

$$\theta = \theta_0 + r \begin{pmatrix} cos\ \alpha \\ sin\ \alpha \end{pmatrix},$$

$$\lim_{r \to \infty} \binom{n}{2}^{-1} R(\theta) \to_P \frac{\tau}{2} \begin{pmatrix} cos\ \alpha \\ sin\ \alpha \end{pmatrix}.$$

The approximative boundary then is the sphere

$$\{\ \frac{\tau}{2} u\ \ :\ \ u^T u = 1\ \}.$$

For observations from an elliptical distribution

$$PCx_1, ..., PCx_n,$$

where P is orthogonal and $C = diag(c1, c2)$ diagonal, the asymptotic boundary is then the ellipse

$$\{ \frac{T}{2} PC^* u \ : \ u^T u = 1 \},$$

where $C^* = diag(c2, c1)$. Major and minor axes give the principal components for the original bivariate distribution.

4.3 Slopes of rank regions

In the rank plot, the rank region of a data point is the order $(n - 1)$ parallelogram of rank values attributable to that point, whose vertices are the rank values of all the tiles in the direct diagram which abut at the point. There is considerable information in the shape of a rank region as to the position of a data point within a data cloud : The lengths and directions of the $n - 1$ chords surrounding the rank region give the distances and directions of the other $n - 1$ points in the direct plot. (The direction is perpendicular $(\pm\frac{\pi}{2})$ to the direction of the chord.) Figure 4 illustrates how the rank region of an outlier will be elongated in a direction perpendicular to the direction of the rest of the data.

If a rank region has sides predominantly of one direction, the rest of the data is mostly oriented in a perpendicular direction from the data point in question. The lengths of the sides of a rank region are proportional to distances to other data points. Thus an outlier is distinguished by having a rank region with long sides, all with a similar direction.

Other remarks can be made.

(i) If all rank regions tend to have sides of similar direction, then the whole data cloud is elongated in a perpendicular direction.

(ii) Data points towards the center of a data cloud tend to have rank regions whose sides are of mixed lengths. The directions will reflect the general orientation of the data cloud.

(iii) Other data cloud patterns will have corresponding rank plot features. For instance, two separated mini-clouds yield a rank plot whose rank regions have sides tending to be distinctly short, in assorted directions, or distinctly long, in a definite direction perpendicular to the direction between the mini clouds.

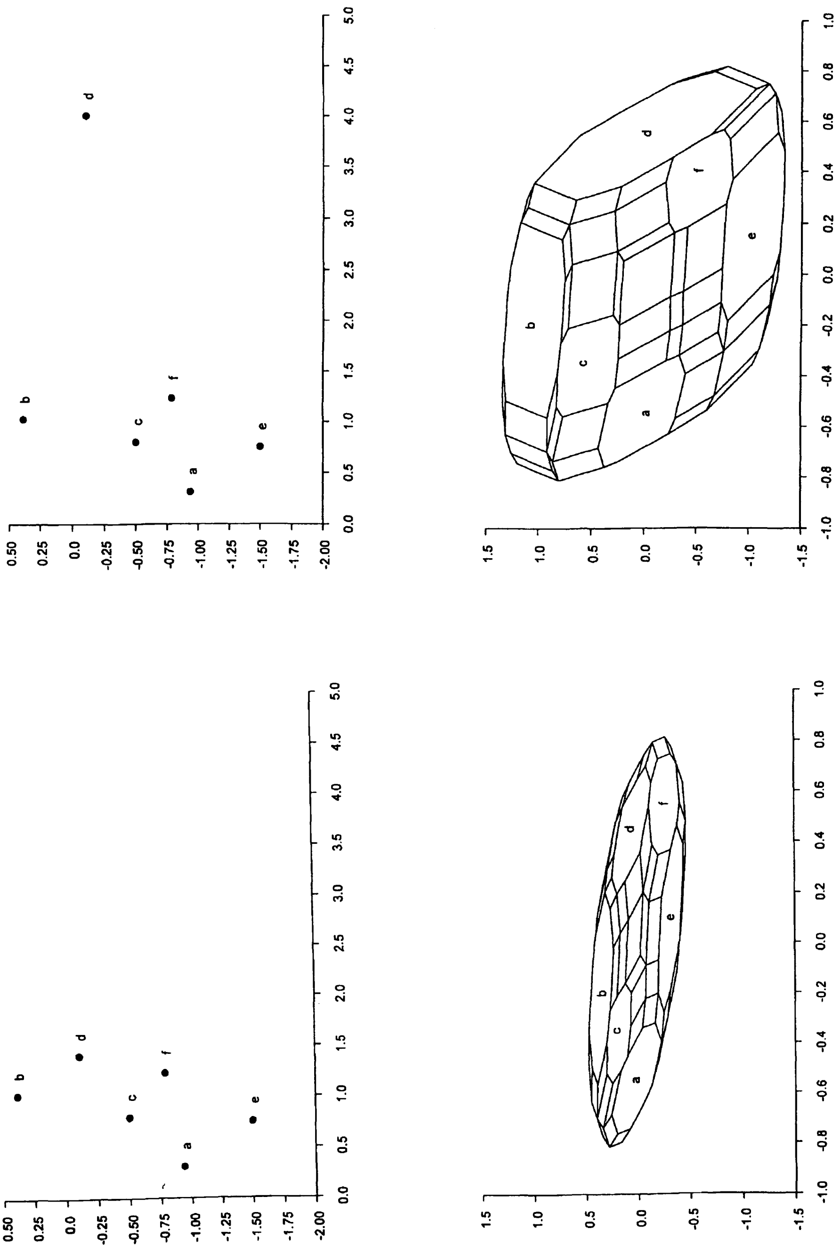

Figure 4: Direct diagrams and data plots for six observations: In the upper case observation *d* is moved to be an outlier.

References

[1] Brown, B. M. and Hettmansperger, T. P. (1987a). Affine invariant rank methods in the bivariate location model. *J. R. Statist. Soc.* B **49**, 301–310.

[2] Brown, B. M. and Hettmansperger, T. P. (1987b). Invariant tests in bivariate models and the L_1 criterion. In *Statistical data analysis based on the L_1-norm and related methods*, Ed. Y. Dodge, pp. 333–344. Amsterdam: North Holland.

[3] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1997a). Affine invariant multivariate one-sample signed-rank tests. *J. Am. Statist. Assoc.* To appear.

[4] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1997b). Affine invariant multivariate rank tests for several samples. *Statistica Sinica.* Conditionally accepted.

[5] Hettmansperger, T. P., Nyblom, J. and Oja, H. (1992). On multivariate notions of sign and rank. In *L_1-statistical Analysis and Related Methods*, Ed Y. Dodge, pp. 267–278. Amsterdam: North-Holland.

[6] Liu, R.Y. (1990). On a notion of data depth based upon random simplices. *Ann. Statist.* **18**, 405-414.

[7] Liu, R.Y. (1992). Data depth and multivariate rank tests. In *L_1-Statistical Analysis and Related Methods*, Ed. Y. Dodge, pp. 279-302. Amsterdam: North-Holland.

[8] Liu, R.Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Am. Statist. Ass.* **88**, 252-260.

[9] Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparam. Statist.* **5**, 201-213.

[10] Niinimaa, A. and Oja, H. (1997). Multivariate median. In *Encyclopedia of Statistical Sciences.* To appear.

[11] Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1**, 327–332.

[12] Small, G. (1990). A survey of multidimensional medians. *Inter. Statist. Rev.* **58**, 263-277.