The Bagplot: A Bivariate Boxplot

Peter J. ROUSSEEUW, Ida RUTS, and John W. TUKEY

We propose the bagplot, a bivariate generalization of the univariate boxplot. The key notion is the halfspace location depth of a point relative to a bivariate dataset, which extends the univariate concept of rank. The "depth median" is the deepest location, and it is surrounded by a "bag" containing the $n/2$ observations with largest depth. Magnifying the bag by a factor 3 yields the "fence" (which is not plotted). Observations between the bag and the fence are marked by a light gray loop, whereas observations outside the fence are flagged as outliers. The bagplot visualizes the location, spread, correlation, skewness, and tails of the data. It is equivariant for linear transformations, and not limited to elliptical distributions. Software for drawing the bagplot is made available for the S-Plus and MATLAB environments. The bagplot is illustrated on several datasets—for example, in a scatterplot matrix of multivariate data.

KEY WORDS: Algorithms; Depth contours; Graphical display; Location depth; Ranks.

1. THE UNIVARIATE BOXPLOT

The univariate boxplot (or box-and-whiskers plot) was proposed by Tukey (1977) as a tool for exploratory data analysis. Two examples of univariate boxplots are shown outside the frame of Figure 1, which displays the car weight x_i and engine displacement y_i of 60 cars (Chambers and Hastie 1993, pages 46–47). Along the x-axis we see a horizontal boxplot of the x_i. It consists of a *box* from the lower quartile of the x_i to their upper quartile, with a crossbar at the median of the x_i. Outside of the box, the upper fence is given by $Q_2 + 4(Q_3 - Q_2)$ and the lower fence by $Q_2 + 4(Q_1 - Q_2)$, where Q_j is the jth quartile hence Q_2 is the median. (The fences are not drawn.) The *whiskers* are the horizontal lines going from the box to the most extreme values inside the fences. For car weight, no x_i lies outside the fences.

Along the y-axis we see the vertical boxplot of the engine displacements. The four y_i lying outside of the fences are flagged as outliers. (Note that the Camaro and the Caprice share the same engine displacement, as do the Mustang and the Victoria, hence the vertical boxplot shows only two outliers.)

2. THE BIVARIATE CASE

The univariate boxplot is based on ranks since the box goes from the observation with rank $\lfloor \frac{n}{4} \rfloor$ to that with rank $\lceil \frac{3n}{4} \rceil$, and the central bar of the box is drawn at the median. A natural generalization of ranks to multivariate data is the notion of halfspace depth (Tukey 1975), which we will explain in the next section. Using this concept, we propose a bivariate version of the boxplot. Its main components are a *bag* that contains 50% of the data points, a *fence* that separates inliers from outliers, and a *loop* indicating the points outside the bag but inside the fence. The resulting graph is called a *bagplot*.

Consider the scatterplot in Figure 1. The depth median—that is, the point with highest halfspace depth—lies in the center and is indicated by a cross. The bag is the polygon drawn as a full line, with dark gray interior. The observations that lie outside of the bag but inside of the fence are indicated by a light gray loop. The fence itself is not plotted

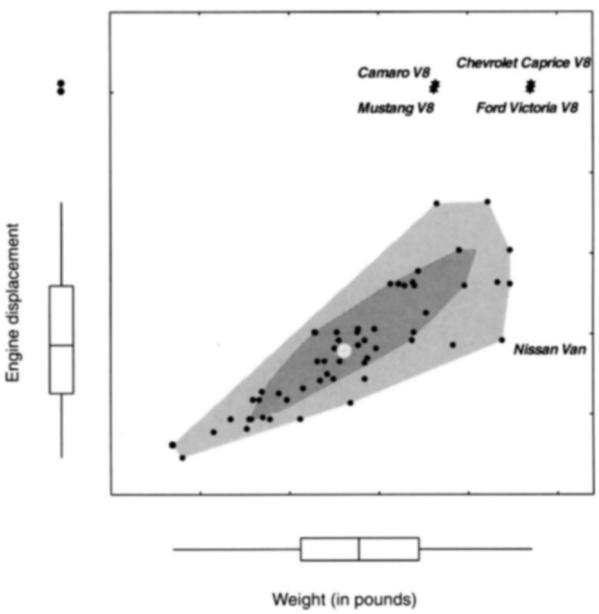

Figure 1. Car weight and engine displacement of 60 cars.

Peter J. Rousseeuw is Professor, and Ida Ruts is Assistant, Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (UIA), Universiteitsplein 1, B-2610 Wilrijk, Belgium. John W. Tukey is Emeritus Professor, Princeton University. We are grateful to the referees, the associate editor, and the editor for their helpful suggestions.

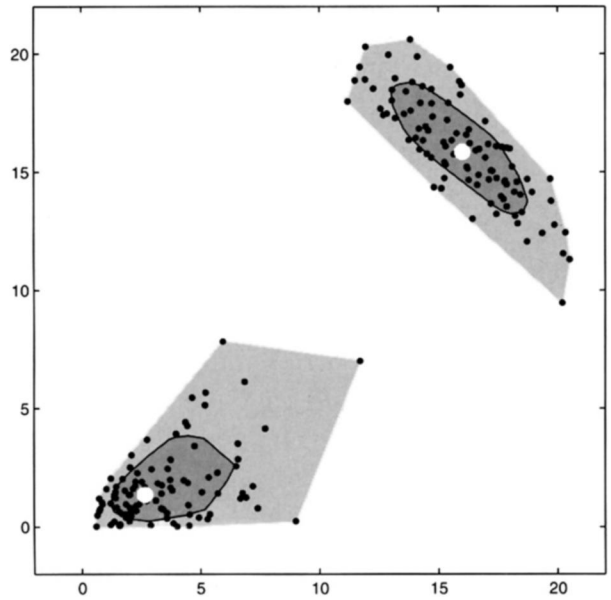

Figure 2. Bagplots of two datasets.

because it would draw the attention away from the data. We also see four observations outside the fence. These outliers are indicated by black stars and labeled. We also labeled the Nissan Van because it came close to the fence, so it is a boundary case.

Note that the bagplot generalizes the spine of the boxplot: For a very "flat" bivariate dataset (e.g., all $y_i \approx 0$) the bag becomes a box. The light gray loop plays the same role as the two whiskers in one dimension, so we could call Figure 1 a "bag-and-bolster plot" to stress the analogy with the term "box-and-whiskers plot."

Like the univariate boxplot, the bagplot also visualizes several characteristics of the data: its location (the depth median), spread (the size of the bag), correlation (the orientation of the bag), skewness (the shape of the bag and the loop), and tails (the points near the boundary of the loop and the outliers).

To illustrate these characteristics, Figure 2 contains the bagplots of two generated datasets, each with 100 points. Their medians (indicated by crosses) are far apart. The bags are of roughly the same size (area), so the datasets have a similar spread. But the bags have a different orientation: the left one slopes upward (positive correlation) and the other slopes downward. We also see that the first dataset is very skewed because the median lies in the lower left part of the bag, where the loop is also narrow, whereas the right part of the bag is wider and has a much wider loop. By contrast, the bagplot of the second dataset is nicely balanced, and its form suggests an elliptic distribution. Finally, both datasets are medium-tailed, judging from the size of the loop and the absence of outliers.

When showing several (possibly overlapping) datasets in one display, it is convenient to plot the bags in different colors. For instance, one bag may be plotted in blue with a light blue loop and dark blue stars for the outliers, whereas the other bag may be red with a light red loop and dark red stars.

3. CONSTRUCTION OF THE BAGPLOT

The halfspace location depth $\text{ldepth}(\boldsymbol{\theta}, Z)$ of some point $\boldsymbol{\theta} \in \mathbb{R}^2$ relative to a bivariate data cloud $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ was introduced by Tukey (1975); see also Eddy (1985) and Green (1985). It is the smallest number of $\mathbf{z}_i$ contained in any closed halfplane with boundary line through $\boldsymbol{\theta}$. A time-efficient algorithm for $\text{ldepth}(\boldsymbol{\theta}, Z)$ was provided by Rousseeuw and Ruts (1996). The depth region D_k is the set of all $\boldsymbol{\theta}$ with $\text{ldepth}(\boldsymbol{\theta}, Z) \geq k$, and was algorithmically constructed by Ruts and Rousseeuw (1996). The depth regions are convex polygons, and $D_{k+1} \subset D_k$. (Note that these regions are different from those generated by *convex hull peeling*, which first removes the vertices of the convex hull of the data cloud, then repeats this on the remainder of the dataset, and so on.)

The depth median $\mathbf{T}^*$ of Z (Donoho and Gasko 1992) is defined as the $\boldsymbol{\theta}$ with highest $\text{ldepth}(\boldsymbol{\theta}, Z)$ if there is only one such $\boldsymbol{\theta}$. Otherwise, $\mathbf{T}^*$ is defined as the center of gravity of the deepest region. An algorithm for $\mathbf{T}^*$ was provided by Rousseeuw and Ruts (1998).

We now construct the bag B as follows. Let $\#D_k$ denote the number of data points contained in D_k. One first determines the value k for which $\#D_k \leq \lfloor n/2 \rfloor < \#D_{k-1}$ and then interpolates linearly between D_k and D_{k-1} (relative to the point $\mathbf{T}^*$) to obtain the set B. The bag B is thus again a convex polygon. Appendix A of Rousseeuw and Ruts (1997) describes the construction of the bag in more detail.

The fence is obtained by inflating B (relative to $\mathbf{T}^*$) by a factor 3. The choice of the value 3 is based on simulations (see Rousseeuw and Ruts 1997, sec. 5). The points outside the fence are flagged as outliers. The loop contains all the data points between the bag and the fence. To be precise, its outer boundary is the convex hull of the bag and the nonoutliers.

When the observations $\mathbf{z}_i = (x_i, y_i)$ are subjected to a translation and/or a nonsingular linear transformation (e.g., a rotation), their bagplot is transformed accordingly. This is because the halfspace depth is invariant under such mappings, and convex polygons are mapped to convex polygons. Therefore the points inside the bag remain inside, the outliers remain outliers, and so on.

4. EXAMPLES

Figure 3a plots the concentration of plasma triglycerides against that of plasma cholesterol for $n = 320$ patients with evidence of narrowing arteries (Hand, Daly, Lunn, McConway, and Ostrowski 1994, p. 221). We see the depth median (marked by a cross), the dark gray bag, the light gray loop, and five outliers highlighted by black stars. Figure 3a illustrates the option of not plotting the points inside the bag, for people who prefer to avoid overplotting. (Our default is to plot all points.) Looking at this bagplot we see much skewness, which suggests taking the logarithm of both variables (also because they are both chemical concentrations). The result is shown in Figure 3b. We see that four of the five previous outliers are now inside the loop, whereas two new outliers arise at the bottom of the figure.

Figure 4a plots the abundance of butterflies versus their altitudinal range, in a mountain area in the northern Iberian peninsula (Gutiérrez and Menéndez 1995). This figure illustrates an alternative representation of the bagplot. The bag is the polygon drawn as a full line, and now the boundary of the loop is indicated by a dotted line. In Figure 4a we note three outliers. When we plot the logarithm of the y-variable in Figure 4b, the outliers disappear.

5. ALGORITHM AND IMPLEMENTATION OF THE BAGPLOT

The entire display is constructed by the algorithm BAGPLOT, which draws on computational tools developed in earlier papers. Two components are the subroutines LDEPTH (Rousseeuw and Ruts 1996) which computes the location depth of an arbitrary point in $O(n \log n)$ time, and ISODEPTH (Ruts and Rousseeuw 1996) which constructs the vertices of a depth contour in $O(n^2 \log n)$ time. For computing the depth median, the $O(n^2(\log n)^2)$ subroutine HALFMED (Rousseeuw and Ruts 1998) is called. Since these times increase quickly with n, we currently use an approximation when n is larger than 150. In that case we compute the depth median and the bag for a random subsample of size 150, and perform the other computations (whose time is linear in n) on the full dataset. In this way we can easily deal with datasets of a few thousand points.

Our S-Plus code for the bagplot is available from the website http://win-www.uia.ac.be/u/statis/index.html. Several options are available. For instance, the user may choose whether or not to plot the observations inside the bag. The outliers—that is, the observations outside the

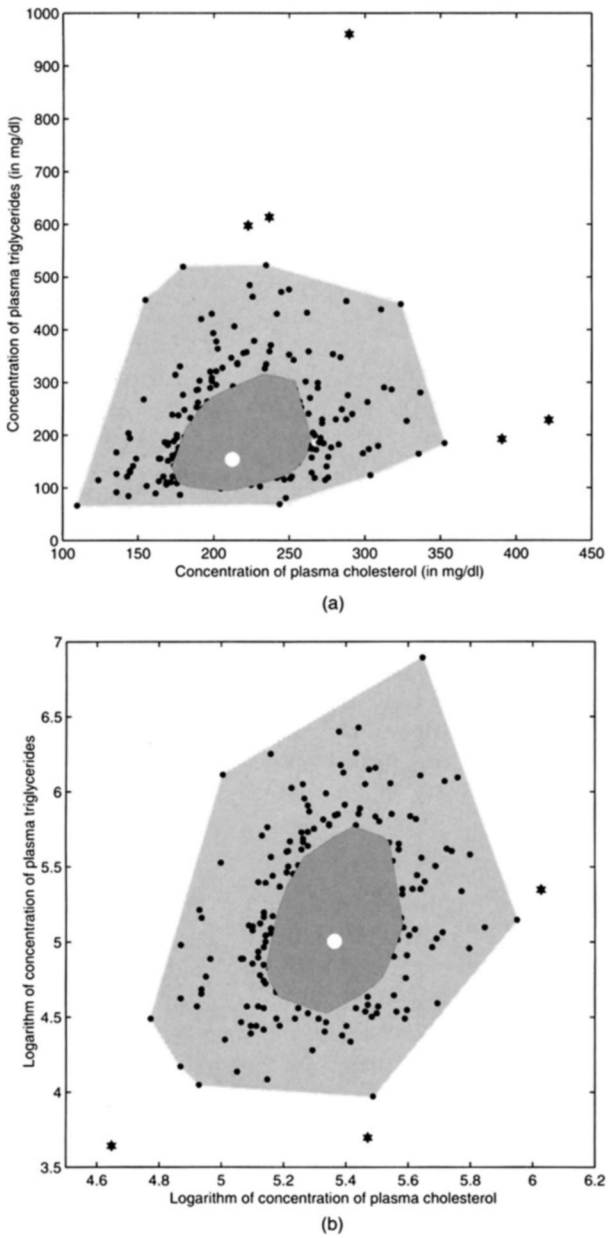

Figure 3. Part (a) shows the concentrations of cholesterol and triglycerides in the plasma of 320 patients. In Part (b) logarithms are taken of both variables.

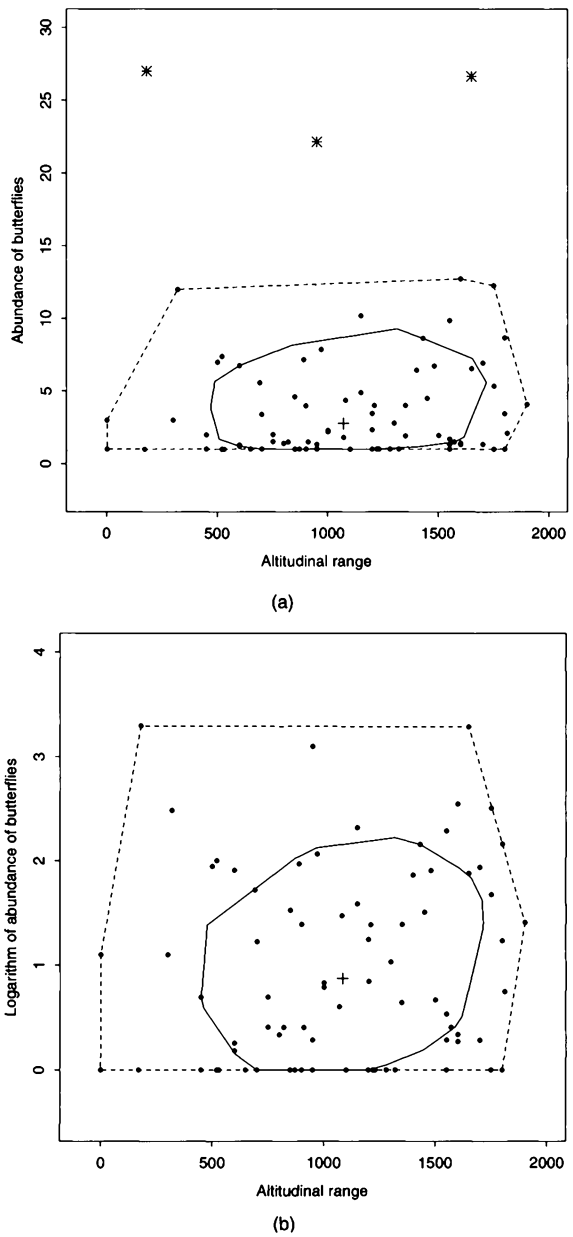

Figure 4. Part (a) shows the altitudinal range and abundance of butterflies. In part (b) the logarithm of the abundance is plotted.

fence—are always plotted. Any observation can be identified (e.g., labeled) by clicking on it.

Our MATLAB code for the bagplot is also available from the website mentioned earlier. Again several options are available—for example, the shading inside the bag may be omitted. Figures 3a and 3b illustrate the option of not plotting the points inside the bag. The user may also choose to plot the fence. We prefer the bagplot as in Figures 1 and 2—that is, with dark gray shading for the bag, light gray shading for the loop, no fence, and plotting all the points.

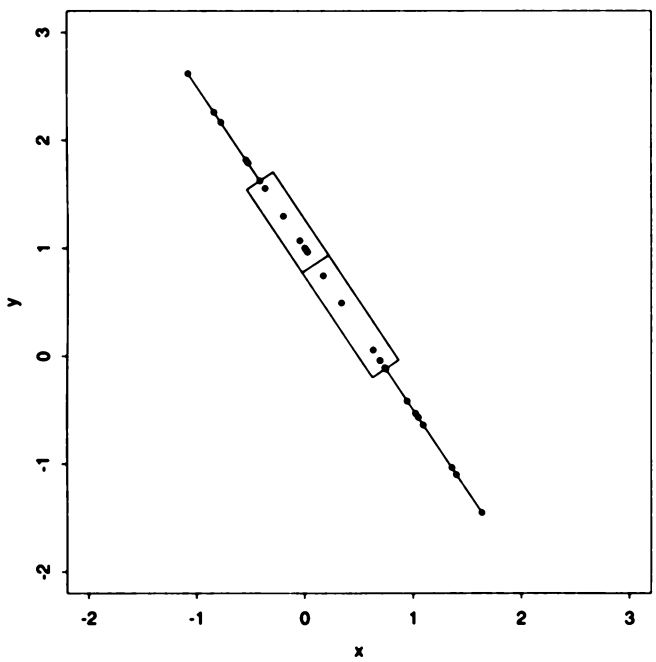

Figure 5. Bagplot for linear data.

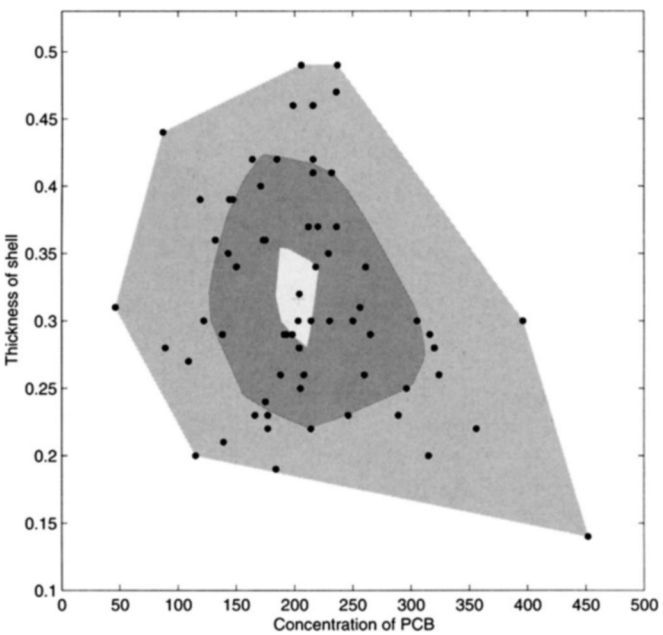

Figure 6. Blotched bagplot of PCB concentration and shell thickness of 65 pelican eggs. The white blotch inside the bag is a 95% confidence region for the depth median of the population.

By showing all the points we preserve the advantages of the scatterplot, because we can still see the local structure and note phenomena like a grid-like appearance (involvement of counts), curvature, clustering, holes in the data, and so on.

For large datasets the computation time can be kept down in several ways. Recently, Johnson, Kwok, and Ng (1998) constructed a fast exact algorithm for depth contours which outperforms ISODEPTH as soon as $n \geq 500$. Work is underway to construct algorithms for depth contours and the depth median with lower time complexity (Ileana Streinu, personal communication), but these are not yet available.

For small datasets, Rousseeuw and Ruts (1997, sec. 5) found that the variability of the fence is too large to reliably detect outliers (this is unavoidable for the outlier detection problem in two dimensions). Therefore, when $n < 15$ we draw only the depth median $\mathbf{T}^*$ and line segments between $\mathbf{T}^*$ and the data points.

If the dataset is linear, the bagplot reduces to a univariate boxplot. In that case our software draws it with a rectangular box, as in Figure 5. In a univariate boxplot the upper fence is given by $Q_2 + 4(Q_3 - Q_2)$ and the lower fence by $Q_2 + 4(Q_1 - Q_2)$, where Q_j is the jth quartile. (This version is well-suited for both symmetric and asymmetric distributions.) Note that the factor 4 in the univariate boxplot differs from the factor 3 used in the bivariate bagplot. Both values were obtained by simulations and experience, so the difference is not accidental but due to the dimensionality of the plots. (In Figure 5, using the factor 3 or 4 makes no difference because this example contains no points outside the fence by either definition.)

6. BLOTCHED BAGPLOTS

It is possible to incorporate a confidence region for the depth median into the bagplot. In this way we can extend one idea of notches (McGill, Tukey, and Larsen 1978) to the bivariate case. Consider the example of Figure 6. For 65 Anacapa pelican eggs, the concentration in parts per million of PCB (polychlorinated biphenyl, an industrial pollutant) was measured, along with the thickness of the egg shell (Hand et al. 1994, p. 131). All the observations have been plotted, and no outliers are flagged. Around the depth median we have drawn the *blotch*, which is a 95% confidence region for the depth median of the population. It was obtained by means of formula (2) in Rousseeuw, Van Aelst and Hubert (1999). This formula allows us to find the largest value k for which $P(\text{ldepth}(\tilde{\boldsymbol{\theta}}, X_n) \geq k) \geq .95$ when X_n comes from a distribution with population median $\tilde{\boldsymbol{\theta}}$. We then use the algorithm ISODEPTH to construct the corresponding depth region $D_k = \{\boldsymbol{\theta} \in \mathbb{R}^2; \text{ldepth}(\boldsymbol{\theta}, Z) \geq k\}$ which we call the blotch. For the pelican data we obtained $k = 21$ yielding the white blotch in Figure 6.

7. OTHER BIVARIATE DISPLAYS

Another type of bivariate boxplot was proposed by Becketti and Gould (1987). They consider the univariate boxplot of x (as in Figure 1) from which they keep the median, the quartiles, and the endpoints of both whiskers. They do

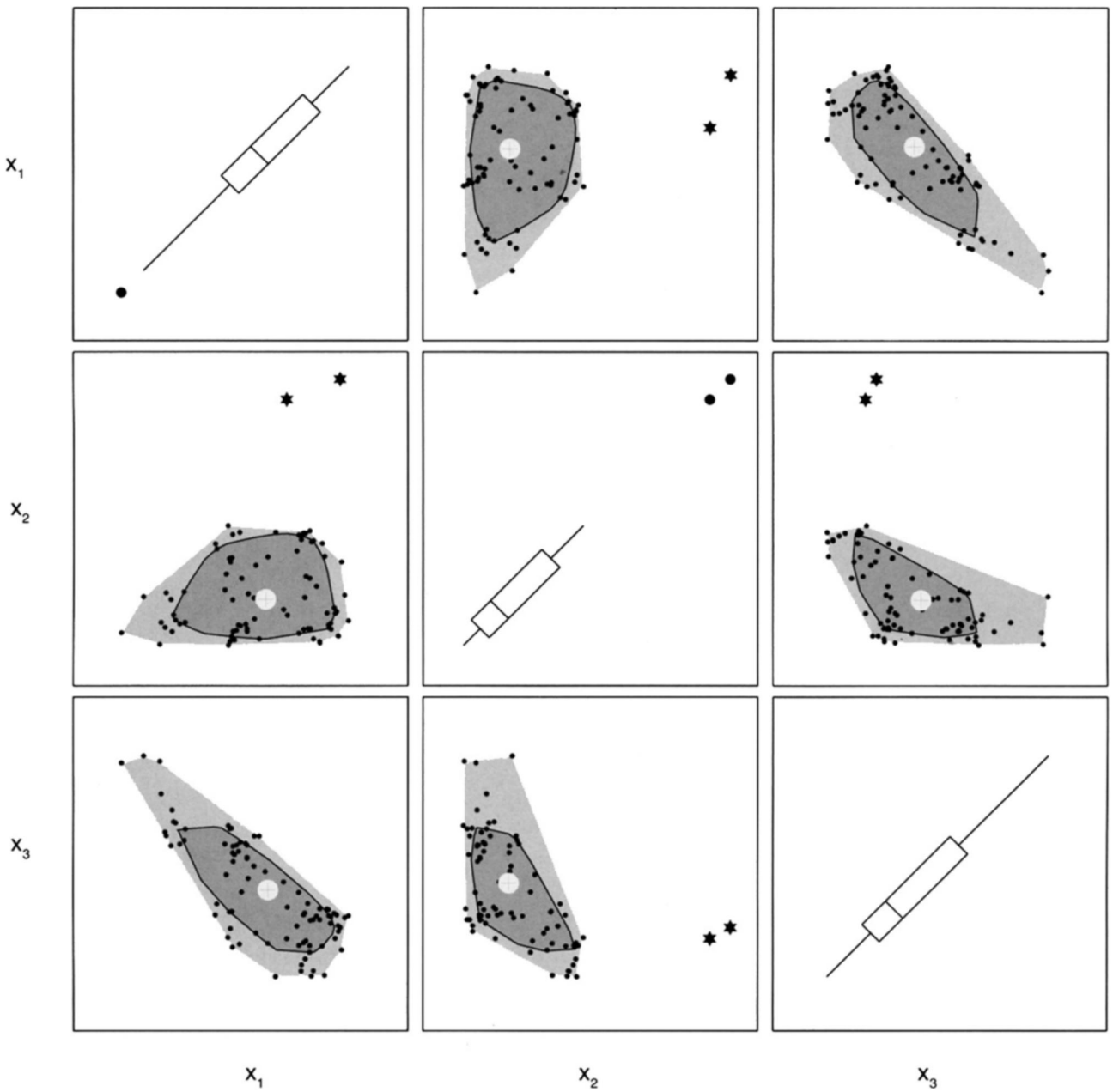

Figure 7. Bagplot matrix of the three-dimensional aquifer data with 85 data points.

the same for the y-variable, and then use these numbers to draw vertical and horizontal lines on the scatterplot, thereby forming a cross and a rectangle. Lenth (1988) modified this plot to put more emphasis on the univariate quartiles. However, neither version reflects the bivariate shape and correlation of the data.

Goldberg and Iglewicz (1992) proposed two generalizations of the boxplot which are truly bivariate. When the data can be assumed to be elliptically symmetric, they construct a robust elliptic plot (relplot). Here the "box" is an ellipse, obtained by a robust method such as the minimum volume ellipsoid estimator proposed by Rousseeuw (1984). For asymmetric data Goldberg and Iglewicz (1992) constructed a quarter elliptic plot (quelplot) where the "box" consists of four quarter ellipses, computed by a kind of M-estimator.

The bagplot differs from the relplot and the quelplot in that its shape is more general. Whereas the relplot/quelplot approach estimates parameters of (nearly) elliptical models, the bagplot is model-free because the halfspace depth is. Other variants of the bagplot have recently been constructed by Romanazzi (1997) and Liu, Parelius, and Singh (in press). Zani, Riani, and Corbellini (1998) applied convex hull peeling (see, e.g., Green 1985), which is somewhat less robust than halfspace depth, as shown by Donoho and Gasko (1992).

Recently, Hyndman (1996) constructed a plot of highest density regions (HDR's). First the bivariate density of

the data is estimated—for example, using a kernel method. Then the 50% HDR is given by the density contour that encompasses 50% of the mass. Typically, the 50% HDR and the 99% HDR are superimposed on the scatterplot of the data. Note that a HDR need not be convex, or even connected. This makes the HDR plot particularly useful to display multimodal distributions, because it focuses on local properties. The HDR plot is not a generalization of the univariate boxplot; for instance, its one-dimensional version may contain several "boxes." The basic notion of the HDR plot is data density, whereas the bagplot is based on ranking. Both approaches are appealing in different ways. The HDR approach is well-suited for multiple modes, but depends on the choice of a density estimator and a bandwidth. On the other hand, the bagplot is equivariant for linear transformations, needs fewer data points, and extends the familiar univariate boxplot. Note that both approaches have different (but equally important) views of what constitutes an outlier: for the HDR plot it is a point lying in an empty area (this could be called a "thinlier"), whereas for the bagplot it is a point lying far away from the bulk of the data. It should be noted that the depth median $\mathbf{T}^*$ and the bag B are robust, which makes the bagplot particularly suited for detecting (the latter type of) outliers.

8. MORE THAN TWO VARIABLES

The halfspace depth and the depth median exist in any dimension, so the bag can still be defined. For instance, there is now a fast approximate algorithm for the multivariate depth median (Struyf and Rousseeuw in press). This algorithm takes three minutes for 1,000 points in five dimensions. Currently, algorithms for depth contours in three or more dimensions are not yet available. In three dimensions the bag is a convex polyhedron, which in more dimensions becomes hard to visualize.

However, in any dimension one can draw the *bagplot matrix* which contains the bagplot of each pair of variables, as in Figure 7. This figure shows the aquifer data (Hand et al. 1994, p. 215–216) in which we changed the first two x_2 values to create outliers. (The variables x_1 and x_2 are coordinates measured in miles, and the water level x_3 is measured in feet above sea level.) Each diagonal cell is the bagplot of a variable against itself, where all the points lie on the $45°$ line. By construction such a bagplot reduces to a univariate boxplot, as explained in Section 5. Here we have drawn these boxplots with their usual factor 4 (instead of the factor 3 used in the bivariate bagplots).

An advantage of having all the data points in a bagplot matrix like Figure 7 is that we can still interactively color and brush as in the usual scatterplot matrices. The light gray zones of the bagplots should not detract from this, but rather aid the interpretation.

[Received May 1998. Revised May 1999.]

REFERENCES

Becketti, S., and Gould, W. (1987), "Rangefinder Box Plots," *The American Statistician*, 41, 149.

Chambers, J. M., and Hastie, T. J. (1993), *Statistical Models in S*, London: Chapman and Hall.

Donoho, D. L., and Gasko, M. (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, 20, 1803–1827.

Eddy, W. F. (1985), "Ordering of Multivariate Data," in *Computer Science and Statistics: Proceedings of the 16th Symposium on the Interface*, ed. L. Billard, Amsterdam: North-Holland, pp. 25–30.

Goldberg, K. M., and Iglewicz, B. (1992), "Bivariate Extensions of the Boxplot," *Technometrics*, 34, 307–320.

Green, P. J. (1985), "Peeling Data," in *Encyclopedia of Statistical Sciences* (Vol. 6), eds. S. Kotz and N. Johnson, New York: Wiley, pp. 660–664.

Gutiérrez, D., and Menéndez, R. (1995), "Distribution and Abundance of Butterflies in a Mountain Area in the Northern Iberian Peninsula," *Ecography*, 18, 209–216.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994), *A Handbook of Small Data Sets*, London: Chapman and Hall.

Hyndman, R. J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician*, 50, 120–126.

Johnson, T., Kwok, I., and Ng, R. (1998), "Fast Computation of 2-Dimensional Depth Contours," Technical Report, AT&T Research Center, Florham Park, NJ.

Lenth, R. (1988), Comment on "Rangefinder Box Plots" by S. Becketti and W. Gould (with reply by Becketti), *The American Statistician*, 42, 87–88.

Liu, R. Y., Parelius, J. M., and Singh, K. (in press), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference," *Annals of Statistics*.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978), "Variations of Box Plots," *The American Statistician*, 32, 12–16.

Romanazzi, M. (1997), "A Schematic Plot for Bivariate Data," *Student*, 2, 149–158.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and Ruts, I. (1996), "AS 307: Bivariate Location Depth," *Applied Statistics (JRSS-C)*, 45, 516–526.

——— (1997), "The Bagplot: A Bivariate Box-and-Whiskers Plot," Technical Report, Universitaire Instelling Antwerpen, Belgium. Available at http://win-www.uia.ac.be/u/statis/.

——— (1998), "Constructing the Bivariate Tukey Median," *Statistica Sinica*, 8, 827–839.

Rousseeuw, P. J., Van Aelst, S., and Hubert, M. (1999), "Rejoinder to the Discussion of Regression Depth," *Journal of the American Statistical Association*, 94, 419–433.

Ruts, I., and Rousseeuw, P. J. (1996), "Computing Depth Contours of Bivariate Point Clouds," *Computational Statistics and Data Analysis*, 23, 153–168.

Struyf, A., and Rousseeuw, P. J. (in press), "High-Dimensional Computation of the Deepest Location," *Computational Statistics and Data Analysis*.

Tukey, J. W. (1975), "Mathematics and the Picturing of Data," *Proceedings of the International Congress of Mathematicians*, 2, 523–531.

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Zani, S., Riani, M., and Corbellini, A. (1998), "Robust Bivariate Boxplots and Multiple Outlier Detection," *Computational Statistics and Data Analysis*, 28, 257–270.