ORIGINAL ARTICLE

Two-dimensional box plot

Phattrawan Tongkumchum

Abstract

Tongkumchum, P.
Two-dimensional box plot
Songklanakarin J. Sci. Technol., 2005, 27(4) : 859-866

In this paper we propose a two-dimensional box plot, a simple bivariate extension of the box plot and the scatter plot. This plot comprises a pair of trapeziums oriented in the direction of a fitted straight line, with symbols denoting extreme values. The choice for the fitted straight resistant line showing the relationship between the two variables is Tukey's resistance line. The main components of the plot are an inner box containing 50% of the projection points of the observations on the fitted line, a median point inside the inner box, and an outer box that separates outliers. The two-dimensional boxplot visualises the location, spread, correlation and skewness of the data.

Key words : two-dimensional box plot, bivariate graphical display, bivariate box plot, Tukey's line

Ph.D.(Statistics), Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani, 94000 Thailand.
E-mail: tphattra@bunga.pn.psu.ac.th
Received, 30 August 2004 Accepted, 19 November 2004

บทคัดย่อ

ภัทราวรรณ ทองคำชุม

บ๊อกพล๊อต 2 มิติ

ว. สงขลานครินทร์ วทท. 2548 27(4) : 859-866

การศึกษานี้นำเสนอบ๊อกพล๊อต 2 มิติซึ่งเป็นกราฟของ 2 ตัวแปรที่พัฒนามาจากบ๊อกพล๊อตและสแคตเตอร์พล๊อต กราฟนี้มีลักษณะเป็นสี่เหลี่ยมผืนผ้า 2 รูปวางซ้อนกันอยู่ในแนวของเส้นตรงที่แทนความสัมพันธ์ของ 2 ตัวแปร และมีสัญลักษณ์แทนค่าสุดโด่ง เส้นตรงที่แสดงความสัมพันธ์ของ 2 ตัวแปรที่เลือกใช้ในที่นี้คือเส้นทูคีย์ ส่วนประกอบหลักของกราฟคือ สี่เหลี่ยมรูปในซึ่งมี 50% ของภาพฉายของข้อมูลบนเส้นตรงอยู่ภายใน จุดแทนมัธยฐานของตัวแปรทั้งสองอยู่ภายในสี่เหลี่ยมรูปใน และสี่เหลี่ยมรูปนอกซึ่งแยกค่าสุดโด่งออก กราฟนี้แสดงตำแหน่ง การกระจาย สหสัมพันธ์ และความเบ้ของข้อมูล

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสงขลานครินทร์ อำเภอเมือง จังหวัดปัตตานี 94000

The scatter plot is widely used as a tool to summarize bivariate numerical data because it visualizes the relation between the two variables. However, for a large data set the scatter plot can be uninformative due to over plotting of the points. The box plot overcomes this problem in one dimension, by abbreviating the whole sample of points with a graph that only shows extreme values individually. There have been various attempts to construct bivariate plots (see, for example, Becketti and Gould, 1987; Lenth, 1988; Goldberg and Iglewicz, 1992; Hyndman 1996 and Rousseeuw *et al*., 1999).

In this paper we suggest an alternative way of constructing a bivariate association using rectangles to extend the univariate box plot to deal with points in a scatter plot.

Characteristics of the bivariate box plot

The proposed bivariate box plot comprises a pair of trapeziums oriented in the direction of a fitted straight line, with symbols denoting extreme values. Generalising the univariate box plot, in which the box goes from the lower to the upper quartile and the central bar of the box is at the median, a robust quartiles is needed, and Tukey's resistant line (McNeil, 1977) is appropriate because it is similarly defined in terms of the sample quartiles.

The main components of our bivariate box plot are an *inner box* containing 50% of the projection points of the observations on the fitted line, a *median point* that is inside an inner box, and an *outer box* that separates outliers. The resulting graph is called a *two-dimensional box plot*.

Like the univariate box plot, the two-dimensional box plot shows several characteristics of the data: its location (the median point), spread (the size of the box), correlation (the orientation of the box) and skewness (the positions of the median point and the outliers) in the box plot. The outer box plays the same role as the two whiskers in one dimension.

Figures 1 and 2 illustrate these characteristics. Figure 1 shows the two-dimensional box plot of two artificially generated data sets, each with 100 points. Their median points are indicated by circles. The boxes differ in size, indicating that the data sets have different spread. The boxes have different orientations: the left one slopes upward (positive correlation) and the other slopes downward. The median points lie approximately in the center of each box, so each sample is approximately symmetrical. Figure 2 (a) shows a plot of the concentration of plasma triglycerides against those of plasma cholesterol for 320 patients with evidence of narrowing arteries (Hand *et al*., 1994). We see the median point (marked by dots), the

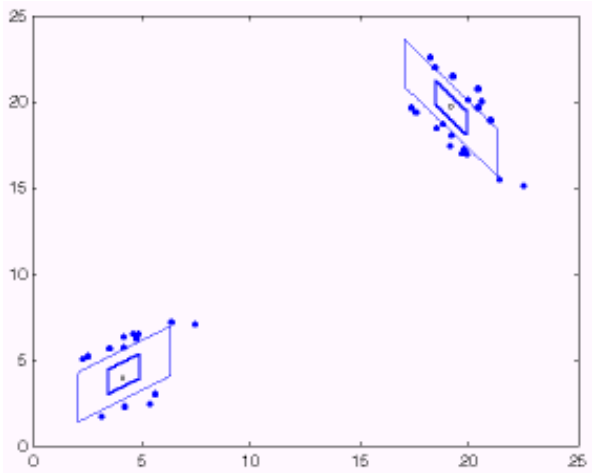
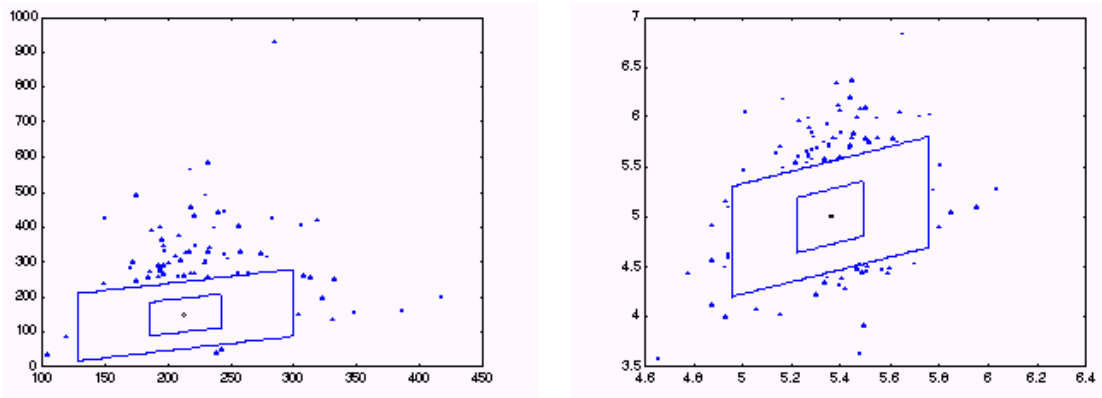


Figure 1. Two-dimensional box plots of two data sets.



(a) **(b)**

Figure 2. Part (a) shows the two-dimensional box plot of the concentration of cholesterol and concentration of triglycerides in the plasma of 320 patients. In the part (b) logarithm are taken of both variables.

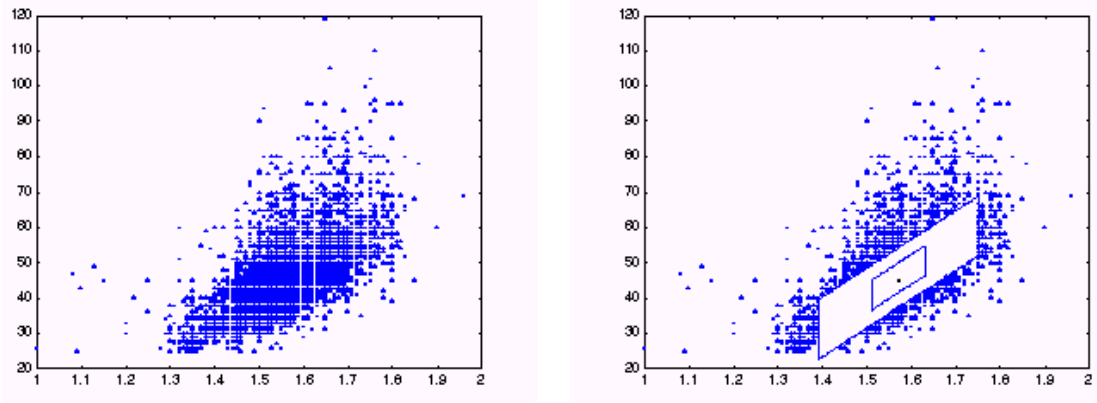
inner box, the outer box and outliers highlighted by dark dots. Looking at this plot we see skewness (more outliers above and to the right of the box), suggesting the desirability of a data transformation. The result after taking logarithms of each variable is shown in Figure 2 (b). We see that the boxes are now balanced with some outliers around them. Figure 3 shows a further example, based on a much larger set of data. These comprise weights and heights of 6316 children aged 9-21 years in secondary school in Pattani province. Figure 4 shows the same graph after taking logarithms of each variable.

Construction of the two-dimensional box plot

A basic concept of a two dimensional box plot is to fit a robust line to the scatter plot, and then to construct a box surrounding the fitted line. In order to construct a two dimensional box plot we use the following procedure:

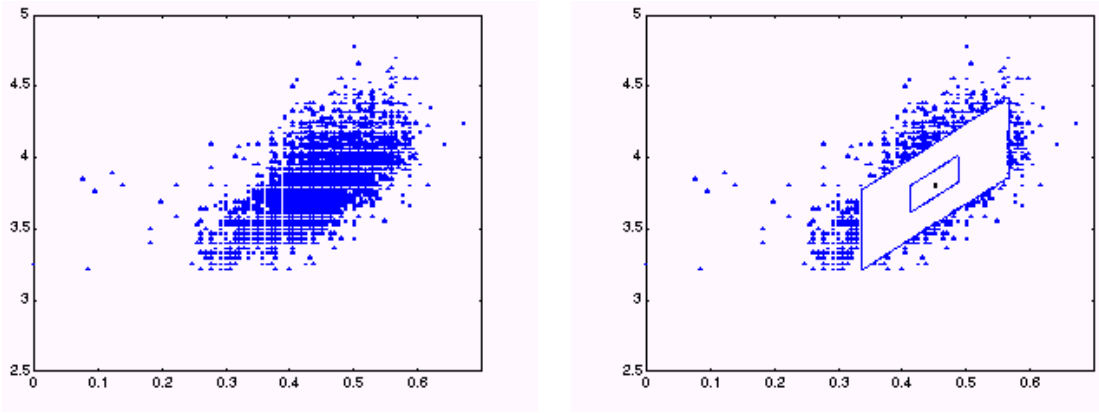
Step 1: Fit the robust line

Let (x_i, y_i) denote the data points, where i =



(a) **(b)**

Figure 3. Part (a) shows the scatter plot of weight and height of the 6316 secondary school children. Part (b) shows the two-dimensional box plot of the same data set.



(a) **(b)**

Figure 4. Part (a) shows the scatter plot of (the natural logarithm of) weight against height of the 6316 secondary school children. Part (b) shows the two-dimensional box plot of the same data set.

1, 2, 3, ..., n. Divide these points into three non-overlapping regions, according to their x-values, in such a way that each region contains equal or nearly equal numbers of points, approximately one third of the observations. Compute the median of the x-values and the median of the y-values in each of the outer regions. These values are denoted as (x_B, y_B) and (x_T, y_T), respectively. The slope (b) of the line joining the points (x_B, y_B) and (x_T, y_T) is $b = (y_T - y_B)/(x_T - x_B)$, and the intercept ($a$) of the line is a = median $\{y_i - bx_i\}$. Thus, Tukey's fitted line $\hat{y} = a + bx$ is obtained. Figure 5 shows the result of this step with the median point of the data set indicated by a larger dot and median points in the outer regions marked as crosses. The data set is 20 observations of weights and heights of 7-year-old school children in Hat Yai in 1992 (Tongkumchum, 2003).

Step 2. Compute quartile lines and fences based on $\{x_i\}$, parallel to the y-axis

Let $(\tilde{x}_i, \tilde{y}_i)$ denote the projection of the point (x_i, y_i) on the line $\hat{y} = a + bx$. $Q_{x(j)}$ is the j^{th} quartile

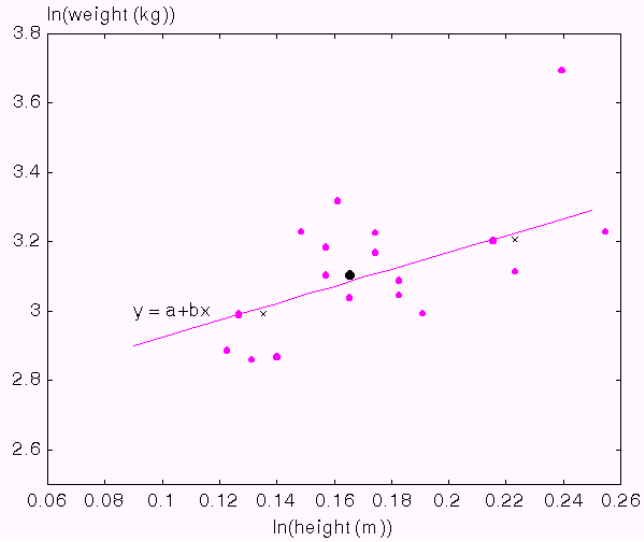


Figure 5. Robust line and medians.

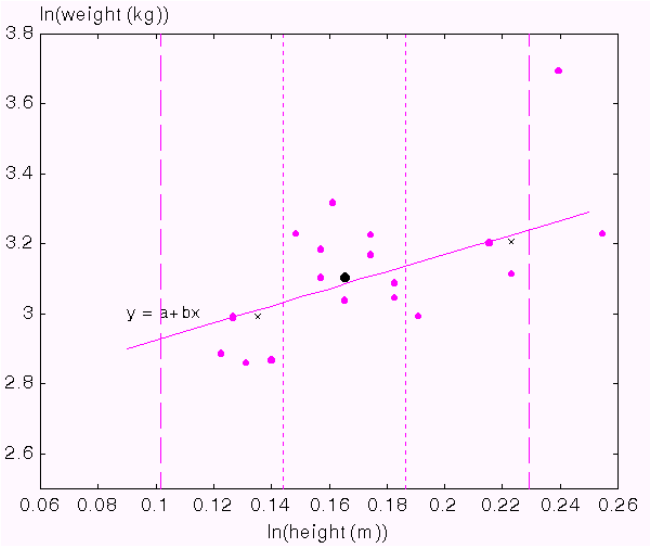


Figure 6. Quartile lines and fences parallel to the y-axis.

of the x_i. The quartile lines are defined as $x = Q_{\tilde{x}(1)}$ and $x = Q_{\tilde{x}(3)}$, where $Q_{\tilde{x}(1)}$ and $Q_{\tilde{x}(3)}$ are the lower and upper quartile, respectively, of the $\tilde{x}_i$ variable. Let D denote the inter-quartile range of the x_i. The *fences* are defined as $x = Q_{\tilde{x}(3)} + D$ and $x = Q_{\tilde{x}(1)} - D$. Figure 6 shows the quartile lines indicated by dotted lines and the fences indicated by dashed lines, for this data set.

Step 3. Compute quartile lines and fences based on the residuals $\{y_i - \hat{y}_i\}$, parallel to the fitted line $\hat{y} = a + bx$.

The residuals are defined by subtracting the fitted values from the y_i. For the observation (x_i, y_i) the residual is thus $e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$. $Q_{e(j)}$ is the j^{th} quartile of the residuals. Let D^* denote the inter-quartile range of the set of residuals. The

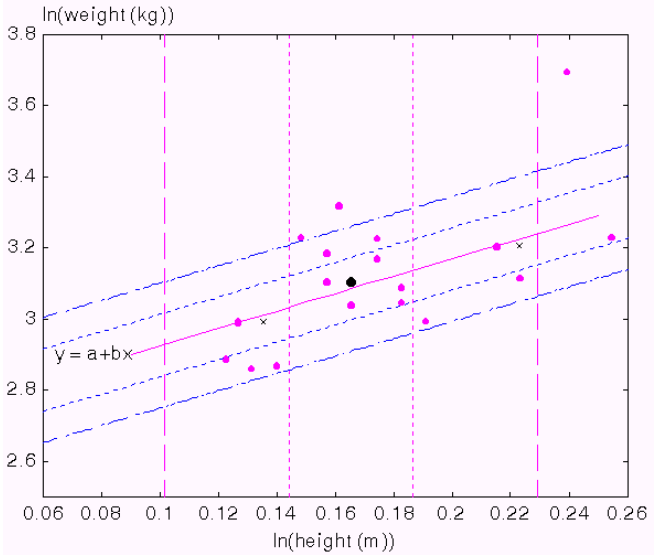


Figure 7. Quartile lines and fences parallel to the fitted line $\hat{y} = a + bx$.

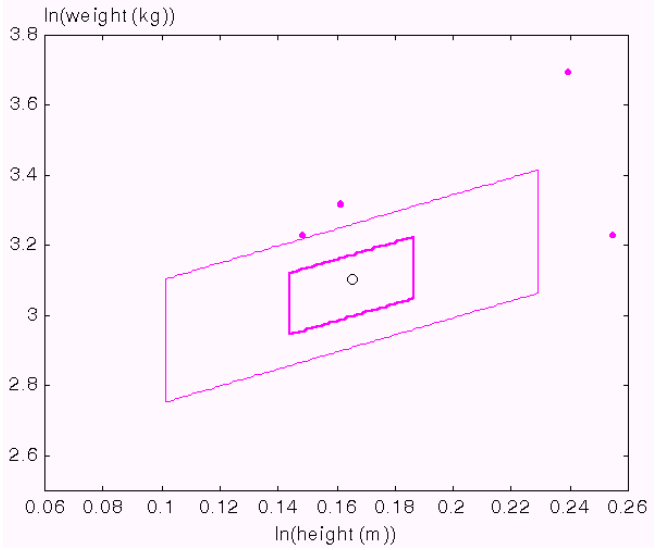


Figure 8. The two-dimensional box plot.

quartile lines are defined as $\hat{y} = a + bx - 0.5D^*$ and $\hat{y} = a + bx + 0.5D^*$. The fences are defined as $\hat{y} = a + bx - D^*$ and $\hat{y} = a + bx + D^*$. Figure 7 shows the quartile lines indicated by dotted lines and the fences indicated by dashed lines.

Step 4: Tidy up

The main components of the two-dimensional box plot are an inner box, an outer box, a median point and outliers. From Figure 8 the inner box comprises the four dotted lines and the outer box comprises the four dashed lines. The median point is the larger dot inside the inner box. The outliers are the dots outside the outer box.

After removing all the unnecessary lines,

we obtain the two-dimensional box plot, shown in Figure 8. The inner box is a small rectangle drawn with thick solid lines. The median point lies inside the inner box and is indicated by a circle. The outer box is a larger trapezium. The outliers are indicated by dots outside the outer box. Note that, as in the univarate box plot, the criterion for outliers in the two-dimensional box plot is arbitrary to some extent, depending on the definition of an outlier we want to show individually.

Other bivariate graphs

Various bivariate plots have been studied. Becketti & Gould (1987) proposed a "rangefinder box plot" that simply superimposes six line segments, two of which form a cross, on the scatter plot. Starting with the univariate box plots of the determinant (x) and the outcome (y), the medians, quartiles, and end points of each set of whiskers are used to draw horizontal and vertical lines on the scatter plot. An enhanced rangefinder graph was suggested by Lenth (1988), in which box plots of the x and y variables are incorporated along the axes. However, these two plots do not show the bivariate relations, shape, or correlation in the data.

Goldberg & Iglewicz (1992) proposed other types of bivariate box plots, namely a "relplot" (a robust elliptic plot) and a "quelplot" (a quarter elliptic plot). They constructed the relplot for data assumed to be elliptically symmetrical. In this case ellipses are obtained from fitting a bivariate Gaussian distribution. For nonsymmetric data the ellipses are replaced by four separate quarter ellipses matched on their major and minor axes, based on robust estimates of location and scale. In contrast to the rangefinder plots, these graphs are truly bivariate plots. They show the bivariate shape of the data using a pair of concentric ellipses or "quels" that serve as a hinge and a fence. The relplot is centered at the mean, whereas the quel is centered at "the centre of probability". The interior region contains 50% of the data and outer region delineates potential outliers. The plots show the location and scale of the data by two intersecting line segments, either on the regression lines or on the major and minor axes. However, this approach estimates parameters based on statistical model assumptions, rather than being distribution-free.

Hyndman (1996) proposed a plot of "highest density regions" (HDR). The idea is to summarise the probability distribution by a region of the sample space covering a specified probability and selecting the region containing relatively high probability density. The bivariate HDR is essentially a contour plot. It is constructed by using the 50% and 99% HDRs, with points lying outside the 99% HDR displayed as in a scatter plot. It is centered at the mode of the data. For constructing the HDR plot, the bivariate density of the data needs to be estimated (using a kernel method, for example), and then the 50% and 99% HDRs are superimposed on the scatter plot of the data. The 50% and 99% HDRs are given by the density contours that encompass 50% or 99% of the probability mass. This type of graph is very suitable for displaying multiple mode data.

Rousseeuw *et al*. (1999) proposed another bivariate extension of the univariate box plot, a "bag plot". The key concept is the halfspace location depth of a point, which extends the univariate concept of rank. The half space location depth ldepth (θ, Z) of some point $\theta \in IR^2$ relative to a bivariate data cloud $Z = \{z_1, z_2, ..., z_n\}$ was introduced by Tukey (1975). It is the smallest number of z_i contained in any closed half space with boundary line through θ. The depth region D_k is the set of all θ with ldepth (θ, Z) $\geq k$. The depth regions are convex polygons, and $D_{k=1} \subset D_k$. The deepest location is the "depth median". The depth median of Z is defined as the θ with highest ldepth, if there is only one such θ. Otherwise it is defined as the center of gravity of the deepest region. To obtain the bag plot, we first determine the value of k for which the number of data points in $D_k \leq [n/2] <$ the number of data point in D_{k-1} and then interpolate linearly between D_k and D_{k-1} relative to the depth median. The bag contains the $n/2$ observations with greatest depth surrounding the depth median. The "fence" is defined by magnifying a bag by a factor of 3. The points outside the fence are marked as outliers.

Our box is different from a region generated by "convex hull peeling", which first removes the vertices of the convex hull of the data cloud, then repeat this on the remainder of the data set, and so on (Rousseeuw, 1999). Our box is also different from the 50% HDR idea, for which the bivariate density of the data first needs to be estimated, and then the 50% HDR is superimposed on the scatter plot of the data.

Conclusion

In this paper we have suggested a simple method for constructing a bivariate box plot based on fitting a robust line to the scatter plot, and then constructing a box surrounding the fitted line. Our two-dimensional box plot comprise a pair of trapeziums oriented in the direction of a fitted straight line, with symbol denoting extreme values. One way of fitting a straight line uniquely to a bivariate data set is Tukey's line. The main components of our two-dimensional box plot are an "inner box" containing 50% of the projection points of observations on the fitted line, "a median point" that is inside the inner box, and an "outer box" that separates outliers. The two-dimensional box plot visualises the location, spread correlation and skewness of the data. This graph is suitable for showing linear relationships. Theoretical questions, such as stability with respect to subsamples and behaviour for data having a non-linear association, are worth pursuing but are beyond the scope of this paper.

Acknowledgements

I would like to thank Professor Don McNeil from Department of Statistics, Macquarie University, Australia, for his guidance and comments.

References

Becketti, S. and Gould, W. 1987. Rangefinder Box Plots: A note, Amer. Statist. 41(2): 149.

Goldberg, K.M. and Iglewicz, B. 1992. Bivariate Extension of the Boxplot, Technometrics. 34(3): 307-320.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. 1994. A hand book of Small Data Sets, Chapman and Hall, London.

Hyndman, R.J. 1996. Computing and Graphing Highest Density Regions, Amer. Statist. 50(3): 120-126.

Lenth, R.V. 1988. Comment on Becketti and Gould (with reply from Becketti), Amer. Statist. 42(1): 87-88.

McNeil, D.R. 1977. Interactive Data Analysis, John Wiley & Sons, New York.

Rousseeuw, P.J., Ruts, I. and Tukey, J.W. 1999. The Bagplot: A Bivariate Boxplot, Amer. Statist. 53(4): 382-387.

Tongkumchum, P. 2003. Modelling Adiposity: A Large Cohort Study in Hat Yai, Thailand. Ph.D. Dissertation Macquarie University.

Tukey, J.W. 1975. Mathematics and the Picturing of Data, Proceedings of the International Congress of Mathematicians, 2, 523-531.

Tukey, J.W. 1977. Exploratory Data Analysis, Addision-Wasley, Reading: MA.