Statistica Sinica **8**(1998), 813-826

BIVARIATE QQ-PLOTS AND SPIDER WEB PLOTS

John I. Marden

University of Illinois

Abstract: QQ-plots are extremely useful in univariate data analysis. In this article, Koltchinskii (1997) and Chaudhuri's (1996) definition of multivariate quantile is used to develop analogous plots for bivariate data. Bivariate qq-plots are exhibited for comparing a sample to a given population distribution (the bivariate normal), and for comparing two or more bivariate samples. The plots are based on drawing arrows from the quantiles in one distribution to the corresponding quantiles in the other. These plots can reveal differences in location, scale and skewness, as well as outliers. Spider web plots are introduced for plotting a systematic set of quantiles for a single sample without having to specify a reference population distribution.

Key words and phrases: Arrow plots, bivariate qq-plots, bivariate quantiles, bivariate ranks, spider web plots.

1. Introduction

QQ-plots are popular and useful diagnostic tools in univariate analysis. They provide graphical assessment of the fidelity of a sample to a particular distribution F, or of the differences between two independent samples. The idea behind a qq-plot is to match quantiles, that is, choose a set of quantiles $q_1, \ldots, q_L$, then plot the q_l^{th} quantile of the distribution F versus the q_l^{th} quantile of the sample (or the q_l^{th} quantile of one sample versus the other) for each l. The closer the plotted points are to the 45°-line, the closer the two distributions.

When moving to two or more dimensions, there are three questions that must be answered: What is a multivariate quantile? How are the quantiles to be plotted chosen? How are the corresponding quantiles in the two distributions plotted against each other? The purpose of this paper is to present some possible answers to these questions, at least in the bivariate case. The next section defines the multivariate quantiles we use. The definition is that of Koltchinskii (1997) and Chaudhuri (1996). Section 3 looks at qq-plots for assessing bivariate normality, and Section 5 uses qq-plots to compare bivariate samples. We generally let the data choose the quantiles of interest. Arrow plots connect the quantiles of the two distributions by drawing an arrow from the q_l^{th} quantile of one distribution to the q_l^{th} quantile of the other. These plots are effective at revealing bivariate location, scale, skewness, and outlier characteristics. In Section 4 we consider the case in which one has a single sample but no particular reference distribution to compare it to. We take a systematic set of quantiles, and connect them with line segments, creating a so-called *spider web plot.* These plots simplify the scatter plot while keeping the main features intact.

The procedures in this paper are rotationally equivariant but not fully affine equivariant. Chakraborti (1997) has developed affine-equivariant versions of these procedures. Liu, Parelius, and Singh (1997) survey various measures of data depth, and introduce several graphical approaches for assessing and comparing multivariate distributions based on these depths. Some of the latter techniques are akin to those in this paper, in particular their DD plots perform a similar role as our qq-plots, and their sunburst plots are analogous to our spider web plots. These procedures are also fully affine invariant.

Friedman and Rafsky (1981) propose a different but very interesting and novel approach to qq-plots for multidimensional data. They use minimal spanning trees fit to the individual samples, and rank the points within each sample one-dimensionally according to their locations on the tree. The usual one-dimensional qq-plots can then be used.

2. Multivariate Quantiles

For a univariate distribution, quantiles are typically defined to range from 0 to 1. We will shift the definition of quantile slightly, so that q has range $(-1, 1)$, and call the new quantiles g-quantiles (for geometric) in order to emphasize the difference. The q^{th} g-quantile is the value η_q such that

$$r(\eta_q) \equiv E_F\left(SIGN(\eta_q - X)\right) = q, \tag{1}$$

where $X \sim F$ and $SIGN$ is the sign function, $SIGN(z) = -1, 0, 1$ as $z < 0, = 0, > 0$. For given q, the g-quantile may not exist or be unique. Note that $r(\eta_q) = P(\eta_q > X) - P(\eta_q < X) = 2F(\eta_q) - 1$ if F is continuous at η_q, so that the q^{th} g-quantile equals the $[(1+q)/2]^{th}$ quantile under the usual definition. In particular, η_0 is the median. One can interpret the q in (1) as being the average direction one must go to move from X to η_q, averaging over the X's. For p-dimensional data, Chaudhuri (1996) and Koltchinskii (1997) have made a thorough examination of geometric quantiles. They are indexed by q in the interior of the p-dimensional unit disk, $\mathbb{Q}^p \equiv \{q \in \mathbb{R}^p \mid \|q\| < 1\}$. The q^{th} geometric quantile is the point η_q in $\mathbb{R}^p$ for which the average unit vector pointing from X to η_q is q, again averaging over the X's. Thus the definition is the same as (1) where for a vector z the sign is given by $SIGN(z) = z/\|z\|$ if $z \neq 0$, and 0 if $z = 0$. The equation may not be satisfied exactly. However, the precise definition is that η_q is the minimizer over η of

$$E_F\left(\|\eta - X\| - (\eta - X)'q\right), \tag{2}$$

which does satisfy (1) as long as the function in (2) is differentiable at η_q. This definition is similar to that of regression quantiles in Koenker and Bassett (1978). An interesting and fortunate property is that unless F concentrates all its mass along a single straight line, for each q, η_q exists and is unique.

The discussion above is based on a distribution function F, but the definitions also work for a sample $x_1, \ldots, x_n$ of p-dimensional vectors, where F is replaced by the empirical distribution function. Thus the sample q^{th} g-quantile $\widehat{\eta}_q$ minimizes $(1/n) \sum (\|\eta - x_i\| - (\eta - x_i)'q)$ over η, and (at least approximately) satisfies

$$r(\widehat{\eta}_q) = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{\eta}_q - x_i}{\|\widehat{\eta}_q - x_i\|} = q. \tag{3}$$

The r in this equation and equation (1) denotes "rank". It is the multivariate rank function that corresponds to the geometric quantiles we are using. See Koltchinskii (1997), Chaudhuri (1996), and Möttönen, Oja, and Tienari (1997) about this. In the multivariate case, $\widehat{\eta}_0$ is the geometric or *spatial* median. See Small (1990) for a comprehensive discussion of this and other multivariate medians. An important property of the r function is its uniqueness, that is, the values $r(x)$ for all x uniquely determine the distribution function F. See Koltchinskii (1997) in this regard.

QQ-plots are used to compare two distributions F_1 and F_2. It may be that both are population distributions, or one is a population distribution and one is an empirical distribution based on a sample, or both are empirical distributions. In any case, we choose a set of indices $q_1, \ldots, q_L$ in $\mathbb{Q}^p$, and find the corresponding g-quantiles for the two distributions:

$$\eta_1^{(1)}, \ldots, \eta_L^{(1)}, \text{ and } \eta_1^{(2)}, \ldots, \eta_L^{(2)}, \tag{4}$$

where $\eta_l^{(k)}$ is the q_l^{th} g-quantile for F_k. If the two distributions are close, then their corresponding g-quantiles should be close, $\eta_l^{(1)} \approx \eta_l^{(2)}$. The way in which the corresponding g-quantiles differ can give insight into the differences between the distributions. We visualize the comparisons for the $p = 2$ case by drawing an arrow from $\eta_l^{(1)}$ to $\eta_l^{(2)}$ for each l. Such plots we call bivariate qq-plots. Sections 3 and 5 contain examples of their use.

3. Bivariate Normal Plots

Given a sample $y_1, \ldots, y_n$ of bivariate observations, we would like to determine how close it is to a bivariate normal distribution. The reference distribution in this case is $F_1 = N(0, I_2)$, the bivariate normal distribution with mean zero and covariance matrix the 2×2 identity. The second distribution F_2 is based on the data, but we take a linear transformation of the y_i's so that the observations are centered and scaled to conform with F_1. We could subtract the sample mean and multiply by the inverse square root of the sample covariance matrix, but instead we use a procedure that is a bit more robust. The exact details of this are given in Section 1 of the Appendix. For each i, let x_i denote the transformed observation y_i, and F_2 be the empirical distribution function of the x_i's. We wish to compare each x_i with what would be its value if the sample actually were normal. In one dimension, this is accomplished by taking the rank of x_i, say j, and finding the usual $[j/(n+1)]^{th}$ (or something similar) quantile of the standard normal. The procedure for the bivariate case is the same. The g-quantiles of interest are the bivariate ranks $r(x_i)$ as in (3), so that $L = n$ and $q_l = r(x_l)$ for $l = 1, \ldots, n$. The g-quantiles for F_1 are thus the $\eta_l^{(1)}$ that satisfy (1) for each q_l, and the g-quantiles for F_2 are simply the sample observations: $\eta_l^{(2)} = x_l$. The qq-plot then draws an arrow from $\eta_l^{(1)}$ to x_l for each l. One can think of the arrow as pointing from where the observation "should" be if the sample were bivariate normal to where the observation actually is. Section 2 of the Appendix shows how to calculate the g-quantiles of the spherical normal distribution. The next examples exhibit bivariate normal qq-plots.

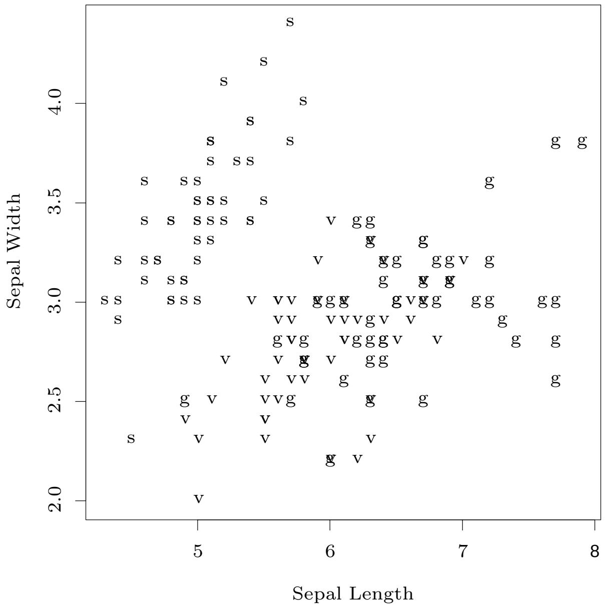
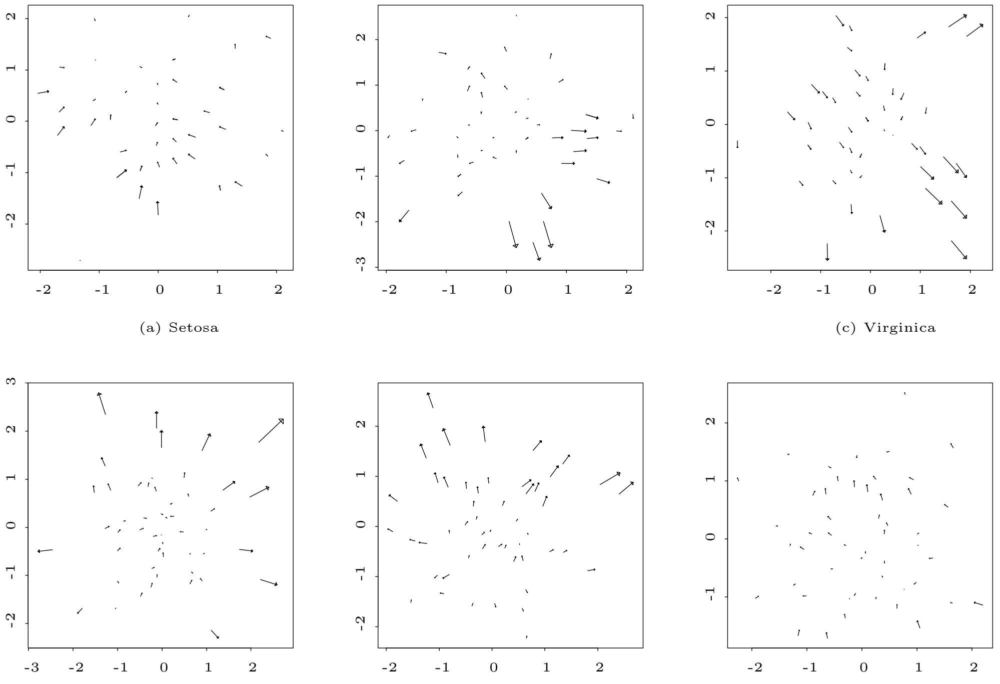


Figure 1. Iris data. s = Setosa, v = Versicolor, g = Virginica.

Example 3.1. Fisher's iris data

Fisher (1936) analyzed data on three species of iris, Setosa, Versicolor, and Virginica. There are 50 specimens from each species, and four variables measured on each. We will focus on two variables: Sepal Length and Sepal Width, which are measured in centimeters. Figure 1 contains a scatter plot of the data, where “s” denotes Setosa, “v” denotes Versicolor, and “g” denotes Virginica. For each species, we take the linear transformation of the bivariate data as described in the Appendix, Section A.1. Figure 2 (a), (b), and (c) have the bivariate normal qq-plots for the individual samples. For Setosa, the arrows are all quite small, revealing nothing to lead one to suspect that the data is not bivariate normal. Versicolor has some longer arrows. The longest are four near the bottom. They are pointing down and slightly to the right, which means that they are farther from the bulk of the data than they should be in a normal sample. There is also a cluster of smaller arrows at the right, pointing rightward. For the Virginica sample, the arrows tend to be larger than for Versicolor. There are a couple of points that are outlying to the upper right, and several to the lower right. The arrows in the upper left are not very long, but are pointing to the lower right, which suggests a skewness towards the lower right.



(a) Setosa

(c) Virginica

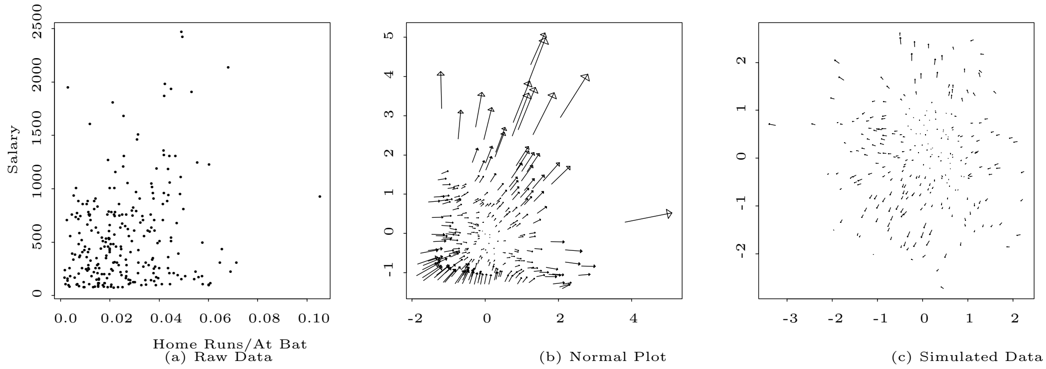
(e) Simulated

Figure 2. Normal plots for iris data

Figure 2 (e), (f), and (g) contain normal plots for randomly generated data from $N(0, I_2)$, each based on fifty points, to help calibrate what was seen in the other plots. Based on the lengths of the arrows in these three random plots, it appears that the Setosa sample does have small enough arrows to believe normality, while Versicolor and Virginica may be borderline too long. To give a rough numerical assessment of the arrow lengths, we calculated the average arrow length for the three iris plots, obtaining 0.092 for Setosa, 0.139 for Versicolor, and 0.169 for Virginica. For the three plots of simulated data, the average lengths are 0.140, 0.139, and 0.060. We then simulated 100 more sets of 50 bivariate normals, and calculated the proportion of samples for which the average arrow was larger than each of the iris species' average. These proportions are estimated p-values for testing bivariate normality. We obtained .85 for Setosa, .52 for Versicolor, and .24 for Virginica. Thus, at least using this measure, all three species have sepal dimensions that behave reasonably like a bivariate normal.

Example 3.2. Baseball data

This example uses the data set on major league baseball players discussed in American Statistical Association (1988). Plot (a) in Figure 3 graphs 263 players' career home runs per at bat through 1986 versus their 1987 salaries in thousands of dollars.



(a) Raw Data (b) Normal Plot (c) Simulated Data

Figure 3. Baseball data

Plot (b) is the bivariate normal plot. Plot (c) is a normal plot based on a set of 263 simulated observations from $N(0, I_2)$. There is no question the data are not bivariate normal. The points to the lower left are too close to the center, and the points to the upper right are too far, indicating skewness towards the upper right. There are several very large outliers. In order to see if the data can be transformed to a sample that is closer to bivariate normality, we look at the variablewise Box and Cox (1964) power transformations. That is, we choose powers (λ_1, λ_2), and assess the bivariate normality of the transformed data $(y_{i1}^{\lambda_1}, y_{i2}^{\lambda_2}), i = 1, \ldots, n$, where $y_i = (y_{i1}, y_{i2})'$, and the power 0 means logarithm.

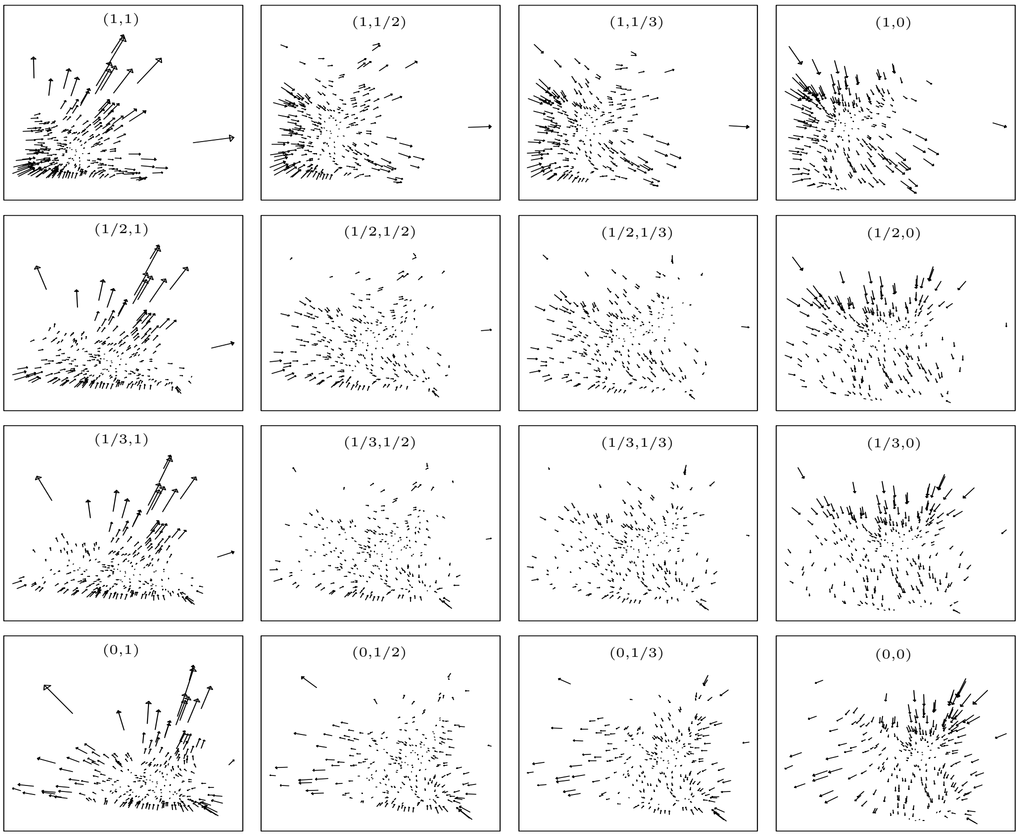


Figure 4. Baseball data – power transforms

Figure 4 contains the normal plots for λ_j's taking the values $1, 1/2, 1/3$, and 0. The parenthetical numbers in the plots are (λ_1, λ_2)'s. The table below contains the average lengths of the arrows for these plots. For each cell, the number in parentheses is the proportion among 100 normal plots using simulated sets of data (263 observations from $N(0, I_2)$ for each set) whose average arrow length exceeded that for that cell.

$\lambda_1 \downarrow$; $\lambda_2 \rightarrow$	1	1/2	1/3	0
1	0.247 (0)	0.161 (0)	0.155 (0)	0.174 (0)
1/2	0.175 (0)	0.081 (.10)	0.082 (.10)	0.121 (.01)
1/3	0.161 (0)	0.061 (.43)	0.066 (.29)	0.114 (.01)
0	0.185 (0)	0.084 (.07)	0.092 (.04)	0.139 (0)

The first column of the plots are those for which the first variable, Salary, is untransformed. In these plots, there are long arrows pointing upward, indicating a positive skewness in the vertical direction. As we go down that column, the arrows shift from pointing to the right to pointing to the left, showing that the transformations on the first variable, Home Runs per At Bat, are changing the horizontal skewness from positive to negative. The top row of plots shows the complementary pattern, where the horizontal skewness stays positive, but the arrows shift from pointing upwards to pointing downwards as the transformation on the second variable becomes stronger. Similar considerations hold for the other plots, where the log-log plot shows skewness towards the lower left.

It is clear that the variables should be transformed, but the log transformation is too strong. The best plot, both visually and according to the average arrow lengths, is that with $(\lambda_1, \lambda_2) = (1/3, 1/2)$. The estimated p-value for this choice is quite reasonable, .43. The transformation is effective in moving substantially towards normality.

4. Spider Web Plots

The qq-plots in the previous section were based on a reference distribution (bivariate normal) of interest. An alternative approach to graphing a single bivariate sample is to first choose a small, systematic set of bivariate g-quantiles, and connect the g-quantiles with line segments. Our goal is to choose enough g-quantiles to capture the essence of the data, but not so many that they are more complicated than the data. The actual g-quantiles we choose may depend on the data, because it is not very informative to take g-quantiles outside the range of the particular set of data. If the data happens to be approximately spherically symmetric, then it makes sense to take a reasonably symmetric set of g-quantiles. Thus we concentrate on the following g-quantiles q. Take a small number of radii, $0 < \rho^{(1)} < \cdots < \rho^{(K)} < 1$, and angles, $\theta^{(j)} = 2\pi j/L, j = 0, \ldots, L-1$. The corresponding set of g-quantiles consists of $\eta^{(0)}$, the median, plus

$$\eta^{(ij)} = \eta_{q^{(ij)}}, \text{ where } q^{(ij)} = \rho^{(i)}(\cos(\theta^{(j)}), \sin(\theta^{(j)}))',$$

for $i = 1, \ldots, K$ and $j = 0, \ldots, L-1$. We connect the points along the spokes, $\eta^{(0)} \rightarrow \eta^{(1j)} \rightarrow \cdots \rightarrow \eta^{(Kj)}$ for each j, as well as around, $\eta^{(i0)} \rightarrow \eta^{(i1)} \rightarrow \cdots \rightarrow \eta^{(i,L-1)} \rightarrow \eta^{(iL)}$ for each i. The result is the *spider web plot*.

Figure 5 exhibits three spider web plots for the Baseball data in Example 3.2. The plots have eight spokes and the radii .1, .2, ..., .8. The first plot is for the raw data. (We apply the linear transformation towards sphericity as in Section 3.) Skewness towards the upper right is evident. The second plot is for the best transformation found in Example 3.2, that is, $(\lambda_1, \lambda_2) = (1/3, 1/2)$. This spider web looks reasonably spherically symmetric. The third plot is the log-log plot, $(\lambda_1, \lambda_2) = (0, 0)$. There we see skewness towards the lower left. Thus these spider web plots smooth the scatter plots to show certain characteristics.

They have the advantage of not needing a specific distribution for comparison, although they are not as detailed as the QQ plots from the previous section.

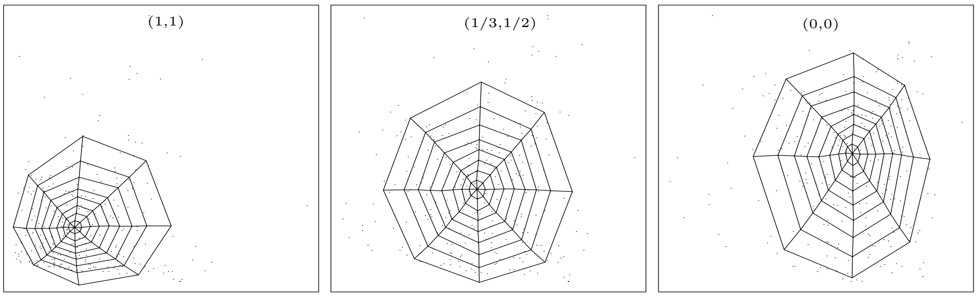
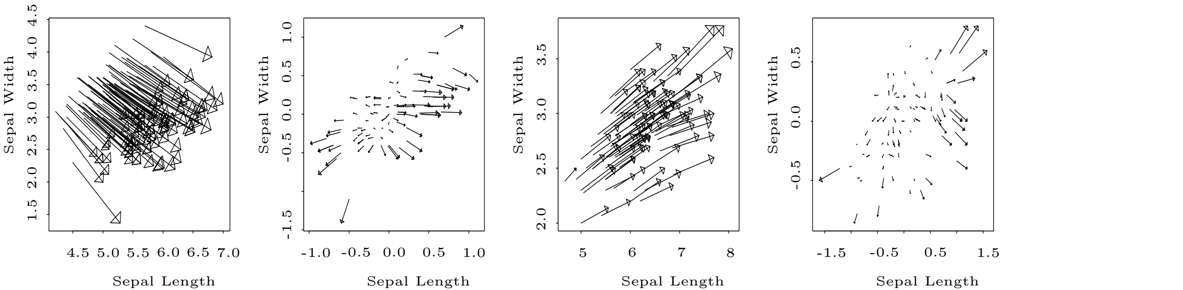


Figure 5. Baseball data – spider web plots for power transforms

5. Comparing Samples

We assume we have two samples, $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$. Now F_1 and F_2 are the empirical distribution functions for the two samples, respectively. To compare the samples we again choose a set of q's, then draw arrows from the q^{th} sample g-quantile among the x_i's to the q^{th} sample g-quantile among the y_i's. In the next example the q's we use consist of the ranks of all the observations, where for each observation the rank is calculated with respect to its own group. Thus each point will either be the base or the tip of one of the arrows. That is,

$$q_l = \frac{1}{m}\sum_{i=1}^{m}\frac{x_l - x_i}{\|x_l - x_i\|}, l = 1, \ldots, m, \text{ and } q_{m+l} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_l - y_i}{\|y_l - y_i\|}, l = 1, \ldots, n. \tag{5}$$



(a) Setosa→Versicolor (b) Setosa→Versicolor: Centered (c) Versicolor→Virginica (d) Versicolor→Virginica: Centered

Figure 6. Iris data – two-sample plots

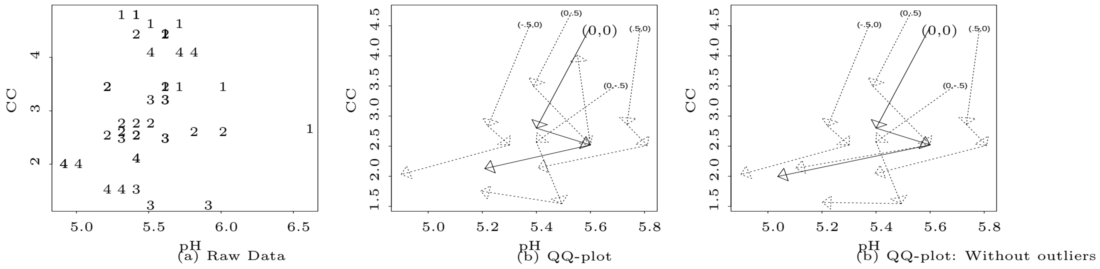
Example 5.1. Fisher's iris data

Plot (a) in Figure 6 is the qq-plot obtained by drawing arrows from the g-quantiles of the Setosa sample to the g-quantiles of the Versicolor sample. All the arrows are pointing towards the lower right, and they are generally of similar lengths and nearly parallel. Thus the Versicolor sample is close to a straight location shift of the Setosa sample, where Sepal Length is shifted up about one centimeter and Sepal Width is shifted down about the same. (The average of the arrows is (.92,-.65).) The location difference is so large in this plot, it may hide other potential differences between the samples. Thus we recalculated the qq-plot after centering each sample by subtracting the sample's spatial median from each observation. The result is Plot (b). The arrows are generally pointing away from the center, primarily in the directions from upper right to lower right to lower left. This pattern suggests that the Versicolor sample is more spread out than the Setosa sample, and more skewed towards the right.

Plots (c) and (d) are the corresponding two plots for comparing Versicolor to Virginica. Plot (c) shows primarily a shift towards the upper right. (The average arrow is (.66,.20).) However, the arrows emanating from the upper right are generally longer than those emanating from the lower left. Thus even in this plot there is evidence of scale differences. Plot (d) for the centered samples shows more clearly that Virginica is more spread out, the difference most noticeable towards the upper-right.

Example 5.2. Biomedical data

Smith, Gnanadesikan, and Hughes (1962) present a data set with thirteen variables measured on urine samples of 45 men. The men are classified into four groups, based on weight, with 12, 14, 9, and 10 men in the four groups, going from lightest (Group 1) to heaviest (Group 4) (see also Seber (1984)). Hettmansperger and Oja (1994) use the first two variables, pH level and modified creatinine coefficient (CC), to test the differences among the four weight groups using their median test. We will use the same two variables to create a four-way qq-plot. Plot (a) in Figure 7 shows the data, where the numbers indicate which weight group the point is from.



(a) Raw Data

(b) QQ-plot

(b) QQ-plot: Without outliers

Figure 7. Biomedical data

For the qq-plot, we find the $q_1, \ldots, q_L$ g-quantiles within each group, then draw arrows from the Group 1 g-quantiles to the Group 2 g-quantiles, then from Group 2 to Group 3, and from Group 3 to Group 4. Thus each quantile can be traced as we move through the groups. In order to prevent too many arrows being plotted, we take five fixed q's: $q_1 = (0,0)$ (which corresponds to the spatial median), $q_2 = (.5,0)$, $q_3 = (0,.5)$, $q_4 = (-.5,0)$, and $q_5 = (0,-.5)$. These last four are arbitrary but are chosen to range over the possible values of q without being too numerous. Plot (b) is the qq-plot. The solid line connects the four spatial medians, $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$. The dotted lines correspond to the other g-quantiles. There is a clear pattern. Group 2 is down and to the left of Group 1; Group 3 is slightly down and to the right of Group 2; then Group 4 is down and to the left of Group 3. There is one glitch, however. Unlike the other g-quantiles, the $(0,.5)^{th}$ g-quantile from Group 3 to Group 4 goes almost straight up.

Looking at Plot (a), we see that three of the Group 4 observations have a high value of CC, 4.12, while the other seven values for this group are either 2.14 or 2.03. Setting aside those three with high values, we obtain the qq-plot in Plot (c). This plot is like Plot (b), but now all five g-quantiles follow a similar route from Group 1 to Group 4.

6. Conclusion

The g-quantiles are location and rotation equivariant, and scale equivariant if the same factor is applied to all variables. That is, for any $a \neq 0$, orthogonal matrix Γ, and vector b, if we transform all the data points by $x \rightarrow ax\Gamma + b$, then all g-quantiles are similarly transformed, $\eta_q \rightarrow a\eta_q\Gamma + b$. The main drawback is that the g-quantiles are not affine equivariant, that is, we cannot replace Γ with an arbitrary nonsingular matrix. For the normal plots in Section 3, the scaling was dictated by the reference distribution, hence there was no ambiguity in the goal for transforming the original data, although there are many reasonable methods for estimating the transformation. The spider web plots of Section 4 and the between-sample qq-plots of Section 5 are dependent on the relative scaling of the variables, much as relative scaling matters in principle components. In the examples we gave the variables were transformed to have similar scales, or already had similar scales, so that scaling was not a problem, but in general one may wish to adjust the scales of the variables before proceeding to the g-quantiles.

We hope that the examples show the potential of these bivariate qq-plots. The directions, placements, and lengths of the arrows are effective in revealing location shifts, scaling differences, skewness, and outliers. The plots can also lead to inference procedures, such as tests for bivariate normality based on arrow lengths as in Section 3, as well as estimates of location shift as in Example 5.1.

Acknowledgements

This work was supported in part by National Science Foundation grants DMS 93-04244 and DMS 93-04525. I thank the referees and associate editor for many fine comments that led to a much more focussed presentation.

Appendix

A.1. Transforming the variables

Start with the data $y_1, \ldots, y_n$, where $y_i = (y_{i1}, y_{i2})'$. We first regress the second variable from the first using the Sen-Theil estimator of the regression slope. This slope b is the median of the pairwise slopes between the points in the sample:

$$b = median\Big\{\frac{y_{i2} - y_{j2}}{y_{i1} - y_{j1}} \mid 1 \leq i < j \leq n\Big\}, \tag{A.1}$$

where pairs for which $y_{i1} = y_{j1}$ are ignored. Let $w_i = (y_{i1}, y_{i2} - by_{i1})'$ for each i. Next we scale each variable in w so that it matches the univariate standard normal's in terms of the difference between two selected univariate quantiles t_1 and t_2. The quantiles we select are the γ_1^{th} and γ_2^{th}, where $\gamma_1 = [(1-\alpha)/2]$ and $\gamma_2 = [(1+\alpha)/2]$ for some α between 0 and 1. With these choices, the proportion of univariate observations between t_1 and t_2 is about α for each variable, and if the two variables are independent, the proportion of bivariate observations for which both variables are between t_1 and t_2 is about α^2. We will take $\alpha = 1/\sqrt{2}$, so that about half the observations have both variables between t_1 and t_2. For the normal, $t_2 = -t_1 = 1.052$. Let t_{1j} and t_{2j} be the γ_1^{th} and γ_2^{th} sample quantiles, respectively, of the observations $y_{1j}, \ldots, y_{nj}$ on the j^{th} variable. Then scale the w_i's so that these quantiles are as for the normal, i.e., let $v_i = 2 \times 1.052 \times (w_{i1}/(t_{21} - t_{11}), w_{i2}/(t_{22} - t_{12}))'$. Finally, center the v_i's by subtracting their bivariate median, η_0. The results are the x_i's. To summarize, $x_i = Ay_i - \eta_0$, where

$$A = \begin{pmatrix} 2 \times 1.052/(t_{21} - t_{11}) & 0 \\ 0 & 2 \times 1.052/(t_{22} - t_{12}) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -b & 1 \end{pmatrix}. \tag{A.2}$$

A.2. Finding the ranks and quantiles of the spherical normal

Suppose $F = N(0, I_p)$ where $p \geq 2$. For $\eta \in \mathbb{R}^p$, in (1), $r(\eta) = SIGN(\eta) f(\|\eta\|)$, where

$$f(\delta) = \frac{\delta}{\sqrt{2}} e^{-\delta^2/2} \frac{\Gamma(\frac{p}{2} + \frac{1}{2})}{\Gamma(\frac{p}{2} + 1)} \, {}_1F_1\left(\frac{p}{2} + \frac{1}{2}; \frac{p}{2} + 1; \frac{\delta^2}{2}\right), \tag{A.3}$$

and ${}_1F_1(a; b; z)$ is the confluent hypergeometric function, ${}_1F_1(a; b; z) = \sum_{k=0}^{\infty} c_k z^k / k!$ with $c_k = \Gamma(a+k)\Gamma(b)/(\Gamma(a)\Gamma(b+k))$. (See Abramowitz and Stegun (1972) for further information about the confluent hypergeometric function. Möttönen, Oja, and Tienari (1997) find the function r for the multivariate Student's t family as well as the normal. They refer also to Möttönen and Oja (1994).)

Proof. Let $X \sim N(0, I_p)$, so that $r(\eta) = E((\eta - X)/\|\eta - X\|)$. If $\eta = 0$, then the symmetry in the normal shows that $r(0) = 0$. Suppose $\eta \neq 0$, and let G be a $p \times p$ orthogonal matrix whose first column is $SIGN(\eta)$, so that $G'\eta = \|\eta\|(1, 0, \ldots, 0)' \equiv \eta^*$. Because GX is also $N(0, I_p)$, $r(\eta) = Gr(G'\eta) = Gr(\eta^*) = G\gamma(1, 0, \ldots, 0)' = \gamma SIGN(\eta)$, where

$$\gamma = E\Big(\frac{\|\eta\| - X_1}{((\|\eta\| - X_1)^2 + X_2^2 + \cdots + X_p^2)^{1/2}}\Big). \tag{A.4}$$

Note that $T \equiv (\|\eta\| - X_1)/((X_2^2 + \cdots + X_p^2)/(p-1))^{1/2}$ is a noncentral Student's t variable with noncentrality parameter $\|\eta\|$ and degrees of freedom $p - 1$. Thus $\gamma = E(T/(T^2 + p - 1)^{1/2})$. Straightforward but involved calculations will show that $\gamma = f(\|\eta\|)$, which proves the result.

The g-quantile η_q of the spherical normal for a specified $q \in \mathbb{Q}^p$ is then given by $\eta_q = SIGN(q) f^{-1}(\|q\|)$. The function f in (A.3) can be inverted using Newton's method.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions.* Dover, New York.

American Statistical Association (1988). *ASA Proceedings of Statistical Graphics Section* 76-137. The data can be found in StatLib at *http://lib.stat.cmu.edu/ datasets/baseball.data* and *http://lib.stat.cmu.edu/datasets/baseball.corr*

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Stat. Soc. Ser. B* **26**, 211-252.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **91**, 862-872.

Chakraborti, B. (1997). On affine equivariant multivariate quantiles. Technical Report 17/97, Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, Calcutta.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179-188.

Friedman, J. H. and Rafsky, L. C. (1981). Graphics for the multivariate two-sample problem. *J. Amer. Statist. Assoc.* **76**, 277-295.

Hettmansperger, T. P. and Oja, H. (1994). Affine invariant multivariate multisample sign tests. *J. Roy. Statist. Soc. Ser. B* **56**, 235-249.

Koltchinskii, V. (1997). *M*-estimation, convexity and quantiles. *Ann. Statist.* **25**, 435-477.

Liu, R., Parelius, J. M. and Singh, K. (1997). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. Technical Report, Rutgers University.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

Möttönen, J. and Oja, H. (1994). Efficiency of the spatial signed-rank test. Unpublished manuscript.

Möttönen, J., Oja, H. and Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Ann. Statist.* **25**, 542-552.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.* **63**, 1379-1389.

Small, C. G. (1990). A survey of multidimensional medians. *Internat. Statist. Rev.* **58**, 263-277.

Smith, H., Gnanadesikan, R. and Hughes, J. B. (1962). Multivariate analysis of variance (MANOVA). *Biometrics* **18**, 22-41.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis I. *Nederl. Akad. Wetensch. Proc.* 386-392.

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street #101, Champaign, IL 61820, U.S.A.

E-mail: marden@stat.uiuc.edu

(Received September 1996; accepted March 1998)