Statistical Science
2004, Vol. 19, No. 4, 606–614
DOI 10.1214/088342304000000512

Positions and QQ Plots

John I. Marden

Abstract. Quantile–quantile (QQ) plots for comparing two distributions are constructed by matching like-positioned values (i.e., quantiles) in the two distributions. These plots can reveal outliers, differences in location and scale, and other differences between the distributions. A particularly useful application is comparing residuals from an estimated linear model to the normal. A robust estimate, such as the Sen–Theil estimate, of the regression line is important. Extensions to two-dimensional QQ plots are presented, relying on a particular notion of multivariate position.

Key words and phrases: Quantile, QQ plot, spatial rank, spatial quantile, Sen–Theil estimate.

1. INTRODUCTION

The objective of this paper is to present some simple nonparametric graphical methods for comparing a sample to a theoretical distribution or for comparing two samples. We concentrate on the quantile–quantile (QQ) plot, which, when comparing two distributions, matches quantiles of one with the same quantiles of the other. These plots are good at revealing location and scale differences, as well as identifying outliers; see Section 3.

The QQ plots are often used to compare residuals to the normal distribution in linear regression. To find a good estimate of the residuals, we need a robust estimate of the regression line. We use the Sen–Theil estimate for this task (presented in Section 3.2).

We end (Section 5) by extending QQ plots to multivariate data. Throughout, we use the notion of "position" of a point among other points, which we define in Section 2 in the univariate case to be a centering and scaling of the usual ranks of the data. Section 4 extends positioning to multivariate data.

2. UNIVARIATE POSITIONS

One concept that will be useful throughout this paper is the position of a point z relative to a set of points or relative to a distribution F. Consider a sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ of points on the real line and another point z. Imagine that you are standing at the point z. Imagine also that each of the observations is equipped with an arrow of length 1 pointing toward you, so that the observations to the left are pointing up $(+1)$ and the observations to the right are pointing down (-1). Your position among the data is then the average of those arrows. The formal definition follows.

DEFINITION 1. The position of $z \in \mathbb{R}$ relative to the sample $\mathbf{X}_n$ is

$$\text{pos}(z; \mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \text{Sign}(z - x_i), \tag{1}$$

where $\text{Sign}(w)$ is the sign of the number w, that is,

$$\text{Sign}(w) = \begin{cases} -1, & \text{if } w < 0, \\ 0, & \text{if } w = 0, \\ 1, & \text{if } w > 0. \end{cases}$$

If F is a distribution function, then the position of z relative to F is

$$\text{pos}(z; F) = E[\text{Sign}(z - X)], \quad \text{where } X \sim F. \tag{2}$$

Note that $\text{pos}(z; \mathbf{X}_n) = \text{pos}(z; \widehat{F}_n)$, where $\widehat{F}_n$ is the empirical distribution function of the x_i's. The summation in (1) counts $+1$ for each point z is above and -1 for each point z is below, so $\text{pos}(z; \mathbf{X}_n)$ is the proportion z is above minus the proportion z is below:

$$\text{pos}(z; \mathbf{X}_n) = \frac{\#\{x_i | z > x_i\} - \#\{x_i | z < x_i\}}{n}.$$

The $\text{pos}(x; F)$ is similar, replacing proportions with probabilities:

$$\text{pos}(z; F) = P[z > X] - P[z < X].$$

John I. Marden is Professor, Department of Statistics, University of Illinois at Urbana–Champaign, Champaign, Illinois 61820, USA (e-mail: jimarden@uiuc.edu).

In either case, $-1 \le \text{pos} \le 1$, where the larger z is, the closer its position is to $+1$. The median has position 0; values above the maximum have position $+1$ and values below the minimum have position -1.

The positions of the x_i's themselves are centered and scaled versions of their ranks. That is, $\text{Rank}(x_j) = k$ if x_j is the kth smallest of the x_i's and the ranks are related to the positions via

$$\text{pos}(x_j; \mathbf{X}_n) = \frac{2 \cdot \text{Rank}(x_j) - n - 1}{n}. \tag{3}$$

If there are ties among the x_i's, then (3) holds for Rank being the midrank. The analogous formula that relates position relative to F and F itself is

$$\text{pos}(z; F) = 2 \cdot F(z) - 1. \tag{4}$$

3. UNIVARIATE QQ PLOTS

We wish to compare a sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ to a theoretical distribution F or compare one sample to another. (One could also compare two theoretical distributions, but we do not specifically address this issue.) We present the QQ plot, which compares like-positioned elements from the two distributions. For example, to compare the sample $\mathbf{X}_n$ to a theoretical distribution F, we could compare each x_j to the value η_j, where η_j has the same position relative to F that x_j has relative to the sample $\mathbf{X}_n$. That is,

$$\text{pos}(x_j; \mathbf{X}_n) = \text{pos}(\eta_j; F). \tag{5}$$

Using (4), and assuming that F is continuous and strictly increasing, the η_j in (5) is

$$\eta_j = F^{-1}\left(\frac{\text{pos}(x_j; \mathbf{X}_n) + 1}{2}\right),$$

which means that η_j is the q_jth quantile of F for $q_j = (\text{pos}(x_j; \mathbf{X}_n) + 1)/2$. Also, x_j is the q_jth quantile of the sample. The *QQ plot* plots the η_j's versus the x_j's. If the sample is a reasonable facsimile of the distribution F, then the $x_j \approx \eta_j$, so that the points should be close to the line $x = \eta$.

To illustrate, we use data on $n = 136$ students who were asked the fastest speed they had ever driven a car. Figure 1 is a stem-and-leaf plot of the data. The slowest speed was 50 mph and the fastest was 150 mph. Notice most of the values are rounded to a 5. The distribution looks a little skewed to the faster speeds and there is a small bump around 140–150 mph.

We wish to compare the sample to a normal distribution. One should first decide on which normal distribution, because it is clear that we do not expect these data to be $N(0, 1)$. Because we are looking at a location–scale family, we can use the fact that if ξ is the qth quantile of a $N(0, 1)$, then $\eta = \mu + \sigma\xi$ is the qth quantile of a $N(\mu, \sigma^2)$. Thus if the data are like a normal sample, the observations should lie approximately along a straight line, $x_j \approx \mu + \sigma\xi_j$, where ξ_j is the q_jth quantile of a $N(0, 1)$.

```
 5 | 0
 6 |
 7 | 55555
 8 | 000000000004555555555555556
 9 | 00000000000000000000000000014555555555555555555
10 | 00000000000000035555
11 | 00000000000002555
12 | 00000000005
13 | 5
14 | 04557
15 | 0
```

FIG. 1. *Stem-and-leaf plot of* 136 *students' responses to the question, "What is the fastest you have ever driven a car?" (The decimal point is* 1 *digit(s) to the right of the* |.)

Table 1 shows the calculations for four selected observations, $x_j = 50, 90, 100, 150$. The second column has their positions, the third has the corresponding q_j and the fourth has the q_jth quantiles for the $N(0, 1)$. Figure 2 plots the ξ_j versus the x_j's for the entire data set. The straight line in the plot is $x = \widehat{\mu} + \widehat{\sigma}\xi$, where $\widehat{\mu} = \text{median}(\mathbf{X}_n) = 95$ and $\widehat{\sigma} = \text{mad}(\mathbf{X}_n) = 14.826$. The mad is the median absolute deviation, defined as $\text{mad}(\mathbf{X}_n) = 1.4826 \cdot \text{median}|x_i - \widehat{\mu}|$. The constant 1.4826 ensures that if the data are normal, then $\widehat{\sigma}$ is a consistent estimator of σ. One could also estimate the parameters by fitting a straight line to the QQ plot; see Barnett (1975). The last column in Table 1 contains the estimates $\widehat{\eta}_j = \widehat{\mu} + \widehat{\sigma}\xi_j$. If the data are approximately normal, then the x_j's should be close to their respective $\widehat{\eta}_j$'s. The first three are fairly close, but the fourth is not as close.

TABLE 1

x_j	$\text{pos}(x_j; \mathbf{X}_n)$	q_j	ξ_j	$\widehat{\eta}_j$
50	−0.993	0.004	−2.680	55.26
90	−0.316	0.342	−0.407	88.96
100	0.301	0.651	0.387	100.74
150	0.993	0.996	2.680	134.74

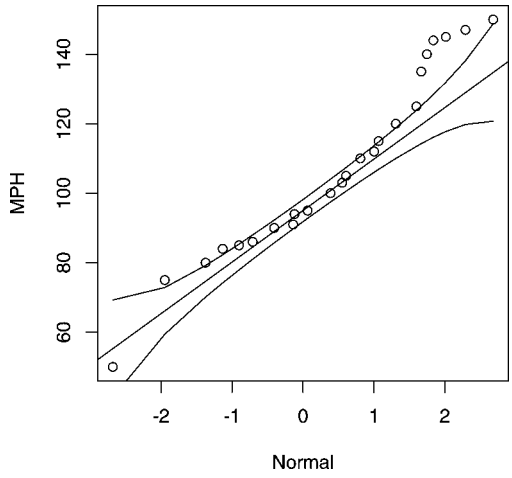

FIG. 2. *The QQ plot that compares the driving speed data with the* $N(0,1)$. *The curved lines indicate the estimates line* $\pm 2 \cdot$ (*standard error*).

We have added approximate error bars in Figure 2 as well. If the data are a sample from the distribution F, and if $x_{\lceil nq \rceil}$ (for a real number z, $\lceil z \rceil$ is the smallest integer greater than or equal to z) and η_q are the qth sample and theoretical quantiles, respectively, then as $n \to \infty$,

$$\sqrt{n}(x_{\lceil nq \rceil} - \eta_q) \to N\left(0, \frac{q(1-q)}{f(\eta_q)^2}\right), \tag{6}$$

where f is the density for F, and we assume that f is continuous and positive at η_q. The two curved lines in the figure are plots of $\widehat{\mu} + \widehat{\sigma}\xi_j \pm 2 \cdot \mathrm{se}_j$, where

$$\mathrm{se}_j = \sqrt{\frac{q_j(1-q_j)}{n}} \frac{\widehat{\sigma}}{\phi(\xi_j)}$$

and ϕ is the $N(0,1)$ density.

Consider the plot. The points generally follow the line fairly well, except for the upper tail, where there are seven or so points that are substantially larger than the line. Thus it appears that the data deviate from the normal by having too many very large values around 140–150. The rest of the points are basically between the error bars, although there is some curvature in the middle.

In addition to being good at identifying outliers and heavy tails, QQ plots can reveal characteristics such as skewness and bimodality, and can be effective even for small samples. Wilk and Gnanadesikan (1968) gave a thorough introduction to QQ plots and Cleveland (1993) provided numerous examples of their use.

3.1 Comparing Two Samples

Consider the driving speed data again, but now separate the values by sex. There are $n_1 = 100$ women and $n_2 = 36$ men. Denote these subsamples by $\mathbf{X}_{n_1}^{(1)} = \{x_1^{(1)}, \ldots, x_{n_1}^{(1)}\}$ for the women and $\mathbf{X}_{n_2}^{(2)} = \{x_1^{(2)}, \ldots, x_{n_2}^{(2)}\}$ for the men. A QQ plot chooses some values for the men, say $\eta_1^{(2)}, \ldots, \eta_K^{(2)}$, and then finds the values $\eta_j^{(1)}$ which have the same position relative to the women that the $\eta_j^{(2)}$ have relative to the men, that is,

$$\mathrm{pos}(\eta_j^{(1)}; \mathbf{X}_{n_1}^{(1)}) = \mathrm{pos}(\eta_j^{(2)}; \mathbf{X}_{n_2}^{(2)}), \quad j = 1, \ldots, K. \tag{7}$$

Because of the lack of continuity of the pos function for a sample, it is likely that (7) cannot be solved exactly, so we use linear interpolation. The choice of the $\eta_j^{(2)}$'s is arbitrary, but typically one chooses the data so that $K = n_2$ and $\eta_j^{(2)} = x_j^{(2)}$. We are also consciously using the smaller sample, because using more points than there are observations in the smaller sample seems shaky.

For each $x_j^{(2)}$, we find the position among the men, $\mathrm{pos}(x_j^{(2)}; \mathbf{X}_{n_2}^{(2)})$, and then find (approximately) the $\eta_j^{(1)}$ that solves $\mathrm{pos}(\eta_j^{(1)}; \mathbf{X}_{n_1}^{(1)}) = \mathrm{pos}(x_j^{(2)}; \mathbf{X}_{n_2}^{(2)})$. For example, $\mathrm{pos}(80; \mathbf{X}_{n_2}^{(2)}) = -0.9444 = \mathrm{pos}(68.98; \mathbf{X}_{n_1}^{(1)})$ and $\mathrm{pos}(120; \mathbf{X}_{n_2}^{(2)}) = 0.4444 = \mathrm{pos}(91.60; \mathbf{X}_{n_1}^{(1)})$; thus a man with 80 mph is in approximately the same position among men as a woman driving 68.98 mph is among women. Similarly, 120 among men is matched with 91.60 among women.

The QQ plot plots the $x_j^{(2)}$'s versus the $\eta_j^{(1)}$'s. Figure 3 has the plot, along with the line fit to the data using the Sen–Theil procedure introduced in Section 3.2,

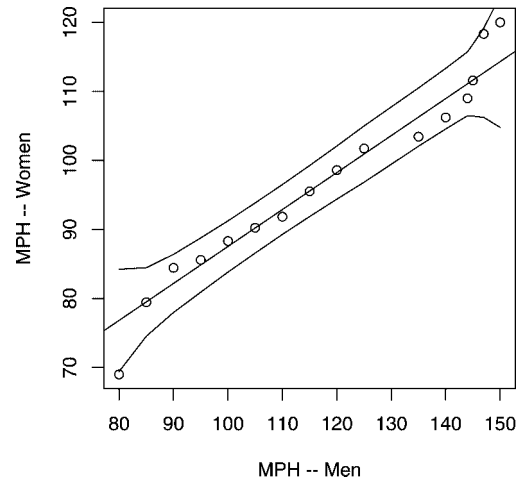

FIG. 3. *The QQ plot that compares the men to the women on their driving speeds.*

$\eta^{(1)} = 33.96 + 0.5356x^{(2)}$. The fact that this line is far from $\eta^{(1)} = x^{(2)}$ shows that the two samples are different in (at least) location and scale. If the samples are independent samples from distributions that differ by only location and scale (i.e., both distributions have the same shape, such as normal), then the QQ plot should be close to a straight line $\eta^{(1)} = a + bx^{(2)}$, where the slope b is the ratio of scales, and the difference in location can be measured by the differences in medians. For our data, the slope 0.5346 suggests that the women's distribution is about half as spread out as the men's. The sample median of the men is 110, so the estimate of the difference in medians, women minus men, is $a + (b - 1)110 = 33.96 - 0.4644(110) = -17.12$. The women (reportedly) drive much more slowly and their speeds are much less spread out than those of the men.

The closer the points are to the straight line, the more similar the shapes of the distributions. The points in the plot are fairly close to the line, suggesting that there is not much reason to believe the shapes of the distributions are much different, at least compared to the $\pm 2 \cdot$ (standard error) lines. The standard errors are calculated assuming the two distributions are normal. This assumption tends to be conservative in the tails, that is, the standard errors for the tail quantiles are larger than they would be for a heavier tailed distribution. We need to take into consideration uncertainty in both samples; hence, the standard errors are

$$\text{se}_j = \widehat{\sigma}_1 \frac{\sqrt{q_j(1-q_j)}}{\phi(z_j)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $\widehat{\sigma}_1$ is an estimate of the standard deviation of the women's distribution, which we take to be the $\text{mad}(\mathbf{X}^{(1)}_{n_1}) = 7.413$, $q_j = (\text{pos}(x_j) + 1)/2$ and z_j is the q_jth quantile of the $N(0, 1)$.

3.2 QQ Plots for Residuals

A common use for QQ plots is to check the residuals in linear regression to see if they are reasonably normal. The simple linear regression model that relates the explanatory variable x to the dependent variable y is

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n,$$

where α and β are the (unknown) intercept and slope parameters, respectively, and the e_i's are the residuals. We assume that the x_i's are fixed and the residuals e_i are independent and identically distributed, with $E[e_i] = 0$ and $\text{Var}[e_i] = \sigma^2$, and entertain the additional assumption that the e_i's are normally distributed.

To check the residuals, we first need estimates $\widehat{\alpha}$ and $\widehat{\beta}$ of the intercept and slope. Then the residual e_i is estimated by $\widehat{e}_i = y_i - \widehat{\alpha} - \widehat{\beta}x_i$. It is important to use fairly robust estimates of the parameters; otherwise the estimated residuals may not adequately reflect the structure of the e_i's. In keeping with the nonparametric approach, we base estimation of the slope β on the pairwise slopes, that is, the slopes between pairs of points:

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} \quad \text{if } x_i \neq x_j.$$

Because $b_{ij} = b_{ji}$, we look at just the slopes b_{ij} for $i < j$. The *Sen–Theil* estimate of the slope is the median of these pairwise slopes. See Theil (1950) and Sen (1968) for development of this estimator and an associated confidence interval procedure. The estimate of the intercept we use is the median of the $y_i - \widehat{\beta}x_i$'s:

$$\widehat{\beta} = \text{median}\{b_{ij}\} \quad \text{and} \quad \widehat{\alpha} = \text{median}\{y_i - \widehat{\beta}x_i\}.$$

To illustrate, we use data from Rousseeuw and Leroy (1987) on the average body weight (in kilograms) and brain weight (in grams) for 28 animal species. Figure 4 plots the logs of the data. Here there are $\binom{28}{2} = 378$ pairwise slopes, with median $\widehat{\beta} = 0.6739$. The $\widehat{\alpha} = 2.2827$. We also find the least squares estimate of (α, β), which is the (a, b) that minimizes $\sum(y_i - a - bx_i)^2$, in this case being $(2.5549, 0.4960)$. The figure indicates the Sen–Theil and least squares estimates of the line. Note the three points to the lower right. They are dinosaurs, who have relatively small brains for their body size. One can see how these points

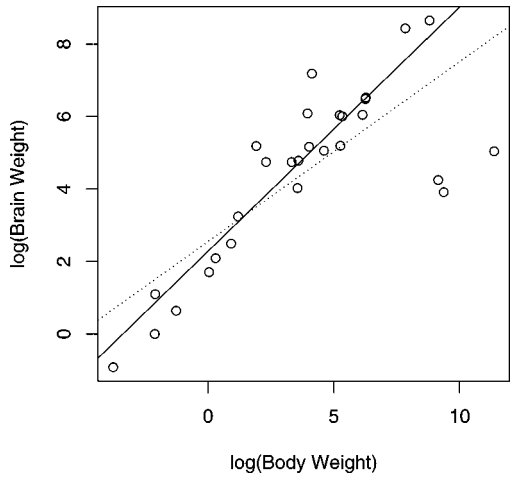

FIG. 4. *The scatterplot of the animals data. The solid line is the Sen–Theil estimate of the regression line; the dotted line is least squares.*

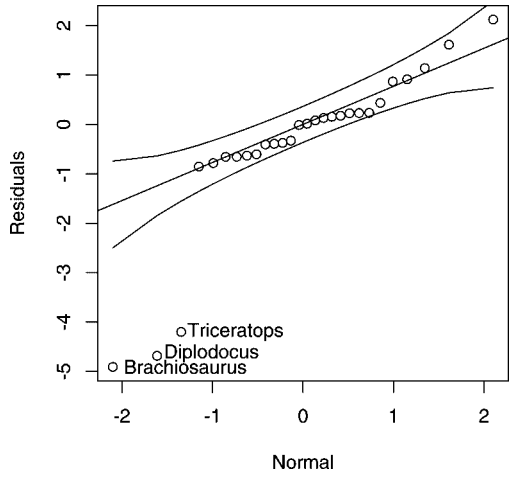

FIG. 5. *The QQ plot of the residuals from the body and brain weight data.*

pull the least squares line toward them, relative to the Sen–Theil line.

Figure 5 shows the QQ plot of the residuals from the Sen–Theil fit, comparing the residual to the normal distribution. We can see clearly the three outliers. The plot from the least squares line also shows these three, though not quite as vividly.

Next we remove the three dinosaurs from the data set, leaving us with mammals. Figure 6 shows the QQ plot from the resulting Sen–Theil fit, $1.999 + 0.7514x_i$. Now there are four outliers, all primates, suggesting that they have relatively large brains for their bodies. There is one more primate, the gorilla, that does not appear as an outlier. We note that without the dinosaurs, the Sen–Theil and least squares lines are very similar.

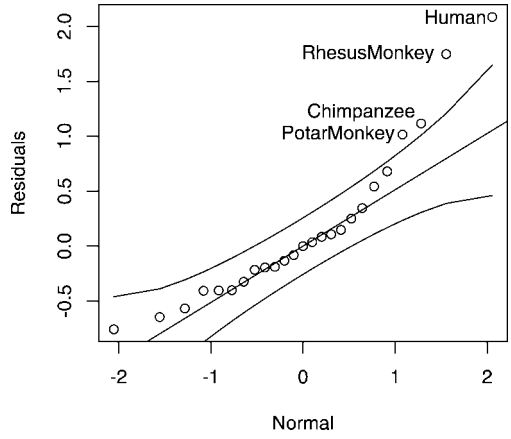

FIG. 6. *The QQ plot of the residuals from the body and brain weight data without the dinosaurs.*

It turns out that the least squares estimate of β can be written as a weighted average of the b_{ij}'s,

$$\widehat{\beta}_{\mathrm{LS}} = \frac{\sum\sum_{1\le i<j\le n} b_{ij}(x_i - x_j)^2}{\sum\sum_{1\le i<j\le n}(x_i - x_j)^2}.$$

(To be precise, the summations are over only those i, j with $x_i \neq x_j$.) The slopes are heavily weighted if their corresponding x_i's are far apart. Thus the least squares estimate is sensitive to outliers in y_i's that are associated with x_i's at the extremes, leading to lack of robustness. Another estimator more robust than least squares is the least absolute deviation estimate, which chooses the (a, b) to minimize $\sum |y_i - a - bx_i|$. It is interesting that at least one solution to the minimization occurs at $b = b_{ij}$ for some (i, j), and $a = \text{median}\{y_i - bx_i\}$.

In simple linear regression, it is usually not too hard to spy outliers on the original plot. The use of robust estimators of the parameters coupled with QQ plots of the residuals is particularly useful in multiple regression. See Rousseeuw and Leroy (1987) and Cook and Weisberg (1997) for general regression graphics.

4. BIVARIATE POSITIONS

So far, comparisons between distributions have been based on one variable. The QQ plots can be extended to multivariate data. For example, the data set on the students' driving speed includes other variables such as height, weight and age. One may wish to compare the two distributions (male and female) on height and weight simultaneously. The idea is the same as in the univariate example: We match up each male, say, with the like-positioned female. First, we have to define position in a bivariate sense. There are many approaches, but we present just one, which is fairly simple. See Liu, Parelius and Singh (1999) and Rousseeuw, Ruts and Tukey (1999) for other multivariate graphical techniques. Serfling (2002) reviewed possible multivariate quantiles and gave some criticism of the one we use.

A bivariate analog of the univariate position is to imagine you are at point z in the plane in the midst of some bivariate data. For each observation, you look at the arrow (of length 1) pointing from the observation toward you. Your position is then the average of these arrows. To illustrate, Figure 7 shows the plot of 37 men's performances on written assignments and exams in a statistics course. (These are different stu-

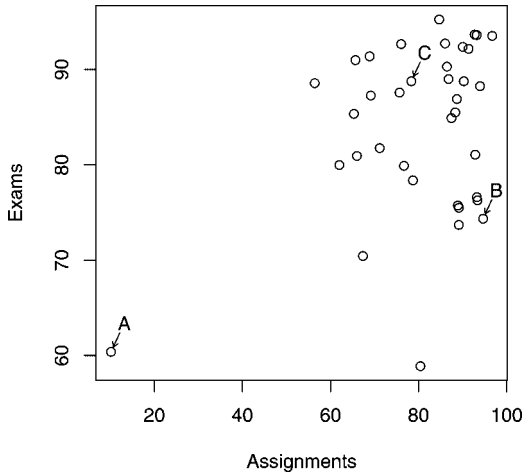

FIG. 7. *Assignments versus exams for the men.*

dents from those in Section 3.) Three points are noted for future reference. Figure 8 repeats the data, adding the point $z = (60, 70)$ and arrows pointing from each of the observations to z. Because z is relatively low, most of the arrows are pointing to the lower left. "Averaging" the arrows means averaging the unit vectors parallel to the arrows, so that the arrow from the point x_j to z has corresponding unit vector $(z - x_j)/\|z - x_j\|$, where for a vector $y = (y_1, \ldots, y_d)$, $\|y\| = \sqrt{y_1^2 + \cdots + y_d^2}$. For this z, the average is $(-0.6916, -0.5161)$. Thus the position of this point relative to the data is in the lower left. The "spatial median" is the z that has position 0. Figure 9 adds the arrows pointing to the spatial median for these data, which is $(84.5539, 85.8202)$. One can imagine that those arrows do indeed average to 0.

Formally, the position for bivariate or, in general, multivariate data is directly analogous to the univariate

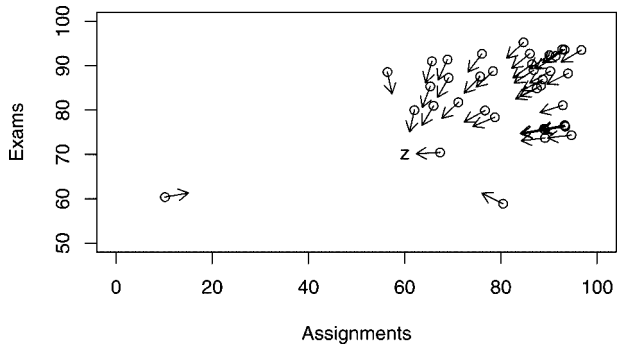

FIG. 8. *Assignments versus exams for the men, adding the arrows pointing toward the point* $z = (60, 70)$.

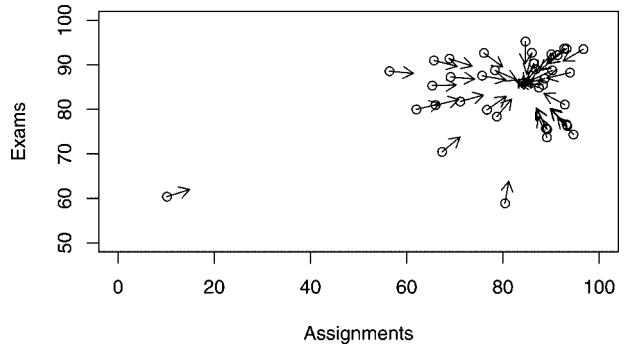

FIG. 9. *Assignments versus exams for the men, adding the arrows pointing toward the spatial median* $(84.5539, 85.8202)$.

definition (1), once we define the multivariate version of the sign function (2) to be the unit vector above.

DEFINITION 2. The position of $z \in \mathbb{R}^d$ among the points $\mathbf{X}_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ is

$$\text{pos}(z; \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \text{Sign}(z - x_i), \tag{8}$$

where $\text{Sign}(w)$ is the multivariate sign of the vector w given by

$$\text{Sign}(w) = \begin{cases} \dfrac{w}{\|w\|}, & \text{if } w \neq 0, \\ 0, & \text{if } w = 0. \end{cases}$$

If F is a distribution function, then the position of z relative to F is

$$\text{pos}(z; F) = E[\text{Sign}(z - X)], \quad \text{where } X \sim F.$$

See Small (1990), Koltchinskii (1997), Chaudhuri (1996) and Möttönen, Oja and Tienari (1997) for more details.

Notice that if the dimension $d = 1$, then this definition is the same as the original Definition 1. The position function in Definition 2 is often called the *spatial rank* or *geometric rank*. We prefer the term "position" here because "rank" may connote a univariate ordering of bivariate data.

Figure 10 plots the positions of the data points. Comparing the points labeled A, B and C in Figures 7 and 10, note that the outlier A in the raw data is still on the periphery in the positions plot, but has been brought in relative to the other points. Point B stays on the edge and point C is still somewhere in the middle. Generally, the points that were close together are spread out in the positions plot and the ones far away have joined the rest.

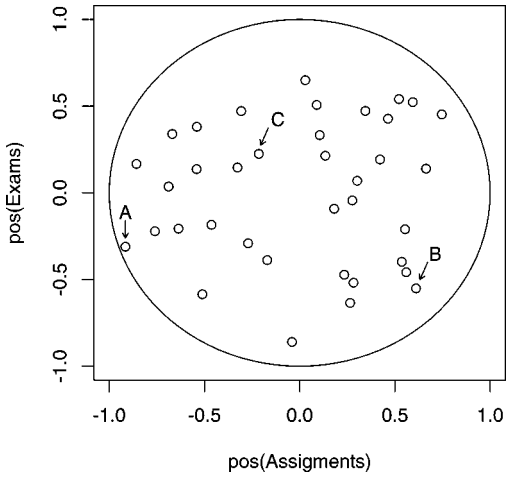

FIG. 10. *Bivariate positions of the assignments versus exams data for the men. The labeled points are the positions of the same-labeled points in Figure* 7.

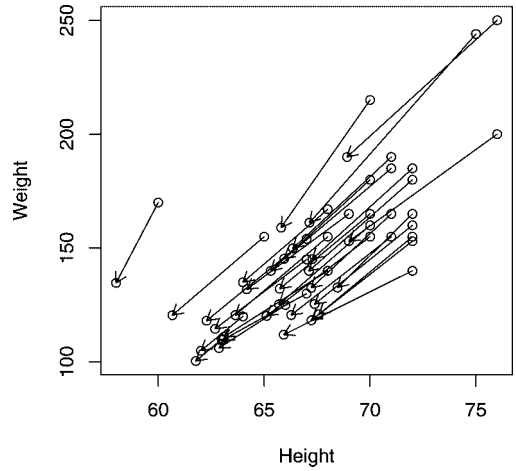

FIG. 11. *Bivariate QQ plot that compares men and women on height and weight.*

5. BIVARIATE QQ PLOTS

Using the bivariate positions from the previous section, it is straightforward to match observations in one sample with the observations with the same positions in another. To illustrate, consider the data from Section 3, but use the variables height in inches and weight in pounds. Let $\mathbf{X}_{n_2}^{(2)} = \{x_1^{(2)}, \ldots, x_{36}^{(2)}\}$ be the height and weight vectors for the 36 men. As in Definition 2, we find the positions of these observations relative to the sample, $q_j = \text{pos}(x_j^{(2)}; \mathbf{X}_{n_2}^{(2)})$, where the q_j's are two-dimensional vectors. For the women, we find the corresponding positions, that is, we solve for $\eta_j^{(1)}$ in the equation $q_j = \text{pos}(\eta_j^{(1)}; \mathbf{X}_{n_1}^{(1)})$. [The solution η to $q = \text{pos}(\eta; \mathbf{X}_n)$ is often called the qth spatial quantile; see Koltchinskii (1997) and Chaudhuri (1996) for theoretical properties.] Now we have each $x_j^{(2)}$ matched with its $\eta_j^{(1)}$. Table 2 shows four such matchups. The women tend to be smaller than the men, but the amount smaller is not constant. Thus the small man (5 ft 4 inches, 120 pounds) is matched with a women 2 inches shorter and 15 pounds lighter, while the big man (6 ft 3 inches, 244 pounds), is matched with a woman almost 8 inches shorter and over 80 pounds lighter.

Figure 11 graphs the matchings by drawing arrows pointing from the men's $x_j^{(2)}$'s to the women's $\eta_j^{(1)}$'s. Indeed, the arrows are generally pointing to the lower left, where the larger the men, the longer the arrows. The differences in lengths suggest that men and women may differ proportionally rather than additively, which leads to the QQ plot in Figure 12 based on the logs of the data. In this plot, the arrows are all very similar, both in direction and length, which suggests that in log terms, the women and men differ primarily in location.

The average of these arrows is $(0.06730, 0.2380)$, which means on average the women are 0.067 smaller in log height and 0.238 in log weight. If we augment the women's values by adding 0.067 to each

TABLE 2

Men		Women	
Height	**Weight**	**Height**	**Weight**
64	120	62.02	104.97
68	167	64.00	135.00
72	140	65.92	111.99
75	244	67.14	161.14

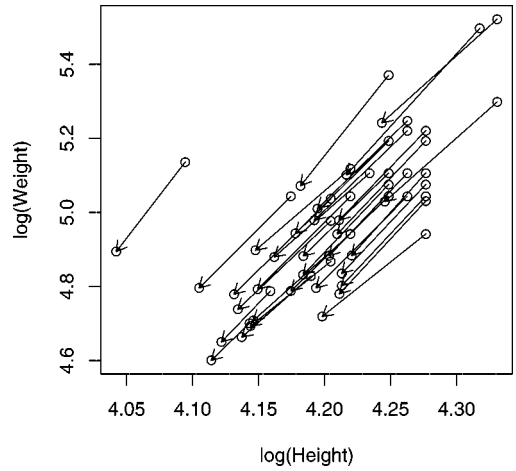

FIG. 12. *Bivariate QQ plot that compares men and women on* log(*height*) *and* log(*weight*).

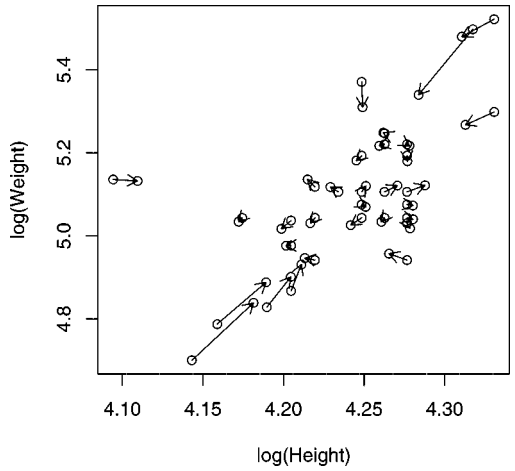

FIG. 13. *Bivariate QQ plot that compares men and women on* log(*height*) *and* log(*weight*)*, with the women shifted.*

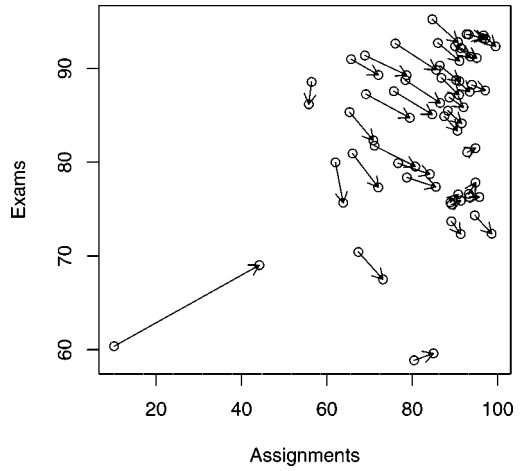

FIG. 14. *Bivariate QQ plot that compares men and women on assignments and exams.*

woman's log height and 0.238 to log weight, we obtain Figure 13. Now the bulk of the arrows are very small, suggesting that the difference between men and women is just a shift for those values, but for the extremes, the women are less extreme than the men. That is, the smaller men are associated with larger augmented women and the larger men are associated with smaller augmented women. Thus the tails of the men's distribution are larger than the tails of the women's distribution.

As suggested by this example, if one distribution is simply a shift of the other, then the arrows on the QQ plot will all be the same length and pointing in the same direction. If the two distributions differ only in scale, then the arrows from the larger distribution to the smaller will be pointing in toward the center, with longer arrows emanating from the points farther from the center.

Another example uses the assignment and exam data. Figure 14 shows the QQ plot, with the arrows pointing from the men to the women. Overall, the arrows tend to point to the lower right, which means that the women tend to do better on the assignments and worse on the exams. However, the pattern is more complicated than a direct shift. The arrows are quite small when the men's assignment scores are above about 85, while much larger when those scores were less than 85. Table 3 extracts what the plot suggests, comparing the median scores for the men and women split on the men's assignment scores. For the higher assignment scores, the men do only 1 point better on exams and 3 points worse on assignments, while for the lower assignment scores, the men do 3 points higher on the exams, but 10 points worse on the assignments. Additional examples can be found in Marden (1998), including QQ plots that compare samples to the bivariate normal.

We remark that the multivariate positions depend on the relative scales of the variables. That is, changing the scales of the variables will change the positions and the spatial quantiles. For the data we used, the variables did not have wildly differing scales, but in practice it may be a good idea to first normalize the data, for example, divide each variable by its mad. Alternatively, one could use coordinatewise positions: for vector w, the ith component in its position vector is the univariate position of the w_i among the values of the ith variable for the observations. This vector has the same relationship with the coordinatewise ranks given in (3) for the univariate position. [It also arises from (8) when using coordinatewise signs.] The coordinatewise positions are invariant under separate scale changes for the variables, but unlike our positions (Definition 2), do not preserve the spatial arrangement of the data. That is, multiplying the data by an orthogonal matrix changes the picture. Chakraborty (2001) gave an approach that handles scale as well as rotation changes by using affine invariant multivariate quantiles for the QQ plots; see also Visuri et al. (2003).

TABLE 3

	Men's assignment > 85		**Men's assignment ≤ 85**	
	Assignments	**Exams**	**Assignments**	**Exams**
Women	93.55	86.51	79.41	82.34
Men	90.16	87.58	69.15	85.34

6. CONCLUDING REMARKS

The purpose of this paper is to present QQ plots for comparing distributions. The main tool is positioning, used for QQ plots as well as for obtaining the Sen–Theil estimate of the slope in linear regression. At each stage, there are many choices to be made. We make what we believe are reasonable choices, but certainly not the only reasonable ones. Other procedures for estimating parameters robustly and defining multivariate ranks or positions are certainly available, and everyone is encouraged to try a variety of methods.

The data sets we use to illustrate the methods are quite small, both in dimension and number of observations. The methods here easily apply to large sample sizes, but especially for bivariate QQ plots, one should be circumspect in the number of points and arrows plotted. That is, plotting more than 40 or 50 arrows will likely make the plot unreadable, so if both sample sizes are quite large, one should make a judicious choice of $\eta_j^{(2)}$'s in (7) to use. One could randomly choose 40, say, from one of the samples, but a more systematic choice that covers the range of the data more uniformly would be preferable.

For dimensions $d > 2$, the arrows will be in d space, so will be more difficult to visualize. One could project the arrows to the various bivariate planes determined by pairs of variables, create three-dimensional plots (preferably dynamic) or use projection pursuit ideas to project to the plane which maximizes the lengths (say) of the projected arrows.

In any case, there is definite room for further work in this area.

ACKNOWLEDGMENT

Partial support for this research was provided by NSF Grant DMS-00-71757.

REFERENCES

BARNETT, V. (1975). Probability plotting methods and order statistics. *Appl. Statist.* **24** 95–108.

CHAKRABORTY, B. (2001). On affine equivariant multivariate quantiles. *Ann. Inst. Statist. Math.* **53** 380–403.

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

CLEVELAND, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

COOK, R. D. and WEISBERG, S. (1997). Graphics for assessing the adequacy of regression models. *J. Amer. Statist. Assoc.* **92** 490–499.

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

LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858.

MARDEN, J. I. (1998). Bivariate QQ-plots and spider web plots. *Statist. Sinica* **8** 813–826.

MÖTTÖNEN, J., OJA, H. and TIENARI, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Ann. Statist.* **25** 542–552.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

ROUSSEEUW, P. J., RUTS, I. and TUKEY, J. W. (1999). The bagplot: A bivariate boxplot. *Amer. Statist.* **53** 382–387.

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

SERFLING, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Statist. Neerlandica* **56** 214–232.

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

THEIL, H. (1950). A rank-invariant method of linear and polynomial regression analysis. I. *Nederl. Akad. Wetensch. Proc.* **53** 386–392.

VISURI, S., OLLILA, E., KOIVUNEN, V., MÖTTÖNEN, J. and OJA, H. (2003). Affine equivariant multivariate rank methods. *J. Statist. Plann. Inference* **114** 161–185.

WILK, M. B. and GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55** 1–17.