Information Visualization (2008), 1–18

www.palgrave-journals.com/ivs

Judging correlation from scatterplots and parallel coordinate plots

Jing Li[1]
Jean-Bernard Martens[2]
Jarke J van Wijk[1]

[1] *Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands;* [2] *Department of Industrial Design, Eindhoven University of Technology, the Netherlands*

Correspondence:
Jing Li, Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.
Tel: +31(0)40 247 3883;
Fax: +31(0)40 246 8508;
E-mail: J.Li@tue.nl

Abstract
Scatterplots and parallel coordinate plots (PCPs) can both be used to assess correlation visually. In this paper, we compare these two visualization methods in a controlled user experiment. More specifically, 25 participants were asked to report observed correlation as a function of the sample correlation under varying conditions of visualization method, sample size and observation time. A statistical model is proposed to describe the correlation judgment process. The accuracy and the bias in the judgments in the different conditions are established by interpreting the parameters in this model. A discriminability index is proposed to characterize the performance accuracy in each experimental condition. Moreover, a statistical test is applied to derive whether or not the human sensation scale differs from a theoretically optimal (i.e., unbiased) judgment scale. Based on these analyses, we conclude that users can reliably distinguish twice as many different correlation levels when using scatterplots as when using PCPs. We also find that there is a bias towards reporting negative correlations when using PCPs. Therefore, we conclude that scatterplots are more effective than parallel plots in supporting visual correlation analysis.
Information Visualization advance online publication, 1 May 2008;
doi:10.1057/palgrave.ivs.9500179

Keywords: Correlation visualization; scatterplots; parallel coordinate plots; evaluation of visualization; perception of correlation; statistical graphs

Introduction

Multivariate data analysis is concerned with finding patterns and relationships in data that contain measurements on more than one variable. Linear association is the most basic and simple relationship between two variables and correlation is the most widespread measure for quantifying the strength and direction of such bivariate linear associations. Assessing correlation is often an important first step in more advanced data analyses, such as factor analysis. Moreover, correlation analysis was identified as one of the 10 low-level analysis tasks that are important within information visualization.[1] These low-level tasks are fundamental to the cognitive processes involved in analyzing data, and describe key user activities while employing visualization for understanding data.

In statistical analysis, correlation can be defined in different ways, but most common is to use r, Pearson's product-moment coefficient.[2] If x_i and y_i, for $i = 1, \ldots, n$, are the sample data for the two variables under consideration, then the *sample correlation* coefficient r is equal to

$$r = \frac{S_{xy}}{S_x S_y}, \tag{1}$$

where $S_x = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$, $S_y = \sqrt{\sum (y_i - \bar{y})^2/(n-1)}$ are the sample standard deviations of x and y, and $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)$ is the sample covariance between x and y. The value of r ranges from -1

Received: 12 December 2007
Revised: 11 March 2008
Accepted: 16 March 2008

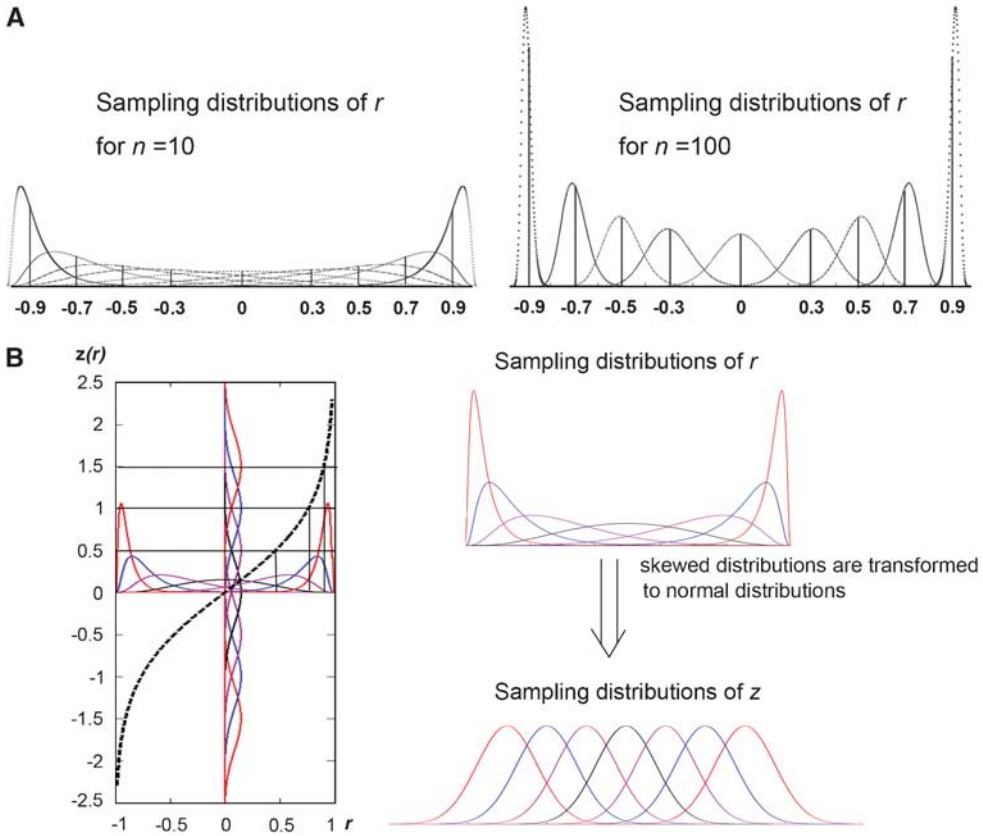

Figure 1 (A) Distributions of the sample correlation coefficient r, for sample size $n = 10$ (larger variance) and $n = 100$ (smaller variance). The distribution is positively skewed for negative correlations, unskewed for zero correlation and negatively skewed for positive correlations. (B) Illustration of the Fisher z transformation on r for $n = 10$. The skewed sampling distributions with varying standard deviations on r are transformed into normal distributions with constant standard deviation on z.

(perfect negative correlation) via 0 (no correlation) to 1 (perfect positive correlation). If the data pairs (x_i, y_i), for $i = 1, \ldots, n$, are drawn from a population with *population correlation* coefficient ρ, then the sample correlation coefficient r will vary around ρ. The variances in r will decrease with increasing sample size n. Figure 1(A) shows these *sampling distributions* of r for nine different values of ρ and two different sample sizes ($n = 10$ and $n = 100$). When the absolute value of ρ is close to zero, then the sampling distribution of Pearson's r is approximately normal, but when the absolute value of ρ is close to one, then the sampling distribution is highly skewed and has noticeable smaller standard deviation. This change in the shape of the distribution of r as a function of ρ implies that the sensitivity of the estimation Eq. (1) is not constant. Otherwise phrased, the estimation equation is biased, in terms of accuracy, towards high correlation coefficients.

In statistics, there is a well-known scale with constant sensitivity, that is, the Fisher z transformation of r:

$$z = Z(r) = \frac{1}{2} \log \frac{1+r}{1-r}. \qquad (2)$$

Fisher z has the property that the estimated values for $Z(r)$ are normally distributed around the true value $Z(\rho)$, while the standard deviation of this normal distribution $\sigma = 1/\sqrt{n-3}$ only depends on the sample size n, and not on ρ. In other words, the sensitivity is constant across the z scale for a fixed sample size n. Figure 1(B) illustrates this by means of an example for $n = 10$.

Although a single measure for correlation is easy to interpret, much information is left out. This may lead to misinterpretation of data, especially when the underlying assumption of a joint normal distribution of both variables is violated, as in the example of Figure 2. Visualization can complement calculation in that it aims at showing all data, instead of only a summary statistic. Visualization is used increasingly in data analysis, especially when setting up hypotheses and identifying characteristic patterns.

Scatterplots are used most often to visualize bivariate data, and are routinely used to estimate correlation visually. The x and y variables are mapped to Cartesian coordinates, and sample items (x_i, y_i), for $i = 1, 2, \ldots, n$, are depicted by points. The strength of the correlation is usually associated with the degree to which the points aggregate around a line. This is not fully correct however, as we will explain in more detail in the section 'Scatterplots'. The direction of the correlation is indicated

by the dominant direction of the cloud of points, see Figure 3(A) (direction southwest–northeast for positive correlation, and direction southeast–northwest for negative correlation).

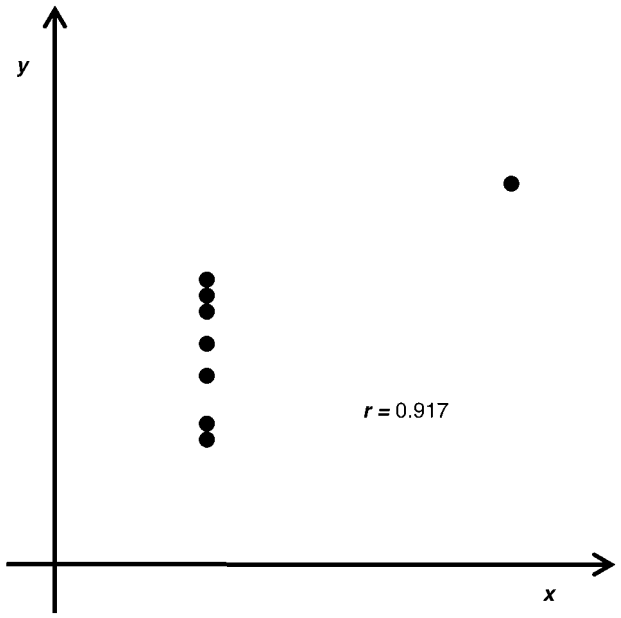

Figure 2 The value of *r* (0.917) suggests a high degree of correlation, while *x* and *y* are not linearly associated.

Parallel coordinates plots (PCPs) constitute another, somewhat less well-known visualization method. Here, the *x* and *y* variables are mapped to two vertical parallel axes with sample items being depicted by line segments between these axes. In other words, the projective line-point duality[3,4] from geometry is used, so that a scatterplot can be transformed into a PCP and vice versa. Thus, it is also possible to visualize correlation by PCPs.[4–6] Positive and negative correlations lead to very different patterns in PCPs.[6,7] The three characteristic patterns, associated with the extreme values of *r*, are shown in Figure 3(B). Positive correlation leads to a pattern with parallel line segments; negative correlation gives a diabolo-like pattern, with line segments intersecting in one point. For data sets with an intermediate correlation the direction of correlation can be estimated by judging which of these patterns is approximated best, and the strength of correlation by judging how close this approximation is.

For multivariate correlation analysis, it is common practice to place correlation coefficients for each pair of variables in a matrix with one row and one column for every variable.[2] In the same way, a matrix of scatterplots can be used as a combined view of all bivariate scatterplots (Figure 4(A)).[6] However, it may be difficult to get an overview of all the relationships in such a matrix, especially when the number of variables is large. PCPs were particularly invented to deal with multivariate data. Bivariate PCPs are easily extended into multivariate visualizations by adding a parallel axis for each variable

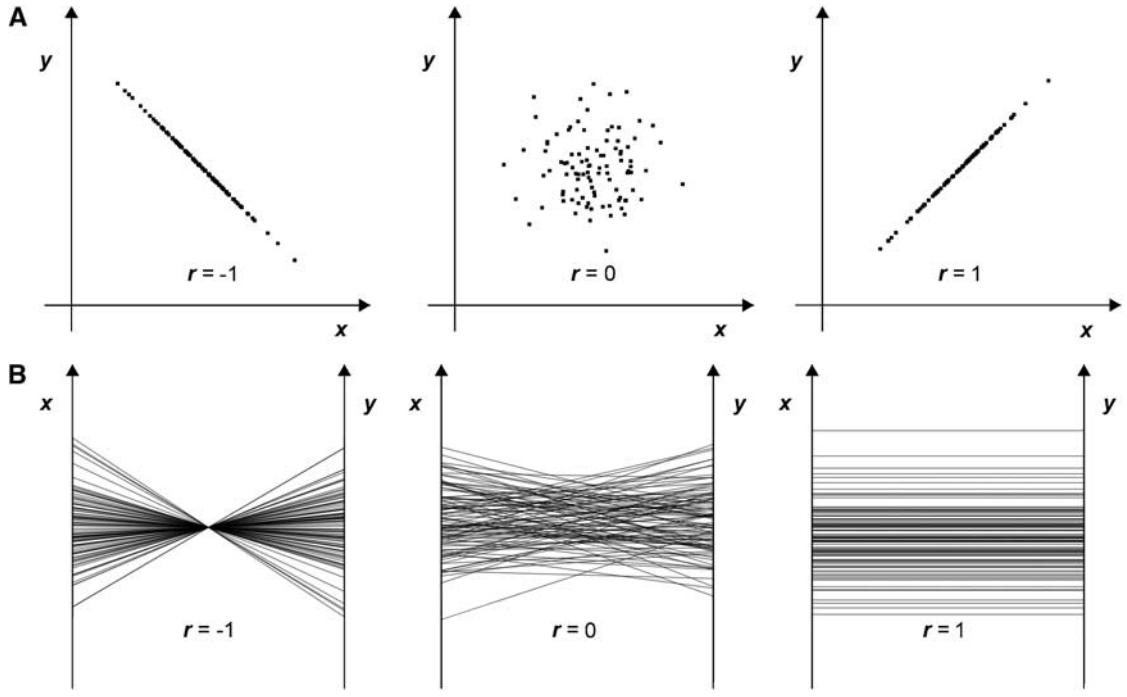

Figure 3 The three extreme values of *r* visualized with (A) scatterplots and (B) parallel coordinate plots.

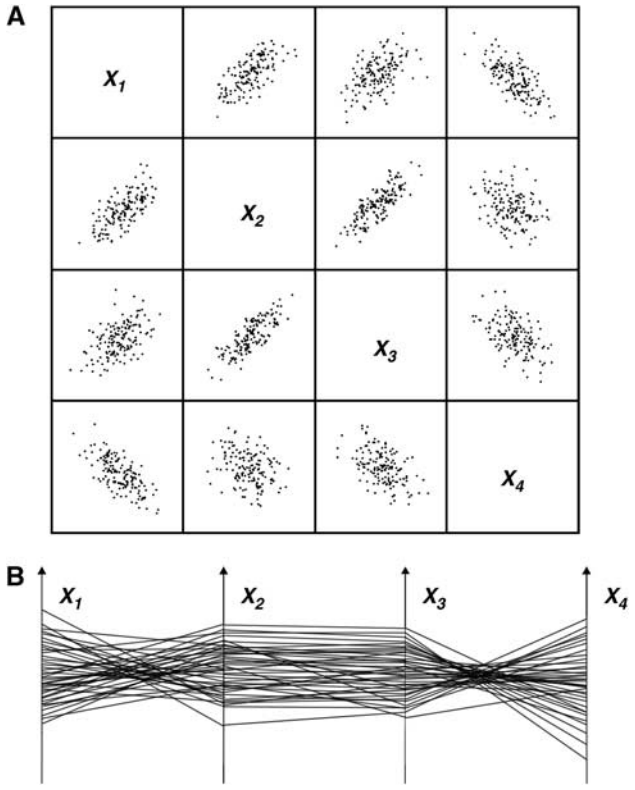

Figure 4 Visualization of multivariate data: (A) scatterplot matrix and (B) parallel coordinate plot.

(Figure 4(B)). Sample items are depicted as polygonal lines across the set of parallel axes. The pattern of those polygonal lines can disclose relationships for multiple pairs of variables. However, to explore relationships between variables exhaustively, different orders of the axes have to be examined.

Scatterplots are used intensively in practice, and statistical studies and psychological experiments have suggested how to use scatterplots for correlation analysis. PCPs support simultaneous viewing of relations between multiple pairs of variables, but we are not aware of any studies to evaluate how effective they are in conveying relevant information. We focus here on correlation analysis, as a central issue in multivariate data analysis, but we acknowledge that other tasks, such as cluster detection, are also highly relevant. We limit ourselves to bivariate data analysis, because this is a common activity and also because scatterplot matrices as well as higher dimensional PCPs consist of combinations of bivariate displays.

Our aim is to evaluate the effectiveness of these two visualization methods for correlation analysis and to deduce whether or not observers sense correlation differently in both visualization methods. To this end, we have performed a user experiment in which the observers were asked to report perceived correlation as a function of population correlation, sample size, visualization method and observation time. A statistical model will be proposed to describe the experimental data, and variations in the model parameters will be used to highlight differences in perception in the distinct conditions.

In the following sections, we discuss related work, the experimental design, the data analysis by means of the proposed statistical model, followed by conclusions.

Background

Results from statistics, psychology and information visualization provide useful background information for our experiment. In statistics, there is a long history of using scatterplots to study correlation. Restrictions and problems of using scatterplots have been clearly identified. Also, assessing bivariate association has interested psychologists since it is a basic cognitive task. However, not many guidelines can be found for PCPs. In information visualization, many applications of PCPs have been presented for multivariate data analysis, and suggestions for improvements have been made. Finally, we consider the evaluation methodology to be used in our comparative study.

Scatterplots

Although scatterplots are routinely used, they can be seriously misleading when used to judge correlation.[8] Before explaining the problem, we define some visual features of scatterplots.

The SD-line is defined as the line passing through the centroid of the point cloud with slope S_y/S_x, where S_x and S_y are the sample standard deviations, as defined in Eq. (1). The major axis D is defined as the major diameter of an ellipse fitting the point cloud and the minor axis d is the minor diameter of the ellipse. The direction of D is approximately the direction of the SD-line. The ratio d/D is a measure for how closely the point cloud approximates a line (Figure 5).

One common mistake is to consider the linear degree of the point cloud (d/D) as representative of $|r|$. However, $|r|$ measures the degree of linear association between two normally distributed random variables, and not the degree to which the point cloud is linear; $|r|$ is actually determined by both d/D and the direction of the SD-line. Indeed, rotation of a point cloud yields different $|r|$ values,[8] while the linear degree of the point cloud obviously remains the same (Figure 6). If the axes in scatterplots are rotated continuously, the value of $|r|^2$ varies smoothly from 0 to its maximum value.[8] The maximum of $|r|^2$ is attained when $S_y/S_x = 1$. Visually combining the two features in scatterplots is difficult and a recommendation[9] is therefore to fix S_y/S_x to 1, such that the $|r|$ of different sample data sets is visually comparable by comparing d/D. In this case, the direction of the SD-line is fixed to be parallel to the line $y = x$. We will follow this recommendation and use sample data sets with $S_y/S_x = 1$ in our experiment.

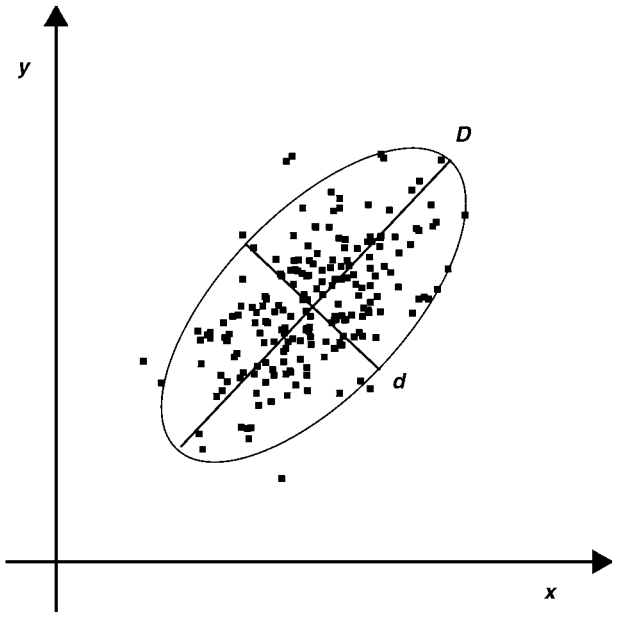

Figure 5 The major diameter D and minor diameter d of the ellipse fitting the point cloud.

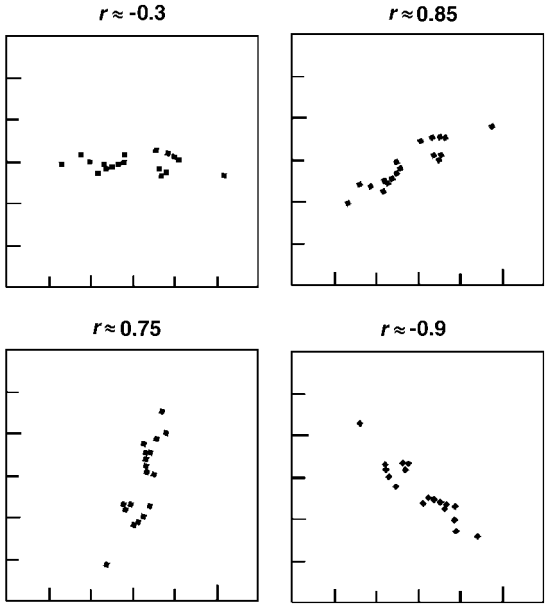

Figure 6 Effect on correlation of rotating a point cloud in a scatterplot.

Another perceptual problem of scatterplots was demonstrated by Cleveland *et al*.[9] In this study, it was established that variables looked more highly correlated when the size of the point cloud size was decreased. The direction of D was fixed to be $y=x$, which is approximately equal to the condition that $S_y/S_x = 1$. The smaller the point cloud size, the more points fuse (Figure 7). According to this observation, the scales of the axes in the scatterplots should be kept constant when comparing correlation coefficients across different samples.

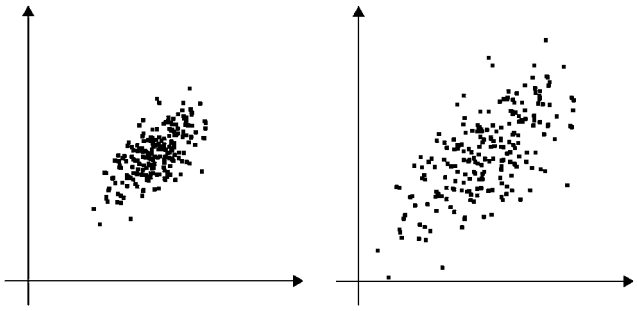

Figure 7 Two scatterplots with $r = 0.66$ and different scales for which the visually estimated correlation tends to vary.

An interesting issue is how the perceived correlation relates to the actual sample correlation r. Several experiments have demonstrated that the perceived correlation is often different from r in case of scatterplots, particularly underestimating $|r|$ for smaller values.[9–11] So the relationship between perceived and statistical correlation is not linear. The experiments by Cleveland *et al*[9] showed that neither of two proposed functions of r: $w(r) = 1 - \sqrt{1 - r^2}$ and $g(r) = 1 - \sqrt{(1-r)/(1+r)}$, succeed in describing the perceived correlation very well. The geometric interpretation of $w(r)$ is d/D. The geometric interpretation of $g(r)$ is the ratio of the elliptical area of the point cloud to the rectangle area that contains the point cloud. Therefore, they concluded that, despite the fact that both geometrical cues might be employed during the perceptual process of judging correlation, neither $w(r)$ nor $g(r)$ describe the human sensation scale correctly. They hypothesized that this might be due to the fact that area and length perceptions may be non-linearly related to their physical quantities (more specifically, they proposed a power-law relationship with a power less than 1). Their hypothesis may also explain the failure of one alternative proposal in Strahan and Hansen[10] for describing the non-linear perception of correlation.

A recent intensive study on the physiological processing involved when viewing scatterplots was carried out by Best *et al*.[12] No explicit response was required from participants, as different brain sites were monitored while subjects were observing scatterplots with 11 different correlation levels ($r = 0, \pm0.1, \pm0.3, \pm0.5, \pm0.7, \pm0.9$). The result showed that perceiving correlation needs visual coding of spatial patterns as well as verbal coding to label the relationships. This indicates that visual correlation perception is not a pre-attentive activity. It was also observed that positive and negative correlations are processed in different ways, which might explain the positive bias when observing negative correlations.[13] Three levels of correlation strength (weak, moderate and strong) were treated differently during the perception process.

The mental effort increased when judging moderate correlations (particularly, in case of $|r| = 0.3$ and 0.5), which was explained by means of an efficiency hypothesis.[12] In case of moderate correlations, it was less obvious for the brain to decide which spatial areas were most relevant for the cognitive task and should be activated.

Parallel coordinate plots

We have only found few experimental results[14,15] for PCPs and the question of what exactly humans can deduce from PCPs is largely open. There however seem to be two rather common agreements in the field of information visualization. One agreed advantage of using PCPs for viewing multivariate data is that the linked sequence of bivariate displays supports both a continuous view and a comparative view for more than one pair of variables. Most applications of PCPs have been developed based on this advantage.[16–18] An agreed disadvantage of using PCPs is that the noisy intersections of many lines obscure the underlying patterns and cause visual clutter, especially when visualizing a large amount of data. Several efforts have been made to reduce visual clutter in PCPs. For instance, algorithms have been developed to focus on the main trends and filter out the less relevant ink.[19,20] Alternatively, sample data can be preprocessed into clusters, and different rendering methods can be used to highlight these clusters.[21,22] An interesting method with curved lines was also proposed for better clustering in PCPs.[23] We focus here on standard PCPs, but acknowledge that such methods can be used to reduce clutter. The visual stimuli used in our experiment will be limited to a range of sample sizes where clutter is not a dominant issue (see the section 'Sample size n').

Regarding the advantage of using PCPs for multivariate data, we agree that an integrated view may help to obtain a good overview of the data. However, we argue also that finding correlations in PCPs is ultimately based on finding correlations between pairs. Hence, effectively capturing the decomposed view of every bivariate display is a prerequisite for an effective overview or comparison among different pairs of variables. We therefore evaluate the simplest form of PCPs, where just one pair of variables is visualized.

Comparative evaluation in information visualization

The standard approach in information visualization for the assessment of new techniques is to carry out a task with an existing and a new method, and to measure the time needed and the errors made.[24–27] Such a comparative evaluation was carried between parallel coordinates and stardinates (a glyph-based technique with multiple dimensions represented on asterisk-like axes).[24] The efficiency and effectiveness of the two visualization methods were compared based on the task completion time, the number of correct answers, the number of hypotheses and on how subjective statements were in agreement with expert views. The results demonstrated that stardinates seem less informative at first glance to individuals but are more appropriate for interpreting highly structured data in detail. In the study, measuring insight was addressed with open-ended protocols and domain relevance.[28] These experimental findings however provided very little insight into how the perceptual processes developed.

An alternative approach that we will advocate in this paper is to describe the perceptual process using a statistical model. Such a model relates the perceived inputs to the observed user responses.[29] More specifically, the probabilities of different outputs in response to a specific input are expressed in terms of the parameters in the model. The parameter values are estimated using the ML (Maximum Likelihood) principle, that is, they are iteratively optimized until the observed frequencies of responses are the most likely outputs of the model. We adopt this approach here, because such a model provides detailed insight in the perceptual process, and because estimated model parameters can be easily compared across experimental conditions. In the following section, we describe the statistical model and the design of the actual experiment.

Experimental design

In an experiment, independent variables refer to variables that are manipulated by the experimenter and dependent variables are those being measured.[30] In order to keep the required experimental effort limited, we need to identify the most interesting aspects of the task. This has resulted in the variation of four independent variables in our user experiment: the visualization method V (scatterplots *sc vs* PCPs *pc*), the observation duration T (limited display time *ld vs* unlimited display time *ud*), sample size n and population correlation coefficient ρ. The user judgment of correlation U is the dependent variable. Our experiment aims to find out how the independent variable V influences the dependent variable U in case of different settings of the other three independent variables T, n and ρ.

Model

The experiment involves only one user task, namely judging correlation. The proposed model for the perceptual and judgment process is shown schematically in Figure 8. The outputs u_{ki} provided by the subjects in response to the correlation level i ($i = 1, \ldots, I$, where $I = 7$ in our experiment) and the experimental condition k (unique combinations of n, V and T) are the observables in the experiment. The mean p_{ki} and standard deviation σ_k of the normally distributed internal visual sensations s_{ki} cannot be observed directly but need to be deduced from the pattern of observed responses. More specifically, we assume that the relationship between the continuous sensation s_{ki} and the discrete response u_{ki} is described by a uniform quantizer. This implies that the expected frequencies of responses in the available output categories ($-2, -1, 0, +1, +2$, which we will explain in more detail

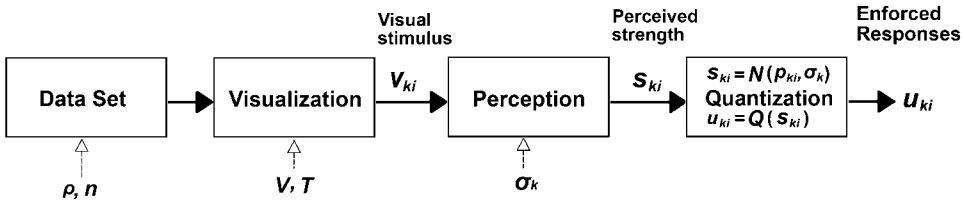

Figure 8 The proposed statistical model describing the human perception and judgment process of visual correlation. The parameters ρ and n specify the data set that is visualized, while V and T identify the visualization method being used and the time available for the participant to observe the visualization. The visual stimulus v_{ki} triggers a perceptual process that is statistical in nature, that is, the perceived signal strength s_{ki} is assumed to be normally distributed around an average value p_{ki} that depends on both the correlation strength ρ (coded by index *i)* and the independent parameters n, V and T (coded by the index k), and a standard deviation σ_k that is independent of correlation strength. The mapping from the internally perceived signal strength s_{ki} to the externally observed user response u_{ki} is modeled by a quantization process.

in the section 'Response scale of dependent variable U') can be established as described below. If we denote the Gaussian distribution by

$$\varphi(u; p_{ki}, \sigma_k) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(u-p_{ki})^2}{2\sigma_k^2}\right) = \frac{1}{\sigma_k}\,\varphi\left(\frac{u-p_{ki}}{\sigma_k}\right), \quad (3)$$

where $\varphi(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$ is the Gaussian distribution with zero mean and unit standard deviation, then the expected frequency of response $u = 2$ can be expressed as

$$P(u=2) = \frac{1}{\sigma_k}\int_{1.5}^{+\infty} \varphi\left(\frac{u-p_{ki}}{\sigma_k}\right) \mathrm{d}u = \int_{1.5/\sigma_k}^{+\infty} \varphi\left(\frac{u}{\sigma_k} - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}\left(\frac{u}{\sigma_k}\right) = \int_{1.5/\sigma_k}^{+\infty} \varphi\left(u^* - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}u^*. \quad (4)$$

Similarly, the expected frequencies of responses $u = 1$, 0, -1, and -2 can be expressed as

$$P(u=1) = \int_{0.5/\sigma_k}^{1.5/\sigma_k} \varphi\left(u^* - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}u^*, \quad (5)$$

$$P(u=0) = \int_{-0.5/\sigma_k}^{0.5/\sigma_k} \varphi\left(u^* - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}u^*, \quad (6)$$

$$P(u=-1) = \int_{-1.5/\sigma_k}^{-0.5/\sigma_k} \varphi\left(u^* - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}u^*, \quad (7)$$

and

$$P(u=-2) = \int_{-\infty}^{-1.5/\sigma_k} \varphi\left(u^* - \frac{p_{ki}}{\sigma_k}\right) \mathrm{d}u^*. \quad (8)$$

By varying the model parameters p_{ki} and σ_k we can adjust these expected frequencies to the actually observed distribution of responses in the experiment. More specifically, maximum likelihood estimation is used to find the optimal parameter values. This estimation process is illustrated in Figure 9.

The model assumes the existence of a sensation scale for correlation that has constant sensitivity (or noise) across its entire range. This sensitivity reflects the accuracy of the underlying physical processes (of viewing and processing the visual stimulus), and is supposed to be independent of the type of information being processed. The practical argument for being interested in a scale with constant sensitivity (i.e., with interval properties) is that it helps to assess the differences (or intervals) between stimuli. More specifically, if a scale has interval properties then we can deduce that the perceptual difference between stimulus 1 and 2 is bigger or smaller than the perceptual difference between stimulus 2 and 3, if the corresponding distances between the stimulus representations on the interval scale behave accordingly.

Independent variables – stimulus specification

The visual stimuli in the experiment are determined by the sample data sets, the visualization method V and display time condition T. The combination of V and T results in four test sessions: (*sc*, *ld*), (*pc*, *ld*), (*sc*, *ud*), (*pc*, *ud*). Furthermore, in each test session we use 42 sample data sets. We use three different levels of n and seven different levels of ρ (Figure 10), and for each pair of (n, ρ), two random samples are generated. Examples of the resulting visual stimuli for different values of ρ, n and V are shown in Figure 10. Below, we explain how the specific stimulus levels were selected.

Specification of correlation – Fisher-z transformation In agreement with an earlier study,[12] we decided for three levels of correlation strength, that is, weak, moderate and strong. We wanted to explore both positive, negative and zero correlations, which resulted in seven different levels for ρ. Since earlier experiments have revealed that these

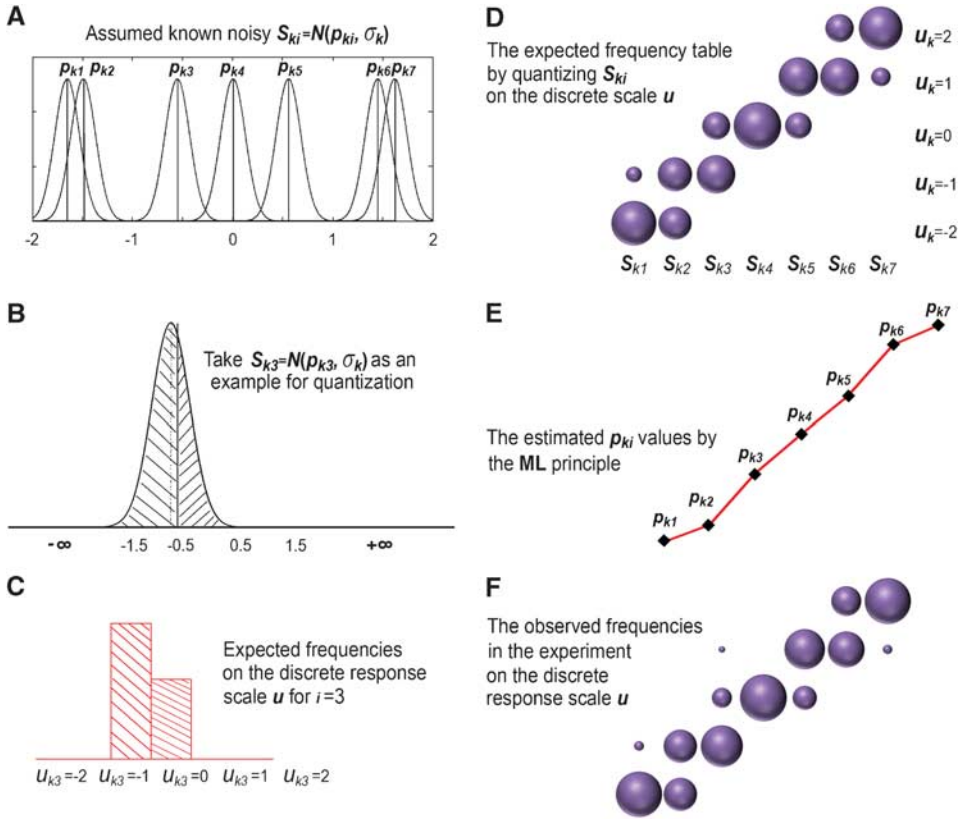

Figure 9 The quantization mechanism and estimation process based on the ML principle for condition $k = \{(sc, ud) \,\&\, n = 160\}$: (A) the estimated parameters p_{ki} and σ_k determine the distributions of perceived strength s_{ki}; (B) quantization of s_{k3} into a discrete response u_{k3} (s_{ki} in $[-\infty, -1.5]$ results in $u_{ki} = -2$, s_{ki} in $[-1.5, -0.5]$ results in $u_{ki} = -1$, s_{ki} in $[-0.5, 0.5]$ results in $u_{ki} = 0$, s_{ki} in [0.5,1.5] results in $u_{ki} = 1$ and s_{ki} in $[1.5, \infty]$ results in $u_{ki} = 2$); (C) producing response frequencies by integrating distributions over quantization intervals; (D) expected response frequencies presented by a bubble chart; (E) average stimulus strengths p_{ki}; (F) observed response frequencies presented by a bubble chart with the same scales as in (D).

levels of correlation are treated differently by humans, we expected to be able to reveal differences in the perception and judgment process with such stimuli. Next, we needed to decide on the precise values of ρ. The aim was to select values that were about equally distant on a human judgment scale. As discussed earlier, equidistance on the scale of ρ does not map to equidistance on the human sensation scale of correlation.[9,10] Hence equal intervals of ρ did not seem appropriate. Since alternative functions like ρ^2 (Strahan and Hansen[10]) and $w(\rho)$ and $g(\rho)$,[9] have also not been proven to represent human perception, we decided to select our stimuli such that they were equidistant on the theoretically optimal Fisher z scale. This is equivalent to assuming that human perception is close to optimal. We therefore adopted equal intervals on z ($z = 0, \pm 0.5, \pm 1.0, \pm 1.5$) to select the average correlation coefficients in our experiment (corresponding to $\rho = 0, \pm 0.462, \pm 0.762, \pm 0.905$). The correlation strength levels range from weak to strong (Figure 11). In the section 'Data analysis', we can consider to what extent the actual human sensation differs from this theoretical optimum z.

Sample size n For both scatterplots and PCPs, we should use the same sample sizes for comparison purposes. Empirically, n can be varied over a wide range for scatterplots without problems. However, PCPs become cluttered for large n. We carried out a small user test to obtain more insight into the problem. Three subjects (two males and one female), aged between 25 and 30 years, were involved in the test. Random sample sets with increasing size were presented to them as PCPs. In the first round, the increment was 50 data items ($n = 50$, 100, 150, 200, 250). We stopped increasing n when the subject reported that the image shown made no sense anymore due to visual clutter. One subject stopped at 150, the other two stopped at 200. In the second round, the increment was 10 items ($n = 150$, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250). Now two subjects stopped at $n = 200$ and one at $n = 180$. Since our experiment is not intended to expose the well-known problem of visual clutter in case of PCPs, we adopted the following three sample sizes to be used in the experiment: 10, 40 and 160, which correspond to small sample size, medium sample size and large sample size.

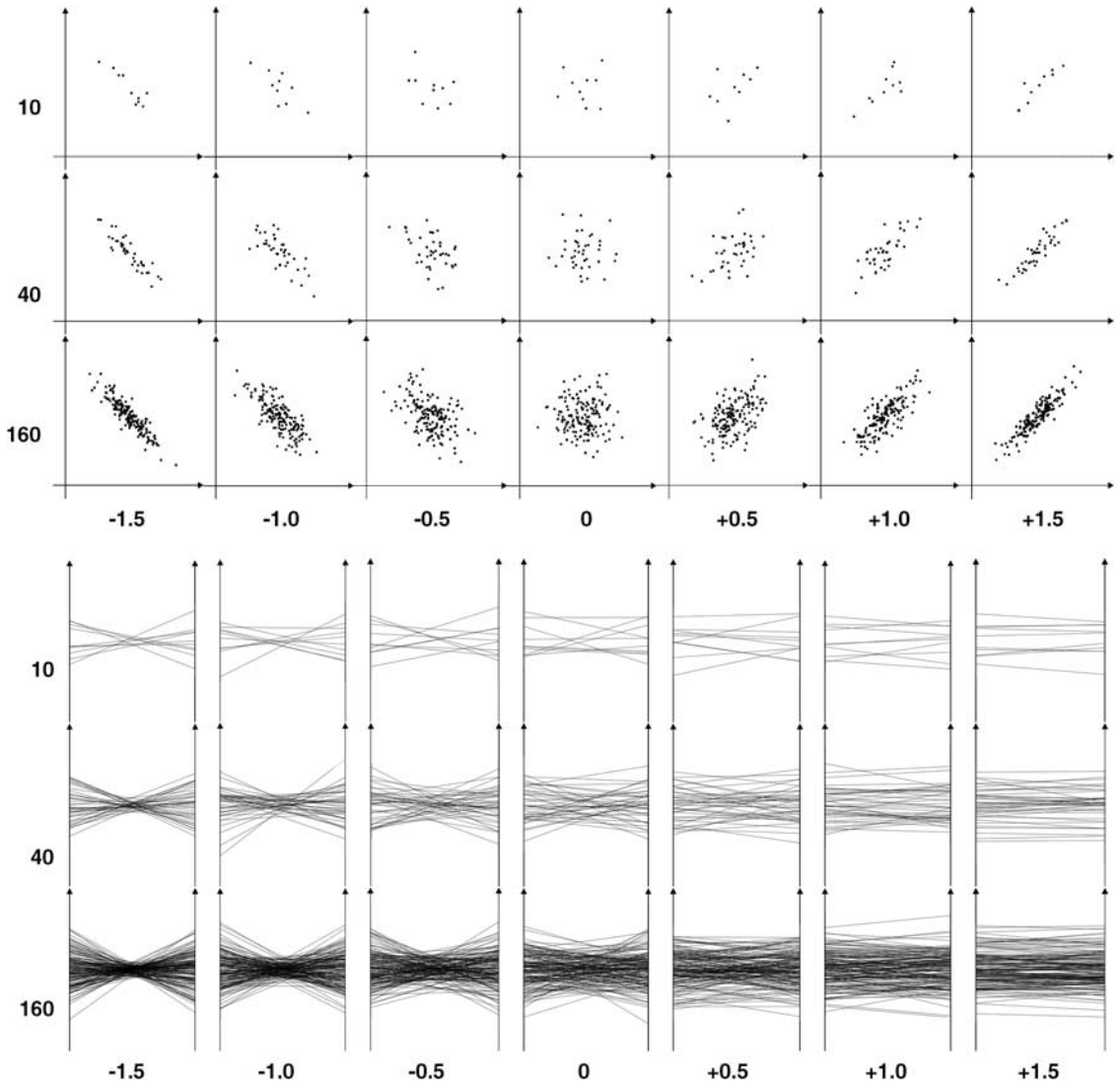

Figure 10 Visual stimuli: scatterplots (top) and PCPs (bottom) with controlled correlations defined by z (in columns) and sample size n (in rows).

Display time condition T We also wanted to investigate the difference between limited and unlimited inspection (display) time, in order to better understand the difference between occasional observation (for instance in scatterplot matrices) and careful studying. The limited display time is supposed to be short enough to only allow a quick impression of the data. At the same time, since correlation analysis is not a pre-attentive activity, it should not be too short either. A small pilot study was conducted to make an informed choice about the limited time condition. Three subjects were involved here (two female, one male), aged between 20 and 35 years. Random sample data were presented to them for increasingly long display times. In the first round we used an increment of 300 ms (t=500 ms, 800 ms, 1100 ms, 1400 ms). All subjects reported increased difficulty of judgment for time settings below 1100 ms. They complained that the visual image could no longer be retained properly in memory. In the second round, we used an increment of 100 ms (t = 900 ms, 1000 ms, 1100 ms, 1200 ms). All subjects declared that a display time of 1000 ms was marginally acceptable. Based on this result, we selected 1 s for the limited time condition.

Response scale of dependent variable *U*

For the response scale, we used five semantic levels: strong negative, negative, not correlated, positive and strong positive. The scale is restricted to five levels, which are organized symmetrical around no correlation. Past experience[29] with categorical scaling has shown that such limited scales allow for an approximately linear mapping between sensation strength and response rating.

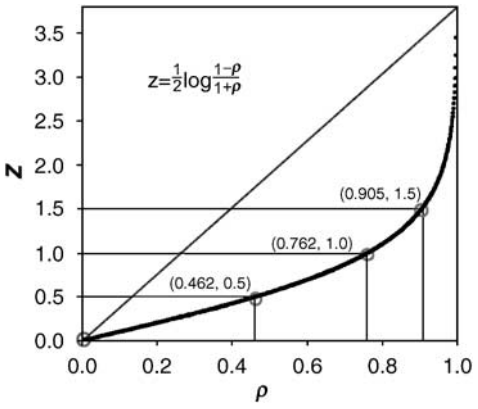

Figure 11 Fisher *z*–Pearson's ρ transform, with the four equidistant levels chosen for *z* and the corresponding values for ρ. Note that *z* has an infinite range.

We abstained from a continuous scale, because a discrete scale is faster to work with and matches better with the task at hand. No scores or values were indicated to users in order to avoid associations of categories with specific correlation values. In the section 'Data analysis', we use the frequencies of different responses in the same condition to deduce the actual sensation strengths. For the sake of the data analysis, the categorical responses are coded as numerical values −2, −1, 0, 1, and 2.

Test procedure and participants

The experiment started with a small tutorial which briefly introduced scatterplots and PCPs and how to use them to analyze correlation. Characteristic images of both visualization methods for $r=-1$, $r=0$ and $r=1$ were shown on paper. Next, there was a trial session in which participants could familiarize themselves with the test environment and the test interface. Trial samples were presented with both visualization methods and displayed in both time conditions. Afterwards, the formal test sessions started.

We wanted to use the same images for both the limited and unlimited time conditions, to enable direct comparison. In a pilot study (with 10 participants, simulating the formal test but not included in the data analysis), we found that participants spent quite a long time to investigate patterns in the unlimited time condition. As a result, patterns could be remembered, which led us to decide to always start in the limited time condition. In order to avoid an effect on the order in which the two visualization methods were tested, we used both orders. Consequently, there were two orders of the four sessions: A. (*sc*, *ld*)→(*pc*, *ld*)→(*sc*, *ud*)→(*pc*, *ud*); B. (*pc*, *ld*)→(*sc*, *ld*)→(*pc*, *ud*)→(*sc*, *ud*). In the formal test, participants were randomly assigned to one of these two orders.

For each session and subject, the 42 stimuli created by combining pre-generated sample data sets and visualization conditions were displayed in random order. This arrangement aimed to average out learning effects. During the test sessions, participants were allowed to refer to the characteristic graphs from the tutorial since most of them were not familiar with PCPs. Finally, participants were interviewed to give their subjective comments. The experiment was performed by 25 participants. They were university Ph.D. students or faculty from different departments and between 24 and 45 years old. All of them knew the concept of correlation in statistics. Most of them had experienced scatterplots before, and one-fourth of them knew the concept of PCPs, but none of them had actually used PCPs to analyze correlations. All subjects however commented that the pre-test tutorial and trial session supplied enough information for them to do correlation analysis with the help of PCPs.

Test environment

The experiment was carried out in an office environment during daytime. The evaluation program was running on a desktop PC. The graphs were displayed on the left-hand side of the PC screen in an area of $24 \times 26\,\text{cm}^2$ with white background and black objects (Figure 12). One point in a scatterplot occupied an area of $1 \times 1\,\text{mm}^2$. The same axis scale was used for all graphs. The resolution of the plotting area was 788×854 pixels.

The interface for collecting user responses was displayed on the right-hand side of the PC screen (Figure 12). Participants were asked to push one of five buttons, corresponding to five response levels. They could change their rating up to the point of clicking the 'Next' button. Since the 'Next' button also triggered the next stimulus, the participants were reminded to be prepared for this next stimulus, especially in the limited time condition. In this condition, the visualized pattern appeared right after the 'Next' button was clicked, but disappeared after 1 s. After 1 s, the graph area was blanked, while the right-hand-side input interface remained visible, waiting for an input from the test person. In this way, we were able to limit the time of the visualization, without limiting the response time. We observed in the pilot study that participants sometimes clicked the 'Next' button before they had finalized their response. Although we explicitly reminded our participants of this potential mistake before the formal test, there were five judgments of three subjects that required correction afterwards by the experiment leader who kept track of such occurrences.

Data analysis

The effectiveness of the different visualization methods for correlation assessment can be characterized in a number of different ways. One option is to estimate the discriminative power (performance accuracy), which we define as the number of levels of correlation that users can distinguish reliably in a specific condition. Another option is to establish how closely the actual human perception approximates the theoretically optimal performance of the Fisher *z* scale. A third option is to establish

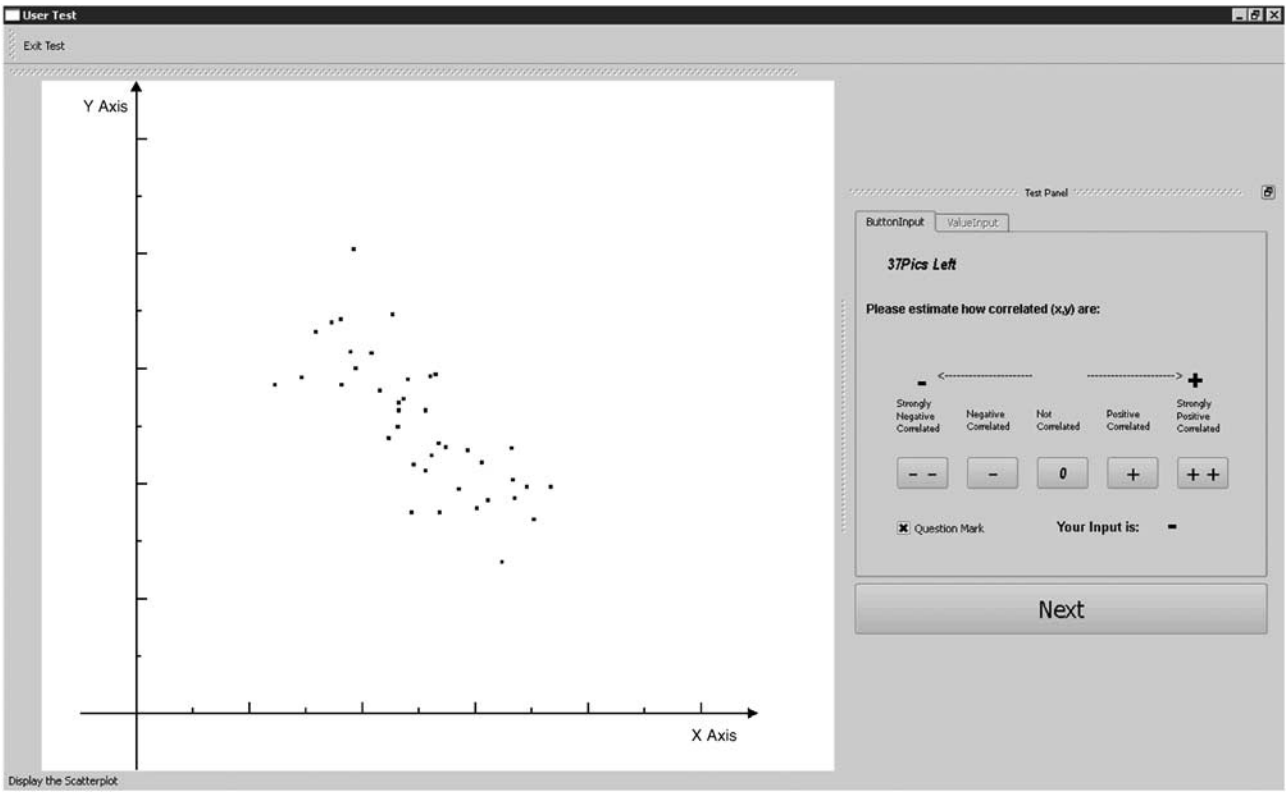

Figure 12 The interface of the experimental program.

whether or not there are systematic biases in the correlations being reported, such as a bias towards negative over positive correlations.

In order to quantitatively study these aspects, we need to derive the actual sensations (expressed by the average sensation values p_{ki} in the model of Experimental Design (see the section 'Model') and the internal noise σ_k from the observed response data u_{ki}. We report the performance accuracy as a function of visualization method, inspection duration and sample size below in the subsection. We also establish whether or not sensation values that are proportional to the Fisher z scale are able to explain the observed response data. Any deviation of the sensation scale from the Fisher z scale and any offset from the origin will be interpreted as signs of perception bias.

Figure 13 shows the distributions of the response data in the four different test sessions. The x-axis present the z values of the stimuli, while the y-axis presents the corresponding user response value u. The area of the bubbles visualizes the number of observations for all combinations of z score and response value u. Observations are accumulated over all subjects and sample sizes n. An approximate linear relationship passing though the origin seems to exist between z and u in all cases. Therefore we can hypothesize linearity between the average subjective sensations p_{ki} and the Fisher z scores. The results of statistical tests of this hypothesis are reported in the section 'Bias analysis and comparing the sensation scale with z'.

Accuracy under different conditions

The response data per condition k (i.e., every unique combination of visualization method, observation duration and sample size) were analyzed using the statistical modeling tool XGMS,[29] in the sense that the relevant parameters of the statistical model that we introduced in the model of Experimental Design (see the section 'Model'), that is, the average sensation values p_{ki}, for $i = 1, \ldots, I$, and the noise standard deviation σ_k, were estimated from the observed response frequencies.

In order to describe how the estimated parameters can be used to derive a performance index, we need to borrow some well-known facts from statistical detection and estimation theory.[29] Detection theory describes how well two conditions can be distinguished based on noisy measurements. As shown in Figure 14(A), a large overlap between the distributions of the measurement variable in both conditions gives rises to a substantial percentage of errors. In case both distributions are normal with equal standard deviation σ but different average values μ_1 and μ_2, then the percentage of errors P_ε is determined by the *discriminability index* $d' = (\mu_1 - \mu_2)/\sigma$ (d-prime). A $d' = 1$ corresponds to a just noticeable difference (JND) which

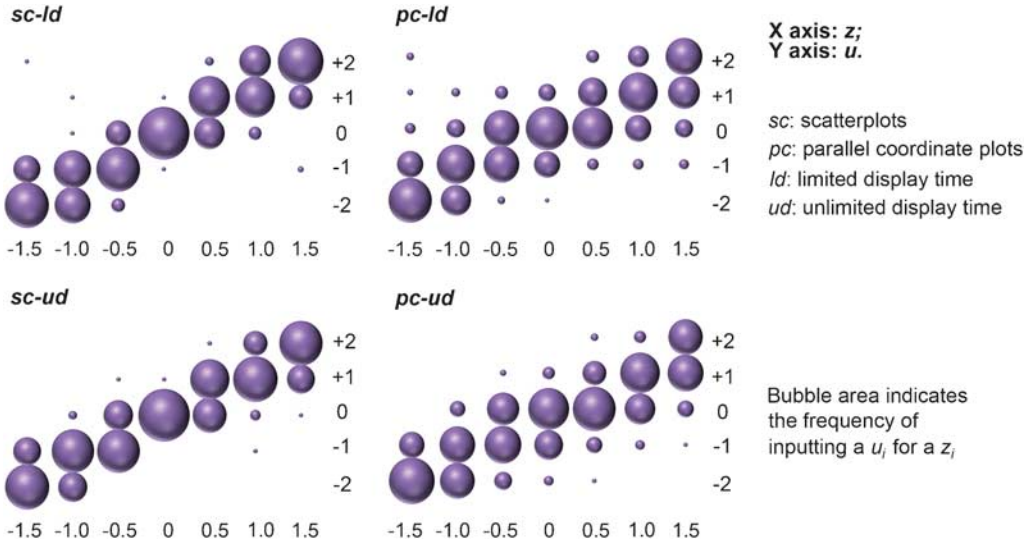

Figure 13 Distributions of observed responses in the four test sessions.

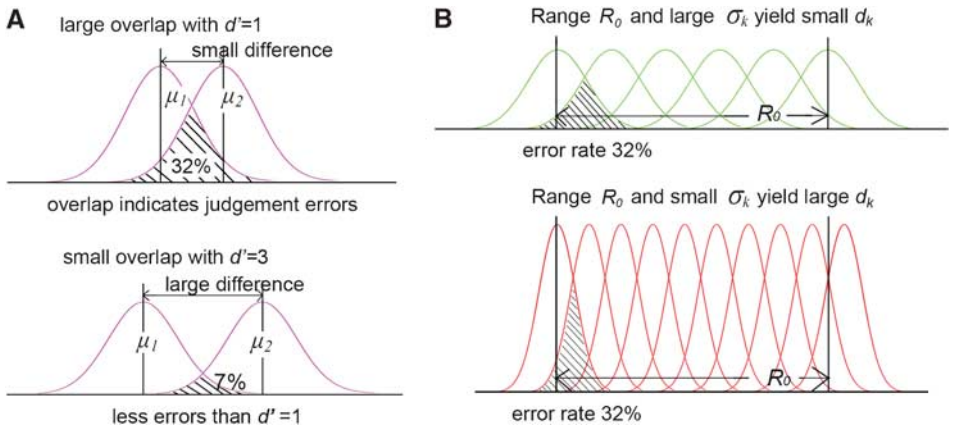

Figure 14 (A) Discriminability index *d*-prime from signal detection theory: if the overlap between the distributions of the measurement variables on which detection is substantial (such as when $d' = 1$), then a substantial percentage of errors is expected; if the overlap is small (such as when $d' = 3$), then few errors are expected; (B) For the same uniform overlaps (i.e., the same error rates 32% are expected as when $d' = 1$), large σ_k (top) yield fewer discriminable levels (corresponding to a smaller d_k) than when σ_k is small (bottom).

still gives rise to a substantial error rate of $P_\varepsilon = 32\%$, while $d' = 3$ corresponds to a difference that can be judged reliably, that is, with a small error rate of $P_\varepsilon = 7\%$. In case several levels need to be distinguished, we can add discriminability indices between successive levels to obtain a discriminability (accuracy) index for the entire range. In our specific case, this results in

$$d_k = R_k/\sigma_k, \tag{9}$$

where $R_k = \max_i(p_{ki}) - \min_i(p_{ki})$ $(i=1,\ldots,I)$ is the range of the average sensations for the correlation strengths used in the experiment. The index d_k is the normalized sensation range (normalized by σ_k) and indicates the number of JNDs under condition k (i.e., the number of levels that can be distinguished with a *d*-prime of one). Higher values of d_k indicate that more levels can be distinguished reliably, as is illustrated in Figure 14(B). Tables 1 and 2 give an overview of the obtained d_k values in the different experimental conditions.

There exists a theoretical upper limit for d_k. More precisely, since the z scale is the theoretically optimal scale (with constant variance), and z has a known normal distribution $N(Z(\rho, 1/(n-3)^2)$, the expected value for the discriminability index is

$$d_z = \frac{Z_{max} - Z_{min}}{\sigma_z} = \frac{Z_{max} - Z_{min}}{1/\sqrt{n-3}} = 3\sqrt{n-3}, \tag{10}$$

Table 1 Accuracy index d_k for scatterplots

	n		
T	*10*	*40*	*160*
Limited	7.70	13.14	14.59
Unlimited	8.84	11.36	22.25
All time conditions	7.99	10.99	15.45

Table 2 Accuracy index d_k for PCPs

	n		
T	*10*	*40*	*160*
Limited	3.51	5.34	6.11
Unlimited	5.17	6.01	6.50
All time conditions	4.20	5.57	6.18

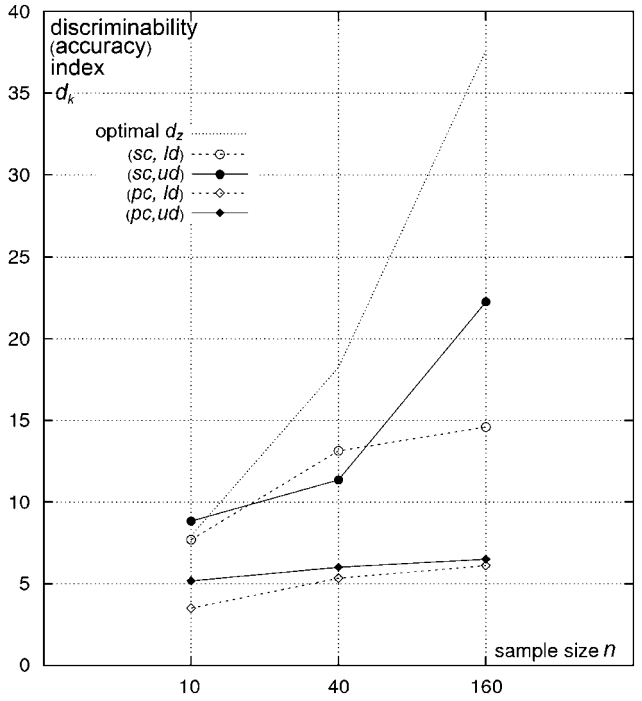

Figure 15 Summary of accuracy under different conditions.

where $Z_{max}=1.5$ and $Z_{min}=-1.5$ in our case. This optimal index d_Z only depends on the sample size n. The upper limits for the discriminability index are hence expected to be at 7.94 for $n=10$, 18.25 for $n=40$ and 37.59 for $n=160$. Figure 15 plots discriminability indices, together with the theoretical upper limits, for the different conditions considered in the experiment. By comparing indices d_k for different conditions, we can establish the influence of different independent variables on the accuracy.

Effect of visualization method In Table 1, d_k values are provided for scatterplots under different conditions. The lowest value is 7.70 under limited display time and sample size $n = 10$. If we adopt $d' = 3$, for a reliable distinction between successive levels, then this result implies that only three levels can be reliably distinguished in this condition (indicated by $d_k/3$, since d_k is calculated as when $d' = 1$). The highest d_k value is 22.25, or seven clearly distinguishable levels, for an unlimited display time and a sample size equal to $n = 160$. Figure 16 shows 22 levels of correlation, with uniform sampling of the z value, in scatterplots with $n = 160$. Note that these levels are indeed close to what is just distinguishable.

From Figure 15, we can see that for $n=10$, scatterplots allow to fully exploit the information in the data. For $n=40$ and $n=160$, the performance increases, but stays well below the theoretical optimal limit. A confusing case might be (*sc*, *ud*) and $n=10$, where the estimation of d_k is slightly above the optimal limit. This case is discussed later about confidence intervals of d_k in the section 'Confidence interval on d_k and d_Z'.

In Table 2, d_k values are provided for PCPs under different conditions. The lowest value is 3.51 for limited display time and sample size 10. It indicates that users cannot make very reliable judgments for correlation. The highest value is 6.50 for unlimited display time and sample size 160, which demonstrates that reliably distinguishing three levels of correlation is what can be expected at best.

Comparing both visualization methods, the highest performance of PCPs is even lower than the lowest performance of scatterplots and the average performance of scatterplots is at least twice as high as that of PCPs. All the evidence hence implies that the accuracy of judgment for PCPs is much lower than that for scatterplots.

Effect of display time and sample size For both visualization methods (Figure 15), d_k values for unlimited viewing $\{T = ud\}$ are almost always larger than d_k values for restricted viewing $\{T = ld\}$ in case of identical sample size. This indicates that giving users more time to observe helps them to distinguish more levels of correlation, especially in case of scatterplots. The only observed exception is for scatterplots with $n=40$.

We have also measured the time to complete each test session per participant. The average times for all participants and the differences between different sessions are shown in Table 3. A paired *t*-test was used to compare the mean of the completion times of the four test sessions. The mean completion time of (*sc*, *ld*) was 62 s lower (33% faster) than that of (*pc*, *ld*); and the mean completion time of (*sc*, *ud*) was 92 s lower (39% faster) than that of (*pc*, *ud*). Both cases are statistically significant with *P*-values smaller than 0.001. It is also interesting to observe that for PCPs, the mean completion time in the limited display time session (*pc*, *ld*) is significantly lower than that in the unlimited display time session (*pc*, *ud*), while for

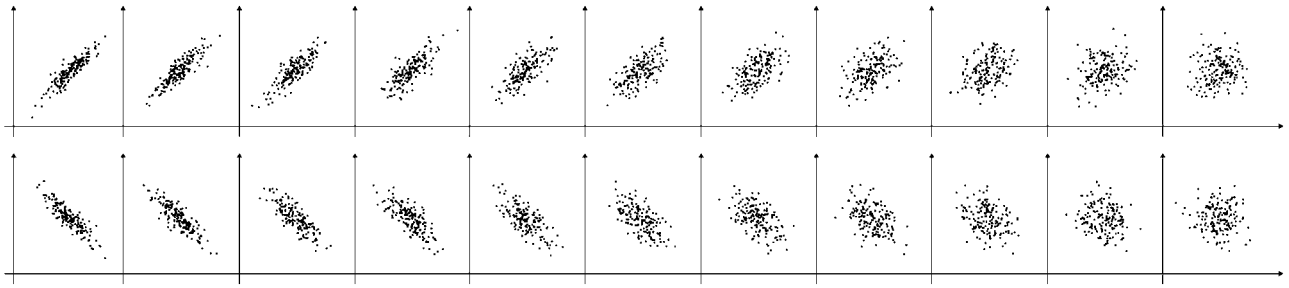

Figure 16 An example of 22 levels of correlation visualized by scatterplots. According to our estimation, the discriminability between two neighboring scatterplots is similar as when $d' = 1$. If we assume that $d' = 3$ is required to be clearly distinguishable, then the clearly discernable levels should be such that there are two scatterplots in between them in this figure.

scatterplots, the mean completion time of *ld* is not significantly lower than that of *ud*. These results seem to indicate that people can perceive correlation easily and quickly with scatterplots. For PCPs, it is more difficult both to sense signals and to perceive correlations. When display time is limited, people spend more time to interpret the perceived signal and to generate an output response when using PCPs instead of scatterplots. When display time is unlimited, people also spend more time for viewing stimuli and reporting responses, which seems to point at an increase in cognitive load when using PCPs.

For both visualization methods, the judgment accuracy increases with increasing sample size within the experimental range (for our case, up to 160 items). This indicates that a larger sample size leads to a more accurate judgment. For scatterplots, the accuracy improves substantially for larger sample sizes, while for PCPs the improvement is very modest.

Confidence interval on d_k ***and*** d_z Although we have estimated the values of d_k from our experimental data, we need to recognize that this estimate is actually a stochastic variable. We would very likely obtain a different set of responses in case we would repeat the same experiment. Using such new data would result in (slightly) different values for the model parameters, and hence also for the deduced performance variable d_k. In order to properly judge differences in this performance variable, the (for instance, 90%) confidence intervals (CI) should hence be computed for this variable. However, to compute the CI for d_k by means of experiment repetition is not feasible. A common statistical practice is to perform re-sampling of the obtained experimental data. We consider the CI obtained through re-sampling for two special cases: (*sc*, *ud*), $n = 10$ where the d_k value is slightly above the optimal value d_z; and (*sc*, *ud*), $n = 160$ where the d_k value seems surprisingly high. Table 4 shows the 90% confidence intervals that were obtained for d_k using the re-sampling procedure. Note that the optimal boundary d_z is also a stochastic variable that should be treated in a similar way. The corresponding CI of d_z is also shown in Table 4 (although it was obtained by re-sampling, it could also be derived from our knowledge of the average standard deviation for this theoretical case).

Table 3 Means of completion times

In seconds	*sc*	*pc*	*pc-sc*
ld	127	189	62
ud	144	236	92
ud-ld	17	47	

Table 4 Confidence intervals at 90%

	n		
Var	*10*	*40*	*160*
d_z	[4.01, 11.85]	[14.95, 21.53]	[34.28, 40.85]
d_k (sc, ud)	[8.16, 9.55]		[20.67, 23.20]

We can see that the CI of d_z for $n = 10$ contains the CI of d_k for (*sc*, *ud*) and $n = 10$. This can explain why the d_k value from our experiment could be slightly above the average optimal value for d_z. The CI of d_z for $n = 160$ is completely above the CI of d_k for (*sc*, *ud*) and $n = 160$, which indicates the discriminability of human sensors is substantially below that of an optimal estimator when the sample size is large. At the same time, the CI of d_k for (*sc*, *ud*) and $n = 160$ is small and precise, which provides confidence in the accuracy of our estimate.

Bias analysis and comparing the sensation scale with z

The perception and judgment bias can also be assessed by comparing the human sensation scale with the Fisher z scale. This deviation can be assessed by observing the non-linearity between the estimated sensation values and the optimal z values. A lack of symmetry or an offset from the origin in the sensation scale points at a different treatment of positive and negative correlations. We first report the results of a statistical test on linearity, before providing the actual sensation values and assessing the evidence for asymmetric behavior.

Table 5 Chi square χ^2 (*P*-value) for scatterplots

	n		
T	*10*	*40*	*160*
Limited	10.50 (0.062)	78.06 (0.000...)*	31.02 (0.000...)*
Unlimited	10.33 (0.067)	22.77 (0.000...)*	13.93 (0.016)
All time conditions	15.55 (0.008)	84.43 (0.000...)*	31.11 (0.000...)*

**P*-value is smaller than 0.001.

Table 6 Chi square χ^2 (*P*-value) for PCPs

	n		
T	*10*	*40*	*160*
Limited	6.45 (0.265)	7.81 (0.167)	19.04 (0.002)
Unlimited	7.22 (0.205)	9.44 (0.093)	16.64 (0.005)
All time conditions	13.42 (0.020)	12.54 (0.028)	28.55 (0.000...)*

**P*-value is smaller than 0.001.

Non-linearity bias Hypothesizing the existence of a linear relationship between sensation values and *z* scores is equivalent to assuming a simplified model where the sensation values in the statistical model of Figure 8 can be expressed as

$$p_{ki} = f_k z_i + c_k \ (i = 1, \ldots, I). \tag{11}$$

In the original model, there are $I = 7$ parameters p_{ki}, where each parameter specifies the sensation value for one of the seven correlation levels. In the simplified model, there are only two parameters f_k and c_k required to specify the sensation values, since the z_i values are known. Hence, the simplified model has five parameters less than the more general model. Since we use a maximum likelihood estimator, we can compare the likelihood of both models, using a chi-squared test (with five degrees of freedom).[29] More precisely, if the difference in likelihood exceeds the threshold set by the chi-squared test, then we can reject the hypothesis that both models are equivalent. Chi-squared tests most often choose the boundary such that the confidence of false rejection is below 0.05. If the difference in likelihood is below the threshold, then we have insufficient evidence for rejecting the hypothesis that the sensation values are linearly related to the *z* scores. Tables 5 and 6 present the χ^2 statistics obtained from the comparisons between both models under different conditions. A small *P*-value (the most widely used threshold for the *P*-value is 0.05) stands for a significant difference between the optimal model and the linear regression model, thus indicating a significant non-linearity. In case we adopt 0.05 as the threshold value, any *P*-value larger than 0.05 indicates insufficient evidence for non-linearity between sensation values and *z* scores.

For scatterplots, the *P*-values are larger than 0.05 when $n = 10$ which indicates no evidence for a non-linear relationship. In case of scatterplots with $n = 40$ or $n = 160$, the *P*-values are quite small, which is evidence for a significant non-linearity. For PCPs, *P*-values are larger than 0.05 in case of $n = 10$ and $n = 40$. The sensation scales for PCPs hence seem to be more linearly related to *z* scores than in the case of scatterplots. Given that the range of sensation values is much smaller in case of PCPs, this should not be considered very unexpected.

In cases where there is evidence for a non-linear relationship between *z* scores and sensation values, we of course would like to obtain more insight into the nature of this non-linearity. This can be obtained by plotting sensation values as a function of *z* scores in the different experimental conditions. Such a visualization will reveal not only deviations from linearity but also evidence for any asymmetric behavior (a different treatment of positive and negative correlations). In order to properly compare the sensation values p_{ki} across different experimental conditions *k*, we normalize them by dividing them by the standard deviation σ_k of the estimated noise:

$$\bar{p}_{ki} = \frac{p_{ki}}{\sigma_k} \quad (i = 1, \ldots, I). \tag{12}$$

This definition is also in agreement with our earlier discussion, since

$$d_k = \frac{\max_i(p_{ki}) - \min_i(p_{ki})}{\sigma_k} = \max_i\left(\frac{p_{ki}}{\sigma_k}\right) - \min_i\left(\frac{p_{ki}}{\sigma_k}\right)$$
$$= \max_i(\bar{p}_{ki}) - \min_i(\bar{p}_{ki}). \tag{13}$$

Figs. 17 and 18, respectively, present the relationship between $\bar{p}_{ki}$ and z_i under different conditions of sample size and viewing time.

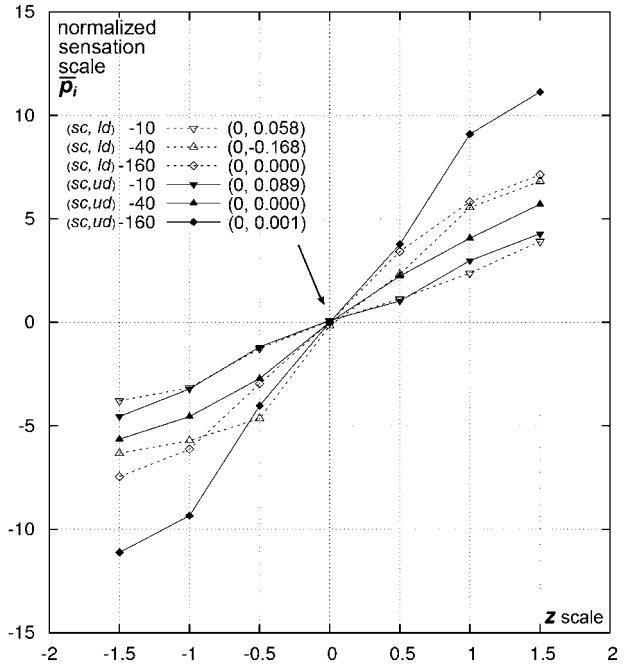

Figure 17 Relationships between the normalized average sensation values and the Fisher *z* scores for scatterplots. The arrow indicates the offsets in judgment for zero correlation.

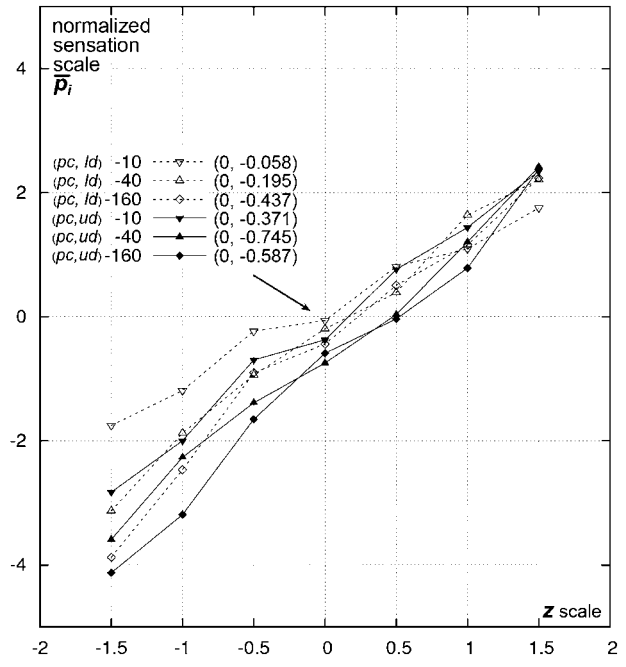

Figure 18 Relationships between the normalized average sensation values and the Fisher *z* scores for parallel coordinate plots. The arrow indicates the offsets in judgment for zero correlation.

Comparing Figure 17 with Figure 18, we observe that the normalized sensation scores for PCPs are indeed more linearly related to *z* scores than in case of scatterplots.

Table 7 Corrected binomial test $Z(P$-value) for PCPs

	n		
T	*10*	*40*	*160*
Limited	−0.14 (0.444)	−0.71 (0.239)	−1.70 (0.045)
Unlimited	−0.99 (0.161)	−2.55 (0.005)	−2.12 (0.017)

A compression of the sensation scale, such as observed for the extreme *z* scores in case of scatterplots, implies that the visualization method is more sensitive for discriminating small correlations than for discriminating high correlations. For PCPs, the most obvious characteristic is the offset at a *z* score of zero, which is discussed in more detail in the next section. It indicates a bias towards reporting negative correlations, even in cases where the true correlation value is (slightly) positive.

Another interesting issue here is that the *z* scores have an infinite range while human judgments can be assumed to be restricted to a finite range. Therefore, obtaining a sensation scale that is linearly related to *z* scores over the entire range of *z* scores is not possible. In other words, linearity inevitably needs to be broken when the range of *z* scores increases. The evidence for this is obviously most pronounced in case of scatterplots.

Offset bias Negative offsets for PCPs are shown in Figure 18 while no discernible offsets can be observed for scatterplots. To test the significance of the offset, we extract the frequency tables of u_i when $z_i = 0$ under different conditions. Any significant skew of the distribution will indicate a significant offset from the origin. For scatterplots, almost 100% of the responses are that $u = 0$. Therefore, there is no evidence for any offset. For PCPs, the frequencies are not symmetrically distributed. A corrected binomial test[31] is used to test the skew significance and Table 7 presents the *P*-values for a statistical test of negative skew.

There is evidence for significant negative skew in case of $\{T = ld\} \cup \{n = 160\}$, $\{T = ud\} \cup \{n = 40\}$ and $\{T = ud\} \cup \{n = 160\}$. Furthermore, 90% of the subjects commented during the post-test interview that the intersection pattern of PCPs is a dominant visual cue for them. When this pattern is clearly present, they can easily make up their decision on strong negative correlation; however, when this pattern coexists with a parallel pattern, they feel quite irritated because of its interference. Therefore, the offset bias, particularly the bias towards negative correlation, does exist for PCPs. Moreover, the bias tends to be more serious when people get more time to view the graph and more samples are visualized with PCPs.

Conclusions

The effectiveness of different visualization methods for the correlation judgment has been assessed in a number

of different, but related aspects. Firstly, it was assessed by establishing a performance index for the number of correlation levels that can be distinguished reliably; secondly, by looking for evidence of a non-linear or asymmetric relationship between sensation values and (optimal) Fisher z scores.

The discriminability index d_k is proposed as a performance measure. It is a combined indicator of all the parameters in our statistical model (the sensation values p_{ki} for the different levels i of correlation and the noise factor σ_k that describes the accuracy of the judgments). Roughly speaking, d_k is the normalized range of the sensation values, and can be interpreted as the number of correlation levels that people can judge. We have established that, for all combinations of sample size n and observation time T, scatterplots allow people to distinguish at least twice as many different correlation levels as PCPs and the performance is even approaching the theoretically optimal expressed by the Fisher z scores when sample size is small.

Another relevant observation is the deviation of the human sensation scale from the optimal z scale. The deviation has been examined in two ways: by visualizing and statistically testing the non-linear relationship between the sensation scale and the z scale; and by estimating the offset in sensation for zero correlation. The non-linearity is more pronounced in case of scatterplots than in case of PCPs. Furthermore, the non-linearity seems to be most pronounced at high correlation levels for scatterplots. For PCPs, the sensation scale is more linear with z, but negative offset at the origin is observed. People underestimate correlation in PCPs by perceiving strongly positive correlated data as positively correlated, uncorrelated (and slightly positively correlated) data as negatively correlated, and negatively correlated data as strongly negatively correlated. This result matches with the subjective comments we collected after the experiment that the intersection pattern for $r = -1$ is perceptually stronger than the parallel pattern for $r = 1$. We call this phenomenon the 'diabolo effect' in PCPs. Moreover, in the unlimited time condition with larger sample sizes, the offset is more serious, which means that the longer people look at it, the stronger the diabolo effect is evoked. We also found that people tend to overestimate negative correlations. This might also be explained by the diabolo effect, since strongly negatively correlated data lead to a more pronounced intersection pattern.

Generally speaking, scatterplots are more effective in supporting visual correlation analysis between two variables than PCPs. For PCPs, the judgment is less accurate. At the same time, a diabolo effect is introduced into the perception process of PCPs, which causes a bias towards reporting negative correlations. Since none of our subjects had previous experience with PDPs for correlation analysis, the poor performance of PCPs could be an effect of unfamiliarity. However, we could also argue that the poor performance of PCPs might lead to the unfamiliarity among users.

The statistical model of the perceived correlation established in this paper enabled us to compare and evaluate two different visualization methods. In the future we intend to study if a similar approach can also be used for other visualization aspects, such as the relation between cluster detection and visual attributes of icons.

Acknowledgements

This research is supported by the VIEW program of the Netherlands Organization for Scientific Research (NWO) under research grant no. 643.100.502.

References

1 Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization. *Proceedings of the IEEE Symposium on Information Visualization* 2005 (InfoVis'05, Minneapolis, USA), IEEE Computer Society Press: Washington, DC, USA, 2005; 111–117.
2 Anderson TW, Finn JD. *The New Statistical Analysis of Data*. Springer-Verlag: New York, 1996; 139pp.
3 Inselberg A, Dimsdale B. Parallel coordinates: a tool for visualizing multidimensional geometry. *Proceedings of the IEEE Visualization Conference* 1990 (VIS'90 San Franscisco, CA), IEEE Computer Society Press: Los Alamitos, CA, USA, 1990; 361–378.
4 Inselberg A. The plane with parallel coordinates. *The Visual Computer* 1985; **1**: 69–91.
5 Wegman EJ. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association* 1990; **85**: 664–675.
6 Slocum TA, McMaster RB, Kessler FC, Howard HH. *Thematic Cartography and Geographic Visualization*, 2nd edn. Prentice Hall Series in Geographic Information Science. Pearson Prentice Hall: New Jersey, 2005; 40pp.
7 Siirtola H. Direct manipulation of parallel coordinates. *Proceedings of International Conference on Information Visualization* 2000 (IV'00, London, UK), IEEE Computer Society Press: London, England, UK, 2000; 373–378.
8 Loh WY. Does the correlation coefficient really measure the degree of clustering around a line?. *Journal of Educational Statistics* 1987; **12**: 235–239.
9 Cleveland WS, Diaconis P, McGill R. Variables on scatterplots look more highly correlated when the scales are increased. *Science, New Series* 1982; **216**: 1138–1141.
10 Strahan RF, Hansen CJ. Underestimating correlation from scatterplots. *Applied Psychological Measurement* 1978; **2**: 543–550.
11 Erlick DE, Mills RG. Perceptual quantification of conditional dependency. *Journal of Experimental Psychology* 1967; **73**: 9–14.
12 Best LA, Hunter AC, Stewart BM. Perceiving relationships: a physiological examination of the perception of scatterplots. In: Barker-Plummer D, Cox R and Swoboda N (Eds). *Diagrammatic Representation and Inference, Proceedings of Fourth International Conference, Diagrams*. 2006 (LNAI 4045), Springer-Verlag: Berlin, Heidelberg, 2006; 244–257.
13 Kareev Y. Positive bias in the perception of covariation. *Psychological Review* 1995; **102**: 490–502.
14 Johansson J, Forsell C, Lind M, Cooper M. Perceiving patterns in parallel coordinates: determining thresholds for identification of relationships. *Information Visualization* (advance online publication 31 January 2008) http://www.palgrave-journals.com/ivs/journal/vaop/ncurrent/abs/9500166a.html (accessed 11 March 2008).
15 Forsell C, Johansson J. Task-based evaluation of multi-relational 3D and standard 2D parallel coordinates. *Proceedings of Electronic Imaging* 2007 (San Jose, CA, USA), Vol. 6495, Copublished by SPIE and IS&T: Bellingham, WA, USA, 2007; 64950C-1–12.
16 Wegenkittl R, Loffelmann H, Groller E. Visualizing the behavior of higher dimensional dynamical systems. *Proceedings of the IEEE*

Visualization Conference 1997 (VIS'97, Phoenix, AZ, USA), IEEE Computer Society Press: Washington, DC, USA, 1997; 119–125.
17 Fanea E, Carpendale S, Isenberg T. An interactive 3D integration of parallel coordinates and Star Glyphs. *Proceedings of the IEEE Symposium on Information Visualization* 2005 (InfoVis'05, Minneapolis, MN, USA), IEEE Computer Society Press: Washington, DC, USA, 2005; 149–156.
18 Tory M, Potts S, Moller T. A parallel coordinates style interface for exploratory volume visualization. *IEEE Transactions on Visualization and Computer Graphics* 2005; **11**: 71–80.
19 Ellis G, Dix A. Enabling automatic clutter reduction in parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics* 2006; **12**: 717–723.
20 Novotny M, Hauser H. Outlier-preserving Focus+Context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics* 2006; **12**: 893–900.
21 Johansson J, Ljung P, Jern M, Cooper M. Revealing structure within clustered parallel coordinates displays. *Proceedings of the IEEE Symposium on Information Visualization* 2005 (InfoVis'05, Minneapolis, MN, USA), IEEE Computer Society Press: Washington, DC, USA, 2005; 125–132.
22 Artero AO, Ferreira de Oliveira MC, Levkowitz H. Uncovering clusters in crowded parallel coordinates visualizations. *Proceedings of the IEEE Symposium on Information Visualization* 2004 (InfoVis'04, Austin, TX, USA), IEEE Computer Society Press: Washington, DC, USA, 2004; 81–88.
23 Graham M, Kennedy J. Using curves to enhance parallel coordinate visualizations. *Proceedings of the International Conference on Information Visualization* 2003 (IV'03, London, UK), IEEE Computer Society Press: London, England, UK, 2003; 10–16.
24 Lanzenberger M, Miksch S, Pohl M. Exploring highly structured data – a comparative study of stardinates and parallel coordinates. *Proceedings of the International Conference on Information Visualization* 2005 (IV'05, Greenwich, UK), IEEE Computer Society Press: London, England, UK, 2005; 3–9.
25 Kobsa A. User experiment with tree visualization systems. *Proceedings of the IEEE Symposium on Information Visualization* 2004 (InfoVis'04, Austin, TX, USA), IEEE Computer Society Press: Washington, DC, USA, 2004; 9–16.
26 Ghoniem M, Fekete J, Castagliola P. A comparison of the readability of graphs using node-link and matrix-based representations. *Proceedings of the IEEE Symposium on Information Visualization* 2004 (InfoVis'04, Austin, TX, USA), IEEE Computer Society Press: Washington, DC, USA, 2004; 17–24.
27 Irani P, Slonowsky D, Shajahan P. Human perception of structure in shaded space-filling visualizations. *Information Visualization* 2006; **5**: 47–61.
28 North C. Visualization Viewpoints: Toward Measuring Visualization Insight. *IEEE Computer Graphics and Applications* 2006; **26**: 6–9.
29 Martens J. *Image Technology Design*: *A Perceptual Approach*. The International Series in Engineering and Computer Science. Kluwer Academic Publisher: Dordrecht, 2003; 193pp.
30 McGrath RE. *Understanding Statistics – A Research Perspective*. Addison-Wesley: Reading, MA, 1996; 26pp.
31 Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd edn. McGraw-Hill: 1988, 42pp.
32 Wilson P, Tanner Jr, Theodore GB. Definitions of d' and η as psychophysical measures. In: Swets JA *et al*. (Ed). *Signal Detection and Recognition by Human Observers*. John Wiley & Sons, Inc.: New York, 1964; 147–163.
33 Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 2001; **24**: 87–185.