Visualising Concordance

Gaj Vidmar[1] and Nino Rode[2]

[1] Institute of Biomedical Informatics, Faculty of Medicine, University of Ljubljana, Vrazov trg 2, SI-1000 Ljubljana, Slovenia
[2] Faculty of Social Work, University of Ljubljana, Topniska 33, SI-1000 Ljubljana, Slovenia

Summary

Concordance describes the agreement between *m* rankings of *k* objects. Despite the long history of measures of concordance and the recently revived interest in comparison of concordance (c.f. Legendre 2005), the task of visualising concordance remained virtually unaddressed. We first show how to depict concordance by simply plotting raw data in parallel coordinates. Then we review further possibilities for depicting concordance using the recently developed plots of inter-rater variability in ordinal ratings (Nelson & Pepe 2000) and plots of correlation matrices (Trosset 2005). Next, we propose two novel concordance plots. The concordance bubble-plot is based on raw rank data, while the pin-cushion plot depicts rank differences in polar coordinates. We present visualisations of artificial and real-life datasets with different degree of concordance and identify strong and weak points of the proposed plots. In conclusion, we review some other work related to visualisation of concordance and discuss some other options for constructing novel concordance plots.

Keywords: Concordance coefficients, Graphical methods, Ranks, Parallel coordinates, Polar coordinates, Bubble plot, Pin-cushion plot

1 Introduction

The problem of concordance pertains to m rankings of k objects. Such data are frequently gathered in human resources management, education, marketing and elsewhere, when job applicants, promotion candidates, products, political parties or other subjects or objects are ranked by executives, experts, focus groups, other human observers or even automated algorithms. There are different definitions and measures of concordance, so before considering the issue of visualising concordance, we briefly review the various concordance coefficients.

A number of sources can serve as the basis for such a review (Palachek & Schucany 1984, Siegel & Castellan 1988, Legendre & Lapointe 2004). Several concordance measures have been proposed, but with the exception of top-down correlation (Iman & Conover 1987), they are all based on bivariate ordinal correlation – either Spearman's ρ or Kendall's τ. The best-known and most widely used solution is the one proposed by Kendall & Babington-Smith (1939). Their coefficient of concordance W is basically the average ρ of all possible pairs of rankings (1), and it is related to the Friedman's test (repeated-measures analysis of variance using ranks) in the sense that Friedman's χ^2 statistic provides the means for testing significance of W against the null-hypothesis of no concordance:

$$W = [(m - 1)\, \mathrm{E}(\rho) + 1] / m;\ \chi^2 = W\, m\, (k - 1),\quad df = k - 1. \tag{1}$$

Kendall's coefficient of agreement is obtained the same way as W except that τ is used instead of ρ, which can be written as $u_K = W_\tau$. Ehrenberg (1952) proposed the simple uncorrected average of τ's as a coefficient of agreement, so the resulting coefficient (2) is linearly related to u_K:

$$u_E = \mathrm{E}(\tau_{\text{ranking-ranking}}) = [u_K\, (m/(m - 1)) - 1] / (m - 1). \tag{2}$$

As first proposed by Lyerly (1952), by changing τ back to ρ (and hence u_K back to W) in the above formula, one obtains an analogous concordance measure as the expected value of ρ's ($u_L = \mathrm{E}(\rho_{\text{ranking-ranking}})$). Finally, if a selected ranking is considered a target or benchmark with respect to the other rankings, the appropriate measure is the correlation between a group of judges and a criterion, which is defined as the average of individual judge-criterion correlations ($T_C = \mathrm{E}(\tau_{\text{ranking-criterion}})$). This coefficient is also computationally linked to some of the methods for comparing concordance between groups, reviewed in Section 3.

2 Existing Possibilities to Depict Concordance

Having absorbed the statistical notion of concordance, one can start looking for its graphical representation. However, we did not find any indication that the task of visualising concordance has been directly addressed so far. Nevertheless, without inventing any new statistical graph type, the rank data for concordance analysis can be directly depicted using parallel coordinates (Inselberg 1985). As shown by Wilkie (1980), Kendall's τ can be calculated from and depicted by the number of intersections of lines between corresponding points in the two rankings plotted in parallel coordinates. Similarly, the more intersections we see if we plot concordance data (with k lines) in parallel coordinates with each axis (of the m axes) representing one judge, the less there is concordance, and vice versa.

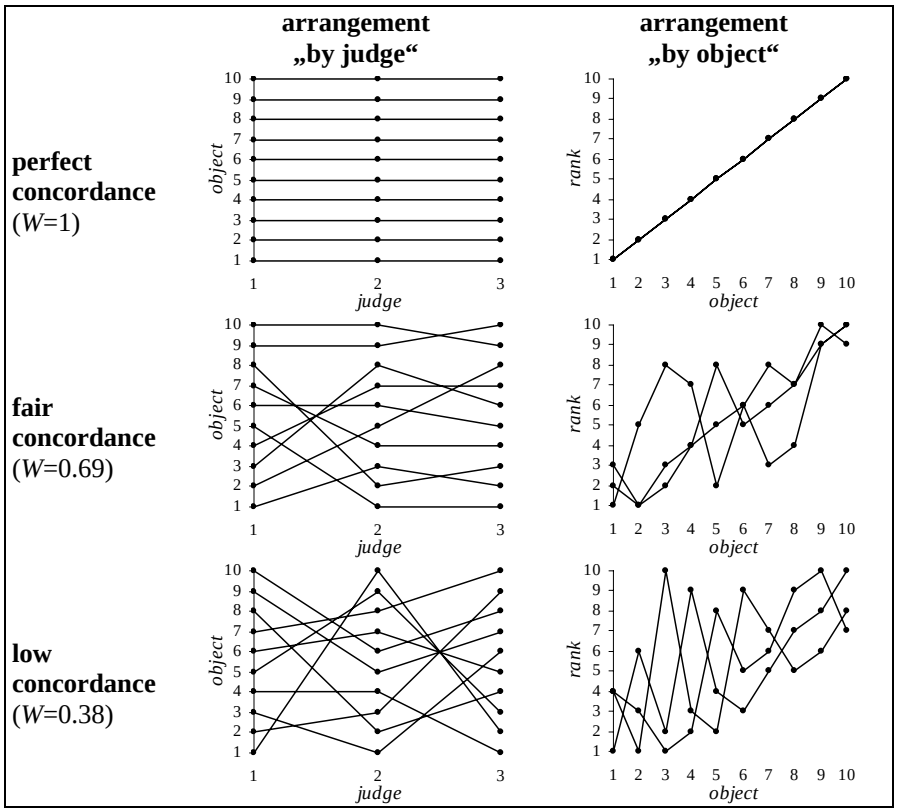

Figure 1: Representation of concordance data (ranking of k=10 objects by m=3 judges) using parallel coordinates for different levels of concordance (perfect, fair and low): parallel axes can be judges (left) or objects (right, sorted by increasing mean rank).

Such arrangement „by judge“ is one of the two possibilities for depicting raw concordance data in parallel coordinates; the other is „by object“, i.e., using one line (of the *m* lines) for each judge with each axis (of the *k* axes) representing one object, whereby perfect concordance corresponds to completely overlapping lines. In the arrangement „by judge“, judges can be ordered on the basis of mutual similarity determined by *a posteriori* tests and clustering proposed by Legendre (2005), whereas in the arrangement „by object“ it is useful to order objects by mean rank. Different degrees of concordance are depicted using both arrangements in Figure 1.

The closest match to concordance plots among special-purpose visualisation is graphical description of inter-rater variability in ordinal categorical ratings in medical setting (Nelson & Pepe, 2000), but that method has slightly different scope. Its main idea, i.e., that the mean-variance relationship of ordinal data requires rating distributions to be examined separately for different mean rating levels in order to get a full description of inter-rater variability, does also apply to mean ranking levels and inter-judge variability, and we do apply it in the concordance bubble-plots introduced in Section 4.1. However, the principle to simply add a third dimension for the mean rating level to the simple histogram display of total variability for all objects is more effective for sample sizes much larger than those usually encountered in concordance studies. Furthermore, the method extends naturally to more possible values (i.e., continuous data, with the distributions possibly smoothed), rather than to fewer possible values (i.e., rank data, especially in case of ties).

The third existing possibility to depict concordance data is offered by visualisation of correlation matrices proposed by Trosset (2005). He based his method on angular representation of product-moment correlation using *h*-plots (Corsten & Gabriel 1976, Seber 1984), and extended it to arbitrary correlation matrices using multidimensional scaling. If applied to matrices of Spearman rank-correlation coefficients, the method seems to be the natural graphical companion to Legendre’s method of testing concordance (i.e., clustering of judges), but it turns out to have notable drawbacks. The method depicts perfect concordance correctly with overlapping lines at three o’clock position (and practically unambiguously, since the lines always overlap and the attained minimum of the objective function is always below 10^{-8}), but the algorithm fails to converge to a correct and unique representation of zero concordance. We found that with the default parameters, the code (http://www.math.wm.edu/~trosset/Research/MDS/vc.s), which we ported to R (R Development Core Team, 2004), also failed to converge even with statistically significant concordance (W=0.675, p=0.033) for correlation matrix based on k=m=4. It kept converging, though, for lower and non-significant concordance with a somewhat larger dataset (W=0.367, p=0.088; objective function 2.692; k=4, m=6; data on French wine experts from Schucany & Frawley 1973, presented in more detail in Section 4.1). In any case, moderate concordance

with small k and n is to be expected in real-life applications, and the method can fail with such data. Moreover, the angular correlation-matrix plots provide no information on objects and their mean ranks. Nevertheless, our approach shares a basic idea with them, namely polar co-ordinates, which we use in pin-cushion plots (Section 4.2).

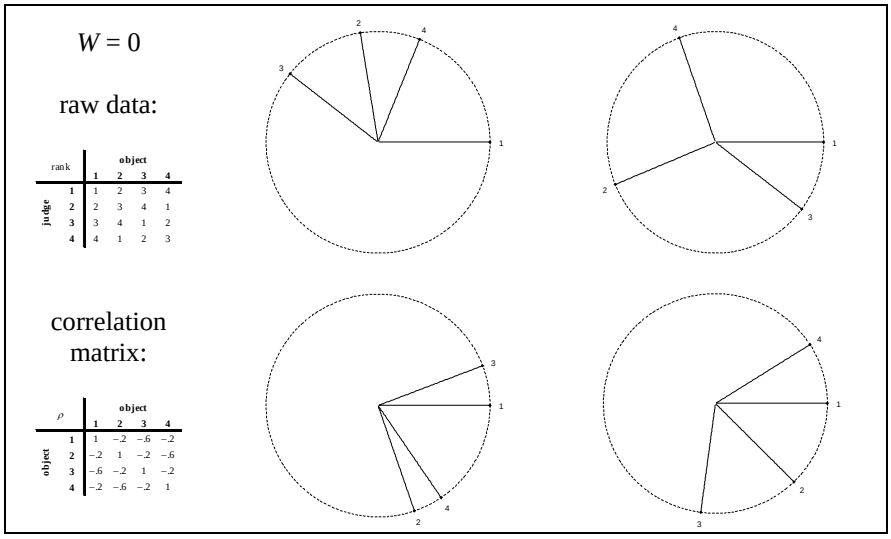

Figure 2: Representation of concordance data (k=m=4) using visualisation of Spearman rank-correlation matrices according to Trosset (2005): for zero concordance, the method fails to converge (raw data, correlation matrix and four examples of resulting plots given).

The deficiencies of the exisitng possibilities motivated us to develop new ways of visualising concordance. A further motive was that in addition to depicting a dataset corresponding to a single concordance coefficient, visualisation of concordance should also serve as a means for quickly and efficiently comparing concordance between groups or samples, or at least to clarify the associated statistics, and none of the exisitng possibilities is particularly suitable for that purpose. Like with concordance coefficients, we briefly review the statistics for concordance comparison first, and then proceed with visualisation.

3 Comparison of Concordance between Groups

The basis of the majority of the multi-group concordance analysis methods is the two-group $\mathcal{L}$ statistic (3), introduced by Schucany & Frawley (1973). It is a generalization of the logic of the test for ordered alternatives and its L statistic

(Page 1963). The goal is to test concordance both within and between two groups of judges (of size m_1 and m_2, respectively) ranking k objects. If the judges (rankings) are represented by m_1 and m_2 rows and the objects are represented by k columns of the two respective tables of ranks, and $\boldsymbol{s}$ and $\boldsymbol{t}$ denote vectors of column sums in those two tables (with elements S_i and T_i, respectively), the statistic is $\mathcal{L} = \Sigma_{i=1..k} S_i T_i = \boldsymbol{s}^{\mathrm{T}} \boldsymbol{t}$. The appropriate statistic for testing the null hypothesis of no concordance between groups is $\mathcal{L}^*$ (3), which is based upon the expected value and variance of $\mathcal{L}$ (4), which, in turn, are derived from $\mathrm{E}(S_i)$, $\mathrm{Var}(S_i)$ and $\mathrm{Cov}(S_i, S_j)$:

$$\mathcal{L}^* = [\mathcal{L} - \mathrm{E}(\mathcal{L})] / \mathrm{Var}(\mathcal{L})^{1/2}; \tag{3}$$

$$\mathrm{E}(\mathcal{L}) = m_1 m_2 k (k + 1)^2 / 4; \; \mathrm{Var}(\mathcal{L}) = m_1 m_2 (k - 1) k^2 (k + 1)^2 / 144. \tag{4}$$

After appraising the normal approximation of $\mathcal{L}^*$, Schucany and Frawley proposed a standardization of $\mathcal{L}$ due to its finite range (6). The resulting statistic (5) is confined to the usual interval of a correlation measure: $\mathcal{W} = 0$ indicates no concordance within and between groups; $\mathcal{W} = 1$ when there is complete agreement within each group on the same rank ordering; and $\mathcal{W} = -1$ when there is perfect agreement within each group but the two groups manifest opposite ordering:

$$\mathcal{W} = [\mathcal{L} - \mathrm{E}(\mathcal{L})] / [\max(\mathcal{L}) - \mathrm{E}(\mathcal{L})] \quad \in [-1, 1]; \tag{5}$$

$$\min(\mathcal{L}) = m_1 m_2 k (k + 1) (k + 2) / 6; \; \max(\mathcal{L}) = m_1 m_2 k (k + 1) (2k + 1) / 6. \tag{6}$$

The two-group test was soon extended to multiple groups: Beckett & Schucany (1975, 1979) introduced ANACONDA, multi-group analysis of concordance, named on the basis of the similarity of the concept to ANOVA. However, the method found very little application. Recently, an attempt was made to revive and improve it (Vidmar & Cernigoj 2004), and we refer the interested reader to that paper (as well as the two original papers) for explanation and discussion of the ANACONDA procedure.

Legendre (2005) took the most modern approach towards concordance analysis, motivated by the problem of species association in ecology. He devised a permutation test of W that allows for *a posteriori* tests to determine which of the individual judges are concordant with one or several of the other judges. To preserve a (approximately) correct experiment-wise error rate, p-value adjustment using the Holm (1979) procedure is applied in these tests. The paper provides a clear and detailed explanation of the permutation testing procedures with an excellent illustrative example, and it is accompanied by publicly available software (from http://www.bio.umontreal.ca/casgrain/en/labo/kendall_w.html).

4 New Plots

We introduce two new plots. The concordance bubble-plot is based on raw rank data, while the pin-cushion plot depicts rank differences. Each plot is described in a separate section below.

4.1 Concordance Bubble-Plot

Analogously to the inter-rater variability plots of Nelson & Peppe (2000), concordance bubble-plots display the frequency of assigned ranks as a function of mean rank (i.e., object). They are therefore a particular type of scatter-plots with circle size used to code the number of identical points and object names used as labels for the mean ranks. Representations of different degrees of concordance with concordance bubble-plots are given in Figure 3.

Perfect concordance is represented with all the circles on the main diagonal (upper left panel), and the main diagonal is therefore drawn in each plot (dashed line) to facilitate judgment of discordance among judges. Without tied ranks, for many combinations of m and k the minimum possible concordance is greater than zero, so the mean ranks are not equal, but they are still very close to each other (upper right panel) compared to the situation of perfect concordance in which they are as far apart as possible. The plots in the bottom two panels depict the now classic concordance data of Schucany & Frawley (1973) on two groups of wine experts ranking four wines, whereby the nine American experts tend to agree more among themselves (W=0.60, bottom left panel) than the six French experts (W=0.37, bottom right panel). In general, concordance bubble-plots have a high data-ink ratio and provide crucial additional information for concordance data.

The issue of circle size deserves some additional consideration. Namely, had the area of the circles been made proportional to the number of rank-pairs, the radius for frequency f should have been square root of f times the radius for frequency 1, while with the radius proportional to f, one might dismiss the resulting graph as a typical example of the „lie factor“ (Tufte 1998). However, it is well known from the huge body of research in perception and psychophysics (c.f. Gescheider 1985) that for area as a visual stimulus, a concave increasing function is a good first approximation for the general relationship between actual and perceived magnitude (call it Fechner’s Law, or Stevens’ Law with power less than 1), so a circle with radius 2 is actually perceived as having not much more than twice the area of a circle with unit radius. Hence, without resorting to a complex psychophysical scaling study of questionable benefit, we produced the sample concordance bubble-plots with radius proportional to f, while it is up to the user to decide what radius scaling to apply in visualisation practice.

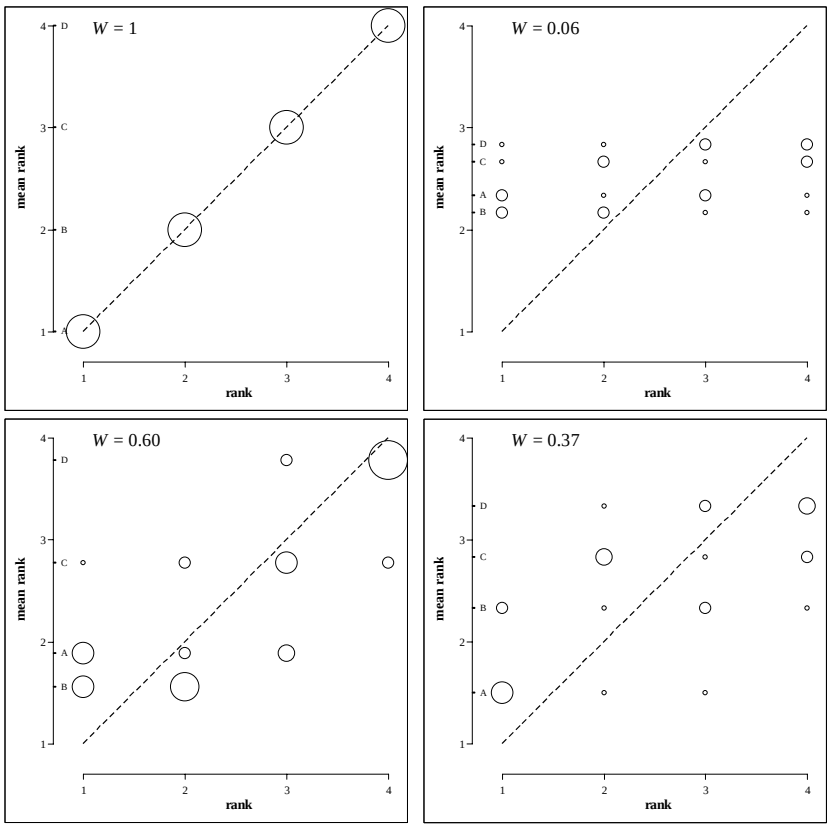

Figure 3: Different degrees of concordance (perfect, minimum possible, fair, low) depicted with concordance bubble-plots (k=*4* for all plots; *m*=9 with *W*=0.60, otherwise *m*=6); artificial data for top row, data from Schucany & Frawley (1973) for bottom row.

4.2 Pin-Cushion plot

Since each of the k objects is assigned $m\,(m–1)\,/\,2$ pairs of ranks, there are $k\,m\,(m–1)\,/\,2$ pairs of same-object ranks in a concordance dataset. Concordance plot can be constructed from the differences between rankings within each of those pairs. We propose representing them by angle, using vertical line as the basis representing no difference (i.e., a pair of identical ranks assigned to an object), and plotting rank difference clockwise (the bigger the difference, the larger the angle). Dependence on the order in which the sets of ranks are placed in the data matrix is avoided by plotting the absolute values of the differences. The maximum possible absolute difference, which is $k-1$, is represented by the

horizontal axis. The angle from the horizontal axis (ϕ) representing a given absolute rank difference is computed as $\phi = 90° \, (1 - |d| / (k - 1))$.

If a difference is not present in the dataset, the corresponding line is not drawn; otherwise, the number of occurrences of a difference is represented by line length. Hence, larger concordance is represented by longer upright pins and shorter inclined pins with respect to smaller concordance. A base shape can provide space for the value of concordance coefficient, or a rectangle is drawn to anchor the perception of line length by means of its height matching the line of length one (i.e., $|d|=1$) and its width matching $|d|=2$ (as in Figure 4).

The resulting graph resembles a pin-cushion, as can be seen in Figure 4, which depicts experimental data from social psychology (Vidmar & Cernigoj 2004). Putting aside the original research design and hypothesis testing for brevity, one can assess the efficacy of concordance visualisation through visual impression by considering that the value of $\mathcal{W}$ for comparing conditions X and Y within group A is 0.75, while $\mathcal{W}$ for comparing groups X and Y within B is 0.56. In general, pin-cushion plot is compact and has a high data-ink ratio, so we consider it particularly suitable for visually comparing concordance between a relatively large number of groups of judges.

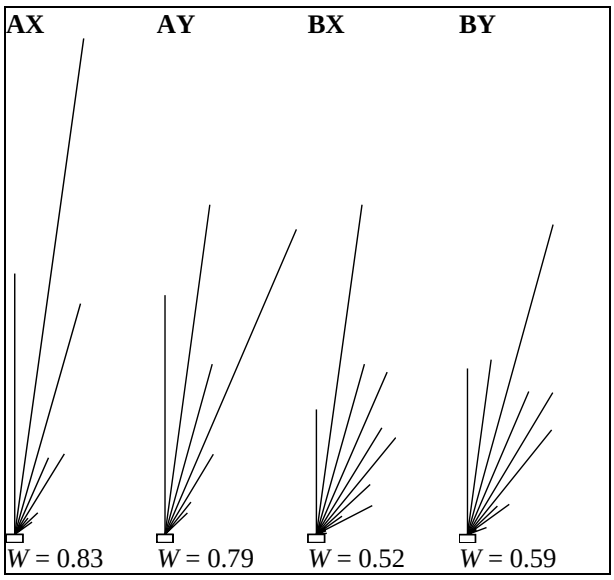

Figure 4: Pin-cushion plots of concordance data (k=12, m=5) from two-factor (groups A&B under conditions X&Y) experiment (Vidmar & Cernigoj 2004).

5 Discussion

Firstly, there are at least two further types of concordance plots that we can think of. They are not as successful as the presented two, but they deserve to be mentioned because of novelty and because they shed additional light on the qualities and downsides of concordance bubble-plot and pin-cushion plot.

- A straight-forward way to construct a concordance graph from the same-object rank pairs is to plot them, preferably in parallel coordinates. The key principle of such a plot would be that for each pair of ranks for a given object, the smaller of the two ranks (labelled R<) would be projected onto the first axis and the larger (R>) onto the second axis, with both axes oriented vertically. The number of identical {R<, R>} pairs could be represented by line thickness. We produced a number of such plots and found that perfect and minimum concordance form easily recognisable patterns, but the in-between levels of concordance are difficult to distinguish. Furthermore, the data-ink ratio is low, and special attention must be paid to axes in order to understand what is being plotted. On the positive side, such „concordance parallel-coordinates plots" have an aesthetic appeal, and they share the special properties of parallel coordinates.

- The shortest summary of the same-object rank pairs is the distribution of absolute differences ($|d|$) of rank pairs for any given object, which is naturally visualised by a bar-plot. To allow comparison between datasets with different k and/or m, it is preferable to plot relative frequencies. In general, the less concordance there is among the judges, the larger the centre of the distribution and the smaller its skewness. The major deficiency of such „concordance bar-plots" is that they – like angular correlation-matrix plots, concordance parallel-coordinates plots and pin-cushion plots – do not provide information on the mean ranks.

For completeness, we must also mention visualisation of frequency distributions of ranked data using permutation polytopes (Thompson 1994, Baggerly 1995). Though that field far exceeds the scope and level of our paper, we would just like to note that despite their theoretical and representational complexity, such visualisations completely represent concordance data only with three or four objects. Nevertheless, we refer the intrigued reader to the references, since it is a fascinating and far-reaching topic.

A possible critique of the presented plots is that area, direction, length and angle are the „elementary perceptual tasks" upon which they are based, and these do not rank high in the hierarchy of accuracy of observer judgments (Cleveland 1994, Cleveland & McGill 1984). However, it is an established fact that perception and therefore also comparison of attributes devoid of context is impossible (Lockhead

1992, 1995), and that understanding good scientific graphics requires time and attention (Tufte 1998). Hence, we designed concordance plots with the aim of presenting a clear and distinguishable pattern for a given degree of concordance, while also providing the necessary elements for subsequent detailed examination. Furthermore, at least in the concordance bubble-plot, the context elements (axes, labels for mean object ranks and diagonal line) actually transform the generally less accurate perceptual tasks into the generally more accurate one, namely judging position along a common scale.

A general concern with both proposed plots is that they become less clear with a large number of objects, but that is not a major limitation because more than a dozen objects or judges are seldom used in actual concordance studies. Considering concordance bubble-plots, had we used jitter instead of circle size to represent frequency, clutter would appear much sooner.

We designed all plots in black and white, since according to good graphical practice guidelines (e.g. Wainer & Thissen 1981), only information dimensions intended to be presented but left unpictured or indiscernible would have required the use of colour. As a final point, even though we have not depicted tied ranks, it is evident that they do not present a problem to the proposed plots.

6 Conclusion

Even though examples of highly contested and seldom used inventions from the history of data visualisation call for caution, we believe that concordance bubble-plot and pin-cushion plot can find application in the various fields where concordance is studied, ranging from ecology to marketing.

Their software implementation is simple – we produced Figure 3 with Microsoft® Excel, and Figure 4 with the freely available graphical package jsplot (http://ourworld.compuserve.com/homepages/jsieberer). The files are available for public download (from http://www.mf.uni-lj.si/ibmi-english/biostat-center, Software section). We are planning to implement concordance visualisation in R.

References

Baggerly, K.A. (1995), *Visual Estimation of Structure in Ranked Data*, PhD thesis, Rice University, Houston.

Beckett, J. & Schucany, W. R. (1975), ‚ANACONDA: Analysis of concordance of g groups of judges', *Proceedings of Social Statistics Section of the American Statistical Association*, 311-313.

Beckett, J. & Schucany, W. R. (1979), ‚Concordance among categorized groups of judges', *Journal of Educational Statistics* **4**(2), 125-137.

Cleveland, W. S. (1994), *The Elements of Graphing Data*, Hobart Press, Summit.

Cleveland, W. S. & McGill, R. (1984), ‚Graphical perception: Theory, experimentation, and application to the development of graphical methods', Journal of the American Statistical Association **79**(387), 531-554.

Corsten, L. C. A., & Gabriel, K. R. (1976), ‚Graphical exploration in comparing variance matrices', *Biometrics* **32**, 851-863.

Ehrenberg, A. S. C. (1952), ‚On sampling from a population of rankers', *Biometrika* **39**, 82-87.

Gescheider, G. A. (1985), *Psychophysics: Method, Theory, and Application*, 2nd ed., Lawrence Erlbaum, Hillsdale.

Holm, S. (1979), ‚A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**, 65-70.

Inselberg, A. (1985), ‚Plane with parallel coordinates', *Visual Computer* **1**, 69-97.

Iman, R. L. & Conover, W. J. (1987), ‚A measure of top-down correlation', *Technometrics* **29**(3), 351-357.

Kendall, M., & Babington Smith, B. (1939), ‚The problem of *m* rankings', *Annals of Mathematical Statistics* **10**, 275-287.

Legendre, P. & Lapointe, F-J. (2004), ‚Assessing congruence among distance matrices: Single-malt Scotch whiskies revised', *Australian & New Zealand Journal of Statistics* **46**(4), 615-629.

Legendre, P. (2005), ‚Species associations: The Kendall coefficient of concordance revisited', *Journal of Agricultural, Biological, and Environmental Statistics* **10**(2), 226-245.

Lockhead, G. R. (1992), ‚Psychophysical scaling: Judgment of attributes or objects?', *Behavioral and Brain Sciences* **15**, 543-558.

Lockhead, G. R. (1995), ‚Psychophysical scaling methods reveal and measure context effects', *Behavioral and Brain Sciences* **18**, 607-612.

Lyerly, S. B. (1952), ‚The average Spearman rank correlation coefficient', *Pychometrika* **17**, 421-428.

Nelson, J. C. & Pepe, M. S. (2000), ‚Statistical description of interrater variability in ordinal ratings', *Statistical Methods in Medical Research* **9**, 475-496.

Page, E. B. (1963), ‚Ordered hypotheses for multiple treatments: A significance test for linear ranks', Journal of the American Statistical Association **58**, 216-230.

Palachek, A. D. & Schucany, W. R. (1984), ‚On approximate confidence intervals for measures of concordance', *Psychometrika* **49**(1), 133-141.

R Development Core Team (2004), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, http://www.r-project.org.

Schucany, W. R. & Frawley, W. H. (1973), ‚A rank test for two group concordance', *Psychometrika* **38**(2), 249-258.

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

Siegel, S. & Castellan, J. (1988), *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed., McGraw-Hill, New York.

Thompson, G. L. (1994), ‚Visualising frequency distributions of ranked data'. *Computational Statistics* **9**(1), 1-10.

Trosset, M. W. (2005), ‚Visualizing correlation', *Journal of Computational and Graphical Statistics* **14**(1), 1-19.

Tufte, E. R. (1998), *The Visual Display of Quantitative Information*, 16th printing, Graphics Press, Chesire.

Vidmar, G. & Cernigoj, M. (2004), ‚Studying norms in small groups by means of multi-group concordance analysis', *Horizons of Psychology* **13**(4), 55-66.

Wainer, H. & Thissen, D. (1981), ‚Graphical data analysis', *Annual Review of Psychology* **32**, 191-241.

Wilkie, D. (1980), ‚Pictorial representation of Kendall's rank correlation coefficient', *Teaching Statistics* **2**, 76-78.