NARDI FUNDULEA, ROMANIA
www.incda-fundulea.ro

ROMANIAN AGRICULTURAL RESEARCH, NO. 28, 2011
Print ISSN 1222-4227; Online ISSN 2067-5720

INSPECTING ASSOCIATIONS IN MULTIVARIATE DATA SETS WITH AN INTERACTIVE MODIFIED BLAND-ALTMAN PLOT

Marcin Kozak and Agnieszka Wnuk

Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

Corresponding author: Marcin Kozak, e-mail: nyggus@gmail.com

ABSTRACT

This paper offers a new method of inspecting associations in multivariate data sets, which is important when one wants to present a correlation table, frequent situation in agricultural sciences. We propose to use a simple visualization technique called the Bland-Altman plot for that, with a minor modification. The plot compares Pearson's and Spearman's correlation coefficients, on which basis quite important information can be extracted concerning linearity among the variables. We offer a simple R function for such a plot along with its interactive version. Thanks to such interactive visualization one can very quickly and easily study linearity of associations even in very large data sets, which would be very troublesome for example with the help of a scatterplot matrix.

Key words: interactive graphing, Pearson correlation, Spearman correlation, visualization.

INTRODUCTION

Multivariate data are very common in agricultural sciences; equally common are applications of simple Pearson's correlations, often reported by means of correlation tables. Practice shows that all too often this is an oversimplification of reality because it is rather an uncommon situation that all variables in a multivariate data set are linearly associated. It is thus wise *always* to check whether the assumption of linearity is valid.

One such check might be the comparison of Pearson's and Spearman's correlation coefficients for each pair of variables. Indeed, Spearman's correlations are sometimes used to measure monotonic rather than linear relationships (e.g., Roel and Plant, 2004; Sahramaa et al., 2004; Yamada et al., 2004; Zhang and Mergoum, 2007; Fuentes et al., 2005; Cox and Gerard, 2007; Hu et al., 2008; Nicodemo et al., 2009); they are robust to outliers and cover all monotonic associations, which include also linear ones. If the relationship is linear, then Spearman's and Pearson's correlations will be roughly the same. If the absolute value of Spearman's correlation is noticeably higher than Pearson's, the relationship is monotonic but not linear. If the opposite is the case (the absolute value of Spearman's correlation is noticeably smaller than Pearson's), the relationship may be neither monotonic nor linear, yet Pearson's coefficient can be strongly influenced by some outlier value(s). Worth noting is that even when both types of correlations are similarly weak, this does not mean that there is no relationship between the two variables but this does indicate that there is no monotonic relationship.

Such a comparison of correlation coefficients will not be easy for many variables, because for a p-variate data set there are $p(p-1)/2$ pairs of variables, reaching for example 45 such pairs for $p = 10$ and 190 pairs for $p = 20$. The aim of this paper is to propose a simple interactive visualization method of inspecting associations among many variables. The R (R Development Core Team, 2010) code is also presented.

MATERIAL AND METHODS

Interpretation of scatterplots among many variables can be done by means of scagnostics measures (Wilkinson, 2005), but these are not easy to understand and interpret for most

Received 9 September 2010; accepted 25 January 2011

users. In this paper we propose to use the Bland-Altman plot (Altman and Bland, 1983; Bland and Altman, 1986), also called the Tukey mean-difference plot (Cleveland, 1994), for Pearson's and Spearman's correlation coefficients for all pairs of variables. Hereafter we will suggest a small modification to facilitate reading the plot.

The Bland-Altman plot in the present context would be obtained by graphing a scatterplot with the coordinates of

$$x_{\text{tmd},ij} = \frac{r_{Pij} + r_{Sij}}{2} \text{ and } y_{\text{tmd},ij} = r_{P_{ij}} - r_{S_{ij}} \quad (1)$$

where $x_{\text{tmd},ij}$ and $y_{\text{tmd},ij}$ are the corresponding coordinates for the Bland-Altman plot, and $r_{P_{ij}}$ and $r_{S_{ij}}$ are Pearson's and Spearman's correlations between the *i*th and *j*th variables.

However, we think that it will be better to modify such a plot, with Pearson's correlation forming the x-axis, so with coordinates of

$$x_{\text{tmd},ij} = r_{Pij} \text{ and } y_{\text{tmd},ij} = r_{P_{ij}} - r_{S_{ij}} \quad (2)$$

In that way one immediately receives information about Pearson's correlation and its difference from Spearman's correlation.

One might want to quickly learn which associations the particular points in the plot represent. This can be very efficiently done through interactive visualization, in which pressing a mouse button causes the required information about the correlation (variable names and Pearson's and Spearman's correlations) corresponding to the closest point to the pointer be displayed; this point could also be focused by re-plotting in a different colour for a short while.

Below we present the code for R (R Development Core Team 2010) functions `spot.nonlinear()`, which constructs a modified Bland-Altman plot as discussed before, and `spot.nonlinear.identify()`, which produces interactive version of this plot with the help of the `tcltk` package of R. The `identifyA()` function is a modification of the `identify()` function that utilizes `tcltk` boxes.

Note that some control of the appearance of the plot (such as plotting symbol, its size and color) can be made through the "..." argument, which is then passed to the `plot()` function. It can include for example arguments `pch` and `col` to control the plotting symbol and its color; `cex.lab` and `cex.axis` to control the font size for the axis labels and tick mark labels, respectively; `cex` to decrease the size of plotting symbols; and so on.

However, for interactive plotting it will be best not to change the colour of plotting symbols because after identification of a point, it is plotted in red for three seconds, after which it is plotted back in black. Note that in case of overlap, open circles – used here as default – are the most efficient plotting symbols (Cleveland, 1994).

The plot in figure 1 was obtained by pasting the function `spot.nonlinear()` (see below) into the R console and running the following command:

```
> data(soil, package = "agricolae")
> spot.nonlinear(soil[, 2:23])
```

To run the interactive version of the plot, simply paste the three functions below into the R console and type

```
> spot.nonlinear.identify(soil[, 2:23])
```

Now it suffices to press the left mouse button after placing the pointer near a point of interest to obtain the desired information in the box showed in figure 3. Pressing the `End` button closes this box, while the Escape key ends interactive identification of points.

After that, the information (names of variables and Pearson's and Spearman's correlations) about which points were identified is printed within the console. Note that the function automatically abbreviates names of variables to at least as many characters as is needed to differentiate the names, but now fewer than two.

The invisible function within the `spot.nonlinear()` function is used to return (only after being assigned to an object) the information about data points presented in the Bland-Altman plot. It is then used in the `identifyA()` function. Users will seldom utilize this feature. The code was tested in R 2.10.1 under Windows XP and Windows 7.

```
spot.nonlinear <- function(x, ...) {
 aa <- as.dist(cor(x))
    cor.P <- as.vector(aa)
    cor.S <- as.vector(as.dist(cor(x,method = "spearman")))
    labs <- vector()
     names(x) <- abbreviate(names(x), 2)
     variable.names = names(x)
 for (i in 1:ncol(x)) for (j in 1:ncol(x)) if (j > i)
   labs <- c(labs, paste (colnames(x)[i], colnames(x)[j],
              sep = " and "))
 plot(x = cor.P, y = cor.P - cor.S,
     ylab = expression(paste("Pearson's ",
          italic(r) - "Spearman's ", italic(r))),
     xlab = "Pearson's correlation",
 xlim = c(-1.05, 1.05), las = 1, ...)
   axis(3, lab = F)
   axis(4, lab = F)
   abline(h = seq(-1, 1, 0.2), col = "grey", lty = "dashed")
   abline(h = 0)
   invisible(list(cor.P = cor.P, diff = cor.P - cor.S, labs = labs,
      variable.names = variable.names))
}
identifyA <- function(x, y = NULL, labs = 1:length(x),
   n = length(x)) {
   library(tcltk)
   xy <- xy.coords(x, y)
   x <- xy$x
   y <- xy$y
   n.checked <- 0
   while (n.checked < n) {
      ans <- identify(x, y, n = 1, plot = FALSE)
      if (!length(ans))
         break
      points(x[ans], y[ans], col = "red")
          n.checked <- n.checked + 1
      tt <- tktoplevel()
      tkwm.title(tt, labs[ans])
      tkgrid(tklabel(tt, text = paste("Correlation between",
           labs[ans], "\n")))
      tkgrid(tklabel(tt, text = paste("  Pearson's r = ",
           format(round(x[ans], 2), nsmall = 2), sep = "")))
      tkgrid(tklabel(tt, text = paste("Spearman's r = ",
           format(round(-y[ans] + x[ans], 2), nsmall = 2),
           "\n", sep = "")))
     exit.but <- tkbutton(tt, text = "End",
                          command = function() tkdestroy(tt))
         tkgrid(exit.but)
     cat(paste(labs[ans],    ";    Pearson's    r    =    ",
      format(round(x[ans],
          2), nsmall = 2), "; Spearman's r = ", format(round(-
               y[ans] + x[ans], 2), nsmall = 2), "\n", sep = ""))
     Sys.sleep(3)
         points(x[ans], y[ans], col = "black")
   }
}
spot.nonlinear.identify <- function(x, ...) {
   aa <- spot.nonlinear(x, ...)
   identifyA(aa$cor.P, aa$diff, labs = aa$labs)
   dev.off()
}
```

RESULTS

We will present the method for soil analysis data, originating from International Potato Center – Lima, Peru, which are available through the `soil` data set of the `agricolae` package (De Mendiburu, 2010) of R (R Development Core Team, 2010). Besides the variable indicating the locations of 13 soil samples, the data set contains 22 variables of soil analysis. So there are 231 pairs of variables and the same number of associations to study. Note that with just 13 observations, the associations can be strongly affected by outlier values.

The scatterplot matrix would include 462 panels, so we decided not to present it in this paper. It can be constructed in R with the following command:

```
> data(soil, package = "agricolae")
> pairs(soil[, 2:23], gap = 0)
```

From the scatterplot one can notice that there are some outlier values for some of the traits, but trying to dig deep into the associations is very difficult.

Figure 1 represents a modified Bland-Altman plot as discussed above. We can immediately see that there are many associations that could not be considered linear. After running the interactive version of this plot (see the code in Material and Methods section), we have detected many of them, of which four are presented in figure 2.

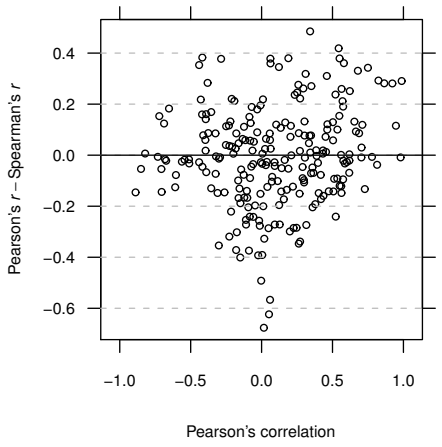

Figure 1. A modified Bland-Altman plot comparing Pearson's and Spearman's sample correlation coefficients for the soil data set. One can quickly notice that for many associations the correlations differ

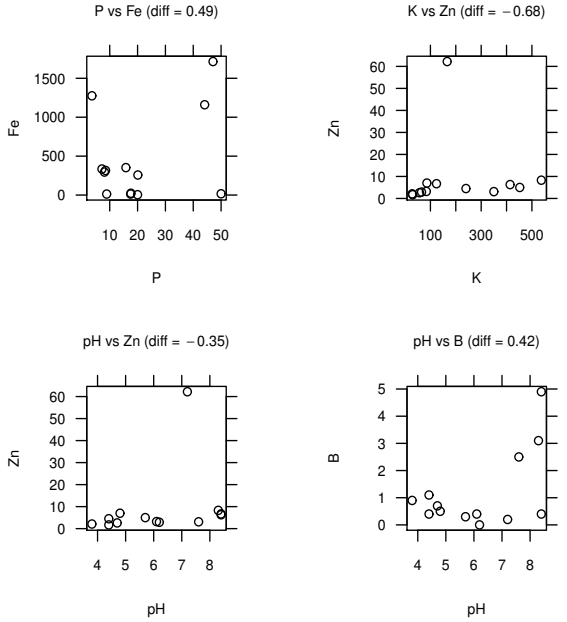

Figure 2. Four examples of scatterplots indicated as untypical by Figure 1. "diff" stands for the difference between Pearson's and Spearman's correlation

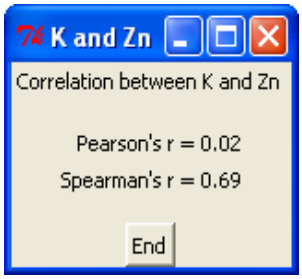

Figure 3. Appearance of the tcltk box after a plotting symbol representing the association between K and Zn has been identified.

CONCLUSIONS

The technique proposed in this paper to a much extent facilitates detecting nonlinear associations among many traits. What we have done in the example in several minutes would require a long time and a lot of effort with a scatterplot matrix or any other technique, so we believe that the interactive modified Bland-Altman plot may found its niche in agricultural research.

REFERENCES

Altman, D.G., Bland, J.M., 1983. *Measurement in medicine: The analysis of method comparison studies*. Statistician, 32: 307-317.

Bland, J.M., Altman, D.G., 1986. *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1 (8476): 307-310.

Cleveland, W.S., 1994. *The elements of graphing data*. 2nd Ed. Hobart Press, Summit, New Jersey, USA, 1994.

Cox, M.S., Gerard, P.D., 2007. *Soil management zone determination by yield stability analysis and classification*. Agronomy J., 99: 1357-1365.

Fuentes, R.G., Mickelson, H.R., Busch, R.H., Dill-Macky, R., Evans, C.K., Thompson, W.G., Wiersma, J.V., Xie, W., Dong, Y., Anderson, J.A., 2005. *Resource allocation and cultivar stability in breeding for fusarium head blight resistance in spring wheat*. Crop Sci., 45: 1965-1972.

Hu, W., Shao, M.A., Wang, Q.J., Reichardt, K., 2008. *Soil water content temporal-spatial variability of the surface layer of a loess plateau hillside in China.* Scientia Agricola, 65: 277-289.

De Mendiburu, F., 2010. Agricolae: *Statistical Procedures For Agricultural Research*. R Package Version 1.0-9. http://Cran.R-Project.Org/Package=Agricolae

Nicodemo, D., Couto, R.H.N., Malheiros, E.B., De Jong, D., 2009. *Honey Bee As An Effective Pollinating Agent of Pumpkin*. Sci. Agric. (Piracicaba, Braz.), 66: 476-480.

R Development Core Team, 2010. *R : A language and environment for statistical computing*. R Foundation For Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-Project.Org

Roel, A., Plant, R.E., 2004. *Factors underlying yield variability in two California rice fields*. Agronomy J., 96: 1481-1494.

Sahramaa, M., Hömmö, L., Jauhiainen, L., 2004. *Variation in seed production traits of reed canarygrass germplasm.* Crop Sci., 44: 988-996.

Wilkinson, L., 2005. *The grammar of graphics*. 2nd Ed. Springer-Verlag, New York.

Yamada, T., Jones, E.S., Cogan, N.O.I., Vecchies, A.C., Nomura, T., Hisano, H., Shimamoto, Y., Smith, K.F., Hayward, M.D., Forster, J.W., 2004. *QTL analysis of morphological, developmental, and winter hardiness-associated traits in perennial ryegrass*. Crop Sci., 44: 925-935.

Zhang, G., Mergoum, M., 2007. *Developing evaluation methods for kernel shattering in spring wheat*. Crop Sci., 47: 1841-1850.