Where's Waldo: Visualizing Collinearity Diagnostics

Michael Friendly* Ernest Kwan

Abstract

Collinearity diagnostics are widely used, but the typical tabular output used in almost all software makes it hard to tell what to look for and how to understand the results. We describe a simple improvement to the standard tabular display, a graphic rendition of the salient information as a "tableplot," and graphic displays designed to make the information in these diagnostic methods more readily understandable.

In addition, we propose a visualization of the contributions of the predictors to collinearity through a *collinearity biplot*, which is simultaneously a biplot of the smallest dimensions of the correlation matrix of the predictors, $\boldsymbol{R}_{XX}$, and the largest dimensions of $\boldsymbol{R}_{XX}^{-1}$, on which the standard collinearity diagnostics are based.

Key words: condition indices, collinearity biplot, diagnostic plots, effect ordering, multiple regression, tableplot, variance inflation

1 Introduction

> Q: (Collinearity diagnostics and remedies): "Some of my collinearity diagnostics have large values, or small values, or whatever they're not supposed to have. Is this bad? If so, what can we do about it?"
>
> from: `http://www.sociology.ohio-state.edu/people/ptv/faq/collinearity.htm`

Problems in estimation in multiple regression models that arise from influential observations and high correlations among the predictors were first described in a comprehensive way in Belsley, Kuh, and Welsch's (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. This book would prove to be a highly influential point on the landscape of diagnostic methods for regression, but not always one of high leverage, at least in graphical methods for visualizing and understanding collinearity diagnostics.

Later, David Belsley wrote *A guide to using the collinearity diagnostics* (Belsley, 1991b), that seemed to promise a solution for visualizing these diagnostics. For context, it is worth quoting the abstract in total:

*Michael Friendly is Professor, Psychology Department, York University, Toronto, ON, M3J 1P3 Canada, E-mail: friendly@yorku.ca. Ernest Kwan is Assistant Professor, Sprott School of Business, Carleton University, E-mail: ernest_kwan@carleton.ca. This work is supported by Grant 8150 from the National Sciences and Engineering Research Council of Canada. We are grateful to John Fox for comments on an initial draft and to the referees for helping us to clarify the presentation.

> The description of the collinearity diagnostics as presented in Belsley, Kuh, and Welsch's, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, is principally formal, leaving it to the user to implement the diagnostics and learn to digest and interpret the diagnostic results. This paper is designed to overcome this shortcoming by describing the different graphical displays that can be used to present the diagnostic information and, more importantly, by providing the detailed guidance needed to promote the beginning user into an experienced diagnostician and to aid those who wish to incorporate or automate the collinearity diagnostics into a guided-computer environment.

Alas, the "graphical displays" suggested were just tables– of eigenvalues, condition numbers and coefficient variance proportions associated with the collinearity diagnostics. There is no doubt that Belsley's suggested tabular displays have contributed to the widespread implementation and use of these diagnostics. Yes, as the initial quote for this introduction indicates, users are often uncertain about how to interpret these tabular displays.

To make the point of this paper more graphic, we liken the analyst's task in understanding collinearity diagnostics to that of the reader of Martin Hansford's successful series of books, *Where's Waldo* (titled *Where's Wally* in the UK, *Wo ist Walter* in Germany, etc.). These consist of a series of full-page illustrations of hundreds of people and things and a few Waldos— a character wearing a red and white striped shirt and hat, glasses, and carrying a walking stick or other paraphernalia. Waldo was never disguised, yet the complex arrangement of misleading visual cues in the pictures made him very hard to find. Collinearity diagnostics often provide a similar puzzle.

The plan of this paper is as follows: We first describe a simple example that illustrates the current state of the art for the presentation of collinearity diagnostics. Section 2 summarizes the standard collinearity diagnostics in relation to the classical linear regression model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. In Section 3 we suggest some simple improvements to the typical tabular displays and a graphic rendering called a "tableplot" to make these diagnostics easier to read and understand. Section 4 describes some graphic displays based on the the biplot (Gabriel, 1971) that helps interpret the information in collinearity diagnostics.

It is also important to say here what this paper does *not* address: collinearity in regression models is a large topic, and the answer to the last question in the initial quote, *If so, what can we do about it?* would take this article too far afield. The methods described below do provide a means to see what is potentially harmful, whether large or small and thus, answers the question *Is this bad?* More importantly, the graphic displays related to these questions can often say *why* something about the data is good or bad.

1.1 Example: Cars data

As a leading example, we use the Cars dataset from the 1983 ASA Data Exposition (`http://stat-computing.org/dataexpo/1983.html`) prepared by Ernesto Ramos and David Donoho. This dataset contains 406 observations on the following seven variables: **MPG** (miles per gallon), **Cylinder** (# cylinders), **Engine** (engine displacement, cu. inches), **Horse** (horsepower), **Weight** (vehicle weight, lbs.), **Accel** (time to accelerate from 0 to 60 mph, in sec.), and **Year** (model year, modulo 100). An additional categorical variable identifies the region of origin (American, European, Japanese), not analyzed here. For this data, the natural questions concern how well MPG can be explained by the other variables.

Collinearity diagnostics are not part of the standard output of any widely-used statistical software; they must be explicitly requested by using options (SAS), menu choices (SPSS) or other packages (R: `car`, `perturb`).

As explained in the next section, the principal collinearity diagnostics include: (a) variance inflation factors, (b) condition indices, and (c) coefficient variance proportions. To obtain this output in SAS, one can use the following syntax, using the options VIF and COLLINOINT.

```
proc reg data = cars;
  model mpg = weight year engine horse accel cylinder / vif collinoint;
  run;
```

Another option, COLLIN, produces collinearity diagnostics that include the intercept. But these are useless *unless* the intercept has a real interpretation and the origin on the regressors is contained within the predictor space, as explained in Section 2.2. See Fox (1997, p. 351) and the commentary surrounding Belsley (1984) for discussion of this issue. We generally prefer the intercept-adjusted diagnostics, but the choice is not material to the methods presented here.

The model specified above fits very well, with an $R^2 = 0.81$; however, the t-tests for parameters shown in Table 1 indicates that only two predictors— Weight and Year are significant. Table 1 also shows the variance inflation factors. By the rules-of-thumb described below, four predictors— Weight, Engine, Horse and Cylinder have potentially serious problems of collinearity, or at least cause for concern. The condition indices and coefficient variance proportions are given in Table 2. As we describe below, we might consider the last two rows to show evidence of collinearity. However, the information presented here hardly gives rise to a clear understanding.

Table 1: Cars data: parameter estimates and variance inflation factors

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr $> \|t\|$	Variance Inflation
Intercept	1	−14.63175	4.88451	−3.00	0.0029	0
Weight	1	−0.00678	0.00067704	−10.02	<.0001	10.857
Year	1	0.76205	0.05292	14.40	<.0001	1.253
Engine	1	0.00848	0.00747	1.13	0.2572	20.234
Horse	1	−0.00290	0.01411	−0.21	0.8375	9.662
Accel	1	0.06121	0.10366	0.59	0.5552	2.709
Cylinder	1	−0.34602	0.33313	−1.04	0.2996	10.658

Table 2: Cars data: Condition indices and variance proportions, in the form displayed by most statistical software

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Weight	Year	Engine	Horse	Accel	Cylinder
1	4.25623	1.00000	0.00431	0.00968	0.00256	0.00523	0.00922	0.00457
2	0.83541	2.25716	0.00538	0.85620	0.00114	0.00003952	0.00396	0.00296
3	0.68081	2.50034	0.01278	0.05358	0.00177	0.00244	0.42400	0.00515
4	0.13222	5.67358	0.08820	0.00581	0.01150	0.29168	0.06140	0.31720
5	0.05987	8.43157	0.71113	0.06882	0.00006088	0.66021	0.49182	0.11100
6	0.03545	10.95701	0.17819	0.00592	0.98297	0.04040	0.00961	0.55912

2 Background: Notation and collinearity diagnostics

Consider the classical linear regression model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ vector of responses; $\boldsymbol{X}$ is an $n \times p$ full-rank matrix of predictors, the first column of which consists of 1s; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated, where, by convention, the first element (intercept) is denoted β_0; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors, with $\mathcal{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathcal{V}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$. The usual least squares estimates of the parameters $\boldsymbol{\beta}$ are given by $\boldsymbol{b} \equiv \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and $\mathcal{V}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, whence the standard deviations of the parameters, which inversely reflect the precision of estimation, are given by $[\operatorname{diag}\mathcal{V}(\boldsymbol{b})]^{1/2}$.

2.1 Variance inflation

It can be shown (e.g., Fox (1984)) that the sampling variances of the non-intercept parameter estimates can be expressed as

$$\mathcal{V}(b_j) = \frac{\sigma^2}{(n-1)s_j^2} \times \left(\frac{1}{1 - R^2_{j\,|\,\text{others}}}\right) \quad , \tag{1}$$

where s_j^2 is the sample variance of the j-th column, $\boldsymbol{X}_j$, and $R^2_{j\,|\,\text{others}}$ is the squared multiple correlation from the regression of $\boldsymbol{X}_j$ on the other predictors. It is easily seen that the second term in Eqn. (1) is a factor that multiplies the parameter variances as a consequence of correlations among the predictors. This term, called the *variance-inflation factor* (VIF) by Marquandt (1970) has become a standard collinearity diagnostic. When the predictors are all uncorrelated, all $R_j^2 = 0$ and all VIF_j have their minimum value of 1. As any R_j^2 approaches 1 (complete linear dependence on the other predictors), VIF_j approaches ∞.

In the linear regression model with standardized predictors, the covariance matrix of the estimated intercept-excluding parameter vector $\boldsymbol{b}^\star$ has the simpler form,

$$\mathcal{V}(\boldsymbol{b}^\star) = \frac{\sigma^2}{n-1}\boldsymbol{R}_{XX}^{-1} \quad , \tag{2}$$

where $\boldsymbol{R}_{XX}$ is the correlation matrix among the predictors. It can then be seen that the VIF_j are just the diagonal entries of $\boldsymbol{R}_{XX}^{-1}$.

2.2 Condition indices and variance proportions

Large VIF_j indicate predictor coefficients whose precise estimation is degraded due to large $R^2_{j\,|\,\text{others}}$. To go further, we need to determine (a) how many dimensions in the space of the predictors are associated with nearly collinear relations; (b) which predictors are most strongly implicated in each of these.

In the predictor space, the linear relations among the variables can be seen most easily in terms of the principal component analysis of the standardized predictors, or, equivalently, in terms of the eigen decomposition of $\boldsymbol{R}_{XX}$ as $\boldsymbol{R}_{XX} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, where $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ are the ordered eigenvalues of $\boldsymbol{R}_{XX}$ and $\boldsymbol{V}$ is the $p \times p$ matrix whose columns are the corresponding eigenvectors. By elementary matrix algebra, the eigen decomposition of $\boldsymbol{R}_{XX}^{-1}$ is then

$$\boldsymbol{R}_{XX}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}' \quad . \tag{3}$$

Thus, $\boldsymbol{R}_{XX}$ and $\boldsymbol{R}_{XX}^{-1}$ have the same eigenvectors, and the eigenvalues of $\boldsymbol{R}_{XX}^{-1}$ are just λ_i^{-1}. Using Eqn. (3), the variance inflation factors may be expressed as

$$VIF_j = \sum_{k=1}^{p} \frac{V_{jk}^2}{\lambda_k} \ , \tag{4}$$

which shows that only the small eigenvalues contribute to variance inflation, but only for those predictors that have large eigenvector coefficients on those small components. These facts lead to the following diagnostic statistics for collinearity:

Condition indices : The smallest of the eigenvalues, those $\lambda_k \approx 0$, indicate collinearity and the number of small values indicates the number of near collinear relations. Because the sum of the eigenvalues, $\Sigma\lambda_i = p$ which increases with the number of predictors, it is useful to scale them all in relation to the largest. This leads to *condition indices*, defined as $\kappa_k = \sqrt{\lambda_1/\lambda_k}$. These have the property that the resulting numbers have common interpretations regardless of the number of predictors. By common conventions (Belsley, 1991a) condition indices from 10–30 are considered values to be wary of; > 30 indicates trouble; and > 100 is a sign of potential disaster in estimation.

Coefficient variance proportions : Large VIFs indicate variables that are involved in *some* nearly collinear relations, but they don't indicate *which* other variable(s) each is involved with. For this purpose, Belsley *et al.* (1980) and Belsley (1991a) proposed calculation of the proportions of variance of each variable associated with each principal component as a decomposition of the coefficient variance for each dimension. These may be expressed (Fox, 1984, §3.1.3) as

$$p_{jk} = \frac{V_{jk}^2}{VIF_j \ \lambda_k} \tag{5}$$

Note that for orthogonal regressors, all condition indices $\kappa_k = 1$ and the matrix of coefficient variance proportions, $\boldsymbol{P} = (p_{jk}) = \boldsymbol{I}$. Thus, Belsley (1991a,b) recommended that the sources of collinearity be diagnosed (a) only for those components with large κ_k, *and* (b) for those components for which p_{jk} is large (say, $p_{jk} \geq 0.5$) on two or more variables.

3 Improved tabular displays: How many Waldos?

The standard tabular display of condition indices and variance proportions in Table 2 suffers mainly from the fact that the most important information is disguised by being embedded in a sea of mostly irrelevant numbers, arranged by inadvertent design to make the user's task harder. One could easily nominate this design as a *Where's Waldo* for tabular display.

The first problem is that the table is sorted by decreasing eigenvalues. This would otherwise be typical and appropriate for eigenvalue displays, but for collinearity diagnostics, it hides Waldo among the bottom rows. It is more useful to sort by decreasing condition numbers.

A second problem is that the variance proportions corresponding to *small* condition numbers are totally irrelevant to problems of collinearity. Including them, perhaps out of a desire for completeness, would otherwise be useful, but here it helps Waldo avoid detection. For the purpose for which these tables are designed, it is better simply to suppress the rows corresponding to low condition indices. In

this article we adopt the practice of showing one more row than the number of condition indices greater than 10 in tables, but there are even better practices.

Software designers who are troubled by incompleteness can always arrange to make the numbers we would suppress perceptually less important, e.g., smaller in size or closer to background in texture or shading, as in Table 3. A simple version of this is to use a 'fuzz' value, e.g., `fuzz=0.5` such that all variance proportions $<$ `fuzz` are blanked out, or replaced by a place holder (`'.'`). For example, the `colldiag` function in the R package `perturb` (Hendrickx, 2008) has a print method that supports a `fuzz` argument.

The third problem is that, even for those rows (large condition numbers) we want to see, the typical displayed output offers Waldo further places to hide among numbers printed to too many decimals. Do we really care that the variance proportion for Engine on component 5 is 0.00006088? For the large condition numbers, we should only be concerned with the variables that have large variable proportions associated with those linear combinations of the predictors.

For the Cars data, we need only look at the rows corresponding to the highest condition indices. Any sharp cutoff for the coefficient variance proportions to display (e.g., $p_{jk} \geq 0.5$) runs into difficulties with borderline values, so in Table 3, we highlight the information that should capture attention by (a) removing all decimals and (b) distinguishing values ≥ 0.5 by font size and style; other visual attributes could also be used. From this, it is easily seen that there are two Waldos in this picture. The largest condition index corresponds to a near linear dependency involving Engine size and number of cylinders; the second largest involves weight and horsepower, with a smaller contribution of acceleration. Both of these have clear interpretations in terms of these variables on automobiles: engine size is directly related to number of cylinders and, orthogonal to that dimension, there is a relation connecting weight to horsepower.

Table 3: Cars data: Revised condition indices and variance proportions display. Variance proportions > 0.5 are highlighted.

Condition			Proportion of Variation ($\times 100$)					
Number	Index	Eigenvalue	Weight	Year	Engine	Horse	Accel	Cylinder
6	10.96	0.03545	18	1	**98**	4	1	**56**
5	8.43	0.05987	**71**	7	0	**66**	49	11

3.1 Tableplots

Table 2 and even the collinearity-targeted version, Table 3, illustrate how difficult it is to display quantitative information in tables so that what is *important* to see— patterns, trends, differences, or special circumstances— are made *directly* visually apparent to the viewer. Tukey (1990) referred to this property as **interoccularity**: the message hits you between the eyes.

A *tableplot* is a semi-graphic display designed to overcome these difficulties, developed by Ernest Kwan (2008a). The essential idea is to display the numeric information in a table supplemented by symbols whose size is proportional to the cell value, and whose other visual attributes (shape, color fill, background fill, etc.) can be used to encode additional information essential to direct visual understanding. The method is far more general than described here and illustrated in the context of collinearity diagnostics. See Kwan (2008b) for some illustrations of the use for factor analysis models.

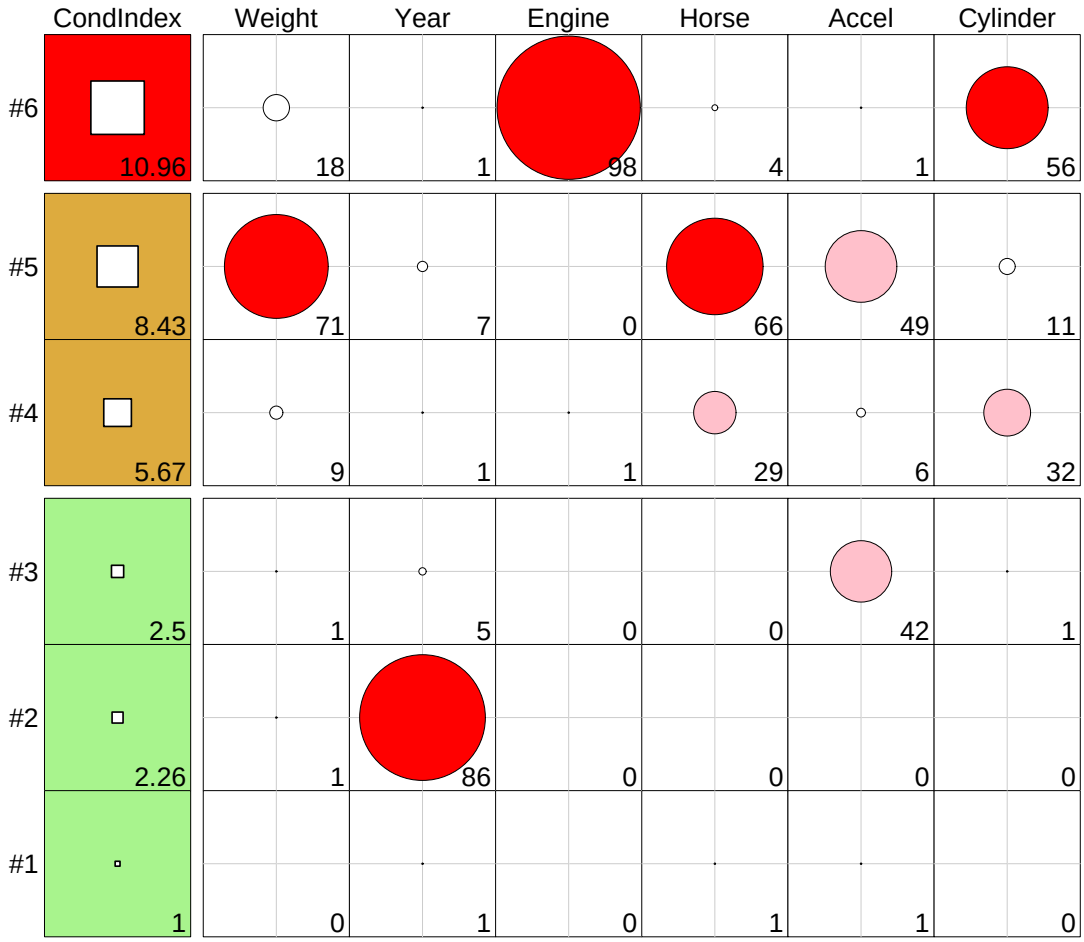

Figure 1: Tableplot of condition indices and variance proportions for the Cars data. In column 1, the square symbols are scaled relative to a maximum condition index of 30. In the remaining columns, variance proportions ($\times 100$) are shown as circles scaled relative to a maximum of 100.

Figure 1 is an example of one tableplot display design for collinearity diagnostics. All the essential numbers in Table 2 are shown, but encoded in such a way as to make their interpretation more direct. The condition indices are shown by the area of the white squares in column 1. In this example these are scaled so that a condition index of 30 corresponds to a square that fills the cell. The background color of the cells in this column indicate a reading of the severity of the condition index, κ_k. In this example, green ("OK") indicates $\kappa_k < 5$, yellow ("warning") indicates $5 \leq \kappa_k < 10$ and red ("danger") indicates $\kappa_k \geq 10$. The remaining columns show the collinearity variance proportions, $100 \times p_{jk}$, with circles scaled relative to a maximum of 100, and with color fill (white, pink, red) distinguishing the ranges $\{$ 0–20, 20–50, 50–100$\}$.

The details of the particular assignments of colors and ranges to the condition indices and variance proportions are surely debatable, but the main point here is that Figure 1 succeeds where Table 2 does not and Table 3 only gains by hiding irrelevant information.

4 Graphical displays: biplots

As we have seen, the collinearity diagnostics are all functions of the eigenvalues and eigenvectors of the correlation matrix of the predictors in the regression model, or alternatively, the SVD of the $\boldsymbol{X}$ matrix in the linear model (excluding the constant). The standard biplot (Gabriel, 1971, Gower and Hand, 1996) can be regarded as a multivariate analog of a scatterplot, obtained by projecting a multivariate sample into a low-dimensional space (typically of 2 or 3 dimensions) accounting for the greatest variance in the data. With the symmetric (PCA) scaling used here, this is equivalent to a plot of principal component scores of the mean-centered matrix $\tilde{\boldsymbol{X}}$ of predictors for the observations (shown as points or case labels), together with principal component coefficients for the variables (shown as vectors) in the same 2D (or 3D) space.

The standard biplot, of the first two dimensions, corresponding to the largest eigenvalues of $\boldsymbol{R}_{XX}$ is shown in Figure 2. This is useful for visualizing the principal variation of the observations in the space of the predictors. It can be seen in Figure 2 that the main dimension of variation among the automobile models is in terms of engine size and power; acceleration time is also strongly related to Dimension 1, though of course inversely. The second dimension (accounting for an additional 13.8% of predictor variance) is largely determined by model year.

In these plots: (a) The variable vectors have their origin at the mean on each variable, and point in the direction of positive deviations from the mean on each variable. (b) The angles between variable vectors portray the correlations between them, in the sense that the cosine of the angle between any two variable vectors approximates the correlation between those variables (in the reduced space). (c) Similarly, the angles between each variable vector and the biplot axes approximate the correlation between them. (d) Because the predictors were scaled to unit length, the relative length of each variable vector indicates the proportion of variance for that variable represented in the low-rank approximation. (e) The orthogonal projections of the observation points on the variable vectors show approximately the value of each observation on each variable. (f) By construction, the observations, shown as principal component scores are uncorrelated;

But the standard biplot is less useful for visualizing the relations among the predictors that lead to nearly collinear relations. Instead, biplots of the smallest dimensions show these relations directly, and can show other features of the data as well, such as outliers and leverage points.

4.1 Visualizing variance proportions

As with the tabular display of variance proportions, Waldo is hiding in the dimensions associated with the smallest eigenvalues (largest condition indices). As well, it turns out that outliers in the predictor space— high leverage observations— can often be seen as observations far from the centroid in the space of the smallest principal components.

Figure 3 shows the biplot of the Cars data for the smallest two dimensions— what we can call the *collinearity biplot*. The projections of the variable vectors on the Dimension 5 and Dimension 6 axes are proportional to their variance proportions in Table 2. The relative lengths of these variable vectors can be considered to indicate the extent to which each variable contributes to collinearity for these two near-singular dimensions.

Thus, we see that (as in Table 3) Dimension 6 is largely determined by Engine size, with a substantial relation to Cylinder. Dimension 5 has its' strongest relations to Weight and Horse. In the reduced tabular display, Table 3, we low-lighted all variance proportions < 0.5, but this is unnecessary in the graphical representation.

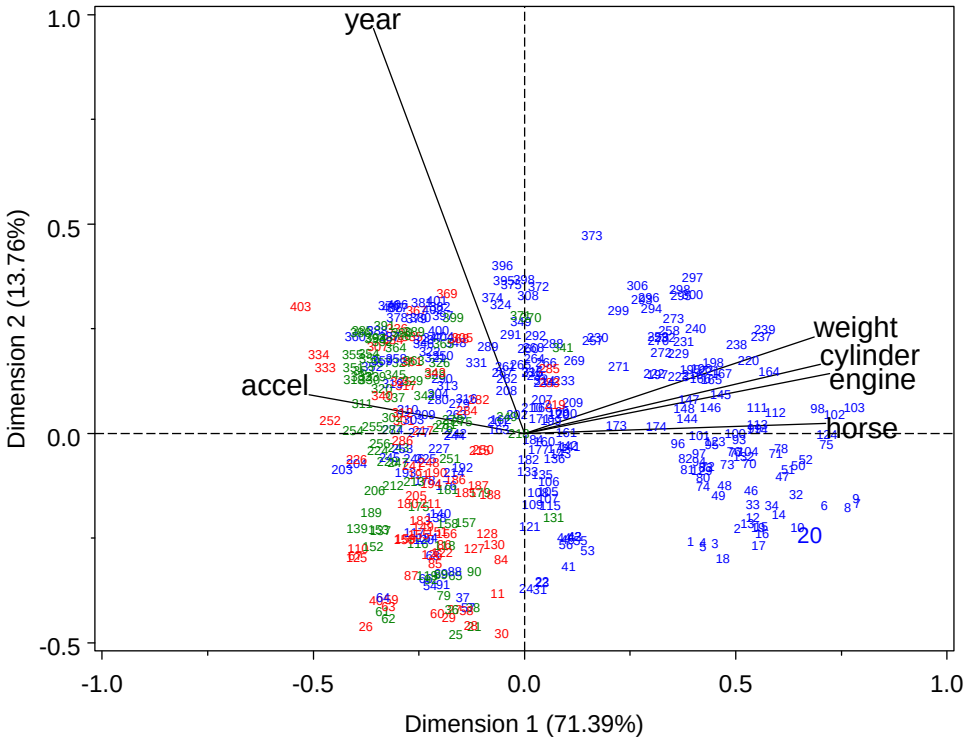

Figure 2: Standard biplot of the Cars data, showing the first two dimensions. The observations are labeled with their case numbers and colored according to region of origin. The right half of the plot consists almost entirely of US cars (blue).

Moreover, there is one observation, #20, that stands out as an outlier in predictor space, far from the centroid. It turns out that this vehicle, a Buick Estate wagon, is an early-year (1970) American behemoth, with an 8-cylinder, 455 cu. in, 225 horse-power engine, and able to go from 0 to 60 mph in 10 sec. (Its MPG is only slightly under-predicted from the regression model, however.)

This and other high-leverage observations may be seen in other graphical displays; but it is useful to know here that they will often also be quite noticeable in what we propose here as the collinearity biplot. The web page of supplementary materials for this article (Section 5.2) shows a robust outlier-detection QQ plot and an influence plot of these data for comparison with more well-known methods.

4.2 Visualizing condition indices

The condition indices are defined in terms of the ratios λ_1/λ_k (on a square root scale). It is common to refer to the maximum of these, $\sqrt{\lambda_1/\lambda_p}$ as the *condition number* for the entire system, giving the worst near-singularity for the entire model.

A related visualization, that potentially provides more information than just the numerical condition indices can be seen in biplots of Dimension 1 vs. Dimension k, where, typically only the k corresponding to the smallest eigenvalues are of interest. To visualize the relative size of λ_k to λ_1 it is useful to overlay this plot with a data ellipse for the component scores.

Figure 4 shows the condition number biplot for the Cars data, where the condition number can be approximately seen as the ratio of the horizontal to the vertical dimensions of the data ellipse. As the

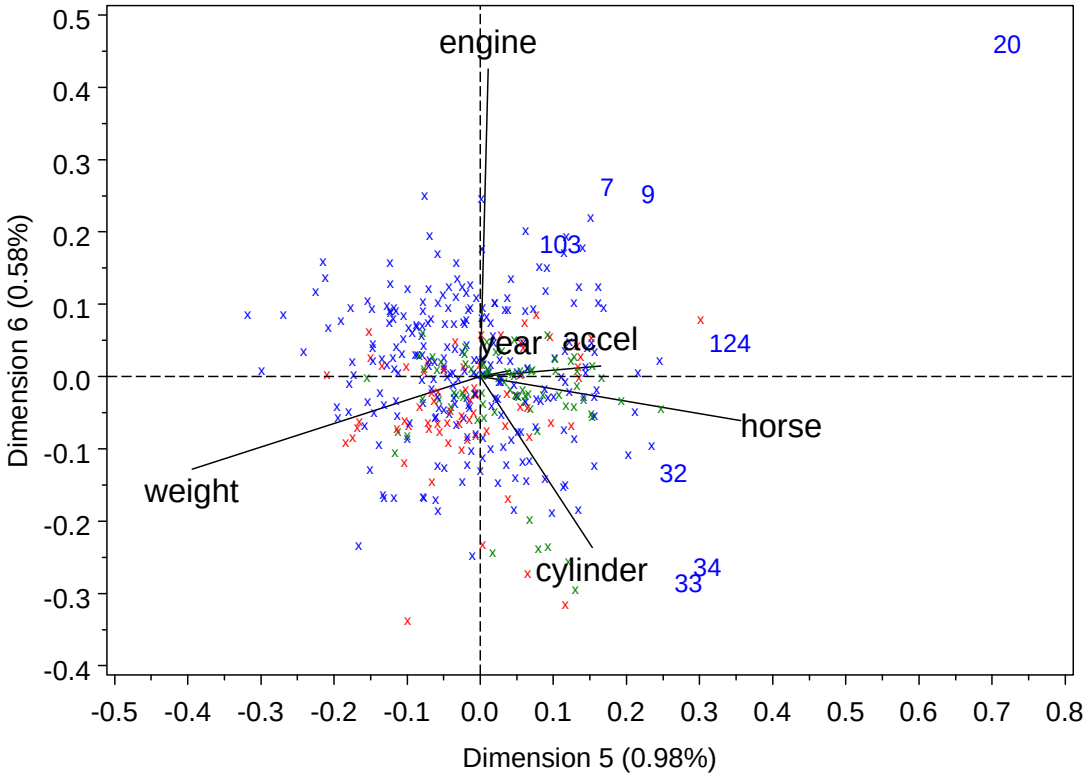

Figure 3: Collinearity biplot of the Cars data, showing the last two dimensions. The projections of the variable vectors on the coordinate axes are proportional to their variance proportions. To reduce graphic clutter, only the eight most outlying observations in predictor space are identified by case labels. An extreme outlier (case #20) appears in the upper right corner.

numerical values of the condition indices in Table 2 also indicate, ill-conditioning in this example is not particularly severe. In general, however, the relative lengths of the variable vectors approximately indicate those variables that contribute to ill-conditioning.

The observations can also be interpreted in this display in relation to their projections on the variable vectors. Case 20, the main high-leverage point is seen as having high projections on Dimension 1, which in this view is seen as the four variables with high VIFs: Engine, Horse, Weight, Cylinder. The observations that contribute to large condition indices are those with large projections on the smallest component, Dimension 6.

5 Other examples

5.1 Biomass data

Rawlings (1988, Table 13.1) described analyses of a data set concerned with the determination of soil characteristics influencing the aerial biomass production of a marsh grass, *Spartina alterniflora* in the Cape Fear estuary of North Carolina. The soil characteristics consisted of 14 physical and chemical properties, measured at nine sites (three locations $\times$ three vegetation types), with five samples per site, giving $n = 45$ observations. The data were collected as part of a Ph.D. dissertation by Richard Linthurst.

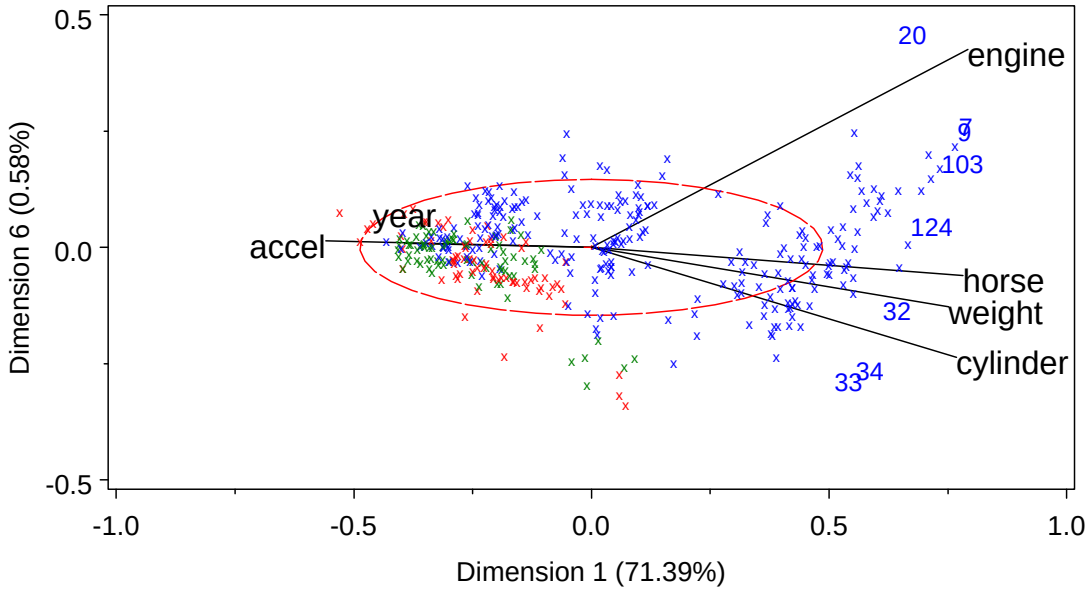

Figure 4: Condition number biplot of the Cars data, showing the first and last dimensions, with a 68% data ellipse for the component scores. The square root of the ratio of the horizontal and vertical axes of the ellipse portrays the condition number.

The quantitative predictors (i.e., excluding location and type) were: **H2S**: free sulfide; **sal**: salinity; **Eh7**: redox potential at pH 7; **pH**: soil pH; **buf**: buffer acidity at pH 6.6; and concentrations of the following chemical elements and compounds— **P**: phosphorus; **K**: potassium; **Ca**: calcium; **Mg**: magnesium; **Na**: sodium; **Mn**: manganese; **Zn**: zinc; **Cu**: copper, **NH4**: ammonium.

It should be noted that if the goal of the study was simply to *describe* biomass for the given timeframe and locations, collinearity would be less of an issue. But here, the emphasis in analysis focuses on identifying the "important" variable determinants of biomass.

The model with all 14 quantitative predictors fits very well indeed, with an $R^2 = .81$. However, the parameter estimates and their standard errors shown in Table 4 indicate that only two variables, K and Cu, have significant t statistics, all of the others being quite small in absolute value. For comparison, a stepwise selection analysis using 0.10 as the significance level to enter and remove variables selected the four-variable model consisting of pH, Mg, Ca and Cu (in that order), giving $R^2 = 0.75$; other selection methods were equally inconsistent and perplexing in their possible biological interpretations. Such results are a clear sign-post of collinearity. We see from Table 4 that many of the VIFs are quite large, with six of them (pH, buf, Ca, Mg, Na and Zn) exceeding 10. How many Waldos are hiding here?

The condition indices and variance proportions for the full model are shown, in our preferred form, in Table 5, where, according to our rule of thumb, 1+#(Condition Index> 10), we include the 4 rows with condition indices > 7. Following Belsley's (1991a) rules of thumb, it is only useful to interpret those entries (a) for which the condition indices are large and (b) where two or more variables have large portions of their variance associated with a near singularity. Accordingly, we see that there are two Waldos contributing to collinearity: the smallest dimension (#14), consisting of the variables pH, buf, and Ca, and the next smallest (#13), consisting of Mg and Zn. The first of these is readily interpreted as indicating that soil pH is highly determined by buffer acidity and calcium concentration.

Alternatively, Figure 5 shows a tableplot of the largest 10 condition indices and associated variance

Table 4: Linthall data: Parameter estimates and variance inflation factors

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr $>$ $\lvert t \rvert$	VIF
Intercept	1	2909.93	3412.90	0.85	0.401	0
H2S	1	0.42900	2.9979	0.14	0.887	3.02
sal	1	−23.980	26.170	−0.92	0.367	3.39
Eh7	1	2.5532	2.0124	1.27	0.214	1.98
pH	1	242.528	334.173	0.73	0.474	62.08
buf	1	−6.902	123.821	−0.06	0.956	34.43
P	1	−1.702	2.640	−0.64	0.524	1.90
K	1	−1.047	0.482	−2.17	0.038	7.37
Ca	1	−0.1161	0.1256	−0.92	0.363	16.66
Mg	1	−0.2802	0.2744	−1.02	0.315	23.76
Na	1	0.00445	0.02472	0.18	0.858	10.35
Mn	1	−1.679	5.373	−0.31	0.757	6.18
Zn	1	−18.79	21.780	−0.86	0.395	11.63
Cu	1	345.16	112.078	3.08	0.004	4.83
NH4	1	−2.705	3.238	−0.84	0.410	8.38

Table 5: Linthall data: Condition indices and variance proportions ($\times 100$). Variance proportions > 0.4 are highlighted.

#	Cond Index	Eigen value	H2S	sal	Eh7	pH	buf	P	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
14	22.78	0.0095	22	25	3	**95**	**70**	1	8	**60**	16	21	34	2	4	17
13	12.84	0.0298	0	9	2	0	12	8	28	1	**67**	25	16	**45**	0	21
12	10.43	0.0453	0	16	3	4	13	5	0	16	15	28	2	9	**43**	24
11	7.53	0.0869	19	3	0	0	4	1	7	18	1	5	14	32	9	2

proportions, using the same assignments of colors to ranges as described for Figure 1. Among the top three rows shaded red for the condition index, the contributions of the variables to the smallest dimensions is more readily seen than in tabular form (Table 5).

The standard biplot (Figure 6) shows the main variation of the predictors on the two largest dimensions in relation to the observations; here, these two dimensions account for 62% of the variation in the 14-dimensional predictor space. As before, we can interpret the relative lengths of the variable vectors as the proportion of each variable's variance shown in this 2D projection, and the (cosines of) the angles between the variable vectors as the approximation of their correlation shown in this view.

Based on this evidence it is tempting to conclude, as did Rawlings (1988, p. 366) that there are two clusters of highly related variables that account for collinearity: Dimension 1, having strong associations with five variables (pH, Ca, Zn, buf, NH4), and Dimension 2, whose largest associations are with the variables K, Na and Mg. This conclusion is wrong!

The standard biplot does convey useful information, but is misleading for the purpose of diagnosing collinearity because we are only seeing the projection into the 2D predictor space of largest variation and inferring, indirectly, the restricted variation in the small dimensions where Waldo usually hides. In the principal component analysis (or SVD) on which the biplot is based, the smallest four dimensions

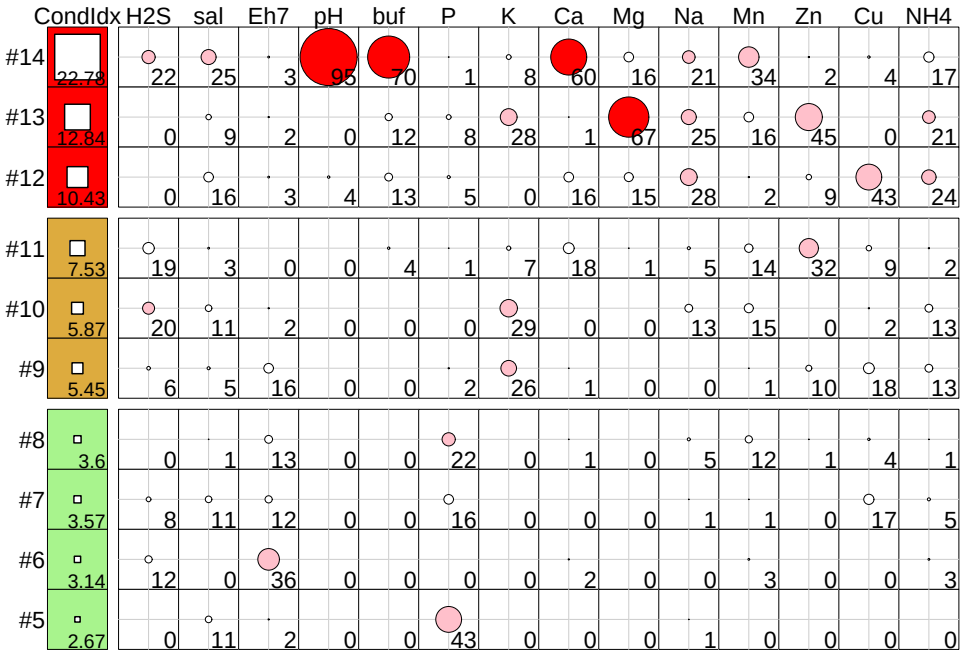

Figure 5: Tableplot of the 10 largest condition indices and variance proportions for the Linthall data. In column 1, the symbols are scaled relative to a maximum condition index of 30. In the remaining columns, variance proportions ($\times 100$) are scaled relative to a maximum of 100.

(shown in Table 5) account for 1.22% of the variation in predictor space. Figure 7 shows the collinearity biplot for the smallest two dimensions of $\boldsymbol{R}_{XX}$, or equivalently, the largest dimensions of $\boldsymbol{R}_{XX}^{-1}$, comprising only 0.28% of predictor variation.

Again, this graph provides a more complete visualization of the information related to collinearity than is summarized in Table 5, if we remember to interpret only the long variable vectors in this plot and their projections on the horizontal and vertical axes. Variables pH, buf, and Ca to a smaller degree stand out on dimension 14, while Mg and Zn stand out on dimension 13. The contributions of the other variables and the observations to these nearly singular dimensions would give the analyst more information to decide on an effective strategy for dealing with collinearity.

5.2 Further examples and software

The data and scripts, for SAS and R, for these examples and others, will be available at `www.math.yorku.ca/SCS/viscollin`, where links for biplot and tableplot software will also be found.

6 Discussion

As we have seen, the standard collinearity diagnostics— variance inflation factors, condition indices, and the variance proportions associated with each eigenvalue of $\boldsymbol{R}_{XX}$— *do* provide useful and relatively complete information about the degree of collinearity, the number of distinct near singularities, and the variables contributing to each. However, the typical tabular form in which this information is provided to the user is perverse— it conceals, rather than highlights, what is important to understand.

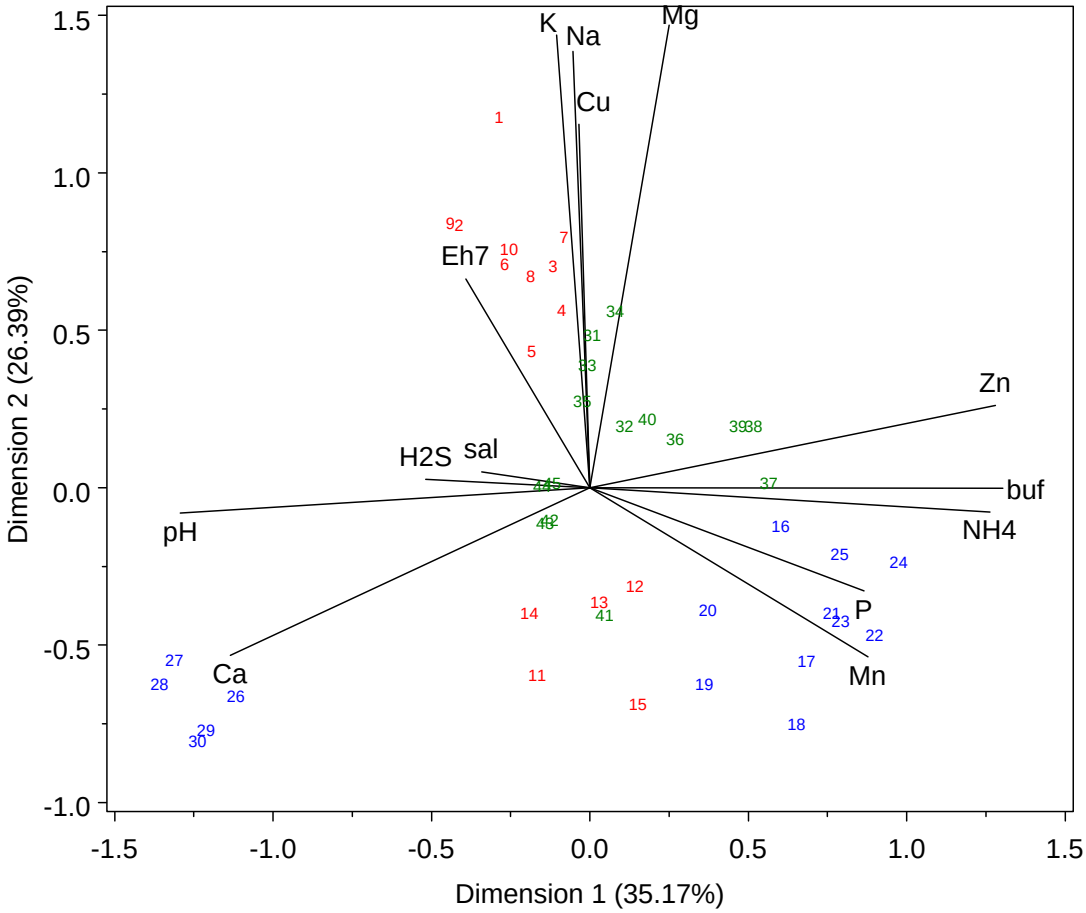

Figure 6: Standard biplot of the Linthurst data, showing the first two dimensions. The point labels refer to case numbers, which are colored according to the location of the site.

We have illustrated some simple modifications to the usual tabular displays designed to make it easier to see the number of Waldos in the picture and the variables involved in each. The re-ordered and reduced forms of these tables are examples of a general principle of *effect ordering for data display* (Friendly and Kwan, 2003) that translates here as "in tables and graphs, make the important information most visually apparent." Tableplots take this several steps further to highlight what is important to be seen *directly* to the eye. The collinearity biplot, showing the smallest dimensions, does this too, but also provides a more complete visualization of the relations among the predictors and the observations in the space where Waldo usually hides.

References

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, 38(2), 73–77. [with discussion, 78–93].

Belsley, D. A. (1991a). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York, NY: Wiley.

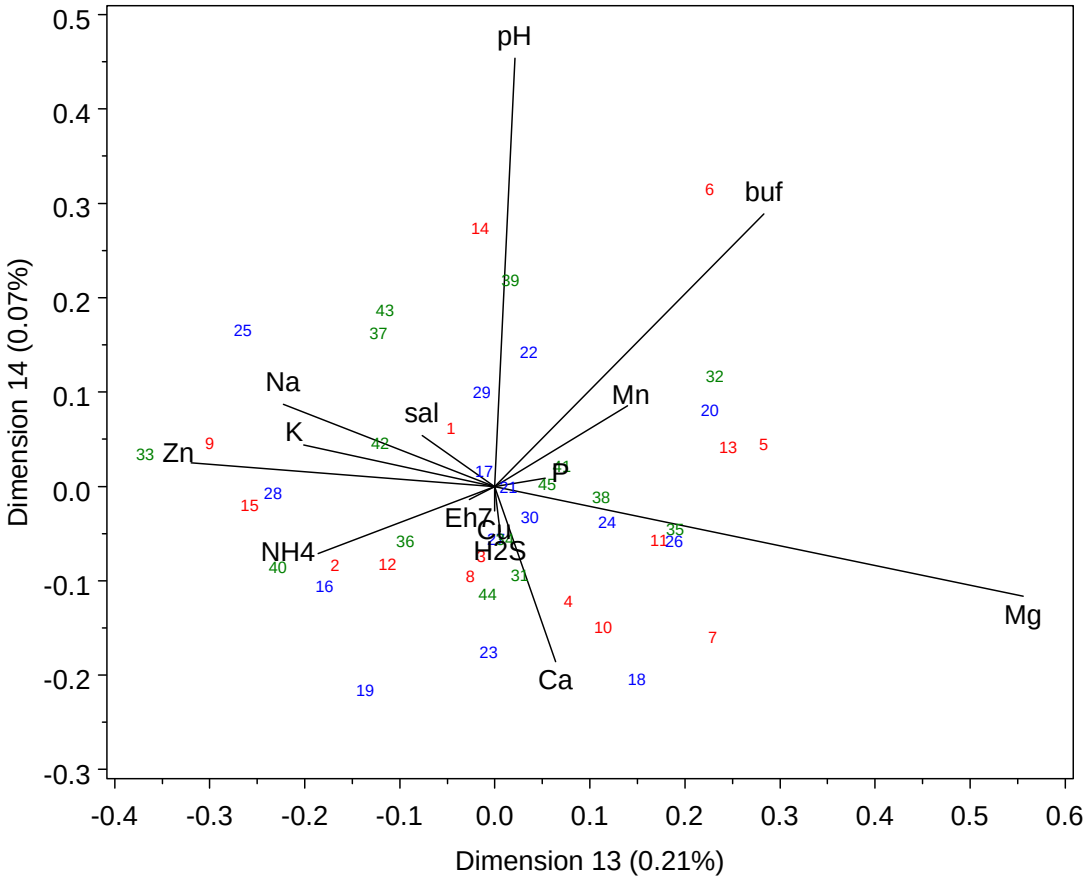

Figure 7: Collinearity biplot of the Linthurst data, showing the last two dimensions

Belsley, D. A. (1991b). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4, 33–50.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons.

Fox, J. (1984). *Linear Statistical Models and Related Methods*. New York: John Wiley and Sons.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage Publications.

Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4), 509–539.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics*, 58(3), 453–467.

Gower, J. C. and Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.

Hendrickx, J. (2008). `perturb`*: Tools for evaluating collinearity*. R package version 2.03.

Kwan, E. (2008a). *Improving Factor Analysis in Psychology: Innovations Based on the Null Hypothesis Significance Testing Controversy*. Unpublished doctoral dissertation, York University, Toronto, ON.

Kwan, E. (2008b). Tableplot: A new display for factor analysis. (in preparation).

Marquandt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12, 591–612.

Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool*. Pacific Grove, CA: Brooks/Cole Advanced Books.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, 5(3), 327–339.