Predictive nonlinear biplots: maps and trajectories

Journal Item

Predictive nonlinear biplots: maps and trajectories

S. K. Vines
Department of Mathematics and Statistics,
The Open University,
Walton Hall,
Milton Keynes, UK
E-mail: s.k.vines@open.ac.uk

January 27, 2015

Abstract

When the difference between samples is measured using a Euclidean-embeddable dissimilarity function, observations and the associated variables can be displayed on a nonlinear biplot. Furthermore, a nonlinear biplot is predictive if information on variables is added in such a way that it allows the values of the variables to be estimated for points in the biplot. In this paper an r dimensional biplot which maps the predicted value of a variable for every point in the plot, is introduced. Using such maps it is shown that even with continuous data, predicted values do not always vary continuously across the biplot plane. Prediction trajectories that appropriate for summarising such non-continuous prediction maps are also introduced. These prediction trajectories allow information about two or more variables to be estimated even when the underlying predicted values do not vary continuously.

Keywords: Euclidean-embeddable dissimilarity function, Nonlinear biplot, normal projection, prediction, prediction region, predictive trajectory

1 Introduction

A biplot is a plot in which information about samples and variables is simultaneously displayed. The term 'biplot' was first coined in Gabriel (1971), with the prefix 'bi-' intended to reflect that two different modes are displayed rather than the number of dimensions used for the display. Typically biplots are 2-dimensional, making them easy to display on paper or on computer screens,

although this does not have to be the case.

This paper will deal with biplots in which samples are represented by points on the plot whose positions are determined using classical scaling. Thus this method falls within the class of multidimensional techniques that produce low-dimensional representations of points based on their dissimilarities.

Furthermore, it will be assumed that the functional form of the underlying dissimilarity function is known, and is in the class of Euclidean-embeddable functions. This class of dissimilarity functions includes the familiar Pythagorean distance, along with other dissimilarity functions such as the square-root of the City Block (Manhattan) distance and Clark's distance (Gower and Legendre, 1986). This means that the low-dimensional representation of the points is a projection of a high-dimensional configuration that exactly represents the dissimilarities between points instead of rather an low-dimensional approximate representation of dissimilarities obtained directly by, for example, minimising Stress or S-Stress.

The aim will be to add information about variables to the plot in such a way that values of the variables can be associated with the configuration of the points. This will primarily be done by adding trajectories to the plot, one for each variable. The trajectories will in general be nonlinear, and hence such biplots are known as nonlinear biplots (Gower and Harding, 1988). These trajectories will be calculated by adding points that correspond to positions along an axis.

In multidimensional scaling, different approaches are available to add points to an existing configuration of points (see for example Borg and Groenen (2005)). In this paper, an approach which matches the construction of the existing configuration will be followed. That is, knowledge of the form of the dissimilarity function will be used to calculate the exact position of the extra points in a sufficiently high-dimensional space. Projection is then used to place these points on the biplot.

In linear (PCA) biplots (biplots that are produced when Pythagorean distance is chosen to be the dissimilarity function), the position of marker points on the trajectories representing variables depends on whether trajectories are to be used for interpolation or prediction (Gower and Hand, 1996, p.15). That is, on whether the trajectory is going to be used to placing a new observation in the most appropriate place on the biplot (interpolation) or to be used to determine what values of the original variables are best associated with a point on the biplot, usually one of those already plotted (prediction). In nonlinear biplots the trajectories themselves also generally depend on whether they are going to be used for interpolation or prediction (see for example Gower and Ngouenet (2005)). Here the focus will be on prediction trajectories. That is, trajectories complete with marker points, suitable for prediction purposes.

On nonlinear biplots, prediction trajectories also depend on the method by which points in the biplot are to be projected on to the trajectory. Here the focus will be on normal projection prediction trajectories. With normal projection prediction trajectories, a projection P^* of any point P in the biplot on the

trajectory is where the line PP^* intersects the trajectory orthogonally. The position of P^* along the trajectory then indicates the predicted value to be associated with P.

In the next section, existing methodology that has been used to calculate normal projection prediction trajectories for nonlinear biplots will be described. This existing methodology relies on the assumption that the dissimilarity function is smooth. This assumption is not always appropriate as there are Euclidean-embeddable dissimilarity functions that are not smooth everywhere. Hence the existing methodology cannot be applied to all such dissimilarity functions. So in Section 3 an alternative approach to prediction in nonlinear biplots is introduced so that normal projection prediction trajectories can be calculated regardless of whether the dissimilarity function is smooth.

As Section 3 will also show, the alternative approach to prediction introduced in this paper will allow prediction maps for each variable to be created – that is, plots where every point is coloured according to the value it predicts. Such maps can be used to explore how predicted values vary across the biplot plane. For example in Subsection 3.3 prediction maps will used to illustrate a new observation about nonlinear biplots: the dimension of prediction regions (regions on the biplot that all predict the same value of a variable) depends on whether the dissimilarity function is smooth. A mathematical explanation for this observation will be given in Subsection 3.4.

Prediction maps, by colouring every point in the biplot, effectively preclude the depiction of more than one variable on the same biplot. So, in Section 4 for the special case of 2-dimensional nonlinear biplots with 2-dimensional prediction regions, a new method of calculating a prediction trajectory through the biplot space to approximate the prediction regions is proposed. Then, by superimposing the prediction trajectories for the different variables on the same plot, the ability to compare different variables on the same biplot is restored.

2 Displaying variables in nonlinear biplots

2.1 Preliminaries

Let $\boldsymbol{X}$ represent an $n \times p$ data matrix of n samples and p variables, with its ith row vector $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$ representing the ith sample. Further suppose that the dissimilarity between two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is measured by the dissimilarity function, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which is Euclidean-embeddable. That is, it is possible to find a configuration of n points in m-dimensional space such that the Euclidean distance between the points representing samples i and j is $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Let $\boldsymbol{\Delta}$ be the doubly-centered matrix of dissimilarities multiplied by $-\frac{1}{2}$,

$$\boldsymbol{\Delta} = -\frac{1}{2}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{D}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')$$

where the (i,j)th element of $\boldsymbol{D}$ is $d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is the $n \times 1$ vector of ones. A consequence of Euclidean-embeddability is that $\boldsymbol{\Delta}$ is a positive semi-definite matrix (Schoenberg, 1935). So, via the spectral decomposition of $\boldsymbol{\Delta}$, it is possible to find a $n \times m$ real matrix $\boldsymbol{Y}$ such that $\boldsymbol{Y}\boldsymbol{Y}' = \boldsymbol{\Delta}$ and that $\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is a diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ in the main diagonal. Let $\boldsymbol{y}_i'$ be the ith row of $\boldsymbol{Y}$. Then $\boldsymbol{y}_i$ can be regarded as the location of the ith sample in m-dimensional space such that for $j = 1, \ldots n$, the distance between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ matches $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Usually $m = (n-1)$ meaning that the exact correspondence between inter-point distances and dissimilarities cannot normally be directly plotted on a low dimensional plot. However, as a result of least squares properties of spectral decompositions, the best rank r approximation of $\boldsymbol{\Delta}$ is obtained by simply using the first r columns of $\boldsymbol{Y}$ as the positions of the samples (see, for example, Gower and Harding (1988)).

Suppose now that we are interested in a new point $\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_p)$. Let $\boldsymbol{d}(\boldsymbol{\mu})$ be the $n \times 1$ vector of squared dissimilarities between $\boldsymbol{\mu}$ and the samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. That is, $\boldsymbol{d}'(\boldsymbol{\mu}) = (d^2(\boldsymbol{x}_1, \boldsymbol{\mu}), \ldots, d^2(\boldsymbol{x}_n, \boldsymbol{\mu}))$. Then, setting $\boldsymbol{z}'(\boldsymbol{\mu}) = (z_1(\boldsymbol{\mu}), \ldots, z_m(\boldsymbol{\mu}), z_{m+1}(\boldsymbol{\mu}))'$ where

$$(z_1(\boldsymbol{\mu}), \ldots, z_m(\boldsymbol{\mu}))' = \frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{Y}'\left(\frac{1}{n}\boldsymbol{D}\mathbf{1} - \boldsymbol{d}(\boldsymbol{\mu})\right) \tag{1}$$

and, given the values of $z_1(\boldsymbol{\mu}), \ldots, z_m(\boldsymbol{\mu})$, the value of $z_{m+1}(\boldsymbol{\mu})$ is taken so that

$$\boldsymbol{z}'(\boldsymbol{\mu})\boldsymbol{z}(\boldsymbol{\mu}) = \frac{1}{n}\mathbf{1}'\boldsymbol{d}(\boldsymbol{\mu}) - \frac{1}{2n^2}\mathbf{1}'\boldsymbol{D}\mathbf{1} \tag{2}$$

ensures that the (Euclidean) distance between the $(m+1)$ dimensional column vector $(\boldsymbol{y}_i', 0)'$ and $\boldsymbol{z}(\boldsymbol{\mu})$ matches $d(\boldsymbol{x}_i, \boldsymbol{\mu})$ (see, for example, Appendix A.7, Gower and Hand (1996)). So it is possible to add an extra point to represent $\boldsymbol{\mu}$ to the plot whilst preserving the correspondence between dissimilarities and distances on the plot, though in general doing so requires the dimensionality of the plot to be increased from m to $m+1$.

2.2 Prediction trajectories via normal planes

In Section 2.1, no mention was made about the process by which information about variables is added to the biplot. Gower and Harding (1988) show that this can be done by considering particular sequences of new points added to the plot.

For the kth variable, the sequence of new points is chosen to mimic the sequence of points along the kth axis of a p-dimensional scatterplot. That is, for variable k the particular sequence of points corresponds to $\mu_k \boldsymbol{e}_k$, where μ_k varies over the range the kth variable and $\boldsymbol{e}_k$ is the $p \times 1$ vector with all elements zero except for the kth element which has value 1.

The locations of these new points map out a trajectory on a $(m+1)$ dimensional plot, the kth "pseudoaxis". Technically the space in which the trajectory actually has a dimension that goes up with the number of points used to form it, because a new point added to a biplot requires one more dimension in order to represent all dissimilarities exactly. However, for practical purposes, it is sufficient to assume that the extra dimension added is the same dimension for all new points. See for example, Gower and Harding (1988).

Then, like axes on traditional scatterplots, the value of the kth variable that is associated with any particular point P in the $m+1$ dimensional space is determined by where P projects on to the kth pseudoaxis. It turns out that providing that the pseudoaxis is smooth, the plane $\mathcal{N}(\mu_k^*)$, normal to the kth pseudoaxis at the point P_k^* corresponding to $\mu_k = \mu_k^*$, includes all the positions in $(m+1)$-dimensional space which project back to P_k^* on the kth pseudoaxis. Thus, for any point in $\mathcal{N}(\mu_k^*)$, it is reasonable to associate ('predict') the value μ_k^* for the kth variable. Focusing on just the r dimensions of the plotted biplot space, this means that there is a just an $(r-1)$-dimension linear subspace for which the value μ_k^* for the kth variable is predicted. Furthermore it can be shown that for an r-dimensional biplot, this subspace corresponds to the subspace $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_r)$ where

$$\left(\boldsymbol{\alpha}' \boldsymbol{\Lambda}_r^{-1} \boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right) \frac{\partial \boldsymbol{d}(\mu_k^* \boldsymbol{e}_k)}{\partial \mu_k} = 0$$

(Gower and Ngouenet, 2005). In particular this means that for 2-dimensional biplots, the points $\boldsymbol{\alpha}$ associated with the value $\mu_k = \mu_k^*$ — the 'prediction region' for μ_k^* — all lie on a straight line. However, normal planes for different values of μ_k are in general not parallel so a point $\boldsymbol{\alpha}$ in the r-dimensional biplot might lie in more than one prediction region for variable k.

Note that when the dissimilarity function is additive, that is when it can be written in the form

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} d_k^2(x_{ik}, x_{jk}),$$

further simplification is possible. Then the prediction region for μ_k^* corresponds to all the points $\boldsymbol{\alpha}$ such that

$$\left(\boldsymbol{\alpha}' \boldsymbol{\Lambda}_r^{-1} \boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right) \frac{\partial \boldsymbol{d}_k(\mu_k^*)}{\partial \mu_k} = 0$$

where $\boldsymbol{d}_k'(\mu_k) = (d_k^2(x_{1k}, \mu_k), \ldots, d_k^2(x_{nk}, \mu_k))$. Thus when the dissimilarity function is additive, the prediction region for a value μ_k of the kth variable does not depend on what values are assumed for the other $(p-1)$ variables.

On the r-dimensional biplot, a prediction trajectory for the kth variable is then formed by joining the prediction regions for the different values of μ_k in such a way that all the points in the prediction region for μ_k^* project on to the same position on the trajectory.

Three different strategies for projecting on the prediction trajectory have been proposed: normal projection, circular projection and back projection (Gower

and Ngouenet, 2005). Of these, normal projection is arguably the most natural. With normal projection, the projection of a point P on the biplot, is a point P^* on the trajectory where the line PP^* intersects the trajectory orthogonally. In general normal projection prediction trajectories are not straightforward to compute as the calculation involves integration. Nevertheless numerical integration routines allow general implementations to be written (for example the R function `Nonlinbipl` described in Gower et al. (2011), which is included in the R package `UBbiplot`).

As has already been noted, the above theory all relies on the pseudoaxis being smooth. This in turn means requiring that the dissimilarity function is smooth for all values of μ_k. Unfortunately not all Euclidean-embeddable dissimilarity functions are smooth everywhere. For example, consider the square root of the Manhattan distance function:

$$d^2(\boldsymbol{x}, \boldsymbol{\mu}) = \sum_{k=1}^{p} |x_k - \mu_k|.$$

This dissimilarity function is Euclidean-embeddable (Gower and Legendre, 1986) and additive, but it is not smooth when $\mu_k = x_{ik}$, $i = 1, \ldots, n$. That is, this Euclidean-embeddable dissimilarity function is not smooth for values of μ_k that occur in the data. Therefore in the following section an alternative way of calculating prediction regions that does not depend on the smoothness of dissimilarity functions is pursued.

3 Prediction regions via least squares

In the previous section, planes normal to pseudoaxes were used to determine the prediction regions and hence normal projection prediction trajectories. However, as was pointed out in that section, the use of normal planes makes the implicit assumption that the dissimilarity function is smooth everywhere. This assumption does not cover all additive Euclidean-embeddable dissimilarity functions, let alone some non-additive Euclidean-embeddable dissimilarity functions. Thus an approach that does not depend on the smoothness of the dissimilarity function is required.

3.1 Prediction at a point via least squares

Consider a point $\boldsymbol{\alpha}$ in a biplot of dimension r (typically $r = 2$). A question of interest is what values of the original p variables should be associated with $\boldsymbol{\alpha}$?

In the previous section, consideration of normal planes was used to try to answer this question. An alternative approach is via least squares: finding the value $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ such that the (Euclidean) distance between $\boldsymbol{\alpha}^* = (\alpha_1, \ldots, \alpha_r, 0, \ldots, 0)'$ and $\boldsymbol{z}(\boldsymbol{\mu})$ is minimised. That is, for a given $\boldsymbol{\alpha}$ finding the value $\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}$ which minimises $g(\boldsymbol{\mu}|\boldsymbol{\alpha}) = (\boldsymbol{\alpha}^* - \boldsymbol{z}(\boldsymbol{\mu}))'(\boldsymbol{\alpha}^* - \boldsymbol{z}(\boldsymbol{\mu}))$.

Denoting $(z_1(\boldsymbol{\mu}), \ldots, z_r(\boldsymbol{\mu}))'$ by $\boldsymbol{z}_r'(\boldsymbol{\mu})$, the first r columns of $\boldsymbol{Y}$ by $\boldsymbol{Y}_r$ and the $r \times r$ submatrix formed from the first r columns and rows of $\boldsymbol{\Lambda}$ by $\boldsymbol{\Lambda}_r$ then $g(\boldsymbol{\mu}|\boldsymbol{\alpha}) = (\boldsymbol{\alpha} - \boldsymbol{z}_r(\boldsymbol{\mu}))'(\boldsymbol{\alpha} - \boldsymbol{z}_r(\boldsymbol{\mu})) + \boldsymbol{z}'(\boldsymbol{\mu})\boldsymbol{z}(\boldsymbol{\mu}) - \boldsymbol{z}_r'(\boldsymbol{\mu})\boldsymbol{z}_r(\boldsymbol{\mu})$. Thus using equations (1) and (2),

$$\begin{aligned} g(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= \boldsymbol{\alpha}'\boldsymbol{\alpha} - 2\boldsymbol{\alpha}'\boldsymbol{z}_r(\boldsymbol{\mu}) + \boldsymbol{z}'(\boldsymbol{\mu})\boldsymbol{z}(\boldsymbol{\mu}) \\ &= \boldsymbol{\alpha}'\boldsymbol{\alpha} - \boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r'\left(\frac{1}{n}\boldsymbol{D}\mathbf{1} - \boldsymbol{d}(\boldsymbol{\mu})\right) + \frac{1}{n}\boldsymbol{d}'(\boldsymbol{\mu})\mathbf{1} - \frac{1}{2n^2}\mathbf{1}'\boldsymbol{D}\mathbf{1} \\ &= \boldsymbol{\alpha}'\boldsymbol{\alpha} - \frac{1}{2n^2}\mathbf{1}'\boldsymbol{D}\mathbf{1} - \frac{1}{n}\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r'\boldsymbol{D}\mathbf{1} + (\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}')\boldsymbol{d}(\boldsymbol{\mu}) \\ &= \text{constant} + (\boldsymbol{\alpha}^T\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}')\boldsymbol{d}(\boldsymbol{\mu}) \end{aligned}$$

where the constant is not dependent on $\boldsymbol{\mu}$. Thus, for fixed $\boldsymbol{\alpha}$, $g(\boldsymbol{\mu}|\boldsymbol{\alpha})$ is a constant plus a weighted average of the squared dissimilarities between each of the samples $\boldsymbol{x}_i$ and $\boldsymbol{\mu}$.

Further simplification is possible for additive dissimilarity functions. In such cases

$$g(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \sum_{k=1}^{p} g_k(\mu_k|\boldsymbol{\alpha})$$

where

$$g_k(\mu_k|\boldsymbol{\alpha}) = \text{constant} + \left(\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right)\boldsymbol{d}_k(\mu_k). \tag{3}$$

This means that for additive dissimilarity functions (smooth or not), predictions for the kth variable are not conditional on the values assumed for any other variable at any point in the biplot.

When the dissimilarity function is smooth, local minima of $g_k(\mu_k|\boldsymbol{\alpha})$ at any given point $\boldsymbol{\alpha}$ can be found by solving

$$\frac{\partial g_k(\mu_k|\boldsymbol{\alpha})}{\partial \mu_k} = \left(\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right)\frac{\partial \boldsymbol{d}_k(\mu_k)}{\partial \mu_k} = 0.$$

In other words local minima of $g_k(\mu_k|\boldsymbol{\alpha})$ correspond to values of μ_k for which $\boldsymbol{\alpha}$ lies in the prediction regions obtained in Section 2.2. However, explicit use of the least squares principle clarifies what should be done when a point $\boldsymbol{\alpha}$ lies in more than one prediction region for a variable: the value of μ_k that corresponds to the global minimum of $g(\mu_k|\boldsymbol{\alpha})$ should be selected.

Equation (3) leads to simple way of finding the predicted value of μ_k at $\boldsymbol{\alpha}$ which does not depend on the smoothness of the dissimilarity function. The function $g_k(\mu_k|\boldsymbol{\alpha})$ can evaluated over a range of values of μ_k. Then it is just a question of identifying the value of μ_k that produces the smallest value of $g_k(\mu_k|\boldsymbol{\alpha})$. Provided that that the global minimum lies within the range evaluated, this approach also reduces the risk of the local minimum that happens to be the global minimum is missed. For this reason, the range of values of μ_k for which $g_k(\mu_k|\boldsymbol{\alpha})$ is evaluated should always include the values observed in the data.

3.2 Producing prediction maps

Subsection 3.1 describes a straightforward way to obtain the predicted value of μ_k for any value of $\boldsymbol{\alpha}$, regardless of whether the dissimilarity function is smooth or not.

Repeating this process for a grid of points across the biplot plane leads naturally to a pixel-colouring algorithm for mapping predictions. Covering the biplot plane with pixels and colouring each according to its predicted value leads to a prediction map for variable k.

Although this simple pixel-colouring approach is computationally intensive, there is a short cut. The form of (3) means that the values of $\boldsymbol{d}_k(\mu_k)$ for a range of μ_k only have to be evaluated once regardless of the number of positions in the r dimensional plane for which predictions are required. So the form of the pixel-colouring algorithm for a prediction map of μ_k is as follows.

Step 1 For a range of values of μ_k, evaluate $\boldsymbol{d}_k(\mu_k)$.

Step 2 For each grid point α in the biplot plane, minimise (3), making use of the values calculated in Step 1.

For biplots of dimension $r = 2$, the pixel-colouring algorithm has been implemented in an R package by the author.

3.3 Example: Fighter aircraft

Table 2.3.1 in Cook and Weisberg (1982) contains data on fighter aircraft abstracted from Stanley and Miller (1979). For each of 21 fighter aircraft, four variables were retained: SPR, the specific power, proportional to the power per unit weight; RGF, the flight range factor; PLF, payload as a fraction of gross weight of aircaft; and SLF, sustained load factor.

Various biplots of these data have been published by Gower and Hand (1996), Gower and Ngouenet (2005) and Gower et al. (2011). Following Gower and Ngouenet (2005), suppose initially that the appropriate dissimilarity function to use with these data is Clark's distance:

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}} \right)^2 .$$

When $x_{ik} = x_{jk} = 0$ the contribution to the dissimilarity function is defined to be 0. This dissimilarity function is additive and downweights discrepancies between high valued observations relative to low valued ones. Furthermore, this is a scale-free dissimilarity function because the dissimilarities are invariant to multiplication of any variable by a positive constant, but it is not invariant to translation of the data points.

Prediction maps for all four variables are given in Figure 1. In these maps the variation of predicted values over the biplot becomes clear. There is a strong correlation between predicted values for RGF and SLF as the colour pattern is very similar. The value of SPR increases steadily from top left to bottom right of the plot whereas there is a sharp change in predicted PLF with aircraft 'r' and 'g' being in the region with much lower predicted values. On the maps individual prediction regions correspond to straight lines as expected. Furthermore in the maps for SPR and SLF, prediction regions for different values of μ_k are clearly not parallel

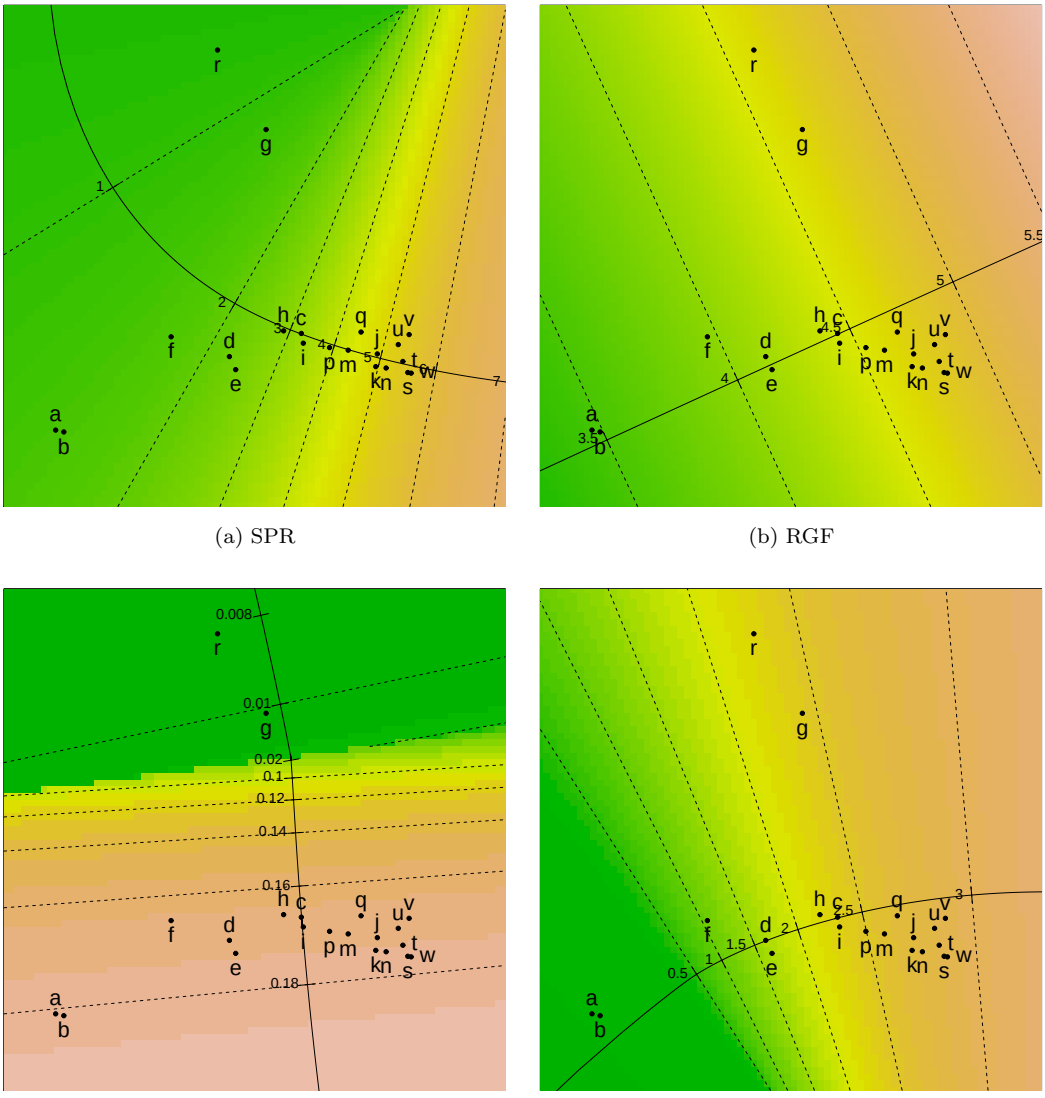
In the prediction map for PLF some of the prediction regions are curtailed. For example, the prediction region for $\mu_3 = 0.02$ is curtailed: it is not present on the left hand side of the map. This curtailment means that for points along the line corresponding to $\mu_3 = 0.02$ on the left hand side of the plot, $\mu_3 = 0.02$ is a local minimum of $g_3(\mu_3|\boldsymbol{\alpha})$ but it is not the global minimum of $g_3(\mu_3|\boldsymbol{\alpha}_2)$.

The form of (3) makes it easy to investigate the effect of changing the dissimilarity function. Most of the code can be written in terms of a generic d, only the functional form of d needs to be specifically tailored. For example, an alternative dissimilarity function is the square-root of the Canberra distance:

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$$

(see, for example, Cox and Cox (2001)) and when $x_{ik} = x_{jk} = 0$ the contribution to the dissimilarity function is defined to be 0. Like Clark's distance, this dissimilarity function is additive, scale-free and down-weights discrepancies between high-valued observations. However, this dissimilarity measure is based on the L_1 norm. So although $d(\boldsymbol{x}_i, \boldsymbol{\mu})$ is a continuous function, it is not smooth everywhere. Instead each function $d_k(x_{ik}, \mu_k)$, $k = 1, \ldots, p$ has a corner at each $\mu_k = x_{ik}$, $i = 1, \ldots, n$. Although the Euclidean-embeddability of this dissimilarity function has not been proven for all data sets, for these data $\boldsymbol{\Delta}$ is positive semi-definite. Hence square-root of the Canberra distance is, at the very least, Euclidean-embeddable for these data.

The resultant prediction maps are shown in Figure 2. Overall these plots reflect the similar functional forms of Clark's distance and square-root of the Canberra distance. For each variable, the patterns of colours are similar in Figures 1 and 2, indicating similarity in the variation of predicted values across the maps. However, there is one noticeable difference between the prediction maps in Figure 1 and Figure 2. The predicted values do not vary smoothly over the biplot maps in Figure 2. Instead all the biplot maps appear to be split into a series of 2-dimensional prediction regions according to the value predicted. The values associated with the 2-dimensional prediction regions turn out to correspond to values observed in the data. It is as though when the square root of the Canberra distance is used as the dissimilarity function the data are being treated as categorical for the purposes of prediction.



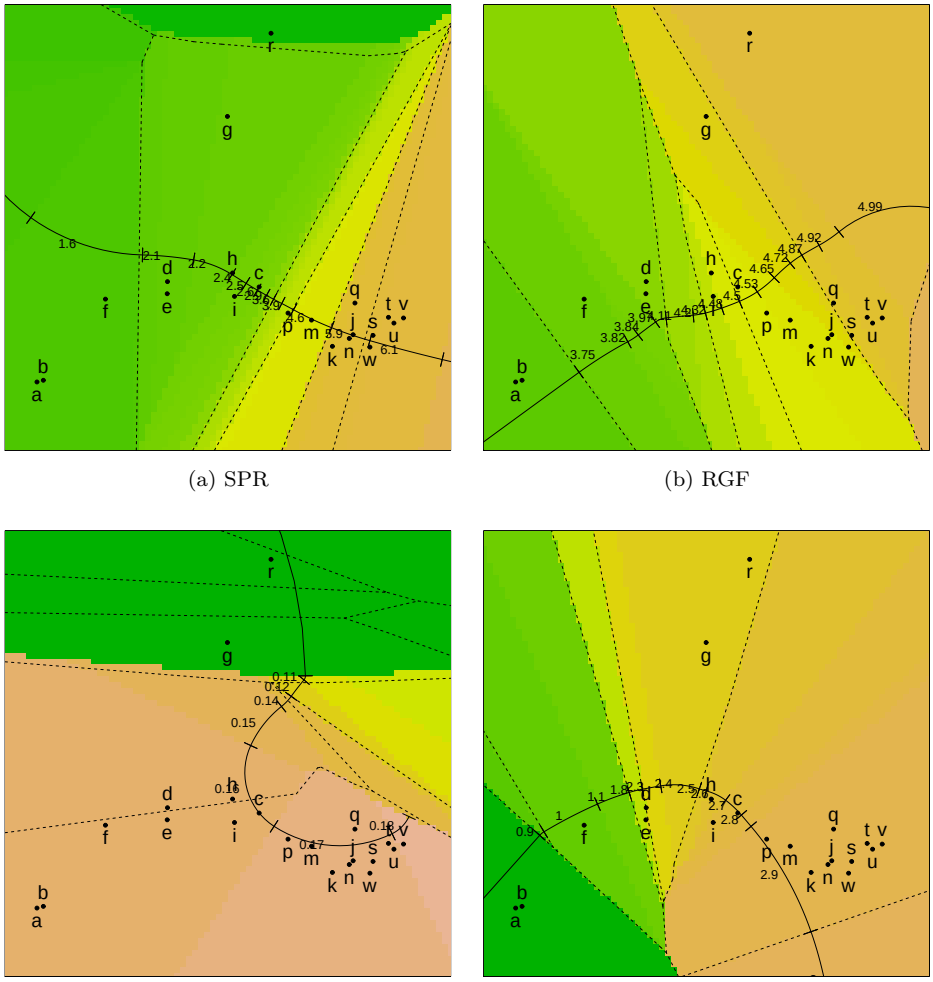
(a) SPR

(b) RGF

(c) PLF

(d) SLF

Figure 1: Prediction maps for the aircraft data using Clark's distance as the dissimilarity function. On each plot a normal projection prediction trajectory (solid line) is shown, along with selected prediction regions (dotted lines). The colour of each pixel indicates the predicted value, with green pixels corresponding to low values and yellow pixels corresponding to high values.



(a) SPR (b) RGF

(c) PLF (d) SLF

Figure 2: Prediction maps for the aircraft data using square-root of the Canberra distance as the dissimilarity function.

3.4 Dimension of prediction regions

Section 3.3 suggested that the dimension of prediction regions depends on the dissimilarity function. In this section it will be shown that the dimensionality can be attributed to the smoothnesss of the dissimilarity function. For simplicity of exposition, the focus will be on additive dissimilarity functions. However, similar arguments will apply to non-additive dissimilarity functions.

For variable k, let $L_k(\mu_0)$ denote the region of the biplot for which $\mu_k = \mu_0$ is a local minimum of $g_k(\mu_k|\boldsymbol{\alpha})$, let $\frac{\partial_- f(x)}{\partial x}$ denote the partial derivative of $f(x)$ with respect to x when approaching from the left-hand side of the function and let $\frac{\partial_+ f(x)}{\partial x}$ denote the partial derivative of $f(x)$ with respect to x when approaching from the right-hand side of x. Assuming that $g_k(\mu_k|\boldsymbol{\alpha})$ is continuous and sufficiently well-behaved, at any local minimum with respect to μ_k, we must then have $\frac{\partial_- g_k(\mu_k|\boldsymbol{\alpha})}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k} \leq 0$ and $\frac{\partial_+ g_k(\mu_k|\boldsymbol{\alpha})}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k} \geq 0$. Thus we must have the following:

$$\left(\boldsymbol{\alpha}\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right) \frac{\partial_- \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k} \leq 0 \text{ and } \left(\boldsymbol{\alpha}\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right) \frac{\partial_+ \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k} \geq 0.$$

Setting

$$\boldsymbol{s}^- = \left(\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r'\right) \frac{\partial_- \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k}, \qquad \boldsymbol{s}^+ = \left(\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r'\right) \frac{\partial_+ \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k}$$

$$c^- = \frac{1}{n}\mathbf{1}' \frac{\partial_- \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k} \qquad c^+ = \frac{1}{n}\mathbf{1}' \frac{\partial_+ \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}\Big|_{\mu_k=\widehat{\mu}_k}$$

this means that at a local minimum we require that

$$\boldsymbol{\alpha}\boldsymbol{s}^- \leq -c^- \text{ and } \boldsymbol{\alpha}\boldsymbol{s}^+ \leq -c^+.$$

This means that half the biplot plane meets the necessary condition of the minimum based on the left derivative. Similarly half the biplot plane meets the necessary condition for a minimum based on the right derivative. In general the two half planes will intersect. Thus, $L_k(\mu_0)$ will in general to correspond to an unbounded region in $\mathbb{R}^r$.

However, when the underlying dissimilarity function is smooth at μ_k,

$$\frac{\partial_- \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k} = \frac{\partial_+ \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k} = \frac{\partial \boldsymbol{d}_k'(\mu_k)}{\partial \mu_k}.$$

This means that $\boldsymbol{s}^- = \boldsymbol{s}^+ = \boldsymbol{s}$ and $c^- = c^+ = c$. Thus, at $\boldsymbol{\alpha}$, μ_k is a local minimum of $g_k(\mu_k|\boldsymbol{\alpha})$ when $\boldsymbol{\alpha}\boldsymbol{s} \leq -c$ and $\boldsymbol{\alpha}\boldsymbol{s} \geq -c$. This is only possible if in fact $\boldsymbol{\alpha}$ lies in the plane defined by $\boldsymbol{\alpha}\boldsymbol{s} = -c$. So, when the dissimilarity function is smooth with respect to μ_k, $L(\mu_k)$ can at most be a subspace in $\mathbb{R}^{r-1}$.

On prediction maps, such as those shown in Figures 1 and 2, each pixel displays the value of μ_k that corresponds to the global minimum of $g_k(\mu_k|\boldsymbol{\alpha})$. When the

dissimilarity function is smooth, the global minimum for a particular pixel must also be a local minimum. So the visible prediction regions on the prediction map are also $(r-1)$-dimensional.

For dissimilarity functions such as the square root of the Canberra distance, the dissimilarity function is smooth for most values of μ_k, and hence the prediction regions for these values of μ_k are $(r-1)$-dimensional. However, at the points $\mu_k = x_{1k}, \ldots, x_{nk}$ the dissimilarity function is continuous but not smooth. Thus for these values of μ_k, the corresponding prediction regions are r-dimensional. Furthermore, in the case of the aircraft data, it is the prediction regions for $\mu_k = x_{1k}, \ldots, x_{nk}$ that dominate. This implies that only the values $\mu_k = x_{1k}, \ldots, x_{nk}$ tend to minimise $g_k(\mu_k|\boldsymbol{\alpha})$ globally.

4 Prediction trajectories to represent 2-dimensional prediction regions in a 2-dimensional biplot

In Section 3.4 it was shown that for values of μ_k where the dissimilarity function is continuous but not smooth, the dimension of the prediction region is r. In such situations it is not possible to cross the prediction region for μ_k orthogonally whilst remaining in the biplot. Hence the process described in Section 2.2 for generating normal prediction trajectories needs modification for such values of μ_k.

As noted in the previous section, in theory over the plotted region, a mixture of r-dimensional and $(r-1)$-dimensional predictions regions might appear on the prediction map. In the case of a 2-dimensional biplot this means that the prediction regions are a mixture of 2-dimensional and 1-dimensional prediction regions. However initial experience so far suggests that the 2-dimensional prediction regions tend to dominate. On prediction maps, the proportion of the area of the plot in which non-data values appear as predicted values has so far been negligibly small. Indeed, when the dissimilarity function is chosen to be the square root of the City Block distance the predicted value at a point will always correspond to a data value (see Appendix). Thus the focus will be only on values of μ_k that correspond to data values, the so-called 'basic points' (Gower and Hand, 1996). Furthermore, the dissimilarity function is assumed to be additive, though the same principles are likely to apply to non-additive dissimilarity functions.

Consider two distinct values μ_{Ak} and μ_{Bk}. In terms of prediction there will be nothing to choose between two values when $g_k(\mu_{Ak}|\boldsymbol{\alpha}) = g_k(\mu_{Bk}|\boldsymbol{\alpha})$. That is when

$$\boldsymbol{\alpha}\left(\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r'(\boldsymbol{d}_k(\mu_{Ak}) - \boldsymbol{d}_k(\mu_{Bk}))\right) = \frac{1}{n}\mathbf{1}'(\boldsymbol{d}_k(\mu_{Ak}) - \boldsymbol{d}_k(\mu_{Bk})).$$

This means that, on the biplot, the region of points for which μ_{Ak} and μ_{Bk} are equivalent predictively is an 1-dimensional line $\{l_{A,B}(\boldsymbol{\alpha}) : c_1\alpha_1 + c_2\alpha_2 = c_0\}$

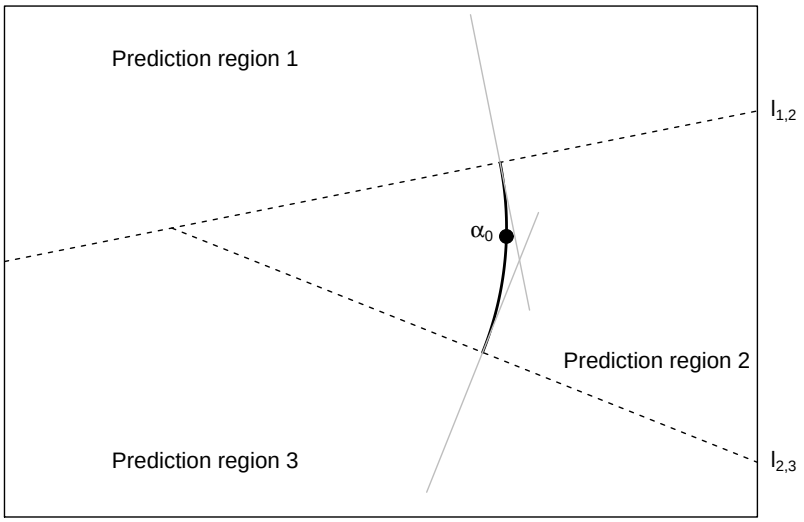


Figure 3: Illustration of the calculation of a trajectory segment when the prediction region is 2-dimensional. The solid black line is arc a_2. The three prediction regions are for μ_1, μ_2 and μ_3 respectively.

where the coefficients c_0, c_1, c_2 are easily calculated using the dissimilarity function.

4.1 Obtaining prediction trajectories for a simple prediction map

Suppose first that there are just three unique values, μ_{1k}, μ_{2k} and μ_{3k} for the kth variable in $\boldsymbol{X}$. Further suppose that the prediction region for μ_{2k} lies between those for μ_{1k} and μ_{3k}. Then for any point $\boldsymbol{\alpha}_0$ in the prediction region for μ_{2k}, it is possible to define the arc $a_2(\boldsymbol{\alpha})$ as the part of the circle bounded by the lines $l_{2,1}$ and $l_{2,3}$, which goes through $\boldsymbol{\alpha}_0$ and whose centre is the intersection of the lines $l_{2,1}$ and $l_{2,3}$ (Figure 3). Thus, by construction, arc a_2 meets both lines $l_{2,1}$ and $l_{2,3}$ perpendicularly. The line segment, a_1, can also be defined as the line in the prediction region for μ_{1k} that meets the line $l_{2,1}$ perpendicularly at the intersection of $a_2(\boldsymbol{\alpha})$ with $l_{2,1}$. Similarly the line segment, a_3, can also be defined as the line in the prediction region for μ_{3k} that meets the line $l_{2,3}$ perpendicularly at the intersection of $a_2(\boldsymbol{\alpha})$ with $l_{2,3}$.

Then any point in the prediction region for μ_{2k} will project orthogonally on to a_2. Equally points in the prediction region for μ_{1k} will usually orthogonally project on to a_1 and points in the prediction region for μ_{3k} will usually orthogonally project on to a_3. Thus the smooth trajectory formed by joining a_1, a_2 and a_3 acts as a type of normal projection prediction trajectory. Predictive values are found by projecting orthogonally on to the trajectory. However, along

the prediction trajectory, the predicted values are ascribed to entire segments instead of varying smoothly within a segment.

4.2 Obtaining prediction trajectories when there are more than three prediction regions

In Subsection 4.1 it was assumed that there were only three unique values for μ_k. Now assume that there are n_k unique values for the kth variable: $\mu_k(1), \ldots, \mu_k(n_k)$ ordered so that $\mu_k(i+1) > \mu_k(i)$ for $i = 1, \ldots, (n_k - 1)$. The principle of stitching together arcs to form a prediction trajectory with similar properties to a normal projection prediction axis can be applied when $n_k > 3$, starting from a point $\boldsymbol{\alpha_0}$ whose predicted value is known to be $\mu_k(i)$.

However, extra care needs to be taken because not all points along the arc between $l_{i-1,i}$ and $l_{i,i+1}$ will necessarily predict $\mu_k(i)$; it is just that they will predict $\mu_k(i)$ ahead of $\mu_k(i-1)$ and $\mu_k(i+1)$. So at the point at which the arc crosses $l_{i,i+1}$ the predicted value $\mu_k(j)$ is calculated. If $j = i+1$, no adjustment is needed. However, if $j \neq i+1$ this indicates that on the prediction map it is the prediction region for $\mu_k(j)$ that borders $\mu_k(i)$, not $\mu_k(i+1)$. In this case the arc for $\mu_k(i)$ is recalculated to go between $l_k(i-1, i)$ and $l_k(i, j)$. Furthermore, if $j > i+1$, the next arc calculated is taken to be for $\mu_k(j)$, running initially between $l_{i,j}$ and $l_{j,j+1}$. The arcs for $\mu_k(i+1), \ldots, \mu_k(j-1)$ are missed out. If $j < i-1$, the trajectory is ended and no further arcs calculated. This leads to the following algorithm for determining the arcs that correspond to $\mu_k(j)$, $j \geq i$.

Step 1: Set $\boldsymbol{\alpha}_f = \boldsymbol{\alpha}_0$, $r = i-1$, $s = i$, $t = i+1$.

Step 2: Calculate the arc a^* that goes through the fixed point $\boldsymbol{\alpha}_f$ and that crosses the lines $l_{r,s}$ and $l_{s,t}$ perpendicularly. (In the special case that $t = n_k + 1$, just calculate the line segment that goes through fixed point $\boldsymbol{\alpha}_f$ and that crosses the lines $l_{r,s}$.)

Step 3: Calculate the predicted value(s) for μ_k at the intersection between a^* and $l_{s,t}$.

Step 4: If the predicted value(s) in Step 3 do not include μ_s, set t to be such that μ_t is one of the predicted values in Step 3. (In practice it is assumed that in this situation there will be only one predicted value identified in Step 3.) Return to Step 2.

Step 5: Take the arc a^* to the next segment of the trajectory.

Step 6: If $t > s$ and $s < n_k$, set $\boldsymbol{\alpha}_f$ to be the intersection between a^* and $l_{s,t}$, $s = t$ and $t = s+1$. Return to Step 2.

Else if $t < s$ or $s = n_k$, stop.

Thus the prediction trajectory is defined by a sequence of arcs, such that the corresponding value of μ_k monotonically increases. The sequence ends when

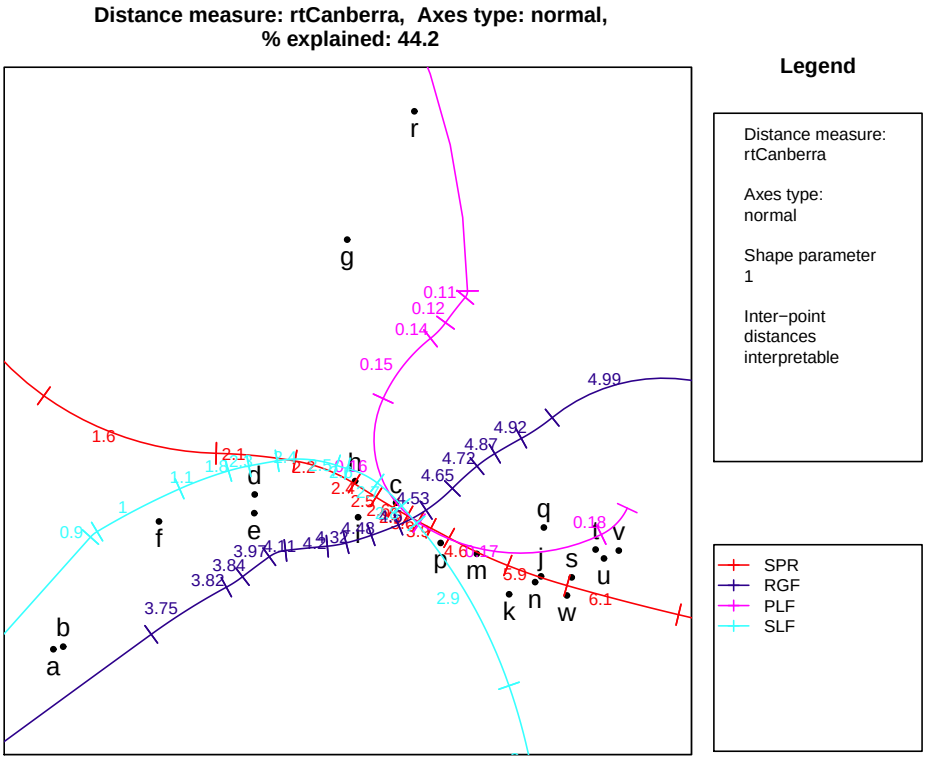


Figure 4: A predictive biplot, using normal axes, of the aircraft data using square-root of the Canberra distance dissimilarity function.

either the arc for $\mu_k(n_k)$ is added or that the prediction trajectory reaches the boundary of a prediction region $\mu_k(t)$ where $\mu_k(t)$ is lower than $\mu_k(s)$, the value of μ_k corresponding to the arc last added.

A similar procedure can then be used to determining the arcs that correspond to $\mu_k(j)$, $j \leq i$. The steps only have to be altered so that arcs corresponding to monotonically decreasing values of μ_k are considered.

The calculation of such prediction trajectories is included in the R package produced by the author. The resulting prediction trajectories for the fighter aircraft data calculated using this package for the biplot based on the square root of the Canberra dissimilarity function are given in Figure 4. Notice that although each prediction trajectory is made up of a series of arcs, a general pattern of variation over the biplot is depicted. For example, the predicted values of SPR increase from left to right.

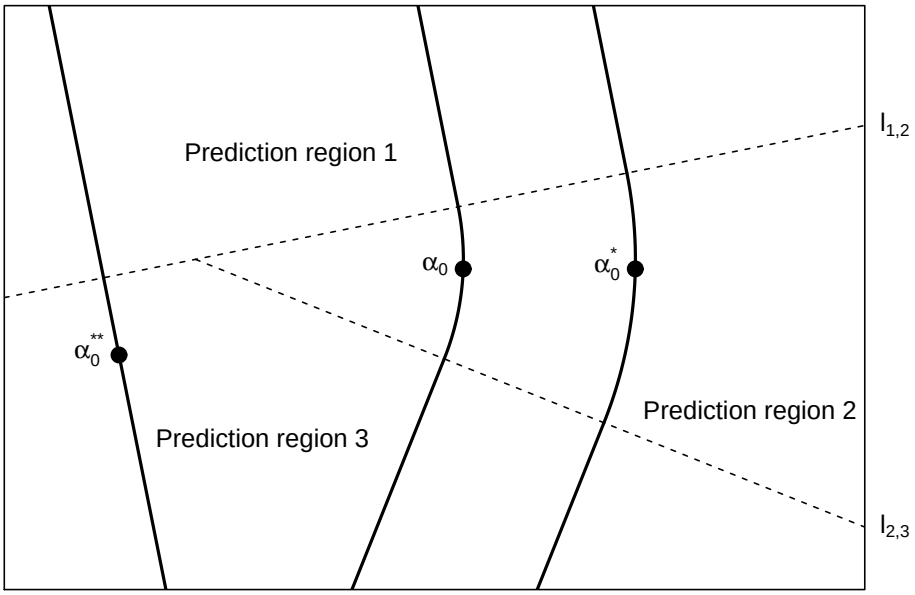


Figure 5: Illustration of trajectories obtained when different origins are selected

Comparison of predicted values obtained directly from the minimisation of (3) and via the trajectories also suggest that the trajectories provide a good summary of the information in the prediction map. For the 21 points representing the fighter aircraft, the prediction trajectories indicates a different value from that which would be obtained from the prediction map for only one fighter aircraft, aircraft 'r', for the variable SPR. Furthermore even for this aircraft, an orthogonal projection of the point representing aircraft 'r' does occur in the correct arc. The incorrect prediction is a result of there being a shorter orthogonal projection on to a different arc.

4.3 Influence of the origin

In Subsections 4.1 and 4.2, the trajectories were calculated to go through the point $\boldsymbol{\alpha_0}$. This initial point represents the origin of the trajectory, and as such is essentially arbitrary. Any point in the biplot plane could in theory be selected.

As Figure 5 shows, selection of a different position for $\boldsymbol{\alpha_0}$ can effect the resultant trajectory. On this figure, the prediction trajectories obtained based on three different origins $\boldsymbol{\alpha_0}$, $\boldsymbol{\alpha_0^*}$ and $\boldsymbol{\alpha_0^{**}}$ are depicted in a situation where there are just three prediction regions.

Points $\boldsymbol{\alpha_0}$ are $\boldsymbol{\alpha_0^*}$ both located in Prediction region 2. In this simple case this leads to the resultant trajectories being parallel. Thus both trajectories will lead to the same predictions being obtained.

Point $\boldsymbol{\alpha}_{\mathbf{0}}^{**}$ does not lie in Prediction region 2. Furthermore its position is such that in order to meet the line $l_{1,2}$ perpendicularly the resulting prediction trajectory does not go through Prediction region 2. So using this prediction trajectory points in Prediction region 2 will be given the same prediction are that in Prediction region 3.

Generally when the prediction map is more complex the shape of the trajectory depends on which prediction regions it happens to cross. However it is expected that the main impact on predictions obtained by changing the origin is whether or not a particular value is represented on the resulting trajectory at all. Otherwise it is to be expected that the values predicted for points in the biplot plane will not generally change.

However in all cases, points that lie exactly on a prediction trajectory are correctly predicted. For example, in Figure 5, the points on the prediction trajectory going through $\boldsymbol{\alpha}_{\mathbf{0}}^{**}$ that lie above $l_{1,2}$ would be correctly judged to correspond to a value of μ_{1k} and values below would correctly judged to correspond to a value of μ_{2k}. This property of prediction trajectories is different to prediction trajectories constructed using normal planes. The use of normal planes just means that for a particular value of μ_k, the prediction trajectory orthogonally intersects the corresponding prediction region. The same point might also lie on another prediction region, and one that is preferred according to the least squared criterion introduced in Section 3.1. So in this case, reading a value off for a point that lies exactly on the trajectory does not guarantee that it is the best predicted value for that point.

5 Discussion

Nonlinear biplots occur when Euclidean-embeddable, but not Pythagorean, dissimilarity functions are used to measure the difference between observations. This paper has shown how prediction maps can be used to display the variation in the predicted values of a variable over a low-dimensional biplot space. These prediction maps offer a visual way to assess the impact of the dissimilarity function on the biplot low-dimensional space.

Although novel for nonlinear biplots, the use of prediction maps is not new for another type of biplot: the generalized biplot. In generalized biplots, the data contains at least one variable that is categorical. Prediction maps then arise as a natural way to display the predictive regions for the levels of the categorical variables.

The properties of predictive regions for categories have been previously explored and an efficient algorithm for calculating them proposed (Gower, 1993). This algorithm is based on computing the boundaries between pairs and triples of predicted values, boundaries that are easy to calculate in the nonlinear biplot setting. Thus, potentially such an algorithm can also be used to compute prediction regions for nonlinear biplots.

The prediction trajectories proposed for 2-dimensional biplots constructed with non smooth dissimilarity functions provide a mechanism for displaying information for predicting the values of several variables on the same plot. The principle of stitching together a series of simple curves to form the trajectory is similar to that followed in Groenen et al. (2014).

In that paper, the authors stitch together a series of splines to form prediction trajectories. These spline-based prediction trajectories are calculated so that the predicted value for a point in the biplot is the value corresponding to the closest position on the trajectory to that point. Thus like normal projection prediction trajectories, the mechanism by which predicted values are read off is a natural one.

The calculation of the spline-based trajectories requires knowledge of what the predicted values should be at selected points. In Groenen et al. (2014), this is achieved by assuming that the predicted values at data points are simply the same as the observed values, an assumption that is also used when regression biplot axes are calculated. This has the advantage of avoiding the need to have a mathematical description of the dissimilarity function, greatly widening the class of multidimensional scaling plots to which the trajectories can be added. However for the nonlinear biplots described in this paper, it means that the spline-based trajectories are not coherent. That is, the method used to estimate the trajectories does not match the method by which the configuration of points was obtained. In particular it means that the spline-based trajectories may not accurately reflect the structure of the underlying prediction maps.

As yet few implementations of nonlinear predictive biplots are available. Two notable implementations are the R packages `BiplotGUI` (available from the Comprehensive R Archive Network) and `UBbiplot` (available as part of Gower et al. (2011)). However neither of these packages currently produce prediction maps, nor do they provide any mechanism for the production of normal projection predictive trajectories when the dissimilarity function is not smooth. The best that can be managed using these packages for such dissimilarity functions is circular projection predictive trajectories. Furthermore, the visual impression created by such circular predictive trajectories can be misleading. For example, Figure 6 is a nonlinear biplot with circular predictive trajectories for the aircraft data when the dissimilarity function is the square root of the City Block distance measure produced using `UBbiplot`.

The impression given is that predicted values vary smoothly along the trajectory. However this impression is incorrect. Inspection of the plotting positions of the trajectories reveals that for each of the variables, most of the plotting points along the trajectory coincide. Thus there is no movement along the trajectory for most of the values. This also explains the angular nature of the trajectories. A prediction map would quickly reveal the discrete range of predicted values possible for each variable. However neither package currently provided a facility for producing such maps.

In Subsection 4.3 it was noted that the origin for normal projection prediction trajectories constructed for biplots based on non-smooth dissimilarity functions

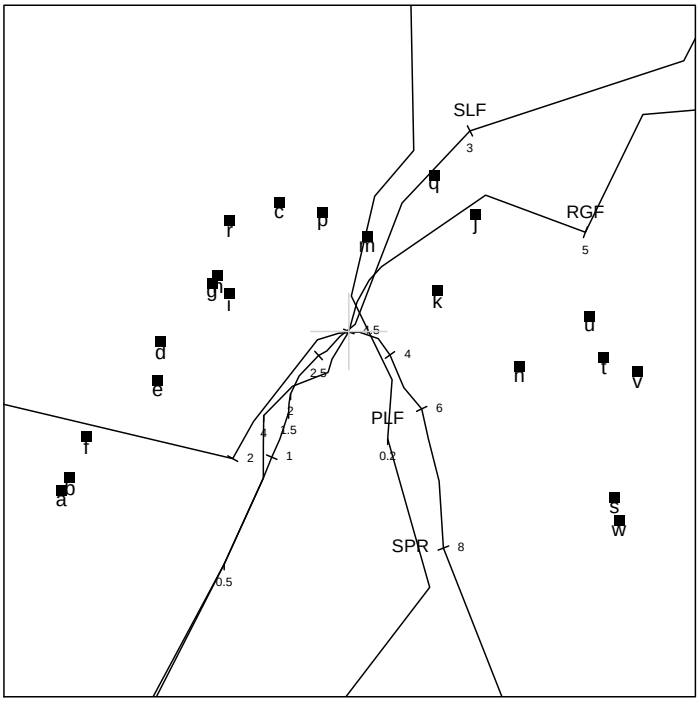


Figure 6: A predictive biplot, using circular projection axes, of the aircraft data using square-root of the City Block distance dissimilarity function produced using the R package `UBbiplot`.

of these prediction trajectories can be altered. Thus the origin can be placed away from the centre of the data points as suggested in Blasius et al. (2009). However given that points lying on such a trajectory are, by construction, guaranteed to have correct predictions given by the trajectory there is an argument for keeping trajectories as close as possible to data points.

When the dissimilarity function is smooth, the procedure described in Section 4 can still be used to generate prediction trajectories. In this case, the trajectories will only be approximate as predicted values that correspond to data values, and not any intermediate values, would be indicated by the trajectories. Further, the calculation of the prediction trajectories proposed in Section 4 merely require an ordering to be placed on the values of each variable. Thus there appears to be no reason why such prediction trajectories cannot be placed on generalized biplots to represent the prediction maps for ordinal variables.

Acknowledgements

I am very grateful to John Gower for greatly improving my understanding of biplots and for all the assistance he has given me during the preparation of this manuscript.

References

J. Blasius, P. H. C. Eilers, and J. C. Gower. Better biplots. *Computational Statistics & Data Analysis*, 53(8):3145–3158, 2009.

I. Borg and P. J. F. Groenen. *Modern multidimension scaling: Theory and applications, second edition.* Springer, 2005.

R. D. Cook and S. Weisberg. *Residuals and influence in regression.* Chapman & Hall, London, 1982.

T. F. Cox and M. A. A. Cox. *Multidimensional scaling, second edition.* Chapman & Hall, London, 2001.

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

J. C. Gower. The construction of neighbour-regions in two dimensions for prediction with multi-level categorical variables. In Otto Opitz, Berthold Lausen, and Rüdiger Klar, editors, *Information and Classification: Concepts, Methods and Applications*, pages 174–189. Springer-Verlag, 1993.

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

J. C. Gower and S. A. Harding. Nonlinear biplots. *Biometrika*, 75(3):445–455, 1988.

J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification*, 3:5–48, 1986.

J. C. Gower and R. F. Ngouenet. Nonlinearity effects in multidimensional scaling. *Journal of Multivariate Analysis*, 94:344–365, 2005.

J. C. Gower, S. Gardner-Lubbe, and N. J. R. le Roux. *Understanding Biplots*. Wiley, 2011.

P. J. F. Groenen, N. J. le Roux, and S. Gardner-Lubbe. Spine-based nonlinear biplots. *Advances in Data Analysis and Classification*, 2014.

I.J. Schoenberg. Remarks to Maurice Frechet's article 'Sur la definition axiomatique d'une classe d'espace distances vectoriellement applicable sure l'espace de Hilbert'. *Annals of Mathematics*, 36(3):724–732, 1935.

W. Stanley and M. Miller. Measuring technological change in jet fighter aircraft. Technical Report R-2249-AF, Rand Corporation, Santa Monica, 1979.

Appendix

In Section 3.1 it was shown that when the dissimilarity function is additive for any $\boldsymbol{\alpha}$ in a biplot of dimension r, the predicted value μ_k for the kth variable is chosen so that

$$g_k(\mu_k|\boldsymbol{\alpha}) = \text{constant} + \left(\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1}'\right)\boldsymbol{d}_k(\mu_k)$$

is minimised. This means that

$$g_k(\mu_k|\boldsymbol{\alpha}) = \sum_{i=1}^{n} w_i d_k^2(x_{ik}, \mu_k)$$

where w_i is the ith element of $(\boldsymbol{\alpha}'\boldsymbol{\Lambda}_r^{-1}\boldsymbol{Y}_r' + \frac{1}{n}\mathbf{1})$. Note that $\sum_{i=1}^{n} w_i = 1$.

When the dissimilarity function is chosen to be the square root of the City Block distance

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|.$$

So for the kth variable, the predicted value is the value of μ_k that minimises

$$g_k(\mu_k|\boldsymbol{\alpha}) = \text{constant} + \sum_{i=1}^{n} w_i |x_{ik}, \mu_k|.$$

To make notation easier, assume without loss of generality that for the kth variable the observations are ordered such that $x_{1k} \leq x_{2k} \leq \ldots \leq x_{nk}$.

For μ_k to be a global minimum, it must be a minimum in one of the following $(n+1)$ ranges.

- $\mu_k \leq x_{1k}$
- $x_{sk} \leq \mu_k \leq x_{(s+1)k}$ for $s = 1, \ldots, (n-1)$
- $\mu_k \geq x_{nk}$.

For values of μ_k such that $\mu_k \leq x_{1k}$. Then

$$\begin{aligned} g_k(\mu_k|\boldsymbol{\alpha}) &= \text{constant} + \sum_{i=1}^{n} w_i(x_{ik} - \mu_k) \\ &= \text{constant} - \sum_{i=1}^{n} w_i\mu_k = \text{constant} - \mu_k. \end{aligned}$$

This is minimised by taking μ_k as large as possible. That is, by setting $\mu_k = x_{1k}$.

For values of μ_k such that $x_{sk} \leq \mu_k \leq x_{(s+1)k}$ for a value of s in the range $1, 2, \ldots, (n-1)$. Then

$$\begin{aligned} g_k(\mu_k|\boldsymbol{\alpha}) &= \text{constant} + \sum_{i=1}^{s} w_i(\mu_k - x_{ik}) + \sum_{i=s+1}^{n} w_i(x_{ik} - \mu_k) \\ &= \text{constant} + \left(\sum_{i=1}^{s} w_i - \sum_{i=s+1}^{n} w_i \right) \mu_k. \end{aligned}$$

Now if $\left(\sum_{i=1}^{s} w_i - \sum_{i=s+1}^{n} w_i\right) > 0$, this means that $g_k(\mu_k|\boldsymbol{\alpha})$ is miminised in this range by setting μ_k as small as possible. That is by setting $\mu_k = x_{sk}$.

Similarly if $\left(\sum_{i=1}^{s} w_i - \sum_{i=s+1}^{n} w_i\right) < 0$, this means that $g_k(\mu_k|\boldsymbol{\alpha})$ is miminised in this range by setting μ_k as large as possible. That is by setting $\mu_k = x_{(s+1)k}$.

And if $\left(\sum_{i=1}^{s} w_i - \sum_{i=s+1}^{n} w_i\right) = 0$, this means that $g_k(\mu_k|\boldsymbol{\alpha})$ is constant throughout the range $x_{sk} \leq \mu_k \leq x_{(s+1)k}$.

Finally for values of μ_k such that $\mu_k \geq x_{nk}$. Then

$$\begin{aligned} &= \text{constant} + \sum_{i=1}^{n} w_i(\mu_k - x_{ik}) \\ &= \text{constant} + \mu_k. \end{aligned}$$

This is minimised by taking μ_k as small as possible. That is, by setting $\mu_k = x_{nk}$.

So the global minimum of $g_k(\mu_k|\boldsymbol{\alpha})$ is either $\mu_k = x_{ik}$, or $\mu_k = x_{ik}$, and $\mu_k = x_{(i+1)k}$ and all the values inbetween for some value i in the range $i = 1, 2, \ldots, n$. Thus the predicted value of the kth variable for a point $\boldsymbol{\alpha}$ in the biplot always includes one of the observed data values for that variable.