Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 12 Offprint

Constrained Biplots and Triplots

First published: September 2010
ISBN: 978-84-923846-8-6

Supporting websites:
http://www.fbbva.es
http://www.multivariatestatistics.org

CHAPTER 19

Constrained Biplots and Triplots

The pervading theme of this book has been the visualization of the maximum amount of information in a rectangular data matrix through a graphical display of the rows and columns, called the biplot. Often the rows are cases, displayed as points, and the columns are variables, displayed as vectors, and thanks to the scalar product property the projections of row points onto axes defined by the column vectors lead to approximations of the original data. Up to now no condition has been imposed on the solution apart from certain normalization conditions on the coordinates because of the indeterminacy of the matrix decomposition. In this final chapter we look at several ways of constraining the biplot display to have some additional condition on its solution. Imposing restrictions on a biplot necessarily makes it sub-optimal in representing the original data matrix, but in many situations such constraints add value to the interpretation of the data in relation to external information available about the rows or the columns.

Contents

More than a supplementary point

The idea of a constrained biplot can be illustrated using the "morphology" data set, the measurements of the 75 *Arctic charr* fish, and the log-ratio (LRA) biplot of Exhibit 7.3. The LRA biplot explained 37.5% of the variance (20.9% on the first axis, 16.6% on the second) of the 75 × 26 data matrix, which was logarithmically transformed and double-centred, called the *log-ratio transformation*. The body weight of each fish was also available, and it would be interesting to see if the body weight is related to the solution. This is achieved using the regression biplot of Chapter 2, where continuous variables can be added to an existing plot using

their regression coefficients on the dimensions of the map. A regression analysis is performed, of body weight as response and the fish coordinates on the two axes as explanatory variables, and the two standardized regression coefficients are used to draw this supplementary variable's direction. The coefficients turn out to be 0.116 and 0.203, with an R^2 of 0.055, and we could add a short vector to Exhibit 7.3 pointing towards the upper right (0.116 on the first axis, 0.203 on the second) to show the relationship of this additional variable to the biplot. An explained variance of 5.5% of the variable body weight, however, shows that this variable has little relationship with the biplot. The idea of adding the body weight variable was to see if there was any relationship between this variable and the shape of the fish (remember that it is the shape that the biplot is visualizing, not the size, thanks to the log-ratio transformation). For example, perhaps fish that are generally wider than they are longer may be heavier (this is frequently the case for humans!).

A more direct way of investigating this possible relationship is to *constrain* the first dimension of the biplot to be linearly related to body weight, so that body weight will coincide exactly with the first axis—that is, it will be 100% explained by the first axis—while the second axis will be the optimal axis not related to body weight but still trying to visualize the morphometric data as accurately as possible. As a spin-off we obtain a measure of exactly how much variance in the (log-ratio transformed) morphometric data is explained by body weight. Before we look at how the solution is obtained technically, let us look at the result of imposing the constraint, shown in Exhibit 12.1—body weight is now perfectly correlated with the first axis, pointing to the right. Body weight explains 4.0% of the variance of the morphometric variables (in the Computational Appendix we shall show that this percentage is highly significant statistically, with a *p*-value of 0.001), while the second axis (which is the first axis of the unconstrained space) explains 20.7%. A log-ratio link that is lying in this horizontal direction and which is long suggests the ratio *Bcw/Ed*, caudal body width relative to eye diameter—one might say the fat fish are heavy-tailed and beady eyed! Plotting body weight against this ratio does show a significant correlation of 0.489, and the slope of the relationship estimates a 1.84% increase in the ratio *Bcw/Ed* for every 10g increase in body weight (since $\exp(0.00182 \times 10) = 1.0184$). The variable *Bd*, body width at dorsal fins, is also in the same direction as *Bcw*, again supporting the not surprising result that heavier fish are wider. In a separate analysis a much weaker relationship was found with the morphological variables and body length.

Constraining by a categorical variable

Suppose that we want to constrain the biplot to be related to an external categorical variable; for example, the four-category sex-habitat variable for the fish data again.

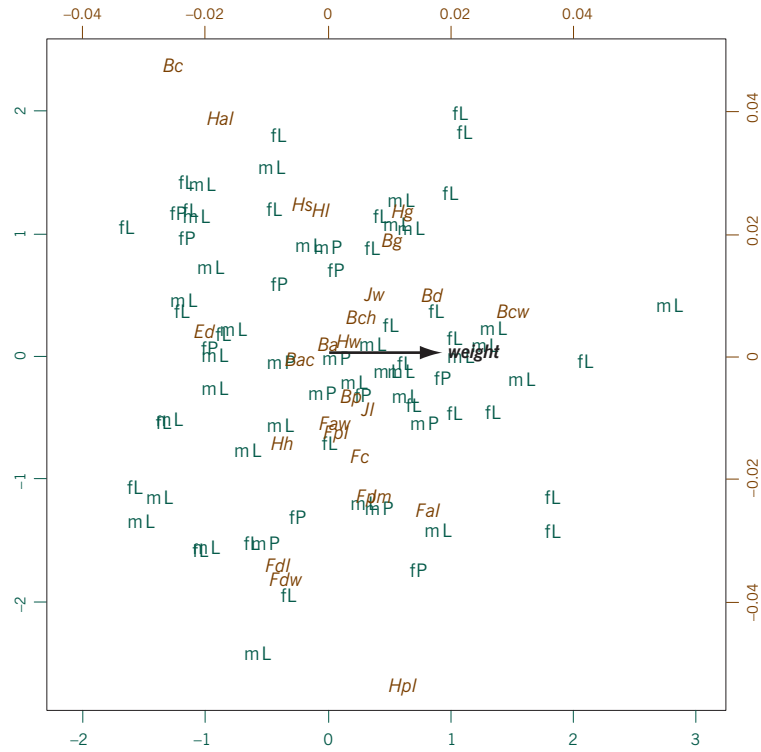


Exhibit 12.1:
The log-ratio biplot with the first axis constrained to be body weight. Rows (fish) are in standard coordinates, columns (morphometric variables) in principal coordinates. The constraining variable body weight follows the scale of the rows

This turns out to be equivalent to the discriminant analysis of Chapter 11. The sex-habitat variable is coded as four dummy variables, and it can be shown that the constrained space is exactly the space of the four centroids, thus the LRA-DA of Exhibit 11.2 is obtained. In a similar way, if the "women" data set is analyzed at an individual level, that is the indicator matrix **Z** is analyzed, and if the solution is constrained to be linearly related to the five dummy variables for marital status, this results in the CA (also called CA-DA) of the aggregated data in Exhibit 11.3.

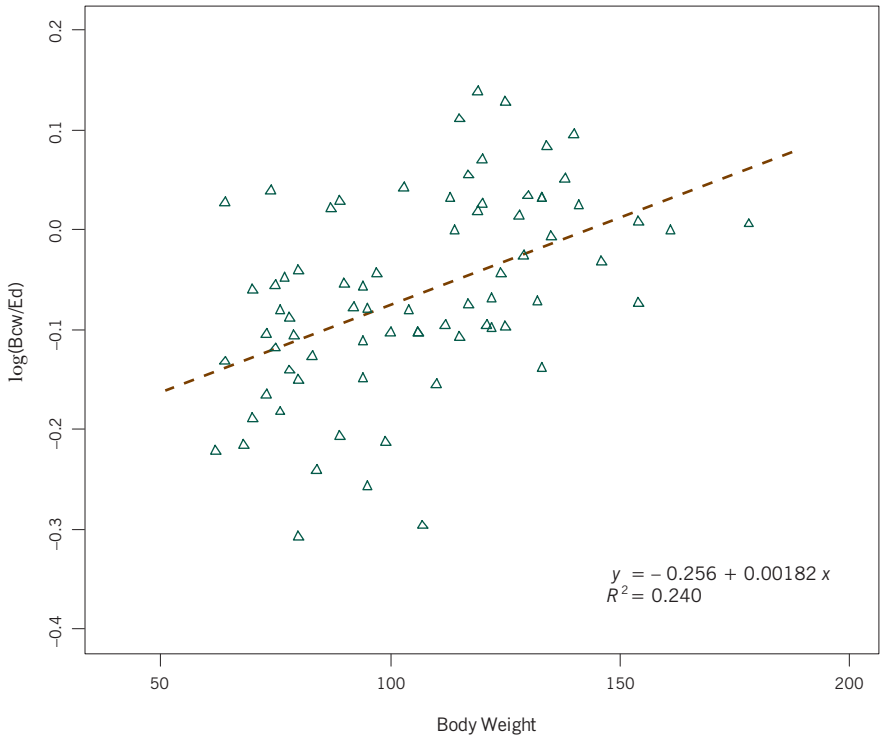
Constrained biplots

Whether it is a PCA, LRA or CA that is being performed, the imposition of the constraint is done in the same way. We give the theory for constrained CA, also known as *canonical correspondence analysis* (CCA), and then explain how this applies to the other methods. Regular unconstrained CA involves the SVD of the matrix of standardized residuals—see (8.2):

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathsf{T}} \tag{12.1}$$

Suppose that $\mathbf{X}$ denotes the $I \times M$ matrix of external constraining variables, and suppose that the columns of $\mathbf{X}$ have already been standardized. Notice that, be-

Exhibit 12.2:
The possible relationship between the log-ratio of Bcw to Ed and body weight that was diagnosed in the biplot



cause the rows are weighted by the masses in **r**, all calculations of mean and variance are performed using these masses, so the columns of **X** have weighted mean of 0 and weighted variance (inertia) of 1. Constraining the solution linearly means projecting **S** onto the space of **X**. The projection matrix is defined as follows (again, the masses are taken into account):

$$\mathbf{Q} = \mathbf{D}_r^{1/2}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_r^{1/2} \tag{12.2}$$

(one can easily check that **Q** satisfies the condition of a projection matrix: **QQ** = **Q**, i.e. applying the projection twice is the same as applying it once). The constrained (or restricted) version of **S** is then:

$$\mathbf{S}^* = \mathbf{QS} \tag{12.3}$$

From here on the calculations continue just as for CA, first calculate the SVD and then the principal and standard coordinates—see (8.2) to (8.4):

SVD: $$\mathbf{S}^* = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \tag{12.4}$$

Principal coordinates of rows: $\mathbf{F}^* = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G}^* = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (12.5)

Standard coordinates of rows: $\mathbf{\Phi}^* = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma}^* = \mathbf{D}_c^{-½}\mathbf{V}$ (12.6)

The above solution has as many principal axes as there are variables (or one less in the case of dummy variables).

There is a similar sequence of calculations to find the principal axes of the unconstrained space. Projection takes place onto the space orthogonal to (i.e., uncorrelated with) the variables in $\mathbf{X}$. This projection matrix is just $\mathbf{I} - \mathbf{Q}$, so the unconstrained (or unrestricted) part of $\mathbf{S}$ is now:

$$\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S} \tag{12.7}$$

(hence $\mathbf{S}$ has been split into two parts: $\mathbf{S} = \mathbf{S}^* + \mathbf{S}^\perp$). The same steps now proceed, where we re-use the same notation $\mathbf{U}$, $\mathbf{D}_\alpha$ and $\mathbf{V}$ for the SVD components, although they are numerically different here, of course:

SVD: $$\mathbf{S}^\perp = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top \tag{12.8}$$

Principal coordinates of rows: $\mathbf{F}^\perp = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G}^\perp = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (12.9)

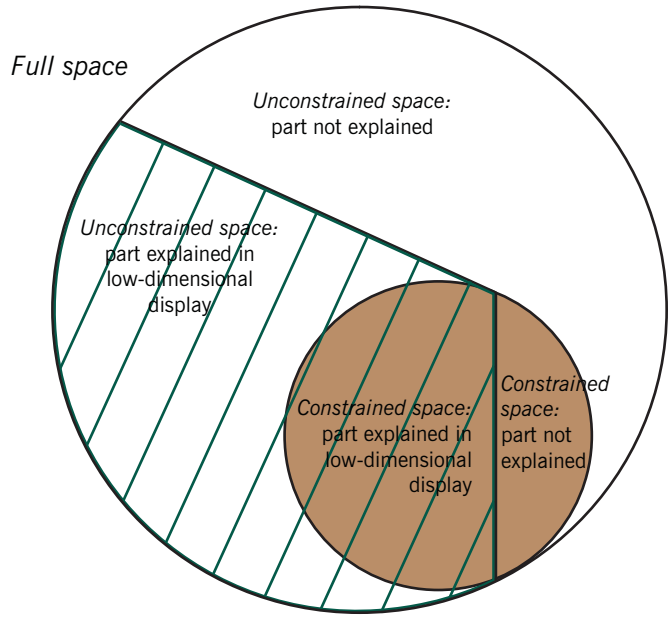
Standard coordinates of rows: $\mathbf{\Phi}^\perp = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma}^\perp = \mathbf{D}_c^{-½}\mathbf{V}$ (12.10)

Constrained LRA is almost identical to the above, starting with the double-centred matrix of log-transformed data, and using the same row and column masses as weights (in Chapter 15 we give the exact formulation). For unweighted LRA, these weights would just be $1/I$ for the rows and $1/J$ for the columns. Similarly for PCA, where the weights are generally equal, implementing the constraints involves starting with the centred (and optionally standardized) matrix, and applying the above steps using weights $r_i = 1/I$ and $c_j = 1/J$. Linearly constrained PCA has also been called *redundancy analysis* in the literature.

Decomposition of variance

When there are two constraining variables, they will still be perfectly explained by the plane of the first two constrained axes, but neither variable will necessarily be identified exactly with a principal axis. For three or more constraining variables the two-dimensional constrained space of representation does not display the constraining variables perfectly. In this case there are two levels of approximation of the data matrix, as depicted in Exhibit 12.3. First the data matrix is split into two parts: the part which is linearly related to the constraining variables, and the part that is not (i.e., $\mathbf{S} = \mathbf{S}^* + \mathbf{S}^\perp$ in the formulation above). Then dimension reduction takes place just as before, but in the constrained space (i.e., the principal

Exhibit 12.3:
The full space decomposition into the constrained space (brown) and unconstrained space (white). Within each space there is a part of the variance (or inertia) that is explained in the respective low-dimensional displays (area with green shading)



axes of $\mathbf{S}^*$ are identified), with constraining variables being displayed in the usual regression biplot style. Dimension reduction can similarly be performed in the unconstrained space by identifying the principal axes of $\mathbf{S}^{\perp}$. This decomposition scheme is illustrated in Exhibit 12.4 for the fish morphology analysis, where the first dimension is constrained by body weight. Since there is only one constraining variable, no dimension reduction is performed in the constrained space. Body weight is represented perfectly on the first dimension, and the second axis of the solution is the optimal first dimension of the unconstrained data space.

Triplots

In a constrained biplot there are three sets of points and the display is called a *triplot*. The third set of points added to the biplot consists of the constraining variables, and they are usually displayed in terms of their regression coefficients with respect to the dimensions of the biplot. Their directions will then be biplot axes onto which the sample points (usually rows) can be projected to give estimates of their values, as before. If the rows have been displayed in standard coordinates, then the constraining variables have directions equal to their correlation coefficients with the axes.

An application to the data set "benthos" illustrates the triplot when there are several explanatory variables. For each site the levels of six variables were measured: total hydrocarbon content (THC), total organic material (TOM), barium (Ba),

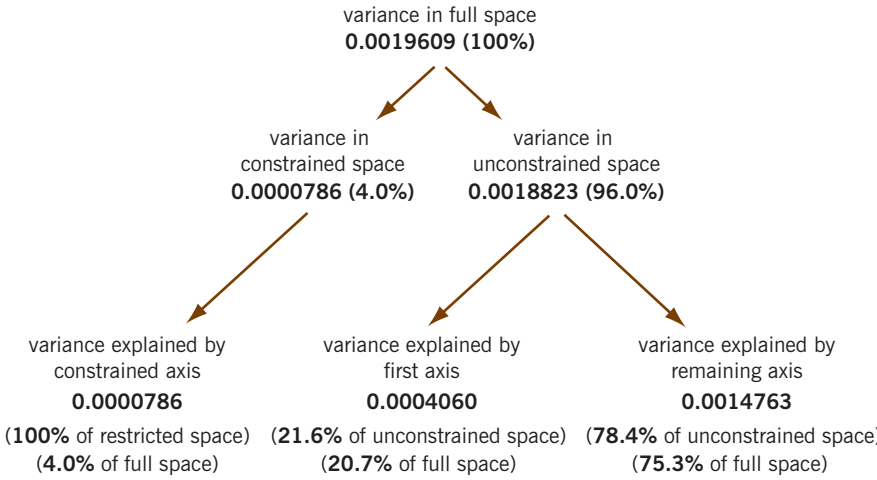


Exhibit 12.4:
The decomposition of variance (or inertia), first into the one-dimensional constrained space of body weight and the unconstrained space uncorrelated with body weight. The constrained space forms the first dimension of the biplot, which is only 4.0% of the total variance, and the first dimension of the unconstrained space forms the second dimension of the biplot, explaining 20.7% of the total variance

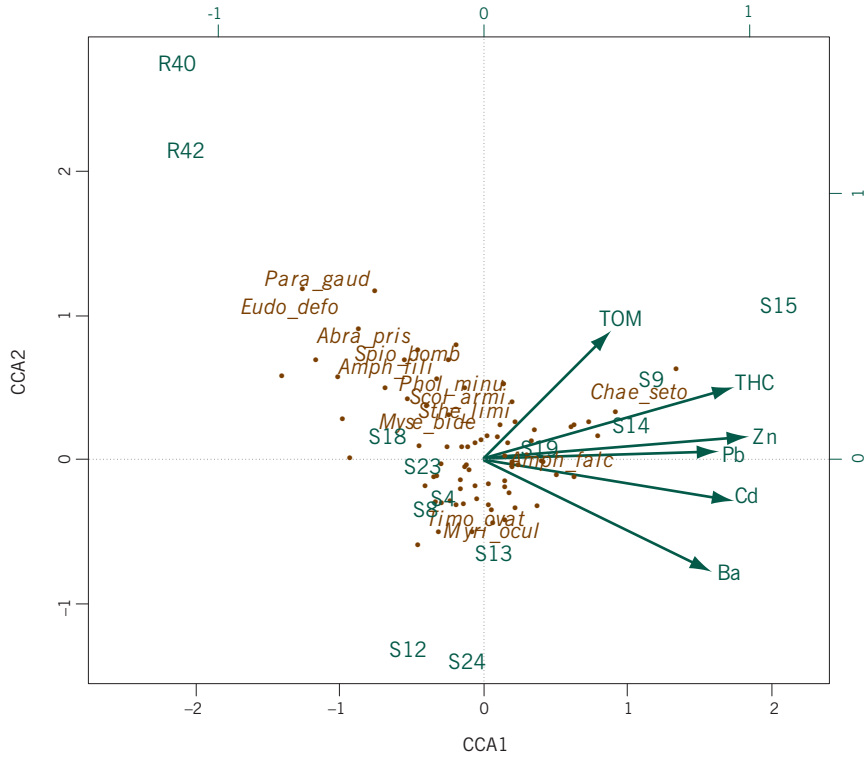
cadmium (Cd), lead (Pb) and zinc (Zn). It was preferable to log-transform these variables because of some very large values. Exhibit 12.5 shows the resultant triplot of the CCA restricted to the space of these six explanatory variables (in other words, dimension reduction has been performed from a six-dimensional space to a two-dimensional one). The sites in the triplot are in standard coordinates, and the species are at weighted averages of the sites. The explanatory variables are shown as vectors with coordinates equal to their regression coefficients on the axes (notice the different scale for these vectors). The reference stations are much more separated from the polluted stations now that the solution is constrained by variables that essentially measure pollution. Barium appears to be the variable that lines up the most with the separation of the reference stations from the others, pointing directly away from the unpolluted reference stations. The variable least associated with the unpolluted versus polluted contrast appears to be total organic material.

The corresponding decomposition of inertia is shown in Exhibit 12.6, showing that the six explanatory variables explain 65% of the total inertia in the data. Of this, 72.5% is explained by the two dimensions of the triplot. Because the explanatory variables are displayed using standardized regression coefficients, their lengths are related to how much of their variance is explained by the axes: the least is TOM (46% variance explained), and the most is Zn (96% variance explained).

Stepwise entry of the explanatory variables

As in multiple regression, the more explanatory variables that enter, the more variance is explained. When the number of explanatory variables equals the number of cases 100% of the variance would be explained and then there is effective-

Exhibit 12.5:
Triplot of the "benthos" data, showing the six constraining variables. Of the total inertia (0.7826) of the species abundance data, 65% is in the constrained space, of which 72.5% is displayed in the triplot



ly no constraint on the data and the analysis would be a regular biplot. To reduce the number of explanatory variables in such an analysis, a stepwise entry of explanatory variables is often performed, which ensures that only variables that explain a significant part of the variance are entered. At each step the variable that explains the most additional variance is entered and this additional variance is tested using a permutation test. The process continues until no variables entering produce a significant increase in explained variance. This procedure is illustrated in the case study of Chapter 15.

SUMMARY: Constrained Biplots and Triplots

1. Biplots, whether they are based on PCA, CA or LRA, display the data in a reduced dimensional space, usually a plane, with the objective of approximating the original data as closely as possible.
2. Often the data matrix can be regarded as responses to be explained by some explanatory variables that are available. The original biplot dimensions are not necessarily related to these explanatory variables, but an alternative approach constrains the principal axes of the biplot to be specifically related to these variables.

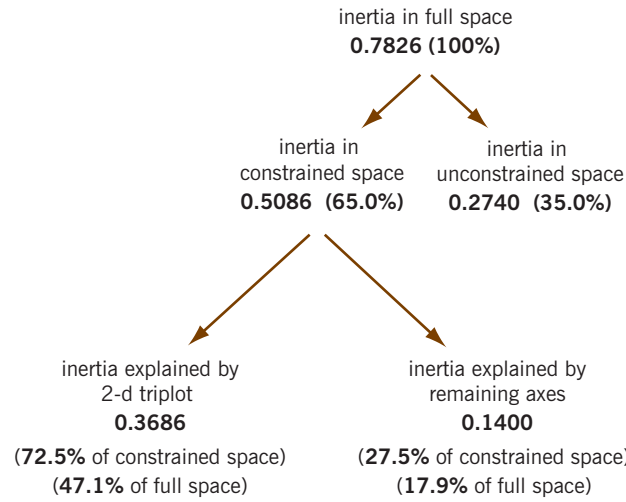


Exhibit 12.6:
The decomposition of inertia, first into the six-dimensional constrained space of the explanatory environmental variables and the unconstrained residual space that is uncorrelated with the explanatory variables. In the constrained space the first two dimensions explain 72.5% of the constrained inertia, which is 47.1% of the original grand total

3. The constraint is usually a linear one: the data are projected first into the constrained space which is linearly correlated with the explanatory variables, and then dimension reduction takes place as before.
4. The result of such an analysis with constraints is a triplot, showing the rows and columns of the original data matrix of interest, plus vectors indicating directions for the explanatory variables.
5. The dimensions of the residual, or unconstrained space, may also be of interest. In this space variance or inertia is explained in biplots that are uncorrelated with the explanatory variables.
6. The initial total variance or inertia of the data matrix is decomposed first into a constrained part (linearly related to the explanatory variables) and a residual unconstrained part (uncorrelated with the explanatory variables). Biplots can also be constructed for the unconstrained part of the data.
7. Explanatory variables are often entered stepwise, where the entering variable is the one that explains the most additional variance in the data, and this added variance can be tested for statistical significance.
8. For a single categorical variable as an explanatory variable, where the categories are coded as dummy variables, the constrained analysis is equivalent to a discriminant analysis between the categories.