A brief tour through the operation of PRIMER v7

A brief tour through the operation of PRIMER v7

- Opening the examples
- Reading data in from Excel
- Basic MVA wizard
- Pre-treatment of data
- Matrix display wizard
- Environmental data
- Resemblance calculation
- ANOSIM tests
- CLUSTER analyses
- MDS & PCA ordinations
- Species analyses
- Other analyses

Opening the examples

Opening the examples

After launching the PRIMER desktop by clicking on its icon, the first step is to open a worksheet of multivariate data, e.g. species abundances over a number of samples. The user's own data will typically be read into the program from Excel (*.xls or *.xlsx), though various text format input options are also provided (or you can type entries into a newly created PRIMER worksheet and edit it directly – though this is not commonly done). However, the PRIMER 7 installation comes with a number of real data sets, in a folder \Examples v7, needed as examples for this manual. You can access this with **Get Examples V7** on the **Help** menu, which prompts you for a directory in which to locate the \Examples v7 folder. It is assumed (for brevity) throughout this manual that this is simply the top level C:\ directory. So, C:\Examples v7 contains sub-directories for each study, in which the data files have usually been saved in PRIMER 7's internal binary format (*.pri, which is unreadable by other software or earlier versions of PRIMER). To open such a species $\times$ samples matrix, e.g. of nematode species abundances in marine sediments from 27 sites over five creeks of the Fal estuary, SW England (whose sediments are contaminated by heavy metals from historic mining), take **File>Open** from the main menu, navigate to the \Examples v7\Fal benthic fauna directory, select the file Fal nematode abundance.pri, and click **Open** to display the species data matrix in the desktop. Taking **Edit>Properties** you will see that PRIMER-format *.pri files carry other information on Title, Data type, Array size, whether Samples are found in •Columns or •Rows, and a Description. With **Edit>Factors** a subsidiary sheet of three factors is also seen to be linked to this worksheet: *Creek*, a creek abbreviation, the full *Creek name* and a numeric *Position* factor of the sampling sites' location down the creek – other factors could be typed in with **Add**.

Reading data in from Excel

Reading data in from Excel

As an example of reading in data from Excel, first open and examine the file Fal environment.xls, to note the simple format of a title in box A1, column headings (unique) for the samples in row 2, row headings (also unique) for the variables in column A, with only numeric entries in the array itself (rows 3 to 14). This is followed by a blank row, followed by the same three factors as above. This format must be adhered to precisely, with no extra blank rows or columns, or extra headers. After **File>Open**, you need to find the drop-down list (bottom right of the Open dialog) and select Excel Files, which should display the Fal environment.xls sheet. Select it, and **Open** now takes you through a File Wizard, for which you take the defaults – but look what they are – other than specifying (Data type•Environmental). If this worksheet were now to be saved in the default format (**File>Save Data As**), the result would be the Fal environment.pri file already in the workspace.

Basic MVA wizard

Basic MVA wizard

To cater for users completely unfamiliar with the basic outputs from a multivariate analysis, e.g. of the species abundance matrix opened above, PRIMER 7 now has a **Wizards>Basic multivariate analysis** menu item, which automatically generates robust outputs from some core routines, using knowledge of the Data type and with the opportunity for the user to alter some inputs from their defaults. Run this routine with the Fal nematode abundance sheet as the active matrix (click on it to make it active – its header bar will then be a slightly darker colour than other open worksheets). Take all the defaults on the Basic analysis wizard dialog box, i.e. just click on **Finish** – having first looked closely at the choices it has made for you! – and several results and graphic windows will appear in the display area of the PRIMER desktop (to the right). The Explorer tree area, to the left, shows the sequence of Data worksheets and Graph outputs created by the Wizard, interspersed with Results windows (the notebook icon) with names which describe the routine that has been run, and the tree shows the relationships among these analyses (what they start from and what they produce) A Wizard is just a bundled version of single routines which appear on PRIMER's other menus or sub-menus, so click on each row of the Explorer tree to display the sequence of steps involved and outputs produced. The final graphical step is a *Multiplot* output of four graphs, 'rolled up' in the tree, shown by the + sign. Clicking on the + (or on any of the plots in the multiplot) unrolls these names. It is now instructive to run through the individual analyses that this Wizard corresponds to.

Pre-treatment of data

Pre-treatment of the data (sometimes in more than one way) is usually desirable. For assemblage data, transformations will reduce the dominant contribution of abundant species to Bray-Curtis similarities. Though not usually needed for controlled (‘quantitative’) sampling, standardising of samples to relative composition (so sample totals are all 100%) can be achieved, with Fal nematode abundance active, by **Pre-treatment>Standardise**>(Standardise•Samples) & (By•Total) – the Wizard default was not to standardise but it did give that option, where % composition is desired.

Transformation of all values (which should be after standardisation, if the latter is appropriate) is obtained by, for example, **Pre-treatment>Transform (overall)**>(Transformation: Square root). A more severe transform would have been by Fourth root or Log(X+1) or by the ultimate in severity of transformation – reduction of the quantitative data to purely Presence/absence of each species. Since the purpose of transforming is to avoid the ensuing analysis becoming dominated by just one or two species with very large abundances, and bring more species into the definition of similarity of two assemblages – whilst at the same time avoiding giving sporadic, singleton species too much weight – the effects of competing choices can be assessed by running the second **Wizards** item.

Matrix display wizard

On active sheet Fal nematode abundance, run **Wizards>Matrix display**, not taking all the defaults in this case but unticking/unchecking the (Reduce species set) box so that all species are retained, and taking (Transformation: Square root) & (✓Retain sample groups>By Factor: Creek). A quite complex set of steps are then carried out, culminating in a run of **Shade Plot** from the **Plots** menu (fully described in Section 10) but, for our current purposes, all that needs to be understood is that the resulting shade plot is simply an image of the data matrix, in which the abundance for each species is represented by the shade of grey, from white (absent) to black (the largest count in the worksheet). Replicates from the 5 creeks are kept together along the *x* axis and the species on the *y* axis have been clustered and ordered in such a way that species with similar distribution across these samples are placed together in the re-ordering. (Multivariate analysis does not use the order of species in the matrix but it helps the human eye to visualise data structures by performing such re-arrangements). Apart from it being clear that some creeks contain a rather different set of species – or at least different abundances of the same species – an observation which is formally tested by the **ANOSIM** routine, e.g. as part of the **Basic multivariate analysis** wizard, the other message is that no one species will dominate an assessment of similarity of samples (columns) to each other. Equally clearly, quite a number of the less frequently occurring species have (transformed) values which are still sufficiently small in relation to the main players that they are almost invisible to the ensuing similarity calculation. This is probably desirable, and suggests we may have a reasonable transformation here. Contrast this with **Wizards>Matrix display** run again on the Fal nematode abundance sheet, but this time with (Transformation: None) – you can ignore the warning that PRIMER gives you (it is trying to tell you this is a bad idea!) – and it is clear from the resulting shade plot that only a few species will now contribute to the similarity computations. So an assessment of biotic differences among creeks, and how this relates to differences in heavy metal levels will really only be about a few numerically dominant species and not broadly community-based. At the other extreme, if you try the severest pres/abs transform, the rare species are now having far too much of an effect and will dilute genuine patterns from species sampled in reasonable numbers.

Section 4 also discusses an alternative approach to balancing contributions from different species, that of **Pre-treatment>Dispersion Weighting**, which downweights species with highly variable counts in replicates, which the sampling device captures in clumps rather than single individuals –relatively more weight is therefore given to species with consistent numbers over replicates of the same condition and these will be more reliable for assessment. If you try that pre-treatment and put the resulting rebalanced matrix into the **Matrix display** wizard, the shade plot gives a matrix image not unlike that for the square root transform, and this is certainly a possible pre-treatment here.

Environmental data

For environmental-type data, such as the Fal environment sheet, it is often appropriate to transform individual variables selectively, rather than all in the same way, since they may be of very disparate types. Here, the main objective is to avoid strong skewness in the distribution over samples, since large outliers will dominate both computation of (normalised) Euclidean distances and the Principal Component Analysis (**Analyse>PCA**), which is often the multivariate analyses chosen for abiotic data. The degree of skewness, or presence of outliers, is visually assessed using **Plots>Histogram Plot** or **Plots>Draftsman Plot** on active sheet Fal environment (you may wish to increase symbol size on the draftsman plot – do this by **Graph>Sample Labels & Symbols** and Size: 150, say). If there is strong right-skewness, those variables might need a log transform by highlighting them and taking **Pre-treatment>Transform (individual)**>(Expression: log(V+1)), Section 4. Alternatively take the rank transform, **Tools>Rank Variables**, which certainly gets rid of outliers! Although there is skewness here, there are no strong outliers and, for this demo, omit any transformation. So, run **Wizards>Basic multivariate analysis** on Fal environment and take all the defaults, examining the different choices made for this environmental-type matrix (e.g. normalising variables onto a common dimensionless scale; Euclidean distance resemblance; PCA ordination, see Section 12).

Resemblance calculation

Resemblance is the general term in PRIMER used to cover (dis)similarity or distance coefficients. The next stage in both the Fal nematode and environment runs of the **Basic multivariate analysis** wizard was to create an appropriate triangular resemblance matrix between all pairs of samples. This is a run of **Analyse>Resemblance** on the pre-treated (transformed or normalised) worksheet. Relevant defaults will be suggested, given the Data type, i.e. (Measure•Bray-Curtis) for biota and (Measure•Euclidean distance) for environmental variables, and (Analyse between•Samples) in both cases. There are, however, nearly 50 other possible choices on this dialog, see Section 5.

ANOSIM tests

The wizard then runs, for both biotic and abiotic data, **Analyse>ANOSIM**>(Model: One-way - A) & (Factors A: Creek)>(Type Unordered) on the respective resemblance matrices as active sheets. This tests for statistically significant differences overall among the 5 creeks in terms of their biota (or environmental data), and follows it up with pairwise tests between pairs of creeks, using the 5 (or in one case 7) locations in each creek as the replicate level. The Results window (eg *ANOSIM1*) shows the ANOSIM R statistic is large (0.82 for biota, 0.71 for environmental variables), close to its maximum value of 1, implying very good clear separation of the creeks, and highly significantly different from the null hypothesis R = 0, of no creek differences – the same is true of the pairwise tests. The associated plot (*Graph1*) is of the null hypothesis values of R under random permutations and shows that values not much more than R = 0.2 would be expected here if creeks did not differ.

ANOSIM tests can be much more extensive. PRIMER 7 introduces the idea of ordered ANOSIM tests, in which a numerical factor can be defined for the groups *a priori* (perhaps testing for simple time trend, or spatial gradient of change). Two-way crossed or nested, and three-way crossed, nested, or mixed crossed and nested, designs can be defined, with any factor ordered or unordered and analyses are then often possible without replicates as well as with them – see Section 9.

CLUSTER analyses

The Basic MVA wizards then run a cluster analysis, again on the respective resemblance matrices. This component routine is **Analyse>Cluster>CLUSTER**>(Cluster mode•Group average), without taking the (✓SIMPROF test) option since the latter is the appropriate test (rather than ANOSIM) when an *a priori* group structure is not defined. That is, if we had chosen to ignore the structure of sites within 5 creeks and simply treated the 27 samples as just 27 Fal estuary locations, the primary thrust of the analysis would not have been the ANOSIM tests and MDS display (see below) of those creek groups. Instead, it would have been a more exploratory analysis of whether the sites fell into clusters of similar communities (or environmental variables) at all – and, if so, which sites constituted those groups. The SIMPROF test is then important in deciding which sub-clusters in the hierarchical group-average cluster analysis (UPGMA) we are entitled to interpret as distinguishable groups, statistically – and, if we did not tick the ✓ANOSIM (1-way) box in the **Basic multivariate analysis** wizard, it would instead run a series of SIMPROF tests on the nodes of the cluster analysis dendrogram (Section 6) to determine this. As it is, the clustering in this Fal example is secondary and *Graph2* simply displays the dendrogram of the 27 sites, without SIMPROF tests. However, it is interesting to note that the dendrogram does largely divide the 27 samples into the 5 creeks, with an exception or two, which is consistent with the clear distinction among creeks seen in ANOSIM. You might like to accentuate this point by **Graph>Sample Labels & Symbols**>(Symbols✓Plot)> (✓By factor Creek) and look also at **Graph>Special** options, e.g. re-orienting the dendrogram.

PRIMER has other clustering tools (Section 6): a hierarchical binary divisive cluster analysis in unconstrained, **Analyse>Cluster>UNCTREE**, or constrained form, **>LINKTREE** (in which only divisions which have an 'explanation' in terms of a threshold on an environmental variable, say, are permitted). Both these methods share a common structure, consistent with the non-parametric treatment of resemblance matrices (which applies to tests such as ANOSIM, RELATE, BEST and ordinations such as non-metric MDS etc), namely each group is successively sub-divided so as to maximise the ANOSIM R statistic (PRIMER's key measure of group separation in multivariate space) between the two groups formed. A further non-hierarchical clustering method is available in the **Analyse>Cluster>kRCLUSTER** routine, a generalisation of classical *k-means* clustering to any resemblance matrix but again using only ranks. SIMPROF tests can be applied to all methods.

MDS & PCA ordinations

The Basic MVA wizard next produces non-metric MDS (*n*MDS) plots in 2-d and 3-d, together with their associated Shepard diagrams, which show how well (or badly) these distances among samples in the low-d ordination plots approximate the high-d resemblances. If the stress (Section 8) is not too large (it is only 0.10 here), *n*MDS plots give a powerful representation of the sample patterns.

The wizard is here running (again on the resemblance matrix) **Analyse>MDS>Non-metric MDS (nMDS)** under default conditions, but taking this directly, there are options to choose higher-d solutions and, more entertainingly, to watch the iterative process of trying to obtain the lowest stress 2-d solution (say) from different restarts of sample points thrown randomly into 2-d, which you can activate – and even record as an *.mp4 file – by (✓Animate) on the dialog (Section 8). There are other recordable animations possible also, of spinning 3-d ordination plots and showing a dynamic trajectory of, for example, a time series of samples on an MDS or other ordination plot.

Where the data matrix is environmental and (usually) variables normalised, there is a choice of ordination by PCA (**Analyse>PCA** run on the normalised data sheet) or *n*MDS on the Euclidean distance resemblances. These are both offered by the wizard but, running directly, a third option is **Analyse>MDS>Metric MDS (mMDS)**, which fits a straight line to the Shepard diagram of MDS (low-d) distances *vs* original (high-d) distances, and in one respect improves on PCA here, giving a more faithful preservation of the high-d distances by avoiding the PCA projection into low-d.

Species analyses

The final step in the Basic MVA wizard is to break down the dissimilarities (or distances) between pairs of creeks into their contributions from each of the species (or abiotic variables), in the tables of *SIMPER1* (or *SIMPER2*), see at the end of Section 10. This is equivalent to running **Analyse> SIMPER** on the transformed data matrix for biota (or normalised data matrix for abiotic variables). There are, however, several other ways in which PRIMER examines variable relationships to each other, or species relationships to the sample patterns (Section 10). We have already seen the power shade plots potentially have for interpretation. Another possibility is Bubble plots of individual species values on the sample *n*MDS ordination: the larger the bubble the greater the abundance of that species at that site – or abundances, because PRIMER does multiple (segmented) bubbles of different colours and circle sectors for different species. For the Fal *n*MDS, try this with **Graph> Special>**(✓Bubble plot) & (Worksheet: Fal nematode abundance) & (Variables>**Change**), moving *Metachromadora vivipara*, *Tripyloides gracilis* and *Leptolaimus limicolus* to the Include box and all other species in the Available box, and ticking (✓3D effect) & (Saturation: 75).

Calculating similarities (index of association) among species – not samples – or correlations among environmental variables, in their pattern of response across the samples, opens up another field of analyses, which we have already seen used to cluster species in the shade plot. Adapting SIMPROF tests to operate on variable clusters (Type 3) rather than sample clusters (Type 1) permits definition of *coherent variable sets*, which within the sets are not statistically distinguishable but across sets have significantly different response patterns over the samples. Run **Wizards>Coherence plots** on Fal environment, the heavy metal levels (and silt/clay ratio) at these 27 sites, with significance set at 0.5%, and strikingly similar metal concentration profiles are seen in the resulting **Line Plot sets**.

Other analyses

A further bubble plot you might like to try on the Fal *n*MDS is to superimpose abiotic variables from the Fal environment worksheet, and we have already referred to the constrained LINKTREE clustering that tries to explain community groupings in terms of particular environmental variables, but PRIMER also has another generic way of looking at the relation of community structure to potentially explanatory variables – in combination, rather than individually, see **BEST** in Section 13. There is an overall hypothesis test for the significance of such a link, and the mechanism of non-parametric matrix correlations (which also includes PRIMER's **RELATE** tests) can be applied to other contexts in which multivariate data sets are compared (Section 14).

PRIMER calculates a range of univariate diversity-related indices through the **DIVERSE** menu including ones based on taxonomic or genetic/functional relatedness of the taxa (**TAXDTEST**), see Section 15, and a range of diversity curves (e.g. dominance plots, species accumulation, Section 16).

Other main menus (e.g. **Select**, **Edit**, **Tools**, **Plots**) offer a wide variety of data manipulations and standard plotting functions (**Histogram**, **Means**, **Box**, **Bar**, **Surface**, **Line** and **Scatter Plot**).

The final Section (17) deals with region estimates for means in multivariate studies, e.g. average communities for each of the Fal creeks, plotted on a 2- or 3-d MDS together with an approximate measure of the uncertainty about these means, from bootstrapping. You might like to finish this brief excursion through PRIMER by running **Analyse>Bootstrap Averages** on the Bray-Curtis resemblance matrix from the Fal biota, taking all the defaults, to get the multivariate *means plot*.

And you can save all this in a PRIMER workspace file *.pwk with **File>Save Workspace As**.