The PCO-MC Short Bus Manual

Warning

Some European users may need to set their system "localization" to present decimals like 1,234.56, instead of 1.234,56, to ensure compatibility with SAS. This can be done in the 'Formats' tab of the 'International' or 'Languages & Text' setting within System Preferences. Software tested using 'Region' = United States.

Introduction

The PCO-MC Short Bus is a Mac OS X utility for performing a PCO-MC (principal coordinate-modal clustering) analysis of genetic polymorphism data following Reeves and Richards (2009). Performing an analysis using *The PCO-MC Short Bus* requires a series of steps that, for the time being, necessitate the use of two commercial software packages (*NTSYS* and *SAS*) and two operating systems (Windows and Mac OS). Source code (in the form of an *Xcode* project) is available upon request for those wishing to modify or streamline the procedure.

Performing an analysis

There are four basic steps required to perform PCO-MC. Two example data sets are provided with the software. It is highly recommended that users work through the example data sets before trying their own. One data set contains 159 AFLP genotypes from native hops individuals. This data set is strongly structured. A second, essentially unstructured, data set containing 500 wild apple individuals genotyped at 7 microsatellite loci is provided for comparison.

1) Generate Principal Coordinates. Principal coordinates are generated using the batch processing capability of *NTSYS* (testing was performed using *NTSYS 2.11x*). Two files must be created, a ".nts" file (containing the genotypes), and a ".ntb" file (containing the batch commands). As always (and especially when multiple operating systems are involved) users should be mindful of the type of line breaks used in input files. All input files must use DOS line breaks. Also, the format of the *NTSYS* input files should follow the example files closely.

Some explanation of *NTSYS* file formats is in order. In an ".nts" file, the first line contains 5 variables that describe the matrix. The second number specifies the number of samples (individuals); the third number indicates the number of characters. For dominant data the third number is the same as the number of loci; for codominant data, it is 2 times the number of loci. For codominant data, a second header line that describes the relationship between locus names and scores is included when the "L" option is invoked after the third number. In a ".ntb" file, commands are preceded by "*". Creating a ".ntb" file for a new data set can be as simple as replacing the word "Hops" or "Kaz" in the example with a name reflecting the new data set, and setting the "n=159" or "n=500" command in the "*eigen" line to the number of samples in your data set. This command tells *NTSYS* how many principal coordinates to calculate.

Start *NTSYS*. Click the button with the image of a person running to select a batch file to run. Browse to the ".ntb" file and open it. The ".nts" file must be in the same folder as the ".ntb" file. In the batch mode dialog, click the “Run” button. *NTSYS* can’t handle long path names, so if you get an error saying “The system cannot find the file specified” chances are that moving your folder closer to the root will solve it. *NTSYS* will write three output files to the folder where you found the “.ntb” file. The one with the prefix “PROJ” contains the principal coordinates. You will need this file later.

Before quitting *NTSYS*, you will need to acquire information for performing the weighting step in PCO-MC. To do this, first cancel the batch mode dialog. Click File > View listing notebook. Save the information in the “eigen” tab of the Report listing by clicking File > Save > Current section... The file name used is not important, but you might call it something like “Weights.txt”.

So, you should come away from *NTSYS* with two files: the “PROJ” file and the “Weights.txt” file.

2) Produce a *SAS* input file. For this step you will need to run *The PCOMC Short Bus* from your Mac. You can either transfer the “PROJ” and “Weights.txt” files to your Mac, or you can leave them where they are and access them over the network (assuming you’ve mounted the relevant PC volume on your Mac).

Start *The PCOMC Short Bus*. It is a simple application with just two functions: “Pre-Process” (converting the *NTSYS* files into a *SAS* input file), and “Post-Process” (analyzing the *SAS* output). By clicking on the triangles, various options can be set by the user. For now, use the default settings. With the “Pre-Process” window activated, press the “Make SAS file from NTSYS output” button. (Note that the “Make SAS file from distance matrix” button is grayed-out and not available. Eventually, we hope to integrate code for calculating principal coordinates directly into *The PCO-MC Short Bus* so that the *NTSYS* step can be eliminated.) You will receive a prompt to choose a file. The first file requested is the “PROJ” file. Select it. Another prompt follows, requesting the “Weights.txt” file. After selecting the second file, the status drawer will open, indicating what is going on. There is some matrix multiplication involved here, so it may take awhile to produce the *SAS* input file if there are a large number of individuals in the data set. The *SAS* input file, with the suffix “.SAS.txt”, will be saved in the same directory as *The PCO-MC Short Bus* application.

3) Run the *SAS* analysis. Move back to the PC, taking the *SAS* input file with you or making it available over the network. Start *SAS* (testing was performed using *SAS 9.1* on Windows XP Pro). *SAS* is sort of inscrutable in terms of the many possible ways it can be configured and run. Here is how I deal with it: Drag and drop the “.SAS.txt” input file on the *SAS* output window to execute it. The title of the output window should give you some hints on the status of the analysis, e.g. “DATA STEP running” or “PROC MODECLUS running”, and after a short period two sample plots should appear in a GRAPH window. These are provided for quick inspection of clustering along the first 2 or 3 principal coordinate axes. You can save these if you want, but they are not necessary for proceeding with the PCO-MC analysis. The file you want will have the suffix “.out.txt” and *SAS* may fling it in any number of locations depending on your configuration. For me, the output file is saved in the folder C:\Documents and Settings\username where “username” refers to the login name of the current user. If it isn’t there, perform a file search using the string “.out.txt” to find it.

4) Analyze the *SAS* output file. Return to the Mac, bringing along the *SAS* output file. Start *The PCO-MC Short Bus*, and select the “Post-Process” window. Press the “Summarize SAS output” button, and select the *SAS* output file. The status drawer will open, and when the procedure reaches the phase called “Checking scan efficiency...”, a dialog will pop up, telling whether the search was incomplete, or, if complete, the percent of R-values examined that were informative. A complete scan includes results from cluster analyses where Rmin was small enough to produce the maximum possible number of clusters (the number of individuals, or 100, whichever is smaller), and Rmax was large enough to result in a single cluster containing all individuals. In general, the percent of R-values that were informative should exceed 25-30%. If your result is lower, or if you want to improve the precision of the analysis, click "Yes" when the software asks "Would you like to try a better range of R-values?" Clicking "Yes" will return you to the "Pre-Process" window and will automatically enter values for Rmin and Rmax in the Options drawer that have been optimized based on the results from the initial *SAS* analysis. You then need to repeat the analysis starting at step 2 above. Usually one round of optimization will result in a scan efficiency >90%, which is more than adequate.

Once you are happy with the scan efficiency, click "No" when the software asks "Would you like to try a better range of R-values?" Some dialogs will appear describing whether there were any limitations with the analysis. Click "Yes" rather than "Cancel" to proceed through these notifications.

The PCO-MC Short Bus produces a maximum of 5 output files (only 3 are produced when there is no structure). The following 3 files will be useful for most users:

1. "Unique Clusters" contains a list of the clusters that were found to be statistically significant (when the p-value associated with the cluster was less than the p-value cutoff). Clusters are presented in ranked order according to their stability value (the proportion of informative R-space in which the cluster was found to be statistically significant). All significant clusters are shown in this file, even if their stability value was lower than the stability cutoff. Note that the p-value cutoff will often be set to a very high value (default = 0.9999). This is by design so that decisions about which clusters are meaningful can be made based upon the stability cutoff rather than the p-value. When no population structure is found, this file will not contain any clusters, just a sentence saying "There are 0 distinct, statistically significant, clusters", and the two files described below will not be created.
2. "Assignment" contains a table describing the assignment of individuals to clusters, after the stability cutoff is applied. The first row contains stability values. The first column contains the sample names. The other columns define cluster membership of the individuals using numbers (assignment to, e.g., cluster #1) and ?'s (not a member of the cluster defined in that column). Note that PCO-MC does not produce a partition, therefore overlapping cluster assignments and unassigned individuals (?'s in all columns) may occur.
3. "Supertree" contains a NEXUS file that can be executed in *PAUP*4.0* (Swofford, 1999) to automatically produce an MRP supertree to graphically represent the hierarchical assignment produced by PCO-MC analysis.

Two additional files, "Cluster summary" and "Pvalue summary", can be used to produce plots like Figure 2a-d in Reeves and Richards (2009). A batch processing option is available in the "Post-Process" window that automatically produces summary files for all *SAS* output files in a folder. This function is useful for evaluating simulation results or for sensitivity analyses. When starting a batch analysis, the target folder must contain only *SAS* output files. Prompts are suppressed during batch processing. A summary file called "BatchLog" is produced instead.

Contact

For help in applying PCO-MC to your data set, for questions about using *The PCO-MC Short Bus*, for technical information about the procedure, or to report bugs, contact:

Pat Reeves, reevesp@lamar.colostate.edu

Citations

Reeves, P.A., and C. M. Richards. 2009. Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. PLoS ONE 4: e4269.

Swofford, D. L. 1999. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.