PCA Randomization Tests Program

Steve Tonsor
Department of Biological Sciences
University of Pittsburgh
Pittsburgh PA 15260
Voice: 412-624-5491
Fax: 412-624-5863
tonsor@pitt.edu

Contact Steve Tonsor with any questions regarding the rather cumbersome output. **Please acknowledge me if you use this or any derivation of it.

This program tests for significant relationships among variables using randomization. It tests for significant correlations/covariances and significant eigenvalues and eigenvector coefficients from Principal Components Analyses. It does this by shuffling the actual trait values Among the observations in a data set. For a data set with n observations and m traits, m data sets are created, each containing one of the variables. For each of the m data sets, the order of the observations is randomly permuted. The data sets are then re-merged, but in the order of their independent random permutations. This has the effect of randomly associating one of the actual measured values of trait 'a' with a randomly drawn actually measured value of trait 'b' and assigning them to an observation. The covariance matrix and principle component eigenvectors and eigenvalues are produced by PROC PRINCOMP. This is repeated NITERAT (value set by user) times, with the values output to a data set. The values are used with proc univariate to calculate randomization-based 99% confidence intervals on the null hypothesis distributions of each of the elements of the "output out=" data set from proc princomp. The confidence limits are comparedto the actual values, and a table is produced in the SAS output window, reporting significance or lack thereof.

THE OUTPUT TABLE CONTAINS THE ACTUAL VARIABLE VALUE FLANKED BY THE LO AND HI 99% CONFIDENCE LIMITS, FOLLOWED BY ** WHEN SIGNIFICANT.
CAUTION SHOULD BE EXERCISED IN INTERPRETING THE NULL HYPOTHESIS CONFIDENCE INTERVALS FOR THE EIGENVECTOR COEFFICIENTS WHEN THE MAGNITUDE OF THE EIGENVALUE IS RELATIVELY SMALL, EVEN IF SIGNIFICANT. WHEN EIGENVALUES FOR TWO PCs ARE CLOSE IN VALUE, THEY CAN SWITCH IN MAGNITUDE, HENCE RANK, BETWEEN RESAMPLE ITERATIONS. THE PC IS ALSO PRONE TO REFLECTION. THE CONFIDENCE INTERVALS CAN BECOME VERY LARGE AND PRONE TO TYPE II ERROR.

SEE Perez-Neto et al. 2003 Ecology 84(9):2347-2363 FOR A DISCUSSION OF THIS ISSUE.

Up to 9 treatments can be handled (these could of course also be populations), with separate hypothesis testing by treatment. Up to 99 traits can be handled, although it may choke at the output to disk step with a ton of traits- I have tested up to 14. The number of observations per treatment can be as large as the user needs it to be (you do have to tell the program the max number of observations per treatment/population).
The first module is a user-settings module. This is the only place in the program in which the user needs to make changes, indicating number of treatments, traits and iterations, as well as directories and file names for input and output. The input and output directories must be the same.

NOTE THAT THE INPUT FILE IS SPECIFIED AS COMMA-DELIMITED. YOU CAN EASILY CONVERT ASCII OR EXCEL FILES TO COMMA-DELIMITED FORMAT USING "SAVE AS" .CSV IN EXCEL.

This avoids problems SAS sometimes has with ordinary excel or txt files, in which SAS sometimes reads two variable values as one string.

I RECOMMEND RUNNING FIRST WITH NITERAT=5 TO DEBUG FILENAMES, ETC. IF YOU DO THIS, DELETE THE OUTPUT FILES, OR CHANGE THE FILENAMES IN THE "SIDEWARD=" AND "FORWARD=" SPECIFICATIONS OF THE USER SETTINGS MODULE, before running a full set of iterations. JUST ABOVE THE USER SETTINGS MODULE, THE PROGRAM SHUTS OFF THE PRINTING OF NOTES AND SOURCE CODE TO THE LOG. IF YOU ARE HAVING TROUBLE GETTING THIS PROGRAM TO RUN, RESTART SAS AND PUT COMMENTS AROUND THIS LINE OF CODE. NOTE ALSO THAT THE LOG FILE IS REDIRECTED TO A DISC FILE TO PREVENT BUFFER OVERFLOWS. IF YOU WANT TO EXAMINE IT, LOOK AT THE LAST LINE OF THE USER SETTINGS MODULE FOR THE NAME OF THE LOG DISK FILE.