VarCan (version 1): Variation Estimation and Partitioning in Canonical Analysis

Pedro R. Peres-Neto
Department of Biology
University of Regina
Regina, SK S4S 0A2, Canada
Phone: (306) 585-4850
Fax: (306) 337-2410
E-mail: Pedro.Peres-Neto@uregina.ca

March 2005

This program is freely distributed and is a windows based program. Please email with questions, comments and suggestions. If you publish a study using the approach, please let me know the reference as I wish to maintain a compilation of studies using this program. Any reprints would be appreciated!

Introduction to the problem:

The search for causes dictating patterns in species distributions in natural and disturbed landscapes is of primary importance in ecological science. Establishing relationships between species distributions and environmental characteristics is a major approach in the search for forces driving species. Habitat models relating habitat characteristics and community structure (species occurrence or abundance) are expected to answer at least two questions: (1) How well is the distribution of a set of species explained by the full set of predictive variables? and (2) which variables are irrelevant or redundant in the sense of failing to strengthen the explanation of patterns after certain other variables have been taken into account? The first question relates to the predictive power of the model that can be used in conservation management, for questions such as estimating habitat suitability, forecasting the effects of habitat change due to human interference, establishing potential locations for species re-introduction, or predicting how community structure may be affected by the invasion of exotic species. The second question is important for heuristic issues such as determining the likelihood of competing hypotheses to explain particular patterns in community structure.

Canonical analyses such as redundancy analysis (RDA, Rao 1964), canonical correspondence analysis (CCA, ter Braak 1986), and distance-based canonical analysis (db-CA, Legendre and Anderson 1999) are invaluable tools for modeling communities through environmental predictors. They provide the means of conducting direct explanatory analyses in which the association among species can be studied with respect to their common and unique relationships with the environmental variables or any other set of predictors of interest. RDA and CCA, which are asymmetric forms of canonical analysis, can be best understood as methods for extending multiple regression, which has a single response Y and multiple predictors X (e.g., several environmental predictors), to multiple regression (linear or not) involving multiple response variables (e.g., several species) and a common matrix of predictors X.

In multiple regression analysis, we can apply variation partitioning (also known as commonality analysis, Kerlinger and Pedhazur 1973) to identify common and unique contributions to model prediction and hence better address the question of the relative influences of the groups of independent variables considered in the model (Mood 1969). When partitioning variation in regression analysis, independent variables are grouped into sets representing broad factors. In that context, variation partitioning is more suitable than analyzing the individual contributions of regressors via their partial correlation coefficients. In this approach, the total percentage of variation explained by the model (R^2) is partitioned into unique and common contributions of the sets of predictors. As ecologists, we are often interested in broad sets of predictors that represent factors such as space, time, local environment and landscape heterogeneity; hence variation partitioning can assist us in addressing the heuristic concerns addressed earlier by estimating the likelihood (or power of explanation) of different hypotheses to explain particular patterns in species distributions (Legendre et al. 2005). Variation partitioning was extended to canonical analysis by Borcard et al. (1992, 1994), which made it possible to partition the explained variation of a species data matrix between two sources of variation, namely environmental and spatial. These two factors were of particular interest due to the fact that spatial processes may influence both species distributions and environmental factors in similar ways, generating apparent species-environment concordance (Legendre 1993). Thus, by applying variation partitioning, we can estimate the unique contribution of environmental variables to species distributions independently of spatial influences, and vice versa. Variation partitioning in canonical analysis has since been extended to three or more sets of predictor matrices (Anderson and Gribble 1998; Økland 2003) and is now routinely used in direct gradient analysis.

What does VarCan do and how does it work?

This program partitions a dependent matrix Y into components of variation related to up to 4 matrices of independent predictors in the case of RDA and up to 2 matrices in the case of CCA. Each data matrix should be entered as a space-delimited text or ascii data files. The rows represent observation (e.g., sampling locations) and the columns are the variables (e.g., species or environmental variables). The rows must be in the same order in all matrices and all files must have the same number of rows. Variables are each separated by a space. No column or row labels are included as the files simply include the data alone. You will be prompted for the data file names.

Y represented the matrix containing response variables and the matrices containing regressors are identified as X (i.e., X1, X2, X3 and X4). The program will calculate all relevant fractions (see Figure at the end of this document). Testing procedures follow Legendre and Legendre (1998, p. 608-612) and will be performed for unique fractions (i.e., the unique contribution of a matrix), for each data matrix separately (e.g., X1, X2, X3, and X4) and for all matrices assembled as one regressor matrix (e.g., X1X2X3X4). The latter represents a test of overall relationship between all predictor matrices and the dependent matrix. Tests are based on permutation procedures. You will be prompted to enter the number of permutations to perform. Permutation test for fractions are based on permuting raw data in the case of RDA and residuals in the case of CCA (reduced model, as described in Legendre and Legendre 1998, p. 608-612).

Results are generated in the form of tables that are saved in a text file. You will be prompted for a name for the result file. The tables will present results using the fraction organization as presented in the Figure at the end of this file. Note that fraction b for two data matrices is not the same as fraction b for three data sets. Therefore, it is important that you interpret results based on the Figure and according to the desired number of data tables.

The percentage of variation explained by a canonical analysis $\mathrm{R}^2_{\mathrm{Y}|X}$ is calculated in the same way as the coefficient of determination (R^2) in multiple regressions having one response variable (i.e., dependent). Remember that canonical analyses are extensions of multiple regressions where we have several response variables. The definition of $\mathrm{R}^2_{\mathrm{Y}|X}$ used here is the one used in ecological applications; it is called the RDA trace statistic in the Canoco 4.5 program (ter Braak and Smilauer 2002) and the proportion of explained variation in Legendre and Legendre (1998).

Recently, Peres-Neto et al. (2006) showed that as the R^2, $\mathrm{R}^2_{\mathrm{Y}|X}$ is biased (i.e., sample values tend to be larger than populations values). They proposed corrections for this bias for both RDA and

CCA and this program follows their implementations. The adjustment for CCA is based on a permutation procedure (see Peres-Neto et al. 2006) and you will be prompted to enter the number of permutations to use in the adjustment. This will be not the case of RDA since the adjustment if based on appropriate degrees of freedom.

VarCan can also test for the difference between 2 fractions of variation in RDA and CCA (see Peres-Neto et al. 2006). In this case, a bootstrap procedure is used and you will be prompted to enter the number of bootstrap samples to be used.

Installation:

First unzip varcanv1.zip in a directory (example: c:\varcanv1). There will be 4 files in the zipped file:

1 – varcanv1.exe
2 – varcanv1.ctf
3 - MCRInstaller.exe
4 – Guide varcanv1.pdf (this document)

Varcanv1.exe is a compiled code in Matlab. To run the executable you have to first install the Matlab Component Runtime. To do that, run the program MCRInstaller.exe. That will install the necessary libraries for the Matlab interpreter. The installation is simple to follow and all is necessary is to press "next" at each step. The installation may take some time (5 minutes or so), but once installed you can run any application compiled in Matlab. Therefore, if you already have this program installed from previous Matlab compiled applications, there is no need to perform this installation.

Once the component is installed, run varcanv1.exe. In the first time you run the application, a directory called varcanv1_mcr will be generated. This directory will be created within the directory you have chosen to unzip the files (c:\varcanv1\varcanv1_mcr). Keep (do not erase after installation) this directory to run varcanv1.exe in future applications. You are now setup to run varcanv1.

References:

Anderson, M.J., and N.A. Gribble. 1998. Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. Australian Journal of Ecology 23: 158-167.

Borcard, D., and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using Oribatid mites (Acari, Oribatei). Environmental and Ecological Statistics 1: 37-61.

Borcard, D., P. Legendre and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73: 1045-1055.

Ezekiel, M. 1930. Methods of Correlation Analysis. John Wiley, New York.

Kerlinger, F.N., and E.J. Pedhazur. 1973. Multiple Regression in Behavioral Research. Holt, Rinehart and Winston, Inc. New York.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74: 1659-1673.

Legendre, P., and M.J. Anderson. 1999. Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. Ecological Monographs 69: 1-24.

Legendre P., D. Borcard and P.R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecological Monographs (*in press*).

Legendre, P., and E. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129: 271-280.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. 2nd English ed. Elsevier Science BV, Amsterdam.

Mood, A.M. 1969. Macro-analysis of the American educational system. Operations Research 17:770-784.

Økland, R.H. 2003. Partitioning the variation in a plot-by-species data matrix that is related to n sets of explanatory variables. Journal of Vegetation Science 14: 693-700.

Peres-Neto, P.R., P. Legendre, P., S. Dray and D. Borcard. 2005. Variation partitioning of species data matrices: estimation and comparison of fractions. (*in review*)

Quinn, G.P., and Keough, M.J. 2002. Experimental Design and Data Analysis for biologists. Cambridge Unversity Press, Cambridge.

Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyaá, Ser. A 26:329-358.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167-1179.

ter Braak, C. J. F., and P. Smilauer. 2002. Canoco reference manual and CanoDraw for Windows user’s guide: software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York.

Zar, J.H. 1999. Biostatistical Analysis. 3rd Ed. Prentice Hall, London.