IMP:PLSMaker

Two-Block Partial Least Squares Analysis of Covariance

H.D. Sheets, 8/27/02, sheets@canisius.edu

This method is designed to display patterns of covariance between two sets (or blocks) of data, one or both of which may be geometric morphometric. The approach differs from a multivariate, multiple regression approach, which treats one set of variables as a function of the other set of variables. In the Two-Block Partial Least Squares (2B-PLS) approach, the two sets are treated symmetrically, without the assumption that one set of variables is the cause of the other, rather that the two blocks covary.

The 2B-PLS method is analogous to a principle components analysis, but applied to the between block covariance (or correlation) matrix, rather to a within block variance-covariance matrix as would be the case in PCA. The goal of PCA is to find a set of axes (the Principal Components) that express the greatest amount of variance within the group, the goal of 2B-PLS is to find pairs of axes, one for each block, that express the greatest pattern of covariance between the axes (called Singular Axes). Like PCA, the 2B-PLS is itself not a statistical test, both of these methods are decomposition methods that find axes that express patterns of variance (the PCA) or covariance (2B-PLS). The are several resampling tests that may be used to assess the statistical significance of the covariance pattern displayed by a 2B-PLS decomposition.

Operation of the 2B-PLS Method

Suppose we have two blocks of variables measured for the same specimens, call these block 1 and block 2. The 2B-PLS method can be used on any type of variable, we assume here that at least one block is a set of geometric morphometric data, specifically partial warp plus uniform component scores for each specimen, after a Procrustes Superimposition of the landmark configuration on a GLS Procrustes Reference. The second block may also be geometric morphometric data, representing the shape of a second set of landmarks, but it could also be traditional morphometric variables (lengths and widths), or life history variables, ecological variables, behaviorial variables, etc. Suppose we have n specimens in the study, k variables in the first block and j variables in the second block.

If we wanted to carry out the familiar Principle Components Analysis of block 1, we would form a variance-covariance matrix S, which would be a k x k matrix, whose diagonal elements would be the variances of the k variables in the first block, and who's off-diagonal elements are the covariances of the variables in the block. To carry out the PCA, we would then calculate the eigenvalue decomposition of the variance-covariance matrix S. This would yield k paired eigenvalues and eigenvectors, each eigenvalue is the variance expressed by the corresponding eigenvector (called the Principal Component Axes). Since the Principle Component axes are the eigenvectors of a matrix, they are orthogonal to one another, and hence the PCA axes are statistically independent of one another, having no covariance. The Principal Component axis expressing the greatest variance is called the first principal component, successively numbered axes express successively less variance. A PCA of block 1 would yield k PC axes, and k corresponding eigenvalues, yielding a useful way of looking at patterns of variance

within block 1, a similar study of variance within block 2 via PCA in a description of variance within block 2 expressed as j PC axes and j eigenvalues.

To study the pattern of covariation between blocks 1 and 2, we first form either the cross block covariance or correlation matrix. When the variables within the blocks are in the same units (as is the case in working with shape data in the form of partial warps scores), it is preferable to use the covariance matrix, as use of the covariance matrix will give the aspects of shape with the largest variance the greatest weight in the analysis. If the second block is not shape, but other variables, the second block should be standardized (to remove the effects of different scales of measurement) if all the variables are not in the same units. The correlation matrix is perhaps best used in the case where both blocks contain variables which are measured in differing units, which will not be the case when one of the blocks represents shape in the form of partial warp plus uniform component scores.

The covariance matrix formed will be a k x j matrix $S_{1\text{-}2}$ ($S_{1\text{-}2}$ indicating covariance between blocks 1 and 2) of covariances between the k variables of the first block and the j variables of the second block. The set of paired axes that express the greatest pattern of covariance between blocks is found using a Singular Value Decomposition (REFS!)

$$S_{1\text{-}2} = U\Sigma V^t$$

where Σ is the k x j singular value matrix, U is a k x k matrix whose columns are the singular axes for block 1 and V is a j x j matrix whose columns are the singular axes for block 2. The SVD method is similar in many ways to eigenanalysis methods (ref!! and

later material in this chapter), although the SVD operates on rectangular matrices. The rectangular singular value matrix Σ has r singular values (σ_i) along the diagonal, r is the rank of the variance-covariance matrix of the smaller of the two blocks, in most cases r is the smaller of j and k. If one block had a variable that was perfectly correlated with another variable in the same block, the rank of the variance-covariance matrix would be less than the number of variables within the block, although this is unlikely to occur in biological systems. The singular axes for block 1 corresponding to the first singular value is the first column of U, the paired singular axes for block 2 is along the first column of V. The singular values (the σ_i) are the square roots of the covariance expressed between the corresponding paired singular axes.

It is now possible to compute singular axes scores for each specimen along each singular axes, in the same manner one would compute PCA axes scores for each specimen. If X_1 is the n x j data matrix for block 1 and X_2 the n x k data matrix for block 2, where each row represents a specimen and each column a variable, then the singular axes scores for block 1 (Y_1) may be calculated as

$$Y_1 = X_1U$$

and the corresponding singular axes for block 2 are

$$Y_2 = X_2V$$

and the cross covariance matrix of Y_1 and Y_2 will be a j x k rectangular matrix who's only non-zero elements are the square singular values which appear along the diagonal. The singular axes are orthogonal, like the PCA axes, and form basis sets for the vector spaces of blocks 1 and 2, in the same manner that PCA axes do for a single block of data. The PCA scores are typically used to reduce the dimensionality of a system when

studying patterns of variance, the 2B-PLS can be used to out the same function when studying patterns of covariance.

When the 2B-PLS method has been applied to partial warp plus uniform component scores representing shape, it is possible to plot the singular axes as patterns of shape variation using all the typical methods of displaying shape changes.

Testing the significance of the Pattern of Covariance revealed by 2B-PLS

Prior to the interpretation of the patterns of covariance presented by the 2B-PLS method, it is necessary to determine if the patterns revealed differ from those expected by chance. At present there appears to be no analytic approach to testing the significance of the 2B-PLS, but it is possible to determine via a permutation or bootstrap test if the singular values produced are consistent with a null hypothesis of no meaningful pattern of covariance between the block.

To carry out a permutation test of the significance of an observed singular value, we repeatedly permute the association of measurements in block 1 with the measurements in block 2 by random reordering or shuffling of the rows in the data matrix. This generates data matrices with the same sample size and variance-covariance structure, but destroys any existing covariance between blocks. The SVD decomposition is carried out for a large number of permuted data sets (typically 1000 permutations) and the range of singular values produced by the permutations is determined. If the observed singular value exceeds the 95^{th} percentile of the singular values produced by permutations, it is possible to claim that the observed singular value is significantly higher than produced by chance with a 5% level of confidence. It is also possible to employ bootstrap sets of the data matrices (where specimens are bootstrapped

independently in the two blocks) rather than permutations, the logic and interpretation of the bootstrap approach is identical to that of the permutation test.

It may also be productive to examine the correlation of the SVD axis scores obtained for the two blocks. The permutation (or independent block bootstrap) test may also be used to determine if the correlation of the SVD scores exceeds that expected by chance.

It is also possible to compute a confidence interval on the correlation of the SVD axis scores, by bootstrapping the specimens in the analysis and repeating the calculations to arrive at a bootstrap estimate of the distribution of correlation coefficients possible given the distribution and sample size of the data sets. Note that in this procedure, the two blocks are not treated independently, specimens are bootstrapped in a way that preserves the covariance structure between blocks, the bootstrap serves to estimate the variance in the overall analysis. This estimate could be used to compare the between block correlations of different sets of specimens to one another {poor wording! Fix}.

Comparing Patterns of Covariance between Different Groups of Specimens