Analysis of structure in datasets

Goal: objectively determine the structure and the relationships among structures of different fields

Main tool: solve eigenvectors and eigenvalues of the variance (EOF) or covariance (MCA) matrix

Methodologies:
- Univariate analysis: Principle component analysis (PCA), Empirical Orthogonal functions (EOF), rotated EOF, complex EOF, singular value decomposition (SVD),
- Multivariate analyses: Maximum covariance analysis (MCA, SVD applied to two variable fields), Canonical correlation analysis (CCA), multivariate EOFs
- References: Hartmann’s notes, Wallace’s notes, Wilks Ch. 12.2.
Multivariate EOF or Principle component analysis:

- Main references: Hartmann’ notes, and singular value decomposition tutorial, Wilks Sect 12.2.2.
- Methodologies:
- Maximum covariance analysis (MCA) previously Singular Value Decomposition (SVD):
- Canonical correlation analysis (CCA)
- Multivariate or extended EOF (MEOF, or EEOF):
• **Maximum covariance analysis (MCA):**
  
  • Identifies the correlation patterns of two two data matrices that are examples of different structures, or state vectors, but which share a common sampling dimension. For example, the fields of water temperature and surface chlorophyll content, measured at the same set of times (does not have to be the same locations). It does so by applying SVD to the covariance matrix of the two fields.

  • Prohaska (1976) first perhaps used MCA in the meteorological literature, although it has long been used in the social sciences. Bretherton et al. (1992) and Wallace et al. (1992) popularized it for meteorological and oceanographic use.
Maximum covariance analysis (MCA):

- How does it work?
- For two data matrixes, \([X]\) with \(M\) spatial points (structure) and \(N\) (temporal) samples, and \([Y]\) with \(L\) spatial points and \(N\) samples, we compute their covariance as shown below:

\[
\frac{1}{N} \times (X)[Y]^T = C_{xy} \quad \text{where}
\]

\[
[X] = M \begin{bmatrix} x_{11} & x_{1N} \\ x_{M1} & x_{MN} \end{bmatrix}, \quad [Y] = L \begin{bmatrix} y_{11} & y_{1N} \\ y_{L1} & y_{LN} \end{bmatrix}, \quad [C_{xy}] = M \begin{bmatrix} C_{11} & C_{1L} \\ C_{M1} & C_{ML} \end{bmatrix},
\]

Apply SVD to \(C_{xy}\), we have

\[
[C_{xy}] = [U][\Sigma][V]^T \quad |C_{xy}|^2 = \sum_{i=1}^{M} \sum_{j=1}^{L} (x_{ij}y_{ij})^2 = \sum_{i=1}^{M} \sigma_i^2 = \sum_{j=1}^{L} (x^*_k y^*_{kj})^2
\]

- The columns of \([U]\) (MXM) are the column space of \([C_{xy}]\), present the structure of covariance field of \([X]\), and the columns of \([V]\) are the row space of \([C_{xy}]\), represent the structure of covariance fields of \([Y]\). \([\Sigma] = [\sigma^2]/N\), represents the squared co-variance for singular vector. \(X^* = [U]^T[X]\) and \(Y^* = [V]^T[Y]\) correspond to the principle component (PC) of the EOFs for \([X]\) and \([Y]\), respectively. \(\sigma_k = x_k^* y_k^* = \text{sum}(x_{kj}^* y_{kj}^*)\) for \(j = 1, N\) is the single value equal to the covariance between \([X]\) and \([Y]\) fields.
- Notice that the PCs or expansion coefficient time series of the left and right fields \([X]\) and \([Y]\) in MCA are not orthogonal. The correlation coefficient between \(X^*_{k}\) and \(Y^*_{k}\) is a measure of the coupling between the two patterns (modes) in the two fields.
How do we display singular vectors in MCA?

- Singular vectors are non-dimensional and normalized, can efficiently represent the covariance structure, instead of variance structures in the two data fields. To show the latter, we can use the regression map between the PC and the original data for $[X]$ and $[Y]$, respectively.

- Heterogeneous regression maps: show the spatial structure of $[X]$ with amplitude and dimension that is correlated with (project onto) the PC of $[Y]$ ($Y^*$), and vice versa, i.e.,

$$
\tilde{u}_k = \frac{1}{N\sigma_k} [X][y^*_k]^T \text{ or } u_{jk} = \frac{1}{N\sigma_k} \sum_{i=1}^{N} x_{ji}y^*_i = \frac{1}{\sigma_k} x_jy_k^* \text{ heterogeneous map for left field } [X]
$$

$$
\tilde{v}_k = \frac{1}{N\sigma_k} [Y][x^*_k]^T \text{ or } v_{jk} = \frac{1}{N\sigma_k} \sum_{i=1}^{N} y_{ji}x^*_i = \frac{1}{\sigma_k} y_jx_k^* \text{ heterogeneous map for right field } [Y]
$$

- The amplitude information in PCs ($x^*$ and $y^*$) can be incorporate into the singular vector patterns.
Homogeneous regression maps:

One can also determine the pattern of variance (with dimension and amplitude) that is correlated with the PC of its own field, by

\[
\tilde{u}_k = \frac{1}{N\sigma_k} [X] x_k^T \text{ or } u_{jk} = \frac{1}{N\sigma_k} \sum_{i=1}^N x_{ji} x_{ik}^* = \frac{1}{\sigma_k} x_j x_k^* \text{ homogeneous map for left field } [X]
\]

\[
\tilde{v}_k = \frac{1}{N\sigma_k} [Y] y_k^T \text{ or } v_{jk} = \frac{1}{N\sigma_k} \sum_{i=1}^N y_{ji} y_{ik}^* = \frac{1}{\sigma_k} y_j y_k^* \text{ homogeneous map for right field } [Y]
\]

Again, the correlation coefficient between PCs of \([X]\) and \([Y]\) \((X^*_k \text{ and } Y^*_K)\) is a measure of the coupling between the two patterns of the singular vectors of the two fields.
Significance of the MCA:
- Check RMSC before apply MCA to determining whether the two fields are well-correlated:

\[
RMSC = \left( \frac{\sum_{i=1}^{M} \sum_{j=1}^{L} (x_i y_j)^2}{\sum_{i=1}^{M} (x_i)^2 \sum_{j=1}^{L} (y_j)^2} \right)
\]

if RMSC \approx 0.1 for well correlated fields

- Carry out EOF analysis of each fields and to see if the two sets of patterns are significantly correlated (Cherry 1997).

- Significance: No clear rules, here are some suggestions:
  - exam the covariance and variance explained by the singular vectors
  - Monte carol test of random numbers matrix with the same dimension as the real data to determine the probability the covariance patterns are obtained by chance.
• Canonical Correlation Analysis (CCA):

• Rationale: sometimes the raw data maybe too noisy. To reduce the chance that correlation patterns emerges from random data and clarify the correlation pattern of coherent signals, Barnett and Preisendorfer (1987) suggest to reconstruct the each of the two data matrixes based on few leading EOFs of the data matrix. Then, one can apply MCA analysis to the reconstructed two data matrixes.

• Often, the input data matrices are comprised of normalized PC time series derived from separate EOF analyses of the x and y fields such that the expansion coefficient time series for each mode have unit variance, so that the correlation (rather than the covariance) between the x and y expansion coefficient time series is maximized.

• CCA is more discriminating than SVD analysis at identifying coupled patterns in two fields, but it is also more susceptible to sampling variability.
How much PCs to retain for reconstructing the data matrix?

- Best determined by experiments various % of the total variance, perhaps >70%.
- The number of degrees of freedom in structure that are retained should be >> than the number of degrees of freedom of the independent samples, especially when structure (spatial) points are more than the sample (temporal) points.
Multivariate or extended EOF:

- One can evaluate the covariance patterns among \( L (>2) \) variables that share common sampling space, e.g., take multiple measurements at the same time, by construction a data matrix as:

\[
X = L \times K \times 1 \times N \times K \times 1 \times Y \times 1 \times N \times Z \times 1 \times N
\]

- Where \( K \) is the dimension of the structure (spatial) points and \( N \) is the number of sampling (temporal) points. One can stack \( M \) rows of variable \( Y \) below the \( M \) rows of variable \( X \), followed by \( M \) rows of \( Z \), etc, until the \( L^{th} \) variable, construct a matrix of \( L \times K \) rows and \( N \) columns as shown above. Applying SVD or solve for eigenvalues and eigenvectors.

- The covariance of matrix of this data matrix has \( L \times K \) by \( L \times K \) dimension. \([C]_{1,1}\) shows the \( K \times K \) covariance matrix of variable \( X \) and \( X \), whereas \([C]_{1,L}\) shows \( K \times K \) covariance between \( X \) and \( L^{th} \) variable, etc.

- The first eigenvector of all \( L \) variables are shown as the first column of the \([E]\) matrix. The first eigenvector of \( X \) is represented by \( 1-K \) rows of the first column. The first eigenvector of the \( Y \) is represented by the \( K+1 \) to \( 2K \)th rows of the first column, and the first eigenvector of the \( L^{th} \) variable is represented by the \((L-1)K+1\) to \( LK \)th rows of the first column.
Assignment:

- Apply SVD to determine the field relationship of the two fields of your interests and compare the pattern of heterogeneous regression maps of the PC of the leading singular mode for the left and right fields. Compare to the pattern of the local regression, i.e., regression coefficient between the two variables of the same location.

- Apple multivariate EOF to sea surface temperature anomalies (SST’), and surface zonal and meridional wind anomalies (u’, v’), respectively. Plot the spatial patterns of the 1st EOF with real amplitude and unit (multiple 1st PC to the SST’, u’ and v’ data matrix). Use vector to show the pattern of leading EOF of the u’ and v’ and either color shades or contours to show the pattern of SST’.

- Apply multivariate EOF to data of your own research.