Lecture 22: : Maximum covariance analysis

Winter 2015

Ref: Hartmann Ch. 4

22.1 Comparing time-dependent spatial fields of data

We often would like to isolate spatial patterns which covary in multiple fields of data. There are several strategies for doing this:

Regression We make an **index time series** isolating an important mode of variability in one field of interest (e. g. its first principal component). Then we regress the other spatially-varying fields on this to get regression maps. See **USTA_regress.html** for an example.

Combined PCA We standardize all fields at all locations and concatenate them into a single time-dependent vector, on which we perform PCA. The spatial patterns can then be partitioned into maps associated with each of the fields, and the PCs give their common time variability. See Bretherton et al. (1992, *J. Climate*)

MCA Maximum covariance analysis (MCA) looks for patterns in two space-time datasets which explain a maximum fraction of the covariance between them. See below.

CCA Canonical correlation analysis (CCA) loots for patterns in two space-time datasets with maximum temporal correlation coefficient. CCA does not necessarily pick patterns which explain much covariance and can be severely affected by random sampling fluctuations. To minimize these issues, the two fields should be **prefiltered** by projection onto a subset of their leading EOFs sufficient to explain most (e. g. 90%) of their variance. See Bretherton et al. (1992).

Regression is easy to implement and understand, but can only describe that part of the covariability between the fields that is related to the index time series. Combined PCA works best if the leading EOFs of the two fields encompass most of their mutual correlation. MCA and preconditioned CCA give similar results, and are best if the patterns of covariability are not well known a priori. Here we describe MCA, since it is simpler than CCA.

22.2 Mathematical setup of MCA

Mathematically we consider two data matrices $\mathbf{X}$ $[m \times n]$ and $\mathbf{Y}$ $[q \times n]$, where n is the number of samples (times) and m and q are respectively the number of x and y measurements at each time. We let $\mathbf{u}$ be an arbitrary unit column m-vector representing a pattern in the x field and $\mathbf{v}$ be an arbitrary unit column m-vector representing a pattern in the y field. Let the time series of their projection onto the data be the $1 \times n$ row vectors,

$$\begin{aligned} \mathbf{a} &= \mathbf{u}^T\mathbf{X}, \\ \mathbf{b} &= \mathbf{v}^T\mathbf{Y}. \end{aligned}$$

Then MCA seeks optimal patterns $\mathbf{u}$ and $\mathbf{v}$ that maximize their covariance

$$\begin{aligned} c &= \mathrm{cov}[\mathbf{a}, \mathbf{b}] \\ &= \mathrm{cov}[\mathbf{u}^T\mathbf{X}, \mathbf{v}^T\mathbf{Y}] \\ &= \frac{1}{n-1}[\mathbf{u}^T\mathbf{X}(\mathbf{v}^T\mathbf{Y})^T] \\ &= \mathbf{u}^T\mathbf{C}_{xy}\mathbf{v}, \end{aligned} \tag{22.2.1}$$

where

$$\mathbf{C}_{xy} = \frac{1}{n-1}\mathbf{X}\mathbf{Y}^T$$

is the covariance matrix between x and y, whose ij'th element is the covariance of $x_i(t)$ with $y_j(t)$.

The maximum c is obtained from the leading mode of the SVD of $\mathbf{C}_{xy}$, with x pattern $\mathbf{u}_1$ (the first left singular vector), y pattern $\mathbf{v}_1$ (the first right singular vector), and $c = \sigma_1$, the first singular value. Succeeding SVD modes maximize c subject to the additional constraint that the patterns be spatially orthogonal to the previous modes.

test the orthogonal fact in u and v vectors

Each SVD mode explains an amount σ_k^2 of the overall squared covariance in $\mathbf{C}_{xy}$. Thus, it is useful to think of the importance of the SVD modes in MCA in terms of their *squared covariance fraction*

$$f_k = \frac{\sigma_k^2}{\sum_{k=1}^{r}\sigma_k^2}$$

22.3 Proof of optimality of leading SVD mode

The proof works analogously to the proof in Lect. 20 for the optimality of PCA in explaining a maximum fraction of the variance of a field. We write the SVD:

$$\mathbf{C}_{xy} = \mathbf{U}\Sigma\mathbf{V}^T$$

We express $\mathbf{u}$ in the basis of left singular vectors, in which it has coordinates

$$\hat{\mathbf{u}} = \mathbf{U}^T\mathbf{u}$$

Because $\mathbf{U}$ is a rotation matrix and $\mathbf{u}$ is a arbitrary unit vector, $\hat{\mathbf{u}}$ is also some other arbitrary unit vector. We expand $\mathbf{v}$ in the basis of right singular vectors, in which it has coordinates

$$\hat{\mathbf{v}} = \mathbf{V}^T \mathbf{v}$$

Again, $\hat{\mathbf{v}}$ is an arbitrary unit vector.

We substitute into (22.2.1). Letting r be the rank of the covariance matrix,

$$\begin{aligned} c &= \mathbf{u}^T \mathbf{C}_{xy} \mathbf{v} \\ &= \mathbf{u}^T \mathbf{U} \Sigma \mathbf{V}^T \mathbf{v} \\ &= \hat{\mathbf{u}}^T \Sigma \hat{\mathbf{v}} \\ &= \sum_{k=1}^{r} \hat{u}_k \sigma_k \hat{v}_k \\ &\leq \left(\sum_{k=1}^{r} \sigma_k^2 \hat{u}_k^2 \right)^{1/2} \left(\sum_{k=1}^{r} \hat{v}_k^2 \right)^{1/2} \\ &\leq \left(\sum_{k=1}^{r} \sigma_k^2 \hat{u}_k^2 \right)^{1/2} \\ &\leq \sigma_1 \left(\sum_{k=1}^{r} \hat{u}_k^2 \right)^{1/2} \leq \sigma_1 \end{aligned}$$

The maximum is achieved by taking $\hat{u}_1 = 1$ and $\hat{u}_k = 0,\ k > 1$, i. e. for $\mathbf{u} = \mathbf{u_1}$. Similarly, from the fourth line in the above equation we see that we must take $\hat{v}_1 = 1$ and other coordinates zero, i. e. $\mathbf{v} = \mathbf{v_1}$.

Similarly, mode $k = N + 1$ explains the maximum fraction of the covariance that is in spatial patterns orthogonal to the first N modes.

22.4 Time series of the MCA patterns

We can define time series of the patterns associated with the k'th SVD mode, which are $1 \times n$ row vectors with covariance σ_k:

$$\mathbf{a}_k = \mathbf{u}_k^T \mathbf{X}$$
$$\mathbf{b}_k = \mathbf{v}_k^T \mathbf{Y}$$

There is a separate time series for each of the two datasets, and the time series associated with different SVD modes are not guaranteed to be uncorrelated.

22.5 MCA example

Matlab script **MCA_PSSTA_USTA** demonstrates the implementation of MCA on the datasets we have previously been using, finding the patterns of maximum covariance between monthly tropical Pacific sea-surface temperature anomalies and U.S. surface temperature anomalies over the period 1970-2002.