Canonical Correlation Analysis (CCA)


These notes are based on Barnett and Preisendorfer, 1987 and have as examples Diaz et al. 1998 and Busuioc et al. 2001.

Three levels to relate variables.

1. Simple Regression
2. Multiple Regression
3. Canonical Correlation Analysis

It is a modelling method that models by finding the optimum linear combination of the predictor that will explain most of the variance in the predictand. Both predictor and predictand are full dimensional (at least 3 dimensional).

One of the most commonly used statistical methods for modelling of fields.

“What patterns tend to occur together in fields “A” and “B” and what is the degree of connection between them?”

If compared to PCA:
PCA – defines the orthogonal coordinate system that describes the maximum variance of one data set.
CCA – describes the coordinate system that describes the maximum cross–covariance between two data sets.
The theory:

Suppose we have two data sets $Y_{t,y}$ and $Z_{t,z}$

What we intend to do using CCA is to find a pair of vectors $U$ and $V$ that are a linear combination from $Y$ and $Z$, respectively, and are maximally correlated, i.e.:

$$U = YR;$$
$$V = ZQ;$$

$$\text{corr}(U, V) = \frac{\sum UV}{\left(\sum U^2 \sum V^2\right)^{1/2}} = \max$$

Physically, this works as if we were projecting the data ($Y$ and $Z$) onto $R$ and $Q$ (the transformation). Then $U$ and $V$ are the projected values of $Y$ and $Z$, respectively.
Solving this equation system for $R$ and $U$

\[
\text{var}(U) = \frac{\sum U^2}{nt - 1} = \frac{U'U}{nt - 1} = \frac{(YR)'(YR)}{nt - 1} = \frac{R'Y'YR}{nt - 1}
\]

\[
\frac{Y'Y}{nt - 1} \text{ is the covariance matrix (S}_{yy} \text{ of Y}
\]

\[
\sum U^2 = R'S_{yy}R \quad \text{(4)}
\]

\[
\sum V^2 = Q'S_{zz}Q \quad \text{(5)}
\]

\[
\sum UV = R'S_{yz}Q \quad \text{(6)}
\]

where: $S_{yy} = \frac{Y'Y}{nt - 1}$; $S_{zz} = \frac{Z'Z}{nt - 1}$; and $S_{yz} = S_{zy} = \frac{Y'Z}{nt - 1}$

Substituting (4), (5), and (6) on (3) gives:

\[
\text{corr}(U, V) = \frac{R'S_{yz}Q}{\left(\frac{R'S_{yy}R \times Q'S_{zz}Q}{1/2}\right)} = \max
\]

To avoid $R$ and $Q$ being arbitrarily large, we set the constraint that the total variance must be 1, i.e.: $R'S_{yy}R = 1$ and $Q'S_{zz}Q = 1$.

So $F(R,Q) = R'S_{yz}Q$ is to be maximized subjected to $R'S_{yy}R = 1$ and $Q'S_{zz}Q = 1$.
Two Lagrange multipliers to solve the problem: \( \mu \) and \( \lambda \), (Lagrange multiplier is used to find the extreme of \( f(x) \) subject to the constraint \( g(x) = c \)) and the function to be maximized becomes:

\[
F(R,Q) = R'S_{yz}Q - \mu(Q'S_{zz}Q-1) - \lambda(R'S_{yy}R-1)
\]

\[
\frac{\partial F}{\partial R'} = 0 \text{ and } \frac{\partial F}{\partial Q} = 0
\]

\[
(S_{yy}^{-1}S_{yz}S_{zz}^{-1}S_{zy} - \mu^2I)R = 0
\]

And can be treated as a eigenvalue problem where:
- \( \mu^2 \) are the eigenvalues of \( S_{yy}^{-1}S_{yz}S_{zz}^{-1}S_{zy} \),
- \( R \) are the eigenvectors
- \( I \) the identity matrix and

\[
Q = \frac{1}{\mu} S_{zz}^{-1}S_{zy}R
\]

Within CCA:
- \( R \) and \( Q \) are the canonical spatial patterns,
- \( \mu^2 \) is the squared canonical correlation coefficients. It contains the squared of the correlation between \( U \) to \( V \).

\( \mu \) is called canonical correlation coefficients. Physically, they represent the level of correlation between patterns of the predictor and patterns of the predictand (that are present in \( U \) and \( V \)).
$R_{ny,nm}$ and $Q_{nz,nm}$ are the **canonical spatial patterns** where $nm$ is the number of **canonical modes**. They contain the weights for transforming $Y$ to $U$ and $Z$ to $V$ (transform).

The new time series (**canonical vectors**) $U$ and $V$ that are the transformation of $Y$ and $Z$ are calculated by:

$$U_{nt,nm} = Y_{nt,ny} R_{ny,nm}$$
$$V_{nt,nm} = Z_{nt,nz} Q_{nz,nm}$$

Notes:

1) Correlation between $U$ and $V$ is given by:

$$<U_i V_j> = \delta_{ij} \mu_i^{1/2}$$
$$<U_i U_j> = \delta_{ij}$$
$$<V_i V_j> = \delta_{ij}$$

2) The number of eigenvalues of $S_{yy}^{-1}S_{yz}S_{zz}^{-1}S_{zy}$ is equal to the minimum between $ny$ and $nz$. 
However, just like linear regression, CCA is subject to instability if the number of time steps (realizations) is not large.

**Solution:** Pre filter the Y and Z with PCA or extended PCA. Apply CCA to the time series resulting from the PCA.

What the PCA does to help?

Remove part of the random noise from the data.
How to do the new PCA/CCA then?

1- Prepare your data set:
   a. Clean missing data (no more than 10% missing data per variable/station).
   b. Normalize/standardize your data
   c. Final input data should be time x variable/space

2- Apply PCA to your 2 sets of data,

3- Identify the significant modes (physically or statistically),

4- Calculate the new time series for the chosen modes (project your data on the new coordinates: $U_y = Y E_y$
   $U_z = Z E_z$

5- Calculate CCA using as indata $U_y$ and $U_z$ instead of $Y$ and $Z$,

6- Calculate the time series (canonical vectors) of the canonical patterns, ($U$ and $V$)

7- Calculate the **canonical maps** ($g$ and $h$ maps):

   These maps show the correlations at specific locations between the new time series (canonical vectors) and the reconstructed time series from the PCA. $g_{nx,nm} = U'U_x/(nt-1)$
   $h_{ny,nm} = V'U_y/(nt-1)$

   The statistical significance of these correlations must be checked.

8- Plot the time series, and the $g$ and $h$ maps or the spatial maps,

9- Analyze all together.
Explained variance:

If the percentage of the variance accounted for by the PCA modes (m) are stored in expvar (as in pca_general.m), then the fraction of variance of Y accounted for by a particular canonical mode is

$$\text{exp var}_m^Y = \sum_{i=1}^{\text{nmy}} R_{i,m}^2 \text{ exp var}_i$$

$$\text{exp var}_m^Z = \sum_{i=1}^{\text{nmz}} Q_{i,m}^2 \text{ exp var}_i$$
Example:
Diaz et al. (1998).
\(X = \) Pacific Ocean sea surface temperature
\(Y = \) seasonal precipitation
$X = \text{Atlantic Ocean sea surface temperature}$

$Y = \text{seasonal precipitation}$
Busuioc et al. (2001)
X = seasonal sea level pressure
Y = precipitation

Fig 8. The patterns of the 1st 3 CCA pairs of monthly mean SLP (left) and monthly total precipitation (right) in Sweden (the eastern islands are not shown) for January. The canonical correlation coefficient between the time coefficient series associated to the patterns of the two parameters as well as the corresponding explained variances are shown.
In practice, not all canonical modes are used to develop a CCA model. Only those modes that are statistically or physically significant are kept for the development of the model.

In this way, we avoid the model being developed on the bases of fitting of noise so that the main features of the temporal and spatial variations of the fields will drive the model.

Choosing the CCA significant modes.

- Subjectively
  Physics

- Objectively
  Statistics – Significant tests

  Jackson (1991) has 11 different options.


1- Calculate CCA.

2- Average $\mu$

3- Cutting value is 70% of the mean $\mu$

4- Compare cutting values with each $\mu_i$

5- Keep only the modes that have $\mu_i > 0.7 \times \text{mean}(\mu)$
Modeling and forecasting with PCA/CCA

What do we have?
Original data:

\( \mathbf{Y} \) is the predictor
\( \mathbf{Z} \) is the predictand

After applying EOF to \( \mathbf{Y} \) and \( \mathbf{Z} \) we have:

For \( \mathbf{Y} \):
- \( \lambda_y \) Eigenvalues
- \( \mathbf{E}_y \) Eigenvector
- \( \mathbf{U}_y \) Time series

For \( \mathbf{Z} \):
- \( \lambda_z \) Eigenvalues
- \( \mathbf{E}_z \) Eigenvector
- \( \mathbf{U}_z \) Time series

After applying CCA to \( \mathbf{U}_y \) and \( \mathbf{U}_z \):

- \( \mathbf{R} \) is the CCA eigenvector for the predictor (\( \mathbf{U}_y \))
- \( \mathbf{Q} \) is the CCA eigenvector for the predictand (\( \mathbf{U}_z \))
- \( \mathbf{U} \) CCA time series for \( \mathbf{U}_y \)
- \( \mathbf{V} \) CCA time series for \( \mathbf{U}_z \)
- \( \mu \) Canonical correlation

We want to find an equation (regression) that estimates \( \mathbf{Z} \) from a particular case of \( \mathbf{Y} \) (original data).

\[ \hat{\mathbf{Z}} = f(\mathbf{Y}) \]
We start by calculating a matrix of regression coefficients \((A)\) for converting the predictor canonical mode temporal amplitudes – time series – \((U)\) to estimates of the predictand PCA time series \((U_z)\).

\[
A_{m,L_z} = \frac{1}{nt} U'U_z \quad \text{predictor CCA time series} \ast \text{predictand PCA time series}
\]

where \(m\) denotes the number of canonical modes and \(L_z\) the number of modes chosen as significant in the PCA of \(Z\).

So that an estimation of the \(Z\) PCA time series \((U_z)\) can be made by:

\[
\hat{U}_z = UA \quad (1)
\]

and then an estimation of \(Z\) by:

\[
\hat{Z} = \hat{U}_z E'_z \quad \text{time series} \ast \text{PCA predictand eigenvector}
\]

Applying (1) here:

\[
\hat{Z} = UAE'_z \quad \text{(projection of the predictor on to the input PCA)}
\]
But we used $U_y$ and $U_z$ (time series from the PCA) as input to the CCA, so that

$$U = U_y \cdot R \quad \text{and then} \quad \hat{Z} = U_y \cdot R \cdot E' \cdot Z$$

now,

$$U_y = Y \cdot E_y \quad \text{(if your initial data is time x variable/space)}$$

so,

$$\hat{Z} = Y \cdot E_y \cdot R \cdot E' \cdot Z$$

but

$S_F = E_y \cdot R$ is the spatial function of the predictor, so that

$$\hat{Z} = Y \cdot S_F \cdot A \cdot E' \cdot Z$$
Report

You are expected to read through at least Barnett and Preisendorfer, 1987 (find it on fire or ftp it from air) before you write your report on CCA.

The report should contain at least:
Description of the data.
Application of CCA to your data set in analysis mode. Physical interpretation of the results.
   How many modes you consider important, why?
   Can you explain physically the modes you consider important? What do they mean?

Develop a CCA model with your data set.
   Analyse the skill of the model (ability of the model to estimate the predictand variable).
References: