## Ch.3 Canonical correlation analysis (CCA) [Book, Sect. 2.4]

With 2 sets of variables $\{x_i\}$ and $\{y_j\}$, *canonical correlation analysis* (CCA), first introduced by Hotelling (1936), finds the linear modes of maximum correlation between $\{x_i\}$ and $\{y_j\}$.

CCA is a generalization of the Pearson correlation between two variables $x$ and $y$ to two sets of variables $\{x_i\}$ and $\{y_j\}$.

CCA: Find $\mathbf{f}_1$ and $\mathbf{g}_1$, so that the correlation between $\mathbf{f}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_1^{\mathrm{T}}\mathbf{y}$ is maximized.

Next find $\mathbf{f}_2$ and $\mathbf{g}_2$ so that the correlation between $\mathbf{f}_2^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_2^{\mathrm{T}}\mathbf{y}$ is maximized, with $\mathbf{f}_2^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_2^{\mathrm{T}}\mathbf{y}$ uncorrelated with both $\mathbf{f}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{g}_1^{\mathrm{T}}\mathbf{y}$.

And so forth for the higher modes.

## 3.1 CCA theory [Book, Sect. 2.4.1]

Consider two datasets

$$\mathbf{x}(t) = x_{il}, \qquad i = 1, \cdots, n_x, \quad l = 1, \cdots, n_t\,, \tag{1}$$

and

$$\mathbf{y}(t) = y_{jl}, \qquad j = 1, \cdots, n_y, \quad l = 1, \cdots, n_t\,. \tag{2}$$

i.e. **x** and **y** need not have the same spatial dimensions, but need the same time dimension $n_t$. Assume **x** and **y** have zero means. Let

$$u = \mathbf{f}^{\mathrm{T}}\mathbf{x}, \quad v = \mathbf{g}^{\mathrm{T}}\mathbf{y}\,. \tag{3}$$

The correlation

$$\rho = \frac{\mathrm{cov}(u, v)}{\sqrt{\mathrm{var}(u)\,\mathrm{var}(v)}} = \frac{\mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{g}^{\mathrm{T}}\mathbf{y})}{\sqrt{\mathrm{var}(u)\,\mathrm{var}(v)}} = \frac{\mathbf{f}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}, \mathbf{y})\mathbf{g}}{\sqrt{\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x})\,\mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y})}}\,, \tag{4}$$

where we have invoked

$$\mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{g}^{\mathrm{T}}\mathbf{y}) = \mathrm{E}[\mathbf{f}^{\mathrm{T}}\mathbf{x}\,\mathbf{g}^{\mathrm{T}}\mathbf{y}] = \mathrm{E}[\mathbf{f}^{\mathrm{T}}\mathbf{x}\,\mathbf{y}^{\mathrm{T}}\mathbf{g}] = \mathbf{f}^{\mathrm{T}}\mathrm{E}[\mathbf{x}\mathbf{y}^{\mathrm{T}}]\mathbf{g}\,. \tag{5}$$

We want *u* and *v*, the two *canonical variates* or *canonical correlation coordinates*, to have maximum correlation between them, i.e. **f** and **g** are chosen to maximize $\rho$.

We are free to normalize **f** and **g** as we like, because if **f** and **g** maximize $\rho$, so will $\alpha$**f** and $\beta$**g**, for any positive $\alpha$ and $\beta$.
We choose the normalization condition

$$\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x}) = 1 = \mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y})\,. \tag{6}$$

Since

$$\mathrm{var}(\mathbf{f}^{\mathrm{T}}\mathbf{x}) = \mathrm{cov}(\mathbf{f}^{\mathrm{T}}\mathbf{x}, \mathbf{f}^{\mathrm{T}}\mathbf{x}) = \mathbf{f}^{\mathrm{T}}\mathrm{cov}(\mathbf{x}, \mathbf{x})\mathbf{f} \equiv \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f}\,, \tag{7}$$

and

$$\mathrm{var}(\mathbf{g}^{\mathrm{T}}\mathbf{y}) = \mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g}\,, \tag{8}$$

(6) implies

$$\mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f} = 1\,, \quad \mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g} = 1\,. \tag{9}$$

With (6), (4) reduces to

$$\rho = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{g}\,, \tag{10}$$

where $\mathbf{C}_{xy} = \mathrm{cov}(\mathbf{x}, \mathbf{y})$.

Problem is to maximize (10) subject to constraints (9).

Use method of Lagrange multipliers [Book, Appendix B], where we incorporate the constraints into the Lagrange function $L$,

$$L = \mathbf{f}^{\mathrm{T}}\mathbf{C}_{xy}\mathbf{g} + \alpha(\mathbf{f}^{\mathrm{T}}\mathbf{C}_{xx}\mathbf{f} - 1) + \beta(\mathbf{g}^{\mathrm{T}}\mathbf{C}_{yy}\mathbf{g} - 1)\,, \tag{11}$$

where $\alpha$ and $\beta$ are the unknown Lagrange multipliers.

To find the stationary points of $L$, we need

$$\frac{\partial L}{\partial \mathbf{f}} = \mathbf{C}_{xy}\mathbf{g} + 2\alpha \mathbf{C}_{xx}\mathbf{f} = 0\,, \tag{12}$$

and

$$\frac{\partial L}{\partial \mathbf{g}} = \mathbf{C}_{xy}^{\mathrm{T}}\mathbf{f} + 2\beta \mathbf{C}_{yy}\mathbf{g} = 0\,. \tag{13}$$

Hence

$$\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{g} = -2\alpha\mathbf{f}\,, \tag{14}$$

and

$$\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\mathbf{f} = -2\beta\mathbf{g}\,. \tag{15}$$

Substituting (15) into (14) yields

$$\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\,\mathbf{f} \equiv \mathbf{M}_f\mathbf{f} = \lambda\mathbf{f}\,, \tag{16}$$

with $\lambda = 4\alpha\beta$. Similarly, substituting (14) into (15) gives

$$\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\mathrm{T}}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\,\mathbf{g} \equiv \mathbf{M}_g\mathbf{g} = \lambda\mathbf{g}\,. \tag{17}$$

Both these equations can be viewed as eigenvalue equations, with $\mathbf{M}_f$ and $\mathbf{M}_g$ sharing the same non-zero eigenvalues $\lambda$.

As $\mathbf{M}_f$ and $\mathbf{M}_g$ are known from the data, **f** can be found by solving the eigenvalue problem (16).

$\beta\mathbf{g}$ can then be obtained from (15). Since $\beta$ is unknown, the magnitude of **g** is unknown, and the normalization conditions (9) are used to determine the magnitude of **g** and **f**.

Alternatively, one can use (17) to solve for **g** first, then obtain **f** from (14) and the normalization condition (9).

The matrix $\mathbf{M}_f$ is of dimension $n_x \times n_x$, while $\mathbf{M}_g$ is $n_y \times n_y$, so one usually picks the smaller of the two to solve the eigenvalue problem.

From (10),

$$\rho^2 = \mathbf{f}^{\mathrm{T}} \mathbf{C}_{xy} \mathbf{g}\, \mathbf{g}^{\mathrm{T}} \mathbf{C}_{xy}^{\mathrm{T}} \mathbf{f} = 4\alpha\beta \left(\mathbf{f}^{\mathrm{T}} \mathbf{C}_{xx} \mathbf{f}\right) \left(\mathbf{g}^{\mathrm{T}} \mathbf{C}_{yy} \mathbf{g}\right), \tag{18}$$

where (12) and (13) have been invoked. From (9), (18) reduces to

$$\rho^2 = \lambda. \tag{19}$$

The eigenvalue problems (16) and (17) yield $n$ number of $\lambda$s, with $n = \min(n_x, n_y)$.

Assuming the $\lambda$s to be all distinct and nonzero, we have for each $\lambda_j$ ($j = 1, \ldots, n$), canonical variates, $u_j$ and $v_j$, with correlation $\rho_j = \sqrt{\lambda_j}$ between the two, and eigenvectors, $\mathbf{f}_j$ and $\mathbf{g}_j$.

It can be shown that

$$\mathrm{cov}(u_j, u_k) = \mathrm{cov}(v_j, v_k) = \delta_{jk}, \quad \text{and} \quad \mathrm{cov}(u_j, v_k) = 0 \quad \text{if} \quad j \neq k\,. \tag{20}$$

Write the forward mappings from the variables $\mathbf{x}(t)$ and $\mathbf{y}(t)$ to the canonical variates $\mathbf{u}(t) = [u_1(t), \cdots, u_n(t)]^{\mathrm{T}}$ and $\mathbf{v}(t) = [v_1(t), \cdots, v_n(t)]^{\mathrm{T}}$ as

$$\mathbf{u} = [\mathbf{f}_1^{\mathrm{T}}\mathbf{x}, \cdots, \mathbf{f}_n^{\mathrm{T}}\mathbf{x}]^{\mathrm{T}} = \mathcal{F}^{\mathrm{T}}\mathbf{x}, \quad \mathbf{v} = \mathcal{G}^{\mathrm{T}}\mathbf{y} \tag{21}$$

Next, find the inverse mapping from $\mathbf{u} = [u_1, \cdots, u_n]^{\mathrm{T}}$ and $\mathbf{v} = [v_1, \cdots, v_n]^{\mathrm{T}}$ to the original variables $\mathbf{x}$ and $\mathbf{y}$. Let

$$\mathbf{x} = \mathbf{F}\mathbf{u}, \quad \mathbf{y} = \mathbf{G}\mathbf{v}\,. \tag{22}$$

We note that

$$\mathrm{cov}(\mathbf{x},\mathbf{u}) = \mathrm{cov}(\mathbf{x},\mathcal{F}^{\mathrm{T}}\mathbf{x}) = \mathrm{E}[\mathbf{x}(\mathcal{F}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}] = \mathrm{E}[\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\mathcal{F}] = \mathbf{C}_{xx}\mathcal{F}\,, \tag{23}$$

and

$$\mathrm{cov}(\mathbf{x},\mathbf{u}) = \mathrm{cov}(\mathbf{F}\,\mathbf{u},\mathbf{u}) = \mathbf{F}\,\mathrm{cov}(\mathbf{u},\mathbf{u}) = \mathbf{F}\,. \tag{24}$$

Eqs. (23) and (24) imply

$$\mathbf{F} = \mathbf{C}_{xx}\mathcal{F}\,. \tag{25}$$

Similarly,

$$\mathbf{G} = \mathbf{C}_{yy}\mathcal{G}\,. \tag{26}$$

Hence the inverse mappings **F** and **G** (from the canonical variates to **x** and **y**) can be calculated from the forward mappings $\mathcal{F}^T$ and $\mathcal{G}^T$.

The matrix $\mathbf{F}$ is composed of column vectors $\mathbf{F}_j$, and $\mathbf{G}$, of column vectors $\mathbf{G}_j$. $\mathbf{F}_j$ and $\mathbf{G}_j$ are the *canonical correlation patterns* associated with $u_j$ and $v_j$.

In general, orthogonality of vectors within a set is not satisfied by any of the four sets $\{\mathbf{F}_j\}$, $\{\mathbf{G}_j\}$, $\{\mathbf{f}_j\}$ and $\{\mathbf{g}_j\}$, while

$$\mathrm{cov}(u_i, u_j) = \mathrm{cov}(v_i, v_j) = \mathrm{cov}(u_i, v_j) = 0, \quad \text{for } i \neq j\,. \tag{27}$$

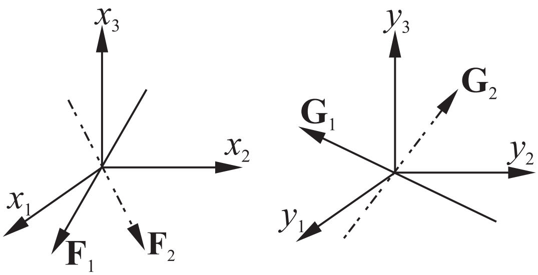


Figure : The CCA solution in the **x** and **y** spaces. Vectors $\mathbf{F}_1$ and $\mathbf{G}_1$ are the canonical correlation patterns for mode 1, and $u_1(t)$ is the amplitude of the "oscillation" along $\mathbf{F}_1$, and $v_1(t)$, the amplitude along $\mathbf{G}_1$. Vectors $\mathbf{F}_1$ and $\mathbf{G}_1$ have been chosen so that the correlation between $u_1$ and $v_1$ is maximized. Next $\mathbf{F}_2$ and $\mathbf{G}_2$ are found, together with $u_2(t)$ and $v_2(t)$. The correlation between $u_2$ and $v_2$ is again maximized, but with $\mathrm{cov}(u_1, u_2) = \mathrm{cov}(v_1, v_2) = \mathrm{cov}(u_1, v_2) = \mathrm{cov}(v_1, u_2) = 0$.

Unlike PCA, $\mathbf{F}_1$ and $\mathbf{G}_1$ need not be not oriented in the direction of maximum variance.

Solving for $\mathbf{F}_1$ and $\mathbf{G}_1$ is analogous to performing rotated PCA in the $\mathbf{x}$ and $\mathbf{y}$ spaces separately, with the rotations determined from maximizing the correlation between $u_1$ and $v_1$.

## 3.2 Pre-filter with PCA [Book, Sect. 2.4.2]

When $\mathbf{x}$ and $\mathbf{y}$ contain many variables, it is common to use PCA to pre-filter the data to reduce the dimensions of the datasets, i.e. apply PCA to $\mathbf{x}$ and $\mathbf{y}$ separately, extract the leading PCs, then apply CCA to the leading PCs of $\mathbf{x}$ and $\mathbf{y}$.

Using Hotelling's choice of scaling for the PCAs, we express the PCA expansions as

$$\mathbf{x} = \sum_j a'_j \mathbf{e}'_j, \quad \mathbf{y} = \sum_j a''_j \mathbf{e}''_j \,. \tag{28}$$

CCA is then applied to

$$\tilde{\mathbf{x}} = [a'_1, \cdots, a'_{m_x}]^{\mathrm{T}}, \quad \tilde{\mathbf{y}} = [a''_1, \cdots, a''_{m_y}]^{\mathrm{T}}, \tag{29}$$

where only the first $m_x$ and $m_y$ modes are used.

Another reason for using the PCA pre-filtering is that when the number of variables is not small relative to the sample size, the CCA method may become *unstable* (Bretherton et al., 1992).

Why? In the relatively high-dimensional **x** and **y** spaces, among the many dimensions and using correlations calculated with relatively small samples, CCA can often find directions of high correlation but with little variance, thereby extracting a spurious leading CCA mode, as illustrated.

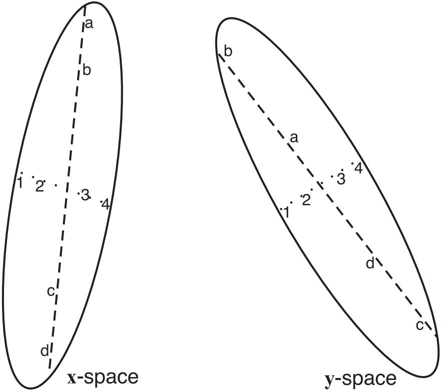


Figure : With the ellipses denoting the data clouds in the two input spaces, the dotted lines illustrate directions with little variance but by chance with high correlation (as illustrated by the perfect order in which the data points 1, 2, 3 and 4 are arranged in the **x** and **y** spaces). Since CCA finds the correlation of the data points along the dotted lines to be higher than that along the dashed lines (where the data points a, b, c and d in the **x**-space are ordered as b, a, d and c in the **y**-space), the dotted lines are chosen as the first CCA mode.

Maximum covariance analysis (MCA), which looks for modes of maximum covariance instead of maximum correlation, would select the dashed lines over the dotted lines since the length of the lines do count in the covariance but not in the correlation, hence MCA is stable even without pre-filtering by PCA.

The instability problem can also be avoided by pre-filtering using PCA, as this avoids applying CCA directly to high-dimensional input spaces (Barnett and Preisendorfer, 1987).

With Hotelling's scaling,

$$\mathrm{cov}(a'_j, a'_k) = \delta_{jk}, \quad \mathrm{cov}(a''_j, a''_k) = \delta_{jk}\,, \tag{30}$$

leading to

$$\mathbf{C}_{\tilde{x}\tilde{x}} = \mathbf{C}_{\tilde{y}\tilde{y}} = \mathbf{I}\,. \tag{31}$$

Eqs.(16) and (17) simplify to

$$\mathbf{C}_{\tilde{x}\tilde{y}}\mathbf{C}_{\tilde{x}\tilde{y}}^{\mathrm{T}}\,\mathbf{f} \equiv \mathbf{M}_f\,\mathbf{f} = \lambda\mathbf{f}\,, \tag{32}$$

$$\mathbf{C}_{\tilde{x}\tilde{y}}^{\mathrm{T}}\mathbf{C}_{\tilde{x}\tilde{y}}\,\mathbf{g} \equiv \mathbf{M}_g\,\mathbf{g} = \lambda\mathbf{g}\,. \tag{33}$$

Q1: Prove that $\mathbf{M}_f$ and $\mathbf{M}_g$ are positive semi-definite symmetric matrices.

---

As $\mathbf{M}_f$ and $\mathbf{M}_g$ are positive semi-definite symmetric matrices, the eigenvectors $\{\mathbf{f}_j\}$ $\{\mathbf{g}_j\}$ are now sets of orthonormal vectors. Eqs.(25) and (26) simplify to

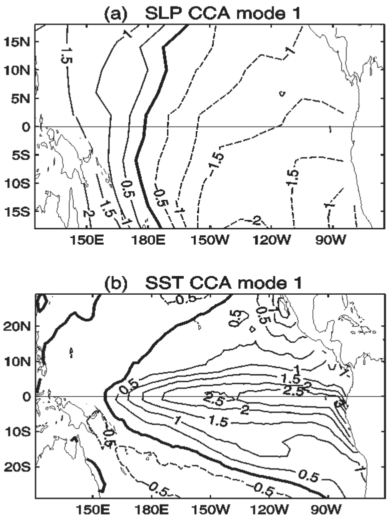
$$\mathbf{F} = \mathcal{F}, \quad \mathbf{G} = \mathcal{G}\,. \tag{34}$$

Hence $\{\mathbf{F}_j\}$ and $\{\mathbf{G}_j\}$ are also two sets of orthonormal vectors, and are identical to $\{\mathbf{f}_j\}$ and $\{\mathbf{g}_j\}$, respectively.

Because of these nice properties, pre-filtering by PCA (with the Hotelling scaling) is recommended when $\mathbf{x}$ and $\mathbf{y}$ have many variables (relative to the sample size).

However, the orthogonality only holds in the reduced dimensional spaces, $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. If transformed into the original space $\mathbf{x}$ and $\mathbf{y}$, $\{\mathbf{F}_j\}$ and $\{\mathbf{G}_j\}$ are in general not two sets of orthogonal vectors.

CCA mode 1 of the tropical Pacific sea level pressure (SLP) field and the SST field:



(a) SLP CCA mode 1

(b) SST CCA mode 1

Figure : The CCA mode 1 for (a) the SLP anomalies and (b) the SST anomalies of the tropical Pacific. As $u_1(t)$ and $v_1(t)$ fluctuate together from one extreme to the other as time progresses, the SLP and SST anomaly fields, oscillating as standing wave patterns, evolve from an El Niño to a La Niña state. The pattern in (a) is scaled by $\tilde{u}_1 = [\max(u_1) - \min(u_1)]/2$, and (b) by $\tilde{v}_1 = [\max(v_1) - \min(v_1)]/2$. Contour interval is 0.5 hPa in (a) and 0.5°C in (b).

The canonical variates $u$ and $v$ (not shown) fluctuate with time, both attaining high values during El Niño, low values during La Niña, and neutral values around zero during normal conditions.

**3.3 Maximum covariance analysis (MCA)** [Book, Sect. 2.4.3]
Instead of maximizing the correlation as in CCA, one can maximize the covariance between two datasets. This alternative method is often called the *singular value decomposition* (SVD). However, von

Storch and Zwiers (1999) proposed the name *maximum covariance analysis* (MCA) as more appropriate.

MCA is identical to CCA except that it maximizes the covariance instead of the correlation.

CCA can be unstable when working with relatively large number of variables, in that directions with high correlation but negligible variance may be selected by CCA, hence the recommended pre-filtering of data by PCA before applying CCA.

MCA, by using covariance instead of correlation, does not have the unstable nature of the CCA, so no need for pre-filtering by PCA.

In MCA, perform SVD on the data covariance matrix $\mathbf{C}_{xy}$,

$$\mathbf{C}_{xy} = \mathbf{USV}^{\mathrm{T}} \, , \tag{35}$$

where the matrix $\mathbf{U}$ contains the left singular vectors $\mathbf{f}_i$, $\mathbf{V}$ the right singular vectors $\mathbf{g}_i$, and $\mathbf{S}$ the singular values. Maximum covariance between $u_i$ and $v_i$ is attained (Bretherton et al., 1992) with

$$u_i = \mathbf{f}_i^{\mathrm{T}}\mathbf{x}, \quad v_i = \mathbf{g}_i^{\mathrm{T}}\mathbf{y} \, . \tag{36}$$

The inverse transform is given by

$$\mathbf{x} = \sum_i u_i \mathbf{f}_i, \quad \mathbf{y} = \sum_i v_i \mathbf{g}_i \, . \tag{37}$$

For most applications, MCA yields rather similar results to the CCA (with PCA pre-filtering) (Bretherton et al., 1992; Wallace et al., 1992).

## CCA functions in Matlab

`www.mathworks.com/help/toolbox/stats/canoncorr.html`

[A, B, rho, U, V] = canoncorr(X,Y)

X and Y are the *transpose* of my data matrices, the transpose of U and V have columns giving the vectors **u** and **v**, respectively, and the transpose of A ad B are the matrices $\mathcal{F}$ and $\mathcal{G}$, respectively.

## References:

Barnett, T. P. and Preisendorfer, R. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, 115(9):1825–1850.

Bretherton, C. S., Smith, C., and Wallace, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, 5:541–560.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28:321–377.

Strang, G. (2005). *Linear Algebra and Its Applications*. Cole Brooks.

von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*. Cambridge Univ. Pr., Cambridge.

Wallace, J. M., Smith, C., and Bretherton, C. S. (1992). Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J Climate*, 5(6):561–576.