GI12/4C59: Information Theory

Lectures 28-30: Eigenvalue methods for signal representation and compression

Massimiliano Pontil

About these lectures

Theme: We introduce canonical correlation analysis (CCA), a method for representing an input/output signal by a smaller dimensional vector which captures most correlations in the data. When the data distribution is Gaussian, CCA is equivalent to maximize the mutual information of the input and output. We also discuss and contrast CCA to other eigenvalue methods such as principal component analysis and partial least squares.

Outline

1. Canonical correlation analysis
2. Generalized eigenvalue problem
3. Principal component analysis
4. Partial least squares
5. Reduced rank regression

Statement of the problem

Consider two jointly input/output r.v. $(X, Y) \sim p(\mathbf{x}, \mathbf{y})$, where $\mathbf{x} \in \mathbb{R}^{\ell}$, $\mathbf{y} \in \mathbb{R}^{q}$.

We wish to find a description of both the input and output which

- can efficiently represent/compress these signals.
- reveal "interesting" relations/features among inputs and outputs.
- be used for supervised learning.

We focus on **linear** descriptions. The approaches we discuss consist in iteratively searching for a pair of vectors $\mathbf{w}_{\mathbf{x}} \in \mathbb{R}^{\ell}$, $\mathbf{w}_{\mathbf{y}} \in \mathbb{R}^{q}$ which maximize an objective function.

Correlation

We always assume X and Y have been preprocessed to have zero mean.

Our first method searches for a pair of vectors $\mathbf{w_x} \in \mathbb{R}^\ell$, $\mathbf{w_y} \in \mathbb{R}^q$ which maximize

$$corr(\mathbf{w_x}^\top \mathbf{x}, \mathbf{w_y}^\top \mathbf{y}) = \frac{E[(\mathbf{w_x}^\top \mathbf{x})(\mathbf{w_y}^\top \mathbf{y})]}{\sqrt{E[(\mathbf{w_x}^\top \mathbf{x})^2] E[(\mathbf{w_y}^\top \mathbf{y})^2]}}$$

where $corr(\cdot, \cdot)$ denotes the correlation between two r.v.. This is a number in the interval $[-1, 1]$.

Remark: to simplify the notation, we have denoted by $E[\mathbf{x}]$ (and not $E[X]$) the average of a random variable X. Below, we also denote by $\widehat{\mathbf{w}}_\mathbf{x}$ and $\widehat{\mathbf{w}}_\mathbf{y}$ the normalized unit vectors.

Mutual information

Suppose (X, Y) is normally distributed,

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp(-(\mathbf{x}, \mathbf{y}) C^{-1} (\mathbf{x}, \mathbf{y})^\top)$$

where

$$C = \begin{bmatrix} C_{\mathbf{xx}} & C_{\mathbf{xy}} \\ C_{\mathbf{yx}} & C_{\mathbf{yy}} \end{bmatrix}.$$

Then, the mutual information between X and Y is

$$I(X; Y) = \frac{1}{2} \ln \left(\frac{|C_{\mathbf{xx}}||C_{\mathbf{yy}}|}{|C|} \right).$$

To see this, use the formula for the differential entropy of a Gaussian distributed vector-valued r.v. with covariance K,

$$H(X) = \frac{1}{2} \ln ((2\pi e)^n |K|) \qquad \text{(here } |K| = det(K)\text{)}$$

and the property $I(X; Y) = H(X) + H(Y) - H(X, Y)$.

Correlation and mutual information (scalar case)

For scalar inputs/outputs ($\ell = q = 1$), we have

$$C = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

where (remember X, Y are zero mean) $\sigma_x^2 = E[X^2]$, $\sigma_y^2 = E[Y^2]$, $\sigma_{xy} = E[XY]$.

In this case

$$\rho_{xy} = \frac{E[XY]}{\sqrt{E[X^2]E[Y^2]}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

is related to the mutual information of X and Y,

$$I(X;Y) = \frac{1}{2} \ln \left(\frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2} \right) = \frac{1}{2} \ln \left(\frac{1}{1 - \rho_{xy}^2} \right)$$

Thus maximizing the correlation is equivalent to maximize the mutual information.

Canonical correlation analysis (CCA)

In practice $p(\mathbf{x}, \mathbf{y})$ is an unknown law of nature.

We wish to discover $\mathbf{w_x}, \mathbf{w_y}$ based on a finite sample D of n i.i.d. pairs $(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})$,

$$D = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}.$$

We can approximate (estimate) the above correlation by

$$corr(\mathbf{w_x}^\top \mathbf{x}, \mathbf{w_y}^\top \mathbf{y}) = \frac{E[(\mathbf{w_x}^\top \mathbf{x})(\mathbf{w_y}^\top \mathbf{y})]}{\sqrt{E[\mathbf{w_x}^\top \mathbf{x}]}\sqrt{E[\mathbf{w_y}^\top \mathbf{y}]}} \approx \frac{\sum_{i=1}^n (\mathbf{w_x}^\top \mathbf{x}_i)(\mathbf{w_y}^\top \mathbf{y}_i)}{\sqrt{\sum_{i=1}^n (\mathbf{w_x}^\top \mathbf{x}_i)^2 \sum_{i=1}^n (\mathbf{w_y}^\top \mathbf{y}_i)^2}}.$$

Some algebra...

We denote by $\widehat{E}$ the expectation w.r.t. the empirical distribution on the sample D and observe that

$$corr(\mathbf{w_x}^\top \mathbf{x}, \mathbf{w_y}^\top \mathbf{y}) \approx \frac{\widehat{E}[(\mathbf{w_x}^\top \mathbf{x})(\mathbf{w_y}^\top \mathbf{y})]}{\sqrt{\widehat{E}[(\mathbf{w_x}^\top \mathbf{x})^2]\widehat{E}[(\mathbf{w_y}^\top \mathbf{y})^2]}}$$

$$= \frac{\mathbf{w_x}^\top \widehat{E}[\mathbf{x}\mathbf{y}^\top]\mathbf{w_y}}{\sqrt{(\mathbf{w_x}^\top \widehat{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{w_x})(\mathbf{w_y}^\top \widehat{E}[\mathbf{y}\mathbf{y}^\top]\mathbf{w_y})}} = \frac{\mathbf{w_x}^\top C_{\mathbf{xy}}\mathbf{w_y}}{\sqrt{(\mathbf{w_x}^\top C_{\mathbf{xx}}\mathbf{w_x})(\mathbf{w_y}^\top C_{\mathbf{yy}}\mathbf{w_y})}}$$

Hence, the problem we wish to solve is

$$\max_{\mathbf{w_x},\mathbf{w_y}} \frac{\mathbf{w_x}^\top C_{\mathbf{xy}}\mathbf{w_y}}{\sqrt{(\mathbf{w_x}^\top C_{\mathbf{xx}}\mathbf{w_x})(\mathbf{w_y}^\top C_{\mathbf{yy}}\mathbf{w_y})}}$$

Solution

We observe that the solution of (*) or (**) is not affected by rescaling $\mathbf{w_x}$ and/or $\mathbf{w_y}$. In order to have a unique solution we maximize the numerator under a normalization constraint, that is

$$\max_{\mathbf{w_x},\mathbf{w_y}} \{\mathbf{w_x}^\top C_{\mathbf{xy}}\mathbf{w_y} \ : \ \mathbf{w_x}^\top C_{\mathbf{xx}}\mathbf{w_x} = 1, \ \mathbf{w_y}^\top C_{\mathbf{yy}}\mathbf{w_y} = 1\}$$

The corresponding Lagrangian function is

$$\mathcal{L}(\mathbf{w_x}, \mathbf{w_y}, \lambda_\mathbf{x}, \lambda_\mathbf{y}) = \mathbf{w_x}^\top C_{\mathbf{xy}}\mathbf{w_y} - \frac{\lambda_\mathbf{x}}{2}(\mathbf{w_x}^\top C_{\mathbf{xx}}\mathbf{w_x} - 1) - \frac{\lambda_\mathbf{y}}{2}(\mathbf{w_y}^\top C_{\mathbf{yy}}\mathbf{w_y} - 1)$$

and the Euler-Lagrange equations are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w_x}} = C_{\mathbf{xy}}\mathbf{w_y} - \lambda_\mathbf{x} C_{\mathbf{xx}}\mathbf{w_x} = 0, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{w_y}} = C_{\mathbf{yx}}\mathbf{w_x} - \lambda_\mathbf{y} C_{\mathbf{yy}}\mathbf{w_y} = 0.$$

Solution (cont.)

Using

$$C_{xy}\mathbf{w}_y - \lambda_x C_{xx}\mathbf{w}_x = 0 \qquad (1)$$

$$C_{yx}\mathbf{w}_x - \lambda_y C_{yy}\mathbf{w}_y = 0 \qquad (2)$$

and the normalization constraints, we obtain

$$\lambda_x = \lambda_y =: \lambda.$$

Now, assuming C_{yy} is non-singular, (1) gives

$$\mathbf{w}_y = \frac{C_{yy}^{-1}C_{yx}\mathbf{w}_x}{\lambda} \qquad (\diamond)$$

and, placing ($\diamond$) in equation (2), we obtain

$$C_{xy}C_{yy}^{-1}C_{yx}\mathbf{w}_x = \lambda^2 C_{xx}\mathbf{w}_x \qquad (\diamond\ \diamond)$$

Generalized eigenvalue problem

($\diamond\ \diamond$) is an example of generalized eigenvalue problem

- Let A, B be squared matrices and B positive definite. We are interested in the problem

  $$A\mathbf{w} = \lambda B\mathbf{w}$$

  We can efficiently transform this problem into a standard eigenvalue problem by setting $B = R^\top R$ (e.g., using a Cholesky decomposition where R is lower triangular) and $\mathbf{u} = R\mathbf{w}$,

  $$A\mathbf{w} = \lambda R^\top R\mathbf{w} \;\Rightarrow\; AR^{-1}R\mathbf{w} = \lambda R^\top(R\mathbf{w}) \;\Rightarrow\; (R^{-1})^\top AR^{-1}\mathbf{u} = \lambda\mathbf{u}.$$

Thus, if we write $C_{xx} = R_{xx}^\top R_{xx}$ and $R_{xx}\mathbf{w}_x = \mathbf{u}$, eq. ($\diamond\diamond$) becomes

$$(R_{xx}^{-1})^\top C_{xy}C_{yy}^{-1}C_{yx}R_{xx}^{-1}\mathbf{u} = \lambda^2\mathbf{u}$$

Once we have found $\mathbf{w}_x$, we simply compute $\mathbf{w}_y$ by ($\diamond$).

Generalized eigenvalue problem (cont)

The generalized eigenvalue problem is closely related to finding the optimal points of the *Rayleigh quotient*

$$r(\mathbf{w}) = \frac{\mathbf{w}^\top A\mathbf{w}}{\mathbf{w}^\top B\mathbf{w}}.$$

In fact,

$$\frac{\partial r}{\partial \mathbf{w}} = \frac{2}{\mathbf{w}^\top B\mathbf{w}}(A\mathbf{w} - rB\mathbf{w}) = 0 \;\Rightarrow\; A\mathbf{w} = rB\mathbf{w}$$

which is like the generalized eigenvalue problem with $r = \lambda$.

Generalized eigenvalue problem (cont.)

The above observations tell us that the optimal values/points of the Rayleigh quotient are the eigen-values/vectors of the generalized eigenvalue problem.

Let $\{(\lambda_i, \mathbf{w}_i)\}_{i=1}^n$ be such a system of eigen-values/vectors with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. Then, it is easy to see that

- λ_1 is a global maximum for r. If $\lambda_1 > \lambda_2$, this maximum is unique (modulo a scale transformation).
- λ_n is a global minimum for r. This is unique if $\lambda_{n-1} > \lambda_n$
- All extremum values for r except λ_1 and λ_n are saddle points. This can verified by inspecting the Hessian matrix

$$H_{ij} = \frac{\partial^2 r}{\partial w_i \partial w_j}.$$

Finding the largest eigenvalue

A solution can be found by means of the following iterative algorithm

- Set $\mathbf{w} = \mathbf{w}^{(0)}$ and choose $\eta > 0$.
- For $t = 1, \ldots, T$
  1. Compute $\nabla r(\mathbf{w}) = \frac{2}{\mathbf{w}^\top B\mathbf{w}}(A\mathbf{w} - rB\mathbf{w})$.
  2. Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\nabla r(\mathbf{w}^{(t)})$.

If η is not too large and T large enough (maybe infinite?), this algorithm converges to the largest eigenvalue $\lambda_1 = \max_{\mathbf{w}} r(\mathbf{w})$ and $\mathbf{w}^{(t)}$ converges to a corresponding eigenvector.

Alternatively, we can use the updating rule:

$$\eta(A\hat{\mathbf{w}} - B\mathbf{w}).$$

In this case we also have that $\|\mathbf{w}^{(t)}\| \to \lambda_1$.

Finding the next eigenvalues

The next eigenvalues/vectors can be found iteratively by applying the same algorithm to a reduced matrix A.

Let λ_1 be the largest eigenvalue with adn $\mathbf{w}_1$ a corresponding eigenvector. Assume $\mathbf{w}_1$ has norm 1 in the metric induced by B (ie, $\mathbf{w}_1^\top B\mathbf{w}_1 = 1$) and set $\mathbf{u}_1 = R\mathbf{w}_1$.

We subtract the contribution of the largest eigenvector to the eq. $(R^{-1})^\top AR^{-1}\mathbf{u}_2 = \lambda_2\mathbf{u}_2$, obtaining

$$(R^{-1})^\top AR^{-1} \to (R^{-1})^\top AR^{-1} - \lambda_1\mathbf{u}_1\mathbf{u}_1^\top.$$

Then, using $\mathbf{u}_1 = R\mathbf{w}_1$, after some algebra we see that this transformation is equivalent to keeping B unchanged and transforming A as

$$A \to A - \lambda_1 B\mathbf{w}_1\mathbf{w}_1^\top B.$$

Remark: It is also possible to write the above transformation using the normalization $\|\mathbf{w}_1\| = \lambda_1$.

Back to CCA

We got a full system of generalized eigenvectors $\mathbf{w}_{\mathbf{x}j} \in \mathbb{R}^{\ell}$, $\mathbf{w}_{\mathbf{y}j} \in \mathbb{R}^{q}$ and eigenvalues $\lambda_j > 0$, $j = 1, \ldots, N = \min(\ell, q)$ such that

$$\mathbf{w}_{\mathbf{x}i} C_{\mathbf{xy}} \mathbf{w}_{\mathbf{y}i} = \lambda_i, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$$

with $\mathbf{w}_{\mathbf{x}i} C_{\mathbf{xx}} \mathbf{w}_{\mathbf{x}j} = \mathbf{w}_{\mathbf{y}i} C_{\mathbf{yy}} \mathbf{w}_{\mathbf{y}j} = \delta_{ij}$, $i, j = 1, \ldots, N$ (here $\delta_{ij} = 1$ is the delta of Kronecker).

- $\mathbf{w}_{\mathbf{x}1}, \mathbf{w}_{\mathbf{y}1}$ is the solution to the CCA problem.
- $\mathbf{w}_{\mathbf{x}2}, \mathbf{w}_{\mathbf{y}2}$ is the solution to the same problem when the directions $\mathbf{w}_{\mathbf{x}1}, \mathbf{w}_{\mathbf{y}1}$ are deflated from the inputs and output respectively.
- In general, $\mathbf{w}_{\mathbf{x}(j+1)}, \mathbf{w}_{\mathbf{y}(j+1)}$ is the solution when the data components along the directions $\mathbf{w}_{\mathbf{x}1}, \ldots, \mathbf{w}_{\mathbf{x}j}$ and $\mathbf{w}_{\mathbf{y}1}, \ldots, \mathbf{w}_{\mathbf{y}j}$ are deflated from the inputs and output respectively.

Alternative formulation

Suppose we have found all canonical pairs $\mathbf{w}_{\mathbf{x}i}, \mathbf{w}_{\mathbf{y}i}$, $i = 1, \ldots, j-1$. Then, the $j-$th canonical pair, $(\mathbf{w}_{\mathbf{x}j}, \mathbf{w}_{\mathbf{y}j})$, is the solution to the problem

$$\text{maximize } \mathbf{w}_{\mathbf{x}}{}^{\top} C_{\mathbf{xy}} \mathbf{w}_{\mathbf{y}}$$

subject to the constraint, for $i = 1, \ldots, j-1$, that

$$\mathbf{w}_{\mathbf{x}i}^{\top} C_{\mathbf{xx}} \mathbf{w}_{\mathbf{x}j} = 0, \quad \mathbf{w}_{\mathbf{y}i}^{\top} C_{\mathbf{yy}} \mathbf{w}_{\mathbf{y}j} = 0, \quad i \neq j$$

$$\mathbf{w}_{\mathbf{x}}^{\top} C_{\mathbf{xx}} \mathbf{w}_{\mathbf{x}i} = 1, \quad \mathbf{w}_{\mathbf{y}i}^{\top} C_{\mathbf{yy}} \mathbf{w}_{\mathbf{y}i} = 1.$$

CCA and mutual information

The canonical correlations are related to mutual information by the formula

$$I(X;Y) = \frac{1}{2} \ln \left(\frac{1}{\prod_i (1-\lambda_i^2)} \right).$$

To see this, remember we saw before that

$$I(X;Y) = \frac{1}{2} \ln \left(\frac{|C_{\mathrm{xx}}||C_{\mathrm{yy}}|}{|C|} \right)$$

and use the formula

$$|C| = |C_{\mathrm{xx}}||C_{\mathrm{yy}} - C_{\mathrm{yx}} C_{\mathrm{xx}}^{-1} C_{\mathrm{xy}}|$$

to obtain

$$I(X;Y) = -\frac{1}{2} \ln \left(|I - C_{\mathrm{yy}}^{-1} C_{\mathrm{yx}} C_{\mathrm{xx}}^{-1} C_{\mathrm{xy}}| \right).$$

where we assumed C_{xx} and C_{yy} are nonsingular.

CCA and mutual information (cont.)

$$I(X;Y) = -\frac{1}{2} \ln \left(|I - C_{\mathrm{yy}}^{-1} C_{\mathrm{yx}} C_{\mathrm{xx}}^{-1} C_{\mathrm{xy}}| \right).$$

We now notice that the eigenvalues of $C_{\mathrm{yy}}^{-1} C_{\mathrm{yx}} C_{\mathrm{xx}}^{-1} C_{\mathrm{xy}}$ are the squared canonical correlations, λ_i, and, since the determinant of a matrix is invariant w.r.t. orthogonal transformations, we obtain

$$\begin{aligned} I(X;Y) &= -\frac{1}{2} \ln \left| diag(1-\lambda_i^2) \right| \\ &= -\frac{1}{2} \ln \left(\prod_i (1-\lambda_i^2) \right) = \frac{1}{2} \ln \left(\frac{1}{\prod_i (1-\lambda_i^2)} \right). \end{aligned}$$

Regularization

If $C_{\mathbf{xx}}$ and/or $C_{\mathbf{yy}}$ are singular (or simply ill-conditioned) we can replace them by $C_{\mathbf{xx}} + \alpha_{\mathbf{x}} I$ and $C_{\mathbf{yy}} + \alpha_{\mathbf{y}} I$, where $\alpha_{\mathbf{x}}, \alpha_{\mathbf{y}} > 0$ and I is the identity matrix with the appropriate dimension.

If the regularization parameters $\alpha_{\mathbf{x}}, \alpha_{\mathbf{y}}$ are chosen with care (e.g. by cross-validation) we will find interesting directions, that is directions which capture not only high sample-correlations, but also true correlations

This fact can be justified using a probabilistic analysis...

Principal component analysis (PCA)

This method seeks for directions along which the variance of an input signal X is maximized,

$$\rho = \frac{\mathbf{w_x}^\top C_{\mathbf{xx}} \mathbf{w_x}}{\mathbf{w_x}^\top \mathbf{w_x}}$$

This is a Rayleigh quotient with $A = C_{\mathbf{xx}}$ and $B = I$.

Since $C_{\mathbf{xx}}$ is symmetric, we have

$$C_{\mathbf{xx}} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

where $C_{\mathbf{xx}} \mathbf{u}_i = \lambda_i \mathbf{u}_i$ and $\lambda_i \geq 0$ since $C_{\mathbf{xx}}$ is positive definite.

Thus, if $\lambda_1 > \lambda_i$, $i = 2, \ldots, n$,

$$\max \rho = \lambda_1, \text{ and } \operatorname{argmax} \rho = \mathbf{u}_1.$$

PCA (cont.)

The direction with maximum variance is also that one along which the mean squared error is minimized

$$\frac{1}{2}E[\|\mathbf{x} - (\mathbf{x}^\top \hat{\mathbf{w}}_\mathbf{x})\hat{\mathbf{w}}_\mathbf{x}\|^2] = E[\mathbf{x}^\top\mathbf{x} - \hat{\mathbf{w}}_\mathbf{x}{}^\top \mathbf{x}\mathbf{x}^\top \hat{\mathbf{w}}_\mathbf{x}] = trace(C_{\mathbf{xx}}) - \frac{\mathbf{w}_\mathbf{x} C_{\mathbf{xx}} \mathbf{w}_\mathbf{x}}{\mathbf{w}_\mathbf{x}{}^\top \mathbf{w}_\mathbf{x}}.$$

By examining the updating rule we see that

$$\mathbf{w}_\mathbf{x}^{(t+1)} = \mathbf{w}_\mathbf{x}^{(t)} + \eta(C_{\mathbf{xx}}\mathbf{w}_\mathbf{x} - \mathbf{w}_\mathbf{x})$$

We can also use a stochastic rule

$$\mathbf{w}_\mathbf{x}^{(t+1)} = \mathbf{w}_\mathbf{x}^{(t)} + \eta(\mathbf{x}_{i(t)}\mathbf{x}_{i(t)}^\top \mathbf{w}_\mathbf{x} - \mathbf{w}_\mathbf{x})$$

where $i(t)$ is a randomly chosen sample at iteration t.

PCA (cont.)

The second direction with largest variance which is orthogonal to $\mathbf{w}_1$ is obtained by subtracting the direction $\mathbf{w}_1$ from the data and applying the above algorithm to it,

$$C_{\mathbf{xx}} \rightarrow C_{\mathbf{xx}} - \lambda_1 \mathbf{w}_1 \mathbf{w}_1^\top.$$

Problems with PCA:

- Mean squared error may be a bad "distortion measure".
- PCA is not a good approach to analyze input/output relations: If we compute separate PCA for the inputs and outputs, then components with high variance may be irrelevant in describing input/output statistics whilst discarded directions with possibly very small variance may be informative.

Partial least squares (PLS)

A better method, alternative to CCA, to describe joint sources is to find *simultaneously* interesting input/output directions which maximize a joint error measure.

Similarly to PCA, we choose a covariance error measure,

$$\rho = E[\hat{\mathbf{w}}_{\mathbf{y}}{}^{\top}\mathbf{x}\hat{\mathbf{w}}_{\mathbf{x}}{}^{\top}\mathbf{y}] = \frac{\mathbf{w}_{\mathbf{y}}{}^{\top}C_{\mathbf{xy}}\mathbf{w}_{\mathbf{x}}}{\sqrt{\mathbf{w}_{\mathbf{y}}{}^{\top}\mathbf{w}_{\mathbf{y}}\mathbf{w}_{\mathbf{x}}{}^{\top}\mathbf{w}_{\mathbf{x}}}}.$$

By differentiating w.r.t. $\mathbf{w}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{y}}$, we get the system of equations

$$\begin{aligned} C_{\mathbf{xy}}\mathbf{w}_{\mathbf{y}} &= \rho\mathbf{w}_{\mathbf{x}} \\ C_{\mathbf{yx}}\mathbf{w}_{\mathbf{x}} &= \rho\mathbf{w}_{\mathbf{y}} \end{aligned}$$

which, in turn, give

$$\begin{aligned} C_{\mathbf{xy}}C_{\mathbf{yx}}\mathbf{w}_{\mathbf{x}} &= \rho^2\mathbf{w}_{\mathbf{x}} \\ C_{\mathbf{yx}}C_{\mathbf{xy}}\mathbf{w}_{\mathbf{y}} &= \rho^2\mathbf{w}_{\mathbf{y}}. \end{aligned}$$

PLS (cont.)

It can be verified that PLS is a Rayleigh problem if we set

$$A = \begin{bmatrix} 0 & C_{\mathbf{xy}} \\ C_{\mathbf{yx}} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & C_{\mathbf{xy}} \\ C_{\mathbf{yx}} & 0 \end{bmatrix}, \quad \mathbf{w} = \begin{pmatrix} \mu_{\mathbf{x}}\hat{\mathbf{w}}_{\mathbf{x}} \\ \mu_{\mathbf{y}}\hat{\mathbf{w}}_{\mathbf{y}} \end{pmatrix}$$

where $\frac{\mu_{\mathbf{x}}}{\mu_{\mathbf{y}}} = \pm 1$.

To find $\mathbf{w}_{\mathbf{x}}$ and $\mathbf{w}_{\mathbf{y}}$ we can compute the SVD for $C_{\mathbf{xy}}$,

$$C_{\mathbf{xy}} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_{\mathbf{x}i}\mathbf{u}_{\mathbf{y}i}^{\top}, \quad \text{or, in matrix notation, } C_{\mathbf{xy}} = U_{\mathbf{x}}\Lambda U_{\mathbf{y}}^{\top}.$$

Since the basis vectors are orthogonal, the problem of maximizing ρ is equivalent to finding the largest singular value, λ_1, that is

$$\lambda_1 = \max_{\mathbf{w}_{\mathbf{y}},\mathbf{w}_{\mathbf{y}}} \rho(\mathbf{w}_{\mathbf{y}},\mathbf{w}_{\mathbf{y}}), \quad (\mathbf{w}_{\mathbf{x},1},\mathbf{w}_{\mathbf{y},1}) = \arg\max_{\mathbf{w}_{\mathbf{y}},\mathbf{w}_{\mathbf{y}}} \rho(\mathbf{w}_{\mathbf{y}},\mathbf{w}_{\mathbf{y}}).$$

Multivariate linear regression (MLR)

The last method we discuss, seeks for directions $\widehat{\mathbf{w}}_{\mathbf{x}}$ and $\widehat{\mathbf{w}}_{\mathbf{y}}$ which minimize a rank-one mean squared regression error

$$\epsilon^2 = \min_a E[\|\mathbf{y} - a\mathbf{x}^\top \widehat{\mathbf{w}}_{\mathbf{x}} \widehat{\mathbf{w}}_{\mathbf{y}}\|^2].$$

The minimizing a is

$$a = \frac{\widehat{\mathbf{w}}_{\mathbf{x}}^\top C_{\mathbf{xy}} \widehat{\mathbf{w}}_{\mathbf{y}}}{\widehat{\mathbf{w}}_{\mathbf{x}}^\top C_{\mathbf{xx}} \widehat{\mathbf{w}}_{\mathbf{x}}} \quad \Rightarrow \quad \epsilon^2 = E[\mathbf{y}^\top \mathbf{y}] - \frac{(\widehat{\mathbf{w}}_{\mathbf{x}}^\top C_{\mathbf{xy}} \widehat{\mathbf{w}}_{\mathbf{y}})^2}{\widehat{\mathbf{w}}_{\mathbf{x}}^\top C_{\mathbf{xx}} \widehat{\mathbf{w}}_{\mathbf{x}}}$$

which is equivalent to maximize

$$\rho(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_{\mathbf{y}}) = \frac{\mathbf{w}_{\mathbf{x}}{}^\top C_{\mathbf{xy}} \mathbf{w}_{\mathbf{y}}}{\sqrt{\mathbf{w}_{\mathbf{x}}{}^\top C_{\mathbf{xx}} \mathbf{w}_{\mathbf{x}} \mathbf{w}_{\mathbf{y}}{}^\top \mathbf{w}_{\mathbf{y}}}}.$$

MLR (cont.)

As before we can differentiate the criterion ρ w.r.t $\widehat{\mathbf{w}}_{\mathbf{x}}, \widehat{\mathbf{w}}_{\mathbf{y}}$ and, after some algebra, we obtain

$$C_{\mathbf{xx}}^{-1} C_{\mathbf{xy}} C_{\mathbf{yx}} \widehat{\mathbf{w}}_{\mathbf{x}} = \rho^2 \widehat{\mathbf{w}}_{\mathbf{x}}$$

$$C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} C_{\mathbf{xy}} \widehat{\mathbf{w}}_{\mathbf{y}} = \rho^2 \widehat{\mathbf{w}}_{\mathbf{y}}.$$

This is equivalent to the Rayleigh problem with

$$A = \begin{bmatrix} 0 & C_{\mathbf{xy}} \\ C_{\mathbf{yx}} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} C_{\mathbf{xx}} & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{w} = \begin{pmatrix} \mu_{\mathbf{x}} \widehat{\mathbf{w}}_{\mathbf{x}} \\ \mu_{\mathbf{y}} \widehat{\mathbf{w}}_{\mathbf{y}} \end{pmatrix}$$

where $\frac{\mu_{\mathbf{x}}}{\mu_{\mathbf{y}}} = \pm 1$.

MLR (cont.)

We can also compute an $N-$rank regression by optimizing the regression coefficients $a = (a_1, \ldots, a_n)$ along the N directions,

$$\rho_N(\widehat{\mathbf{w}}_{\mathbf{x}1}, \widehat{\mathbf{w}}_{\mathbf{y}1}, \ldots, \widehat{\mathbf{w}}_{\mathbf{x}N}, \widehat{\mathbf{w}}_{\mathbf{y}N}) = \min_a E[\|\mathbf{y} - \sum_{i=1}^{N} a_i \mathbf{x}^\top \widehat{\mathbf{w}}_{\mathbf{x}i} \widehat{\mathbf{w}}_{\mathbf{y}i}\|^2].$$

It can be verified that the solution are the first N eigenvectors corresponding to the N largest eigenvalues of the function ρ above.

From linear to nonlinear methods

- Linear projections are very restrictive. Often data are not Gaussian distributed!

Two possible ways to deal with nonlinearities are

1. Use neural networks.
2. Use kernel functions.

Let $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$ the two neural networks

$$f_{\mathbf{x}}(\mathbf{x}) = h(\sum_{k=1}^{N_{\mathbf{x}}} v_{\mathbf{x}k} h(\sum_{i=1}^{\ell} \mathbf{w}_{\mathbf{x}k}^\top \mathbf{x})), \qquad f_{\mathbf{y}}(\mathbf{y}) = h(\sum_{k=1}^{N_{\mathbf{y}}} v_{\mathbf{y}k} h(\sum_{i=1}^{\ell} \mathbf{w}_{\mathbf{y}k}^\top \mathbf{y}))$$

where h is an activation function.

The network parameters can for example be computed by maximizing their mutual information. If there is not hidden layer this is like CCA. (for more information, see the work of Becker and Hinton (1996))

From linear to nonlinear methods (cont.)

The main idea in the second approach is to substitute $\mathbf{x}$ by a nonlinear vector-valued function

$$\varphi : \mathbb{R}^\ell \to H$$

and $\mathbf{x}_1^\top \mathbf{x}_2$ by $K(\mathbf{x}_1, \mathbf{x}_2) = \varphi(\mathbf{x}_1)^\top \varphi(\mathbf{x}_2)$. Here H is an euclidean space, possibly infinite dimensional.

- This approach can still be seen as a generalized eigenvalue problem.

For example, if $\mathbf{x} = (x_1, x_2),\ \mathbf{t} = (t_1, t_2)$ then $K(\mathbf{x}, \mathbf{t}) = (\mu + \mathbf{x}^\top \mathbf{t})^2 - 1,\ \mu \geq 0$ corresponds to

$$\varphi(\mathbf{x}) = (\sqrt{2}\mu x_1, \sqrt{2}\mu x_2, x_1^2, x_2^2, \sqrt{2} x_1 x_2).$$

Mathematically, this approach consists in replacing the spaces $\mathbb{R}^\ell$ and $\mathbb{R}^q$ by two reproducing kernel Hilbert spaces with kernels $K_\mathbf{x}$ and $K_\mathbf{y}$.

Bibliography

M. Borga, *Learning multidimensional signal processing*, PhD thesis, Dept of Electrical Engineering, Linköping University, Sweden 1998 D.R. Available at *http://people.imt.liu.se/∼magnus/publications.html*

See also:

1. Hardoon, S. Szedmak, J. Shaw-Taylor. Canonical correlation analysis; An overview with application to learning methods. Tech. report CSD-TR-03-02, Royal Holloway University of London, 2003.

2. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001. (chapter 3.4 and 14.5)