MTTTS16 Learning from Multiple Sources

5 ECTS credits

Autumn 2019, University of Tampere

Lecturer: Jaakko Peltonen

Lecture 4: Kernel CCA and other variants

On this lecture:

- Probabilistic canonical correlation analysis
- Nonlinear canonical correlation analysis through a “kernel trick”
- Variants of canonical correlation analysis

Canonical Correlation Analysis, recap

Reminder: CCA finds projections of two simultaneously observed data sources (two feature sets for the same samples) so that the projections are maximally correlated.

Used in many tasks and data domains.

Canonical Correlation Analysis, recap

- For **x**, find a projection $w_{x,1}x_1 + w_{x,2}x_2 + ... + w_{x,K}x_K$ where $\mathbf{w_x} = [w_{x,1}, w_{x,2},...,w_{x,K}]$ is the projection basis.
- For **y**, find a projection $w_{y,1}y_1 + w_{y,2}y_2 + ... + w_{y,L}y_L$ where $\mathbf{w_y} = [w_{y,1}, w_{y,2},...,w_{y,L}]$ is the projection basis.
- Find the projection bases by maximizing the correlation between the projections: maximize

$$corr(\mathbf{w}_x^T\mathbf{x}, \mathbf{w}_y^T\mathbf{y}) = \frac{E[\mathbf{w}_x^T\mathbf{x}\,\mathbf{w}_y^T\mathbf{y}]}{(E[(\mathbf{w}_x^T\mathbf{x})^2]E[(\mathbf{w}_y^T\mathbf{y})^2])^{1/2}}$$

with respect to $\mathbf{w_x}$ and $\mathbf{w_y}$.

This definition assumes x and y are zero-mean, otherwise substract the means as in the original correlation definition.

- For a finite data set: maximize the sample correlation

$$\hat{corr}(\mathbf{w}_x^T\mathbf{x}, \mathbf{w}_y^T\mathbf{y}) = \frac{\hat{E}_{ML}[\mathbf{w}_x^T\mathbf{x}\,\mathbf{w}_y^T\mathbf{y}]}{(\hat{E}_{ML}[(\mathbf{w}_x^T\mathbf{x})^2]\hat{E}_{ML}[(\mathbf{w}_y^T\mathbf{y})^2])^{1/2}}$$

Same definition as before

$$\hat{E}_{ML}[x\,y] = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$$

Canonical Correlation Analysis, recap

- CCA can be solved as a generalized eigenvalue equation

$$\hat{C}_{x,y}\hat{C}_{y}^{-1}\hat{C}_{y,x}\boldsymbol{w}_x=\lambda^2\hat{C}_x\boldsymbol{w}_x$$

$$\boldsymbol{w}_y=(1/\lambda)\hat{C}_{y}^{-1}\hat{C}_{y,x}\boldsymbol{w}_x$$

- This is a generalized eigenvalue equation which we can solve to get $\mathbf{w}_x$, and the previous equation then gives $\mathbf{w}_y$ from $\mathbf{w}_x$.

Part 1: Probabilistic Canonical Correlation Analysis

Canonical Correlation Analysis, recap

Reminder: CCA finds projections of two simultaneously observed data sources (two feature sets for the same samples) so that the projections are maximally correlated.

Used in many tasks and data domains.

Canonical Correlation Analysis, recap

- For **x**, find a projection $w_{x,1}x_1 + w_{x,2}x_2 + \ldots + w_{x,K}x_K$ where $\mathbf{w}_x = [\, w_{x,1}, w_{x,2}, \ldots, w_{x,K} \,]$ is the projection basis.
- For **y**, find a projection $w_{y,1}y_1 + w_{y,2}y_2 + \ldots + w_{y,L}y_L$ where $\mathbf{w}_y = [\, w_{y,1}, w_{y,2}, \ldots, w_{y,L} \,]$ is the projection basis.
- Find the projection bases by maximizing the correlation between the projections: maximize

$$corr\left(\mathbf{w}_x^T \mathbf{x}, \mathbf{w}_y^T \mathbf{y}\right) = \frac{E\left[\mathbf{w}_x^T \mathbf{x}\, \mathbf{w}_y^T \mathbf{y}\right]}{\left(E\left[\left(\mathbf{w}_x^T \mathbf{x}\right)^2\right] E\left[\left(\mathbf{w}_y^T \mathbf{y}\right)^2\right]\right)^{1/2}}$$

with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$.

This definition assumes x and y are zero-mean, otherwise substract the means as in the original correlation definition.

- For a finite data set: maximize the sample correlation

$$\hat{corr}\left(\mathbf{w}_x^T \mathbf{x}, \mathbf{w}_y^T \mathbf{y}\right) = \frac{\hat{E}_{ML}\left[\mathbf{w}_x^T \mathbf{x}\, \mathbf{w}_y^T \mathbf{y}\right]}{\left(\hat{E}_{ML}\left[\left(\mathbf{w}_x^T \mathbf{x}\right)^2\right] \hat{E}_{ML}\left[\left(\mathbf{w}_y^T \mathbf{y}\right)^2\right]\right)^{1/2}}$$

Same definition as before

$$\hat{E}_{ML}[x\, y] = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$$

Canonical Correlation Analysis, recap

- CCA can be solved as a generalized eigenvalue equation

$$\hat{C}_{x,y}\hat{C}_{y}^{-1}\hat{C}_{y,x}\boldsymbol{w}_x=\lambda^2\hat{C}_x\boldsymbol{w}_x$$

$$\boldsymbol{w}_y=(1/\lambda)\hat{C}_{y}^{-1}\hat{C}_{y,x}\boldsymbol{w}_x$$

- This is a generalized eigenvalue equation which we can solve to get $\mathbf{w}_x$, and the previous equation then gives $\mathbf{w}_y$ from $\mathbf{w}_x$.

CCA, probabilistic interpretation, motivation

- Probabilistic models are descriptions of data distributions (underlying observed data sets)
- Properties that are strongly connected to a probabilistic model are motivated by the properties of that model (if the model is a good model for data, then the properties involved in the model are likely to be useful).
- Additionally, probabilistic models can be estimated and analyzed in many ways (using all tools of probability theory)
- ------> it is useful to connect the things we compute from data to probabilistic models.
- Can CCA be seen as a probabilistic model for the distribution of data in some data set? Yes!

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation, motivation

- Principal component analysis (PCA) has been shown to be the same as maximum likelihood fitting of a probabilistic model:
  - Assume $x = (x^1, \ldots, x^n)$ are IID observations of random vectors, where $x^j = (x_1^j, \ldots, x_m^j)$ is an individual vector.
  - Sample mean and covariance matrix:

$$\tilde{\mu} = \frac{1}{n}\sum_{j=1}^{n} x^j \qquad \tilde{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(x^j - \tilde{\mu})(x^j - \tilde{\mu})^\top$$

  - PCA tries to find a linear transformation $A \in \mathbb{R}^{d \times m}$ to find orthogonal directions of largest variance. Projecting data onto principal components makes data features uncorrelated.

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation, motivation

- PCA solution for d components: $A = R\Lambda_d^{-1/2}U_d$ where Λ_d is the diagonal matrix of largest eigenvalues, U_d is the matrix of the corresponding eigenvectors, and R is any rotation matrix

- Interpreting the PCA solution: consider maximum likelihood fitting of the following probabilistic model to observations $(x^1, \ldots, x^n)$

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \\ x|z &\sim \mathcal{N}(Wz + \mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md} \end{aligned}$$

  where the parameters are W, μ, and σ^2 This model says data are first distributed along latent axes z, and then noise is independently added to all coordinates.

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation, motivation

- It can be shown the maximum likelihood solution to the model fitting is $\hat{\mu} = \tilde{\mu}$ $\widehat{W} = U_d(\Lambda_d - \sigma^2 I)^{1/2} R$, and $\hat{\sigma}^2 = \frac{1}{m-d} \sum_{i=d+1}^{m} \lambda_i$ where Λ_d is the diagonal matrix of largest eigenvalues, U_d is the matrix of the corresponding eigenvectors, and R is any rotation matrix.

- Given an observation x, the expected value of the latent variable z can be computed from the model as

$$E(z|x) = R^\top (\Lambda_d - \sigma^2 I)^{1/2} \Lambda_d^{-1} U_d^\top (x - \tilde{\mu})$$

- Same subspace as in PCA; same projections if left-out eigenvalues

  are zero

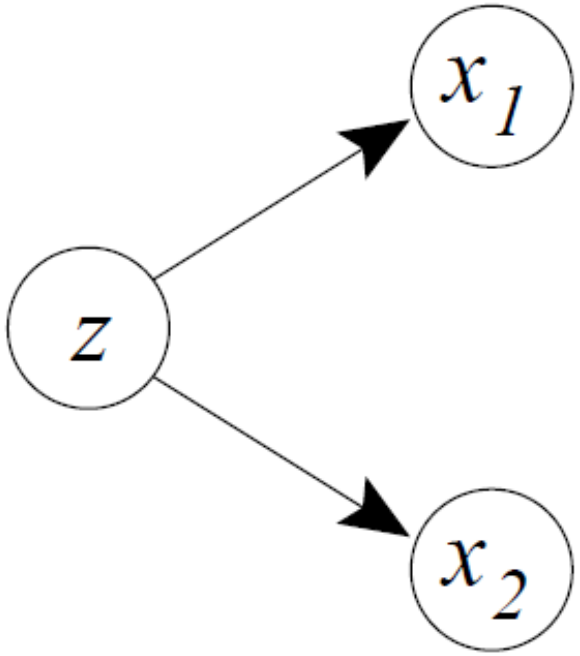
- We will build a probabilistic interpretation for CCA with a similar approach as above

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- We now show that the CCA directions can also be solved by fitting a simple generative model to the data:

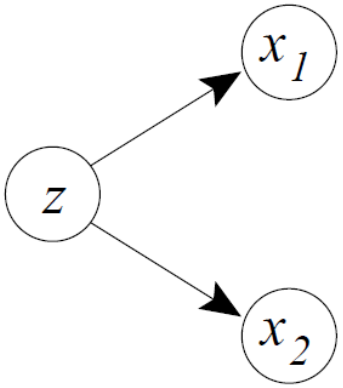
  The model says: there is a single (vector-valued) latent variable z which generates both x_1 and x_2



Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- Model equations:



$$z \sim \mathcal{N}(0, I_d), \quad \min\{m_1, m_2\} \geqslant d \geqslant 1$$

latent variable is normally distributed with p uncorrelated dimensions

$$x_1|z \sim \mathcal{N}(W_1 z + \mu_1, \Psi_1), \quad W_1 \in \mathbb{R}^{m_1 \times d}, \Psi_1 \succcurlyeq 0$$

first observed variable is a projection of the latent variable, with added normally distributed noise (full noise covariance matrix)

$$x_2|z \sim \mathcal{N}(W_2 z + \mu_2, \Psi_2), \quad W_2 \in \mathbb{R}^{m_2 \times d}, \Psi_2 \succcurlyeq 0$$

second observed variable is another projection of the latent variable, with added normally distributed noise (full noise covariance matrix)

- Intuitively, this model makes sense. Next, let's show it really gives the same solution as CCA

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

Normal CCA solution with slightly different notation:

- CCA notation: given feature sets x_1 and x_2 of samples, with dimensionalities m_1 and m_2, find a projection (linear transformation) for each feature set
- Find the projections such that one component within each set of transformed variables is correlated with a single component in the other set.
- CCA reduces the correlation matrix to a block-diagonal matrix, where each block has the form $\begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$ (padded with zeros if the dimensionalities are unequal) and the ρ_i are the canonical correlations; at most *p*=min(m_1 ,m_2) nonzero canonical correlations.
- Denote the sample covariance matrix as $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix}$

covariance among 1st feature set

cross-covariance

covariance among 2nd feature set

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- Then the CCA solution is the set of canonical pairs of projection vectors (u_{1i}, u_{2i}), where $(u_{1i}, u_{2i}) = ((\widetilde{\Sigma}_{11})^{-1/2} v_{1i}, (\widetilde{\Sigma}_{22})^{-1/2} v_{2i})$ and (v_{1i}, v_{2i}) are pairs of left and right singular vectors of the matrix $(\widetilde{\Sigma}_{11})^{-1/2} \widetilde{\Sigma}_{12} (\widetilde{\Sigma}_{22})^{-1/2}$ and the corresponding singular value is the canonical correlation ρ_i for $i = 1 \ldots, p$ and zero otherwise
- If all canonical correlations have different values, the singular vectors have a unique solution.
- Assume the sample covariance matrix is invertible, and denote $U_1 = (u_{11}, \ldots, u_{1m})$ and $U_2 = (u_{21}, \ldots, u_{2m})$. Then
  - $U_1^\top \widetilde{\Sigma}_{11} U_1 = I_m$ projecting the 1st feature set to its projection directions makes the projected features uncorrelated
  - $U_2^\top \widetilde{\Sigma}_{22} U_2 = I_m$ projecting the 2nd feature set to its projection directions makes the projected features uncorrelated
  - $U_2^\top \widetilde{\Sigma}_{21} U_1 = P$ projecting the features makes the cross-correlations diagonal (P = diagonal matrix of the canonical correlations)

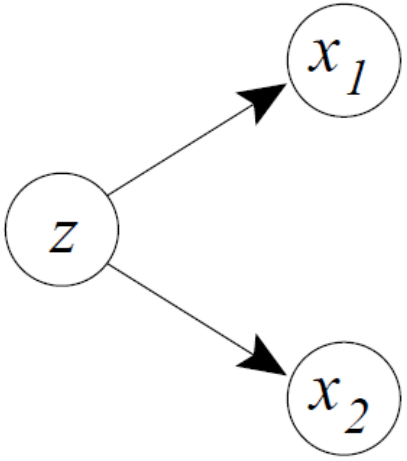
Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- The CCA directions and corresponding canonical correlations can also be obtained from a generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \rho \begin{pmatrix} \widetilde{\Sigma}_{11} & 0 \\ 0 & \widetilde{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

- Next we show that the CCA directions can also be solved by fitting the previously described simple generative model to the data:

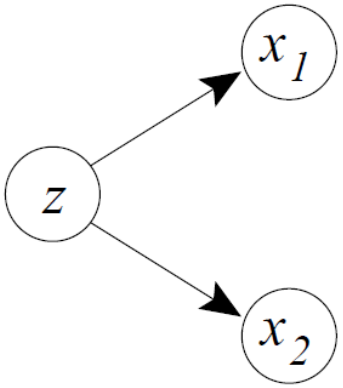


The model says: there is a single (vector-valued) latent variable z which generates both x_1 and x_2

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- Here are the model equations again:



$$z \sim \mathcal{N}(0, I_d), \quad \min\{m_1, m_2\} \geqslant d \geqslant 1$$

latent variable is normally distributed with p uncorrelated dimensions

$$x_1|z \sim \mathcal{N}(W_1 z + \mu_1, \Psi_1), \quad W_1 \in \mathbb{R}^{m_1 \times d}, \Psi_1 \succcurlyeq 0$$

first observed variable is a projection of the latent variable, with added normally distributed noise (full noise covariance matrix)

$$x_2|z \sim \mathcal{N}(W_2 z + \mu_2, \Psi_2), \quad W_2 \in \mathbb{R}^{m_2 \times d}, \Psi_2 \succcurlyeq 0$$

second observed variable is another projection of the latent variable, with added normally distributed noise (full noise covariance matrix)

- It can be shown the maximum likelihood solution is

$$\begin{aligned}
\widehat{W}_1 &= \widetilde{\Sigma}_{11} U_{1d} M_1 \\
\widehat{W}_2 &= \widetilde{\Sigma}_{22} U_{2d} M_2 \\
\widehat{\Psi}_1 &= \widetilde{\Sigma}_{11} - \widehat{W}_1 \widehat{W}_1^\top \\
\widehat{\Psi}_2 &= \widetilde{\Sigma}_{22} - \widehat{W}_2 \widehat{W}_2^\top \\
\hat{\mu}_1 &= \tilde{\mu}_1 \\
\hat{\mu}_2 &= \tilde{\mu}_2
\end{aligned}$$

where $M_1, M_2 \in \mathbb{R}^{d \times d}$ are arbitrary matrices (with spectral norms < 1) such that $M_1 M_2^\top = P_d$.

Columns of U_{1d} , U_{2d} have the first d canonical directions, P_d has the corresponding canonical correlations

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

CCA, probabilistic interpretation

- Given observations of x_1 and/or x_2 , we can use the model to predict the latent variable (mean and variance):

$$
\begin{aligned}
E(z|x_1) &= M_1^\top U_{1d}^\top (x_1 - \mu_1) \\
E(z|x_2) &= M_2^\top U_{2d}^\top (x_2 - \mu_2) \\
\mathrm{var}(z|x_1) &= I - M_1 M_1^\top \\
\mathrm{var}(z|x_2) &= I - M_2 M_2^\top \\
E(z|x_1, x_2) &= \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^\top \begin{pmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{pmatrix} \begin{pmatrix} U_{1d}^\top (x_1 - \mu_1) \\ U_{2d}^\top (x_2 - \mu_2) \end{pmatrix} \\
\mathrm{var}(z|x_1, x_2) &= I - \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^\top \begin{pmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}
\end{aligned}
$$

- The expectation of z given x_1 (or x_2) projects x_1 (or x_2) into the same subspace as in CCA

Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

References

- Becker, S. 1996. Mutual Information Maximization: models of cortical self-organization. Network: Computation in Neural Systems, 7, 7-31.
- Hardoon, D. R., Szedmak, S. and Shawe-Taylor J. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. Neural Computation, 16(12), 2639-2664.
- Magnus Borga. CCA: A Tutorial. http://people.imt.liu.se/~magnus/cca/
- Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California.
- Szedmak, S., De Bie, T., & Hardoon, D. R. (2007). A metamorphosis of canonical correlation analysis into multivariate maximum margin learning. In Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN 2007), Bruges, April 2007.
- Xi Chen, Liu Han, Jaime Carbonell. Structured sparse canonical correlation analysis. Proceedings of AISTATS 2012, JMLR W&CP 22: 199-207, 2012.

Part 2: Nonlinear Canonical Correlation Analysis

Nonlinear CCA by a "kernel trick"

- Suppose we had applied some known nonlinear transformation $f_1(\ )$ from x_1 to a new feature space, and a known nonlinear transformation $f_2(\ )$ from x_2 to a new feature space. Then we could apply CCA to the transformed data. This would give us projections of the transformed data to canonical directions computed from the transformed data. The projections would be nonlinearly related to the original inputs x_1 and x_2 .
- The above is possible if we use a transformation function with a known mathematical form.
- Sometimes, instead of specifying a transformation function, it is easier to specify a similarity measure (inner product) between samples in both feature spaces and across spaces.

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

Nonlinear CCA by a "kernel trick"

- For convenience denote the transformation function by Φ_1 for the first view and Φ_2 for the second. Then the covariance matrices and cross-covariance matrix of transformed data are

$$\Sigma_{11} = E[(\Phi_1(x_1) - \mu_1)(\Phi_1(x_1) - \mu_1)^T]$$
$$\Sigma_{22} = E[(\Phi_2(x_2) - \mu_2)(\Phi_2(x_2) - \mu_2)^T]$$
$$\Sigma_{12} = E[(\Phi_1(x_1) - \mu_1)(\Phi_2(x_2) - \mu_2)^T]$$

- If the data are already centered in the feature space (more on this later) this becomes

$$\Sigma_{11} = E[\Phi_1(x_1)\Phi_1(x_1)^T]$$
$$\Sigma_{22} = E[\Phi_2(x_2)\Phi_2(x_2)^T]$$
$$\Sigma_{12} = E[\Phi_1(x_1)\Phi_2(x_2)^T]$$

estimated as $\Sigma_{12} = \frac{1}{M}\sum_i \Phi_1(x_{1i})\Phi_2(x_{2i})^T$

- We wish to find projections w_1, w_2 to maximize $\boldsymbol{w}_1^T \Sigma_{12} \boldsymbol{w}_2$ under constraint that $\boldsymbol{w}_1^T \Sigma_{11} \boldsymbol{w}_1 = 1$ and $\boldsymbol{w}_2^T \Sigma_{22} \boldsymbol{w}_2 = 1$, without knowing Φ_1 , Φ_2

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

Nonlinear CCA by a "kernel trick"

- The optimal projection vectors will have this form:

$$w_1 = \sum_{i=1}^{M} \alpha_i \Phi_1(x_{1i}) \quad , \quad w_2 = \sum_{i=1}^{M} \beta_i \Phi_2(x_{2i})$$

(where the alpha and beta are some multipliers to be solved)

- The term to be maximized, and the constraint terms, then become

$$w_1^T \Sigma_{12} w_2 = \left(\sum_{i=1}^{M} \alpha_i \Phi_1(x_{1i})^T\right)\left(\frac{1}{M}\sum_{k=1}^{M} \Phi_1(x_{1k})\Phi_2(x_{2k})^T\right)\left(\sum_{l=1}^{M} \beta_l \Phi_2(x_{2l})^T\right)$$

$$= \frac{1}{M}\sum_{i=1}^{M}\sum_{l=1}^{M} \alpha_i \sum_{k=1}^{M} \left(\Phi_1(x_{1i})^T \Phi_1(x_{1k})\right)\left(\Phi_2(x_{2k})^T \Phi_2(x_{2l})^T\right)\beta_l$$

$$w_1^T \Sigma_{11} w_1 = \frac{1}{M}\sum_{i=1}^{M}\sum_{l=1}^{M} \alpha_i \sum_{k=1}^{M} \left(\Phi_1(x_{1i})^T \Phi_1(x_{1k})\right)\left(\Phi_1(x_{1k})^T \Phi_1(x_{1l})^T\right)\alpha_l$$

$$w_2^T \Sigma_{22} w_2 = \frac{1}{M}\sum_{i=1}^{M}\sum_{l=1}^{M} \beta_i \sum_{k=1}^{M} \left(\Phi_2(x_{2i})^T \Phi_2(x_{2k})\right)\left(\Phi_2(x_{2k})^T \Phi_2(x_{2l})^T\right)\beta_l$$

- Denote kernel matrices K_1, K_2 with entries

$$(K_1)_{ij} = \Phi_1^T(x_{1i})\Phi_1(x_{1j}) \quad , \quad (K_2)_{ij} = \Phi_2^T(x_{2i})\Phi_2(x_{2j})$$

- Then the task is to maximize $\boldsymbol{\alpha}^T K_1 K_2^T \boldsymbol{\beta}$ with constraints $\boldsymbol{\alpha}^T K_1 K_1^T \boldsymbol{\alpha} = 1$, $\boldsymbol{\beta}^T K_2 K_2^T \boldsymbol{\beta} = 1$

Importantly, the task definition now **only refers to the kernels**, the actual feature transformations are **no longer needed** in it.

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

Nonlinear CCA by a "kernel trick"

- **Solution of kernel CCA:** first, form the matrix $K = \Gamma_{11}^{-1/2} \Gamma_{12} \Gamma_{22}^{-1/2}$ where $\Gamma_{11} = K_1 K_1^T$, $\Gamma_{22} = K_2 K_2^T$, $\Gamma_{12} = K_1 K_2^T$
- Then form the singular value decomposition of K:

$$K = (\gamma_1, \gamma_2, \ldots, \gamma_k) D (\theta_1, \theta_2, \ldots, \theta_k)^T$$

where D has the singular values,
$\boldsymbol{\gamma}_i$ are eigenvectors of KK^T,
$\boldsymbol{\theta}_i$ are eigenvectors of $K^T K$

- Then the first canonical correlation directions are given by

$$\alpha_1 = \Gamma_{11}^{-1/2} \gamma_1 \quad , \quad \beta_1 = \Gamma_{22}^{-1/2} \theta_1$$

and the second directions are given by the second pair of eigenvectors, and so on

To solve the directions α_1 , β_1 (and so on for the next directions) one only needs to compute the kernel, the actual transformations phi are not needed!

The vectors α_1 , β_1 define the canonical correlation directions but they are not linear projections of the original data features. We show how to project new data onto these directions on the next slide.

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

Nonlinear CCA by a "kernel trick"

- Projections of new data: for a new point $x_{1,new}$ from view 1, the projection to the canonical correlation direction is

$$w_1 \Phi_1^T(x_{1,new}) = \sum_{i=1}^{M} \alpha_i \Phi_1(x_{1i})^T \Phi_1(x_{1,new}) = \sum_{i=1}^{M} \alpha_i K_1(x_{1i}, x_{1,new})$$

where $K_1(x_{1i}, x_{1,new})$ is the kernel (similarity) function value between the new point and training point i in view 1.

- For a new point $x_{2,new}$ from view 2 similarly:

$$w_2 \Phi_2^T(x_{2,new}) = \sum_{i=1}^{M} \beta_i \Phi_2(x_{2i})^T \Phi_2(x_{2,new}) = \sum_{i=1}^{M} \beta_i K_2(x_{2i}, x_{2,new})$$

where $K_2(x_{2i}, x_{2,new})$ is the kernel function in view 2.

To project data one only needs to compute the kernel, the actual transformations phi are not needed!

Thus the whole nonlinear CCA (kernel CCA) can be computed **based on the kernel functions only**, without ever needing to specify what the actual transformations are. Sometimes it is much **easier to specify a kernel function** (inner product function, similarity function) between data than to define good features for data - then kernel CCA is very useful.

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

Technical note

Nonlinear CCA by a "kernel trick"

- Centering (mentioned earlier): if Φ is a transformation to non-zero-mean feature values, then the "centered" transformation $\Phi(x)-(1/N)\sum_i \Phi(x_i)$ has average value zero over the data set
  - If K is the kernel matrix of inner products corresponding to Φ , then $K-(1/N)1K-(1/N)K1+(1/N^2)1K1$, where 1 denotes a matrix where all entries are have value 1, is the kernel matrix corresponding to the centered transformation. The rest of the algorithm can simply be run using these centered kernel matrices.

Following the approach from Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377. Images from that paper.

$$k_1(x,y) = \exp(-(x-y)^T (x-y)*a)$$
$$= f(x)^T f(y)$$
$$(x^T y + 1)^d$$

Part 3: Other variants of Canonical Correlation Analysis

More general “dependent components” of data

Correlation is only a simple measure of dependency: for example, cannot detect correlation between x and y coordinates in a circle.

- Nonlinear kernel transformations help, but then the correlated components become hard to analyze

- Can we find linear transformations (a subspace of each feature set) that would be maximally dependent?

- That is, can we find subspaces that would maximize some more complicated measure of dependency than correlation?

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- One way to measure dependency is to **compare two models of data (hypotheses about the data):** one model that allows dependencies between the two feature sets, and one model that disallows them.
- If the model that includes dependencies is much better at modeling the data, then dependencies are likely to be present in the data.
- Consider a **Bayes factor** between two hypotheses: one says the two feature sets are unrelated, the other says they can be related.
- The Bayes factor is (essentially) the ratio of **likelihoods** of the two hypotheses, given the data:

$$BF = \frac{P(D|H_d)}{P(D|H_i)}$$

- The Bayes factor is a way to compare two hypotheses. How can we use it to find the maximally dependent subspaces from data?

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- Idea: to find two subspaces, we don't care about a model of the whole data, we only care about whether the features in the two subspaces are correlated.
- **For each choice of two subspaces** (a subspace of feature set 1, and a subspace of feature set 2), compare whether data in those two subspaces are related, by **comparing two models**: one model that **allows dependencies between the two subspace**s, and one model that **disallows them**.
- Compute a **Bayes factor** between two hypotheses: one says the two subspaces are unrelated, the other says they can be related:

$$\frac{P(f(D;\theta)|H_d)}{P(f(D;\theta)|H_i)}$$

where f is the mapping from data to the two subspaces by and θ are its parameters

- We will maximize this Bayes factor with respect to the subspaces

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- For a linear transformation, $f_n(D_n;\theta_n) = \theta_n^T D_n$ where Dn is the data matrix for the n:th feature set and θ_n is the projection parameter
- A model is a model of data density: in addition to having the subspaces (linear transformations θ_n), the models typically need **parameters of the density in the subspace**
- When we want to find the best pairs of subspaces, the density parameters are "nuisance parameters" since they change with every pair of subspaces
- We also don't want to overfit the density parameters to small amounts of data (could give exaggerated measures of dependency)
- Idea: there are **nonparametric density estimators** that have very few parameters but are still able to create detailed density estimates: they use the data itself as part of the model.

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- **A leave-one-out Parzen density estimator** evaluates density at each data point by placing a Normal distribution around every other point, and averaging the density. Can be seen as a mixture model with a very large number of components or as a kernel density estimate. Density at training point i:

$$p(x_i;\sigma)=\frac{1}{N-1}\sum_{j=1;\,j\neq i}^{N} N(x_i;x_j,\sigma)=\frac{1}{N-1}\sum_{j=1;\,j\neq i}^{N}\frac{1}{(2\pi\sigma^2)^{d/2}}e^{-(x_i-x_j)^2/2\sigma^2}$$

- The means of the Gaussians come directly from data: the only remaining density parameter is σ, which can be estimated e.g. along with the projections, or by separate heuristics.

- We will use a leave-one-out Parzen density estimator to estimate the density within each subspace, or within each pair of subspaces

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- Model for independent subspaces: features in each subspace are independent of the other subspace---->joint density at a data sample is the product of densities within each subspace.

$$p_{independent}\left(y=\left(\theta_1^T x_1, \theta_2^T x_2\right);\sigma\right)=p\left(\theta_1^T x_1;\sigma\right)p\left(\theta_2^T x_2;\sigma\right)$$

$$=\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\theta_1^T x_i;\theta_1^T x_j,\sigma I_{d_1}\right)\right)\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\theta_2^T x_i;\theta_2^T x_j,\sigma I_{d_2}\right)\right)$$

where y denotes the features of a sample in both subspaces, I_d is an identity matrix of dimension d, and d_1 and d_2 are the dimensionalities of the subspaces (dimensionality after projection)

- Model for dependent subspaces: features in each subspace depend on each other. Density is created over concatenated features of both subspaces. $p_{dependent}\left(y=\left(\theta_1^T x_1, \theta_2^T x_2\right);\sigma\right)$

$$=\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\left[\theta_1^T x_1 \;\; \theta_2^T x_i\right];\left[\theta_1^T x_j \;\; \theta_2^T x_j\right],\sigma I_{d_1+d_2}\right)\right)$$

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

More general "dependent components" of data

- The maximize the Bayes factor

$$\log \frac{\prod_{i=1}^{N} p_{dependent}\left(y=\left(\theta_1^T x_{i,1}, \theta_2^T x_{i,2}\right);\sigma\right)}{\prod_{i=1}^{N} p_{independent}\left(y=\left(\theta_1^T x_{i,1}, \theta_2^T x_{i,2}\right);\sigma\right)}$$

$$=\sum_{i=1}^{N} \log p_{dependent}\left(y=\left(\theta_1^T x_{i,1}, \theta_2^T x_{i,2}\right);\sigma\right) - \log p\left(\theta_1^T x_{i,1};\sigma\right) - \log p\left(\theta_2^T x_{i,2};\sigma\right)$$

$$=\sum_{i=1}^{N} \log\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\left[\theta_1^T x_{i,1}\ \ \theta_2^T x_{i,2}\right];\left[\theta_1^T x_{j,1}\ \ \theta_2^T x_{j,2}\right],\sigma I_{d_1+d_2}\right)\right)$$

$$-\log\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\theta_1^T x_{i,1};\theta_1^T x_{j,1},\sigma I_{d_1}\right)\right)$$

$$-\log\left(\frac{1}{N-1}\sum_{j=1;j\neq i}^{N} N\left(\theta_2^T x_{i,2};\theta_2^T x_{j,2},\sigma I_{d_2}\right)\right)$$

with respect to the projections (and possibly also σ) using for example a gradient descent algorithm.

Following Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005. Pictures from that paper.

References

Part 1:

- Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California.

Part 2:

- Lai, P. L. and Fyfe, C. 2000. Kernel and Nonlinear Canonical Correlation Analysis. International Journal of Neural Systems 10(5), 365-377.
- Leen, G. and Fyfe, C. 2006. A Gaussian Process Latent Variable Model Formulation of Canonical Correlation Analysis. Pages 413-418 of: Proceedings of the 14th European Symposium of Artificial Neural Networks (ESANN).

Part 3:

- Arto Klami and Samuel Kaski. Non-parametric dependent components. In Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing, pages V-209 - V-212, IEEE, 2005.
- Chong Wang. Variational Bayesian Approach to Canonical Correlation Analysis. IEEE Transactions on Neural Networks 18:905-910, 2007.
- Seppo Virtanen, Arto Klami, and Samuel Kaski. Bayesian CCA via group sparsity. In Lise Getoor and Tobias Scheffer, editors, Proceedings of the 28th International Conference on Machine Learning (ICML-11), ICML '11, pages 457–464, New York, NY, 2011. ACM.