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 $w_{x,1}x_1 + w_{x,2}x_2 + \ldots + w_{x,K}x_K$ 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

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

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

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

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 $\mathbf{w}_{\mathbf{x}}$, and the previous equation then gives $\mathbf{w}_{\mathbf{y}}$ from $\mathbf{w}_{\mathbf{x}}$.

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 $\mathbf{w}_\mathbf{x} = [\, w_{x,1}, w_{x,2},\ldots,w_{x,K} \,]$ 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 $\mathbf{w}_\mathbf{y} = [\, w_{y,1}, w_{y,2},\ldots,w_{y,L} \,]$ is the projection basis.
- Find the projection bases by maximizing the correlation between the projections: maximize

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

with respect to $\mathbf{w}_\mathbf{x}$ and $\mathbf{w}_\mathbf{y}$.

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

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

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 $\mathbf{w}_\mathbf{x}$, and the previous equation then gives $\mathbf{w}_\mathbf{y}$ from $\mathbf{w}_\mathbf{x}$.

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 $x^j = (x^j_1, \ldots, x^j_m)$ is an individual vector.
  - Sample mean and covariance matrix:

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

  - 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

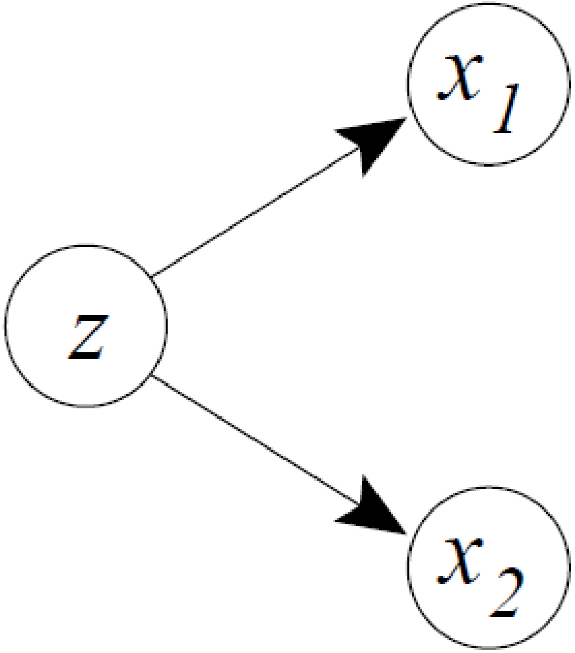
- We will build a probabilistic interpretation for CCA with a similar approach as above

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:

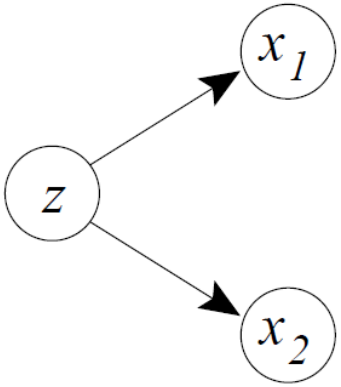
  The model says: there is a single (vector-valued) latent variable z which generates both x_1 and x_2



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)

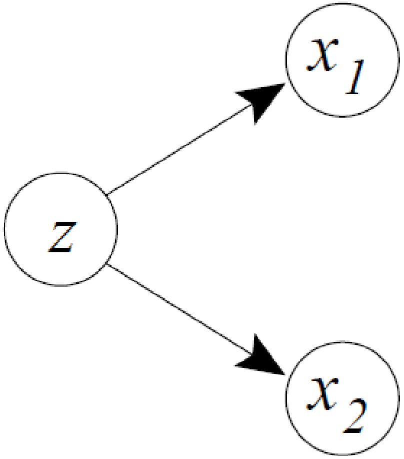
Following the approach from Bach, F. R. and Jordan, M. I. 2005. A Probabilistic Interpretation of Canonical Correlation Analysis. Tech. Report. 688. Dept. of Statistics, University of California. Images from that paper.

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:

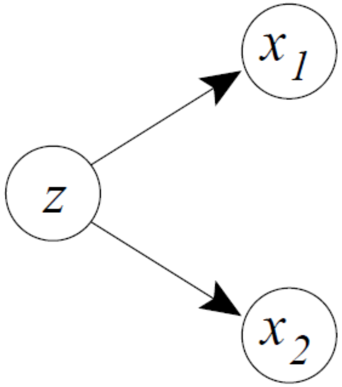


The model says: there is a single (vector-valued) latent variable z which generates both x_1 and x_2

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.