ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

A Gaussian Process Latent Variable Model formulation of Canonical Correlation Analysis

Gayle Leen[1] and Colin Fyfe[1]

1- School of Computing - University of Paisley
PA1 2BE - Scotland
gayle.leen,colin.fyfe@paisley.ac.uk

Abstract. We investigate a nonparametric model with which to visualize the relationship between two datasets. We base our model on Gaussian Process Latent Variable Models (GPLVM)[1],[2], a probabilistically defined latent variable model which takes the alternative approach of marginalizing the parameters and optimizing the latent variables; we optimize a latent variable set for each dataset, which preserves the correlations between the datasets, resulting in a GPLVM formulation of canonical correlation analysis which can be nonlinearised by choice of covariance function.

1 Introduction

We are often interested in finding the relationship between two datasets. A way of achieving this is to project each dataset onto a manifold such that the two projections are maximally correlated. Using a linear projection performs a canonical correlation analysis (CCA) of the data; in this paper we find a probabilistic interpretation of CCA that can be nonlinearised by using nonlinear Gaussian process covariance functions.

2 A latent variable model formulation of CCA

Latent variable models (see e.g. [3]) are defined by a relationship between a set of latent variables $\mathbf{x} = [x_1, ...x_q]^T$ and a set of data variables $\mathbf{y} = [y_1, ...y_D]^T$ probabilistically, and governed by a set of parameters. We consider a latent variable model that is based on a linear mapping between $\mathbf{x}$ and $\mathbf{y}$ with added Gaussian noise: $\mathbf{y} = \mathbf{W}\mathbf{x}+\mathbf{n}$[1], where $\mathbf{n} \sim N_n(0, \mathbf{\Psi})$, and a prior distribution over $\mathbf{x}$: $p(\mathbf{x}) = N_{\mathbf{x}}(0, \mathbf{I}_q)$. For a set of N D-dimensional datapoints $\mathbf{Y} = \{\mathbf{y}_n\}_{n=1}^N$, we obtain the marginal likelihood by integrating over the latent variables and assuming i.i.d. data:

$$p(\mathbf{Y} \mid \mathbf{W}, \mathbf{\Psi}) = \prod_{n=1}^{N} \int p(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{W}, \mathbf{\Psi}) p(\mathbf{x}_n) d\mathbf{x}_n \quad (1)$$

$$= \prod_{n=1}^{N} N_{\mathbf{y}_n}(0, \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}) \quad (2)$$

[1] We assume centred data in this paper, but it is easy to introduce a bias μ

We then find parameter values to maximise the likelihood function in (2) given the data $\mathbf{Y}$. In [4], canonical correlation analysis [5] is formulated as a latent variable model. The model is represented by the graphical model in Figure 1(l), where $\mathbf{y}_1 \in \Re^{m_1}$ and $\mathbf{y}_2 \in \Re^{m_2}$ are the two sets of data variables. Both $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent given the shared latent variable $\mathbf{x}$. The model is defined as follows:

$$\begin{gathered}
\mathbf{x} \sim N(0, \mathbf{I}_q), \min(m_1, m_2) \geq q \geq 1 \\
\mathbf{y} \mid \mathbf{x} \sim N(\mathbf{W}\mathbf{x} + \mu, \mathbf{\Psi}) \\
\text{where } \mathbf{y} = \left(\begin{smallmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{smallmatrix} \right), \mu = \left(\begin{smallmatrix} \mu_1 \\ \mu_2 \end{smallmatrix} \right), \mathbf{W} = \left(\begin{smallmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{smallmatrix} \right), \mathbf{\Psi} = \left(\begin{smallmatrix} \mathbf{\Psi}_1 & 0 \\ 0 & \mathbf{\Psi}_2 \end{smallmatrix} \right) \\
\mathbf{W}_i \in \Re^{m_i x q}, \mathbf{\Psi}_i \in \Re^{m_i x m_i} \succeq 0, i = 1, 2
\end{gathered} \tag{3}$$

The maximum likelihood solutions are given by:

$$\begin{aligned}
\hat{\mathbf{W}}_1 = \tilde{\mathbf{\Sigma}}_{11}\mathbf{U}_{1q}\mathbf{M}_1, \quad \hat{\Psi}_1 = \tilde{\mathbf{\Sigma}}_{11} - \hat{\mathbf{W}}_1\hat{\mathbf{W}}_1^T, \quad \hat{\mu}_1 = \tilde{\mu}_1 \\
\hat{\mathbf{W}}_2 = \tilde{\mathbf{\Sigma}}_{22}\mathbf{U}_{2q}\mathbf{M}_2, \quad \hat{\Psi}_2 = \tilde{\mathbf{\Sigma}}_{22} - \hat{\mathbf{W}}_2\hat{\mathbf{W}}_2^T, \quad \hat{\mu}_2 = \tilde{\mu}_2
\end{aligned} \tag{4}$$

where the columns of $\mathbf{U}_{1q}$ and $\mathbf{U}_{2q}$ are the first q canonical vectors, $\mathbf{M}_1, \mathbf{M}_2 \in \Re^{qxq}$ are arbitrary matrices such that $\mathbf{M}_1\mathbf{M}_2^T = \mathrm{diag}(\rho_1, ..., \rho_q)$ where ρ_i is the ith canonical correlation, and $\tilde{\Sigma} = \left(\begin{smallmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{smallmatrix} \right) = E\left(\left(\begin{smallmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{smallmatrix} \right) \left(\begin{smallmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{smallmatrix} \right)^T \right)$ Using the ML estimates for the parameters, the posterior expectation of $\mathbf{x}$ given the ith dataset $\mathbf{y}_i$ is $E(\mathbf{x} \mid \mathbf{y}_i) = \mathbf{M}_i^T\mathbf{U}_{iq}^T(\mathbf{y}_i - \hat{\mu}_i)$ which lies in the q-dimensional linear subspace of $\Re^{m_i}$, the same as that obtained from CCA.

2.1 Extensions to the latent variable model

This formulates the statistical technique of CCA as a probabilistic model with a log likelihood function given by log $p(\mathbf{y} \mid \mathbf{W}, \Psi)$. However, CCA only finds linearly correlated features between two data sets, and we may want to extract sets of features that share a more complicated relationship. Since we are working within a probabilistic framework, we could extend this simple model by nonlinearising the mapping or creating a mixture of Probabilistic CCA, following [6]. Instead we take a nonparametric approach, to overcome the limitations of parametric models, for which we have to explicitly define the nature of the mapping from latent to data space.

3 From latent variable models to Gaussian Process Latent Variable Models (GPLVM)

Whereas it is more common to marginalise the latent variables and find optimal parameter values to maximise (2), another approach [1] is to marginalise the parameters and optimize the latent variables. The marginalised likelihood is given by placing a prior distribution over the parameters [2], and then integrating

[2] Only $\mathbf{W}$ is integrated out in this model

them out:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{\Psi}) = \prod_{n=1}^{N} \int p(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{W}, \mathbf{\Psi}) p(\mathbf{W}) d\mathbf{W} \tag{5}$$

$$p(\mathbf{Y} \mid \mathbf{X}, \beta) = \frac{1}{(2\Pi)^{\frac{DN}{2}} |\mathbf{K}|^{\frac{D}{2}}} \exp(-\frac{1}{2} \mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)) \tag{6}$$

where $\mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$, $p(\mathbf{W}) = \prod_{i=1}^{D} N_{\mathbf{w}_i}(0, \alpha^{-1}\mathbf{I}_D)$ where $\mathbf{w}_i$ is the ith row of $\mathbf{W}$, and $\mathbf{\Psi} = \beta\mathbf{I}$, following the Probabilistic Principal Component Analysis (PPCA) model as in [7]. The resulting GPLVM [1], [2] is the product of D independent Gaussian processes. The latent coordinate positions $\mathbf{X}$ that maximise the log likelihood function log $p(\mathbf{Y} \mid \mathbf{X}, \beta)$ are given by:

$$\mathbf{X} = \mathbf{U}_q \mathbf{L} \mathbf{V}^T \tag{7}$$

where $\mathbf{U}_q$ is the N by q (the dimension of the latent space) matrix whose columns are the first q eigenvectors of $\mathbf{Y}\mathbf{Y}^T$, $\mathbf{L}$ is a q by q diagonal matrix whose jth element is $l_j = (\frac{\lambda_i}{\alpha D} - \frac{1}{\beta\alpha})^{1/2}$, λ_i is the ith eigenvalue of $\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{V}$ is a rotation matrix. This results in a different probabilistic interpretation of principal component analysis. Note that rather than defining a mapping from latent to data space, the model creates a distribution over the function space based on the locations of the latent variables $\mathbf{X}$ in the latent space. The mapping from latent to data space is implicit in the choice of covariance function K.

3.1 A new latent variable model of CCA

To create a GPLVM version of CCA, we could integrate out the mapping parameter $\mathbf{W}$ from the latent variable model of CCA:

$$p(\mathbf{Y} \mid \mathbf{X}, \mathbf{\Psi}) = \prod_{n=1}^{N} \int p(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{W}, \mathbf{\Psi}) p(\mathbf{W}) d\mathbf{W}$$

with $p(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{W}, \mathbf{\Psi})$ from (3), and using $p(\mathbf{W})$ as above (where $D = m_1 + m_2$). However, this integral in analytically intractable, due to the block diagonal form of $\mathbf{\Psi}$. To overcome this problem, we propose a new latent variable model of CCA (Figure 1(r) and (8). We introduce an intermediate latent variable $\mathbf{z} = \binom{\mathbf{z}_1}{\mathbf{z}_2}$. which is a deterministic transformation of the data $\mathbf{y}$: $\binom{\mathbf{z}_1}{\mathbf{z}_2} = \begin{pmatrix} \mathbf{A}_1^{1/2} & 0 \\ 0 & \mathbf{A}_2^{1/2} \end{pmatrix} \binom{\mathbf{y}_1}{\mathbf{y}_2}$ The model performs CCA through a probabilistic PCA on the transformed data variable $\mathbf{z}$. This is motivated by the idea that we can perform CCA through sphering each data set, and then performing PCA over the two data sets together.

$$\begin{aligned} \mathbf{x} &\sim N(0, \mathbf{I}_q), & min(m_1, m_2) \geq q \geq 1 \\ \mathbf{z}_i \mid \mathbf{x} &\sim N(\mathbf{V}_i\mathbf{x} + \mu_i, \beta\mathbf{I}_{m_i}), & \mathbf{V}_i \in \Re^{m_i x q}, i = 1, 2 \\ \mathbf{y}_i \mid \mathbf{z}_i &\sim \delta(\mathbf{y}_i - \mathbf{A}_i^{-1/2}\mathbf{z}_i), & \mathbf{A}_i \in \Re^{m_i x m_i}, i = 1, 2 \end{aligned} \tag{8}$$

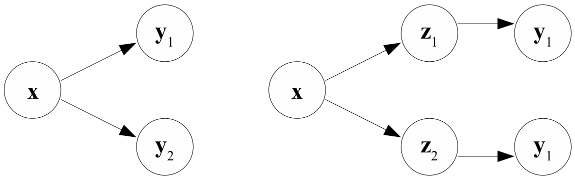

Fig. 1: Two graphical models for canonical correlation analysis. Bach and Jordan's model (left) and our model (right)

where we can rewrite the overall mapping $\mathbf{y} \mid \mathbf{x}$ as:

$$\begin{aligned} \mathbf{y} \mid \mathbf{x} \quad &\sim \quad N(\mathbf{A}^{-1/2}\mathbf{V}\mathbf{x}, \beta\mathbf{A}^{-1}) \\ \text{where } \mathbf{y} \quad &= \quad \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}, \mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix} \end{aligned} \tag{9}$$

which has the same form of the original latent variable model of CCA in (3).

3.2 A GPLVM version of CCA

Our new latent variable model of CCA splits the mapping from latent to data variables into two stages, allowing us to integrate out $\mathbf{V}$ which parameterises the mapping from $\mathbf{x}$ to $\mathbf{z}$, where we define: $p(\mathbf{V}) = \prod_{i=1}^{D} N_{\mathbf{v}_i}(0, \alpha^{-1}\mathbf{I}_D)$ where $\mathbf{v}_i$ is the ith row of $\mathbf{V}$, and $D = m_1 + m_2$

$$\begin{aligned} p(\mathbf{Y} \mid \mathbf{X}, \mathbf{A}, \beta) &= \prod_{n=1}^{N} \int\int p(\mathbf{y}_n \mid \mathbf{z}_n, \mathbf{A}) p(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{V}, \beta) p(\mathbf{V}) d\mathbf{V} d\mathbf{z}_n \\ &= \int \delta(\mathbf{Y} - \mathbf{Z}\mathbf{A}^{-1/2}) \frac{|\mathbf{A}|^{N/2}}{(2\Pi)^{\frac{DN}{2}} |\mathbf{K}|^{\frac{D}{2}}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Z}\mathbf{Z}^T)) d\mathbf{Z} \\ &= \frac{|\mathbf{A}|^{N/2}}{(2\Pi)^{\frac{DN}{2}} |\mathbf{K}|^{\frac{D}{2}}} \exp(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{A}\mathbf{Y}^T)) d\mathbf{Z} \end{aligned} \tag{10}$$

where $\mathbf{Z} = \{\mathbf{z}_n^T\}_{n=1}^N$, $\mathbf{Y} = \{(\mathbf{y}_1)_n^T (\mathbf{y}_2)_n^T\}_{n=1}^N$, $\mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$. The model given in (10) is a GPLVM version of CCA. The matrix $\mathbf{A}$ introduces cross covariance functions between the variables within each data set, similar to $\mathbf{W}$ used in [8] which accounts for different variances in the data dimensions. By accounting for the correlations within each data set with the model, $\mathbf{X}$ should capture similarities between the data sets. The log likelihood function for the model is given by:

$$l = \frac{N}{2}\ln|\mathbf{A}| - \frac{DN}{2}\ln(2\Pi) - \frac{D}{2}\ln|\mathbf{K}| - \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{A}\mathbf{Y}^T) \tag{11}$$

We want to find the latent coordinate set $\mathbf{X}_i$, the projection of the data set $\mathbf{Y}_i$, back into the latent space where $i = 1, ..., 2$. We note that to find $\mathbf{X}$ which underlies both $\mathbf{Y}_1$ and $\mathbf{Y}_2$, we have to find $\mathbf{X}$ that satisfies $\frac{\partial L}{\partial \mathbf{K}}\frac{\partial \mathbf{K}}{\partial \mathbf{X}} = 0$ which is also satisfied by solving $\frac{\partial \mathbf{K}}{\partial \mathbf{X}} = 0$, at which the covariance function $\mathbf{K}$ is given by:

$$\mathbf{K} = \frac{\mathbf{YAY}^T}{D} = \frac{m_1}{D}\frac{\mathbf{Y}_1\mathbf{A}_1\mathbf{Y}_1^T}{m_1} + \frac{m_2}{D}\frac{\mathbf{Y}_2\mathbf{A}_2\mathbf{Y}_2^T}{m_2} = \frac{m_1}{D}\mathbf{K}_1 + \frac{m_2}{D}\mathbf{K}_2 \tag{12}$$

which is the weighted sum of $\mathbf{K}_i$'s where $\mathbf{K}_i$ is the covariance function when we are only considering data set $\mathbf{Y}_i$ and $\frac{\partial L}{\partial \mathbf{X}_i} = 0$. This relates $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}$ through their covariance functions, which allows us to calculate $\mathbf{X}_i$ from:

$$\mathbf{X}_i = \mathbf{U}_{iq}\mathbf{L}_i\mathbf{V}^T \tag{13}$$

where $\mathbf{U}_{iq}$ is the N by q matrix whose columns are the first q eigenvectors of $\mathbf{Y}_i\mathbf{A}_i\mathbf{Y}_i^T$, $\mathbf{L}_i$ is a q by q diagonal matrix whose jth element is $l_j = (\frac{\lambda_i}{\alpha m_i} - \frac{1}{\beta\alpha})^{1/2}$, λ_i is the ith eigenvalue of $\mathbf{Y}_i\mathbf{A}_i\mathbf{Y}_i^T$ and $\mathbf{V}$ is a rotation matrix. We then calculate $\mathbf{K}$ from (12), and find $\mathbf{A}$ to maximise (11) from:

$$\mathbf{A} = N(\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{Y})^{-1} \tag{14}$$

which we then constrain to be of block diagonal form. The other parameters are found by using gradient descent to optimise the log likelihood.

4 Testing the model on a toy dataset

Figure 2 shows a toy data set which comprises two subsets $\mathbf{Y}_1$ (2-dimensional) and $\mathbf{Y}_2$ (1-dimensional). We want to find projections for each subset in latent space such that they are maximally correlated. We can see that y_{12} exhibits a strong correlation with y_{21} whereas y_{11} is independent of y_{21}; therefore the optimal projection for $\mathbf{Y}_1$ would be to project onto y_{12}. Figure 3 (l) shows $\mathbf{X}_2$ against $\mathbf{X}_1$, the projections of $\mathbf{Y}_2$ and $\mathbf{Y}_1$ onto their respective first principal component directions. $\mathbf{Y}_1$ is projected onto y_{11}, capturing the multimodal relationship within $\mathbf{Y}_1$. Figure 3(r) shows our model from (10) trained on the data; it finds 2 sets of linearly correlated 1-dimensional latent coordinates $\mathbf{X}_1$ and $\mathbf{X}_2$ for the respective datasets.

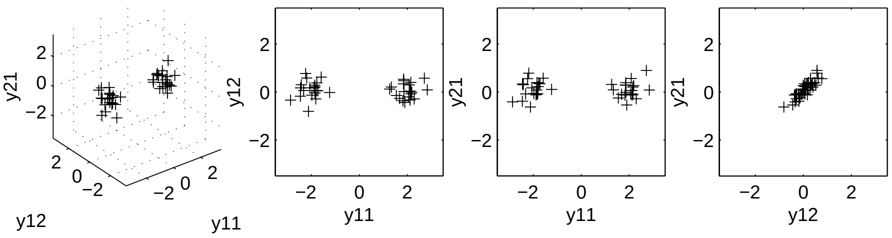

Fig. 2: A toy data set of two data set variables $\mathbf{y}_1 = [y_{11}, y_{12}]^T$ and $\mathbf{y}_2 = [y_{21}]^T$. y_{12} and y_{21} are linearly correlated (far right).

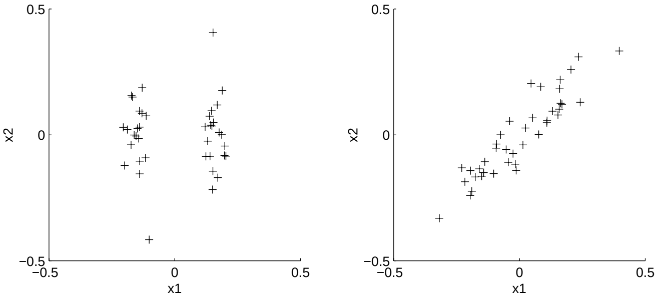

Fig. 3: The latent coordinate positions $\mathbf{X}_2$ vs $\mathbf{X}_1$ using the original GPLVM which uses a PCA projection (l), and using our model in (10)(r)

5 Conclusions and future work

We extended the latent variable model of CCA by a nonparametric approach, and created a GPLVM version of CCA where the mapping is implicit in the choice of covariance function, avoiding the problems associated with parametric modelling. As opposed to Kernel CCA [9], for which a model is not defined, we have a probabilistic model of nonlinear CCA, which is a powerful approach, since we have the benefits of a likelihood function and can select suitable nonlinear functions within a probabilistic framework. Our future work includes incorporating nonlinear covariance functions into the model, and testing it on large and real data sets.

References

[1] N. D. Lawrence. Gaussian process latent variable models for visualization of high dimensional data. *Proc, NIPS 16*, 2004.

[2] N. D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(2005):1783–1816, 2005.

[3] C.M. Bishop. Latent variable models. *In M. I. Jordan (ed.), Learning in Graphical Models*, pages 371–403, 1999.

[4] F.R. Bach and M.I. Jordan. A probabilisic interpretation of canonical correlation analysis. Technical Report 688, Dept of Statistics, University of California, 2005.

[5] H. Hotelling. Relations between two sets of variates. *Biometrika*, (28):312–377, 1936.

[6] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. Technical Report NCRG/97/003, Neural Computing Research Group, Aston University, 1997.

[7] M. Tipping and C. Bishop. Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, 1997.

[8] A. Hertzmann K. Grochow, S.L. Martin and Z. Popovic. Style-based inverse kinematics. *ACM Trans. Graphics*, 23(3):522–531, 2004.

[9] Pei Ling Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2000.