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

Stochastic Processes for Canonical Correlation Analysis

Colin Fyfe and Gayle Leen,
Applied Computational Intelligence Research Unit,
The University of Paisley, Scotland.
email:colin.fyfe,gayle.leen@paisley.ac.uk

Abstract

We consider two stochastic process methods for performing canonical correlation analysis (CCA). The first uses a Gaussian Process formulation of regression in which we use the current projection of one data set as the target for the other and then repeat in the opposite direction. The second uses a Dirichlet process of Gaussian models where the Gaussian models are determined by Probabilistic CCA [1]. The latter method is more computationally intensive but has the advantages of non-parametric approaches.

1 Introduction

A stochastic process $Y(\mathbf{x})$ is a collection of random variables indexed by $\mathbf{x} \in X$ such that values at any finite subset of X form a consistent distribution. A Gaussian Process (GP) therefore is a stochastic process on a function space which is totally specified by its mean and covariance function [7, 6, 4]. A Dirichlet Process (DP) is a stochastic process defined on a space of measures: it can be thought of as an extension to Dirichlet Mixture Models in which the number of models in the mixture tends to ∞.

In this paper, we investigate the use of both types of processes to perform Canonical Correlation Analysis (CCA). Canonical Correlation Analysis is used when we have two data sets which we believe have some underlying correlation. Consider two sets of input data, $\mathbf{x}_1 \in X_1$ and $\mathbf{x}_2 \in X_2$. Then in classical CCA, we attempt to find the linear combination of the variables which gives us maximum correlation between the combinations. Let $y_1 = \mathbf{w}_1^T \mathbf{x}_1$ and $y_2 = \mathbf{w}_2^T \mathbf{x}_2$. Then, for the first canonical correlaton, we find those values of $\mathbf{w}_1$ and $\mathbf{w}_2$ which maximises $E(y_1 y_2)$ under the constraint that $E(y_1^2) = E(y_2^2) = 1$.

2 Gaussian Processes

Consider a stochastic process which defines a distribution, $P(f)$, over functions, f, where f maps some input space, χ to $\Re$. If e.g. $\chi = \Re$, f is infinite dimensional but the $\mathbf{x}$ values index the function, $f(\mathbf{x})$, at a countable number of points and so we use the data at these points to determine $P(f)$ in function space. If $P(f)$ is multivariate Gaussian for every finite subset of X, the process is a GP and is then determined by a mean function $\theta(\mathbf{x})$ and covariance function $\Sigma(\mathbf{x})$. These are often defined by hyperparameters, expressing our prior beliefs on the nature of θ and Σ, whose values are learned from the data.

A commonly used covariance function is $\Sigma : \Sigma_{ij} = \sigma_y^2 \exp(-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2l^2}) + \sigma_n^2 \delta_{ij}$ which enforces smoothing via the l parameter. The σ_y parameter determines the magnitude of the covariances and σ_n enables the model to explain the data, $y = f(\mathbf{x}) + n$, with $n \sim N(0, \sigma_n^2)$.

2.1 GP for Canonical Correlation Analysis

We use a GP to perform CCA in the following manner. Let the input data be $\mathbf{x}_1$ and $\mathbf{x}_2$. Then we define two sets of parameters for the Gaussian Process: let $\theta_i(\mathbf{x}_i), i = 1, 2$, define the mean function of the estimate for CCA and let $\Sigma_i, i = 1, 2$, be the corresponding covariance function. For example, in our first, expository example, we let $\mathbf{x}_1$ and $\mathbf{x}_2$ have a linear relationship so that $\theta_i(\mathbf{x}_i) = b_i \mathbf{x}_i + c_i, i = 1, 2$, with b_i, c_i being the parameters of the process, and $\Sigma^i_{kj} = \sigma^2_{i,y} \exp(-\frac{\|\mathbf{x}_{1,k} - \mathbf{x}_{1,j}\|^2 + \|\mathbf{x}_{2,k} - \mathbf{x}_{2,j}\|^2}{2l_i^2}) + \sigma^2_{i,n} I_N, k, j = 1, ..., N, i = 1, 2$ where N is the number of samples, $\mathbf{x}_{1,j}$ (resp. $\mathbf{x}_{2,j}$) is the j^{th} sample from the first (resp. second) data stream and l_i determines the degree of interaction between the samples. Note that we have continued to index the data stream by i so that $\Sigma^1 \neq \Sigma^2$ since $l_1, \sigma_{1,y}, \sigma_{1,n}$ may evolve differently from $l_2, \sigma_{2,y}, \sigma_{2,n}$.

Then we wish to maximise the covariance in function space of $(\theta_1(\mathbf{x}_1) - \mu_1)(\theta_2(\mathbf{x}_2) - \mu_2)$ under the constraint that $E(\theta_1(\mathbf{x}_1) - \mu_1)^2 = E(\theta_2(\mathbf{x}_2) - \mu_2)^2 = 1$. Let γ_i be a generic parameter of the covariance matrix, Σ^i. Then we use the standard method of gradient descent on the log likelihood with $\theta_2(.)$ as the target for training $\theta_1(.)$,

$$\frac{\partial L}{\partial b_1} = (\theta_2(\mathbf{x}_2) - \theta_1(\mathbf{x}_1))(\Sigma^1)^{-1}\mathbf{x}_1; \quad \frac{\partial l}{\partial c_1} = (\theta_2(\mathbf{x}_2) - \theta_1(\mathbf{x}_1))(\Sigma^1)^{-1} \quad (1)$$

$$\frac{\partial L}{\partial \gamma_1} = -0.5 trace((\Sigma^1)^{-1}\frac{\partial \Sigma^1}{\partial \gamma_1}) + 0.5(\theta_2(\mathbf{x}_2) - \theta_1(\mathbf{x}_1))^T(\Sigma^1)^{-1}\frac{\partial \Sigma^1}{\partial \gamma_1}(\Sigma^1)^{-1}(\theta_2(\mathbf{x}_2) - \theta_1(\mathbf{x}_1)) \quad (2)$$

where

$$\frac{\partial \Sigma_1}{\partial l_1} = 2\Sigma_1 \frac{T^1}{2l_1^2}; \quad \frac{\partial \Sigma_1}{\partial \sigma_{1,y}} = 2\sigma_{1,y} \exp(-\frac{T^1}{2l_1^2}); \quad \frac{\partial \Sigma_1}{\partial \sigma_{1,n}} = 2\sigma_{1,n} I_N \quad (3)$$

where $T^1_{kj} = \frac{\|\mathbf{x}_{1,k} - \mathbf{x}_{1,j}\|^2 + \|\mathbf{x}_{2,k} - \mathbf{x}_{2,j}\|^2}{2l_1^2}$. Thus we are using the current estimates

given by $\theta_2(\mathbf{x}_2)$ as targets for the training of the mean and covariance functions for the estimated functions on $\mathbf{x}_1$. We alternate this training with the equivalent rules for for the estimated functions on $\mathbf{x}_2$ when $\theta_1(\mathbf{x}_1)$ becomes the target. We can view the covariance matrix as the local product of the covariance matrices of X_i, thus creating a covariance matrix[4] for the product space $X_1 \times X_2$. An alternative would be to use the sum of the individual covariances.

We must also heed the constraint that $E(\theta_1(\mathbf{x}_1)-\mu_1)^2 = E(\theta_2(\mathbf{x}_2)-\mu_2)^2 = 1$ during training and so we scale the parameters of $\theta_i()$ after each update to satisfy this constaint.

We create two sets 100 samples of 4 dimensional data in which all values are randomly taken from a $N(0,1)$ distribution except that the first element of each of the two sets of samples is taken from $0.5 * (t + \mu_j) + c_i, i = 1, 2j = 1, ..., 100$, where t is sampled from $N(0,1)$, and is common to both data streams and μ_j is independently drawn from $N(0,1)$.

Figure 1 shows the convergence of the b_i parameters to the correct direction in terms of the cosine between the 4-dimensional vectors and the correct direction. We see quick and reasonably accurate convergence with stability at the converged values. Best results are achieve by annealing the learning rate from 0.01 to 0 during the course of the simulation. There are, however, some problems with this model:

- For this simulation, the true $c_1 = 3, c_2 = 0$, however the estimated values are $c_1 = 1.48, c_2 = -1.49$. We would like some way to ground the simulation in the truth.
- The model is over-confident: the σ_y are too high and the σ_n too small.
- This GP is a parametric approach: it defines the relationship as linear a priori. We would prefer to use a non-parametric approach e.g. with $p(W) = N(0, \sigma^2 I)$, and allow the covariance matrix to determine the nature of the relationship but we then have no target with which we determine the values of hyperparameters.

3 Probabilistic CCA

[1] create a model of CCA based on underlying latent variables. Let

$$\begin{aligned} \mathbf{y} &\sim N(0, I_d) \\ \text{Then } \mathbf{x}_1|\mathbf{y} &\sim N(W_1\mathbf{y}, \Xi_1), \qquad W_1 \in R^{m_1 \times d}, \Xi_1 \succeq 0 \\ \mathbf{x}_2|\mathbf{y} &\sim N(W_2\mathbf{y}, \Xi_2), \qquad W_2 \in R^{m_2 \times d}, \Xi_2 \succeq 0 \end{aligned}$$

where we have assumed zero mean data.

Then the maximum likelihood parameters for the model are given by

$$\begin{aligned} \hat{W_1} = \Sigma_{11} U_{1d} M_1 \quad &, \quad \hat{W_2} = \Sigma_{22} U_{2d} M_2 \\ \hat{\Xi_1} = \Sigma_{11} - \hat{W_1}\hat{W_1}^T \quad &, \quad \hat{\Xi_2} = \Sigma_{22} - \hat{W_2}\hat{W_2}^T \end{aligned}$$

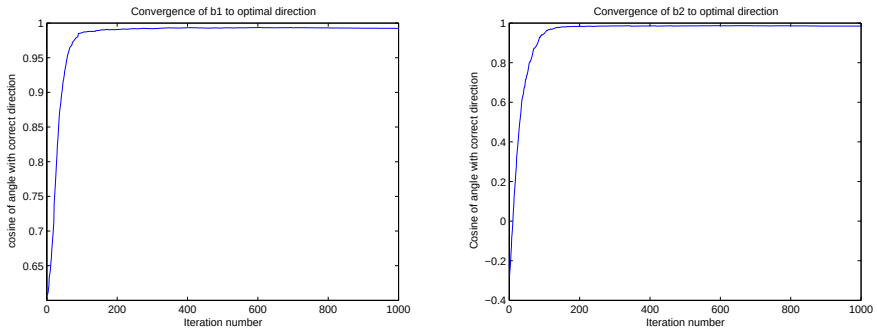

Figure 1: Convergence of $b_i, i = 1, 2$ to the correct direction.

where Σ_{ii} is the covariance matrix of $\mathbf{x}_i$, U_{id} is the first d canonical correlation filters, and $M_1, M_2 \in R^{d \times d}$ are such that $M_1 M_2^T$ gives the diagonal matrix of canonical correlations.

Let $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ and $\Xi = \begin{pmatrix} \Xi_1 & 0 \\ 0 & \Xi_2 \end{pmatrix}$. Let Σ denote the sample covariance matrix. [1] derive an EM algorithm for finding these parameters,

$$W_{t+1} = \Sigma \Xi_t^{-1} W_t M_t (M_t + M_t W_t^T \Xi_t^{-1} \Sigma \Xi_t^{-1} W_t M_t)^{-1} \quad (4)$$

$$\Xi_{t+1} = \begin{pmatrix} (\Sigma - \Sigma \Xi_t^{-1} W_t M_t W_{t+1}^T)_{11} & 0 \\ 0 & (\Sigma - \Sigma \Xi_t^{-1} W_t M_t W_{t+1}^T)_{22} \end{pmatrix} \quad (5)$$

where $M_t = (I + W_t^T \Xi_t^{-1} W_t)^{-1}$.

It is our empirical finding that the above model works well when the data is indeed drawn from that model but that any deviations from this (especially non-Gaussianity), result in incorrect convergence i.e. the model converges to solutions which do not represent anything close to a CCA. Thus we search for a loosening of these constraints.

4 Dirichlet Processes

We will consider symmetric Dirichlet distributions on a space of measures i.e. a sample from this distribution will be a probability density function on the data space. More formally, let the data space, $X \subset R^D$ be partitioned into K disjoint sets, $X_1, X_2, ..., X_K$ such that $X = \cup X_i$ and let $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$ be a probability measure on the data space, so that $P(\mathbf{x} \in X_i) = \theta_i$. Then the Dirichlet distribution is given by

$$P(\Theta | \alpha, \Theta_0) = \frac{\Gamma(\alpha)}{\prod_{i=1}^K \Gamma(\alpha \theta_{0,i})} \prod_{i=1}^K \theta_i^{\alpha \theta_{0,i} - 1} \quad (6)$$

where Θ_0 is the base measure and $\theta_{0,i}$ is the base measure on X_i; these can be thought of as the centres of the process since $E(\Theta(X_i)) = \theta_{0,i}$. α is a posi-

tive scalar which determines the width of the distribution since $\mathrm{Var}(\Theta(X_i)) = \frac{\theta_{0,i}(1-\theta_{0,i})}{\alpha+1}$.

One way [5] to describe the DP is as an extension of the Dirichlet Mixture model as the number of mixtures, K tends to ∞.

$$G \sim DP(.|G_0, \alpha), \qquad \theta_i \sim G(.), \qquad x_i \sim p(.|\theta_i) \tag{7}$$

An alternative description is as the Chinese Restaurant process [2] which emphasises the conjugacy of the DP, analogous to the Dirichlet distribution's conjugacy to the multinomial distribution. Let us have N observations which are multinomially distributed according to Θ. Then

$$\begin{aligned} P(\Theta|\alpha, \Theta_0, \mathbf{x}_1, ..., \mathbf{x}_N) &= \frac{P(\mathbf{x}_1, ..., \mathbf{x}_N)|\alpha, \Theta_0, \Theta)P(\Theta|\alpha, \Theta_0)}{P(\mathbf{x}_1, ..., \mathbf{x}_N)} \\ &= C\prod_{i=1}^{K}\theta_i^{n_i} \times \prod_{i=1}^{K}\theta_i^{\alpha\theta_{0,i}-1} = C\prod_{i=1}^{K}\theta_i^{\alpha\theta_{0,i}+n_i-1} \end{aligned}$$

which is also a Dirichlet process. In the above, n_i is a count of the number of times a sample belonged to the i^{th} set, X_i and so $\sum_{i=1}^{K} n_i = N$. We can view this as a means of recursively updating our estimates of Θ_0 by noting that the new distribution is a Dirichlet distribution with $\alpha^{new} = \alpha + N$ and $\Theta_0^{new} = \frac{\alpha\Theta_0 + N\hat{F}}{\alpha+N}$ where $\hat{F}$ is the empirical distribution i.e. the n_i.

4.1 DP for Canonical Correlation Analysis

We create a DP Mixture of Gaussians in which the parameters of the Gaussians are derived from local Probabilistic CCA while the localness of this operation is as a result of responsiblities found by the DP. Each of the local CCA models returns a local estimate of W and Ξ using the responsibilities. i.e. Σ in (5) becomes

$$\Sigma_k = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^T R_{ki}(\mathbf{x}_i - \mu) \tag{8}$$

where

$$R_{ki} = \begin{cases} 1 \text{ if} & k = \arg\max_j P(\mathbf{x}_i|j)P(j|j^-) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

with $P(j|j^-) = \frac{n_j^-}{N-1+\alpha}$ where n_j^- is a count of the number of other samples (excluding the i^{th} one) allocated to mixture j, a form of Gibbs sampling. With probability $\frac{\alpha}{\alpha+N}$ a new mixture is created and $K = K + 1$. However as more populous classes are more liable to be joined the number of local models is kept low (controlled by the α parameter). We graph in the left diagram of Figure 2 the first element in each of two 4 dimensional data streams which have no other correlations. The right diagrams show the first PCCA directions found by this method.

We see that the DP is also able to capture local linear correlations but approximate the nonlinear relationship between the two sets as a mixture of local

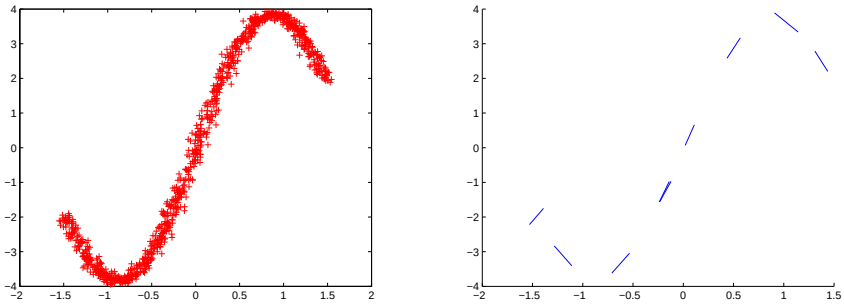

Figure 2: The left diagram maps the first element of $\mathbf{x}_1$ against the first element of $\mathbf{x}_2$ in 1000 samples. The right diagram shows results of 6 iterations with the DP method: we have found 8 local mixtures when α=5 and show the linear correlations found in the same space as the left figure.

probabilistic CCA models. This method is as accurate as the GP method but has the advantages of mixture modelling. Alternative approaches to nonlinear CCA [3] will be investigated in future to determine if non-parametric Gaussian Processes can define such relationships.

References

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

[2] Z. Ghahramani. Non-parametric bayesian methods. Tutorial, Uncertainty in Artificial Intelligence, http://www.gatsby.ucl.ac.uk, 2005.

[3] P.L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(5):365–377, 2001.

[4] D. J. C. MacKay. Introduction to gaussian processes. Technical report, University of Cambridge, http://www.inference.phy.cam.uk/mackay/gpB.pdf, 1997.

[5] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[6] C. E. Rasmussen. *Advanced Lectures on Machine Learning*, chapter Gaussian Processes in Machine Learning, pages 63–71. 2003.

[7] C. K. I. Williams. Prediction with gaussian processes: from linear regression to linear prediction and beyond. Technical report, Aston University, 1997.