Learning shared and separate features of two related data sets using GPLVMs

Gayle Leen
Adaptive Informatics Research Centre
Helsinki University of Technology
gleen@cis.hut.fi

Colin Fyfe
Applied Computational Intelligence Unit
University of the West of Scotland
colin.fyfe@uws.ac.uk

1 Introduction

Dual source learning problems can be formulated as learning a joint representation of the data sources, where the shared information is represented in terms of a shared underlying process. However, there may be situations in which the shared information is not the only useful information, and interesting aspects of the data are not common to both data sets. Some useful features within one data set may not be present in the other and vice versa; this complementary property motivates the use of multiple data sources over single data sources which capture only one type of useful information. For instance, having two eyes (and two streams of visual data) allows us to gain a 3-D impression of the world. This ability of stereo vision combines both shared features and features private to each data stream to form a coherent representation of the world; common shifted features can be used in disparity estimation to infer depths of objects, while some features which may be seen in one view but not in the other, due to occlusions, can provide additional information about the scene.

In this work, we present a probabilistic generative framework for analysing two sets of data, where the structure of each data set is represented in terms of a shared and private latent space. Explicitly modeling a private component for each data set avoids an oversimplified representation of the within-set variation such that the between-set variation can be modeled more accurately, as well as giving insight into potentially interesting features particular to a data set. Since two data sets may have a complex (possibly nonlinear) relationship, we use nonparametric Bayesian techniques - we define Gaussian process priors over the functions from latent to data spaces, such that each data set is modelled as a Gaussian Process Latent Variable Model (GPLVM) [1] where the dependency structure is captured in terms of shared and private kernels.

2 Generative models for two related data sources

Suppose that we have two related data variables $\mathbf{y}_1 \in \Re^{m_1}$ and $\mathbf{y}_2 \in \Re^{m_2}$. We represent each data source as the sum of two independent components, a shared component with the other data source that captures the common information, and a private component which captures the information private to the data source. Assuming that the shared component can be represented in terms of a shared latent variable $\mathbf{s} \in \Re^q, q < min(m_1, m_2)$, and the private components for $\mathbf{y}_1$ and $\mathbf{y}_2$ in terms of latent variables $\mathbf{x}_1 \in \Re^{q_1}, q_1 < m_1$ and $\mathbf{x}_2 \in \Re^{q_2}, q_2 < m_2$ respectively, the data generation process is:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{f}_1(\mathbf{s}) + \mathbf{n}_1(\mathbf{x}_1) \\ \mathbf{y}_2 &= \mathbf{f}_2(\mathbf{s}) + \mathbf{n}_2(\mathbf{x}_2) \end{aligned} \tag{1}$$

where $\mathbf{f}_1(\mathbf{s}) = [f_{1,1}(\mathbf{s}), ..., f_{1,m_1}(\mathbf{s})]^\top$, $\mathbf{f}_2(\mathbf{s}) = [f_{2,1}(\mathbf{s}), ..., f_{2,m_2}(\mathbf{s})]^\top$ are function values that share the same input $\mathbf{s}$, and $\mathbf{n}_1(\mathbf{x}_1) = [n_{1,1}(\mathbf{x}_1), ..., n_{1,m_1}(\mathbf{x}_1)]^\top$, $\mathbf{n}_2(\mathbf{x}_2) =$

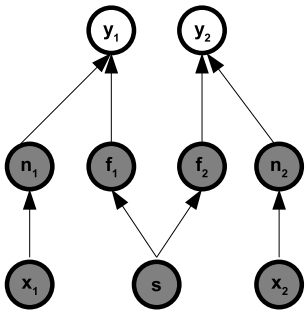

Figure 1: The corresponding graphical model for the unsupervised learning of two related data variables $\mathbf{y}_1$ and $\mathbf{y}_2$. Each data variable consists of two independent components, the shared function $\mathbf{f}$, and the private function $\mathbf{n}$.

$[n_{2,1}(\mathbf{x}_2), ..., n_{2,m_2}(\mathbf{x}_2)]^\top$ are function values of the private latent variables. The corresponding graphical model is shown in Figure 1.

2.1 Gaussian process priors over shared and private functions

Rather than create parametric forms for the functions, we employ the Gaussian process framework to define priors over the functions. Given N pairs of data variables $\mathbf{Y}_1 = [\mathbf{y}_{1,1}, ..., \mathbf{y}_{1,N}]^\top$, $\mathbf{Y}_2 = [\mathbf{y}_{2,1}, ..., \mathbf{y}_{2,N}]^\top$, and defining the underlying function values as $\mathbf{F}_1, \mathbf{F}_2$ evaluated at $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_N]^\top$, and $\mathbf{N}_1, \mathbf{N}_2$ evaluated at $\mathbf{X}_1 = [\mathbf{x}_{1,1}, ..., \mathbf{x}_{1,N}]^\top$ and $\mathbf{X}_2 = [\mathbf{x}_{2,1}, ..., \mathbf{x}_{2,N}]^\top$ respectively, the priors are given by:

$$p(\mathbf{F}_1 \mid \mathbf{S}) = \prod_{i=1}^{m_1} \mathcal{N}(\mathbf{f}_{1,i} \mid 0, \mathbf{K}_{f_1}), p(\mathbf{N}_1 \mid \mathbf{X}_1) = \prod_{i=1}^{m_1} \mathcal{N}(\mathbf{n}_{1,i} \mid 0, \mathbf{K}_{n_1}) \tag{2}$$

$$p(\mathbf{F}_2 \mid \mathbf{S}) = \prod_{i=1}^{m_2} \mathcal{N}(\mathbf{f}_{2,i} \mid 0, \mathbf{K}_{f_2}), p(\mathbf{N}_2 \mid \mathbf{X}_2) = \prod_{i=1}^{m_2} \mathcal{N}(\mathbf{n}_{2,i} \mid 0, \mathbf{K}_{n_2}) \tag{3}$$

where we have used the notation for a function underlying data dimension i as e.g. $\mathbf{f}_{1,i} = [f_{1,i}(\mathbf{s}_1), ..., f_{1,i}(\mathbf{s}_N)]^\top$, $\mathbf{K}_{n_1}$ and $\mathbf{K}_{n_2}$ are the covariance functions for the private function priors, with respective inputs $\mathbf{X}_1$ and $\mathbf{X}_2$, and $\mathbf{K}_{f_1}$ and $\mathbf{K}_{f_2}$ are the covariance functions for the shared function priors, with shared input $\mathbf{S}$.

Given the generative process for the data in (1), and the GP priors in (2) and (3), we integrate over the $\mathbf{F}$'s and $\mathbf{N}$'s to get the resulting model:

$$p(\mathbf{Y}_1 \mid \mathbf{X}, \mathbf{X}_1) = \frac{1}{(2\pi)^{\frac{m_1 N}{2}} |\mathbf{K}_1|^{\frac{m_1}{2}}} \exp\left(-\frac{1}{2} tr(\mathbf{K}_1^{-1} \mathbf{Y}_1 \mathbf{Y}_1^\top)\right) \tag{4}$$

$$p(\mathbf{Y}_2 \mid \mathbf{X}, \mathbf{X}_2) = \frac{1}{(2\pi)^{\frac{m_2 N}{2}} |\mathbf{K}_2|^{\frac{m_2}{2}}} \exp\left(-\frac{1}{2} tr(\mathbf{K}_2^{-1} \mathbf{Y}_2 \mathbf{Y}_2^\top)\right) \tag{5}$$

where $\mathbf{K}_1 = \mathbf{K}_{f_1} + \mathbf{K}_{n_1}$, and $\mathbf{K}_2 = \mathbf{K}_{f_2} + \mathbf{K}_{n_2}$. Each data stream is modelled by a Gaussian Process Latent Variable Variable Model, whose covariance function consists of a shared component (dependent on $\mathbf{S}$) and a private component (dependent on either $\mathbf{X}_1$ or $\mathbf{X}_2$). The dimensions within each data set are modelled as independently and identically distributed, and $\mathbf{X}_1$ and $\mathbf{X}_2$ capture the correlations within $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively. $\mathbf{S}$ captures the correlations between $\mathbf{Y}_1$ and $\mathbf{Y}_2$. We can consider any valid (nonlinear) kernel of the inputs, which imply nonlinear mappings of $\mathbf{S}$, $\mathbf{X}_1$, and $\mathbf{X}_2$ to their respective data spaces. Using a nonparametric Bayesian prior over the private functions underlying each data space is an elegant and flexible prior over underlying private structure of the data sets. The resulting model is a generalisation of probabilistic canonical correlation analysis (PCCA) [2], for which the mappings between latent and data space are linear functions; our model can be viewed as a probabilistic interpretation of nonlinear CCA, where the underlying structure to the within-set variation is modelled explicitly.

2.2 Training the model

Learning the model, given two sets of related data $\mathbf{Y}_1$ and $\mathbf{Y}_2$, consists of finding the latent coordinates $\mathbf{S}, \mathbf{X}_1$, and $\mathbf{X}$ and the hyperparameters $\Theta_{K_{n_i}}, \Theta_{K_{f_i}}, i = 1, 2$, of the two covariance functions $\mathbf{K}_1$ and $\mathbf{K}_2$ to maximise the log likelihood function of (4) and (5). We use scaled conjugate gradients and the GPLVM toolbox available from `http://www.cs.man.ac.uk/~neill/fgplvm/`. The optimisation takes place in two steps; first we jointly optimise $\mathbf{S}$ and the parameters of the shared kernels $\Theta_{K_{f_i}}, i = 1, 2$, then we jointly optimise $\mathbf{X}_1$, $\mathbf{X}_2$ and the private kernel parameters $\Theta_{K_{n_i}}, i = 1, 2$. Since the variation in each data set dimension is effectively shared between the shared latent set $\mathbf{S}$ and the private latent set $\mathbf{X}_1$ or $\mathbf{X}_2$, due to $\mathbf{K}_1 = \mathbf{K}_{\mathbf{f}_1}(\mathbf{S}, \mathbf{S}) + \mathbf{K}_{\mathbf{n}_1}(\mathbf{X}_1, \mathbf{X}_1)$, and $\mathbf{K}_2 = \mathbf{K}_{\mathbf{f}_2}(\mathbf{S}, \mathbf{S}) + \mathbf{K}_{\mathbf{n}_2}(\mathbf{X}_2, \mathbf{X}_2)$, the model is very sensitive to its initialisation: the algorithm may become trapped in a local minimum and fail to recover the true embedded space. In our experiments we use CCA to initialise the positions of $\mathbf{S}$, since $\mathbf{S}$ represents the shared features between $\mathbf{Y}_1$ and $\mathbf{Y}_2$. To initialise the private latent spaces, we calculate the off-subspace variances for $\mathbf{Y}_1$ and $\mathbf{Y}_2$, $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ respectively, which are the noise covariance matrices of probabilistic CCA. We then find $\mathbf{X}_1$ and $\mathbf{X}_2$ by projecting the corresponding data set onto the first q_1 and q_2 dominant eigenvectors of $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ respectively.

3 Experiments

In this section we demonstrate the model's performance on two data sets of images. We separate the images into a set of latent images. The latent images form a basis of prototype images, consisting of three sets of images, a set of images that represent the features common to both sets of data, and two sets of images that represent the features that are only present in their corresponding data set. In our experiments, we use a variation of the bars problem, which is a test problem defined in [3].

3.1 Bars data

The bars problem is a benchmark task for learning independent components from an image. While the original problem consists of decomposing a set of images into a set of underlying features (vertical and horizontal bars), in this experiment we consider a modified version of the problem that illustrates our algorithm's ability to find both shared and private features for two image sets. We create two sets of 8×8 images; 24 examples from each set are shown in Figure 2a. Each image is generated by first instantiating one of the 8 possible horizontal bars, chosen with equal probability. For the first set of images (top three rows of Figure 2a), one of the 4 possible vertical bars in the left half of the image is instantiated with equal probability, and similarly, for the second set of images, (bottom three rows of Figure 2a) one of the 4 possible vertical bars in the right half of the image is instantiated with equal probability. Producing the two image sets involves a shared process in the generation of the horizontal bars, and private processes in generating the vertical bars.

3.2 Parts-based decomposition of bar images

Our aim is to recover the set of eight shared features - the horizontal bars - and the two sets of four private features - the vertical bars. One of the difficulties with the bars data is that each image is nonlinearly related to the underlying features (the bars), since the superposition of the features to form the image results in occlusion, or overlap, of the features. Each image can be thought of as a linear combination of horizontal and vertical bars which is then passed through a nonlinearity which models the overlap i.e. for the ith image of both data sets:

$$\mathbf{Y}_{1,i} = G_{f_1}(\mathbf{S}\mathbf{W}_{f_1}) + G_{n_1}(\mathbf{X}_1\mathbf{W}_{n_1}) \tag{6}$$

$$\mathbf{Y}_{2,i} = G_{f_2}(\mathbf{S}\mathbf{W}_{f_2}) + G_{n_2}(\mathbf{X}_2\mathbf{W}_{n_2}) \tag{7}$$

where G_{f_1}, G_{f_2}, G_{n_1} and G_{n_2} are nonlinear output functions, $\mathbf{W}_{f_1} \in \Re^{q \times m_1}$, $\mathbf{W}_{f_2} \in \Re^{q \times m_2}$, $\mathbf{W}_{n_1} \in \Re^{q_1 \times m_1}$ and $\mathbf{W}_{n_2} \in \Re^{q_2 \times m_2}$ are mixing matrices. For our experiment, we use polynomial covariance functions of degree 2 for each process to reflect our knowledge about the data generation process; the polynomial covariance function is given by:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha \left(w\mathbf{x}_i^\top \mathbf{x}_j + \gamma\right)^2 + \beta^{-1}\delta_{i,j} \tag{8}$$

with hyperparameters $\Theta_{K_{poly}} = \{\alpha, \beta, \gamma, w\}$, where α is a scale parameter, β is the inverse noise variance, w controls the scale of the dot product component, and γ is a bias parameter.

We use an 8-dimensional shared latent space $\mathbf{S}$, and a 4-dimensional private latent spaces $\mathbf{X}_1$ and $\mathbf{X}_2$ (where the columns are the underlying images). We use a training data set of 200 pairs of images such that the 200 columns of $\mathbf{Y}_1 \in \Re^{64\times200}$ and $\mathbf{Y}_2 \in \Re^{64\times200}$ are 8×8 images that contain a vertical bar in the left and right half of the image respectively, and a horizontal bar. We also constrain the latent points' values to lie between 0 and 1, such that they correspond to underlying image pixels. Each latent point $\mathbf{x}$ is reparameterised as $\mathbf{x}'$, using a sigmoid transform $\mathbf{x} = \log(\mathbf{x}'/(1-\mathbf{x}'))$, such that the optimisation takes place in a transformed space. Figure 2b shows the discovered latent images (the columns of $\mathbf{S}$, $\mathbf{X}_1$, and $\mathbf{X}_2$), after training the model on the 200 pairs of training images. As can be seen, the model manages to decompose the training images into the sets of underlying shared and private features.

3.2.1 Reconstruction of the images

In this section, we show how the shared and private latent images which we found in the previous section can be used to reconstruct the original images. This involves finding the posterior distributions of the underlying private and shared functions given the data $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and the latent features $\mathbf{S}$, $\mathbf{X}_1$ and $\mathbf{X}_2$. This investigates how well the algorithm is able to model the overlap between features. Figure 2e shows the first 24 reconstructed images for each data set, given by the posterior means for $\mathbf{Y}_1^* = \mathbf{F}_1^* + \mathbf{N}_1^*$ and $\mathbf{Y}_2^* = \mathbf{F}_2^* + \mathbf{N}_2^*$ The top three rows are reconstructions for the first set, and the bottom three rows are reconstructions for the second set. The reconstructed images are a good approximation to the original images shown in Figure 2a. The reconstructions for the second set model the overlap between bars more accurately than for the first set. Figure 2c and 2d shows the shared and private components of each image. (c) shows the posterior mean of the shared functions $\mathbf{F}_1^*$ (top three rows) and $\mathbf{F}_2^*$ (bottom three rows), and (d) shows the posterior mean of the private functions $\mathbf{N}_1^*$ (top three rows) and $\mathbf{N}_2^*$ (bottom three rows). An interesting observation is that in some of the images, a pixel is missing from one of the bars. This is due to the latent images being put through the nonlinear map implied by the polynomial covariance function. This aids in the successful reconstructions of the original image; the overlap between bars is taken into account by removing a pixel at the point in the image where the bars intersect.

4 Discussion

We have reviewed a probabilistic generative model that models two related data sources in terms of shared and private latent spaces. Each data source is modelled as a GPLVM, which 'maps' from the shared and latent spaces to the data space. The model can be viewed as a nonlinear version of probabilistic canonical correlation analysis. Similar work has been carried out independently in [5]. We demonstrated the model's performance on a parts-based decomposition task on two related sets of images, where the images were decomposed into underlying features that were shared between the data sets, and private to each data set. Other work includes using automatic relevance determination (ARD) methods, as suggested in [6] from the neural networks literature, in the covariance functions to automatically determine the dimensionality of the latent spaces. Future work includes investigation of different initialisation schemes for the model, and experiments with stereo image data sets.

References

[1] 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.

[2] F.R. Bach and M.I. Jordan. A Probabilisic Interpretation of Canonical Correlation Analysis. Technical Report 688, Dept of Statistics, University of California, 2005.

[3] P. Földiák. Forming sparse representations by local anti-Hebbian learning. In *Biological Cybernetics*, number 64, pages 165–170. 1990.

[4] C. H. Ek, J. Rihan, P. H. S. Torr, G. Rogez, and N. D. Lawrence. Ambiguity Modeling in Latent Spaces. In *Machine Learning for Multimodal Interaction*, volume 5237/2008 of *Lecture Notes in Computer Science*, pages 62–73. Springer Berlin/Heidelberg, 2008.

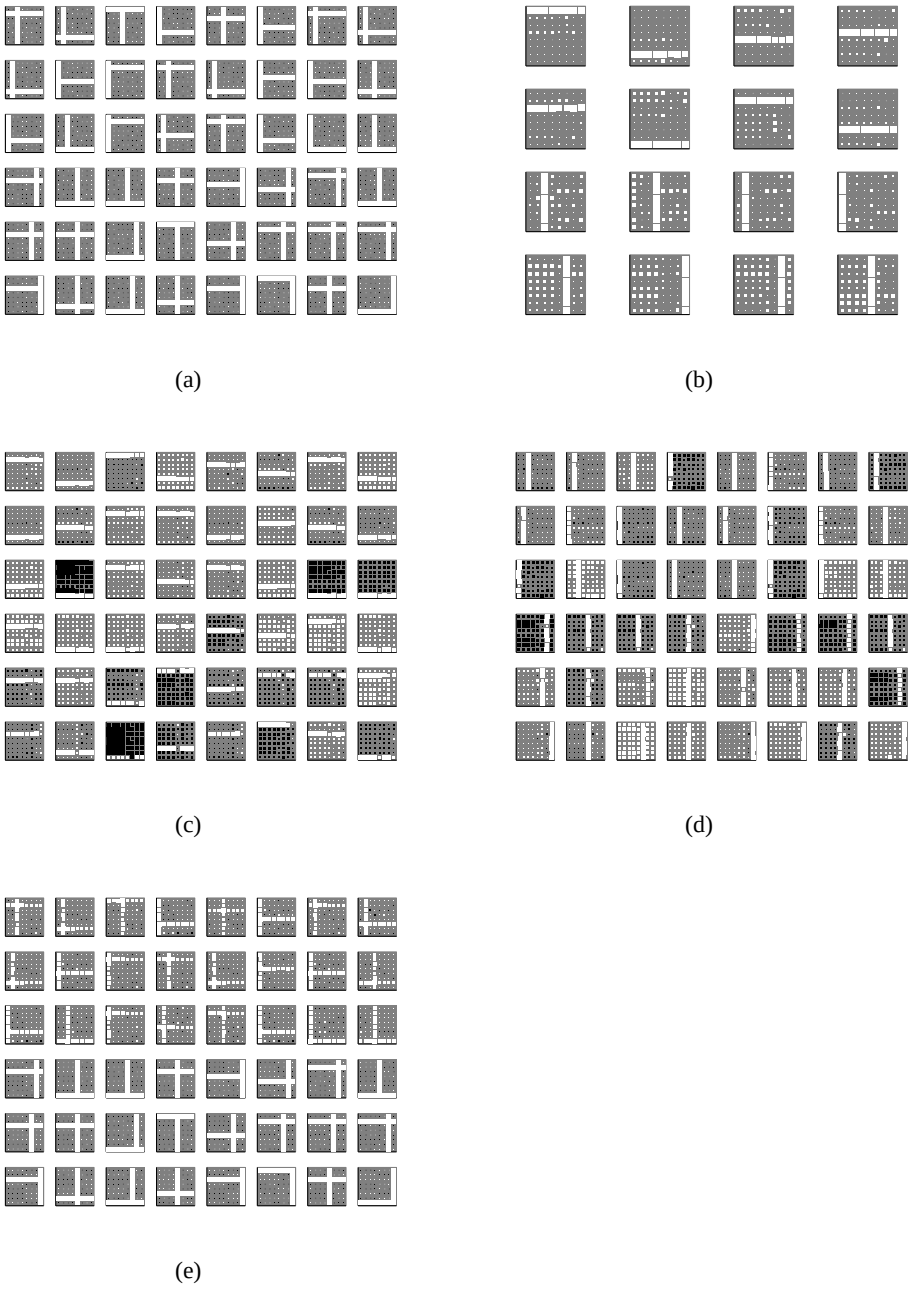

Figure 2: (a) Examples of the training images. The top three rows are from the first data set (the first 24 columns of $\mathbf{Y}_1$), and the bottom three rows are from the second data set (the first 24 columns of $\mathbf{Y}_2$). Each image consists of a horizontal bar chosen at random from the 8 possibilities, which corresponds to the process shared by both sets. The first data set contains a vertical bar chosen at random from the left half of the image, and the second data set contains a vertical bar chosen from the right half of the image. (b) The recovered latent images. The first two rows correspond to the 8 columns of $\mathbf{S}$, and are the shared features i.e. the horizontal bars. The third row corresponds to the 4 columns of $\mathbf{X}_1$, the vertical bars in the left half of the image, and the fourth row corresponds to the 4 columns of $\mathbf{X}_2$,the vertical bars in the right half of the image.The posterior mean of the underlying shared functions is shown in (c) for the first 24 images of $\mathbf{Y}_1$ (top three rows) and $\mathbf{Y}_2$ (bottom three rows). (d) shows the posterior mean of the underlying private functions for $\mathbf{Y}_1$ (top three rows) and $\mathbf{Y}_2$ (bottom three rows). (e) 24 reconstructed images from the first data set $\mathbf{Y}_1$ (top three rows) and the second data set $\mathbf{Y}_2$ (bottom three rows)

[5] D. J. C. MacKay. Probable networks and plausiable predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.