
GP-LVM for Data Consolidation

Carl Henrik Ek
Department of Computing
Oxford Brookes University
cek@brookes.ac.uk

Philip H.S. Torr
Department of Computing
Oxford Brookes University
philiptorr@brookes.ac.uk

Neil D. Lawrence
School of Computer Science
University of Manchester
Neil.Lawrence@manchester.ac.uk

Abstract

Many machine learning tasks are involved with the transfer of information from one representation to a corresponding representation or tasks where several different observations represent the same underlying phenomenon. A classical algorithm for feature selection using information from multiple sources or representations is Canonical Correlation Analysis (CCA). In CCA the objective is to *select* features in each observation space that are maximally correlated compared to dimensionality reduction where the objective is to re-represent the data in a more efficient form. We suggest a dimensionality reduction technique that builds on CCA. By extending the latent space with two additional spaces, each specific to a partition of the data, the model is capable of representing the full variance of the data.

In this paper we suggest a generative model for shared dimensionality reduction analogous to that of CCA.

1 Non-Consolidating Component Analysis

Dimensionality reduction is the task of unsupervised feature selection where the aim is to find a lower dimensional representation of an observed variable. In this paper we are interested in *data-consolidation*. Given multiple sources of information, data consolidation is the task of representing these multiple sources within the same model. Given two sets of corresponding observations $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]^T$ where $\mathbf{y}_n \in \mathbb{R}^{D_Y}$ and $\mathbf{z}_n \in \mathbb{R}^{D_Z}$ we wish to represent both observations using a single variable. Our algorithm is based on a set of assumptions. First we assume that the two data sets can be generated by a smooth noise corrupted function from lower dimensional latent spaces.

$$y_{ni} = f_i^Y(\mathbf{U}^Y) + \epsilon_{ni}^Y, \quad z_{ni} = f_i^Z(\mathbf{U}^Z) + \epsilon_{ni}^Z, \quad (1)$$

We will assume that the two latent spaces \mathbf{U}^Y and \mathbf{U}^Z share a common non-empty subspace $\mathbf{U}^S \subseteq \mathbf{U}^Y \subseteq \mathbf{U}^Z$. We further assume that both the shared subspace \mathbf{X}^S and the remaining completing subspaces $\mathbf{X}^Y = \mathbf{U}^Y - \mathbf{X}^S$ and $\mathbf{X}^Z = \mathbf{U}^Z - \mathbf{X}^S$ can be generated by smooth mappings from the observations.

$$x_{ni}^S = g_i^Y(\mathbf{y}_n) = g_i^Z(\mathbf{z}_n) \quad (2)$$

$$x_{ni}^Y = h_i^Y(\mathbf{y}_n), \quad x_{ni}^Z = h_i^Z(\mathbf{z}_n), \quad (3)$$

Embedding: By applying CCA to the observations we recover a set of directions in each observation space \mathbf{W}^Y and \mathbf{W}^Z explaining the shared variance between \mathbf{Y} and \mathbf{Z} . To recover the partition

specific latent spaces we introduce a novel algorithm called *Non-Consolidating Component Analysis* (NCCA). The directions given by applying CCA already explains a part of the variance in the data. The objective of NCCA is to find further directions, *orthogonal* to the ones given by CCA, explaining the remaining variance. This can be formulated as the following optimization problem,

$$\mathbf{v}_1 = \operatorname{argmax}_{\mathbf{v}_1} \mathbf{v}_1^T \mathbf{C} \mathbf{v}_1$$

subject to: $\mathbf{v}_1^T \mathbf{v}_1 = 1$ and $\mathbf{v}_1^T \mathbf{W} = \mathbf{0}$, (here we have temporarily dropped the partition subscript), \mathbf{W} are the canonical directions and \mathbf{C} is the covariance matrix. The optimal \mathbf{v}_1 is found via an eigenvalue problem,

$$(\mathbf{C} - \mathbf{W} \mathbf{W}^T \mathbf{C}) \mathbf{v}_1 = \lambda_1 \mathbf{v}_1. \quad (4)$$

For successive directions further eigenvalue problems of the form

$$\left(\mathbf{C} - \left(\mathbf{W} \mathbf{W}^T + \sum_{i=1}^{k-1} \mathbf{v}_i \mathbf{v}_i^T \right) \mathbf{C} \right) \mathbf{v}_k = \lambda_k \mathbf{v}_k \quad (5)$$

need to be solved.

The obvious path to non-linearize the suggested algorithm is to apply the kernel-trick to both CCA and NCCA. However, many popular kernels tend to expand the image of the data rather than reducing it. For a close to full-rank kernel matrix this implies that the feature space induced by the kernel are effectively the same which leads to a trivial solution for CCA. Therefore we have as suggested in [3] chosen to first re-represent each observation space by its dominant principal direction in each kernel induced feature space and apply linear CCA and NCCA in this reduced representation.

We follow the approach of [2] and rate the quality of the embeddings for a range of different kernels through the GP likelihood. Intuitively what this means is that if Eq(2) and Eq(3) have correctly “unraveled” the manifold then Eq(1) should also hold and therefore result in a high likelihood.

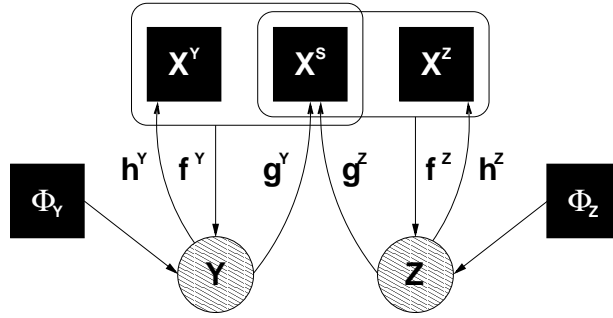


Figure 1: Graphical Model.

2 GP-LVM NCCA

The Gaussian Process Latent Variable Model (GP-LVM) was suggested by Lawrence [4] as a generative model for dimensionality reduction. In the GP-LVM framework the latent variables are treated as the parameters of the model and the location that maximizes the data likelihood is sought.

In [6] a GP-LVM model representing two observation spaces by a single latent variable is presented. The objective is to find the latent locations that maximizes the sum of the data likelihood for each observation space. Even though the model has similarities with CCA the generative formulation of CCA [1], extended to non-linear mappings in [5] is more involved. CCA *selects* and represents portions of variance that is correlated between the different observations. Non-correlated variance is not represented in the CCA model. However, a generative model need to explain the full variance of each observation, correlated and non-correlated. Therefore a generative formulation of CCA also need to “explain away” non-correlated variance. In [1] it is assumed that this non-correlated variance is explained by a Gaussian noise model.

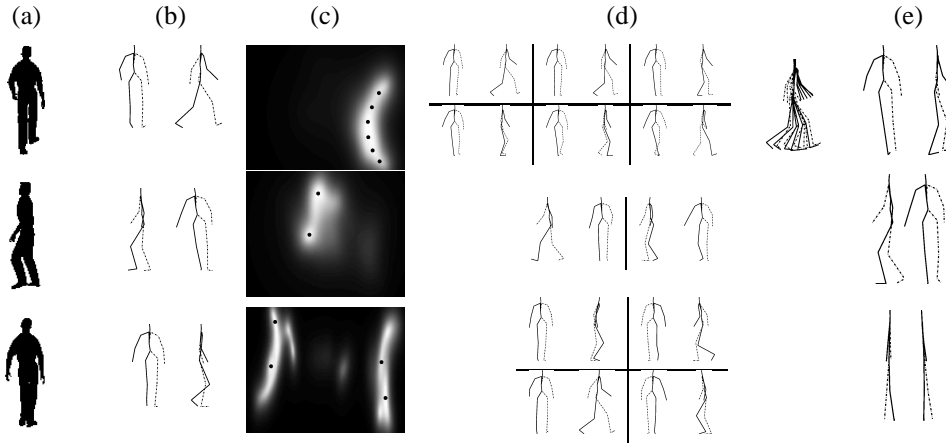


Figure 2: (a) Input Silhouette, (b) Ground Truth Pose, (c) Pose-specific Likelihood, (d) NCCA Prediction, (e) Regression Prediction. *Top Row*: show a silhouette in a face-on-view, in the pose specific likelihood we can see a single mode. Traversing the mode we can see that the pose-specific location with a high likelihood contains poses with different configurations of the legs. *2nd Row*: show a side-one view, from the silhouette it is not possible to disambiguate between the legs. This is reflected by two modes in the pose specific likelihood with each mode representing a different leg configuration. Embedding the data we were surprised that the algorithm was able to disambiguate the heading angle for silhouettes such as the one on the top row. Further examination of the data showed that this ambiguity does not exist in the data as the left arm consistently hangs further away from the body compared to the right throughout the sequence. To confirm that we could capture this ambiguity if present we modified the data to include such silhouettes with the opposite arm configuration. Embedding this modified data it is clear that the algorithm correctly places the information controlling the heading angle in the pose specific latent space. The last column shows as a comparison the results of applying a GP-regression from the silhouette space directly to the pose space. The advantage of our method should be clear compared to a regression based method especially from the output of the regression model for the bottom row.

Constructing a GP-LVM model respecting the subspace structure of the NCCA model is a trivial task. However, for the general case no closed form solution to the GP-LVM objective exists and to proceed the solution is found using gradient based methods. This implies that a initialization of the latent locations \mathbf{X} in a convex region of the global minima is necessary. Therefore a GP-LVM model is reliant on the existence of an analogous method with a convex solution. Using the NCCA solution as initialization the latent locations can be further refined by maximizing the GP-LVM objective.

3 Results

We have applied the suggested model to estimate human pose \mathbf{Z} represented as joint locations in $3D$ space from its corresponding silhouette image \mathbf{Y} . The pose estimation problem from silhouette have been shown to be very hard to model in an efficient way as a silhouette can be associated with several different poses which means it cannot be modeled using a standard regression approach. Due to the high-dimensionality in the data modeling the conditional distribution of silhouettes over poses is very hard. Modeling the data with our model decomposes the estimation in two parts; (1). estimating the shared latent location (2). estimating the pose specific latent location. The shared latent space represent the variance in both data spaces between which there is a functional relationship, it can therefore be found by mapping the silhouette through g^Y . The remaining pose specific latent space contains information about pose which are not present in the silhouette, *e.g.* the different poses that cannot be disambiguated by a specific silhouette. This means is that our latent model allows us to extract as much information as possible from the silhouette and use this information to constrain the possible output poses. After determining the shared latent location \mathbf{x}_*^S corresponding to a specific silhouette \mathbf{y}_* the remaining subspace of the latent representation needs to be determined. This can be done by finding location over the pose specific latent space that maximizes the likelihood in the pose space,

$$\operatorname{argmax}_{\mathbf{x}_*^Z} = p(\mathbf{z} | \{\mathbf{x}_*^S, \mathbf{x}_*^Z\}) = p(\mathbf{z} | \{g_Y(\mathbf{y}_*), \mathbf{x}_*^Z\}).$$

In Figure 1 results are shown of applying NCCA to a benchmark Human Pose data set.

4 Conclusion

We have presented a dimensionality reduction technique that extends the latent space found by Canonical Correlation Analysis with two additional observation specific subspaces. The additional spaces make the model capable of explaining the full variance of each observation space compared to the latent space of CCA which only explains the variance which is shared between the data spaces. We have presented the model both in terms of a convex algorithm and as a analogous GP-LVM model.

References

- [1] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Department of Statistics, University of California, Berkeley, Tech. Rep. 688, 2005.
- [2] S. Harmeling, “Exploring model selection techniques for nonlinear dimensionality reduction,” University of Edinburgh, Tech. Rep. EDI-INF-RR-0960, 2007.
- [3] M. Kuss and T. Graepel, “The geometry of kernel canonical correlation analysis,” Max Planck Institute for Biological Cybernetics, Tübingen, Germany, Tech. Rep. TR-108, 2003.
- [4] N. D. Lawrence, “Probabilistic non-linear principal component analysis with Gaussian process latent variable models,” vol. 6, pp. 1783–1816, 11 2005.
- [5] G. Leen and C. Fyfe, “A Gaussian process latent variable model formulation of canonical correlation analysis,” Bruges (Belgium), 26-28 April 2006 2006.
- [6] A. Shon, K. Grochow, A. Hertzmann, and R. Rao, “Learning shared latent structure for image synthesis and robotic imitation,” *Proc. NIPS*, pp. 1233–1240, 2006.