Learning Alignments from Latent Space Structures

Ieva Kazlauskaite
Department of Computer Science
University of Bath, UK
i.kazlauskaite@bath.ac.uk

Carl Henrik Ek
Faculty of Engineering
University of Bristol, UK
carlhenrik.ek@bristol.ac.uk

Neill D. F. Campbell
Department of Computer Science
University of Bath, UK
n.campbell@bath.ac.uk

Abstract

In this paper we present a model that is capable of learning alignments between high-dimensional data by exploiting low-dimensional structures. Specifically, our method uses a Gaussian process latent variable model (GP-LVM) to learn alignments and latent representations simultaneously. The results show that our model performs alignment implicitly and improves the smoothness of the low dimensional representations.

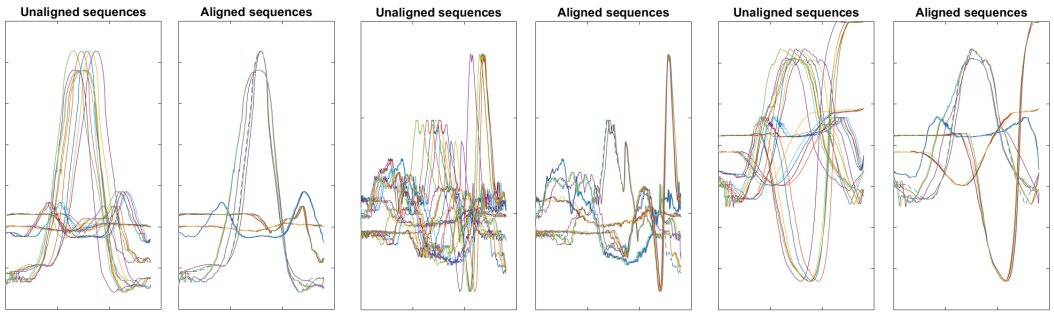
1 Introduction and motivation

Many applications are associated with data that are ordered or sequential. When learning from such data one often has to deal with misalignment due to variations in timing during the capture of each sequence; this is particularly apparent in motion capture data where the events in the sequences are typically not in correspondence. Modelling in such scenarios is challenging since it often requires that the data is aligned as a pre-processing step. Most of the current alignment methods are based on pairwise matching of the given sequences; they suffer from quadratic scaling in cost and exhibit problems of consistency as matching sequences in a loop does not lead to an identity. Another common approach to alignment is constrained energy minimisation which requires an explicit definition of the objective function. In this paper we take a different approach which exploits low-dimensional structures in the data to extract alignment in a completely unsupervised fashion. Specifically, the model learns a probabilistic latent variable model which regularises the alignment of the data, allowing the model to warp each training example so that a simpler structure can be found.

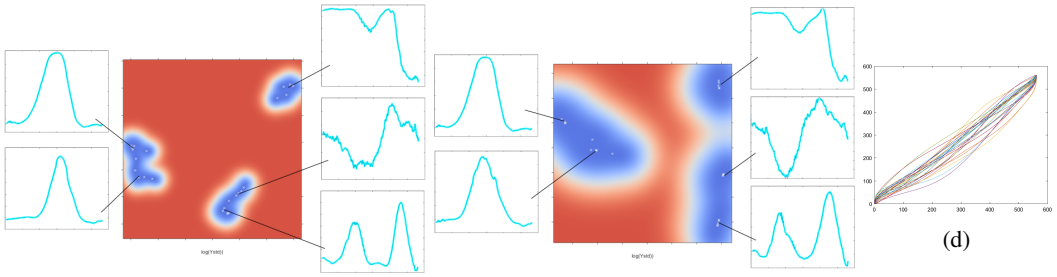
2 Alignment with GP-LVM

Assume that a dataset of observed inputs is stored in a matrix $Y \in \mathbb{R}^{N \times D}$, where N is the number of input points and D is the input dimensionality. The task we are interested in is learning a low-dimensional representation X such that $y_i = f(x_i)$, $X \in \mathbb{R}^{N \times M}$ such that $M \ll D$. The GP-LVM [2] does so by placing a Gaussian process prior over f, a Gaussian prior over X, and seeking a maximum a posteriori estimate of the latent locations and the hyper-parameters of the prior over the mapping. This approach assumes that the observed data is aligned such that each observed dimension is in correspondence. In this paper we explore an approach that is capable of learning alignment of data jointly with the latent variables. In effect it is the regularisation of the low-dimensional representation that facilitates learning alignments as we allow the model to apply monotonic warps to the individual inputs such that the structure of the latent space and the generative mapping becomes as simple as possible.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.



(a) Scaled joint angles against time with and without alignment.



(b) Without alignment.

(c) With alignment.

(d)

Figure 1: Figure (a) shows the alignment results for the 3 joint dimensions, and (d) are the corresponding warping functions. Comparison of manifolds produced without (b) and with (c) alignment in the GP-LVM.

In particular, we define a mapping $\tilde{y}_{nd} = g(y_{nd})$, that is parametrised using a set of K smooth monotonically increasing basis functions, e.g. sigmoid and logarithmic functions, and corresponding weights w_k. The resulting log-likelihood is as follows:

$$L = -\frac{DN}{2}\log(2\pi) - \frac{D}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{Tr}(\mathbf{K}^{-1}g(\mathbf{Y};\mathbf{w})g(\mathbf{Y};\mathbf{w})^T), \tag{1}$$

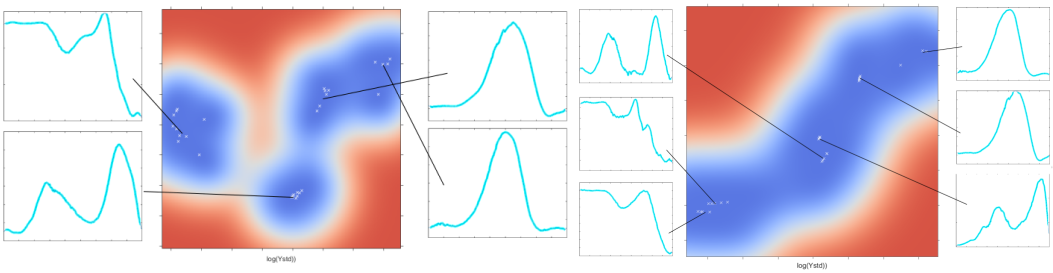
where $\mathbf{K}$ is a chosen covariance matrix, and the mapping $g(\mathbf{Y};\mathbf{w})$ parametrised by $\mathbf{w}$ appears in the data fitting term. Priors over the parameters of the covariance function, the latent variables and the weights of the warping are included to find a MAP estimation. By imposing $\sum_{k\in K} w_k = 1, w_k \geq 0 \,\forall\, k \in K$, we ensure that the resulting mapping g is smooth and monotonic increasing thus allowing the input vectors to be warped but not permuted. The weights are extra parameters of the GP-LVM, and are optimised along with the latent variables and the hyper-parameters. The method is able to find a good minimum given a handful of random initialisations, for which the optimisation is performed in parallel.

In [3] the authors construct a GP with a warped input space to account for differences in observations (e.g. inputs may vary over many orders of magnitude), and show that a warped GP finds the standard preprocessing transforms, such as the logarithm, automatically. In comparison, our approach leads to a warped output space of the GP-LVM, and uses the additional knowledge of possible misalignments in the high-dimensional space to regularise the problem of building a low-dimensional latent space.

3 Experiments

3.1 Toy dataset

We evaluate the performance of our model on a small set of motion capture data from the CMU database [1], where each input vector contains the motion of one joint only. Our first experiment contains five random warpings of five different motions, resulting in five groups of data that contain small random variations within each group. Fig. 1 illustrates how the GP-LVM favours the simplified, i.e. aligned, inputs; the log-likelihood maximisation gives a set of warping weights which produce a fine alignment of the input sequences within each of the groups, and consequently the resulting



(a) Without alignment.

(b) With alignment.

Figure 2: Comparison of manifolds produced without (a) and with (b) alignment in the GP-LVM.

Table 1: Alignment quality for different number of basis functions

Number of basis functions	9	11	17	40
Sum of mean squared errors	98.6	98.0	97.5	94.7

two-dimensional manifold offers a good separation of these groups. We note that the manifold produced using GP-LVM without alignment contains more isolated areas, and the separation of the five groups is not well defined. Therefore, our implicitly aligned model is able to generate smoother transitions in the manifold, producing high quality predicted outputs.

The second experiment uses two specific motion sequences from three types of motions. The model is trained using five random variations of each of the six motions; thus in the manifold we expect to see three groups of latent variables each with two subgroups. Fig. 2 shows precisely such structure in the manifold, and the close positioning of the points within each subgroup is a sign of a successful alignment of the randomly distorted training inputs. In comparison, the traditional GP-LVM fails to find the subgroups within the dataset.

3.2 Comparison to energy minimisation

We consider a set of motion capture clips that include four golf swings, five putts and five motions of placing a ball on a tee (two of these involve bending the legs while the remaining three keep the legs straight). The set of basis functions used in this experiment contains equally spaced instances of sigmoid and logarithmic functions scaled to cover the interval $[0, 1]$. Increasing the number of basis functions from 9 to 40 slightly improves the quality of alignment within each group of motions in terms of the sum of mean squared errors, see Table 1. Large sets of basis functions increase the computational cost as a larger number of parameters need to be optimised; in the following experiments we use 17 basis functions.

Next, we compare the alignments produced by our method and by an energy minimisation algorithm with the same basis functions. In the case of energy minimisation the alignment is performed by comparing the sequences frame by frame, and minimising the objective function which describes how well the sequences match each other using the given temporal warping functions. The energy is defined as the sum of squared differences between the quaternion coordinates of the corresponding joints in the motion sequences.

The GP-LVM recognises four distinct groups of motions in the dataset, automatically separating the motions of placing a ball into two groups, and aligning the motions within each of the groups. The energy alignment approach requires the user to specify the existence of the four distinct motions in the dataset. This requirement is in order to align similar motions in each group as opposed to aligning all the motions to a single sequence. The qualitative comparison of alignments of the four swing motions found by the two methods is shown in Fig. 3. The performance of the two methods is also contrasted by calculating the squared distances between all pairs of sequences within each group. Table 2 shows that the GP-LVM produces as good an alignment of the sequences as the more specialised energy approach without the need for the user to pick an energy function by hand or to provide information on the different groups of motion present in the dataset.

Table 2: Quantitative comparison of alignments.

	Swing	Putt	Placing		Total	
			Together	2 groups	Together	2 groups
Energy alignment	35.7	27.8	203.6	39.4	267.1	**102.9**
GP-LVM	-	-	-	-	-	**103.0**

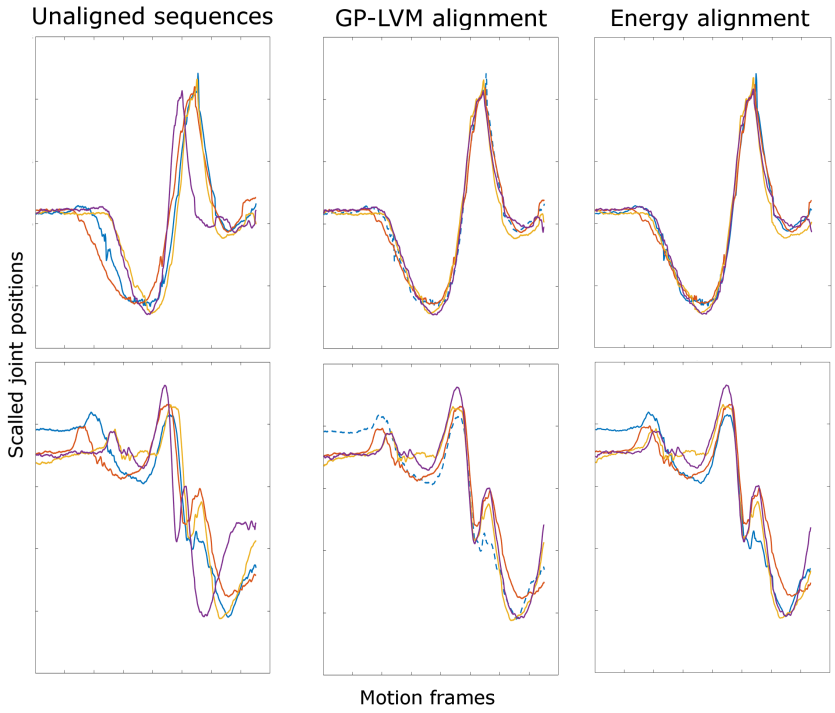


Figure 3: Comparison of alignments produced using the GP-LVM and the energy minimisation. The rows correspond to one dimension of the motion of two distinct joints.

Our tests also demonstrate that the traditional non-convex energy minimisation approach to alignment requires the use of the coarse-to-fine estimation in optimisation. A coarse-to-fine approach is used to try to prevent the optimisation from converging to a poor local minimum; energy minimisation is first performed on the data that has been substantially smoothed, gradually adding high frequency detail; in the final optimisation call the original data is used. On the other hand, the GP-LVM is able to find a good minimum given four random initialisations of the weights vector.

4 Conclusion and future work

We have presented an extension to the traditional GP-LVM that is able to implicitly align the inputs. The corresponding low dimensional manifolds exhibit better smoothness properties and generate more consistent outputs. The proposed approach removes the need for ad hoc pre-processing necessary when the inputs are not in correspondence, and proves beneficial both as part of a dimensionality reduction technique, and as an alignment tool. We demonstrate that on the motion capture dataset the GP-LVM performs as well as the less generic and more laborious energy minimisation approach. In the future we will further explore the convergence properties of our approach, test the framework on additional datasets, including multi-modal data, and aim to extend the class of alignments used by the model.

References

[1] Carnegie Mellon Graphics Lab. Motion Capture Database . "http://mocap.cs.cmu.edu/info.php, 2016. [Online; accessed 12-September-2016].

[2] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16(3):329–336, 2004.

[3] Edward Snelson, Zoubin Ghahramani, and Carl E. Rasmussen. Warped gaussian processes. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 337–344. MIT Press, 2004.