DP-GP-LVM: A Bayesian Non-Parametric Model for Learning Multivariate Dependency Structures

Andrew R. Lawrence [1] **Carl Henrik Ek** [2] **Neill D. F. Campbell** [1]

Abstract

We present a non-parametric Bayesian latent variable model capable of learning dependency structures across dimensions in a multivariate setting. Our approach is based on flexible Gaussian process priors for the generative mappings and interchangeable Dirichlet process priors to learn the structure. The introduction of the Dirichlet process as a specific structural prior allows our model to circumvent issues associated with previous Gaussian process latent variable models. Inference is performed by deriving an efficient variational bound on the marginal log-likelihood of the model. We demonstrate the efficacy of our approach via analysis of discovered structure and superior quantitative performance on missing data imputation.

1 Introduction

Latent variable models provide data-efficient and interpretable descriptions of data. By specifying a generative model, it is possible to achieve a compact representation through exploiting dependency structures in the observed data. Their probabilistic structure allows the model to be integrated as a component in a larger system and facilitates tasks such as data-imputation and synthesis.

Efficient representations can be achieved when the intrinsic dimensionality of the data is much lower than in its observed representation. Traditional approaches, such as probabilistic PCA (Tipping & Bishop, 1999) and the GP-LVM (Lawrence, 2005), assume that the data lie on a single low-dimensional manifold embedded in the high-dimensional space. However, in many scenarios, this assumption is too simplistic as more intricate dependency structures are present in the data. In specific, there are many situations where groups of dimensions co-vary. For a human walking, each limb shares a variation from the direction of travel. One would expect that both arms share information not always present in the lower limbs; however, it is conceivable that the left-side limbs share information not present on the right.

Variations that are not common to *all* dimensions are challenging to model. If included in the representation, a variation only present in a subset of dimensions will *pollute* the representation of the dimensions that do not share this characteristic.

One approach to circumvent this issue is to learn a factorized latent representation where independent latent variables describe each group of variations. An example of such approach is the Inter-Battery Factor Analysis (IBFA) (Tucker, 1958) model, where the latent space consists of dimensions encoding structure shared across all dimensions separately from structure that is private within a group of variates. This model, and the approaches building on this idea, assume that the grouping of the observed dimensions is known a priori. Importantly, this means that the learning task is to recover a latent representation that reflects a given grouping of the dimensions in the observed space.

Even for familiar data, such as human motion, specifying these groupings is challenging, while, in other tasks, extracting the groupings themselves is essential. We refer to these disjoint groupings as *views*. One such example is a medical scenario where each observed variate corresponds to a specific, potentially costly and for the patient intrusive, medical test. If we can learn the groupings of variations we can potentially reduce the range of tests needed for diagnosis.

In this paper, we describe a latent variable model, which we term the DP-GP-LVM, that automatically learns the grouping of the observed data thereby removing the need for a priori specification. By formulating the generative model non-parametrically, our approach has unbounded representative power and can infer its complexity from data. The Bayesian formulation enables us to average over all possible groupings of the observations allowing the structure to emerge naturally from the data. We perform approximate Bayesian inference by optimizing a lower bound on the marginal likelihood of the model.

[1]Dept. of Computer Science, University of Bath, UK [2]Dept. of Computer Science, University of Bristol, UK. Correspondence to: Andrew R. Lawrence <A.R.Lawrence@bath.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

2 Background

Finding latent representations of data is a task central to many machine learning applications. By exploiting dependencies in the data, more efficient low dimensional representations can be recovered. Of specific importance is the work of Spearman (1904) where the interpretation of a latent dimension as a *factor* was introduced. The traditional factor analysis model is unidentifiable meaning additional assumptions need to be incorporated, for example, the Gaussian assumption which leads to PCA (Hotelling, 1933).

A second approach is to introduce groupings of the observed data, as is done in CCA (Hotelling, 1936), where the latent representation that best describes the correlation between the groups is sought. While PCA was described as a model, it is challenging to describe the generative procedure of CCA. IBFA (Tucker, 1958; Kristof, 1967) is a less known model that exploits groupings of the observed data by introducing two different classes of latent factors: *shared* and *private*, where the former represent variations common to all groups while the latter variations only belong to a single group. As discussed in § 1, this factorization is important when specifying a generative model.

A Bayesian formulation of IBFA was proposed by Klami et al. (2013) and a non-linear extension, based on Gaussian processes, called Manifold Relevance Determination (MRD) by Damianou et al. (2012). However, there exists an important distinction between the two models that is rarely highlighted. The linear formulation of IBFA allows the groupings of the data to be inferred while, in the MRD model, the *views* needs to be set a priori.

In this paper, we present a model that combines the benefits of both: a non-linear, Bayesian model that allows the groupings to emerge from data. Our proposed model has commonalities with MRD and, being motivated by its shortcomings, we now proceed to describe it in detail.

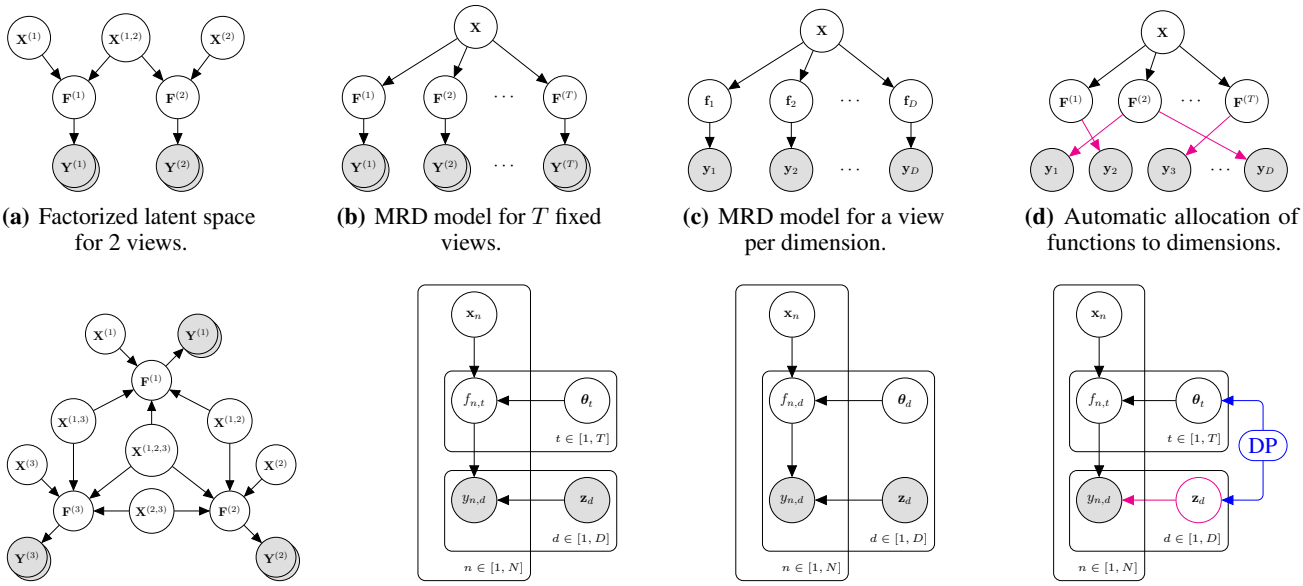
Latent Variable Models The task in unsupervised learning is to learn a latent representation $\mathbf{X} \in \mathbb{R}^{N \times Q}$ from a set of multivariate observations $\mathbf{Y} \in \mathbb{R}^{N \times D}$. We have N as the number of observations while Q and D are the dimensionality of the latent and observed data, respectively, and $Q \ll D$. We denote a single observation $\mathbf{y}_n, n \in [1, N]$, as a D-dimensional vector $\mathbf{y}_n \in \mathbb{R}^D$. The generative model specifies the relationship between the latent space and the observed, $\mathbf{y}_n = f(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$, where the form of the noise $\boldsymbol{\epsilon}_n$ leads to the likelihood of the data.

MRD is a member of a larger class of models called Gaussian Process Latent Variable Models (GP-LVM) (Lawrence, 2005), where a Gaussian process prior (Rasmussen & Williams, 2005) is placed over the generative mapping $f(\cdot)$. Under the assumption of Gaussian noise, it is possible to marginalize over the whole space of functions leading to a rich and expressive model. Due to the non-linearities of $f(\cdot)$, integration of the latent variables cannot be achieved in closed form. Inference of $\mathbf{X}$ can either be achieved through maximum-likelihood or via approximate integration by optimizing a variational lower bound on the marginal likelihood (Damianou et al., 2016b).

Multiple Views We use the term *views* to refer to natural groupings within a set of data; data within a specific view will share a generative structure; in concrete, a set of observed dimensions. Thus, views are observed data that are aligned in terms of samples but of a different modality or a disjoint group of observed dimensions. For example, consider a data set consisting of silhouettes of people, the angles of the joints between their limbs and the background appearance of a room. These would consist of three views (silhouettes, angles and appearance), where the silhouettes and angles have shared (the pose of the body) and private (clothes affect the silhouette but not the joints) information. The appearance of the background is a third view that should be independent of the first two since it has no causal link. The concept of views was first introduced in the GP-LVM framework by Shon et al. (2006) where the observed data was grouped into two sets of views sharing a single set of latent variables. Ek et al. (2008) and Salzmann et al. (2010) extended this, introducing the idea of a factorization with shared and private latent spaces from IBFA.

Model Evolution We now describe the sequence of graphical models in Fig. 1 from the model of Ek et al. (2008) in (a) to our proposed model in (h). The model by Ek et al. (2008) extends naturally beyond two separate groupings. However, due to the fixed structure of the latent space, it leads to a combinatorial explosion in the number of latent variables as illustrated by Fig. 1(e). Further, learning is challenging as the dimensionality of each latent space needs to be known a priori. To circumvent these issues, the MRD model, Fig. 1(b), treats the factorization as part of the GP prior. This GP prior is completely specified by its mean and covariance functions. For most unsupervised tasks a zero mean function is assumed, leaving only the covariance function as its parameterization.

The introduction of Automatic Relevance Determination (ARD) (Neal, 1996) covariance functions allows the MRD to enclose the factorization into the GP prior. In a stationary kernel, the covariance between two latent variables is a function of the distance between the points. Rather than a spherical distance function, the ARD version introduces a parametrized diagonal Mahalanobis distance that is learned independently for each view, represented by $\boldsymbol{\theta}$ in Fig. 1 (f)-(h). The interpretation is that if the distance function *switches off* an axis, this view becomes independent of the corresponding latent dimension; thus, the factorization can be determined by the non-zero ARD weights. However, this approach leads to additional problems as the ARD pa-



(a) Factorized latent space for 2 views.

(b) MRD model for T fixed views.

(c) MRD model for a view per dimension.

(d) Automatic allocation of functions to dimensions.

(e) Factorized latent space for 3 views.

(f) Known allocation version of (b).

(g) Known allocation version of (c).

(h) Unknown allocation version of (d).

Figure 1. Graphical models for the factorized GP-LVM, MRD, and DP-GP-LVM. The term *views*, denoted with $\mathbf{Y}^{(i)}$, refers to a known group of output dimensions. (a) The factorized GP-LVM allows for shared ($\mathbf{X}^{(1,2)}$) and private ($\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$) latent variables. (e) The addition of new observations leads to a combinatorial explosion in the number of latent variables. (b) MRD uses a single latent space with ARD parameters to allow sharing and can be represented in collapsed form (f) using a **known** assignment variable $\mathbf{Z}$ that defines a one-to-one mapping that allocates a *view* to a generative functional prior. (b) and (f) represent MRD with T known views, where each view contains at least one observed dimension; therefore $T \leq D$; while (c) and (g) represent fully independent MRD (fi-MRD), where each observed dimension is assigned to its own view. (d) In an ideal model, we would infer the sharing of functional priors (magenta) between observed dimensions from data. We would also like to infer T, the number of such functional priors, automatically from data. (h) In our DP-GP-LVM, we learn the **unknown** allocation $\mathbf{Z}$ automatically using a Dirichlet process prior (blue).

rameters can also be interpreted as an inverse length scale. This means that a small ARD value for a specific dimension could have two different causes; either that the dimension varies linearly with the view or because it is invariant to the view (Vehtari, 2001). While the MRD only considers the latter cause, we explicitly model both these cases; this is our first extension to the MRD.

Inference of Views (Groupings) In MRD, the groupings of the observed variates must be specified a priori. This (i) restricts the data that can be used, (ii) can be very challenging to specify without supervision, and (iii) means the representation will be sensitive to changes or errors in the grouping provided. One approach to circumvent this is referred to as fully independent MRD (Damianou et al., 2016a), as in Fig. 1 (c) and (g), where a separate functional prior is used for each dimension followed by a post-processing clustering step, which means the model is no longer generative. In addition to the unsatisfactory post-processing, this will lead to a notable increase in the number of parameters as a Mahalanobis metric needs to be learned for each output dimension. Using MRD to infer views then learning a generative model requires three steps: (i) train fully independent MRD to learn ARD weights, (ii) cluster weights to infer views, (iii) retrain MRD using these clusters as views. The requirement of the views to be known a priori is a major limitation of MRD; our proposed model, defined in § 3, is generative and learns the appropriate groupings and latent representation with a single objective function.

In this paper, we introduce a specific unknown indicator variable $\mathbf{z}_d$ that determines which latent dimensions will be associated with each output dimension as in Fig. 1(h). Further, to control the structure of the latent space, we introduce a Dirichlet process (DP) prior to specify prior knowledge of the complexity of the latent representation.

Related Models using Stochastic Processes Previous models have combined elements of GPs with DPs. Mixtures of GP experts place a Gaussian mixture model on the input space then fit each GP to the data belonging to the specific components. An infinite mixture of GP experts uses a DP to determine the number of components. The main works using this approach are by Rasmussen & Ghahramani (2002) and Meeds & Osindero (2006), who use MCMC to approximate the intractable posterior, as well as Yuan & Neubauer (2009) and Sun & Xu (2011), who use variational inference. In addition, Hensman et al. (2015) combines a DP and GP for the purpose of clustering time-series data streams. However, as with the mixture of GP experts, their

model focuses on supervised learning and does not address the unsupervised task that we are studying.

The topic is related to work by Palla et al. (2012) on variable clustering using DPs to infer block-diagonal covariance structures in data. Wood et al. (2006) defined a prior over a number of hidden causes and used reversible jump MCMC to approximate a distribution over causal structures in data.

A related avenue of investigation is the multi-output GP literature, e.g., Álvarez & Lawrence (2009); Álvarez et al. (2010); Álvarez & Lawrence (2011); Dai et al. (2017). These works, looking at transfer learning or filling in missing data, also produce structured models and a number have made use of the Indian Buffet Process (in contrast to a DP) to control complexity and favor sparse explanations.

3 DP-GP-LVM

We now describe the proposed DP-GP-LVM. We assume that the observed data are generated as a function of some unknown latent variables $\mathbf{X} \in \mathbb{R}^{N\times Q}$ where the N observations each come from a lower Q-dimensional latent space such that $Q \ll D$, the number of output dimensions. Thus, we have each output dimension $\mathbf{y}_d = f_d(\mathbf{X}) + \boldsymbol{\epsilon}_d$ where $f_d(\cdot)$ is some function and $\boldsymbol{\epsilon}_d \sim \mathcal{N}(\mathbf{0}, \beta_d^{-1}\mathbf{I}_N)$ is zero mean iid Gaussian noise with precision β_d. We put a standard Gaussian prior over the latent space,

$$p(\mathbf{X}) = \textstyle\prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q|\mathbf{0}, \mathbf{I}_N) \ , \tag{1}$$

and place a zero mean Gaussian process prior over the functions $f_d(\cdot)$ such that:

$$f_d(\mathbf{X}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}'|\boldsymbol{\theta}_d)) \ , \tag{2}$$

$$p(\mathbf{F}|\mathbf{X}) = \textstyle\prod_{d=1}^{D} \mathcal{N}(\mathbf{f}_d|\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}_d}) \ , \tag{3}$$

where $\mathbf{f}_d \in \mathbb{R}^N$ denotes the evaluation of the function at the latent locations $\mathbf{X}$ and $\mathbf{K}_{\boldsymbol{\theta}_d} = k(\mathbf{X}, \mathbf{X}'; \boldsymbol{\theta}_d)$ denotes the evaluation of some covariance function $k(\cdot, \cdot)$ with hyperparameters $\boldsymbol{\theta}_d$. As for the other random variables, we use $\mathbf{F}$ to denote the concatenation of $\mathbf{f}_d$ across d. The observed data is then obtained from these latent functions through a likelihood to model the Gaussian noise,

$$p(\mathbf{Y}|\mathbf{F}) = \textstyle\prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d|\mathbf{f}_d, \beta_d^{-1}\mathbf{I}_N) \ . \tag{4}$$

Sharing Functional Priors As discussed in § 2, we assume that our multivariate observations are not all independent but will potentially share generative structure. From this assumption, there are two properties we seek to encode in our model. Firstly, we would like to encourage the observations to be grouped together, when the data supports it, and share a common generative functional prior; that is,

$$\mathbf{f}_{d'} \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}'|\boldsymbol{\theta}_t)) \quad \forall \quad d' \in \mathcal{D}_t \ , \tag{5}$$

where $\mathcal{D}_t$, $t \in [1, \infty]$ denotes a grouped subset of observed dimensions such that $\bigcup_{t=1}^{\infty} \mathcal{D}_t = [1, D]$.

Secondly, we do not know a priori how many groupings there are nor what these groupings should be; therefore, $\{\mathcal{D}_t\}$ must be inferred from the data itself. In general, there could be an infinite set of potential groupings, however in practice $|\{\mathcal{D}_t\}| \leq D$. We now describe how we achieve the sharing of functional priors and the inference over groupings, a key contribution of our approach.

Function Parameterization The differences in the shared functional priors in (5) are encoded by the hyperparameters of the covariance functions in the GP prior (2). We adopt a covariance function that makes use of ARD to infer a subset of the Q latent dimensions to be used. In particular, we use an exponentiated quadratic covariance function,

$$k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}_t) = {\sigma_t}^2 \exp\Big(-\tfrac{1}{2}\textstyle\sum_{q=1}^{Q} \gamma_{t,q}(x_{i,q} - x_{j,q})^2\Big) \ , \tag{6}$$

where the hyperparameters $\boldsymbol{\theta}_t = [\sigma_t^2, \boldsymbol{\gamma}_t]$ are the signal variance σ_t^2 and the positive ARD weights $\boldsymbol{\gamma}_t \in \mathbb{R}_+^Q$. We observe that if $\gamma_{t,q'} \to 0$ then the function has no dependence on the q' dimension of the latent space; the function is independent of this latent dimension.

Dirichlet Process Prior As the grouping dependence in (5) is encoded in the hyperparameters, we can express our preference for sharing, and perform inference over the groupings, by placing a Dirichlet process prior over the hyperparameters (and noise precision) of the covariance functions for each observed dimension d. The DP consists of a base measure G_0 and a concentration parameter α. We use a wide log-normal distribution as the base measure G_0.

The DP produces a discrete distribution G whose support consists of a countably infinite ($t \in [1, \infty]$) set of kernel hyperparameters $\{\boldsymbol{\theta}_t\}$ and noise precisions $\{\beta_t\}$ independently drawing from G_0. By drawing the kernel hyperparameters and noise precision from a DP,

$$\boldsymbol{\theta}_d, \beta_d \sim G, \ G \sim \mathcal{DP}\big(\alpha, \log \mathcal{N}(\mathbf{0}, \mathbf{I}_{Q+2})\big) \ , \tag{7}$$

the hyperparameters will be clustered; all the output dimensions sharing the same set of hyperparameters are effectively combined to form the set of groupings $\mathcal{D}_t$.

Stick-Breaking Construction To represent the discrete distribution G, we use the stick-breaking construction of a DP. We obtain an infinite set of stick lengths $v_t \in [0, 1]$ through independent draws from a beta distribution,

$$p(v_t|\alpha) = \text{Beta}(v_t|1, \alpha) \ , \tag{8}$$

using the concentration parameter α. From these stick lengths, we obtain a vector of mixing proportions $\pi_t(\mathbf{V}) = v_t \prod_{i=1}^{t-1}(1 - v_i)$. We also obtain an infinite set of kernel hyperparameters and noise precisions $\{\boldsymbol{\theta}_t, \beta_t\}$ through

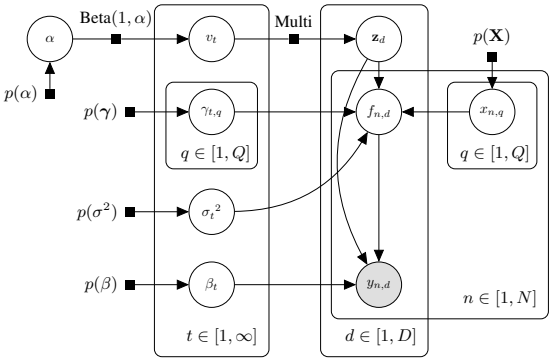


Figure 2. Graphical model of DP-GP-LVM. The grey node $y_{n,d}$ is observed while all the white nodes represent latent random variables. During inference, all the latent variables are marginalized, except the kernel hyperparameters and noise precision $\{\boldsymbol{\gamma}_t, \sigma_t^2, \beta_t\}$, where MAP estimates are used.

independent draws from G_0,

$$p(\boldsymbol{\theta}_t, \beta_t | G_0) = \log \mathcal{N}(\boldsymbol{\theta}_t, \beta_t | \mathbf{0}, \mathbf{I}_{Q+2}) \,. \tag{9}$$

G is constructed as $G = \sum_{t=1}^{\infty} \pi_t(\mathbf{V}) \delta_{\{\boldsymbol{\theta}_t, \beta_t\}}$. We introduce an assignment variable $\mathbf{z}_d$ which associates the kernel hyperparameters and noise precision for observed dimension d to a specific draw from G_0. We obtain the assignment variables through independent draws from a multinomial distribution,

$$p(\mathbf{z}_d | \mathbf{V}) = \text{Mult}(\mathbf{z}_d | \pi(\mathbf{V})) \,, \tag{10}$$

where we may consider $\mathbf{z}_d$ as a one-hot encoding vector of dimension d belonging to set $\mathcal{D}_t$.

The Full Model We combine all these terms to produce the full graphical model of Fig. 2. The full joint distribution factorizes as,

$$\begin{aligned} p(\mathbf{Y}, \mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha, \boldsymbol{\Theta}, \boldsymbol{\beta}) &= p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\beta}, \mathbf{Z})\, p(\boldsymbol{\Theta}, \boldsymbol{\beta}) \\ & p(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z})\, p(\mathbf{X})\, p(\mathbf{Z}|\mathbf{V})\, p(\mathbf{V}|\alpha)\, p(\alpha) \,, \end{aligned} \tag{11}$$

where the individual factors have been defined in (4), (9), (3), (1), (10), (8), and we use a wide gamma prior over the concentration parameter $p(\alpha) = \text{Gamma}(\alpha | s_1, s_2)$. During learning, we deal with the intractable marginalizations using variational inference as detailed in § 3.2.

Although (4) and (3) are not conditioned on $\mathbf{Z}$, they are conditioned on β_d and $\boldsymbol{\theta}_d$, respectively. Given the DP prior, $\boldsymbol{\theta}_d = \prod_{t=1}^{\infty} \boldsymbol{\theta}_t^{[\mathbf{z}_d = t]}$ and $\beta_d = \prod_{t=1}^{\infty} \beta_t^{[\mathbf{z}_d = t]}$, where $[\cdot]$ is the Iverson bracket notation for the indicator function.

3.1 Special Cases of DP-GP-LVM

DP-GP-LVM can be seen as a generalization of both the Bayesian GP-LVM (BGP-LVM) (Damianou et al., 2016b) and the MRD (Damianou et al., 2012) model. We show this with reference to the full model of (11) and Fig. 2. In the BGP-LVM, the observed dimensions are assumed to be iid draws from the same functional prior. This is captured in our model by taking the limiting case of a single cluster from the DP (such that $\alpha \to 0$). In this setting, we have a single set of hyperparameters (and noise precision) shared across all dimensions d. Additionally, latent variables $\{\mathbf{Z}, \mathbf{V}, \alpha\}$ may be removed from the model. In practice, BGP-LVM is recoverable as α can tend towards zero if it fits the data.

Similarly, in the case of MRD, illustrated in Fig. 1 (b), (f), the grouping structure is specified a priori and not inferred from the data. In this instance, the allocation variable $\mathbf{Z}$ becomes observed (dictating the known allocation of dimensions into a finite set of T groups $\{\mathcal{D}_t\}, t = [1, T]$) and the model collapses to that of MRD. The observation of $\mathbf{Z}$ renders the latent variables $\{\mathbf{V}, \alpha\}$ unnecessary.

3.2 Learning

To perform learning of the joint model of (11) we would like to marginalize out the latent variables $\{\mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha\}$ and take MAP estimates over the hyperparameters and noise precisions $\{\boldsymbol{\Theta}, \boldsymbol{\beta}\}$. This corresponds to optimizing the marginal log-likelihood of the observed data,

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}) = \log \int & p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\beta}, \mathbf{Z})\, p(\boldsymbol{\Theta}, \boldsymbol{\beta})\, p(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z}) \\ & p(\mathbf{X})\, p(\mathbf{Z}|\mathbf{V})\, p(\mathbf{V}|\alpha)\, p(\alpha)\, \mathrm{d}\{\mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha\} \,. \end{aligned} \tag{12}$$

Unfortunately, a number of these integrals are intractable and cannot be found in closed form. To make progress, we introduce variational distributions to approximate the posteriors over the latent variables, in a similar manner to Damianou et al. (2016b) and Blei & Jordan (2005), and then optimize the Evidence Lower Bound (ELBO) directly.

Lower Bound We introduce a fully factorized variational distribution $q(\mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha)$ over the latent variables. With $\boldsymbol{\Omega} = \{\mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha\}$ as the latent variables, we have:

$$\begin{aligned} \log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}) &= \log \int q(\boldsymbol{\Omega}) \frac{p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Omega})}{q(\boldsymbol{\Omega})} \, \mathrm{d}\boldsymbol{\Omega} \\ &\geq \mathbb{E}_q \left[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Omega}) \right] + \mathbb{H}[q(\boldsymbol{\Omega})] := \mathcal{L} \,, \end{aligned} \tag{13}$$

with the lower bound as $\mathcal{L}$. We decompose the lower bound into expressions from the GP ($\{\mathbf{F}, \mathbf{X}, \mathbf{Z}\}$) and the DP ($\{\mathbf{Z}, \mathbf{V}, \alpha\}$) such that

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z})} \left[\mathcal{L}_{\mathcal{GP}} \right] + \mathcal{L}_{\mathcal{DP}} + \log p(\boldsymbol{\Theta}, \boldsymbol{\beta}) \,. \tag{14}$$

GP Approximating Distributions As noted by Damianou et al. (2016b), the lower bound on the GP,

$$\int_{\mathbf{F}, \mathbf{X}} q(\mathbf{F})\, q(\mathbf{X}) \, \log \frac{p(\mathbf{Y}|\mathbf{F}, \boldsymbol{\beta}, \mathbf{Z})\, p(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z})\, p(\mathbf{X})}{q(\mathbf{F}) q(\mathbf{X})} \,, \tag{15}$$

is still intractable due to the presence of $\mathbf{X}$ inside the covariance function in $p(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z})$. We make progress by

extending the output space of the GP with a random variable $\mathbf{U}$ drawn from the same GP at some *pseudo input locations* $\mathbf{X}_\text{u}$ such that $\mathbf{u}_d \sim \mathcal{GP}(0, k(\mathbf{X}_\text{u}, \mathbf{X}'_\text{u}|\boldsymbol{\theta}_d)$. These locations are taken as variational parameters and are optimized within the lower bound.

If we assume that the $\mathbf{U}$ form a sufficient statistic for the outputs $\mathbf{F}$, then we have $p(\mathbf{F}, \mathbf{U}|\mathbf{X}, \mathbf{X}_\text{u}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X})p(\mathbf{U}|\mathbf{X}_\text{u})$. Further, if we assume that the approximating distribution factorizes as $q(\mathbf{F}, \mathbf{U}, \mathbf{X}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X})q(\mathbf{U})q(\mathbf{X})$ then we have

$$\int_{\mathbf{F},\mathbf{U},\mathbf{X}} p(\mathbf{F}|\mathbf{U},\mathbf{X})\, q(\mathbf{U})\, q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{F},\boldsymbol{\beta},\mathbf{Z})\, p(\mathbf{U}|\mathbf{X}_\text{u})\, p(\mathbf{X})}{q(\mathbf{U})\, q(\mathbf{X})} , \quad (16)$$

where all the terms are tractable. The optimal $q(\mathbf{U})$ is found to be Gaussian through variational calculus (Damianou et al., 2016b) and marginalized out in closed form. A fully factorized Gaussian form is taken for $q(\mathbf{X}) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{x}_q|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, where $\boldsymbol{\Sigma}_q$ are assumed diagonal.

GP Bound This leads to a GP lower bound of

$$\mathbb{E}_{q(\mathbf{Z})}\left[\mathcal{L}_{\mathcal{GP}}\right] = \sum_{d=1}^{D} \mathcal{F}_d - \text{KL}(q(\mathbf{X}) \,\|\, p(\mathbf{X})) , \quad (17)$$

where the free energy $\mathcal{F}_d$ is given by:

$$\begin{aligned}\mathcal{F}_d = &\frac{N}{2}\log\beta_d + \frac{1}{2}\log\left|\mathbf{K}_{\text{uu}}^{(d)}\right| + \frac{\beta_d}{2}\text{Tr}([\mathbf{K}_{\text{uu}}^{(d)}]^{-1}\boldsymbol{\Psi}_2) \\ &- \frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|\beta_d\boldsymbol{\Psi}_2 + \mathbf{K}_{\text{uu}}^{(d)}\right| - \frac{\beta_d}{2}\psi_0 \quad (18) \\ &- \frac{1}{2}\mathbf{y}_d^T\left[\beta_d\mathbf{I}_N - \beta_d^2\boldsymbol{\Psi}_1\left(\beta_d\boldsymbol{\Psi}_2 + \mathbf{K}_{\text{uu}}^{(d)}\right)^{-1}\boldsymbol{\Psi}_1^T\right]\mathbf{y}_d,\end{aligned}$$

the sufficient statistics for the covariance kernels are

$$\psi_0 = \text{Tr}\left(\mathbb{E}_q[\mathbf{K}_{\text{ff}}^{(d)}]\right), \ \boldsymbol{\Psi}_1 = \mathbb{E}_q[\mathbf{K}_{\text{fu}}^{(d)}], \ \boldsymbol{\Psi}_2 = \mathbb{E}_q[\mathbf{K}_{\text{uf}}^{(d)}\mathbf{K}_{\text{fu}}^{(d)}] , \quad (19)$$

and the kernel hyperparameters $\boldsymbol{\theta}_d$ and noise precision β_d are their expected value with respect to $\mathbf{Z}$. $\boldsymbol{\theta}_d$ and β_d are defined in (22) and (23), respectively. The covariance matrices notation defines how the covariance function was evaluated. The input locations are provided as the subscripts, with f denoting $\mathbf{X}$ and u denoting $\mathbf{X}_\text{u}$, and the superscript denoting the hyperparameters $\boldsymbol{\theta}_d$.

DP Approximating Distributions As specified previously, we use a stick-breaking construction of the DP and introduce variational distributions over the assignment variables $\mathbf{Z}$, the stick lengths $\mathbf{V}$, and the concentration parameter α as a factorized distribution $q(\mathbf{Z}, \mathbf{V}, \alpha) = q(\mathbf{Z})q(\mathbf{V})q(\alpha)$ in a similar manner to Blei & Jordan (2005). In order to deal with the infinite support of the DP, we artificially truncate the number of components to $T < \infty$. We note that this is not a particular limitation of our approach, since, in general, the number of grouped functional priors will not exceed the number of observed dimensions D.

The truncated stick-breaking representation assumes that the likelihood of the length of the stick drawn at T is 1, therefore $q(v_T = 1) = 1$ and $\pi_t(\mathbf{V}) = 0$ for all $t > T$. This allows a finite approximating distribution to be used over the stick lengths; we use beta distributions such that

$$q(\mathbf{V}) = \textstyle\prod_{t=1}^{T-1} \text{Beta}(v_t|a_t, b_t) , \quad (20)$$

with $\boldsymbol{a}$ and $\boldsymbol{b}$ as variational parameters. The truncation at T also allows us to use a parameterized multinomial for the approximate distribution over $\mathbf{Z}$ as

$$q(\mathbf{Z}) = \textstyle\prod_{d=1}^{D} \text{Mult}(\mathbf{z}_d|\boldsymbol{\phi}_d), \ \sum_{t=1}^{T} \phi_{d,t} = 1 . \quad (21)$$

For the concentration parameter we introduce a gamma approximating distribution $q(\alpha) = \text{Gamma}(\alpha|w_1, w_2)$. We use MAP estimates for the kernel hyperparameters and noise precision $\{\boldsymbol{\theta}_t, \beta_t\}$; therefore, we have T sets of free parameters representing them. The expected values of the kernel hyperparameters and noise precision with respect to the assignment variable $\mathbf{Z}$ are needed in (17). They are defined as the following:

$$\boldsymbol{\theta}_d = \mathbb{E}_{q(\mathbf{z}_d)}\left[\textstyle\sum_{t=1}^{T}[\mathbf{z}_d = t]\cdot\boldsymbol{\theta}_t\right] = \sum_{t=1}^{T}\phi_{d,t}\cdot\boldsymbol{\theta}_t , \quad (22)$$

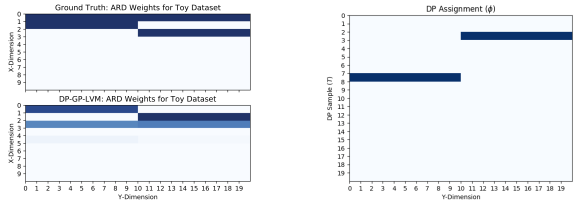
$$\beta_d = \mathbb{E}_{q(\mathbf{z}_d)}\left[\textstyle\sum_{t=1}^{T}[\mathbf{z}_d = t]\cdot\beta_t\right] = \sum_{t=1}^{T}\phi_{d,t}\cdot\beta_t . \quad (23)$$

DP Bound These approximations lead to a tractable DP lower bound of

$$\begin{aligned}\mathcal{L}_{\mathcal{DP}} = &\textstyle\sum_{d=1}^{D}\left[\mathbb{E}_q[\log p(\mathbf{z}_d|\mathbf{V})]\right] + \mathbb{E}_q[\log p(\mathbf{V}|\alpha)] \\ &+\mathbb{E}_q[\log p(\alpha)] + \mathbb{H}[q(\mathbf{V})] + \mathbb{H}[q(\mathbf{Z})] + \mathbb{H}[q(\alpha)] , \quad (24)\end{aligned}$$

where all terms are defined over standard exponential family distributions.

Optimization We optimize directly the objective of (14) with respect to the variational parameters $\{\mathbf{F}, \mathbf{X}, \mathbf{Z}, \mathbf{V}, \alpha\}$ and the hyperparameters and noise precisions $\{\boldsymbol{\Theta}, \boldsymbol{\beta}\}$. We initialize the mean parameters $\boldsymbol{\mu}$ for $q(\mathbf{X})$ with the first Q principal components of the observed data $\mathbf{Y}$ and set all $\boldsymbol{\Sigma}_q = \frac{1}{2}\boldsymbol{I}_N$. The pseudo input locations $\mathbf{X}_\text{u}$ are initialized to a random subset of $\boldsymbol{\mu}$. The stick length parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ are drawn from a standard log-normal. The allocation parameters $\boldsymbol{\Phi}$ are drawn from a standard normal pushed through the soft-max function. The hyperparameters and noise precisions are initialized with draws from their log-normal priors. Finally, the shape and scale for the gamma distribution over α are initialized to their prior $w_1 = s_1 = w_2 = s_2 = 1$. In our experiments, we evaluate the variational lower bound (14) and optimize it directly in TensorFlow (Abadi et al., 2015) using gradient descent with momentum. TensorFlow performs automatic differentiation to calculate gradients through the graph. This allows us to train the model without needing to calculate the partial derivatives of the lower bound with respect to each variational parameter.



(a) ARD Weights **(b)** DP Assigment

Figure 3. DP-GP-LVM latent factorization of synthetic data set. (a) The top panel shows the ground truth grouping while the lower panel indicates the inferred grouping by DP-GP-LVM where a column indicates the latent dimensions in $\mathbf{X}$ allocated to represent the corresponding observed dimension $\mathbf{Y}$. (b) The posterior DP allocations show that only two views are found and provides the correct allocations. The actual indices of the latent variables are not important due to the interchangeable characteristic of the DP.

Prediction and Missing Data After training, prediction from the latent space follows straight forwardly from the BGP-LVM (Damianou et al., 2016b), where each dimension d takes the kernel parameters from their respective posterior distribution. As the model is fully generative, imputation of missing data is also a simple task. The model can be trained neglecting the observations for the missing data and then their value can be predicted from the posteriors conditioned on the observed data. To infer the latent manifold location for a new observation $\mathbf{Y}^*$, we add additional variational parameters for $q(\mathbf{X}^*)$ and optimize the lower bound for the new joint model $\{\mathbf{Y}, \mathbf{Y}^*\}$. The ratio of the lower bounds of the joint model to the original approximates the probability of the new data, as described by Damianou et al. (2016b).

Appendices The appendices provide a derivation of the lower bound (14) (§ A), a stable calculation of the GP lower bound (18) (§ B), and the closed form solutions of the kernel expectations (19) (§ C) and the DP lower bound (24) terms (§ D). § E defines an alternative latent space prior.

4 Experiments

We test the model on four different data sets with the aim of providing intuition to the benefit of our approach in comparison with previous models. Additional results are in appendix § G. As a first experiment we create a synthetic data set, with known groupings, shown in Fig. 3. The data is generated by specifying a GP and using samples from the model as observations. The twenty-dimensional data were generated from three latent variables, where the first ten observed dimensions covary with latent dimensions one and two, and the second ten observed dimensions covary with latent dimensions one and three creating two distinct groups. We observe that the DP-GP-LVM correctly recovers the latent structure underlying the creation of the data.

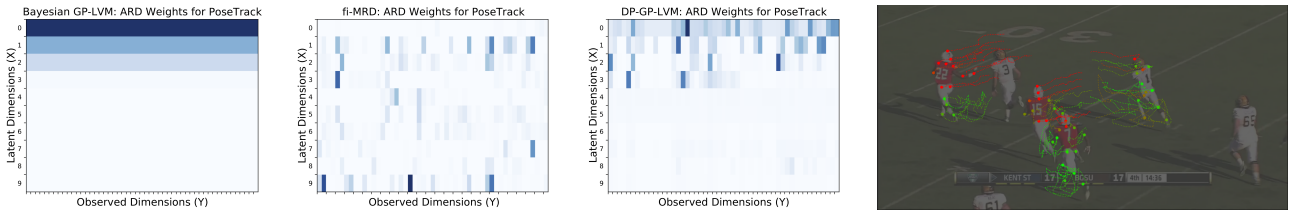
Motion Capture Datasets Motion capture data is represented in high-dimensional vector spaces but due to the underlying structure of the motion and the human body the data resides on a much lower-dimensional manifold. These correlation structures are challenging to specify a priori making this data ideal to demonstrate our approach. PoseTrack (Andriluka et al., 2018) consists of spatial image locations corresponding to an underlying human 3-D motion. We create two separate data sets corresponding to the motion of two and four individuals to allow evaluation of groupings both within and between individuals. We compare the DP-GP-LVM with a model where the observed data is considered as a single group (BGP-LVM) and where each dimension is in its own group (fully independent MRD: fi-MRD). Importantly, our model contains both these two cases but marginalizes over them and all other combinations.

Fig. 4 shows the results for four individuals; an expanded figure and results for two individuals are provided in appendix § G. Unsurprisingly, the fi-MRD model, where each dimension is a group, fails to capture the correlation structure and creates a large number of groups. BGP-LVM, where all dimensions belong to a single group, reduces to three shared dimensions, while the DP-GP-LVM uses combinations of four dimensions. The subtle variations, which the DP-GP-LVM captures with its fourth dimension, will be explained away as noise in BGP-LVM as it is not present in a majority of the dimensions (representation pollution).

Fig. 4(d) shows the mapping of the groupings of the joints superimposed onto the image. The model has grouped joints on the upper-body, which have mainly translational variation, separately from the lower-body, which additionally have significant oscillation due to the leg movement. There is also a third group which is only present in one of the individuals corresponding to a difference in translational movement.

In a third data set, we apply the model to a 3-D motion of a horse (Abson, 2014) shown in Fig. 5 with further details in the appendix (§ G). The skeleton consists of 46 joints leading to 138 observed dimensions. We show the model's ability to learn from small data using only 63 timesteps. Again, the BGP-LVM over simplifies the structure and fi-MRD over complicates it. The inferred structure, in Fig. 5(b), confirms the differences in grouping are from the periodic motion of the limbs compared to the head and torso.

The oi-VAE of Ainsworth et al. (2018) intends to learn interpretable latent representations. For a qualitative comparison in Fig. 6, we learn a representation of the walking sequences used in Fig. 3 in Ainsworth et al. (2018). Following their results, we show the three most significant joints for each latent dimension. Similarly, our method groups together dimensions in a interpretable manner, where the first dimension corresponds to the left and right hands, the second to the arms and the third to the upper body. Importantly, our

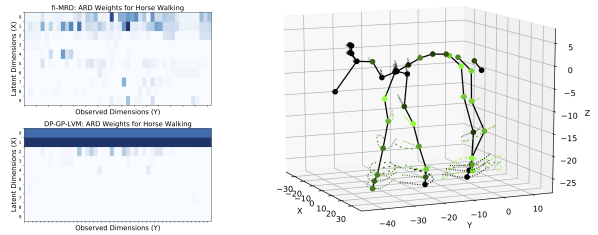


(a) BGP-LVM **(b)** MRD **(c)** DP-GP-LVM **(d)** DP-GP-LVM Inferred Groups

Figure 4. Results for PoseTrack data set with four individuals. (a)-(c) the ARD weights found for the three models (Observed dimensions on the horizontal axis and latent dimensions on the vertical). BGP-LVM increases to three latent dimensions but without the independence structure captured by DP-GP-LVM. (d) Here we show the learned groups (posterior over $\mathbf{Z}$) as the color of the points overlaid on the first frame of the image. The path traces out the future motion. We observe the separation of the clusters into translation and periodic motions.

Table 1. Log-likelihood for imputed distributions against ground truth for the PoseTrack missing data experiments.

Dataset	PoseTrack 2 Person						PoseTrack 4 Person					
$\tilde{N}$ Missing (%)	10	10	10	20	20	20	10	10	10	20	20	20
$\tilde{D}$ Missing (%)	10	20	30	10	20	30	10	20	30	10	20	30
BGP-LVM	-50.02 ± 12.32	-93.39 ± 18.84	-157.67 ± 32.20	-61.64 ± 13.95	-134.08 ± 22.73	-191.90 ± 33.57	-121.06 ± 24.10	-189.52 ± 34.73	-358.36 ± 52.06	-102.98 ± 24.06	-209.79 ± 37.35	-322.79 ± 59.27
fi-MRD	-40.62 ± 10.64	-109.73 ± 17.36	-138.67 ± 14.84	-54.92 ± 8.51	-149.36 ± 11.36	-248.89 ± 36.79	$\mathbf{-28.55 \pm 0.73}$	$\mathbf{-56.45 \pm 1.34}$	-123.47 ± 27.63	-73.89 ± 1.23	-147.42 ± 2.03	$\mathbf{-270.04 \pm 45.07}$
DP-GP-LVM	$\mathbf{-18.11 \pm 0.48}$	$\mathbf{-35.83 \pm 0.44}$	$\mathbf{-54.07 \pm 0.49}$	$\mathbf{-52.28 \pm 0.29}$	$\mathbf{-104.76 \pm 0.47}$	$\mathbf{-158.4 \pm 0.81}$	-28.14 ± 0.92	-55.44 ± 1.69	$\mathbf{-107.41 \pm 2.61}$	$\mathbf{-71.80 \pm 1.87}$	$\mathbf{-143.06 \pm 3.26}$	-311.39 ± 1.31

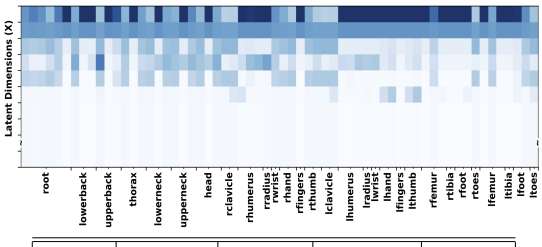


(a) ARD Weights **(b)** Effect of X_3

Figure 5. Horse walking motion capture data set. (a) The ARD weights demonstrate that fi-MRD (top) is unable to learn a reasonable factorization of the latent space while the DP-GP-LVM (bottom) uses three dimensions with the first two shared and the third encoding the oscillating motion at certain joints when the horse is walking. BGP-LVM models the data with two shared latent dimensions, losing the structure. (b) The horse at a middle frame where each joint has a line showing the position in past and future frames with the color modulated by the third ARD weight.

method does so *without* the need to specify the views or the number of latent dimensions, and we use less than 5% of the training data (150 frames vs 3,791).

Quantitative Missing Data Evaluations To quantify the efficacy of our approach, we performed a comparison for imputing missing data for the PoseTrack data sets. Table 1 shows the predictive log-likelihood of the held out ground-truth test data under each model. The $\tilde{N}$ and $\tilde{D}$ parameters indicates the percentage of the data set removed (in terms of samples and dimensions respectively) and held out as missing data. The missing samples and dimensions were taken out at random with 10 repeat samplings to provide the mean and standard error results in the table (see appendix § G for more details). The DP-GP-LVM provides superior estimates for the missing data since it captures the correct dependency structure in the data. The over simplification of the single group model leads to structure explained as noise,



lhand	rhumerus	lclavicle	lowerback	rthumb
rhand	lhumerus	rclavicle	rhumerus	head
lthumb	lradius	head	root	lclavicle

Figure 6. Top: Latent factorization for each joint in walking sequence. *Bottom:* The three joints with the most importance for each latent dimension. The left-most column corresponds to the top latent dimension in the factorization plot.

which cannot be used to constrain the missing data. On the other hand, the fi-MRD, which is much more expensive to compute (see appendix § F), fails to exploit all the dependencies in the data resulting in a reduction in precision for the missing data.

5 Conclusion and Future Work

We presented a generative, non-parametric, latent variable model with the ability to learn dependency structures in multivariate data. Our approach is capable of organizing the observed dimensions into groups that covary in a consistent manner. The model extends previous non-parametric formulations of IBFA by disentangling the factorization of the latent space with the characteristics of the generative mapping.

We intend to investigate further kernel combinations and latent priors for a wide range of applications. In addition, we intend to adapt the inference procedure to improve scalability and allow for online inference where the number of groupings continually evolves with more data.

Acknowledgements

We would like to acknowledge the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665992, the UK's EPSRC Centre for Doctoral Training in Digital Entertainment (CDE, EP/L016540/1), the RCUK-funded Centre for the Analysis of Motion, Entertainment Research and Applications (CAMERA, EP/M023281/1) and the Royal Society for supporting this research.

References

Abadi, M., Agarwal, A., Barham, P., and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Abson, K. Motion capture: capturing interaction between human and animal. 31, 03 2014.

Ainsworth, S. K., Foti, N. J., Lee, A. K. C., and Fox, E. B. oi-VAE: Output interpretable VAEs for nonlinear group factor analysis. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 119–128. PMLR, 2018.

Álvarez, M. and Lawrence, N. D. Sparse convolved gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems 21*, pp. 57–64. 2009.

Álvarez, M., Luengo, D., Titsias, M., and Lawrence, N. Efficient multioutput gaussian processes through variational inducing kernels. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 25–32, 13–15 May 2010.

Álvarez, M. A. and Lawrence, N. D. Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12:1459–1500, July 2011.

Andriluka, M., Iqbal, U., Ensafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and B., S. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018.

Blei, D. M. and Jordan, M. I. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1), 2005.

Dai, Z., Álvarez, M. A., and Lawrence, N. Efficient modeling of latent information in supervised learning using gaussian processes. In *Advances in Neural Information Processing Systems 30*, pp. 5137–5145. 2017.

Damianou, A., Lawrence, N. D., and Ek, C. H. Multi-view learning as a nonparametric nonlinear inter-battery factor analysis. *arXiv preprint arXiv:1604.04939*, April 2016a.

Damianou, A. C., Ek, C. H., Titsias, M. K., and Lawrence, N. D. Manifold relevance determination. In *International Conference on Machine Learning (ICML)*, 2012.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. Variational inference for latent variables and uncertain inputs in Gaussian processes. *Journal of Machine Learning Research*, 17(1):1425–1486, January 2016b.

Ek, C. H., Rihan, J., Torr, P. H. S., Rogez, G., and Lawrence, N. D. Ambiguity modeling in latent spaces. *Int. Conference on Machine Learning for Multimodal Interaction*, 2008.

Hensman, J., Rattray, M., and Lawrence, N. D. Fast nonparametric clustering of structured time-series. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):383–393, 2015.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 28(3/4), 1936.

Klami, A., Virtanen, S., and Kaski, S. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(1):965–1003, Apr 2013.

Kristof, W. Orthogonal inter-battery factor analysis. *Psychometrika*, 32(2):199–227, Jun 1967.

Lawrence, N. D. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(11):1783–1816, November 2005.

Meeds, E. and Osindero, S. An alternative infinite mixture of gaussian process experts. In *Advances in Neural Information Processing Systems 18*, pp. 883–890. 2006.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., 1996. ISBN 0387947248.

Palla, K., Knowles, D., and Ghahramani, Z. A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Rasmussen, C. E. and Ghahramani, Z. Infinite mixtures of gaussian process experts. In *Advances in Neural Information Processing Systems 14*, pp. 881–888. 2002.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN 026218253X.

Salzmann, M., Ek, C. H., Urtasun, R., and Darrell, T. Factorized orthogonal latent spaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 701–708, 2010.

Shon, A., Grochow, K., Hertzmann, A., and Rao, R. P. Learning shared latent structure for image synthesis and robotic imitation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

Spearman, C. "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 1904.

Sun, S. and Xu, X. Variational inference for infinite mixtures of gaussian processes with applications to traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):466–475, June 2011.

Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Tucker, L. R. An inter-battery method of factor analysis. *Psychometrika*, 23(2), Jun 1958. ISSN 1860-0980. doi: 10.1007/BF02289009.

Vehtari, A. *Bayesian model assessment and selection using expected utilities*. Dissertation, Helsinki University of Technology, 2001.

Wood, F., Griffiths, T., and Ghahramani, Z. A non-parametric bayesian method for inferring hidden causes. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2006.

Yuan, C. and Neubauer, C. Variational mixture of gaussian process experts. In *Advances in Neural Information Processing Systems 21*. 2009.