Latent Structure Learning using Gaussian and Dirichlet Processes

Andrew R. Lawrence
University of Bath, UK

Carl Henrik Ek
University of Bristol, UK

Neill D. F. Campbell
University of Bath, UK

1 Introduction and motivation

The widespread adoption of machine learning has created a pressing need for data efficient, interpretable models with efficient inference schemes. At the same time, a profusion of data sources demands an accounting for relationships between, often heterogeneous, data both within and across datasets. These challenges are particularly faced in the context of unsupervised learning; we would like to derive a (latent) description of the observed data that is both generative (recreates the data) and efficient (shares information between the observations).

In this work we propose to build upon recent work that demonstrates the success of Gaussian Process (GP) Bayesian non-parametric approaches to generative models for unsupervised learning [4, 3]. Such models perform well when a structural grouping of the observed data is known a-priori. However, we are often unable to specify this structure a-priori and wish to infer it directly from the data. Our work combines these GP models with a non-parametric clustering Dirichlet Process (DP) model in order to determine these structural groupings during learning in an efficient manner; this removes the need to specify structural relationships in advance and have them inferred from the data.

We refer to our model as GP-DP which has the contribution of simultaneously providing a generative explanation of all the observed data with a mechanism to ensure information sharing in an efficient manner. Our work demonstrates that structured probabilistic dependences between observed data can be recovered in the context of unsupervised learning that provide an explicit, interpretable generative model.

Existing models using GPs and DPs There are other models that have combined elements of GPs with DPs. For example, a mixture of GP experts addresses two of the major problems with GPs: the inverse of the covariance matrix ($\mathcal{O}(n^3)$) and stationary kernels. An infinite mixture of GP experts uses a DP to determine the number of components. The pertinent works that have implemented infinite mixture of GP experts by utilizing a DP are [7] and [9]. In addition, [5] combines a DP and GP for the purpose of clustering time-series data streams. However, as with the mixture of GP experts, their model serves a different purpose than our own.

2 Latent structure learning with the Bayesian GP-LVM and DP

The Gaussian Process Latent Variable Model (GP-LVM), as defined in [6], is an extension to the standard GP model, used for regression, that targets unsupervised learning. With the GP-LVM the input to a GP is unknown and is treated as a continuous latent variable. The Bayesian GP-LVM, as defined in [4], marginalizes these latent variables in order to learn the true latent dimensionality of the observed data. A kernel with automatic relevance determination (ARD) weights, as defined in [8], is used to determine the latent dimensionality. When an ARD weight approaches zero, that latent dimension is unnecessary and may be safely discounted from the model.

When the standard GP-LVM is trained via a maximum a posteriori approach, i.e., the latent variables are simply given a prior distribution, none of the latest dimensions will be reliably shut off; the GP-LVM will utilize all the latent dimensions as provided. Therefore, we implemented the Bayesian GP-LVM as part of our model.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

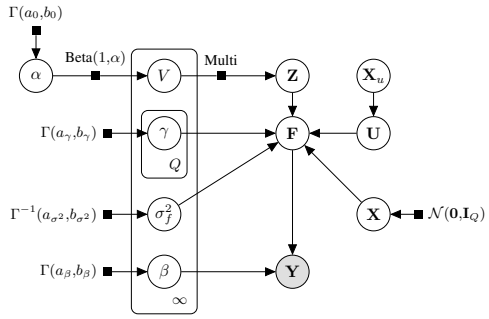

Figure 1: Our GP-DP Model. The Bayesian GP-LVM (of the same form as the MRD model) explains the observed data $\mathbf{Y}$ via the addition of noise β to a function $\mathbf{F}$ of latent variables $\mathbf{X}$ where both $\mathbf{F}$ and $\mathbf{X}$ are marginalized out via a variational approximation that makes use of "inducing points" $\mathbf{U}$, $\mathbf{X}_u$ [4]. The DP ensures that the hyperparameters used in the GP mapping (from $\mathbf{X}$ to $\mathbf{F}$) come from a fixed set of clusters of unknown size. The hyperparameter and noise (σ_f^2,γ,β) clusters are allocated to the dimensions of $\mathbf{F}$ via the multinomial distribution for $\mathbf{Z}$.

The Bayesian GP-LVM has been extended to learn a structured relationship between specified groups of observed dimensions, termed "views" in the form of the Manifold Relevance Determination (MRD) model [3]. Here, each view is considered as a separate version of the Bayesian GP-LVM but with a common, shared latent space. By using kernels based on ARD weights, views can choose to share information (select common latent space dimensions), become independent (select unique latent space dimensions), or some combination of the two. This provides structured unsupervised learning where we infer the relationships between the latent spaces that generate all the observed data.

The GP-DP model Our model is capable of learning a low dimensional manifold that explains the high dimensional observed data while simultaneously learning a structured latent space that specifies both the number of views and the clustering of the observed data into the appropriate view. The DP ensures a sparse representation by acting to reduce the number of possible views which in turn encourages information sharing.

The GP-DP is essentially an extension to MRD with a new view for each output y-dimension but a limited number of groups. It would be similar to treating each y-dimension as its own view and performing MRD then clustering the resulting ARD hyperparameters to limit the number of kernels. However, as the model is trained with a single objective function, GP-DP is less complex than MRD with unknown views and clustering afterwards. The graphical model for GP-DP is captured in Figure 1. We now describe the Bayesian GP-LVM and DP components of the model.

The Bayesian GP-LVM We construct the main generative portion of the model using the standard Bayesian GP-LVM for independent and identically distributed data [4]. We assume the observed, high dimensional, noisy data ($\mathbf{Y} \in \mathbb{R}^{N \times D}$) lies close to a manifold of much lower dimensionality (Q), such that $Q \ll D$ [6]. We assume independence across dimensions of $\mathbf{Y}$, the "output" or "observed" space.

The latent variables, $\mathbf{X} \in \mathbb{R}^{N \times Q}$ capture the locations on the manifold of the observed data $\mathbf{Y}$. A unit Gaussian prior is placed on the latent variables: $\mathbf{x}_n \sim \mathcal{N}(\mathbf{0},\mathbf{I}_Q)$. We assume a non-linear function from the latent space to the observed space and apply a GP prior. The output of this function is defined as $\mathbf{F}$, where $\mathbf{F} \in \mathbb{R}^{N \times D}$ and $\mathbf{f}_d(\mathbf{X}) \sim \mathcal{GP}(\mathbf{0},k(\mathbf{X},\mathbf{X}')|\,\boldsymbol{\gamma},{\sigma_f}^2)$. The observed data $\mathbf{Y}$ is defined by a GP with a mean of $\mathbf{F}$ and a noise precision $\beta \in \mathbb{R}_{\geq 0}$. Therefore, $\mathbf{y}_d \sim \mathcal{N}(\mathbf{f}_d,\beta^{-1}\mathbf{I}_N)$.

The hyperparameters for the Bayesian GP-LVM are the ARD weights and the signal variance, which are defined as $\boldsymbol{\gamma} \in {\mathbb{R}_{\geq 0}}^Q$ and ${\sigma_f}^2 \in \mathbb{R}_{\geq 0}$, respectively. For our model the values of these hyperparameters are allocated by the DP.

In order to marginalize out the latent variables, using variational approximate inference, inducing points are introduced that are assumed to be sufficient statistics to approximate the true function [4]. M inducing inputs ($\mathbf{X}_u \in \mathbb{R}^{M \times Q}$) and outputs ($\mathbf{U} \in \mathbb{R}^{M \times D}$) are used where we take $M \ll N$. The "pseudo function outputs" $\mathbf{U}$ are defined similarly to $\mathbf{F}$ such that $\mathbf{u}_d(\mathbf{X}_u) \sim \mathcal{GP}(\mathbf{0},k(\mathbf{X}_u,\mathbf{X}'_u)|\,\boldsymbol{\gamma},{\sigma_f}^2)$.

The Dirichlet Process We construct the DP portion of the model using the stick-breaking representation [1]. Four infinite sets of independent random variables are drawn as: (i) an infinite number of inde-

pendent stick lengths, $V_i \in [0,1], i = \{1,2,\cdots\}$, are drawn from $V_i \sim B(1,\alpha)$; (ii) hyperparameters for the GP-LVM, $\{\boldsymbol{\gamma}_i, \sigma^2_{f,i}, \beta_i\}, i = \{1,2,\cdots\}$, are drawn from their respective base distributions. Thus we have

$$\gamma \sim \Gamma(a_\gamma, b_\gamma), {\sigma_f}^2 \sim \Gamma^{-1}(a_{\sigma^2}, b_{\sigma^2}), \text{and } \beta \sim \Gamma(a_\beta, b_\beta). \tag{1}$$

The infinite vector of mixing proportions is defined as $\pi_i(\mathbf{V}) = V_i \prod_{j=1}^{i-1}(1-V_j)$, and the assignment of a specific sample (in our case a dimension of $\mathbf{Y}$) is sampled from a multinomial distribution with probabilities $\pi(\mathbf{V})$ as $\mathbf{Z}_d \sim \text{Mult}(\pi(\mathbf{V}))$ [1]. The scaling parameter, $\alpha \in \mathbb{R}_{\geq 0}$, acts as a prior on the number of groups; as α increases the DP allows the data to create more groups. We marginalize α with a broad prior, $\alpha \sim \Gamma(a_0, b_0)$, as recommended in [1].

The hyperparameters and noise precision for the GP-LVM $\{\boldsymbol{\gamma}, {\sigma_f}^2, \beta\}$ are assigned by the DP. Given the assignment variable $\mathbf{Z}$ for a specific dimension of $\mathbf{Y}$, the corresponding set of hyperparameters are assigned to the appropriate GP for the y-dimension. This provides

$$\mathbf{f}_d(\mathbf{X}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{X}, \mathbf{X}') | \gamma_i^{[\mathbf{Z}_d = i]}, {\sigma_f}^{2[\mathbf{Z}_d = i]}) \tag{2}$$

$$\mathbf{y}_d \sim \mathcal{N}(\mathbf{f}_d, \beta_i^{-[\mathbf{Z}_d = i]} \mathbf{I}_N). \tag{3}$$

Training We train our model with a single objective function. This ensures that the low-dimensional manifold, the assignment of observed data into appropriate views, and the latent space structure are learned simultaneously. We utilize mean-field approximation when performing variational inference; thus, the variational distributions are fully-factorized.

The truncated stick-breaking representation of a DP mixture, as proposed in [1], is used to define the variational evidence lower bound (ELBO) of the DP. We introduce two variational distributions $q(\mathbf{V})$ and $q(\mathbf{Z})$, which are beta and multinomial distributions respectively [1]. These distributions are truncated at T, which is a free parameter, rather than allowed to extend to infinity. In order to place a prior on α, we also introduce a variational $q(\alpha)$ using a Gamma distribution.

The ARD based exponentiated quadratic kernel is used for the Bayesian GP-LVM so the $\boldsymbol{\Psi}$ statistics, defined in [4], have an analytic expression. We introduce one variational distribution $q(\mathbf{X})$, which is a diagonal multivariate Gaussian distribution, and a set of inducing inputs to define the ELBO of the Bayesian GP-LVM.

The objective function for GP-DP is defined in (4), where $\mathcal{L}_{\mathcal{GP}}$ is the ELBO for the Bayesian GP-LVM, as defined in [4], and $\mathcal{L}_{\mathcal{DP}}$ is the ELBO for the DP, as defined in [1]. The other terms capture the log-likelihood of the priors on γ, β, and σ^2_f.

$$\mathcal{L} = \mathcal{L}_{\mathcal{GP}} + \mathcal{L}_{\mathcal{DP}} + \sum_{t=1}^{T} \left[\log\Gamma(\beta_t | a_\beta, b_\beta) + \log\Gamma^{-1}(\sigma^2_{f,t} | a_{\sigma^2}, b_{\sigma^2}) + \sum_{q=1}^{Q} \log\Gamma(\gamma_{t,q} | a_\gamma, b_\gamma) \right] \tag{4}$$

We use a gradient based method to jointly optimize the following parameters to maximimize the objective function $\mathcal{L}$. This utilizes $2(T-1)$ parameters for $q(\mathbf{V})$, DT parameters for $q(\mathbf{Z})$, $2NQ$ parameters for $q(\mathbf{X})$, 2 parameters for $q(\alpha)$, MQ values for $\mathbf{X}_u$, QT values for γ, T values for β, and T values for σ^2_f, where D is the dimensionality of the observed data, Q is the dimensionality of the latent space, N is the number of samples, M is the number of inducing points, and T is the truncation level for the DP. Therefore, there are $(4+D+Q)T + (2N+M)Q$ free parameters to optimize over for our GP-DP model.

Prediction Once the model is trained, inference is handled the same way as MRD as defined in Algorithm 1 in [2]. We can define the split between shared and private latent spaces by comparing the learned ARD weights to a threshold (δ) that is close to zero [2]. This is discussed further in Section 3.

3 Experiments

Our GP-DP model was tested against MRD with and without known views. By known views we mean that the MRD was given the correct segmentation of the toy data, while without indicates that each output y-dimension was treated as its own view. Toy data was generated from a known GP. Twenty total draws were made with each draw having one hundred samples. The first ten draws were a function of $\mathbf{X}_{:,0}$ and $\mathbf{X}_{:,1}$ while the second ten were a function of $\mathbf{X}_{:,0}$ and $\mathbf{X}_{:,2}$. This can be seen in the True

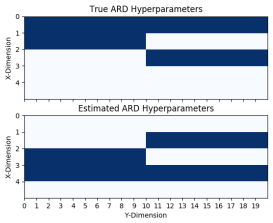

Figure 2: Our GP-DP

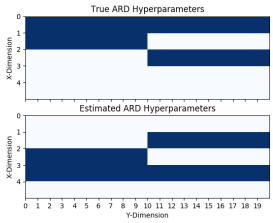

Figure 3: MRD with views

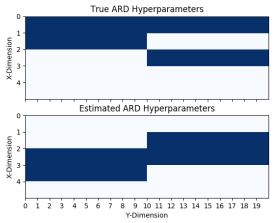

Figure 4: MRD without views

ARD Weights displayed in Figures 2, 3 and 4. The observed data was corrupted with Gaussian noise and none of the three models were given the underlying variables that generated the data.

As seen in Figures 2, 3, and 4, all three models find the correct latent structure of $\mathbf{Y}_{:,0:9} = f(\mathbf{X}_{:,0}, \mathbf{X}_{:,1})$ and $\mathbf{Y}_{:,10:19} = f(\mathbf{X}_{:,0}, \mathbf{X}_{:,2})$. Note the actual indices of the latent variables $\mathbf{X}$ in the results are not important (invariant to permutations). As we are calculating an approximate distribution for $\mathbf{X}$, the indices of the utilized latent dimensions depend on how $q(\mathbf{X})$ is initialized. The number of latent dimensions used and how they are shared and/or private between the various output dimensions are the critical factors.

Figures 2, 3, and 4 show the converged ARD weights for the various models after they have been thresholded. As discussed in [3], an ARD weight can be considered on if it is larger than δ, which is a small number close to zero.

4 Conclusion and future work

We have presented an extension to the MRD model where we learn the number of views and the segmentation of the observed data into the appropriate views automatically. The proposed model removes the need of treating each output dimension as its own view and clustering after training the MRD model. We demonstrate our model and compare it with a toy dataset with known latent dimensionality and shared/private structure. In the future we will test the model with a wider range of data such as motion capture datasets.

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), and the EPSRC Centre for the Analysis of Motion, Entertainment Research and Applications (CAMERA EP/M023281/1) for supporting this research.

References

[1] David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

[2] Andreas Damianou, Neil D. Lawrence, and Carl Henrik Ek. Multi-view learning as a nonparametric nonlinear inter-battery factor analysis. *arXiv preprint arXiv:1604.04939*, April 2016.

[3] Andreas C. Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold relevance determination. In *International Conference on Machine Learning*, 2012.

[4] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. Variational inference for latent variables and uncertain inputs in gaussian processes. *J. Mach. Learn. Res.*, 17(1):1425–1486, January 2016.

[5] James Hensman, Magnus Rattray, and Neil D Lawrence. Fast nonparametric clustering of structured time-series. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):383–393, 2015.

[6] Neil D. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783, 00 2005.

[7] Carl E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of gaussian process experts. In *Advances in Neural Information Processing Systems 14*, pages 881–888. 2002.

[8] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[9] Chao Yuan and Claus Neubauer. Variational mixture of gaussian process experts. In *Advances in Neural Information Processing Systems 21*. 2009.