Bayesian Gaussian Process Latent Variable Model

Michalis K. Titsias
School of Computer Science
University of Manchester

Neil D. Lawrence
School of Computer Science
University of Manchester

Abstract

We introduce a variational inference framework for training the Gaussian process latent variable model and thus performing Bayesian nonlinear dimensionality reduction. This method allows us to variationally integrate out the input variables of the Gaussian process and compute a lower bound on the exact marginal likelihood of the nonlinear latent variable model. The maximization of the variational lower bound provides a Bayesian training procedure that is robust to overfitting and can automatically select the dimensionality of the nonlinear latent space. We demonstrate our method on real world datasets. The focus in this paper is on dimensionality reduction problems, but the methodology is more general. For example, our algorithm is immediately applicable for training Gaussian process models in the presence of missing or uncertain inputs.

1 Introduction

Gaussian processes (GPs) (see e.g. Rasmussen and Williams, 2006) are stochastic processes over real-valued functions. GPs offer a Bayesian nonparametric framework for inference of highly nonlinear latent functions from observed data. They have become very popular in machine learning for solving problems such as nonlinear regression and classification.

The standard application of GP models is to supervised learning tasks where both output and input data are assumed to be given at training time. The application of GPs to unsupervised learning tasks is more involved. One approach to unsupervised learning with GPs is the Gaussian process latent variable model (GP-LVM) proposed by Lawrence (2004, 2005). GP-LVM can be considered as a multiple-output GP regression model where only the output data are given. The inputs are unobserved and are treated as latent variables, however instead of integrating out the latent variables, they are optimized. This trick makes the model tractable and some theoretical grounding for the approach is given by the fact that the model can be seen as a nonlinear extension of the linear probabilistic PCA (PPCA). In PPCA (and in factor analysis (FA)) Bayesian extensions of the model are straightforward (Bishop, 1999b; Ghahramani and Beal, 2000) using variational algorithms based on mean field approximations. An analogous variational method for the GP-LVM is a much more challenging problem which had not been addressed until this paper. The main difficulty is that to apply variational Bayes to GP-LVM we need to approximately integrate out the latent/input variables that appear nonlinearly in the inverse kernel matrix of the GP model. Standard mean field variational methodologies do not lead to an analytically tractable algorithm.

We introduce a framework that allows us to variationally integrate out the latent variables in the GP-LVM and compute a closed-form Jensen's lower bound on the true log marginal likelihood of the data. The key ingredient that makes the variational Bayes approach tractable is the application of variational inference in an *expanded probability model* where the GP prior is augmented to include auxiliary inducing variables. Inducing variables were introduced originally for computational speed ups in GP regression models (Csató and Opper, 2002; Seeger et al., 2003; Csató, 2002; Snelson and Ghahramani, 2006; Quiñonero Candela and Rasmussen, 2005; Titsias, 2009). Our approach builds on, and significantly extends the variational sparse GP method of Titsias (2009) so that a closed-form variational lower bound of the GP-LVM marginal likelihood is computed. This solves a key problem with the GP-LVM: variational inference in the GP-LVM allows for Bayesian training of the model that is robust to overfitting. Furthermore, by using the automatic relevance determination (ARD) squared exponential kernel, the algorithm allows us to automatically infer the dimensionality of the nonlinear latent space without introducing explicit regularizers to enforce this constraint (Geiger et al., 2009).

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

Although, in this paper, we focus on application of the variational approach to the GP-LVM, the methodology we have developed can be more widely applied to a variety of other GP models. In particular, our algorithm is immediately applicable for training GPs with missing or uncertain inputs (Girard et al., 2002). Other possible applications will be discussed as future work.

In the remainder of the paper we first review the GP-LVM and then we introduce our variational approximation. We finish by demonstrating the ability of the new model to automatically determine dimensionality and resist overfitting on real world datasets.

2 Gaussian process latent variable model

Let $Y \in \mathbb{R}^{N \times D}$ be the observed data where N is the number of observations and D the dimensionality of each data vector. These data are associated with latent variables $X \in \mathbb{R}^{N \times Q}$ where, for the purpose of doing dimensionality reduction, $Q \ll D$. The GP-LVM (Lawrence, 2005) defines a forward (or generative) mapping from the latent space to observation space that is governed by Gaussian processes. If the GPs are taken to be independent across the features then the likelihood function is written as

$$p(Y|X) = \prod_{d=1}^{D} p(\mathbf{y}_d|X), \tag{1}$$

where $\mathbf{y}_d$ represents the d^{th} column of Y and

$$p(\mathbf{y}_d|X) = \mathcal{N}(\mathbf{y}_d|\mathbf{0}, K_{NN} + \beta^{-1} I_N). \tag{2}$$

Here, K_{NN} is the $N \times N$ covariance matrix defined by the covariance (or kernel) function $k(\mathbf{x}, \mathbf{x}')$. For the purpose of doing automatic model selection of the dimensionality of latent space, this kernel can be chosen to follow the ARD (see Rasmussen and Williams, 2006) squared exponential form:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{q=1}^{Q} \alpha_q (x_q - x'_q)^2\right). \tag{3}$$

Equation (1) can be viewed as the likelihood function of a multiple-output GP regression model where the vectors of different outputs are drawn independently from the same Gaussian process prior which is evaluated at the inputs X. Since X is a latent variable, we can assign it a prior density given by the standard normal density. More precisely, the prior for X is:

$$p(X) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{0}, I_Q), \tag{4}$$

where each $\mathbf{x}_n$ is the n^{th} row of X. The joint probability model for the GP-LVM model is

$$p(Y, X) = p(Y|X)p(X). \tag{5}$$

The hyperparameters of the model are the kernel parameters $\boldsymbol{\theta} = (\sigma_f^2, \alpha_1, \ldots, \alpha_Q)$ and the inverse variance parameter β. For the sake of clarity, these parameters are omitted from the conditioning of the distribution[1]. Currently, the primary methodology for training the GP-LVM model is to find the MAP estimate of X (Lawrence, 2005) whilst jointly maximizing with respect to the hyperparameters. Here, we develop a variational Bayesian approach to marginalization of the latent variables, X, allowing us to optimize the resulting lower bound on the marginal likelihood with respect to the hyperparameters. The lower bound can also be used for model comparison and automatic selection of the latent dimensionality.

3 Variational inference

We wish to compute the marginal likelihood of the data:

$$p(Y) = \int p(Y|X)p(X)dX. \tag{6}$$

However, this quantity is intractable as X appears nonlinearly inside the inverse of the covariance matrix $K_{NN} + \beta^{-1} I_N$. Instead, we seek to apply an approximate variational inference procedure where we introduce a variational distribution $q(X)$ to approximate the true posterior distribution $p(X|Y)$ over the latent variables. We take the variational distribution to have a factorized Gaussian form over the latent variables,

$$q(X) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, S_n), \tag{7}$$

where the variational parameters are $\{\boldsymbol{\mu}_n, S_n\}_{n=1}^N$ and, for simplicity, S_n is taken to be a diagonal covariance matrix[2]. Using this variational distribution we can express a Jensen's lower bound on the $\log p(Y)$ that takes the form:

$$\begin{aligned} F(q) &= \int q(X) \log \frac{p(Y|X)p(X)}{q(X)} dX \\ &= \int q(X) \log p(Y|X) dX - \int q(X) \log \frac{q(X)}{p(X)} dX \\ &= \widetilde{F}(q) - \text{KL}(q||p), \end{aligned} \tag{8}$$

where the second term is the negative KL divergence between the variational posterior distribution $q(X)$ and the prior distribution $p(X)$ over the latent variables. This term is computed analytically since both distributions are Gaussians. Therefore, the difficult part when estimating the above bound is the first term:

$$\widetilde{F}(q) = \sum_{d=1}^{D} \int q(X) \log p(\mathbf{y}_d|X) dX = \sum_{d=1}^{D} \widetilde{F}_d(q), \tag{9}$$

[1] A precise notation is to write $p(Y, X|\beta, \boldsymbol{\theta}) = p(Y|X, \beta, \boldsymbol{\theta})p(X)$.

[2] This can be extended to non-diagonal within our framework.

where we have used (1). Thus, the computation of $\widetilde{F}(q)$ breaks down to separate computations of each $\widetilde{F}_d(q)$, corresponding to the d^{th} output. Notice that the computation of $\widetilde{F}_d(q)$ involves an analytically intractable integration. This arises because $\log p(\mathbf{y}_d|X)$ contains X in an highly non-linear manner inside the inverse of the covariance matrix, $K_{NN} + \beta^{-1} I_N$. Our main contribution is a mathematical tool that allows us to compute a closed-form lower bound for $\widetilde{F}_d(q)$. As we will see, the key idea is to apply variational sparse GP regression in an augmented probability model.

3.1 Lower bound by applying variational sparse GP regression

The computation in $\widetilde{F}_d(q)$ involves an expectation over the intractable term $\log p(\mathbf{y}_d|X)$. To deal with this, we first compute a Jensen's lower bound on $\log p(\mathbf{y}_d|X)$ by introducing the GP latent function values together with auxiliary inducing variables as those used in sparse GP models.

Sparse approximations have already been applied to speed up the GP-LVM Lawrence (2007). The first step of our approximation is equivalent to applying the new variational approximation of Titsias (2009) to the standard GP-LVM. The likelihood function $p(\mathbf{y}_d|X)$ is just the Gaussian marginal likelihood of a GP regression model. We make this explicit by introducing the GP latent function values $\mathbf{f}_d \in \mathbb{R}^N$ associated with the vector of (noise corrupted) outputs $\mathbf{y}_d$ (the d^{th} column of Y). The "complete" likelihood associated with the marginal likelihood $p(\mathbf{y}_d|X)$ is:

$$p(\mathbf{y}_d, \mathbf{f}_d|X) = p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{f}_d|X), \tag{10}$$

where $p(\mathbf{y}_d|\mathbf{f}_d) = \mathcal{N}(\mathbf{y}_d|\mathbf{f}_d, \beta^{-1} I_N)$ and $p(\mathbf{f}_d|X)$ is the zero-mean GP prior with covariance matrix K_{NN}. Note that the above joint model still contains X inside the inverse of K_{NN} making expectations under distributions over X difficult to compute. We finesse this intractability by introducing auxiliary inducing variables and applying the variational sparse GP formulation of Titsias (2009).

We follow the approach of Lawrence (2007): for each vector of latent function values $\mathbf{f}_d$ we introduce a separate set of M inducing variables $\mathbf{u}_d \in \mathbb{R}^M$ evaluated at a set of inducing input locations given by $Z \in \mathbb{R}^{M \times Q}$. For simplicity, we assume that all $\mathbf{u}_d$s, associated with different outputs, are evaluated at the same inducing locations, however this could be relaxed. The $\mathbf{u}_d$ variables are just function points drawn from the GP prior. Using these inducing variables we augment the joint probability model in eq. (10):

$$p(\mathbf{y}_d, \mathbf{f}_d, \mathbf{u}_d|X, Z) = p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{f}_d|\mathbf{u}_d, X, Z)p(\mathbf{u}_d|Z), \tag{11}$$

where we used the fact that the joint GP prior over function values $\mathbf{f}_d$ and $\mathbf{u}_d$ evaluated at inputs X and Z factorizes as $p(\mathbf{f}_d, \mathbf{u}_d|X, Z) = p(\mathbf{f}_d|\mathbf{u}_d, X, Z)p(\mathbf{u}_d|Z)$ where

$$p(\mathbf{f}_d|\mathbf{u}_d, X, Z) = \mathcal{N}(\mathbf{f}_d|\boldsymbol{\alpha}_d, K_{NN} - K_{NM}K_{MM}^{-1}K_{MN})$$

is the conditional GP prior with $\boldsymbol{\alpha}_d = K_{NM}K_{MM}^{-1}\mathbf{u}_d$. Further, $p(\mathbf{u}_d|Z) = \mathcal{N}(\mathbf{u}_d|\mathbf{0}, K_{MM})$ is the marginal GP prior over the inducing variables. The likelihood $p(\mathbf{y}_d|X)$ can be equivalently computed from the above augmented model by marginalizing out $(\mathbf{f}_d, \mathbf{u}_d)$ and crucially this is true for any value of the inducing inputs Z. This means that, unlike X, the inducing inputs Z are *not* random variables. Neither are they model hyperparameters, they are *variational parameters*. This interpretation of the inducing inputs is key in developing our approximation, it arises from the variational approach of Titsias (2009). Taking advantage of this observation we now simplify our notation by dropping Z from our expressions. We can now apply variational inference to approximate the true posterior, $p(\mathbf{f}_d, \mathbf{u}_d|\mathbf{y}_d, X) = p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{y}_d, X)p(\mathbf{u}_d|\mathbf{y}_d, X)$, with a sparse variational distribution that takes the form

$$q(\mathbf{f}_d, \mathbf{u}_d) = p(\mathbf{f}_d|\mathbf{u}_d, X)\phi(\mathbf{u}_d), \tag{12}$$

where $p(\mathbf{f}_d|\mathbf{u}_d, X)$ is the conditional GP prior that appears in the joint model in (11), while $\phi(\mathbf{u}_d)$ is a variational distribution over the inducing variables $\mathbf{u}_d$. Thus we obtain a lower bound:

$$\log p(\mathbf{y}_d|X) \geq \int \phi(\mathbf{u}_d) \log \frac{p(\mathbf{u}_d)\mathcal{N}(\mathbf{y}_d|\boldsymbol{\alpha}_d, \beta^{-1} I_N)}{\phi(\mathbf{u}_d)} d\mathbf{u}_d - \frac{\beta}{2}\text{Tr}(K_{NN} - K_{NM}K_{MM}^{-1}K_{MN}), \tag{13}$$

where $\boldsymbol{\alpha}_d = K_{NM}K_{MM}^{-1}\mathbf{u}_d$. In the variational sparse GP method (Titsias, 2009), the $\phi(\mathbf{u}_d)$ distribution is computed in an optimal way. Such an optimal choice of this distribution depends on the latent variables X and is not useful in our case. In order to obtain the bound for the GP-LVM we need to take a mean field approach and force independence of the distribution $\phi(\mathbf{u}_d)$ from the random variable X.

So far we have computed a lower bound on $\log p(\mathbf{y}_d|X)$ which is the intractable term in $\widetilde{F}_d(q)$. Using eq. (13) and the definition of $\widetilde{F}_d(q)$ from (9) we have

$$\widetilde{F}_d(q) \geq \int q(X) \Big[\int \phi(\mathbf{u}_d) \log \frac{p(\mathbf{u}_d)\mathcal{N}(\mathbf{y}_d|\boldsymbol{\alpha}_d, \beta^{-1} I_N)}{\phi(\mathbf{u}_d)} d\mathbf{u}_d - \frac{\beta}{2}\text{Tr}(K_{NN}) + \frac{\beta}{2}\text{Tr}(K_{MM}^{-1}K_{MN}K_{NM}) \Big] dX,$$

where we used standard properties of the trace of a matrix. Since (under our factorization assumption) $\phi(\mathbf{u}_d)$ does not depend on the random variable X, we can swap the integrations over X and $\mathbf{u}_d$ and perform firstly the integration with respect to X:

$$\widetilde{F}_d(q) \geq \int \phi(\mathbf{u}_d) \left[\langle \log \mathcal{N}(\mathbf{y}_d|\boldsymbol{\alpha}_d, \beta^{-1} I_N)\rangle_{q(X)} + \log \frac{p(\mathbf{u}_d)}{\phi(\mathbf{u}_d)} \right] d\mathbf{u}_d - \frac{\beta}{2}\text{Tr}\left(\langle K_{NN}\rangle_{q(X)}\right) + \frac{\beta}{2}\text{Tr}\left(K_{MM}^{-1}\langle K_{MN}K_{NM}\rangle_{q(X)}\right),$$

where $\langle \cdot \rangle_{q(X)}$ denotes expectation under the distribution $q(X)$. Now, we can analytically maximize the above lower bound with respect to the distribution $\phi(\mathbf{u}_d)$. The optimal setting of this distribution is $\phi(\mathbf{u}_d) \propto \langle \log \mathcal{N}(\mathbf{y}_d | \boldsymbol{\alpha}_d, \beta^{-1} I_N) \rangle_{q(X)} p(\mathbf{u}_d)$ and the lower bound that automatically incorporates such an optimal setting is obtained easily by reversing Jensen's inequality,

$$\widetilde{F}_d(q) \geq \log \left(\int e^{\langle \log \mathcal{N}(\mathbf{y}_d | \boldsymbol{\alpha}_d, \beta^{-1} I_N) \rangle_{q(X)}} p(\mathbf{u}_d) d\mathbf{u}_d \right)$$
$$- \frac{\beta}{2} \text{Tr} \left(\langle K_{NN} \rangle_{q(X)} \right) + \frac{\beta}{2} \text{Tr} \left(K_{MM}^{-1} \langle K_{MN} K_{NM} \rangle_{q(X)} \right).$$

The r.h.s. in this equation is a lower bound in which the variational distribution $\phi(\mathbf{u}_d)$ has been eliminated optimally. This quantity now can be computed in closed-form since it boils down to computing the statistics $\psi_0 = \text{Tr} \left(\langle K_{NN} \rangle_{q(X)} \right)$, $\Psi_1 = \langle K_{NM} \rangle_{q(X)}$ and $\Psi_2 = \langle K_{MN} K_{NM} \rangle_{q(X)}$. These statistics for certain covariance functions, such as the ARD squared exponential from (3), are computable analytically as discussed in section 3.2. Notice also that $\langle \log \mathcal{N}(\mathbf{y}_d | \boldsymbol{\alpha}_d, \beta^{-1} I_N) \rangle_{q(X)}$ is just a quadratic function of $\mathbf{u}_d$ that depends on the statistics Ψ_1 and Ψ_2. Therefore, the integration involved in the above equation is a standard Gaussian integral. The closed-form of the lower bound on $\widetilde{F}_d(q)$ is:

$$\widetilde{F}_d(q) \geq \log \left[\frac{(\beta)^{\frac{N}{2}} |K_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}} |\beta \Psi_2 + K_{MM}|^{\frac{1}{2}}} e^{-\frac{1}{2} \mathbf{y}_d^T W \mathbf{y}_d} \right]$$
$$- \frac{\beta \psi_0}{2} + \frac{\beta}{2} \text{Tr} \left(K_{MM}^{-1} \Psi_2 \right), \tag{14}$$

where $W = \beta I_N - \beta^2 \Psi_1 (\beta \Psi_2 + K_{MM})^{-1} \Psi_1^T$. We can now compute the closed-from variational lower of the GP-LVM according to equation (8). More precisely, by summing both sides of (14) over the D outputs we obtain on the l.h.s. the term $\widetilde{F}(q)$ (see equation (9)) and on the r.h.s. a lower bound on $\widetilde{F}(q)$. By substituting the latter quantity (in place of $\widetilde{F}(q)$) in (8) we obtain the final GP-LVM lower bound. This bound has an elegant form since it resembles closely the corresponding sparse GP-LVM marginal likelihood (where X is optimized, not integrated out) obtained by applying the variational method of Titsias (2009). The difference is that now (where X is variationally integrated out) we obtain an extra regularization term, i.e. the term KL$(q||p)$ in (8), and also the kernel quantities Tr(K_{NN}), K_{NM} and $K_{MN} K_{NM}$ that contain X are replaced by variational averages, which are the Ψ statistics defined above.

The bound can be jointly maximized over the variational parameters $(\{\boldsymbol{\mu}_n, S_n\}_{n=1}^N, Z)$ and the model hyperparameters $(\beta, \boldsymbol{\theta})$ by applying gradient-based optimization techniques. The approach is similar to the MAP optimization of the objective function employed in Lawrence (2005) with the main difference that now we have an additional set of variational parameters governing the approximate posterior variances in the latent space.

3.2 Computation of the Ψ statistics

To obtain an explicit evaluation of the variational lower bound we need to compute the statistics (ψ_0, Ψ_1, Ψ_2). We can rewrite the ψ_0 statistic as $\psi_0 = \sum_{n=1}^N \psi_0^n$ where

$$\psi_0^n = \int k(\mathbf{x}_n, \mathbf{x}_n) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, S_n) d\mathbf{x}_n. \tag{15}$$

Ψ_1 is an $N \times M$ matrix such that

$$(\Psi_1)_{nm} = \int k(\mathbf{x}_n, \mathbf{z}_m) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, S_n) d\mathbf{x}_n. \tag{16}$$

Ψ_2 is an $M \times M$ matrix which is written as $\Psi_2 = \sum_{n=1}^N \Psi_2^n$ where Ψ_2^n is such that

$$(\Psi_2^n)_{mm'} = \int k(\mathbf{x}_n, \mathbf{z}_m) k(\mathbf{z}_{m'}, \mathbf{x}_n) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, S_n) d\mathbf{x}_n. \tag{17}$$

The above computations involve convolutions of the covariance function with a Gaussian density. For some standard kernels such the ARD squared exponential (SE) covariance and the linear covariance function these statistics are obtained analytically. In particular for the ARD SE kernel, $\psi_0 = N \sigma_f^2$,

$$(\Psi_1)_{nm} = \sigma_f^2 \prod_{q=1}^Q \frac{e^{-\frac{1}{2} \frac{\alpha_q (\boldsymbol{\mu}_{nq} - \mathbf{z}_{mq})^2}{\alpha_q S_{nq} + 1}}}{(\alpha_q S_{nq} + 1)^{\frac{1}{2}}}$$

and

$$(\Psi_2^n)_{mm'} = \sigma_f^4 \prod_{q=1}^Q \frac{e^{-\frac{\alpha_q (z_{mq} - z_{m'q})^2}{4} - \frac{\alpha_q (\mu_{nq} - \bar{z}_q)^2}{2 \alpha_q S_{nq} + 1}}}{(2 \alpha_q S_{nq} + 1)^{\frac{1}{2}}},$$

where $\bar{z}_q = \frac{(z_{mq} + z_{m'q})}{2}$. This gives us all the components we need to compute the variational lower bound for the ARD SE kernel. For the linear covariance function the integrals are also tractable. Suppose the kernel function follows the ARD linear form:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T A \mathbf{x}', \tag{18}$$

where A is a positive definite diagonal covariance matrix. Learning the diagonal elements of A will allow to perform automatic model selection of the dimensionality of the linear latent space in a similar manner to ARD SE covariance function. Thus, the framework provides an alternative method to perform Bayesian probabilistic PCA (Bishop, 1999a; Minka, 2001). For this linear kernel the statistics are such that $\psi_0^n = \text{Tr} \left[A (\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + S_n) \right]$, $(\Psi_1)_{nm} = \boldsymbol{\mu}_n^T A \mathbf{z}_m$ and $(\Psi_2^n)_{mm'} = \mathbf{z}_m^T A (\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + S_n) A \mathbf{z}_{m'}$.

Finally, it is worth noticing that the Ψ statistics are computed in a decomposable way which is useful when a new data vector is inserted into the model. In particular, the statistics ψ_0 and Ψ_2 are written as sums of independent terms where each term is associated with a data point and similarly each column of the matrix Ψ_1 is associated with only one data point. These properties can help to speed up computations during test time as discussed in section 4.

3.3 Summary of the variational method

To summarize the above variational method allows to compute a Jensen's lower bound on the GP-LVM marginal likelihood and the key to obtaining this bound was to introduce auxiliary variables into the model similar to those used in sparse GP regression. Although, we explained the method using a sequence of steps, we could also start by writing the joint probability density over all variables $(Y, X, \{\mathbf{f}_d, \mathbf{u}_d\}_{d=1}^D)$ and then introduce the full variational distribution to approximate the model at once. This full variational distribution that gives rise to the lower bound obtained earlier is given by

$$q(\{\mathbf{f}_d, \mathbf{u}_d\}_{d=1}^D, X) = \left(\prod_{d=1}^D p(\mathbf{f}_d|\mathbf{u}_d, X)\phi(\mathbf{u}_d) \right) q(X).$$

This distribution is a mean field approximation with respect to $\mathbf{u}_d$s and X. However, it is not a mean field with respect to $\mathbf{f}_d$s since once $\mathbf{u}_d$s and X are marginalized out, then $\mathbf{f}_d$s become coupled. In addition, the fact that the $q(X)$ distribution is factorized over the latent variables is a consequence of the mean field assumption between $\mathbf{u}_d$s and X. It does not need to be imposed in advance. To see this, notice that $q(X)$ appears only in the Ψ statistics which as explained earlier are computed in a decomposable (across data points/latent variables) way.

4 Prediction and computation of probabilities in test data

In this section, we discuss how we can use the proposed model, from now on referred to as Bayesian GP-LVM, in order to make predictions in unseen data. Firstly, we explain how we can approximately compute the probability density $p(\mathbf{y}_*|Y)$ of some observed test data vector $\mathbf{y}_* \in \mathbb{R}^D$, which is allowed to have missing values. The computation of this probability can allow us to use the model as a density estimator which, for instance, can represent the class conditional distribution in a generative based classification system. We will exploit such a use in section 5. Secondly we discuss how we can predict the function values $\mathbf{y}_*$ given that we have an estimate of the variational distribution $q(\mathbf{x}_*)$ for the latent variable associated with the observation $\mathbf{y}_*$. This can be useful when we wish to predict the missing values of some partially observed test output $\mathbf{y}_* = (\mathbf{y}_*^O, \mathbf{y}_*^U) \in \mathbb{R}^D$ where $\mathbf{y}_*^O$ are observed components in the vector $\mathbf{y}_*$ and $\mathbf{y}_*^U$ are the missing values that we would like to predict. This second prediction task can also be used to remove the noise of a fully observed output.

First we discuss how to approximate the density $p(\mathbf{y}_*|Y)$. By introducing the latent variables X (corresponding to the training outputs Y) and new test latent variables $\mathbf{x}_*$, the previous density is written as

$$p(\mathbf{y}_*|Y) = \frac{\int p(\mathbf{y}_*, Y|X, \mathbf{x}_*)p(X, \mathbf{x}_*)dXd\mathbf{x}_*}{\int p(Y|X)p(X)dX}. \quad (19)$$

Note that this is a ratio of two marginal likelihoods. In the denominator we have the marginal likelihood of the GP-LVM for which we have already computed a variational lower bound. The numerator is another marginal likelihood that is obtained by augmenting the training data Y with the test point $\mathbf{y}_*$ and integrating out both X and the newly inserted latent variable $\mathbf{x}_*$. To approximate the density $p(\mathbf{y}_*|Y)$, we construct a ratio of lower bounds as follows. $\int p(Y|X)p(X)dX$ is approximated by the lower bound $e^{F(q(X))}$ where $F(q(X))$ is the variational lower bound on the log marginal likelihood as computed in section 3. The maximization of this lower bound specifies the variational distribution $q(X)$ over the latent variables in the training data. Then, this distribution remains fixed during test time. $\int p(\mathbf{y}_*, Y|X, \mathbf{x}_*)p(X, \mathbf{x}_*)dXd\mathbf{x}_*$ is approximated by the lower bound $e^{F(q(X, \mathbf{x}_*))}$. To compute this, we need to optimize with respect to the parameters $(\boldsymbol{\mu}_*, S_*)$ of the Gaussian variational distribution $q(\mathbf{x}_*)$. Such optimization is subject to local minima. However, sensible initializations of $\boldsymbol{\mu}_*$ can be employed based on the mean of the variational distributions associated with the nearest neighbours of $\mathbf{y}_*$ in the training data Y. Furthermore, such optimization is fast because we can perform several precomputations in advance. In particular, notice that because the computation of the Ψ statistics decomposes across data, updating these statistics to account for the insertion of the test point, involves only averages over the single-point variational distribution $q(\mathbf{x}_*)$. Finally, the approximation of $p(\mathbf{y}_*|Y)$ is given by

$$q(\mathbf{y}_*|Y) = e^{F(q(X, \mathbf{x}_*)) - F(q(X))}. \quad (20)$$

We now discuss the second prediction problem where a partially observed test point $\mathbf{y}_* = (\mathbf{y}_*^O, \mathbf{y}_*^U)$ is given and we wish to reconstruct the missing part $\mathbf{y}_*^U$. This involves two steps. Firstly, we optimize the parameters of the variational distribution $q(\mathbf{x}_*)$ by maximizing the variational lower bound on $\int p(\mathbf{y}_*^O, Y|X, \mathbf{x}_*)p(X, \mathbf{x}_*)dXd\mathbf{x}_*$ by keeping all the optimized quantities fixed apart from $q(\mathbf{x}_*)$; exactly as explained earlier. To predict now $\mathbf{y}_*^U$, we take the standard GP prediction approach by taking also into account the fact that the input $\mathbf{x}_*$ is uncertain since it has the distribution $q(\mathbf{x}_*)$. Therefore, the problem takes the form of GP prediction with uncertain inputs similar to Girard et al. (2002). More precisely, to predict $\mathbf{y}_*^U$ we first predict its latent function values $\mathbf{f}_*^U$ according to

$$\begin{aligned} q(\mathbf{f}_*^U) &= \int \left(\prod_{d \in U} \int p(f_{*d}^U|\mathbf{u}_d, \mathbf{x}_*)\phi(\mathbf{u}_d)d\mathbf{u}_d \right) q(\mathbf{x}_*)d\mathbf{x}_* \\ &= \int q(\mathbf{f}_*^U|\mathbf{x}_*)q(\mathbf{x}_*)d\mathbf{x}_*, \end{aligned} \quad (21)$$

where $q(\mathbf{f}_*^U|\mathbf{x}_*)$ is a factorized Gaussian distribution where each factor takes the form of the projected process predictive distribution (Csató and Opper, 2002; Seeger et al., 2003; Rasmussen and Williams, 2006). The marginalization of $\mathbf{x}_*$ couples all dimensions of $\mathbf{f}_*^U$ and produces a non-Gaussian fully dependent multivariate density. For squared exponential kernels all moments of the density $q(\mathbf{f}_*^U)$ are analytically tractable. In practice, we will typically need only the mean and covariance of $\mathbf{f}_*^U$. The mean is

$$\mathbb{E}(\mathbf{f}_*^U) = \Lambda^T \boldsymbol{\psi}_1^*.$$

Here, $\Lambda = \beta(K_{MM} + \beta\Psi_2)^{-1}\Psi_1^T Y^U$ where Y^U is the matrix containing the columns of Y corresponding to the missing values of $\mathbf{y}_*$. Also, the vector $\boldsymbol{\psi}_1^* \in \mathbb{R}^M$ is defined by $\boldsymbol{\psi}_1^* = \langle K_{M*}\rangle_{q(\mathbf{x}_*)}$ where $K_{M*} = k(Z, \mathbf{x}_*)$. Similarly,

$$\text{Cov}(\mathbf{f}_*^U) = \Lambda^T \left(\Psi_2^* - \boldsymbol{\psi}_1^*(\boldsymbol{\psi}_1^*)^T\right)\Lambda \\ + \psi_0^* I - \text{Tr}\left(\left[K_{MM}^{-1} - (K_{MM} + \beta\Psi_2)^{-1}\right]\Psi_2^*\right) I,$$

where $\psi_0^* = \langle k(\mathbf{x}_*, \mathbf{x}_*)\rangle_{q(\mathbf{x}_*)}$ and $\Psi_2^* = \langle K_{M*}K_{*M}\rangle_{q(\mathbf{x}_*)}$. Notice that the Ψ statistics (the terms $(\psi_0^*, \boldsymbol{\psi}_1^*, \Psi_2^*)$) involving the test latent variable $\mathbf{x}_*$ appear naturally in these expressions. Using the above expressions, the predicted mean of $\mathbf{y}_*^U$ is equal to $\mathbb{E}(\mathbf{f}_*^U)$ and the predicted covariance is equal to $\text{Cov}(\mathbf{f}_*^U) + \beta^{-1}I$.

5 Experiments

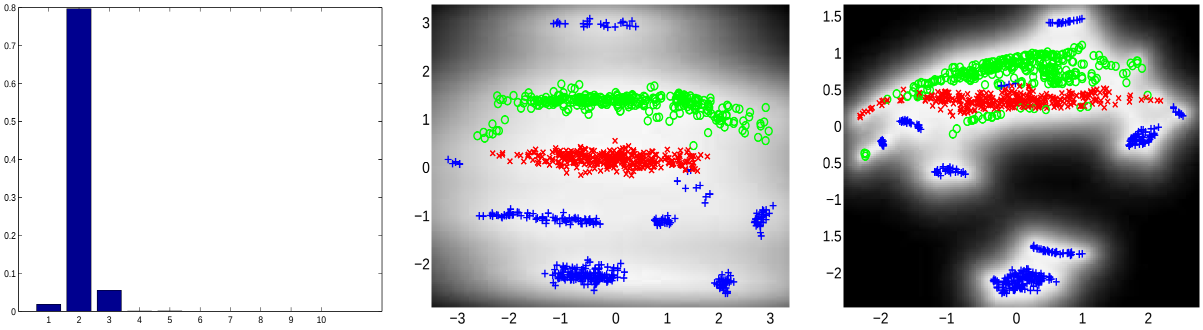
To demonstrate the Bayesian GP-LVM we now consider some standard machine learning data sets. Our aim is to highlight several characteristics of the algorithm: the improved quality of visualizations achieved by the model, the utility of being able to access a lower bound on the marginal likelihood of the data, and the ability of the model to automatically determine the dimensionality of the data.

5.1 Oil flow data

In the first experiment we illustrate the method in the multi-phase oil flow data (Bishop and James, 1993) that consists of 1000, 12 dimensional observations belonging to three known classes corresponding to different phases of oil flow. Figure 1 shows the results for these data obtained by applying the Bayesian GP-LVM with 10 latent dimensions using the ARD SE kernel. The means of the variational distribution were initialized based on PCA, while the variances in the variational distribution are initialized to neutral values around 0.5. As shown in Figure 1(a), the algorithm switches off automatically 7 out of 10 latent dimensions by making their inverse lengthscales zero. Figure 1(b) shows the visualization obtained by keeping only the dominant latent directions (having the largest inverse lengthscale) which are the dimensions 2 and 3. This is a remarkably high quality two dimensional visualization of this data. For comparison, Figure 1(c) shows the visualization provided by the standard sparse GP-LVM that runs by assuming only 2 latent dimensions. Both models use 50 inducing variables, while the latent variables X optimized in the standard GP-LVM are initialized based on PCA. Note that if we were to run the standard GP-LVM with 10 latent dimensions, the model would overfit the data, it would not reduce the dimensionality in the manner achieved by the Bayesian GP-LVM. In these two dimensions, the nearest neighbour error for the different classes (phases of oil flow) in the case of Bayesian GP-LVM is 3 errors from 1000 data points. The number of the nearest neighbour errors made when applying the standard GP-LVM was 26.

5.2 Frey Faces Data

Here, we consider a dataset of faces (Roweis et al., 2002) taken from a video sequence that consists of 1965 images of size 28×20. In this dataset, we would like to exploit the ability of the model to reconstruct partially observed test data. Therefore, we train the model using a random selection of 1000 images and then we consider the remaining 965 images as test data. Furthermore, in each test image we assume that only half of the image pixels are observed. The missing pixels were chosen randomly for each test image. After training on 1000 images, each partially observed test image was processed separately (this involves the optimization of the corresponding variational distribution as discussed in section 4) and the missing pixels were predicted. Figure 2 shows a few examples of reconstructed test images. Each column in this figure corresponds to a test image, where the top plot shows the true test image, the middle one the partially observed image and the bottom image shows the reconstructed image. We also measure the mean absolute reconstruction error over all test images and missing pixels and compare this error with the standard sparse GP-LVM. This standard GP-LVM was applied using several settings of the latent dimensionality: $Q = 2, 5, 10$ and 30. The Bayesian GP-LVM was trained once using 30 latent dimensions. The latent variables X in the standard GP-LVM and the means of the variational distribution in Bayesian GP-LVM were initialized through PCA. The error for Bayesian GP-LVM was 7.4003. For the standard GP-LVM the error was 10.5748, 9.7284, 19.6949 and 19.6961 for $2, 5, 10$ and 30 latent dimensions respectively. Notice that the standard GP-LVM has poor performance for large value of latent dimension and achieves the best error when we consider 5 latent dimensions. Nevertheless, this was still worse than the error from the Bayesian GP-LVM. Finally, Figure 3 shows the values of the inverse lengthscales obtained by the maximization of the variational lower bound. Although, in this case, the algorithm does not shrink some of the dimensions completely to zero, it does force many of them to obtain small values. Note that one of the dimensions (the first from the left) seems to be the most important in explaining the data.



(a) (b) (c)

Figure 1: Panel (a) shows the inverse lengthscales found by applying the Bayesian GP-LVM with ARD SE kernel on the oil flow data. Panel (b) shows the visualization achieved by keeping the most dominant latent dimensions (2 and 3) which have the largest inverse lengthscale value. Dimension 2 is plotted on the y-axis and 3 and on the x-axis. Plot (c) shows the visualization found by standard sparse GP-LVM.

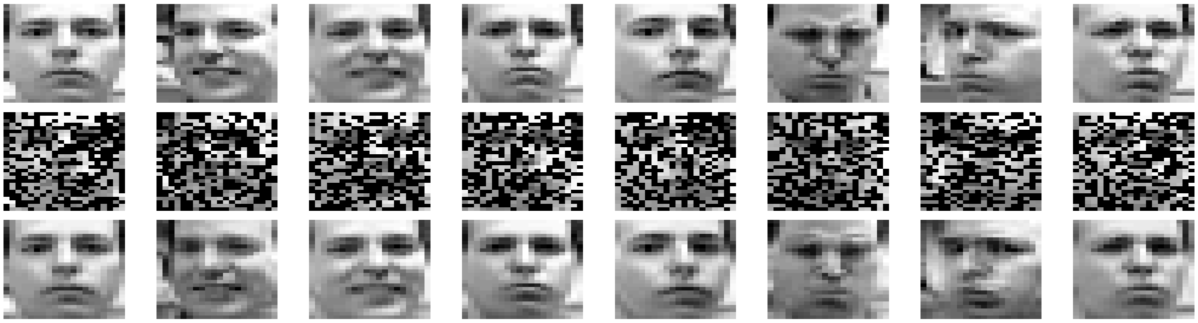


Figure 2: Examples of reconstruction of partially observed test images in Frey faces by applying the Bayesian GP-LVM. Each column corresponds to a test image. In every column, the top panel shows the true test image, the middle panel the partially observed image (where missing pixels are shown in black) and the bottom image is the reconstructed image.

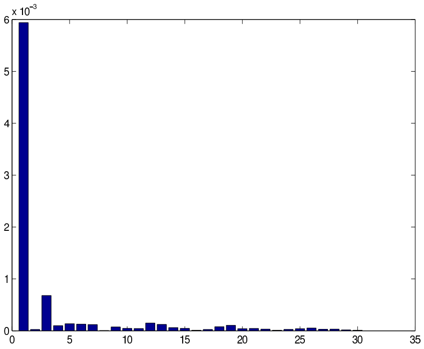


Figure 3: This plot shows the values of the inverse lengthscales found by using the Bayesian GP-LVM with ARD SE kernel in Frey faces.

5.3 Digits Data

In the final experiment we use the Bayesian GP-LVM to build a generative classifier for handwritten digit recognition. We consider the well known USPS digits dataset. This dataset consists of 16×16 images for all 10 digits and it is divided into 7291 training examples and 2007 test examples. We run 10 Bayesian GP-LVMs, one for each digit, on the USPS data base. We used 10 latent dimensions and 50 inducing variables for each model. This allowed us to build a probabilistic generative model for each digit so that we can compute Bayesian class conditional densities in the test data having the form $p(\mathbf{y}_*|Y, \text{digit})$. These class conditional densities are approximated through the ratio of lower bounds in eq. (20) as described in section 4. The whole approach allows us to classify new digits by determining the class labels for test data based on the highest class conditional density value and using a uniform prior over class labels. The test error made by the Bayesian GP-LVM in the whole set of 2007 test points was 95 incorrectly classified digits i.e. 4.73% error.

6 Discussion

We have introduced an approximation to the marginal likelihood of the fully marginalized Gaussian process latent variable model. Our approximation is in the form of a variational lower bound. With the fully marginalized model we can automatically determine the latent dimensionality of a given data set. We demonstrated the utility of this rigorous lower bound on a range of disparate real world data sets.

Our approach can immediately be applied to training Gaussian processes with uncertain inputs where these inputs have Gaussian prior densities. We also envisage several other extensions that become computationally feasible using the same set of methodologies we espouse. Dynamical models based on the GP-LVM have been proposed. It would be straightforward to include a latent space prior with a temporal component. This could be a Kalman filter, a general Gaussian process (Lawrence and Moore, 2007) or an auto regressive Gaussian process (Wang et al., 2006). By using our approach to propagating the Gaussian noise through the dynamics and the latent space a variational lower bound on the likelihood of these models could be derived. The importance of such nonlinear models is clear from the success of unscented Kalman filters and the related ensemble Kalman filter.

The optimization procedure has a similar computational cost to that of previously proposed sparse GP-LVMs. We believe there is scope to improve the speed of the optimization procedure by better exploiting the correlation present in the parameters. A potential strategy would be to use the control points idea used to speed up MCMC in GPs (Titsias et al., 2009) in order to encode the variational posterior, effectively decoupling these correlations and speeding convergence of the optimizer.

Acknowledgements

We wish to thank Mauricio Álvarez for his help with the software implementation. This work is funded by EPSRC Grant No EP/F005687/1 ”Gaussian Processes for Systems Identification with Applications in Systems Biology”.

References

C. M. Bishop. Bayesian PCA. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 482–388, Cambridge, MA, 1999a. MIT Press.

C. M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514, 1999b.

C. M. Bishop and G. D. James. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research*, A327: 580–593, 1993.

L. Csató. *Gaussian Processes — Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.

L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

T. G. Dietterich, S. Becker, and Z. Ghahramani, editors. *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press.

A. Geiger, R. Urtasun, and T. Darrell. Rank priors for continuous non-linear dimensionality reduction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami, FL, 2009.

Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 831–864, Cambridge, MA, 2000. MIT Press.

A. Girard, C. E. Rasmussen, J. Quiñonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - application to multiple-step ahead time series forecasting. In Dietterich et al. (2002), pages 529–536.

N. D. Lawrence. Gaussian process models for visualisation of high dimensional data. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 329–336, Cambridge, MA, 2004. MIT Press.

N. D. Lawrence. Learning for larger datasets with the Gaussian process latent variable model. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*, pages 243–250, San Juan, Puerto Rico, 21-24 March 2007. Omnipress.

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

N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. In Z. Ghahramani, editor, *Proceedings of the International Conference in Machine Learning*, volume 24, pages 481–488. Omnipress, 2007.

T. P. Minka. Automatic choice of dimensionality for PCA. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 598–604, Cambridge, MA, 2001. MIT Press.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

S. T. Roweis, L. K. Saul, and G. E. Hinton. Global coordination of local linear models. In Dietterich et al. (2002), pages 889–896.

M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Weiss et al. (2006).

M. Titsias, N. D. Lawrence, and M. Rattray. Efficient sampling for Gaussian process inference using control variables. In D. Koller, Y. Bengio, D. Schuurmans, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, Cambridge, MA, 2009. MIT Press.

M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.

J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In Weiss et al. (2006).

Y. Weiss, B. Schölkopf, and J. C. Platt, editors. *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.