Hybrid Discriminative-Generative Approach with Gaussian Processes

Ricardo Andrade-Pacheco
acq11ra@sheffield.ac.uk

James Hensman
james.hensman@sheffield.ac.uk

Max Zwießele
m.zwiessele@sheffield.ac.uk

Neil D. Lawrence
n.lawrence@sheffield.ac.uk

Department of Computer Science & Sheffield Institute for Translational Neuroscience
University of Sheffield

Abstract

Machine learning practitioners are often faced with a choice between a discriminative and a generative approach to modelling. Here, we present a model based on a hybrid approach that breaks down some of the barriers between the discriminative and generative points of view, allowing continuous dimensionality reduction of hybrid discrete-continuous data, discriminative classification with missing inputs and manifold learning informed by class labels.

1 Introduction

We consider a framework for modelling with Gaussian processes (GP) which allows us to combine their strengths as both discriminative and generative models. In particular, we extend Gaussian process classification to allow propagation of a generative model through the conditional distribution. This is achieved through a marriage of expectation propagation (EP) [Opper and Winther, 2000, Minka, 2001] with the variational approximations of Titsias and Lawrence [2010]. The resulting framework allows us to deal with mixed discrete-continuous data. We apply it to classification with missing and uncertain inputs, visualization of hybrid binary and continuous data and joint manifold modelling of labelled data.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

2 Discriminative Models

2.1 Overview

From a probabilistic perspective, a *discriminative model* (or *regression model*) represents a conditional density estimate $p(\mathbf{y}|\mathbf{X})$, where some target variables $\mathbf{y} \in \mathbb{R}^{n\times 1}$ are predicted[1] given some known input variables $\mathbf{X} \in \mathbb{R}^{n\times q}$. Hereafter, n represents the number of observations and q the dimensionality of each input. Gaussian process models introduce an additional latent variable $\mathbf{f}$, whose covariance matrix $\mathbf{K_{ff}}$ is computed as a function of the input values. The target points are then related to this latent function through a likelihood function $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$.

Within the GP framework, predictions at a new input position $\mathbf{x}^* \in \mathbb{R}^{1\times q}$ are computed consistently with the training data $\{\mathbf{X}, \mathbf{y}\}$, through the predictive density $p(y^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$. GP models provide an inference engine for non-linear functions, where the marginalization of the prior distribution is tractable. The simplicity of doing inference with them has made GP one of the dominant methods for regression in machine learning. They have also been extended to allow non-linear latent variable models for unsupervised learning [Lawrence, 2005]. However, their tractability is only assured when the likelihood function is Gaussian, i.e., $p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma_i^2)$. Often, it is assumed that $\sigma_i^2 = \sigma^2 \ \forall\, i$, and σ^2 is regarded as the variance of the Gaussian distributed corrupting noise.

2.2 Regression for Non-Gaussian Data

In binary classification, where we take $y_i \in \{0, 1\}$, the realizations of a Gaussian process are normally

[1] For simplicity, we are assuming the target variables to be one-dimensional, although it can be otherwise.

mapped through a *squashing function* $\phi : \mathbb{R} \mapsto (0, 1)$ to provide a set of probabilities $\{\pi_i = \phi(f_i)\}_{i=1}^n$, which can then be used as parameters of a Bernoulli likelihood $p(y_i|f_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$. Such a non-linear transformation over f_i renders exact inference in the resulting model intractable. This led Barber and Williams [1997] to consider the Laplace approximation and Gibbs and MacKay [2000] to adopt a variational lower bound from Jaakkola and Jordan [1996][2] to make progress. The more standard variational approach (often known as variational inference), based on minimizing the Kullback-Leibler divergence between an approximation and the true posterior density, has also been proposed for non-Gaussian data. Seeger [2004] considered this approximation for classification and Tipping and Lawrence [2003] extended the relevance vector machine[3] to heavy tailed data. However, as shown empirically by Kuss and Rasmussen [2005], for the case of classification, standard application of variational inference to *sub-Gaussian* likelihoods can lead to very poor approximations of the marginal likelihood. Instead, the expectation propagation algorithm [Opper and Winther, 2000, Minka, 2001] is generally preferred. Both EP and its variants have been applied to likelihoods that allow semi-supervised learning [Lawrence and Jordan, 2005], ordinal regression [Chu and Ghahramani, 2005] and binary classification [Kuss and Rasmussen, 2005]. However, its application in the context of heavy tailed likelihoods is generally more involved [Jylänki et al., 2011].

3 Generative Models

3.1 Overview

Generative models (or *joint models*) consist of modelling the joint distribution between predictors and response. Gaussian processes have been reformulated as a generative model known as the Gaussian process latent variable model (GP-LVM) [Lawrence, 2005]. In this model, a GP provides a probabilistic mapping between a set of *latent* variables $\mathbf{X} \in \mathbb{R}^{n\times q}$ and a set of observed data variables $\mathbf{Y} \in \mathbb{R}^{n\times p}$, where $q < p$. In the original paper, these latent variables were optimized by maximum likelihood, but Titsias and Lawrence [2010] showed recently that they can be approximately marginalized through a collapsed variational [Hensman et al., 2012] approach. This allows the uncertainty in the latent space to be incorporated in the model and the underlying dimensionality to be determined. Damianou et al. [2012] exploited the ability to determine the latent dimensionality in the context of multi-view learning. Their approach, known as manifold relevance determination (MRD), incorporates multiple views of objects in a model where latent variables are automatically allocated to the *relevant* views, such that some latent dimensions are shared across the views, whilst other are private to a particular view. So far, however, this model has only been applicable to Gaussian data. Here, we extend their approach to non-Gaussian data. The resulting framework allows a range of model extensions including:

1. Classification with uncertain inputs.
2. Dimensionality reduction of non-Gaussian data.
3. Joint modelling of binary labels alongside a data set to form a discriminative latent variable model.

3.2 Joint Models for Non-Gaussian Data

Non-Gaussian data has already been considered in the context of continuous latent variables. The bound of Jaakkola and Jordan [1996] was applied to unsupervised learning of binary data by Tipping [1999] for the principal component analysis (PCA) of binary data (see also Lee and Sompolinsky [1999], Schein et al. [2003]). These models are related to GP models due to the shared challenge of combining a Gaussian prior with a non-Gaussian likelihood. This arises due to the duality between the latent variables (equivalent to our *inputs* $\mathbf{X}$) and desired principal subspace generated by the mapping $\mathbf{W} \in \mathbb{R}^{p\times q}$ in PCA. By associating the j-th column of the mapping matrix $\mathbf{w}_j$ with the j-th output dimension of the data $\mathbf{y}_j$, we can write the associated mapping from the latent variables as $\mathbf{y}_j = \mathbf{X}\mathbf{w}_j$. We induce $\mathbf{w}_j$ to be jointly Gaussian distributed, as in a GP, by defining the usual spherical Gaussian prior independently over the latent variables $x_{ij} \sim \mathcal{N}(0, 1)$. Indeed, marginalizing $\mathbf{w}_j$ with a Gaussian prior leads directly to a GP with a linear covariance function. This was the relation exploited by Lawrence [2005] to generalize PCA in the GP-LVM.

4 GP Variational Inference

4.1 Regression Case

To make GP models feasible for large data sets, the burden of inverting the covariance matrix (computational complexity of $\mathcal{O}(n^3)$ and storage of $\mathcal{O}(n^2)$) must be reduced. Low rank approximations [Quiñonero Candela and Rasmussen, 2005, Snelson and Ghahramani, 2006, Lawrence, 2007, Seeger et al., 2003], in regression problems, make use of variables associated

[2]This variational lower bound exploited the log-convexity of a sigmoidal squashing function, but does *not* follow the standard approach to variational inference.

[3]A sparse Bayesian regression model that can also be expressed as a GP with a degenerate covariance.

with a set of inducing inputs $\mathbf{Z} \in \mathbb{R}^{m \times q}$, where the elements of $\mathbf{Z}$ and $\mathbf{X}$ belong to the same domain. They result in computational complexity[4] of $\mathcal{O}(m^2 n)$ and storage demands of $\mathcal{O}(mn)$.

The deterministic training conditional (DTC) approximation assumes a deterministic relation between the inducing inputs and the latent variables at the observed inputs. In contrast, the fully independent training conditional (FITC) approximation keeps the exact variance of each observation, but assumes independence between them. Unfortunately, neither of these approaches form a lower bound on the marginal likelihood of the Gaussian process. This issue was resolved by Titsias [2009], who introduced a variational approximation that resulted in the lower bound

$$\begin{aligned}\mathcal{L}_T =& \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{Q_{ff}} + \sigma^2 \mathbf{I}\right) \\ &- \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right),\end{aligned} \tag{1}$$

where $\mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\mathbf{K_{uf}}$, $\mathbf{K_{uu}}$ is the covariance function computed between the inducing inputs and $\mathbf{K_{uf}}$ is the covariance function computed across the inducing inputs and the training data. The first term of this lower bound corresponds to the DTC likelihood approximation. The second term can be interpreted as a correction factor that penalizes using $\mathbf{Q_{ff}}$ instead of $\mathbf{K_{ff}}$, depending on how much they differ form each other. A rigorous lower bound on the log-marginal likelihood allows joint optimization of the inducing inputs and hyperparameters without overfitting. This bound was also exploited by Titsias and Lawrence [2010] to allow for approximate variational marginalization of the latent variables in the Bayesian GP-LVM. The success of the EP for approximate inference in non-Gaussian data motivates us to combine EP with this variational bound to provide a general framework for hybrid learning of Gaussian and non-Gaussian data.

4.2 EP for GP Variational Inference

For Gaussian process models, EP combines a Gaussian prior $p(\mathbf{f}|\mathbf{X})$ with a set of site approximations to the likelihood[5] $\{t_i(f_i) \approx p(y_i|f_i)\}_{i=1}^n$. This results in an approximation to the posterior density of $\mathbf{f}$ given by

$$q(\mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{1}{Z_{EP}} p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{n} t_i(f_i) \propto \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

where Z_{EP} is the normalizing constant of $q(\mathbf{f}|\mathbf{y}, \mathbf{X})$ (see Williams and Rasmussen [2006] for notation). The site approximations can be seen as combining to provide a Gaussian-like approximation to the likelihood

$$p(\mathbf{y}|\mathbf{f}) \approx \tilde{\mathbf{Z}} \times \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{3}$$

for some constant $\tilde{\mathbf{Z}}$.

To combine the EP approximation with the variational lower bound in (1), we need to define an EP algorithm based on the DTC low rank approximation. We refer to this algorithm as EP-DTC. Let $\{\tilde{\nu}_i\}_{i=1}^n$ and $\{\tilde{\tau}_i\}_{i=1}^n$ be the natural parameters associated with $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_i)$ and $\tilde{\boldsymbol{\Sigma}} = (\tilde{s}_{ij})$, where $\tilde{s}_{ij} = 0 \; \forall \; i \neq j$. Suppose that at the i-th iteration[6] the natural parameters of the site approximation change by $\Delta\tilde{\nu}_i$ and $\Delta\tilde{\tau}_i$. Then, it can be shown (see Supplementary material) that the updates of the posterior moments are given by

$$\boldsymbol{\Sigma}^{\text{new}} = \mathbf{K_{fu}}(\mathbf{L}\mathbf{L}^\top + \mathbf{k}_i \Delta\tilde{\tau}_i \mathbf{k}_i^\top)^{-1}\mathbf{K_{uf}}, \tag{4}$$

$$\boldsymbol{\mu}^{\text{new}} = \boldsymbol{\mu} + (\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\mu_i)\,\mathbf{s}_i^{\text{new}}, \tag{5}$$

where $\mathbf{k}_i$ is the i-th column of $\mathbf{K_{uf}}$, $\mathbf{s}_i^{\text{new}}$ is the i-th column of $\boldsymbol{\Sigma}^{\text{new}}$ and $\mathbf{L}$ is the Cholesky decomposition of $(\mathbf{K_{uu}} + \mathbf{K_{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K_{fu}})$ from the previous iteration. The derivation follows closely that of Naish-Guzman and Holden [2008], who combined EP with the FITC approximation of Snelson and Ghahramani [2006]. However, for the case when the likelihood is not Gaussian, EP-DTC allows us to approximate (1) as

$$\begin{aligned}\mathcal{L}_E =& \log \mathcal{N}\left(\tilde{\boldsymbol{\mu}}|0, \mathbf{Q_{ff}} + \tilde{\boldsymbol{\Sigma}}\right) \\ &- \frac{1}{2}\mathrm{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}}.\end{aligned} \tag{6}$$

$\mathcal{L}_E$ retains the trace term from the bound of Titsias [2009], only rather than being weighted by the noise variance from the process, the elements of the trace are now weighted by the variances from the site approximations.

It is possible to extend the variational bound in (6) to handle uncertainty on the inputs of the Gaussian process. Girard et al. [2003] and Girard and Murray-Smith [2005] are able to work with noisy inputs in the *predictions* of a GP regression model, by propagating the uncertainty through the covariance. We additionally use variational inference to approximate the *marginal likelihood*, which allow us to incorporate uncertain inputs in the training procedure. This makes possible, within our framwork, to handle uncertain inputs in classification models and to construct hybrid continuous-discrete dimensionality reduction models.

[4] For efficiency, we need to take $m << n$. Mathematically, we find that the original GP is recovered as $m \to n$.

[5] EP can be defined in a more general way, but we will only use this definition for simplicity.

[6] EP is an iterative algorithm in which site approximations are updated one at time, until convergence is achieved. In this case, the i-th iteration refers to the step in which the paremeters of $t_i(f_i)$ are updated.

Table 1: Sparse Binary Classification Models.

				EP-FITC		$\mathcal{L}_E$	
data set	q	m	train/test	error	nlp	error	nlp
synthetic	2	4	250/1000	0.0910	*0.2595*	0.0930	*0.2618*
crabs	5	10	80/120	0.0450	*0.2493*	0.0458	*0.2943*
banana	2	20	400/4900	0.1092	*0.2535*	0.1083	*0.2543*
breast-cancer	9	2	200/77	0.2610	*0.5242*	0.2805	*0.5363*
diabetes	8	2	468/300	0.2273	*0.4789*	0.2290	*0.4922*
flare-solar	9	3	666/400	0.3410	*0.5932*	0.3250	*0.5959*
german	20	4	700/300	0.2470	*0.4985*	0.2637	*0.5114*
heart	13	2	170/100	0.1600	*0.4003*	0.1610	*0.4221*
thyroid	5	6	140/75	0.0560	*0.2087*	0.0560	*0.2164*
titanic	3	2	150/2051	0.2373	*0.5180*	0.2368	*0.5274*
two-norm	20	2	400/7000	0.0239	*0.1273*	0.0241	*0.1682*
waveform	21	10	400/4600	0.0966	*0.2406*	0.0995	*0.2682*

4.3 Sparse GP Binary Classification

Before we proceed to including uncertain inputs in our framework, we first compare the quality of the new bound $\mathcal{L}_E$ with EP-FITC. We applied both approximations to a set of classification benchmarks (12 data sets: two from Ripley's collection[7] and 10 from Gunnar Rätsch's benchmarks[8]). Table 1 shows the error and negative log-probabilities obtained with each model. In each case, the number of inducing inputs used was the same for both models. The covariance functions were all taken to be an exponentiated quadratic with white noise. The values in the table correspond to the average results of 10 folds over the data (except for the synthetic data set, which is already divided into test and training sets). In the case of the crabs data set, we randomly created 10 test/train partitions of size 80/120 ensuring that each training set had equal number of observations per class. Rätsch's benchmark contains 100 training and test splits per data set. In these experiments, we worked with 10 splits randomly chosen. Hyperparameters and inducing inputs were optimized jointly by scale conjugate gradients. For each split, we tried three differently initializations and retained the model with the highest marginal likelihood for testing.

In the tests, the models both exhibited a similar performance, with (if anything) EP-FITC being marginally better. These results give us confidence that our approach is competitive.

[7] `http://www.stats.ox.ac.uk/pub/PRNN/`.

[8] `http://theoval.cmp.uea.ac.uk/~gcc/matlab`.

5 Discriminative-Generative Model

Lasserre et al. [2006] present a general framework for discriminative training of generative models, that relies on a model formulation with an additional set of parameters[9]. We follow a similar approach, by using a variational formulation. So far, we have assumed that we are given a full set of input-output pairs for each data point $\{\mathbf{x}_i, y_i\}_{i=1}^n$. The advantage of extending the variational formulation with EP is that we can now consider distributions over $\mathbf{x}_i$, which allows inference with uncertain inputs and multi-view learning for hybrid data sets. We will assume that we have a Gaussian approximation to the posterior density $q(\mathbf{X})$ in place of $\mathbf{X}$. Given (6), the formulation of a variational bound in the form of the one presented by Titsias and Lawrence [2010] is straightforward (see Supplementary material). Such a bound is formulated as

$$\begin{aligned}\mathcal{L}_H = \log \mathcal{N}\left(\tilde{\boldsymbol{\mu}}|0, \boldsymbol{\Psi}_1^\top \mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Psi}_1 + \boldsymbol{\Lambda} + \tilde{\boldsymbol{\Sigma}}\right) - \tilde{\psi}_0 \\ + \operatorname{tr}\left(\mathbf{K}_{\mathbf{uu}}^{-1}\tilde{\boldsymbol{\Psi}}_2\right) + \mathrm{KL}\left(q(\mathbf{X})\|p(\mathbf{X})\right) + \tilde{\mathbf{Z}},\end{aligned} \tag{7}$$

where $\tilde{\psi}_0 = \operatorname{tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\langle \mathbf{K}_{\mathbf{ff}}\rangle_{q(\mathbf{X})}\right)$, $\boldsymbol{\Psi}_1 = \langle \mathbf{K}_{\mathbf{uf}}\rangle_{q(\mathbf{X})}$, $\tilde{\boldsymbol{\Psi}}_2 = \langle \mathbf{K}_{\mathbf{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K}_{\mathbf{fu}}\rangle_{q(\mathbf{X})}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix such that $\boldsymbol{\Lambda}_{ii} = \operatorname{tr}\left(\tilde{\boldsymbol{\Psi}}_{2(i)}\mathbf{K}_{\mathbf{uu}}^{-1}\right) - \boldsymbol{\Psi}_{1(i)}^\top \mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Psi}_{1(i)}$. The sub-index (i) means that we are only taking the i-th column of the corresponding matrix.

Notice that the first term in the r.h.s. of (7) has no longer the form of the DTC approximation. Instead, its form is closer to the FITC approximation[10],

[9] Additional to the parameters of the discriminative and generative models.

[10] The marginal likelihood in the FITC approximation is given by $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{Q}_{\mathbf{ff}} + \operatorname{diag}\left(\mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}}\right) + \sigma^2\mathbf{I})$.

as it can be expressed as the sum of a diagonal and a non-diagonal matrices. An EP algorithm can be implemented for this new covariance form. Updates computation in this new algorithm resemble those of EP-FITC, but the origin of the terms in the covariance is conceptually different. We start by re-expressing the non-diagonal term in the prior covariance as $\boldsymbol{\Psi}\mathbf{R}^\top\mathbf{R}\boldsymbol{\Psi}^\top$, where $\mathbf{R}$ is the Cholesky decomposition of $\mathbf{K}_{\mathbf{uu}}^{-1}$. Given our Gaussian approximation to the likelihood, the structure of the prior covariance will be kept in the posterior covariance. Hence, the posterior moments can be decomposed as

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Psi}}\mathbf{R}^\top\mathbf{R}\hat{\boldsymbol{\Psi}}^\top + \hat{\boldsymbol{\Lambda}}, \tag{8}$$

$$\boldsymbol{\mu} = \boldsymbol{\omega} + \hat{\boldsymbol{\Psi}}\boldsymbol{\gamma}, \tag{9}$$

where $\hat{\boldsymbol{\Psi}}$ has the same shape of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Lambda}}$ is a diagonal matrix, $\boldsymbol{\omega} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{m\times 1}$. Suppose that at the i-th iteration the natural parameters of the likelihood approximation are increased by $\Delta\tilde{\nu}_i$ and $\Delta\tilde{\tau}_i$. Then, the new posterior covariance and posterior mean can be computed by updating each one of their components (see Supplementary material) as follows:

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \hat{\boldsymbol{\Lambda}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}^2}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\mathbf{e}_i^\top, \tag{10}$$

$$\hat{\boldsymbol{\Psi}}^{\text{new}} = \hat{\boldsymbol{\Psi}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\hat{\boldsymbol{\psi}}_i, \tag{11}$$

$$\delta_i = \frac{\Delta\tilde{\tau}_i}{1+\Delta\tilde{\tau}_i s_{ii}}, \tag{12}$$

$$\mathbf{R}^{\text{new}} = \text{Cholesky}\left(\mathbf{R}^\top\left(\mathbf{I} - \mathbf{R}\hat{\boldsymbol{\psi}}_i\delta_i\hat{\boldsymbol{\psi}}_i^\top\mathbf{R}^\top\right)\mathbf{R}\right), \tag{13}$$

$$\boldsymbol{\omega}^{\text{new}} = \boldsymbol{\omega} + \frac{(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\omega_i)\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i, \tag{14}$$

$$\begin{aligned}\boldsymbol{\gamma}^{\text{new}} &= \hat{\boldsymbol{\Psi}}^{\text{new}}\boldsymbol{\gamma} \\ &+\hat{\boldsymbol{\Psi}}^{\text{new}}\left((\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\tilde{\mu}_i)\mathbf{R}^{\text{new}\top}\mathbf{R}^{\text{new}}\hat{\boldsymbol{\psi}}_i^{\text{new}}\right),\end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\Lambda}} = (\hat{\lambda}_{ij})$, $\hat{\boldsymbol{\psi}}_i$ is the i-th column of $\hat{\boldsymbol{\Psi}}$ and $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^n$.

This gives us a general algorithm that can be used across a range of different applications. We now consider applications of the model in three different domains: classification with uncertain inputs, dimensionality reduction of non-Gaussian data and classification using a hybrid discriminative-generative approach.

6 Experiments

6.1 Classification With Uncertain Inputs

In probabilistic classification, we are not only interested in the class estimates, but also in a measure of the uncertainty about our predictions. If we are aware that there is uncertainty associated to the inputs on which the classification is based, it makes sense to incorporate this uncertainty in our predictions. Even if the class predictions do not change, the confidence intervals may. We present a couple of examples to illustrate how our framework handles such uncertainty.

6.1.1 Toy Example

We show how the decision boundary in a classification model is affected by the increase in the input's uncertainty. We considered an artificial binary classification problem. For an asymmetric increase in the uncertainty (Figure 1a), where only the inputs of one class become more uncertain, the decision boundary becomes more tightly wrapped around the inputs with less uncertainty. In contrast, when uncertainty increases in both sets of input variables (Figure 1b) the decision boundary becomes much smoother overall.

6.1.2 Olivetti Face Data Set

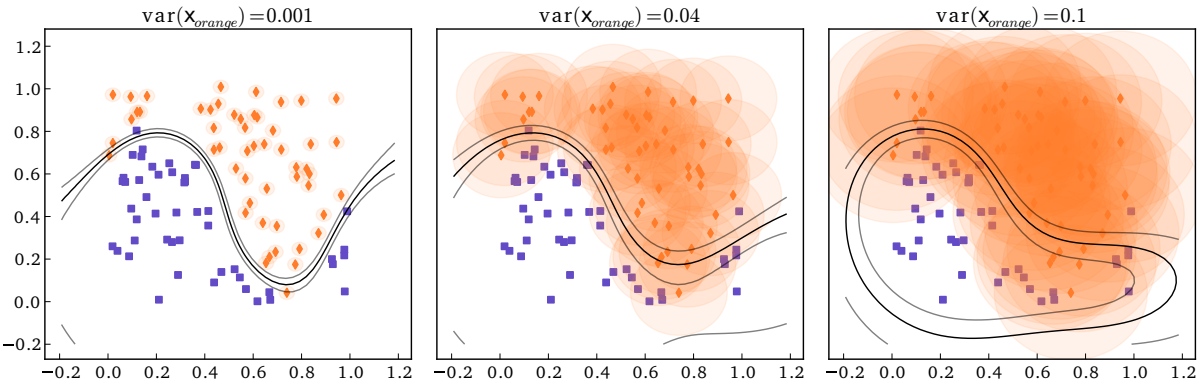
We consider the case of having a trained classifier, but with missing components of the test point $\mathbf{x}^*$. A simple solution would be to replace the missing values with the corresponding means from the training data. Our framework allows us to extend this idea by replacing the missing data with a Gaussian distribution, whose mean and variance matches the training data. We applied this idea using the Olivetti face data set[11] to predict whether or not a person is wearing glasses. We took a random 50/50 split to train two models: a standard GP-EP and a hybrid discriminative-generative model.

On the test data, to simulate missing values, we removed a varying portion of pixels from the images (Figure 2). We then computed the class probability estimates of both models. Notice that, as the proportion of missing values increases, the hybrid model becomes less certain and begin to converge towards the prior probability of an individual wearing glasses (about 30%). In contrast, the standard model just becomes certain that the image is a face with no glasses. Table 2 shows a comparison of the errors and negative log-probabilities obtained after introducing uncertainty.

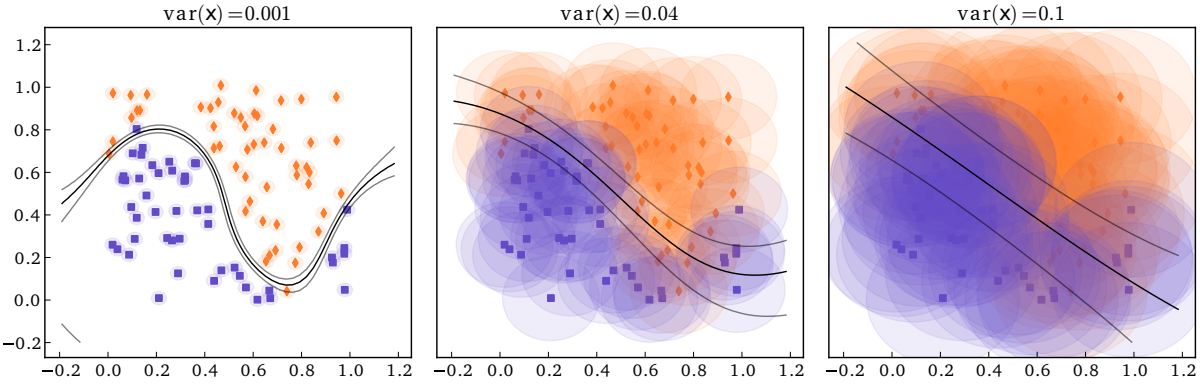
6.2 Dimensionality Reduction of Non-Gaussian Data

Manifold learning techniques model a high dimensional process, by encoding its dominant sources of variation in a latent process of lower dimensionality. Commonly,

[11] `http://www.cs.nyu.edu/~roweis/data.html`.



(a) Asymmetric uncertainty. The uncertainty increase on the inputs of one class only, from left to left, causes the decision boundary to shrink around the class with less uncertainty.



(b) Symmetric uncertainty. The uncertainty increase on the inputs of both classes, from left to right, causes a smoothing out of the decision surface.

Figure 1: Classificaton with uncertain inputs. Class elements are distinguished by colour and marker shape. The shaded ellipses represent 95% confidence intervals for each uncertain input. The contour lines represent the probabilities (bold line 0.5, light lines 0.4 and 0.6) of the points belonging to the *orange class*.

a Gaussian noise model is assumed, for example, in the probabilistic PCA and the Bayesian GP-LVM. By integrating EP to the GP variational framework, we can apply dimensionality reduction techniques on data with non-Gaussian noise. We applied our model on the Zoo data set[12], where 101 animals from 7 categories (mammal, bird, fish, etc.) are described by 15 boolean attributes and 1 numerical attribute. The hybrid approach can model each attribute with a different noise model. We used a Bernoulli and a Gaussian likelihoods for the boolean and numerical attributes, respectively. Figure 3 shows the latent representation of the data.

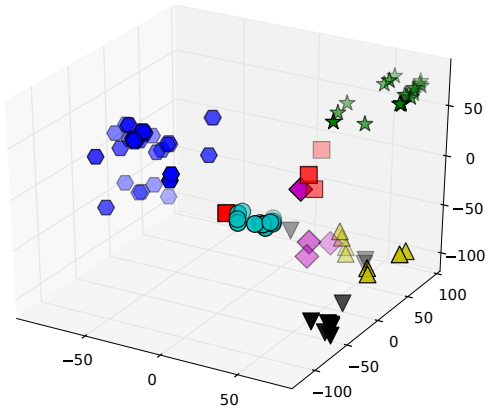
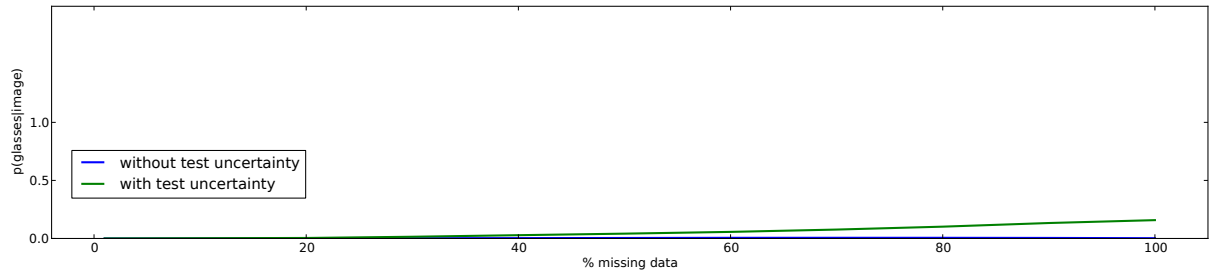


Figure 3: Three dimensional representation of the Zoo data set. The actual labels (unseen by the algorithm) are represented by different colors and bullets: mammals (blue hexagons), birds (green stars), reptiles (red squares), fish (cyan circles), amphibians (purple diamonds), insects (olive-green triangles) and crustaceans (black triangles).

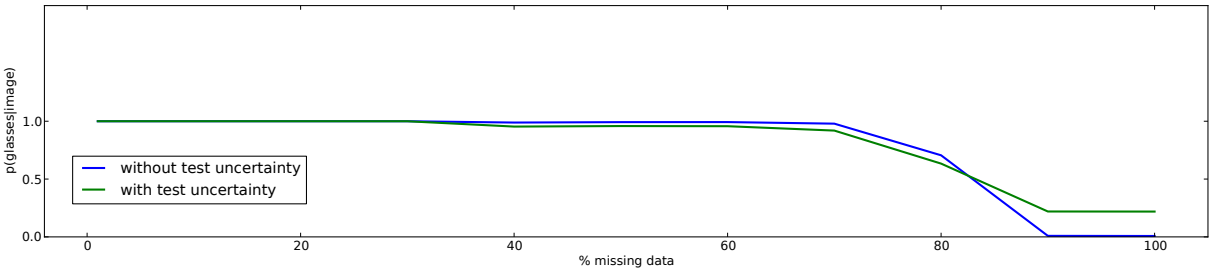
6.3 Discriminative Latent Variable Model

The manifold relevance determination approach of Damianou et al. [2012] considers multiple views of the same data set, allowing each view to be associated with

[12] `http://archive.ics.uci.edu/ml/datasets/Zoo`.



(a) Without glasses.



(b) With glasses.

Figure 2: Graceful failure with missing data. Increasing quantities of missing data are shown for two test cases, with the average (over 100 permutations) classification probability. For the standard GP-EP, missing pixels were replaced with the mean from the training data, for the hybrid model the independent marginal probability of the pixel is used. In the uncertain case, as more data are removed, the model predicts that the image contains glasses with $p = 0.3$, which matches the prior for the data set. Without consideration of the uncertainty, the model will alway predict that the image contains glasses with probability 0, such is the appearance of the mean of the pixels.

Table 2: Olivetti Faces Classification.

	Without uncertainty		**With uncertainty**	
	error	nlp	error	nlp
No missing data	0.0200	*21.0248*		
50% of pixels randomly missing	0.1650	*94.8951*	0.1650	*73.5056*
Half of face occluded	0.1650	*69.1357*	0.1650	*67.0423*

private and shared portions of the latent space. We can construct a discriminative latent variable model, which includes class labels and data points as different views. We considered the 3s and 5s from the USPS digits database. In Figure 4, we show an example where we used 50 observations to train the model and learn a 2-dimensional latent space. Notice that the discrimination occurs across the first latent dimension, whilst the second latent dimension is used to represent non-discriminative variation in the data. The figure shows the position of 100 unlabelled test data points mapped into the latent space alongside the locations of the training data.

We next followed Urtasun and Darrell [2007] in fitting a discriminative manifold model to labelled training sets of varying sizes. The error rates of the resulting models on 100 test points are shown in Figure 5a. Our results are similar to those presented by Urtasun and Darrell [2007] (our data set partitions differ and our error appears to share the same form, but be worse overall). However, when we compared to standard EP-GP (Figure 5b), our performance was significantly worse. This contrasts to the results in Urtasun and Darrell [2007], who found standard GP classification to underperform on this data set. In our experience, standard EP-GP classification *can* perform badly when the initialization is poor and random restarts are not tried. This can explain the discrepancy between our results and theirs. To achieve similar results to EP-GP classification (and therefore exploit the advantages of the hybrid discriminative-generative model) we be-

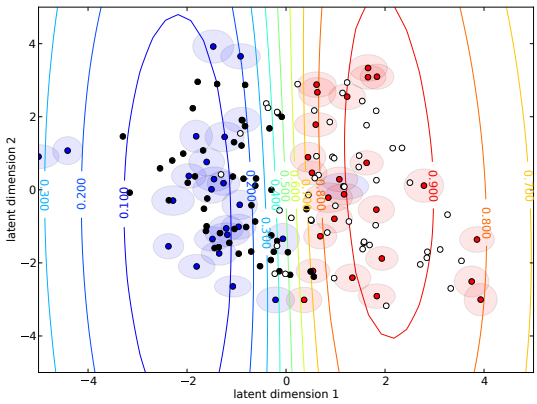
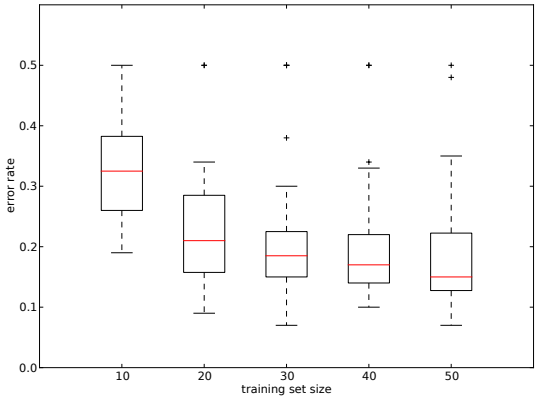
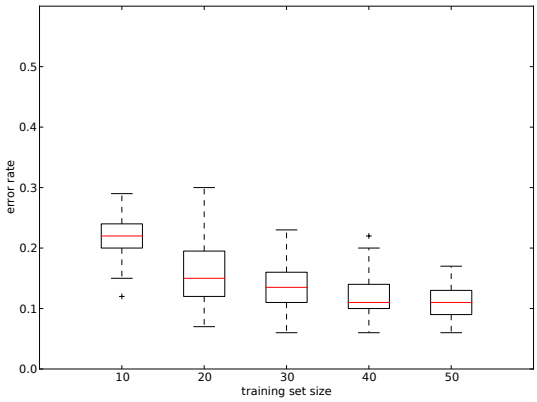


Figure 4: Lower dimensional representation of the USPS digits. The blue and red points represent the examples of 3's and 5's, respectively, in the training set. The shaded ellipses represent the uncertainty of the latent variables. The black and white colors represent the test points (3's and 5's) mapped to the learnt manifold. The contour lines represent the label probabilities (of being five).



(a) Hybrid model.



(b) Standard EP-GP.

Figure 5: *Left:* Classification error rates on the USPS data for the hybrid model as the data set size increases. *Right:* Classification errors rates on the USPS data for standard EP-GP classification. Results are bar and whisker plots summarizing 40 different sub sampled training sets.

lieve that our generative model needs to be more representative of the underlying data. One possible way in which this could be achieved would be through use of the deep GP formalism of Damianou and Lawrence [2013].

7 Conclusions

Gaussian processes have traditionally been used as either discriminitive *or* generative models. By combining the EP approximation with a variational bound on the marginal likelihood, we have developed a framework for building hybrid discriminative-generative models with GP. This required the development of two new EP algorithms for sparse GP. The first algorithm was used to define a model which is comparable with the generalized FITC classification, while the second is able to incorporate estimates of input's uncertainty into the routine. These allowed us to incorporate discriminitive Gaussian processes into a probabilistic model such as the Bayesian GP-LVM.

We have shown how the addition of input's uncertainty leads to well behaved algorithms, in particular, when training on data where such uncertainty is class-dependent and when predicting using missing inputs. We are able to use these techniques to apply the Bayesian GP-LVM on non-Gaussian data and make continuous latent representations of mixed data types.

Finally, in a further contribution in this volume [Hensman et al., 2014] a novel variational approach, tilted variational Bayes (TVB), is proposed for dealing with non-Gaussian likelihoods. This approach appears competitive with expectation propagation. Our next goal is to combine TVB with the low rank approximation of Titsias and Lawrence [2010] to form an efficient hybrid model that provides a rigorous lower bound on the marginal likelihood.

Acknowledgements

RAP is supported by CONACYT and SEP scholarships, MZ by EU FP7-PEOPLE Project Ref 316861, JH by MRC Fellowship "Bayesian models of expression in the transcriptome for clinical RNA-seq" and NL by EU FP7-KBBE Project Ref 289434 and EU FP7-HEALTH Project Ref 305626.

References

D. Barber and C. K. I. Williams. Gaussian processes for Bayesian classification via hybrid Monte Carlo. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.

C. M. Bishop and B. J. Frey, editors. *Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, pages 1019–1041, 2005.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31.

A. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. In J. Langford and J. Pineau, editors, *Proceedings of the International Conference in Machine Learning*, volume 29, San Francisco, CA, 2012. Morgan Kauffman.

M. N. Gibbs and D. J. C. MacKay. Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

A. Girard and R. Murray-Smith. Gaussian processes: Prediction at a noisy input and application to iterative multiple-step ahead forecasting of time-series. In R. Murray-Smith and R. Shorten, editors, *Switching and Learning in Feedback Systems*, volume 3355 of *Lecture Notes in Computer Science*, pages 158–184. Springer Berlin Heidelberg, 2005.

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 S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 529–536, Cambridge, MA, 2003. MIT Press.

J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, Cambridge, MA, 2012.

J. Hensman, M. Zwiessele, and N. D. Lawrence. Tilted variational Bayes. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Workshop on Artificial Intelligence and Statistics*, volume 33, Iceland, 2014. JMLR W&CP 33.

T. S. Jaakkola and M. I. Jordan. Computing upper and lower bounds on likelihoods in intractable networks. In E. Horvitz and F. V. Jensen, editors, *Uncertainty in Artificial Intelligence*, volume 12, San Francisco, CA, 1996. Morgan Kauffman.

P. Jylänki, J. Vanhatalo, and A. Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12: 3227–3257, 2011.

M. J. Kearns, S. A. Solla, and D. A. Cohn, editors. *Advances in Neural Information Processing Systems*, volume 11, Cambridge, MA, 1999. MIT Press.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6: 1679–1704, 2005.

J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006.

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. 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 and M. I. Jordan. Semi-supervised learning via Gaussian processes. In L. Saul, Y. Weiss, and L. Bouttou, editors, *Advances in Neural Information Processing Systems*, volume 17, pages 753–760, Cambridge, MA, 2005. MIT Press.

D. D. Lee and H. Sompolinsky. Learning a continuous hidden variable model for binary data. In Kearns et al. [1999].

T. P. Minka. Expectation propagation for approximate Bayesian inference. In J. S. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence*, volume 17, San Francisco, CA, 2001. Morgan Kauffman.

A. Naish-Guzman and S. Holden. The generalized FITC approximation. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 1057–1064. MIT Press, Cambridge, MA, 2008.

M. Opper and O. Winther. Gaussian processes for classification: Mean field algorithms. *Neural Computation*, 12:2655–2684, 2000.

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.

A. I. Schein, L. K. Saul, and L. H. Ungar. A generalized linear model for principal component analysis of binary data. In Bishop and Frey [2003].

M. Seeger. Gaussian processes for Machine Learning. *International Journal of Neural Systems*, 14(2):69–106, 2004.

M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In Bishop and Frey [2003].

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

M. E. Tipping. Probabilistic visualisation of high-dimensional binary data. In Kearns et al. [1999], pages 592–598.

M. E. Tipping and N. D. Lawrence. A variational approach to robust Bayesian interpolation. In C. Molina, T. Adali, J. Larsen, M. V. Hulle, S. Douglas, and J. Rouat, editors, *Neural Networks for Signal Processing XIII*, pages 229–238. IEEE, 2003.

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 Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.

M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. In Y. W. Teh and D. M. Titterington, editors, *Proceedings of the Thirteenth International Workshop on Artificial Intelligence and Statistics*, volume 9, pages 844–851, Chia Laguna Resort, Sardinia, Italy, 13-16 May 2010. JMLR W&CP 9.

R. Urtasun and T. Darrell. Discriminative Gaussian process latent variable model for classification. In Z. Ghahramani, editor, *Proceedings of the International Conference in Machine Learning*, volume 24. Omnipress, 2007. ISBN 1-59593-793-3.

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