MiDGaP: Mixture Density Gaussian Processes

Jaleh Zand & Stephen Roberts
Machine Learning Research Group &
Oxford-Man Institute of Quantitative Finance
University of Oxford
{jz,sjrob}@robots.ox.ac.uk

Abstract

Gaussian Processes (GPs) have become a core technique in machine learning over the last decade, with numerous extensions and applications. Although several approaches exist for warping the conditional Gaussian posterior distribution to other members of the exponential family, most tacitly assume a unimodal posterior. In this paper we present a mixture density model (MDM) allowing multi-modal posterior distributions from GPs. We make explicit comparison with alternate models, namely the Mixture Density Network (MDN) and Mixture of GP Experts (GPE). Unlike MDN approaches, we allow full probability distributions over the latent variables that encode the mixture posterior, allowing uncertainty to propagate in a principled manner. Unlike the GPE methods, we achieve non-Gaussian posteriors within a single GP model. We showcase the performance of the approach on synthetic and real timeseries data sets. Our results indicate that not only is the approach competitive in terms of error metrics but also provides further insight into the multiplicity of potential paths a timeseries may take in the future.

1 Introduction

Most prediction models focus on inferring a unimodal posterior distribution, often assumed to be Gaussian with a mean estimating the conditional average of the target data conditioned on the input. Although this provides valuable models in a range of problem domains, it fails to entertain the notion that the distribution over predictions, or missing data, might be *multimodal*. In this paper we consider the posterior distribution to be modelled as a (finite) mixture of Gaussians, allowing for a rich variety of possible posterior forms, including multimodality, asymmetric and non-Gaussian, all of which may be approximated by the appropriate mixture model. Such *Mixture Density Networks* (MDNs) have been long known in the literature, since their introduction in Bishop (1994). We extend the approach by placing a Gaussian Process (GP) over the latent variable set which encodes the final posterior distribution, thus allowing full measures of uncertainty to propagate. Furthermore, we allow the model to be dynamic enabling us to look at changes in the posterior over time, identifying regions in a data stream where multiple possible futures are likely. We apply our approach to synthetic data to show its functioning as well as showcasing its operation on several real-world data sets.

There have been several Gaussian Process mixture models presented in the literature. Tresp (2001) developed a mixture of Gaussian Process Experts (GPE) model, which was re-formulated by Rasmussen & Ghahramani (2002). Both approaches were inspired by the the Mixture of Experts framework of Jacobs et al. (1991). All these approaches look to decompose the input domain into a set of regions with either soft or strict borders. The model we present in this paper, however, aims to infer a multimodal posterior distribution, extending the range of posterior models that the Gaussian Process framework can contend with.

The contributions of this paper are as follows. Firstly we present a dynamic model that allows a non-Gaussian, multimodal posterior distribution over forecasts, via a finite mixture of Gaussians.

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

The parameters of the mixture model are themselves uncertain variables, whose posterior variability we propagate upwards into the mixture model itself. Secondly, we apply this approach to synthetic and real problem sets, showing that there are clear regions in which multiple future outcomes are postulated.

The structure of this paper is as follows; in Section 2 we present the structure of the model, discussing the mixture model for the predictive distribution and solutions to its inference. In section 3 we test our approach against other models on a set of real data sets, showcasing its performance. We conclude in Section 4.

2 Model framework

Most generally, we consider a time series $\mathbf{y_t}$ and aim to make a prediction for $\mathbf{y_t}$ using the input variable set $\mathbf{x_t}$. In order to make the prediction we model the conditional density $p(\mathbf{y_t}|\mathbf{x_t})$.

We start with a model using m mixture components, similar to that presented by Bishop (1994):

$$p(\mathbf{y_t}|\mathbf{x_t}) = \sum_{i=1}^{m} \alpha_i(\mathbf{x_t})\phi_i(\mathbf{y_t}|\mathbf{x_t}), \tag{1}$$

in which

$$\phi_i(\mathbf{y_t}|\mathbf{x_t}) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_i(\mathbf{x_t})} \exp\left\{-\frac{\|\mathbf{y_t} - \mu_i(\mathbf{x_t})\|^2}{2\sigma_i^2(\mathbf{x_t})}\right\}. \tag{2}$$

Noting that the mixture model $p(\mathbf{y_t}|\mathbf{x_t})$ in equation 1 offers extreme flexibility in modelling the posterior distribution over $\mathbf{y_t}$.

We assume the parameters of the mixture model, the means $\mu_i(x)$, the mixing coefficients $\alpha_i(x)$ and the variances of each mixture, $\sigma_i(x)$, to be continuous functions of $\mathbf{x_t}$. The parameters of the mixture model are taken to be transformations of a set of latent variables $\mathbf{z}^\alpha, \mathbf{z}^\sigma, \mathbf{z}^\mu$, conditioned on $\mathbf{x_t}$. We make the following assumptions for the priors. The mixture coefficients, α_i, must satisfy $\sum_{i=1}^{m} \alpha_i(\mathbf{x_t}) = 1$. To enforce this we use the softmax transform from the latent variables:

$$\alpha_i = \frac{\exp(\mathbf{z}_i^\alpha)}{\sum_{j=1}^{m} \exp(\mathbf{z}_j^\alpha)}. \tag{3}$$

The standard deviations have to be non-zero and positive and therefore we adopt the following transform, conditioning on latent variables:

$$\sigma_i = \exp(\mathbf{z}_i^\sigma). \tag{4}$$

Finally for the mean priors we use the below:

$$\mu_i = \mathbf{z}_i^\mu. \tag{5}$$

The data likelihood function is hence:

$$L = \prod_{t=1}^{n} p(\mathbf{y_t}, \mathbf{x_t}) = \prod_{t=1}^{n} p(\mathbf{y_t}|\mathbf{x_t})p(\mathbf{x_t}). \tag{6}$$

Noting that the error function is simply the negative-log of the likelihood provides an error for each data point (dropping the final term $-\log p(\mathbf{x_t})$):

$$E_t = -\ln\left\{\sum_{i=1}^{m} \alpha_i(\mathbf{x_t})\phi_i(\mathbf{y_t}|\mathbf{x_t})\right\}. \tag{7}$$

The MDM model that we present here estimates the latent variables $\mathbf{z}$ using GPs. Therefore $\mathbf{z} = \mathcal{GP}(\mathbf{x_t})$. We use a GP with a squared exponential kernel and hyperparameters $\theta = \{\sigma_\mathbf{n}, \sigma_\mathbf{f}, \mathbf{L}\}$, where σ_n is the noise standard deviation, σ_f is the output scale, and L is the lengthscale hyperparameter of the kernel. As we lack direct observations for the latent variables $\mathbf{z}$, we augment the set of unknown parameters in the model with $\mathbf{z}$, which are successively re-inferred to maximise the (log) marginal likelihood of the data conditioned on recent observations. Once inference has taken place, at each step sequentially, we can make a prediction for the successor latent values and infer Equations 1 and

2. Noting that with this method the estimation of the hidden variables $\mathbf{z}$ is a distribution, and in order to infer estimations for the parameters of the model in Equations 3, 4, and 5, we sample from this distribution.

In the following section, we compare our model to a standard GP, the MDN and a mixture of experts model. The MDN model we compare with is the maximum-likelihood approach of Bishop (1994) in which estimates of the latent variables, $\mathbf{z}$, are made using a feed-forward neural network with a single hidden layer, in which we use radial basis functions (we refer to this model as RBFN). The mixture of experts model that we use for comparison is the "fast Bayesian Mixture Expert Model" (fBME) presented in Bo, L., et al. (2012).

3 Results

As a first step we test the model on synthetic data, where we know the solution. Following this we test the model on two real timeseries, the sunspot data, and the Mackey-Glass equation dataset. We further make a comparison of our results with the predictions estimated by standard GP, fBME, and RBFN models. For each time series we use the lagged data as input $\mathbf{x_t}$. The lag is set to 10 for all timeseries and we predict the next 10 data points at each time step with a rolling window. For all time series the number of mixture coefficients m, is set to 3.

Synthetic dataset: We construct a simple synthetic data, which provides regions of multi-model output density, as follows:

$$\mathbf{y_t} = \begin{cases} 0.3 + \epsilon, & p(y_t = 0.3) = \alpha_t \\ 0.7 + \epsilon, & p(y_t = 0.7) = 1 - \alpha_t \end{cases} \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, 0.01)$. The value of α_t changes gradually from 1.0 to 0.0 in a time span of 200 data points. We concatenate $\mathbf{y_t} = 0.5+\epsilon$ for 220 time steps with the above multimodal timeseries, to create the synthetic dataset which is initially unimodal in the target variable then evolves multimodality.

Figure 1 illustrates the predicted posterior distribution for the synthetic timeseries by each model. The heatmap in each subfigure represents the value of the posterior density. We note that our approach (top right panel) predicts the multimodal distribution most accurately, that the RBF model performs fairly well on this data and the standard GP (as a reference) is expected to perform poorly. The mixture of experts approach (fBME) does not manage to capture the underlying duality of the data in this instance.

Over the synthetic data set, as well as two candidate real-world data sets (the classic sunspot data timeseries, as well as a section of the Mackey-Glass chaotic timeseries, each consisting of 400 points) the approach we present here performs well, as measured using the negative log-likelihood of future data conditioned on the model. All methods capable of multimodel predictive posteriors are expected to outperform a standard GP on data where multiple future trajectories are possible. Table 1 illustrates the average errors. We note that our approach does considerably better than the other models on the synthetic data, better than all models on the Mackey-Glass timeseries and similarly to all models on the sunspot data.

Time series	Standard GP	fBME	RBFN	MDM
Synthetic data	10.49	17.60	-0.68	**-1.36**
Sunspot data set	-1.07	**-1.26**	-1.08	-1.10
Mackey-Glass	0.06	0.11	0.78	**0.04**

Table 1: Negative log-likelihood on out-of-sample future data. The lowest error model is shown in bold in each case.

4 Conclusion

We present a mixture density model Gaussian Process, capable of inferring flexible multimodal posterior distributions. We consider the performance of the model on three candidate data sets, one

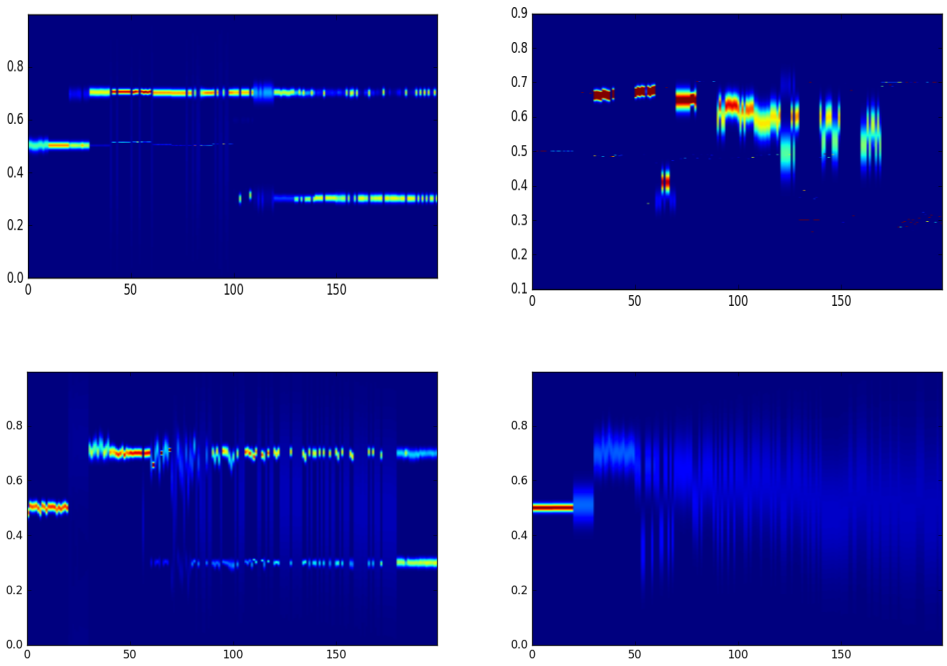

Figure 1: Synthetic data: Comparison of predicted posterior distribution by model. Top left - GP mixture density model. Top right - Mixture of experts. Bottom left - RBF mixture density model. Bottom right - Standard Gaussian Process.

synthetic and two real. In two of the three cases our approach significantly outperforms alternate models, including a standard GP. We see our method as useful in not only offering multimodal forecasts, but also in monitoring the complexity changes in such forecasts, which can be valuable in determining state changes and tipping points in complex systems.

References

[1] Christopher M. Bishop (1994) Mixture Density Networks, *Neural Computing Research Report*: NCRG/94/004.

[2] C. E. Rasmussen, Z. Ghahramani (2002) Infinite Mixtures of Gaussian Process Experts, *Advance in Neural Information Processing Systems*: 14.

[3] V. Tresp (2001) Mixture of Gaussian Process, *Advances in neural information processing systems*: 13.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, & G. E. Hinton (1991) Adaptive mixture of local experts, *Neural Computation*: vol 3.

[5] E. Meeds, S. Osindero (2006) An Alternative Infinite Mixture of Gaussian Process Experts, *Advances in Neural Information Processing Systems*: 18.

[6] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, & S. Aigrain (2012) Gaussian Processes for Timeseries Modelling, *Philosophical Transactions of the Royal Society (Part A)*: vol 371.

[7] Christopher M. Bishop (1995) Neural Networks for Pattern Recognition, *Oxford University Press.*

[8] L. Bo, C. Sminchisescu, A. Kanaujia & D. Metaxas (2008) Fast Algorithms for Large Scale Conditional 3D Prediction, *IEEE International Conference on Computer Vision and Pattern Recognition*.