Nonstationary Gaussian Process Regression using a Latent Extension of the Input Space[*]

Tobias Pfingsten, Malte Kuss and Carl Edward Rasmussen

Robert Bosch GmbH, Corporate Research and Advance Engineering, Stuttgart
Max Planck Institute for Biological Cybernetics, Tübingen
{Tobias.Pfingsten, Malte.Kuss, Carl}@tuebingen.mpg.de

Introduction Gaussian Processes (GPs) can be used to specify a prior over latent functions in non-parametric Bayesian models, e.g. for regression and classification. For this abstract we assume familiarity with the basic concepts of Gaussian Process models, see for example the introduction by Mackay [1]. A GP is defined by a mean and a covariance function, the latter describing dependencies $\hat{k}(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}'))$ between function values as a function of the corresponding inputs $\mathbf{x}$ and $\mathbf{x}'$. A common assumption when specifying a GP prior is stationarity, i.e. that the covariance between function values only depends on the distances $|\mathbf{x} - \mathbf{x}'|$, not on their location. It is far more difficult to specify a GP prior allowing the function to have different properties in different parts of the input space. In this work we describe new techniques for non-parametric Bayesian regression for, e.g. discontinuous, functions where the stationarity assumption does not hold.

Several approaches to the problem of how to specify nonstationary GP models can be found in the literature. Sampson and Guttorp [2] propose to use multidimensional scaling for spatio-temporal Processes to map a nonstationary spatial Process into a latent space in which the problem becomes approximately stationary. Schmidt and O'Hagan [3] pick up the idea and use GPs to implement the mapping. In comparison to a direct definition of a nonstationary covariance function, as proposed by [4], the detour via a latent space is advantageous because it assures positive definiteness of the covariance between observations in the original space and eases an intuitive interpretation of the problem.

In this work we propose to augment the input space $\mathbb{R}^D$ by a latent *extra input* which we infer from the data. When thinking of regression for discontinuous functions, the extra input could tear apart regions of the input space that are separated by abrupt changes of the function values. The idea to add an extra dimension to the input space is strongly related to the use of a so-called *Mixtures of Local Experts* (MoE) as described in [5, 6, 7, 8] where several independent GPs, so called experts, are used to explain the data in different regions of the input space. In this framework a *gating network* assigns responsibilities to certain experts, defining a mapping from the known inputs $\mathbf{x}$ to the class associations. We close the gap between a mixture of independent experts and a single GP using the fact that the latent associations to the experts can be seen as a discretized latent input. In the following we present two approaches for approximate Bayesian inference in GP models, that implement nonstationarity by an augmented input space. The first method is inspired by the MoE view with a discrete latent input and is implemented in an MCMC sampling scheme, whereas the second method estimates a continuous latent mapping by evidence maximization.

Nonstationarity by Augmentation Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ denote N training samples, where $y_i \in \mathbb{R}$ stand for a target and $\mathbf{x}_i \in \mathbb{R}^D$ is the corresponding D-dimensional input vector. The standard GP regression model assumes a relation $y_i = f(\mathbf{x}_i) + \varepsilon$ via a latent function f, where the observational noise is normally distributed. The key idea is to use a Gaussian Process prior on f and to make inference about the latent function directly. Below, all parameters of the covariance function and the likelihood are collected in a vector $\boldsymbol{\theta}$.

Assume for example we use the common quadratic exponential covariance function $\hat{k}(\mathbf{x}, \mathbf{x}') = v^2 \exp\{-\frac{1}{2} \sum_d w_d^{-2} (x_d - x'_d)^2\}$ and extend the inputs by a latent variable ℓ. The covariance function of a GP in this augmented input space $\bar{\mathbf{x}} = (\mathbf{x}^T, \ \ell)^T$ reads

$$k(\bar{\mathbf{x}}, \bar{\mathbf{x}}') = \hat{k}(\mathbf{x}, \mathbf{x}') \exp\Big(- \tfrac{1}{2} \big(\tfrac{\ell - \ell'}{w_o}\big)^2 \Big) \ . \tag{1}$$

To this point we have not specified any assumptions about ℓ. Considering data which is generated by more than one stationary Process, it is adequate to restrict ℓ to discrete values,

[*] Presented as a poster on the ISBA Eighth World Meeting on Bayesian Statistics in Valencia, 2006.

say $\{0,1\}$. While we retain the standard GP for $w_o \to \infty$, small w_o drive the correlation between function values to zero if $\ell \neq \ell'$. In this case we obtain the above above mentioned mixture of independent GPs where the joint covariance matrix of the training examples comes in the form of a block-diagonal matrix [9]. As we infer w_o from the data, also intermediate values are possible which correspond to a mixture of correlated GPs when ℓ is discrete. In the following, we compare two approaches for handling the latent variables. The first approach uses a discrete latent extra input and can be interpreted as a MoE where we use a tailored GP classifier to model the gating network. The second approach directly models ℓ as a continuous function of the inputs $\mathbf{x}$ using a sparse parametric GP regression model [10].

MoE approach with discrete latent input Using Bayes' rule one obtains the posterior distribution for the hyper-parameters $\boldsymbol{\theta}$ and latent inputs $\boldsymbol{\ell}$ of the training instances $p(\boldsymbol{\theta},\boldsymbol{\ell}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta},\boldsymbol{\ell})p(\boldsymbol{\theta})p(\boldsymbol{\ell})$, where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\ell})$ are prior distributions. The likelihood of the hyper-parameters $\boldsymbol{\theta}$ and latent inputs $\boldsymbol{\ell}$, called assignments in the MoE interpretation, is simply the marginal likelihood $p(\mathcal{D}|\boldsymbol{\theta},\boldsymbol{\ell})$. As inference in this model is analytically intractable we resort to an MCMC technique, alternating a Gibbs scheme to draw samples $\boldsymbol{\ell}_n$ of the discrete extra input and hybrid MC to draw samples $\boldsymbol{\theta}_n$ from the posterior distribution $p(\boldsymbol{\theta},\boldsymbol{\ell}|\mathcal{D})$. We use a uniform prior on $\boldsymbol{\ell}$ which is updated using the data $\mathcal{D}$, which means for prediction at $\mathbf{x}^*$ that we have to add up predictions for all possible ℓ^* with equal weight. In order to improve predictions we need to define a gating network $p(\ell^*|\mathcal{D},\mathbf{x}^*)$ which models a dependency of the position $\mathbf{x}^*$. The predictive distribution at points $\mathbf{x}^*$ is given by an average over the unknown ℓ^*,

$$p(f^*|\mathcal{D},\mathbf{x}^*) = \int \mathrm{d}\boldsymbol{\theta} \sum_{\boldsymbol{\ell}} p(\boldsymbol{\theta},\boldsymbol{\ell}|\mathcal{D}) \left(\sum_{\ell^*} p(f^*|\mathcal{D},\mathbf{x}^*,\boldsymbol{\theta},\boldsymbol{\ell},\ell^*)p(\ell^*|\mathcal{D},\mathbf{x}^*,\boldsymbol{\theta},\boldsymbol{\ell}) \right) , \qquad (2)$$

where we identify $p(\ell^*|\mathcal{D},\mathbf{x}^*,\boldsymbol{\theta},\boldsymbol{\ell})$ as the gating network. While a standard classifier only models $p(\ell^*|\mathbf{X},\boldsymbol{\ell},\mathbf{x}^*)$, instead we construct a model for $p(\boldsymbol{\ell}|\mathcal{D})$ at test inputs.

We tailor the gating network to our needs based on a GP classifier [11] with latent function ξ, which assumes some likelihood $p_\ell\left(\ell|\xi,\mathbf{x}\right)$. In our setup we aim at using $\pi_k = p(\ell_k = 1|\mathcal{D},\boldsymbol{\theta})$ as extra information from the regression setup and construct a likelihood $p(\pi_k|\xi,\mathbf{x}_k,\mathcal{D},\boldsymbol{\theta})$ to take this information into account:

$$p\left(\pi_k\,|\xi,\mathbf{x}_k,\mathcal{D},\boldsymbol{\theta}\right) \;\propto\; \pi_k\, p_\ell\left(\ell=1\,|\xi,\mathbf{x}_k\right) + (1-\pi_k)\, p_\ell\left(\ell=0\,|\xi,\mathbf{x}_k\right) \; . \qquad (3)$$

This model shows sensible behavior in both limiting cases: If the MoE clearly assigns a training case $(\mathbf{x}_k, y_k)$ to one of the experts, i.e. $\pi_k \approx 0$ or 1, the sum in (3) collapses and coincides with the likelihood p_ℓ. If, on the other hand, $\pi_k \approx \frac{1}{2}$ we obtain $p\left(\pi_k|\xi,\mathbf{x}_k,\mathcal{D},\boldsymbol{\theta}\right) = \frac{1}{2}$ and the classifier effectively ignores the data point as the regression model does not determine the association ℓ_k.

Direct approach with continuous latent inputs The MoE approach restricts the latent input to discrete values and uses a gating network to model a mapping $p(\ell|\mathbf{x},\mathcal{D})$ from the input space to the latent labels. If, on the other hand, we assume ℓ to be a continuous parametric function of the inputs $\mathbf{x}$ we can see its parameters $\boldsymbol{\theta}_l$ as additional parameters of the covariance function (1) itself and treat them just as the other parameters $\boldsymbol{\theta}$ in a hybrid MC or the computationally cheaper evidence maximization scheme. To obtain a flexible parametric function we use *sparse parametric GPs*, which have recently been introduced by Snelson and Ghahramani [10]. The sparse GPs comprise the flexibility of a full GP while reducing computational complexity through a reduction of the number of support points which are treated as free parameters. The predictive mean $\mu^P(\mathbf{x})$ is a parametric function of these "pseudo" inputs and targets, along with the parameters of the used covariance function. We model the latent extra input $l(\mathbf{x}|\boldsymbol{\theta}_l)$ using the predictive mean of such a sparse GP which we map to $[0,1]$ defining $l(\mathbf{x}|\boldsymbol{\theta}_l) = [1+\exp\{\mu^P(\mathbf{x})\}]^{-1}$. In comparison to the discrete, MoE type latent input, the parametric representation has the virtue that we can spare a gating network and treat the mapping as a part of the covariance function. In principle any parametric function could be used to model ℓ, however, using the flexible sparse GPs we ensure that the model can capture the underlying structure of the data.

Experiments We evaluate the performance of our algorithms using two 2-, and two 4-dimensional data sets which stem from simulations of an inverted pendulum system as it is frequently studied in control theory and reinforcement learning. The problem is to swing up a pendulum attached to a cart which can only be accelerated horizontally. The objective is to swing up the pendulum and balance it while the cart itself is centered. As loss function we use the squared distance between the actual and the target position of the pendulum. The

MSE	ν-SVR	stat. GP	MoE	param. latent
Force 2D	2.619	2.433	2.233	2.120
Loss 2D	0.174	0.107	0.094	0.036
Force 4D	7.701	5.534	5.147	4.246
Loss 4D	9.109	8.360	8.370	3.485

Table 1: Predictive performances, Mean Square Error (MSE)

controller behaves very differently depending on whether the state of the system allows the controller to balance the pendulum. In this balancing region the loss is relatively small, outside the controller applies larger forces and the variability of the losses increases. The controller has a stochastic element that can lead to irregularities in the forces and the loss function has a discontinuity marking the border of the balancing region. We learn the control signal (force) and the corresponding loss accumulated over one second (loss) as functions of the pendulum angle, the cart's position and the corresponding velocities. The two-dimensional data sets were created by setting the velocities to zero.

For all data sets we did a 10 fold cross validation and show the average mean square error in Table 1. For brevity we omit the predictive probabilities. We compare the performance of both proposed methods, the MCMC implementation of the discrete ”MoE” latent input, and the approach of a continuous ”parametric latent” input (evidence maximization), to a stationary GP (MCMC) and the widely used ν-SVR [12] regression as a benchmark.

The experiments show that the extra flexibility introduced by the latent space greatly improves predictions in both implementations, where the parametric approach—while being computationally much more efficient—clearly outperforms the MoE scheme. The implementations differ substantially, both showing benefits and disadvantages. The discrete latent space ”MoE” approach proves to be very efficient in approximating the posterior distribution $p(\boldsymbol{\ell}, \boldsymbol{\theta}|\mathcal{D})$, while crux is a good gating network to correctly classify the test instances. The parametric approach avoids this problem by directly maximizing the evidence of the regression problem to find a mapping to the latent input. The optimization problem, however, is non-convex and it can be hard to find a global maximum.

Synopsis In this work we describe the construction of nonstationary GP models for regression. An attractive way to model nonstationarity is to use a mapping to a latent space where the problem appears approximately stationary. We propose two alternative models which implement the mapping via an augmentation of the input space by a latent extra input and illustrate how approximate Bayesian inference can be implemented for those models. The first approach assumes the latent inputs to be discrete variables and we show that this is effectively an extension to the known MoE approach. The second approach assumes continuous latent variables which we implement using the flexible sparse parametric Gaussian Processes. In various experiments we find the proposed models to give significantly improved performance.

References

[1] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. CUP, 2003.

[2] P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87:108–119, 1992.

[3] A.M. Schmidt and A. O'Hagan. Bayesian inference for nonstationary spatial covariance structure via spatial deformations. *JRSS, series B*, 65:745–758, 2003.

[4] C.J. Paciorek and M.J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *NIPS 16*. MIT Press, 2004.

[5] R.A. Jacobs, M. I. Jordan, S. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:1–12, 1991.

[6] V. Tresp. Mixtures of Gaussian processes. In *NIPS 13*. MIT Press, 2000.

[7] C.E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *NIPS 14*. MIT Press, 2002.

[8] J.Q. Shi, R. Murray-Smith, and D.M. Titterington. Hierarchical Gaussian process mixtures for regression. *Statistics and Computing*, 15(1):31–41, 2005.

[9] Z. Ghahramani. Personal communication, 2004.

[10] E. Snelson and Z. Ghahramani. Sparse parametric gaussian processes. In *NIPS 18*, 2005.

[11] C.K.I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

[12] B. Schölkopf, P.L. Bartlett, A.J. Smola, and R. Williamson. Shrinking the tube: a new support vector regression algorithm. In *NIPS 11*. MIT Press, 1999.