Non-Stationary Gaussian Process Regression with Hamiltonian Monte Carlo

Markus Heinonen[1,2] **Henrik Mannerström**[1] **Juho Rousu**[1,2] **Samuel Kaski**[1,2] **Harri Lähdesmäki**[1]

[1] Department of Computer Science, Aalto University, Finland
[2] Helsinki Institute for Information Technology HIIT, Finland

Abstract

We present a novel approach for non-stationary Gaussian process regression (GPR), where the three key parameters – noise variance, signal variance and lengthscale – can be simultaneously input-dependent. We develop gradient-based inference methods to learn the unknown function and the non-stationary model parameters, without requiring any model approximations. For inferring the full posterior distribution we use Hamiltonian Monte Carlo (HMC), which conveniently extends the analytical gradient-based GPR learning by guiding the sampling with the gradients. The MAP solution can also be learned with gradient ascent. In experiments on several synthetic datasets and in modelling of temporal gene expression, the non-stationary GPR is shown to give major improvement when modeling realistic input-dependent dynamics.

1 Introduction

Gaussian process regression has emerged as a powerful, yet practical class of non-parametric Bayesian models that quantify the uncertainties of the underlying process using Gaussian distributions (Rasmussen and Williams, 2006). Gaussian processes are commonly applied to time-series interpolation, regression and classification, where the GP can provide predictive distributions (Rasmussen and Williams, 2006).

The standard GP model assumes that the model parameters stay constant over the input space. This includes the observational noise variance ω^2, as well as the signal variance σ^2 and the lengthscale ℓ of the covariance function. The signal variance determines the signal amplitude, while the characteristic lengthscale defines the local 'support' neighborhood of the function. In many real world problems either the noise variance or the signal smoothness, or both, vary over the input space, implying a *heteroscedastic* noise model or *non-stationary* function dynamics, respectively (Le et al., 2005; Wang and Neal, 2012). In both cases, the analytical posterior of the GP becomes intractable (Tolvanen et al., 2014). For instance, in biological studies, rapid signal changes are often observed quickly after perturbations, with the signal becoming smoother in time (Heinonen et al., 2015).

Non-stationary models have been introduced from several perspectives. The treed GP model contains multiple piecewise GPs of varying covariances (Gramacy, 2005). Several authors have proposed transforming or warping the input space to achieve effective non-stationarity (Sampson and Guttorp, 1992; Schmidt and O'Hagan, 2003). Snoek et al. (2014) infer parametric warpings for multi-task GPs. In spatial statistics, spatial warpings have been extensively studied (Anderes and Stein, 2008). Another approach is to model the temporal evolution of the GP covariance matrix directly with generalised Wishart processes (Wilson and Ghahramani, 2011).

Several authors have proposed extending GPs directly with input-dependent parameters. These latent parameters are treated as separate Gaussian processes and inferred jointly with the unknown function (Tolvanen et al., 2014). In a heteroscedastic noise GPs, a latent noise variance is inferred in a maximum likelihood (ML) (Kersting et al., 2007) or maximum a posteriori (MAP) fashion (Quadrianto et al., 2009). Fully Bayesian inference methods include MCMC sampling (Goldberg et al., 1997) and variational and expectation propagation approximations of the posterior (Lazaro-Gredilla and Titsias, 2011; Tolvanen et al., 2014). Non-stationarities can also be included in the signal variance or lengthscale by the use of non-stationary variants of kernel functions (Gibbs, 1997). Non-stationary lengthscales for Gaussian processes were introduced by Gibbs (1997) and further extended by Paciorek and Schervish (2004) with MCMC inference. Recently, Tolvanen et al. (2014) introduced a non-stationary signal variance using expectation

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

propagation and approximate variational inference.

In this paper we introduce the first non-stationary and heteroscedastic GP regression framework, in which the three main components (noise variance, signal variance and the lengthscale) can be simultaneously input-dependent, with direct GP priors. We propose an inference method for the exact joint posterior of the underlying signal, under the Gaussian likelihood, and all three latent functions, avoiding the need for introducing variational or expectation propagation approximations (Lazaro-Gredilla and Titsias, 2011; Tolvanen et al., 2014). We use HMC-NUTS, which can effectively sample the posterior guided by the analytical model gradients. Furthermore, an exact MAP solution arises as a simple gradient ascent on the posterior. We dramatically improve the performance of both approaches by posterior whitening using Cholesky decompositions of the latent function priors. Our experiments demonstrate the necessity of non-stationary GPR to model realistic input-dependent dynamics, while in simpler conditions the proposed method performs comparably to conventional stationary or previous non-stationary GPR models.

In Section 2 we introduce the non-stationary GP model. In its subsections we first introduce MAP and HMC inference, discuss model whitening and finally define the predictive distributions. Section 3 presents experimental results on several synthetic and one real biological datasets, and we conclude in Section 4. The implementation is available at `github.com/markusheinonen/adaptivegp`.

2 Heteroscedastic non-stationary GP model

Let $\mathbf{y} = (y_i)_{i=1}^n \in \mathbb{R}^n$ be an observation vector over n inputs $\mathbf{x} = (x_i)_{i=1}^n \in \mathbb{R}^n$. We assume an additive regression model,

$$y(x) = f(x) + \varepsilon(x), \quad \varepsilon(x) \sim \mathcal{N}(0, \omega(x)^2),$$

where both the underlying signal $f(x)$ and the zero-mean observation noise variance $\omega(x)^2$ are unknown functions to be learned[1]. We proceed by first placing a zero mean GP prior on the unknown function $f(x)$,

$$f(x) \sim GP(0, K_f(x, x')), \tag{1}$$

which assumes that $\mathbf{cov}(f(x), f(x')) = K_f(x, x')$. We use a non-stationary generalisation of the squared exponential kernel (Gibbs, 1997),

$$K_f(x, x') = \sigma(x)\sigma(x')\sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}} \times \exp\left(-\frac{(x - x')^2}{\ell(x)^2 + \ell(x')^2}\right), \tag{2}$$

where $x, x' \in \mathbb{R}$, and $\sigma(x)$ and $\ell(x)$ are input-dependent signal variance and lengthscale functions, respectively. The kernel reduces into a standard squared exponential kernel if both are constant. We show the kernel (2) is positive definite in the Supplementary Material.

We model the lengthscale, signal variance and noise variance with *latent functions*. We are interested in smoothly varying latent functions and thus we place separate GP priors on them:

$$\log(\ell(t)) \equiv \tilde{\ell}(t) \sim GP(\mu_\ell, K_\ell(x, x'))$$
$$\log(\sigma(t)) \equiv \tilde{\sigma}(t) \sim GP(\mu_\sigma, K_\sigma(x, x'))$$
$$\log(\omega(t)) \equiv \tilde{\omega}(t) \sim GP(\mu_\omega, K_\omega(x, x')),$$

where we set the priors on the logarithms to ensure their positivity. We select separate standard squared exponential covariances for each,

$$K_c(x, x') = \alpha_c^2 \exp\left(-\frac{(x - x')^2}{2\beta_c^2}\right),$$

where $c \in \{\ell, \sigma, \omega\}$. The model has nine hyper-parameters $\boldsymbol{\theta} = (\mu_\ell, \mu_\sigma, \mu_\omega, \alpha_\ell, \alpha_\sigma, \alpha_\omega, \beta_\ell, \beta_\sigma, \beta_\omega)$ that define the prior for the three latent functions $\tilde{\ell}$, $\tilde{\sigma}$ and $\tilde{\omega}$. The means μ determine latent function means, while the α's are scaling terms. The β's are the characteristic lengthscales of the priors. In practice, the μ's and α's have a small effect on the models, whereas the β's have a large effect on the model by determining the smoothness of the latent functions. They can be set based on prior knowledge or using grid-search over suitable values.

Given a dataset $(\mathbf{x}, \mathbf{y})$, the model can equivalently be written as $\mathbf{f}|\boldsymbol{\ell}, \boldsymbol{\sigma} \sim \mathcal{N}(\mathbf{0}, K_f)$, where $\mathbf{f} = (f(x_i))_{i=1}^n$ is a latent function vector at the observed points $\mathbf{x}$ and $K_f \in \mathbb{R}^{n \times n}$ has elements $[K_f]_{ij} = K_f(x_i, x_j)$ computed using eq. (2) with signal standard deviations $\boldsymbol{\sigma} = (\sigma(x_i))_{i=1}^n$ and lengthscales $\boldsymbol{\ell} = (\ell(x_i))_{i=1}^n$. Finally, the data likelihood is $\mathbf{y}|\boldsymbol{\ell}, \boldsymbol{\sigma}, \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, K_f + \Omega)$, where $\Omega = \operatorname{diag} \boldsymbol{\omega}^2 \in \mathbb{R}^{n \times n}$ is a diagonal noise matrix and $\boldsymbol{\omega}^2 = (\omega(x_i)^2)_{i=1}^n$ are the noise variances. We note that no current method has studied a non-stationary parameterisation with all three input-dependent parameters.

To infer latent functions from the *full posterior* $p(\mathbf{f}, \tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y}, \boldsymbol{\theta})$ we introduce two approaches in the next two Sections[2]. We propose to learn the MAP estimate $p(\mathbf{f}|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\omega}}_{\text{MAP}}, \mathbf{y})$, or infer the full posterior using HMC sampling. Both approaches are based on the analytical gradients of the latent functions.

2.1 Maximum a posteriori estimation

As the first approach, we follow the approaches by Kersting et al. (2007) and Quadrianto et al. (2009), and resort to find-

[1] We assume univariate inputs throughout this paper. See Supplementary Material for a generalisation into multivariate inputs.

[2] In the following we omit the hyperparameters $\boldsymbol{\theta}$ for notational clarity

ing the MAP solution of the *latent posterior* $p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y})$,

$$\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\omega}}_{\text{MAP}} = \arg\max_{\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}} p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y}),$$

where $\mathbf{f}$ has been marginalised out. Using Bayes' theorem this is equivalent to maximizing the marginal likelihood

$$\mathcal{L} = p(\mathbf{y}|\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}) p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}), \tag{3}$$

which evaluates to

$$\mathcal{N}(\mathbf{y}|\mathbf{0}, K_f + \Omega)\mathcal{N}(\tilde{\boldsymbol{\ell}}|\mu_\ell, K_\ell)\mathcal{N}(\tilde{\boldsymbol{\sigma}}|\mu_\sigma, K_\sigma)\mathcal{N}(\tilde{\boldsymbol{\omega}}|\mu_\omega, K_\omega),$$

and whose logarithm we denote as the marginal log likelihood (MLL).

The partial derivatives of the log of marginal likelihood (3) with respect to the latent functions are analytical:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\ell}}_i} &= \frac{1}{2} \operatorname{tr}\left((\mathbf{a}\mathbf{a}^T - K_y^{-1}) \frac{\partial K_y}{\partial \tilde{\boldsymbol{\ell}}_i} \right) - [K_{\tilde{\ell}}^{-1}(\tilde{\boldsymbol{\ell}} - \mu_{\tilde{\ell}})]_i \\
\frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\sigma}}} &= \operatorname{diag}\left((\mathbf{a}\mathbf{a}^T - K_y^{-1}) K_f \right) - K_{\tilde{\sigma}}^{-1}(\tilde{\boldsymbol{\sigma}} - \mu_{\tilde{\sigma}}) \\
\frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\omega}}} &= \operatorname{diag}\left((\mathbf{a}\mathbf{a}^T - K_y^{-1}) \Omega \right) - K_{\tilde{\omega}}^{-1}(\tilde{\boldsymbol{\omega}} - \mu_{\tilde{\omega}})
\end{aligned} \tag{4}$$

where $\mathbf{a} = (K_f + \Omega)^{-1}\mathbf{y}$ and $\frac{\partial K_y}{\partial \tilde{\boldsymbol{\ell}}_i}$ is given in the Supplementary Material.

We perform gradient ascent over the MLL, $\log \mathcal{L}$. The solution is only guaranteed to converge to a local optimum, and hence we perform multiple restarts from random initial conditions. The MAP solution is adequate when the posterior is close to unimodal.

Given the MAP solution, the function posterior $p(\mathbf{f}|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\omega}}_{\text{MAP}}) \sim \mathcal{N}(\mathbf{m}_{\text{MAP}}, \Sigma_{\text{MAP}})$ is a Gaussian with

$$\begin{aligned}
\mathbf{m}_{\text{MAP}} &= K_f^T (K_f + \Omega_{\text{MAP}})^{-1} \mathbf{y} \\
\Sigma_{\text{MAP}} &= K_f - K_f^T (K_f + \Omega_{\text{MAP}})^{-1} K_f,
\end{aligned}$$

where K_f has been computed with eq. (2) using MAP latent vectors $\log(\boldsymbol{\ell}) = \tilde{\boldsymbol{\ell}}_{\text{MAP}}$ and $\log(\boldsymbol{\sigma}) = \tilde{\boldsymbol{\sigma}}_{\text{MAP}}$, and Ω_{MAP} with $\log(\boldsymbol{\omega}) = \tilde{\boldsymbol{\omega}}_{\text{MAP}}$.

2.2 HMC inference

As a second approach we sample the latent posterior $p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y})$ using Hamiltonian Monte Carlo (HMC) (Hoffman and Gelman, 2014; Neal, 2011). In HMC, an additional momentum variable is introduced for each of the model variables, and the extended model is interpreted as a Hamiltonian system. Time evolution of the Hamiltonian dynamics is simulated to produce proposals for the Metropolis algorithm. The latent posterior $p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y})$ is proportional to the marginal likelihood in eq. (3), and thus the HMC sampling of $(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}})$ uses the same gradients $\left(\frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\ell}}}, \frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\sigma}}}, \frac{\partial \log \mathcal{L}}{\partial \tilde{\boldsymbol{\omega}}} \right)$ from eq. (4) as the MAP solution. Thus, we only need to do HMC sampling over the three latent vectors $(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}})$ and the posterior of $\mathbf{f}$ for each sample follows analytically as a Gaussian, leading to a mixture of m Gaussians.

The function posterior $p(\mathbf{f}|\mathbf{y})$ can then be approximated with the HMC samples

$$\begin{aligned}
p(\mathbf{f}|\mathbf{y}) &= \iiint p(\mathbf{f}|\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}, \mathbf{y}) p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y}) d\tilde{\boldsymbol{\ell}} d\tilde{\boldsymbol{\sigma}} d\tilde{\boldsymbol{\omega}} \\
&\approx \frac{1}{m} \sum_{i=1}^{m} p(\mathbf{f}|\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\omega}}_i, \mathbf{y}),
\end{aligned} \tag{5}$$

where

$$\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\omega}}_i \sim p(\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}|\mathbf{y}) \tag{6}$$

are m HMC samples of the latent posterior. The function posterior $p(\mathbf{f}|\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\omega}}_i, \mathbf{y}) = \mathcal{N}(\mathbf{m}_i, \Sigma_i)$ for each HMC sample is a Gaussian with

$$\begin{aligned}
\mathbf{m}_i &= K_{f_i}^T (K_{f_i} + \Omega_i)^{-1} \mathbf{y} \\
\Sigma_i &= K_{f_i} - K_{f_i}^T (K_{f_i} + \Omega_i)^{-1} K_{f_i},
\end{aligned}$$

where K_{f_i} is a non-stationary kernel matrix computed using $\tilde{\boldsymbol{\ell}}_i$ and $\tilde{\boldsymbol{\sigma}}_i$, and Ω_i is the diagonal noise covariance matrix of $\tilde{\boldsymbol{\omega}}_i$.

2.3 Posterior whitening

The posterior of the latent vectors is, by definition, highly correlated due to Gaussian priors, leading to inefficient Monte Carlo sampling. To ease the sampling, we perform the sampling over the whitened latent vectors (Kuss and Rasmussen, 2005),

$$\begin{aligned}
\mathring{\boldsymbol{\ell}} &= L_\ell^{-1} \tilde{\boldsymbol{\ell}}, \quad K_\ell = L_\ell L_\ell^T \\
\mathring{\boldsymbol{\sigma}} &= L_\sigma^{-1} \tilde{\boldsymbol{\sigma}}, \quad K_\sigma = L_\sigma L_\sigma^T \\
\mathring{\boldsymbol{\omega}} &= L_\omega^{-1} \tilde{\boldsymbol{\omega}}, \quad K_\omega = L_\omega L_\omega^T,
\end{aligned}$$

with Cholesky decompositions of the corresponding GP prior covariances, which are fixed based on the hyperparameters $\boldsymbol{\theta}$. The derivatives of the MLL with respect to the whitened parameters can be retrieved analytically. E.g. the lengthscale becomes $\frac{\partial \log \mathcal{L}}{\partial \mathring{\boldsymbol{\ell}}} = \frac{\partial \log \mathcal{L}}{\partial L_\ell \mathring{\boldsymbol{\ell}}} \frac{\partial L_\ell \mathring{\boldsymbol{\ell}}}{\partial \mathring{\boldsymbol{\ell}}} = L_\ell^T \nabla_{\tilde{\boldsymbol{\ell}}} \mathcal{L}$, where the last term is the standard gradient of the non-whitened model defined in eq. (4). The two other parameters follow the same procedure. In practice the whitening leads to several orders of magnitude improvement on inference speed.

2.4 Making predictions

Both the MAP solution and the HMC sampler infer values of the latent functions only at the n observed inputs

$\mathbf{x}$. To extrapolate the values of the unknown function and the latent functions over arbitrary target points $\mathbf{x}_\star \in \mathbb{R}^{n_\star}$, we approximate the *predictive distribution* (Goldberg et al., 1997) by extrapolating the latent functions $\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\omega}}$ to $\tilde{\boldsymbol{\ell}}_\star, \tilde{\boldsymbol{\sigma}}_\star, \tilde{\boldsymbol{\omega}}_\star$ independently of the data $\mathbf{y}$, and then express the function posterior $\mathbf{f}_\star$ with them. That is, we approximate $p(\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \mathbf{y})$ by $p(\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}})$ and $p(\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \mathbf{y})$ by $p(\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}_{\text{MAP}})$, which have analytical forms. With the MAP solution we have

$$
\begin{aligned}
&p(\mathbf{f}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\omega}}_{\text{MAP}}, \mathbf{y}) \qquad (7)\\
&\approx \iint p(\mathbf{f}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\ell}}_\star, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_\star, \tilde{\boldsymbol{\omega}}_{\text{MAP}}, \mathbf{y})\\
&\qquad \times p(\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}) p(\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}_{\text{MAP}}) d\tilde{\boldsymbol{\ell}}_\star d\tilde{\boldsymbol{\sigma}}_\star\\
&\approx \frac{1}{s}\sum_{j=1}^{s} p(\mathbf{f}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\ell}}_{j_\star}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{j_\star}, \tilde{\boldsymbol{\omega}}_{\text{MAP}}, \mathbf{y})
\end{aligned}
$$

where we approximate the integral by drawing s samples $\{\tilde{\boldsymbol{\ell}}_{j_\star}\}_{j=1}^s, \{\tilde{\boldsymbol{\sigma}}_{j_\star}\}_{j=1}^s$ of $n_\star$ dimensions from the conditional Gaussians $\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}$ and $\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}_{\text{MAP}}$ (See Supplementary Material). This results in a mixture of s corresponding Gaussians $\mathcal{N}(\mathbf{m}_{\text{MAP},j_\star}, \Sigma_{\text{MAP},j_\star})$, where

$$
\begin{aligned}
\mathbf{m}_{\text{MAP},j_\star} &= K_{\text{MAP},j_\star}^T (K_{\text{MAP,MAP}} + \Omega_{\text{MAP}})^{-1}\mathbf{y}\\
\Sigma_{\text{MAP},j_\star} &= K_{j_\star,j_\star} - K_{\text{MAP},j_\star}^T (K_{\text{MAP,MAP}} + \Omega_{\text{MAP}})^{-1} K_{\text{MAP},j_\star},
\end{aligned}
$$

and where $K_{j_\star,j_\star} \in \mathbb{R}^{n_\star \times n_\star}$, $K_{\text{MAP},j_\star} \in \mathbb{R}^{n \times n_\star}$ and $K_{\text{MAP,MAP}} \in \mathbb{R}^{n \times n}$ are computed with eq. (2) over the latent vectors $(\tilde{\boldsymbol{\ell}}_{\text{MAP}}, \tilde{\boldsymbol{\sigma}}_{\text{MAP}})$ over inputs $\mathbf{x}$, or using $(\tilde{\boldsymbol{\ell}}_{j_\star}, \tilde{\boldsymbol{\sigma}}_{j_\star})$ over inputs $\mathbf{x}_\star$. The simplest approximation is to denote the conditional means $\tilde{\boldsymbol{\ell}}_{j_\star} = \mathbb{E}[\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}_{\text{MAP}}]$ and $\tilde{\boldsymbol{\sigma}}_{j_\star} = \mathbb{E}[\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}_{\text{MAP}}]$ as the sole samples with $s = 1$. This is a sufficient approximation if the inputs $\mathbf{x}$ are sufficiently dense.

The predictive distribution given the HMC sample $\{\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\omega}}_i\}$ is derived analogously. We average over the m HMC samples instead of a single MAP solution, and over the s samples $\{\tilde{\boldsymbol{\ell}}_{ij_\star}\}_{j=1}^s$ and $\{\tilde{\boldsymbol{\sigma}}_{ij_\star}\}_{j=1}^s$ from the conditionals, resulting in

$$
\begin{aligned}
&p(\mathbf{f}_\star|\mathbf{y}) \approx p(\mathbf{f}_\star|\{\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\omega}}_i\}, \mathbf{y}) \qquad (8)\\
&\approx \frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s} p(\mathbf{f}_\star|\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\ell}}_{ij_\star}, \tilde{\boldsymbol{\sigma}}_i, \tilde{\boldsymbol{\ell}}_{ij_\star}, \tilde{\boldsymbol{\omega}}_i, \mathbf{y})\\
&\approx \frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s} \mathcal{N}(\mathbf{m}_{i,j_\star}, \Sigma_{i,j_\star})
\end{aligned}
$$

where $\mathbf{m}_{i,j_\star} = K_{i,j_\star}^T (K_i + \Omega_i)^{-1}\mathbf{y}$ and $\Sigma_{i,j_\star} = K_{j_\star,j_\star} - K_{i,j_\star}^T (K_i + \Omega_i)^{-1} K_{i,j_\star}$, and where the kernel matrices are computed using $\tilde{\boldsymbol{\ell}}_i, \tilde{\boldsymbol{\sigma}}_i$ and $\tilde{\boldsymbol{\ell}}_{ij_\star}, \tilde{\boldsymbol{\sigma}}_{ij_\star}$.

We note that a slower but perhaps more elegant alternative is to model latent functions jointly over concatenated inputs $\mathbf{x}_t \equiv (\mathbf{x}, \mathbf{x}_\star)$, resulting in $\boldsymbol{\ell}_t \equiv (\boldsymbol{\ell}, \boldsymbol{\ell}_\star)$, and analogously for the other functions. In this case the function posterior contains the predictive posterior with the approximation used in eq. (7), but the latent vector sizes increase to $n + n_\star$.

Table 1: Datasets with varying forms of non-stationarities. The column 'n' defines the total number of data points and the column 'n_{train}' the number of training points.

Dataset	Non-stationary functions	n	n_{train}
$\mathbf{D}_\sigma$	$\sigma(t)$	100	50
$\mathbf{D}_\ell$	$\ell(t)$	150	75
$\mathbf{D}_{\sigma,\omega}$	$\sigma(t), \omega(t)$	100	50
$\mathbf{D}_{\ell,\omega}$	$\ell(t), \omega(t)$	150	75
$\mathbf{D}_{\ell,\sigma,\omega}$	$\ell(t), \sigma(t), \omega(t)$	90	45
$\mathbf{M}_\omega$	$\omega(t)$	133	67
$\mathbf{T}_{\sigma,\omega}$	$\omega(t), \sigma(t)$	500	250
$\mathbf{J}$	N/A	101	50

3 Experiments

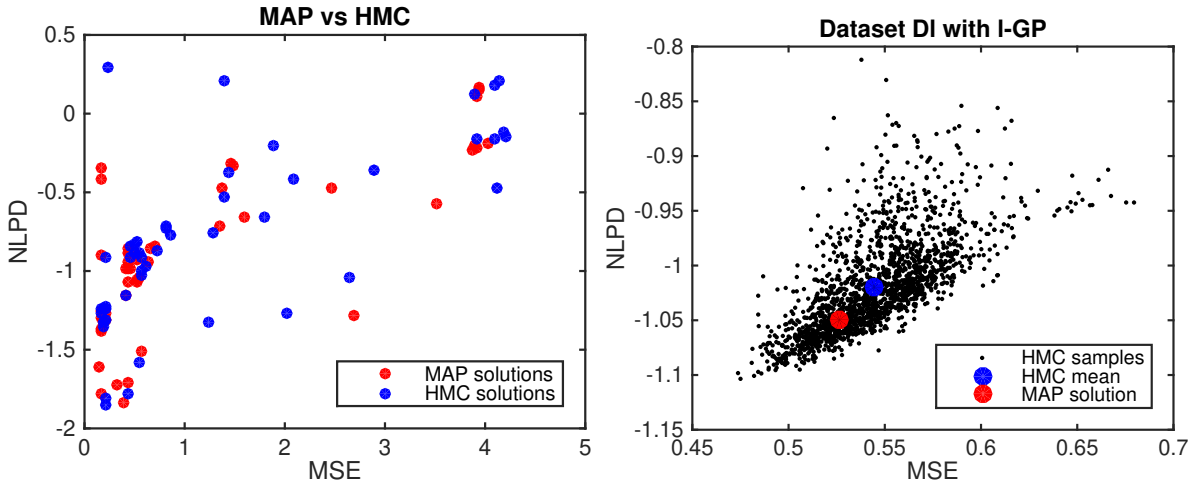
We assess the performance of the proposed method on several synthetic and real datasets. We experiment with five simulated datasets, four empirical datasets and a gene expression time series dataset (Heinonen et al., 2015). The empirical datasets contain the SP500 index **S**, the motorcycle dataset **M** (Silverman, 1985), the 3rd 'jump' dataset **J** from Paciorek and Schervish (2004) and a non-stationary dataset **T** from GPstuff `demo_epinf` (Vanhatalo et al., 2013). The five additional simulated datasets were generated with different combinations of non-stationarities (See Table 1). We expect datasets exhibiting specific types of input-dependent characteristics to require a model with a corresponding input-dependent parameter.

We scale all outputs to range $[-1, 1]$ and the inputs to range $[0, 1]$. For each dataset, we use half of the data as training data and the rest as test data. We will assess the performance on the test data with mean squared error MSE $= \frac{1}{n_{\text{test}}}\sum_i (y_i^{test} - [\mathbf{m}_\star]_i)^2$ and with the mean log-predictive density NLPD $= -\sum_i \log p(y_i^{test}|[\mathbf{m}_\star]_i, [\Sigma_\star]_{ii})$, where smaller value is better. For consistency, we model the stationary parameters as vectors $c\mathbf{1}$ of length n_{train} for $c = \{\omega, \sigma, \ell\}$ whenever a parameter is not set as non-stationary.

We run MAP optimisation from 10 different initial conditions and choose the one with the highest MLL value. We run 10 chains of 1000 samples of HMC-NUTS sampling using model whitening (Algorithm 3 of Hoffman and Gelman (2014), $\epsilon = 0.01$, maximum tree depth 10). For our datasets setting $s = 1$ with the conditional means was sufficient for obtaining an accurate predictive posteriors. We define all hyperparameters μ to be the mean of their corresponding parameter functions. We set all hyperparameters α to a large value of 1, which allows high freedom in the range of the corresponding parameters. We empirically select the hyperparameters β from a set $\{0.05, 0.1, 0.2\}$, and fixed $\beta_\ell = \beta_\sigma = 0.1$ and $\beta_\omega = 0.2$ throughout the experiments, which gave good results on all datasets. The MAP inference is approximately as fast as vanilla GP re-

Table 2: Test MSE and NLPD results on the synthetic datasets over various MAP models. Optimal values are in boldface. The optimal or second to optimal NLPD values follow a diagonal line. Smaller value is better for both quantities.

	$\mathbf{M}_\omega$		$\mathbf{D}_\sigma$		$\mathbf{D}_\ell$		$\mathbf{D}_{\omega,\sigma}$		$\mathbf{D}_{\omega,\ell}$		$\mathbf{D}_{\omega,\sigma,\ell}$		$\mathbf{J}$		$\mathbf{T}_{\omega,\sigma}$	
Method	MSE	NLPD	MSE	NLPD	MSE	NLPD	MSE	NLPD	MSE	NLPD	MSE	NLPD	MSE	NLPD	MSE	NLPD
GP	3.91	0.11	0.21	−1.31	0.71	−0.85	0.44	−0.86	0.54	−1.04	0.16	−1.25	1.48	−0.32	17.83	0.10
(ω)-GP	3.91	**−0.22**	0.21	−1.27	2.46	−0.47	0.45	−0.98	2.69	−1.28	0.33	−1.73	3.51	−0.58	17.35	0.01
(σ)-GP	3.93	0.17	**0.18**	−1.37	0.64	−0.93	0.43	−0.94	0.52	−1.07	0.17	−0.42	1.38	1.04	16.84	0.07
(ℓ)-GP	3.94	0.16	**0.18**	**−1.38**	**0.53**	**−1.05**	0.44	−0.88	0.41	−1.16	0.17	−0.34	**1.35**	**−0.72**	17.63	0.09
(ω,σ)-GP	**3.87**	**−0.23**	**0.18**	−1.30	0.65	−0.86	0.43	**−1.07**	0.56	−1.51	**0.15**	−1.61	1.60	−0.66	16.33	**−0.02**
(ω,ℓ)-GP	4.02	−0.19	0.19	−1.31	**0.53**	−0.93	**0.42**	−0.99	**0.40**	**−1.83**	0.17	−0.90	1.38	−0.47	**9.30**	0.01
(ω,σ,ℓ)-GP	3.90	−0.21	0.19	−1.32	**0.53**	−0.90	0.45	−0.98	0.43	−1.70	0.16	**−1.79**	1.47	−0.32	9.77	−0.00



(a) All datasets and models

(b) $\mathbf{D}_\ell$ dataset with ℓ-GP model

Figure 1: The MSE and NLPD performance of the HMC posterior samples. (a) Comparison of test errors between MAP and HMC mean solutions over all datasets and methods ($8 \times 7 = 56$, x-axis limited to 5 for clarity). (b) Test errors of the HMC samples compared to the HMC mean and MAP solution on a single D_ℓ dataset with ℓ-GP model.

gression, while HMC sampling took several hours on the tested datasets (data not shown).

3.1 Regression performance

Table 2 shows the MSE and NLPD performance on the test folds of the synthetic datasets using MAP GP models with different combinations of non-stationary parameters. For each dataset, the model with the lowest NLPD, or second lowest NLPD value, is the one where the model's non-stationarities match those of the dataset. For instance, the dataset **T** contains heteroscedastic noise and input-dependent signal variance σ. For this, the best NLPD performance is obtained with a matching ω, σ-GP, and with a fully non-stationary ω, σ, ℓ-GP as well. The vanilla (stationary) GP performance is always surpassed by non-stationary models on datasets with non-stationary dynamics.

Adding 'unnecessary' non-stationarities retains or only slightly worsens the performance, with the major exception being the dataset **J**. Here, the lengthscale is clearly input-dependent (NLPD -0.72, optimal), while in contrast the non-stationary signal variance σ is unable to model the data (NLPD 1.04). Adding heteroscedastic weakens the model, giving a strong indication of a homoscedastic noise model.

3.2 HMC performance

We explored the difference between the MAP solution and the HMC sampling. In practice we found the MAP to be slightly better on average regarding the MSE and NLPD values (See Figure 1a). However, the sampling is able to explore the multimodality of the latent posterior (See Figure 5). Figure 1b shows the test errors of the individual HMC samples in comparison to the MAP solution with the $\mathbf{D}_\ell$ dataset using the ℓ-GP model. The HMC solution includes numerous samples that are better, while on average being slightly worse than MAP.

The dataset $\mathbf{D}_\ell$ contains several latent modes (See Figure 4, bottom), which the HMC sampler captures. These modes include latent functions that imply a 'shortcut' or a 'zigzag' signal around timepoints 0.18 or 0.75, or both. The HMC samples are centered mostly around the shortcut profile at the earlier timepoint, while only a few samples

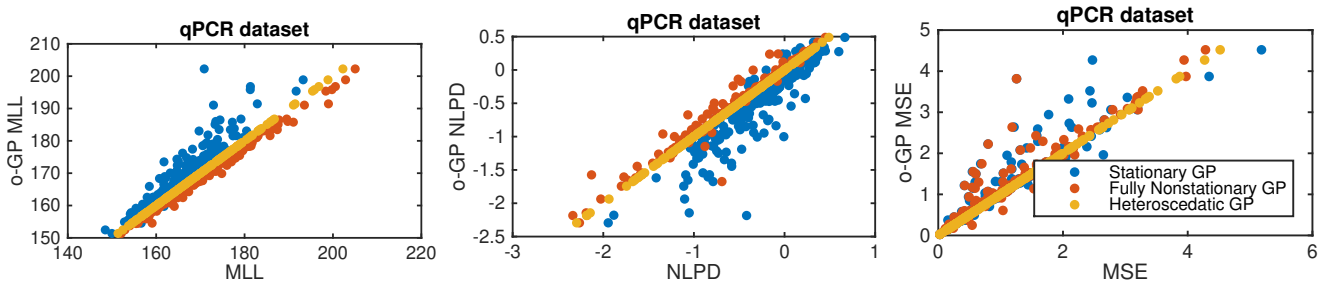


Figure 2: Comparison of the MLL, MSE and NLPD values (x-axis) of the stationary GP, ω-GP and (ω, σ, ℓ)-GP over the 205 gene expression time series (x-axis) against the heteroscedastic GP on the y-axis. Each row contains a triplet of values corresponding to the three GP models of the same time series.

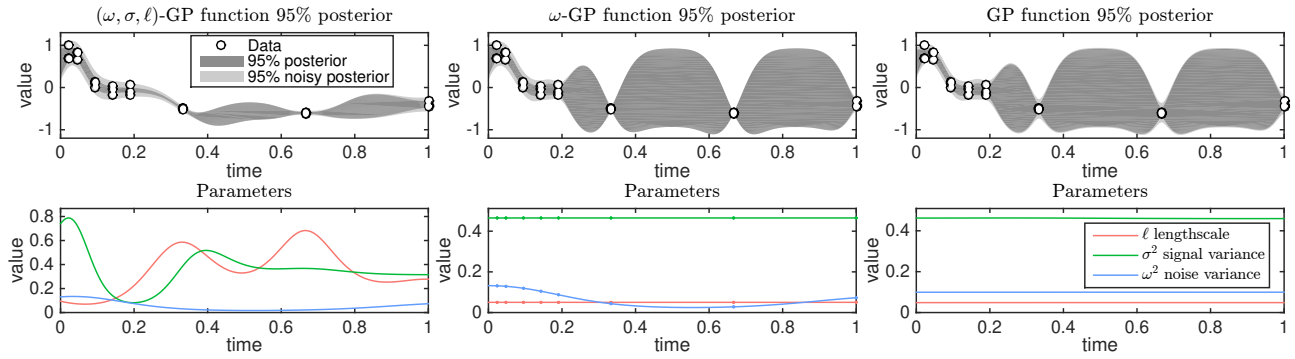


Figure 3: Comparison of three GP models on an example gene expression time series.

with the shortcut profile exist at the later timepoint. The MAP solution has chosen both zigzags. The latent posterior shows largest variance in the signal variance $\sigma(t)$ component, while the lengthscale $\ell(t)$ and noise variance ω have tighter distributions.

3.3 Biological dataset

We demonstrate the method with a biological dataset of 205 gene expression time series measurements of human endothelial cells after irradiation at time $t = 0$. Due to the irradiation the dataset exhibits non-stationary dynamics as the cells try to repair themselves and revert back to steady states. The gene expressions are measured over 8 days $(0.5, 1, 2, 3, 4, 7, 14, 21)$ in three replicates (Heinonen et al., 2015). The goal is to construct a realistic model of the underlying gene expression process and the underlying dynamics with no knowledge of the 'true' expression levels, given only the small number of sparse measurements.

We modeled the dataset using stationary GP, heteroscedastic ω-GP and three non-stationary GPs: (ω, σ)-GP, (ω, ℓ)-GP and with (ω, σ, ℓ)-GP. We found the performance of the three non-stationary GPs to be similar. Figure 2 indicates the MLL, MSE and NLPD values of the 205 time-series under stationary, heteroscedastic or completely non-stationary models. Addition of heteroscedasticity greatly increases the model fits, while also improving the data likelihoods against the function posterior. Finally the completely non-stationary GP still improves model fits, while consistently improving the NLPD values, with similar MSE performance compared to the HGP. Figure 3 compares the three models learned from an example gene expression time series (See Supplementary Material for additional models).

3.4 Latent function reconstruction

An interesting application of the proposed method is to learn the 'true' input-dependent parameters of the data generating process, with only samples of the function f and no samples of the underlying parameters. The key question is how accurately the parameters of the non-stationary model can be inferred in this setting. Due to the lack of empirical datasets with 'gold-standard' input-dependent parameter values, we show promising results on parameter reconstruction error on simulated data.

We simulate a noisy sample where true generating latent parameter functions $\ell(\cdot), \sigma(\cdot), \omega(\cdot)$ are known. We infer both the MAP solution and sample the posterior of the latent parameter processes ℓ, σ, ω and the unknown function f. Figure 5 highlights the MAP and HMC solutions in comparison to the generating parameters, and to the state-

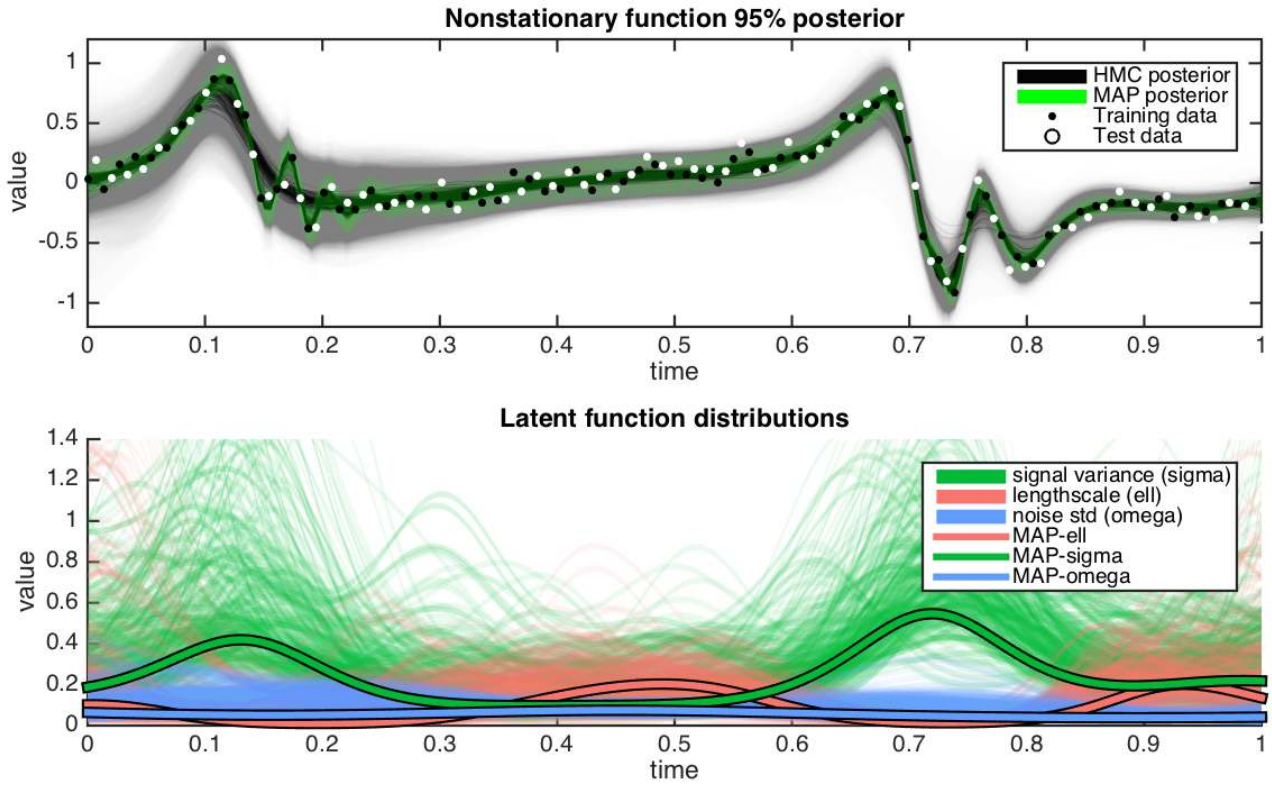


Figure 4: The set of function posteriors corresponding to the latent function samples are plotted in gray, and the MAP solution in green, along with the training and test data (top). The latent function sample is drawn with green (σ), red (ℓ) and blue (ω) colors, with the MAP solution highlighted with bold lines (bottom).

of-the-art σ, ω-GP model of Tolvanen et al. (2014). Given only a single noisy time series from a generating function, a range of matching parameter processes can be inferred. As shown in Figure 5, the obtained HMC parameter samples overlap with the generating parameters, while the MAP solution differs slightly due to the inherent randomness in the noise and in data sampling.

4 Discussion

In this paper, we have proposed a fully non-stationary Gaussian process regression framework, where all three key components can be input-dependent. Our approach uses analytical gradient-based techniques to perform inference with HMC sampling and MAP estimation. We are able to effectively sample from the exact posterior of the latent functions. We have shown that the method is able to infer the underlying latent functions and improve regression performance when the datasets truly are non-stationary, and achieve equivalent performance to a stationary model when they are not.

The interplay between the signal variance and the lengthscale is an interesting topic (Diggle et al., 1998; Zhang, 2004). When modeling the 'jump' dataset the non-stationary signal variance was unable to model dynamics, while a non-stationary lengthscale produced a good model. This is natural since the signal variance serves as a linear amplitude over the function f, while the lengthscale has a possibly non-linear effect on the function model. In addition, the non-stationary squared exponential kernel can be changed into any differentiable non-stationary covariance function with input-dependent parameters, e.g. the non-stationary Matérn kernel could be used (Paciorek and Schervish, 2004).

The gradient-based HMC is a powerful inference tool for Gaussian processes, and could be further enhanced by utilizing natural gradients or position-dependent mass matrices with Riemannian Manifold HMC (Girolami and Calderhead, 2011). We note that the method could be extended by also inferring the hyperparameters $\boldsymbol{\theta}$ using HMC. However, proper care has to be taken to set their priors.

Acknowledgements

This work has been funded in part by Finnish Funding Agency for Innovation Tekes (grant no 40128/14, Living Factories).

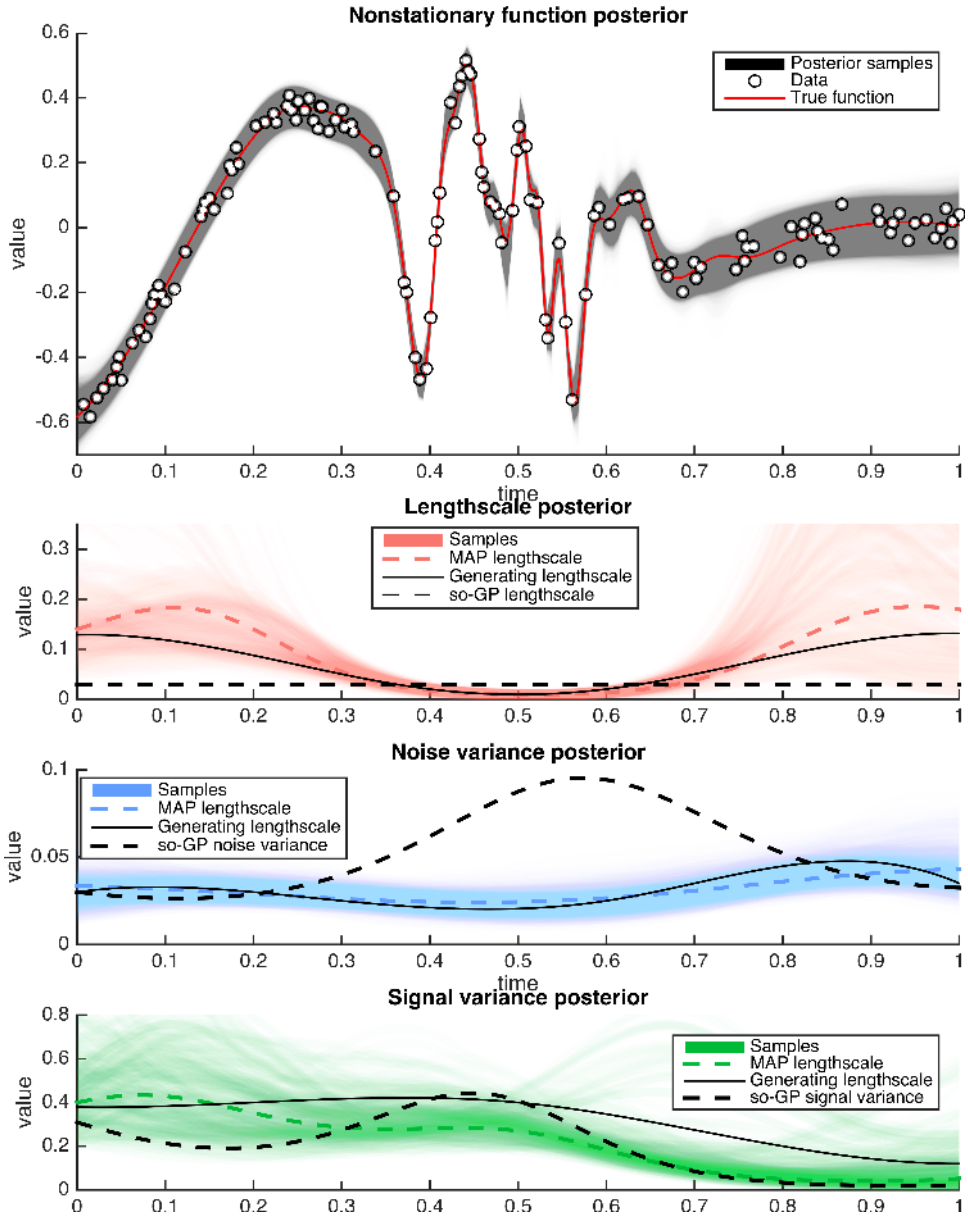


Figure 5: Latent function reconstruction errors (top) over a non-stationary simulated dataset with known non-stationary dynamics. Completely non-stationary GP model (bottom) posteriors, MAP solution (dashed line) and generating latent functions (black solid lines) with 150 data points. The latent lengthscales and noises are estimated correctly, while signal variance is approximately matched. Dashed black lines shows the comparison to the state-of-the-art σ, ω-GP model of Tolvanen et al. (2014).

References

E. Anderes and M. Stein. Estimating deformations of isotropic gaussian random fields on the plane. *The Annals of Statistics*, 36:719–741, 2008.

P. Diggle, J. Tawn, and R. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society, Series C*, 47 (3):299–350, 1998.

M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, Department of Physics, University of Cambridge, 1997.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 73(2):123–214, 2011.

P. Goldberg, C. Williams, and C. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In *NIPS*, pages 493–499, 1997.

R. Gramacy. *Bayesian treed Gaussian process models*. PhD thesis, UC Santa Cruz, 2005.

M. Heinonen, O. Guipaud, F. Milliat, V. Buard, B. Micheau, G. Tarlet, M. Benderitter, F. Zehraoui, and F. d'Alche Buc. Detecting time periods of differential gene expression using gaussian processes: An application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics*, 31:728–735, 2015.

M. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heterscedatic gaussian process regression. In *ICML*, pages 393–400, 2007.

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

M. Lazaro-Gredilla and M. K. Titsias. Variational heteroscedatic gaussian process regression. *ICML*, pages 841–848, 2011.

Q. Le, A. Smola, and S. Canu. Heteroscedastic Gaussian process regression. In *ICML*, pages 489–496, 2005.

R. Neal. *Handbook of Markov Chain Monte Carlo*, chapter 5, MCMC Using Hamiltonian Dynamics. CRC Press, 2011.

C. Paciorek and M. J. Schervish. Nonstationary covariance functions for gaussian process regression. In *NIPS*, pages 273–280, 2004.

N. Quadrianto, K. Kersting, M. Reid, T. Caetano, and W. Buntine. Kernel conditional quantile estimation via reduction revisited. In *IEEE International Conference on Data Mining*, pages 938–943, 2009.

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

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

A. Schmidt and A. O'Hagan. Bayesian inference for nonstationary spatial covariance structures via spatial deformations. *Journal of the Royal Statistical Society Series B*, 65:743–758, 2003.

B. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, 47:1–52, 1985.

J. Snoek, K. Swersky, R. Zemel, and R. Adams. Input warping for bayesian optimization of non-stationary functions. In *ICML*, pages 1674—-1682, 2014.

V. Tolvanen, P. Jylänki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic gaussian process regression. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–6. IEEE, 2014.

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. Gpstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.

C. Wang and R. Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. Technical report, University of Toronto, 2012. URL `http://arxiv.org/abs/1212.6246`.

A. Wilson and Z. Ghahramani. Generalised wishart processes. In *Uncertainty in Artificial Intelligence (UAI)*, pages 736–744, 2011.

H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.