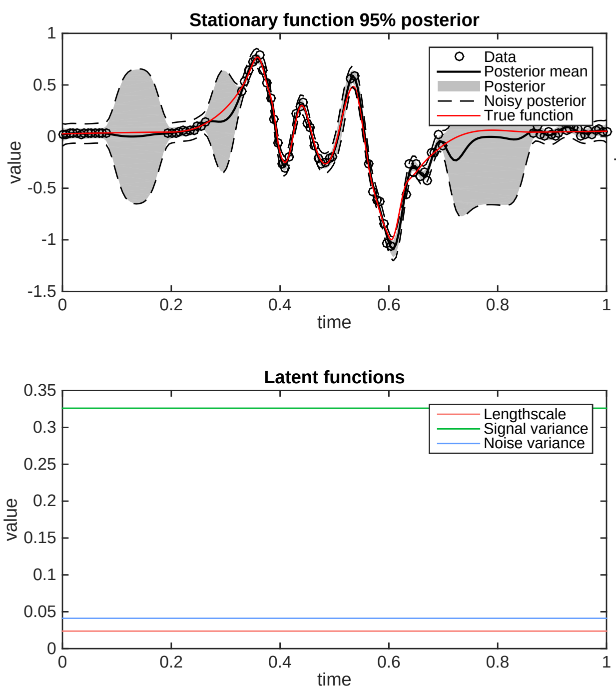
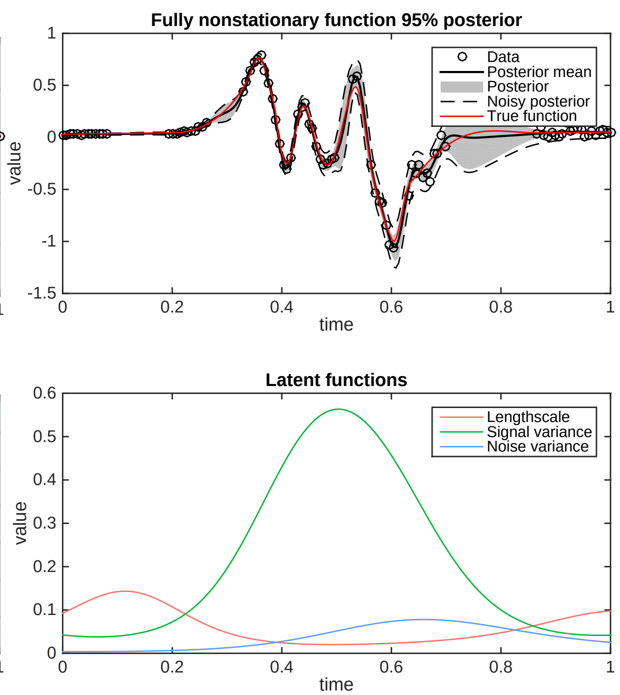
Supplemental information:
Non-Stationary Gaussian Process Regression with Hamiltonian Monte Carlo

Comparison between vanilla and (ω, σ, ℓ)-GP

A comparison between stationary (vanilla) and (ω, σ, ℓ)-GP is shown in Supplemental Figure 1.



(a) Vanilla GP (NLPD -1.95)



(b) Fully nonstationary GP (NLPD -2.47)

Figure 1: Nonstationary, heteroscedastic GP (right) fits nonstationary dataset better than vanilla GP (left). Vanilla GP overestimates confidence intervals at stationary regions, as a result of having to choose the lengthscale to fit the non-stationary regions, and for the same reason produces spurious oscillations in regions where there are no observations. The dataset is the $\mathbf{D}_{\omega,\sigma,\ell}$ from Table 2.

Comparison of GP models on gene expression time series

A comparison between stationary and all 7 non-stationary GP variants on the gene 140 expression time series is shown in Supplemental Figure 2.

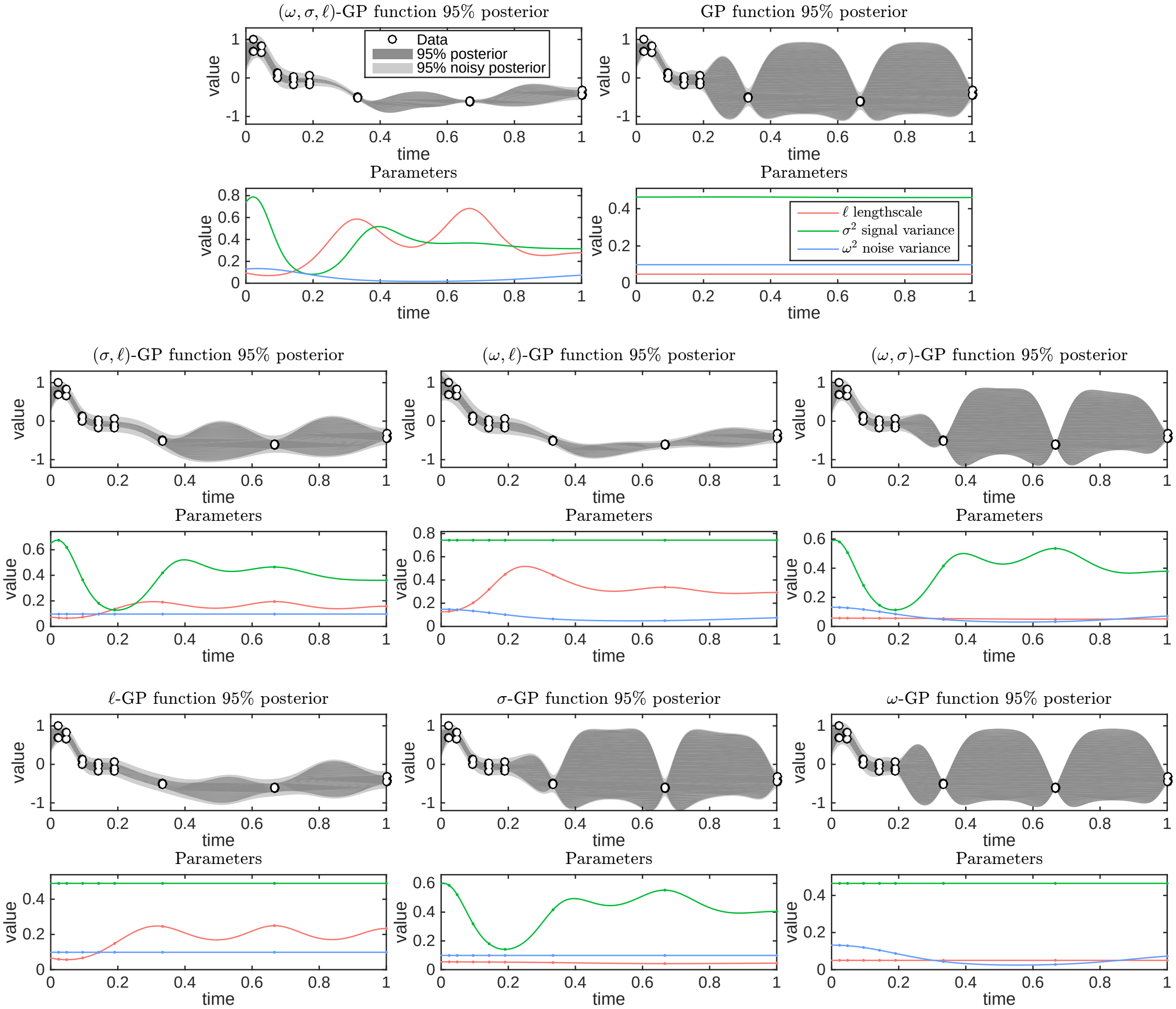


Figure 2: Comparison of all 8 combinations of stationarities and non-stationarities in the parameters on an example gene expression time series (gene 140).

Kernel SDP proof

The kernel

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

is a one-dimensional Gaussian SDP kernel $C^{NS}(\mathbf{x}_i,\mathbf{x}_j)$ of Paciorek and Schervish (2004). The kernel K_f can be stated as

$$K_f(x,x') = \sigma(x)K(x,x')\sigma(x'),$$

which is positive definite for any function $\sigma(\cdot)$ (Shawe-Taylor and Christianini, 2004).

Conditional distributions

The conditional distributions of the latent functions at target timepoints $\mathbf{x}_\star$ given the latent functions at observed timepoints $\mathbf{x}$ are

$$
\begin{aligned}
p(\tilde{\boldsymbol{\ell}}_\star|\tilde{\boldsymbol{\ell}}) &= \mathcal{N}(\tilde{\boldsymbol{\ell}}_\star|K_\ell(\mathbf{x}_\star,\mathbf{x})^T K_\ell(\mathbf{x},\mathbf{x})^{-1}(\tilde{\boldsymbol{\ell}}-\mu_\ell)+\mu_\ell, K_\ell(\mathbf{x}_\star,\mathbf{x}_\star)-K_\ell(\mathbf{x}_\star,\mathbf{x})^T K_\ell(\mathbf{x},\mathbf{x})^{-1}K_\ell(\mathbf{x}_\star,\mathbf{x}))\\
p(\tilde{\boldsymbol{\sigma}}_\star|\tilde{\boldsymbol{\sigma}}) &= \mathcal{N}(\tilde{\boldsymbol{\sigma}}_\star|K_\sigma(\mathbf{x}_\star,\mathbf{x})^T K_\sigma(\mathbf{x},\mathbf{x})^{-1}(\tilde{\boldsymbol{\sigma}}-\mu_\sigma)+\mu_\sigma, K_\sigma(\mathbf{x}_\star,\mathbf{x}_\star)-K_\sigma(\mathbf{x}_\star,\mathbf{x})^T K_\sigma(\mathbf{x},\mathbf{x})^{-1}K_\sigma(\mathbf{x}_\star,\mathbf{x}))\\
p(\tilde{\boldsymbol{\omega}}_\star|\tilde{\boldsymbol{\omega}}) &= \mathcal{N}(\tilde{\boldsymbol{\omega}}_\star|K_\omega(\mathbf{x}_\star,\mathbf{x})^T K_\omega(\mathbf{x},\mathbf{x})^{-1}(\tilde{\boldsymbol{\omega}}-\mu_\omega)+\mu_\omega, K_\omega(\mathbf{x}_\star,\mathbf{x}_\star)-K_\omega(\mathbf{x}_\star,\mathbf{x})^T K_\omega(\mathbf{x},\mathbf{x})^{-1}K_\omega(\mathbf{x}_\star,\mathbf{x})),
\end{aligned}
$$

where K_ℓ, K_σ and K_ω are standard gaussian kernels computed using the hyperparameters $\boldsymbol{\theta}$ (Rasmussen and Williams, 2006).

Partial derivatives

The partial derivatives of the unconstrained latent functions against the marginal log likelihood

$$
\begin{aligned}
\log\mathcal{L} = \log p(\mathbf{y}|\tilde{\boldsymbol{\ell}},\tilde{\boldsymbol{\sigma}},\tilde{\boldsymbol{\omega}},\theta) &= \log p(\mathbf{y}|\tilde{\boldsymbol{\ell}},\tilde{\boldsymbol{\sigma}},\tilde{\boldsymbol{\omega}})p(\tilde{\boldsymbol{\ell}}|\theta)p(\tilde{\boldsymbol{\sigma}}|\theta)p(\tilde{\boldsymbol{\omega}}|\theta)\\
&= \log\mathcal{N}(\mathbf{y}|\mathbf{0},K_f+\Omega)+\log\mathcal{N}(\tilde{\boldsymbol{\ell}}|\mu_{\tilde{\ell}},K_{\tilde{\ell}})+\log\mathcal{N}(\tilde{\boldsymbol{\sigma}}|\mu_{\tilde{\sigma}},K_{\tilde{\sigma}})+\log\mathcal{N}(\tilde{\boldsymbol{\omega}}|\mu_{\tilde{\omega}},K_{\tilde{\omega}})
\end{aligned}
$$

are analytically derived.

The partial derivative for $\tilde{\ell}(t)$ is

$$
\begin{aligned}
\frac{\partial\log\mathcal{L}}{\partial\tilde{\boldsymbol{\ell}}_i} &= \frac{1}{2}\operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})\frac{\partial K_y}{\partial\tilde{\boldsymbol{\ell}}_i}\right)-[K_{\tilde{\ell}}^{-1}(\tilde{\boldsymbol{\ell}}-\mu_{\tilde{\ell}})]_i\\
\frac{\partial[K_y]_{ij}}{\partial\tilde{\boldsymbol{\ell}}_i} &= \frac{S_{ij}E_{ij}}{R_{ij}L_{ij}^3}\ell_i\ell_j(4d\ell_i^2-\ell_i^2+\ell_j^4),
\end{aligned}
$$

$S_{ij}=\sigma_i\sigma_j$, $R_{ij}=\sqrt{\frac{2\ell_i\ell_j}{\ell_i^2+\ell_j^2}}$, $E_{ij}=\exp\left(-\frac{(t_i-t_j)^2}{\ell_i^2+l_j^2}\right)$, and $L_{ij}=\ell_i^2+\ell_j^2$. The derivative matrix $\frac{\partial K_y}{\partial\tilde{\boldsymbol{\ell}}_i}$ becomes a 'plus' matrix where only i'th column and row are nonzero.

The derivatives for $\tilde{\sigma}(t)$ is

$$
\frac{\partial\log\mathcal{L}}{\partial\tilde{\boldsymbol{\sigma}}} = \operatorname{diag}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})K_f\right)-K_{\tilde{\sigma}}^{-1}(\tilde{\boldsymbol{\sigma}}-\mu_{\tilde{\sigma}})
$$

and for $\tilde{\omega}(t)$ is

$$
\frac{\partial\log\mathcal{L}}{\partial\tilde{\boldsymbol{\omega}}} = \operatorname{diag}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})\Omega\right)-K_{\tilde{\omega}}^{-1}(\tilde{\boldsymbol{\omega}}-\mu_{\tilde{\omega}})
$$

where $\boldsymbol{\alpha}=K_y^{-1}\mathbf{y}$.

The partial derivatives of the latent parameters $(\tilde{\ell},\tilde{\sigma},\tilde{\omega})\in\mathbb{R}^3$ in a stationary formulation are for $\tilde{\ell}$ (Rasmussen and Williams, 2006)

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\tilde{\ell}} &= \frac{1}{2}\operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})\frac{\partial K_y}{\partial\tilde{\ell}}\right)-\mathbf{1}^T K_{\tilde{\ell}}^{-1}(\tilde{\ell}\mathbf{1}-\mu_{\tilde{\ell}})\\
\frac{\partial K_y}{\partial\tilde{\ell}} &= \ell^{-2}D\odot K_f,
\end{aligned}
$$

for $\tilde{\sigma}$

$$
\frac{\partial\mathcal{L}}{\partial\tilde{\sigma}} = \operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})K_f\right)-\mathbf{1}^T K_{\tilde{\sigma}}^{-1}(\tilde{\sigma}\mathbf{1}-\mu_{\tilde{\sigma}})
$$

and for $\tilde{\omega}$

$$
\frac{\partial\mathcal{L}}{\partial\tilde{\omega}} = \operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T-K_y^{-1})\Omega\right)-\mathbf{1}^T K_{\tilde{\omega}}^{-1}(\tilde{\omega}\mathbf{1}-\mu_{\tilde{\omega}}).
$$

Multivariate inputs

The method extends directly into multivariate inputs $\mathbf{x} \in \mathbb{R}^d$ by defining the distance functions from all squared exponential kernels into squared norms,

$$K_f(\mathbf{x}, \mathbf{x}') = \sigma(\mathbf{x})\sigma(\mathbf{x}')\sqrt{\frac{2\ell(\mathbf{x})\ell(\mathbf{x}')}{\ell(\mathbf{x})^2 + \ell(\mathbf{x}')^2}} \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{\ell(\mathbf{x})^2 + \ell(\mathbf{x}')^2}\right)$$
$$K_c(\mathbf{x}, \mathbf{x}') = \alpha_c^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\beta_c^2}\right),$$

for $c \in \{\ell, \sigma, \omega\}$.

Two-dimensional data

We demonstrate the feasibility of the proposed method in two dimensions. As a base dataset we use the two-dimensional data[1] from Paciorek and Schervish (2004) with assumed unknown underlying distribution and noise.

We turn the data into non-stationary by adding a high 'peak' at $(0.4, 1.5)$ and a deep 'hole' at $(0.5, -1.0)$. In addition, we remove all datapoints inside a rectangle defined by $-0.6 < x_1 < 0$ and $0 < x_2 < 0.8$ to assess the GP interpolation. The Supplemental Figure 3 shows the resulting model using both stationary and non-stationary GP. The non-stationary GP is necessary to properly model the peak and the hole, while decreasing the posterior variance.

The stationary GP parameters are $\ell = 0.21$, $\sigma = 1.31$ and $\omega = 0.37$. The hyperparameters of the non-stationary GP have a large impact on the resulting model. The underlying data has a varying signal variance, and hence we set the $\beta_\sigma = 0.2$ to a low value to achieve peaked signal variance surface, while keeping the other two parameters smooth with $\beta_\ell = \beta_\omega = 5$. We set the parameter variances to $\alpha_\ell = 0.5, \alpha_\sigma = 3, \alpha_\omega = 1$. It was sufficient to empirically test the hyperparameters from a set $\{0.1, 0.2, 0.5, 1, 3, 5\}$.

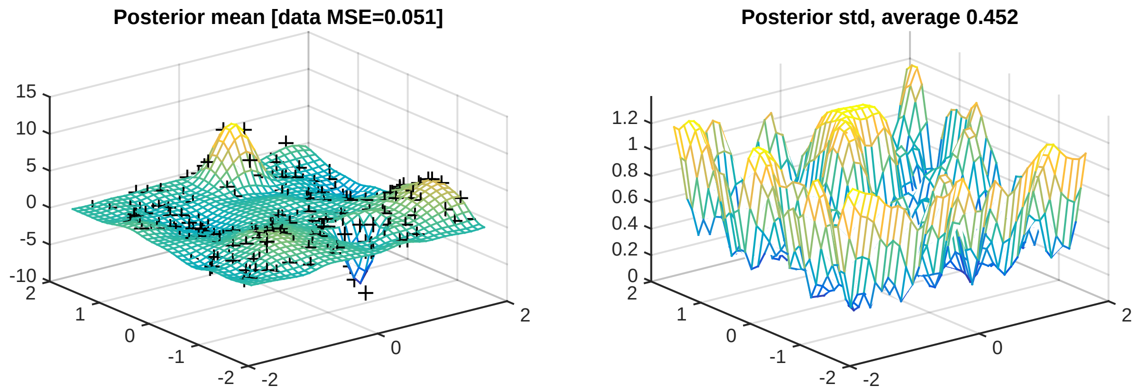
References

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

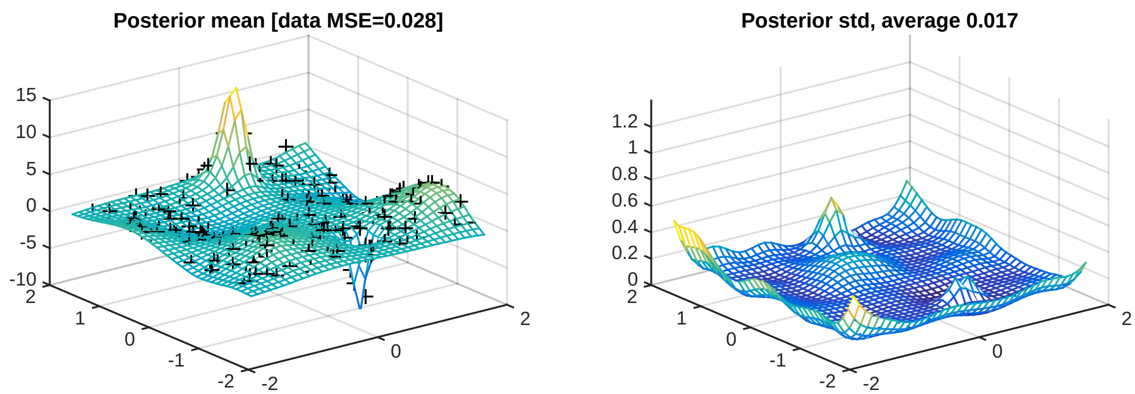
C.E. Rasmussen and K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

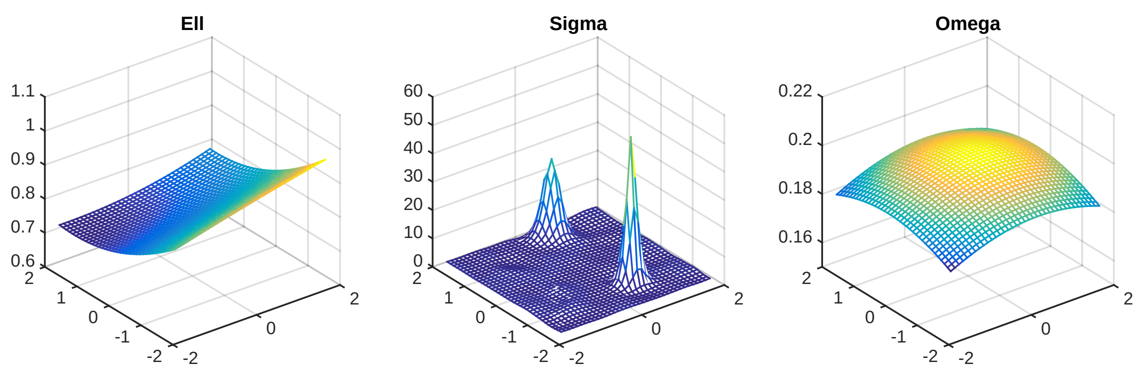
[1]The raw dataset is included in the `demo_regression1` example of GPstuff package



(a) Stationary GP



(b) (ℓ, σ, ω)-GP



(c) (ℓ, σ, ω)-GP parameter surfaces

Figure 3: (a) Stationary GP posterior mean and standard deviation compared against (b) (ℓ, σ, ω)-GP posterior. The non-stationary GP parameter surfaces are shown in subfigure (c).