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

Comparison between vanilla and \((\omega, \sigma, \ell)\)-GP

A comparison between stationary (vanilla) and \((\omega, \sigma, \ell)\)-GP is shown in Supplemental Figure 1.

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 \(D_{\omega,\sigma,\ell}\) 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 \( C^{NS}(x_i, x_j) \) 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 \( x_s \) given the latent functions at observed timepoints \( x \) are

\[
p(\tilde{\ell}, \sigma | x, K, K_f) = N(\tilde{\ell}, \mu_\ell, \Sigma_\ell) = N(\tilde{\ell}, \mu_\ell, K_f)K_f(x, x)^{-1}K_f(x, x)^{-1}K_f(x, x)
\]

where \( K_f, K, \) and \( \sigma \) are standard gaussian kernels computed using the hyperparameters \( \Theta \) (Rasmussen and Williams, 2006).

Partial derivatives

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

\[
\log L = \log p(y | \tilde{\ell}, \sigma, \Theta) = \log p(y | \tilde{\ell}, \sigma, \Theta)p(\tilde{\ell} | \Theta)p(\sigma | \Theta)
\]

are analytically derived.

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

\[
\frac{\partial \log L}{\partial \ell_i} = \frac{1}{2} \text{tr} \left( \left( \alpha \alpha^T - K_y^{-1} \right) \frac{\partial K_y}{\partial \ell_i} \right) - [K_f^{-1} - \mu_\ell]_i,
\]

\[
\frac{\partial[K_y]_{ij}}{\partial \ell_i} = \frac{S_{ij}E_{ij}L_{ij}}{R_{ij}L_{ij}^2}(4d \ell_i^2 - \ell_i^2),
\]

where \( S_{ij} = \sigma_i \sigma_j, R_{ij} = \sqrt{\frac{2d}{\ell_i + \ell_j} E_{ij} = \exp(-\frac{(\ell_i - \ell_j)^2}{\ell_i + \ell_j}), \text{ and } L_{ij} = \ell_i^2 + \ell_j^2. \) The derivative matrix \( \frac{\partial K_y}{\partial \omega} \) becomes a ‘plus’ matrix where only \( i \)th column and row are nonzero.

The derivatives for \( \sigma(t) \) is

\[
\frac{\partial \log L}{\partial \sigma} = \text{diag} \left( \left( \alpha \alpha^T - K_y^{-1} \right) K_f \right) - K_y^{-1} (\tilde{\sigma} - \mu_\sigma)
\]

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

\[
\frac{\partial \log L}{\partial \tilde{\omega}} = \text{diag} \left( \left( \alpha \alpha^T - K_y^{-1} \right) \Omega \right) - K_y^{-1} (\tilde{\omega} - \mu_\omega)
\]

where \( \alpha = K_y^{-1} y \).

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)

\[
\frac{\partial L}{\partial \ell} = \frac{1}{2} \text{tr} \left( \left( \alpha \alpha^T - K_y^{-1} \right) \frac{\partial K_y}{\partial \ell} \right) - 1^T K_f^{-1} (\tilde{1} - \mu_\ell)
\]

\[
\frac{\partial K_y}{\partial \ell} = \ell^{-2} D \circ K_f,
\]

for \( \tilde{\sigma} \)

\[
\frac{\partial L}{\partial \sigma} = \text{tr} \left( \left( \alpha \alpha^T - K_y^{-1} \right) K_f \right) - 1^T K_y^{-1} (\tilde{1} - \mu_\sigma)
\]

and for \( \tilde{\omega} \)

\[
\frac{\partial L}{\partial \tilde{\omega}} = \text{tr} \left( \left( \alpha \alpha^T - K_y^{-1} \right) \Omega \right) - 1^T K_y^{-1} (\tilde{1} - \mu_\omega).
\]
**Multivariate inputs**

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

\[
K_f(x, x') = \sigma(x)\sigma(x')\sqrt{\frac{2\ell(x)\ell(x')}{\ell(x)^2 + \ell(x')^2}} \exp \left( -\frac{|x - x'|^2}{\ell(x)^2 + \ell(x')^2} \right)
\]

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

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**


\(^1\)The raw dataset is included in the demo_regression1 example of GPstuff package
Figure 3: (a) Stationary GP posterior mean and standard deviation compared against (b) \((\ell, \sigma, \omega)\)-GP posterior. The non-stationary GP parameter surfaces are shown in subfigure (c).