Gaussian Process Product Models for Nonparametric Nonstationarity

Ryan Prescott Adams and Oliver Stegle

Inference Group
Cavendish Laboratory
University of Cambridge

ICML July 2008
Gaussian Processes for Regression

- Bayesian regression problem – given a dataset $\mathcal{D} = \{x_n, y_n\}_{n=1}^N$
- Want to learn the latent function $f$ this data comes from.

Gaussian process is a prior over functions $f(x)$

$$p(f(x) | \mathcal{D}) = \frac{p(\mathcal{D}|f(x)) p(f(x))}{p(\mathcal{D})}.$$
Gaussian Processes for Regression

- Bayesian regression problem – given a dataset \( D = \{x_n, y_n\}_{n=1}^N \)
- Want to learn the latent function \( f \) this data comes from.

Gaussian process is a prior over functions \( f(x) \)

\[
p(f(x)|D) = \frac{p(D|f(x))p(f(x))}{p(D)}.
\]
Gaussian processes for regression

- Gaussian process marginal posterior mean and errorbars:

- Belief about smoothness and lengthscale of $f(x)$ expressed via the covariance function $C(x, x')$. 

R. P. Adams and O. Stegle

Gaussian Process Product Models for Nonparametric Nonstationarity

ICML 2008
Gaussian Processes for Regression

- Gaussian process marginal posterior mean and errorbars:

- Belief about smoothness and lengthscale of $f(x)$ expressed via the covariance function $C(x, x')$. 
For a “vanilla GP” the marginal predictive distribution for an unseen input $x^*$ is Gaussian

$$p(y^* | x^*, D) = N(\mu^*, v^*),$$

$$\mu^* = k_T^T C_N^{-1} y_N$$

$$v^* = C(x^*, x^*) - k_T^T C_N^{-1} k_N$$

Hyperparameters $\theta$ can be optimized via LML

$$\mathcal{L}(\theta) = -\frac{1}{2} \ln |C_N(\theta)| - \frac{1}{2} y_N^T C_N^{-1}(\theta) y_N - \frac{N}{2} \ln 2\pi.$$
Gaussian Processes for Regression
Predictions and Hyperparameters

For a “vanilla GP” the marginal predictive distribution for an unseen input $x^*$ is Gaussian

$$p(y^*|x^*, D) = \mathcal{N}(\mu^*, v^*),$$

$$\mu^* = k_N^T C_N^{-1} y_N$$

$$v^* = C(x^*, x^*) - k_N^T C_N^{-1} k_N$$

Hyperparameters $\theta$ can be optimized via LML

$$\mathcal{L}(\theta) = -\frac{1}{2} \ln |C_N(\theta)| - \frac{1}{2} y_N^T C_N^{-1}(\theta) y_N - \frac{N}{2} \ln 2\pi.$$
Covariance function $C(x, x')$ to model how targets at inputs $x, x'$ co-vary.

A popular choice

$$C(x, x') = C_0 \exp \left( -0.5 \frac{|x - x'|^2}{l^2} \right) + \sigma^2 \delta_{x, x'}$$

with hyperparameters $C_0, l, \sigma$

Covariance functions that only depend on the distance between inputs are called stationary.
Covariance function $C(x, x')$ to model how targets at inputs $x, x'$ co-vary.

A popular choice

$$C(x, x') = C_0 \exp \left( -0.5 \frac{|x - x'|^2}{l^2} \right) + \sigma^2 \delta_{x, x'}$$

with hyperparameters $C_0, l, \sigma$

Covariance functions that only depend on the distance between inputs are called stationary.
Gaussian Processes for Regression

Stationarity vs Nonstationarity

Stationary covariance functions

- Stationary covariances yield intuitive interpretation
- But
  - Strong assumption
  - Do real data look like this?

Predefined nonstationary covariance functions

- We could specify a nonstationary $C(x, x')$ a priori
- But...
  - Nonintuitive and difficult task
  - We are still making strong assumptions

Learn nonstationarity

- Introduce additional latent spaces
### Gaussian Processes for Regression

#### Stationarity vs Nonstationarity

<table>
<thead>
<tr>
<th>Stationary covariance functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Stationary covariances yield intuitive interpretation</td>
</tr>
<tr>
<td>• But</td>
</tr>
<tr>
<td>▶ Strong assumption</td>
</tr>
<tr>
<td>▶ Do real data look like this?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predefined nonstationary covariance functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• We could specify a nonstationary $C(x, x')$ a priori</td>
</tr>
<tr>
<td>• But</td>
</tr>
<tr>
<td>▶ Nonintuitive and difficult task</td>
</tr>
<tr>
<td>▶ We are still making strong assumptions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learn nonstationarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Introduce additional latent spaces</td>
</tr>
</tbody>
</table>
Stationary covariance functions

- Stationary covariances yield intuitive interpretation
- But
  - Strong assumption
  - Do real data look like this?

Predefined nonstationarity covariance functions

- We could specify a nonstationary $C(x, x')$ a priori
- But
  - Nonintuitive and difficult task
  - We are still making strong assumptions

Learn nonstationarity

- Introduce additional latent spaces
Gaussian Processes for Regression
Stationarity vs Nonstationarity

Stationary covariance functions
- Stationary covariances yield intuitive interpretation
- But
  - Strong assumption
  - Do real data look like this?

Predefined nonstationarity covariance functions
- We could specify a nonstationary $C(x, x')$ a priori
- But
  - Nonintuitive and difficult task
  - We are still making strong assumptions

Learn nonstationarity
- Introduce additional latent spaces
Gaussian Processes for Regression
Stationarity vs Nonstationarity

**Stationary covariance functions**
- Stationary covariances yield intuitive interpretation
- But
  - Strong assumption
  - Do real data look like this?

**Predefined nonstationary covariance functions**
- We could specify a nonstationary $C(x, x')$ a priori
- But
  - Nonintuitive and difficult task
  - We are still making strong assumptions

**Learn nonstationarity**
- Introduce additional latent spaces
Latent space extensions of stationary covariances

\[ C(x, x') = C_0 \exp \left( -0.5 (x - x')^T W (x - x') \right) + \sigma^2 \delta_{x, x'} \]

- Variable lengthscale, spatial deformations \( W(x) \) (Schmidt & O’Hagan, 2003)
- Input dependent observation noise \( \sigma(x) \) (Goldberg et al., 1998)
- Nonstationary amplitude variations \( C_0(x) \) (Turner & Sahani, 2008)

This work
Latent space extensions of stationary covariances

\[ C(x, x') = C_0 \exp \left( -0.5 (x - x')^T W (x - x') \right) + \sigma^2 \delta_{x,x'} \]

- Variable lengthscale, spatial deformations \( W(x) \) (Schmidt & O’Hagan, 2003)
- Input dependent observation noise \( \sigma(x) \) (Goldberg et al., 1998)
- Nonstationary amplitude variations \( C_0(x) \) (Turner & Sahani, 2008)

This work

R. P. Adams and O. Stegle
Gaussian Process Product Models for Nonparametric Nonstationarity
ICML 2008 7 / 30
Latent space extensions of stationary covariances

\[ C(x, x') = C_0 \exp \left( -0.5 (x - x')^T W (x - x') \right) + \sigma^2 \delta_{x,x'} \]

- Variable lengthscale, spatial deformations \( W(x) \) (Schmidt & O’Hagan, 2003)
- Input dependent observation noise \( \sigma(x) \) (Goldberg et al., 1998)
- Nonstationary amplitude variations \( C_0(x) \) (Turner & Sahani, 2008)

This work
Latent space extensions of stationary covariances

\[ C(x, x') = C_0 \exp \left( -0.5 (x - x')^T W (x - x') \right) + \sigma^2 \delta_{x,x'} \]

- Variable lengthscale, spatial deformations \( W(x) \)
  (Schmidt & O’Hagan, 2003)

- Input dependent observation noise \( \sigma(x) \)
  (Goldberg et al., 1998)

- Nonstationary amplitude variations \( C_0(x) \)
  (Turner & Sahani, 2008)

This work
Outline

1 Motivation
   • Gaussian Process Regression
   • Predictions and Hyperparameters
   • Nonstationarity

2 Gaussian Process Product Model

3 Inference
   • Expectation Propagation
   • Hyperparameters
   • Making Predictions

4 Results
The Gaussian Process Product Model
Varying Amplitudes

- Model data as pointwise product of two latent functions to achieve nonstationary amplitude

\[ y_n \sim \mathcal{N}(f(x_n)e^{g(x_n)}, \sigma^2). \]

- Place independent zero-mean Gaussian process priors on \( f(x) \) and \( g(x) \).

- Exponentiation of \( g(x) \) to reduce multimodality -
  - For \( N \) data there would be at least \( 2^N \) modes due to sign flips.
Model data as pointwise product of two latent functions to achieve nonstationary amplitude

\[ y_n \sim \mathcal{N}(f(x_n)e^{g(x_n)}, \sigma^2). \]

Place independent zero-mean Gaussian process priors on \( f(x) \) and \( g(x) \).

Exponentiation of \( g(x) \) to reduce multimodality -

- For \( N \) data there would be at least \( 2^N \) modes due to sign flips.
The Gaussian Process Product Model
Varying Amplitudes

- Model data as pointwise product of two latent functions to achieve nonstationary amplitude
  \[ y_n \sim \mathcal{N}(f(x_n) e^{g(x_n)}, \sigma^2). \]
- Place independent zero-mean Gaussian process priors on \( f(x) \) and \( g(x) \).
- Exponentiation of \( g(x) \) to reduce multimodality -
  - For \( N \) data there would be at least \( 2^N \) modes due to sign flips.
The Gaussian Process Product Model

Convention

\[ f(x) \sim \mathcal{GP}(0, C_f(\cdot, \cdot), \theta_f) \]

with hyperparameters to capture near-stationary variations,

\[ g(x) \sim \mathcal{GP}(0, C_g(\cdot, \cdot), \theta_g) \]

to capture slowly-varying amplitude nonstationarity.
The Gaussian Process Product Model

Samples from the model

Shorter lengthscale \((l_g = 2.0, l_f = 0.25)\)
The Gaussian Process Product Model

Samples from the model

Short lengthscale \((l_g = 2.0, l_f = 0.5 \text{ noisy})\)
Outline

1 Motivation
   - Gaussian Process Regression
   - Predictions and Hyperparameters
   - Nonstationarity

2 Gaussian Process Product Model

3 Inference
   - Expectation Propagation
   - Hyperparameters
   - Making Predictions

4 Results
Joint posterior on $f$ and $g$

- GP Prior on $f$
- GP Prior on $g$

\[
p(f, g \mid D, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \times \prod_{n=1}^{N} \mathcal{N}(y_n; f_n e^{g_n}, \sigma^2)
\]

- Likelihood term for data $y$.
- Intractable posterior.
Joint posterior on $f$ and $g$

- GP Prior on $f$
- GP Prior on $g$

$$p(f, g | D, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \times \prod_{n=1}^{N} \mathcal{N}(y_n; fn e^{gn}, \sigma^2)$$

- Likelihood term for data $y$.
- Intractable posterior.
GPPM posterior

Joint posterior on $f$ and $g$

- GP Prior on $f$
- GP Prior on $g$

$$p(f, g | D, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \times \prod_{n=1}^{N} \mathcal{N}(y_n; f_n e^{g_n}, \sigma^2)$$

- Likelihood term for data $y$.
- Intractable posterior.
GPPM posterior

Joint posterior on $f$ and $g$

- GP Prior on $f$
- GP Prior on $g$

$$p(f, g | D, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \times \prod_{n=1}^{N} \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2)$$

- Likelihood term for data $y$.
- Intractable posterior.
Joint posterior on $f$ and $g$

- GP Prior on $f$
- GP Prior on $g$

$$p(f, g | \mathcal{D}, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \times \prod_{n=1}^{N} \mathcal{N}(y_n; f_n e^{g_n}, \sigma^2)$$

- Likelihood term for data $y$.
- Intractable posterior.
Approximate Inference

Choice of Approximation

- We expect posterior to be near-Gaussian.
- Likelihood factorizes to $N$ independent terms.
- The likelihood introduces nontrivial dependences between $f$ and $g$ such that a factorized approximation is inappropriate.

Expectation Propagation

- A variational approximation.
- Well suited for such factorized likelihoods.
Approximate Inference

**Choice of Approximation**

- We expect posterior to be near-Gaussian.
- Likelihood factorizes to $N$ independent terms.
- The likelihood introduces nontrivial dependences between $f$ and $g$ such that a factorized approximation is inappropriate.

**Expectation Propagation**

- A variational approximation.
- Well suited for such factorized likelihoods.
EP in a nutshell
(Minka, 2001)

- EP approximates a posterior of the form

\[ P(\theta | D) \propto P(\theta) \prod_n p(D_n | \theta), \]

with a tractable alternative with approximate likelihood terms

\[ Q(\theta | D) \propto P(\theta) \prod_n q_n(\theta) \]

- \( q_n(\theta) \) updated iteratively by minimizing a divergence measure

\[
\text{KL} \left[ P(\theta) \prod_{i \neq n} q_i(\theta) \times \underbrace{P(D_n | \theta)}_{\text{exact factor}} \middle| \middle| \underbrace{P(\theta) \prod_{i \neq n} q_i(\theta) \times q_n(\theta)}_{\text{approximation}} \right].
\]

- Equivalent to moment-matching.
EP in a nutshell
(Minka, 2001)

- EP approximates a posterior of the form

\[ P(\theta | D) \propto P(\theta) \prod_n p(D_n|\theta), \]

with a tractable alternative with approximate likelihood terms

\[ Q(\theta|D) \propto P(\theta) \prod_n q_n(\theta) \]

- \( q_n(\theta) \) updated iteratively by minimizing a divergence measure

\[
\text{KL} \left[ P(\theta) \prod_{i \neq n} q_i(\theta) \times P(D_n|\theta) \left|\left| P(\theta) \prod_{i \neq n} q_i(\theta) \times q_n(\theta) \right\right. \right].
\]

- Equivalent to moment-matching.
EP in a nutshell
(Minka, 2001)

- EP approximates a posterior of the form

\[ P(\theta | D) \propto P(\theta) \prod_{n} p(D_n | \theta), \]

with a tractable alternative with approximate likelihood terms

\[ Q(\theta | D) \propto P(\theta) \prod_{n} q_n(\theta) \]

- \( q_n(\theta) \) updated iteratively by minimizing a divergence measure

\[
\text{KL}
\left[
P(\theta) \prod_{i \neq n} q_i(\theta) \times \underbrace{P(D_n | \theta)}_{\text{exact factor}}
\right|
\left|
\underbrace{P(\theta) \prod_{i \neq n} q_i(\theta)}_{\text{approximation}} \times q_n(\theta)
\right] .
\]

- Equivalent to moment-matching.
Approximate posterior for GPPM

\[ q(f, g \mid D, \theta) \propto \mathcal{N}(f; 0, C_f) \times \mathcal{N}(g; 0, C_g) \prod_{n=1}^{N} \tilde{t}_n(f_n, g_n), \]

\( \Sigma_{GP} \) is the joint prior covariance

\[ \Sigma_{GP} = \begin{bmatrix} C_f & 0 \\ 0 & C_g \end{bmatrix}. \]

\( \tilde{t}_n(\cdot, \cdot) \) are local Gaussian approximations of the \( n \)-th likelihood term:

\[ \tilde{t}_n(f_n, g_n) = \mathcal{N}(f_n, g_n; \cdot) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2) \]

Parameters updated by moment-matching.
EP in the GPPM model

- Approximate posterior for GPPM

\[
q(f, g \mid D, \theta) \propto \mathcal{N}(0, \Sigma_{GP}) \prod_{n=1}^{N} \tilde{t}_n(f_n, g_n),
\]

- \(\Sigma_{GP}\) is the joint prior covariance

\[
\Sigma_{GP} = \begin{bmatrix}
C_f & 0 \\
0 & C_g
\end{bmatrix}.
\]

- \(\tilde{t}_n(\cdot, \cdot)\) are local Gaussian approximations of the \(n\)-th likelihood term:

\[
\tilde{t}_n(f_n, g_n) = \mathcal{N}(f_n, g_n; ) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2)
\]

- Parameters updated by moment-matching.
EP in the GPPM model

- Approximate posterior for GPPM

\[ q(f, g \mid \mathcal{D}, \theta) \propto \mathcal{N}(0, \Sigma_{\text{GP}}) \prod_{n=1}^{N} \tilde{t}_n(f_n, g_n), \]

- \( \Sigma_{\text{GP}} \) is the joint prior covariance

\[ \Sigma_{\text{GP}} = \begin{bmatrix} C_f & 0 \\ 0 & C_g \end{bmatrix}. \]

- \( \tilde{t}_n(\cdot, \cdot) \) are local Gaussian approximations of the \( n \)-th likelihood term:

\[ \tilde{t}_n(f_n, g_n) = Z_n \mathcal{N}(f_n, g_n; \mu_n, \Sigma_n) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2) \]

- Parameters updated by moment-matching.
EP in the GPPM model

- Approximate posterior for GPPM

\[ q(f, g \mid D, \theta) \propto \mathcal{N}(0, \Sigma_{GP}) \prod_{n=1}^{N} \tilde{t}_n(f_n, g_n), \]

- \( \Sigma_{GP} \) is the joint prior covariance

\[ \Sigma_{GP} = \begin{bmatrix} C_f & 0 \\ 0 & C_g \end{bmatrix}. \]

- \( \tilde{t}_n(\cdot, \cdot) \) are local Gaussian approximations of the \( n \)-th likelihood term:

\[ \tilde{t}_n(f_n, g_n) = \mathcal{Z}_n \mathcal{N}(f_n, g_n; \mu_n, \Sigma_n) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2) \]

- Parameters updated by moment-matching.
EP in the GPPM model

Covariance structure of the approximation

\[ \text{Prior} \times \text{Approximate Likelihood} = \text{Global Approximation} \]

A few more details

- Calculation of moments not tractable.
  - We use 2D Gaussian quadrature for numerical moment calculation.
- Approximately 10 EP iterations are sufficient.
- Scheme is practical up to about 1,000 data points.
EP in the GPPM model

Covariance structure of the approximation

\[
\begin{bmatrix}
C_f & \times \\
\times & C_g
\end{bmatrix} = \begin{bmatrix}
& \\
& \\
\end{bmatrix}
\]

Prior  Approximate Likelihood  Global Approximation

A few more details

- Calculation of moments not tractable.
  - We use 2D Gaussian quadrature for numerical moment calculation.
  - Approximately 10 EP iterations are sufficient.
- Scheme is practical up to about 1,000 data points.
We have to choose hyperparameters for two latent GPs and the likelihood: lengthscales, observation noise.

Ideally use the marginal likelihood $Z = P(\theta|\mathcal{D})$.

EP provides an approximation:

$$
\mathcal{N}(f_n, g_n; \mu_n, \Sigma_n) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2)
$$

"zeroth moment".
EP in the GPPM model

Optimizing Hyperparameters

- We have to choose hyperparameters for two latent GPs and the likelihood: lengthscales, observation noise.
- Ideally use the marginal likelihood $Z = P(\theta|\mathcal{D})$.
- EP provides an approximation:
  \[
  Z_n \mathcal{N}(f_n, g_n; \mu_n, \Sigma_n) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2)
  \]

- “zeroth moment“.
We have to choose hyperparameters for two latent GPs and the likelihood: lengthscales, observation noise.

Ideally use the marginal likelihood $Z = P(\theta|D)$.

EP provides an approximation:

$$Z_n \mathcal{N}(f_n, g_n; \mu_n, \Sigma_n) \approx \mathcal{N}(y_n; f_ne^{g_n}, \sigma^2)$$

“zeroth moment“.
EP in the GPPM model
Optimizing Hyperparameters

- **Global Approximation**
  \[
  \ln Z_{EP} = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mu^T \Sigma^{-1} \mu \\
  - \frac{1}{2} \ln |\Sigma_{GP}| - \frac{1}{2} \tilde{\mu}^T \tilde{\Sigma}^{-1} \tilde{\mu} + \sum_{n=1}^{N} \ln \tilde{Z}_n
  \]

- **GP prior**
- **Local (likelihood) approximations**
- **“Zeroth moments”**
EP in the GPPM model
Optimizing Hyperparameters

- Global Approximation

\[
\ln Z_{EP} = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mu^T \Sigma^{-1} \mu
\]

- GP prior
- Local (likelihood) approximations
- “Zeroth moments”
EP in the GPPM model
Optimizing Hyperparameters

- Global Approximation

\[
\ln Z_{EP} = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mu^T \Sigma^{-1} \mu
\]

- GP prior

- Local (likelihood) approximations

  "Zeroth moments"
EP in the GPPM model
Optimizing Hyperparameters

- Global Approximation

\[
\ln Z_{EP} = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mu^T \Sigma^{-1} \mu
\]

- GP prior

- Local (likelihood) approximations

- “Zeroth moments”
EP in the GPPM model

Making predictions

- Joint predictive distribution on $p(f^*, g^*|x^*)$ from EP approximation.
- Given $\mathcal{N}(f^*, g^*; \mu^*, \Sigma^*)$, two approximations for the predictive distribution on $y^*$:
  - Mixture of Gaussians
    - Generate samples from $g^*$
    - Use conditional distribution on $f^*$ to create a mixture of Gaussians
      $$p(y^*|x^*, \mathcal{D}) \approx \sum_i \mathcal{N}(y^*; \mu_{f|g_i}^* e^{g_i^*}, v_{f|g_i}^* e^{2g_i^*}).$$
    - Appropriately heavy-tailed
  - Linearization
    - Linearize around the mean.
    - $p(y^*)$ is Gaussian again.
EP in the GPPM model
Making predictions

- Joint predictive distribution on \( p(f^*, g^*|x^*) \) from EP approximation.
- Given \( \mathcal{N}(f^*, g^*; \mu^*, \Sigma^*) \), two approximations for the predictive distribution on \( y^* \):
  - Mixture of Gaussians
    - Generate samples from \( g^* \)
    - Use conditional distribution on \( f^* \) to create a mixture of Gaussians
      \[
p(y^* | x^*, \mathcal{D}) \approx \sum_i \mathcal{N}(y^*; \mu_{f|g_i}^* e^{g_i^*}, v_{f|g_i}^* e^{2g_i^*}).\]
    - Appropriately heavy-tailed
  - Linearization
    - Linearize around the mean.
    - \( p(y^*) \) is Gaussian again.
Joint predictive distribution on $p(f^*, g^*|x^*)$ from EP approximation.

Given $\mathcal{N}(f^*, g^*; \mu^*, \Sigma^*)$, two approximations for the predictive distribution on $y^*$:

- Mixture of Gaussians:
  - Generate samples from $g^*$
  - Use conditional distribution on $f^*$ to create a mixture of Gaussians

$$p(y^* | x^*, D) \approx \sum_i \mathcal{N}(y^*; \mu_f^*|g_i, v_f^*|g_i e^{2g_i^*}).$$

- Appropriately heavy-tailed

- Linearization:
  - Linearize around the mean.
  - $p(y^*)$ is Gaussian again.
Outline

1 Motivation
   • Gaussian Process Regression
   • Predictions and Hyperparameters
   • Nonstationarity

2 Gaussian Process Product Model

3 Inference
   • Expectation Propagation
   • Hyperparameters
   • Making Predictions

4 Results
Results

- Comparison of 3 models
  - Vanilla GP
  - Sparse Gaussian Process; pseudo inputs (Snelson & Ghahramani, 2006)
  - GPPM

- Evaluation on 3 datasets
  - Motorcycle helmet data
  - SP500 Log Daily Returns
  - Heart rate data

- Hyperparameters
  - GPPM: optimization via grid search
  - SPGP, GP: ML-II
Motorcycle Helmet Data
Classical example of nonstationarity
Motorcycle Helmet Data
Classical example of nonstationarity
SP500 Log Daily Returns

MLP

MSE

Fraction of missing Data

vanilla GP  SPGP  GPPM
Heart Rate Data

EP Posterior $g(x)$

EP Posterior $f(x) \cdot \exp(g(x))$

mean HR
Heart Rate Data

\begin{figure}
\centering
\includegraphics[width=\textwidth]{heart_rate_data.png}
\caption{Graph showing the comparison of different models (MLP, vanilla GP, SPGP, GPPM) for heart rate data with varying fractions of missing data (0.1 to 0.6). The top graph shows the EP Posterior $g(x)$, and the bottom graph shows the EP Posterior $f(x) \cdot \exp(g(x))$ with mean HR.}
\end{figure}
Summary

- GPPM provides a principled way for GP regression learning smoothly varying nonstationary amplitude modulations.
- Expectation Propagation to achieve efficient inference in this model.

**Future work:** refine quadrature-EP to enable gradient based hyperparameter optimization.

Thanks to

- David MacKay for helpful comments.
- Cambridge Gates Trust for funding.
Summary

- GPPM provides a principled way for GP regression learning smoothly varying nonstationary amplitude modulations.
- Expectation Propagation to achieve efficient inference in this model.

Future work: refine quadrature-EP to enable gradient based hyperparameter optimization.

Thanks to
- David MacKay for helpful comments.
- Cambridge Gates Trust for funding.
References


EP in the GPPM model
Making predictions

- Find predictive distribution on output $y^*$ for unseen input $x^*$.
- EP posterior on $f, g$ yields joint Gaussian prediction for latent functions $p(f^*, g^* | D, x^*) = \mathcal{N}(\mu^*, \Sigma^*)$

$$
\mu^* = K^T \left( \Sigma_{GP} + \tilde{\Sigma} \right)^{-1} \tilde{\mu}
$$

$$
\Sigma^* = \kappa - K^T \left( \Sigma_{GP} + \tilde{\Sigma} \right)^{-1} K
$$

- GP Prior
- Local (likelihood) approximations