

Gaussian Process Product Models for Nonparametric Nonstationarity

Ryan Prescott Adams and **Oliver Stegle**

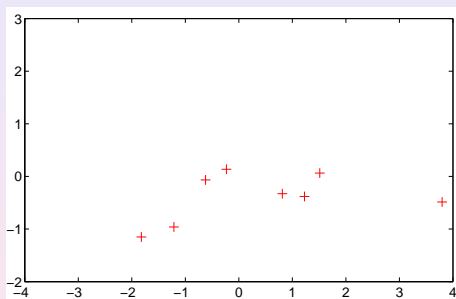
Inference Group
Cavendish Laboratory
University of Cambridge

ICML July 2008



Gaussian Processes for Regression

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

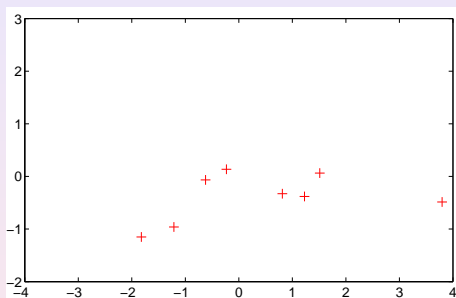


- Gaussian process is a prior over functions $f(\mathbf{x})$

$$p(f(\mathbf{x})|\mathcal{D}) = \frac{p(\mathcal{D}|f(\mathbf{x}))p(f(\mathbf{x}))}{p(\mathcal{D})}$$

Gaussian Processes for Regression

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

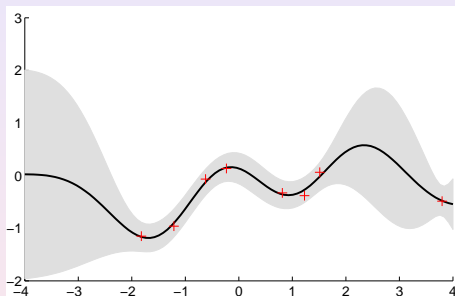


- Gaussian process is a prior over functions $f(\mathbf{x})$

$$p(f(\mathbf{x})|\mathcal{D}) = \frac{p(\mathcal{D}|f(\mathbf{x}))p(f(\mathbf{x}))}{p(\mathcal{D})}.$$

Gaussian Processes for Regression

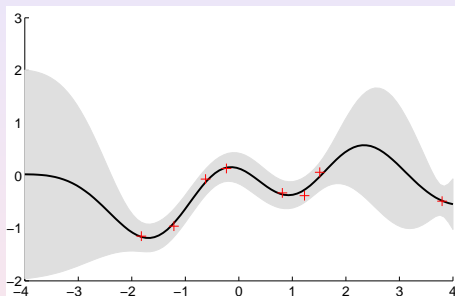
- Gaussian process marginal posterior mean and errorbars:



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

Gaussian Processes for Regression

- Gaussian process marginal posterior mean and errorbars:



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

Gaussian Processes for Regression

Predictions and Hyperparameters

- For a “vanilla GP” the marginal predictive distribution for an unseen input \mathbf{x}^* is Gaussian

$$\begin{aligned}p(y^*|\mathbf{x}^*, \mathcal{D}) &= \mathcal{N}(\mu^*, v^*), \\ \mu^* &= \mathbf{k}_N^\top \mathbf{C}_N^{-1} \mathbf{y}_N \\ v^* &= C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_N^\top \mathbf{C}_N^{-1} \mathbf{k}_N\end{aligned}$$

- Hyperparameters θ can be optimized via LML

$$\mathcal{L}(\theta) = -\frac{1}{2} \ln |\mathbf{C}_N(\theta)| - \frac{1}{2} \mathbf{y}_N^\top \mathbf{C}_N^{-1}(\theta) \mathbf{y}_N - \frac{N}{2} \ln 2\pi.$$

Gaussian Processes for Regression

Predictions and Hyperparameters

- For a “vanilla GP” the marginal predictive distribution for an unseen input \mathbf{x}^* is Gaussian

$$\begin{aligned}p(y^*|\mathbf{x}^*, \mathcal{D}) &= \mathcal{N}(\mu^*, v^*), \\ \mu^* &= \mathbf{k}_N^\top \mathbf{C}_N^{-1} \mathbf{y}_N \\ v^* &= C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_N^\top \mathbf{C}_N^{-1} \mathbf{k}_N\end{aligned}$$

- Hyperparameters $\boldsymbol{\theta}$ can be optimized via LML

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2} \ln |\mathbf{C}_N(\boldsymbol{\theta})| - \frac{1}{2} \mathbf{y}_N^\top \mathbf{C}_N^{-1}(\boldsymbol{\theta}) \mathbf{y}_N - \frac{N}{2} \ln 2\pi.$$

Gaussian Processes for Regression

The Covariance Function

- Covariance function $C(\mathbf{x}, \mathbf{x}')$ to model how targets at inputs \mathbf{x}, \mathbf{x}' co-vary.
- A popular choice

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

with hyperparameters C_0, l, σ

- Covariance functions that only depend on the distance between inputs are called **stationary**.

Gaussian Processes for Regression

The Covariance Function

- Covariance function $C(\mathbf{x}, \mathbf{x}')$ to model how targets at inputs \mathbf{x}, \mathbf{x}' co-vary.
- A popular choice

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

with hyperparameters C_0, l, σ

- 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
 - This is a nontrivial 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

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 $C(\mathbf{x}, \mathbf{x}')$ 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 $C(\mathbf{x}, \mathbf{x}')$ 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 $C(\mathbf{x}, \mathbf{x}')$ 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

$$C(\mathbf{x}, \mathbf{x}') = C_0 \exp\left(-0.5(\mathbf{x} - \mathbf{x}')^\top \mathbf{W}(\mathbf{x} - \mathbf{x}')\right) + \sigma^2 \delta_{\mathbf{x}, \mathbf{x}'}$$

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

This work

Latent space extensions of stationary covariances

$$C(\mathbf{x}, \mathbf{x}') = C_0 \exp\left(-0.5(\mathbf{x} - \mathbf{x}')^\top \mathbf{W}(\mathbf{x} - \mathbf{x}')\right) + \sigma^2 \delta_{\mathbf{x}, \mathbf{x}'}$$

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

This work

Latent space extensions of stationary covariances

$$C(\mathbf{x}, \mathbf{x}') = C_0 \exp\left(-0.5 (\mathbf{x} - \mathbf{x}')^\top \mathbf{W} (\mathbf{x} - \mathbf{x}')\right) + \sigma^2 \delta_{\mathbf{x}, \mathbf{x}'}$$

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

Latent space extensions of stationary covariances

$$C(\mathbf{x}, \mathbf{x}') = C_0 \exp\left(-0.5(\mathbf{x} - \mathbf{x}')^\top \mathbf{W}(\mathbf{x} - \mathbf{x}')\right) + \sigma^2 \delta_{\mathbf{x}, \mathbf{x}'}$$

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

This work

Outline

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

$$y_n \sim \mathcal{N}(f(\mathbf{x}_n)e^{g(\mathbf{x}_n)}, \sigma^2).$$

- Place independent zero-mean Gaussian process priors on $f(\mathbf{x})$ and $g(\mathbf{x})$.
- Exponentiation of $g(\mathbf{x})$ 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

$$y_n \sim \mathcal{N}(f(\mathbf{x}_n)e^{g(\mathbf{x}_n)}, \sigma^2).$$

- Place independent zero-mean Gaussian process priors on $f(\mathbf{x})$ and $g(\mathbf{x})$.
- Exponentiation of $g(\mathbf{x})$ 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

$$y_n \sim \mathcal{N}(f(\mathbf{x}_n)e^{g(\mathbf{x}_n)}, \sigma^2).$$

- Place independent zero-mean Gaussian process priors on $f(\mathbf{x})$ and $g(\mathbf{x})$.
- Exponentiation of $g(\mathbf{x})$ to reduce multimodality -
 - ▶ For N data there would be at least 2^N modes due to sign flips.

The Gaussian Process Product Model

- Convention

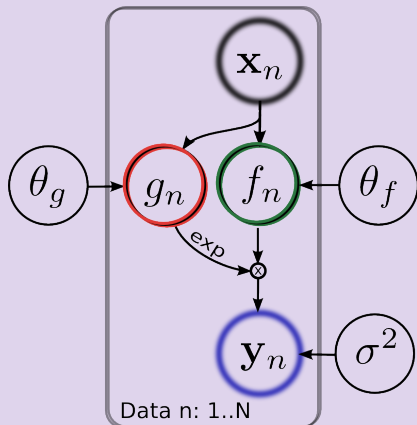
$$f(\mathbf{x}) \sim \mathcal{GP}(0, C_f(\cdot, \cdot), \boldsymbol{\theta}_f)$$

with hyperparameters to capture near-stationary variations,

$$g(\mathbf{x}) \sim \mathcal{GP}(0, C_g(\cdot, \cdot), \boldsymbol{\theta}_g)$$

to capture slowly-varying amplitude nonstationarity.

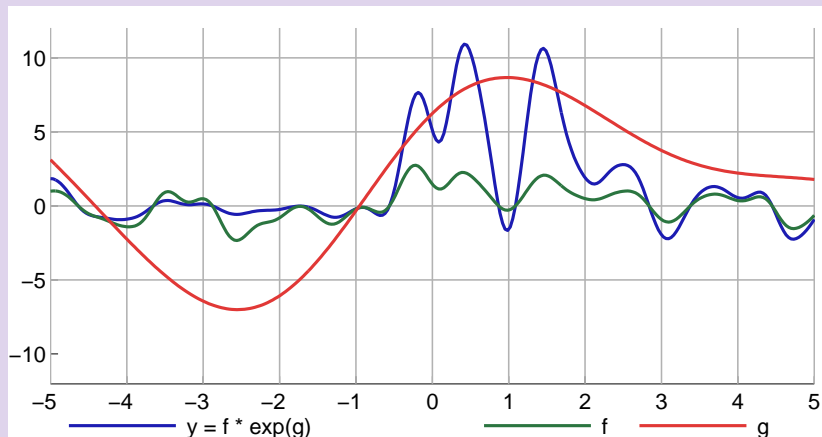
Graphical model



The Gaussian Process Product Model

Samples from the model

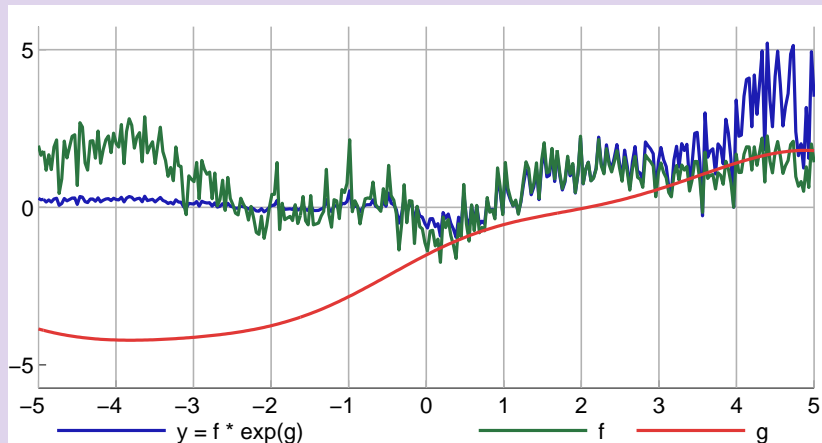
Shorter lengthscale ($l_g = 2.0, l_f = 0.25$)



The Gaussian Process Product Model

Samples from the model

Short lengthscale ($l_g = 2.0, l_f = 0.5$ *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 \mathbf{f} and \mathbf{g}

- GP Prior on \mathbf{f}
- GP Prior on \mathbf{g}

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

- Likelihood term for data \mathbf{y} .
- Intractable posterior.

Joint posterior on \mathbf{f} and \mathbf{g}


- GP Prior on \mathbf{f}
- GP Prior on \mathbf{g}

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

- Likelihood term for data \mathbf{y} .
- Intractable posterior.

Joint posterior on \mathbf{f} and \mathbf{g}

- GP Prior on \mathbf{f}
- GP Prior on \mathbf{g}

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


- Likelihood term for data \mathbf{y} .
- Intractable posterior.

Joint posterior on \mathbf{f} and \mathbf{g}

- GP Prior on \mathbf{f}
- GP Prior on \mathbf{g}

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

- Likelihood term for data \mathbf{y} .
- Intractable posterior.

Joint posterior on \mathbf{f} and \mathbf{g}

- GP Prior on \mathbf{f}
- GP Prior on \mathbf{g}

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

- Likelihood term for data \mathbf{y} .
- Intractable posterior.

Choice of Approximation

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

Expectation Propagation

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

Choice of Approximation

- We expect posterior to be near-Gaussian.
- Likelihood factorizes to N independent terms.
- The likelihood introduces nontrivial dependences between \mathbf{f} and \mathbf{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

$$P(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n p(\mathcal{D}_n | \boldsymbol{\theta}),$$

with a tractable alternative with approximate likelihood terms

$$Q(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n q_n(\boldsymbol{\theta})$$

- $q_n(\boldsymbol{\theta})$ updated iteratively by minimizing a divergence measure

$$\text{KL} \left[P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \overbrace{P(\mathcal{D}_n | \boldsymbol{\theta})}^{\text{exact factor}} \parallel P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \underbrace{q_n(\boldsymbol{\theta})}_{\text{approximation}} \right].$$

- Equivalent to moment-matching.

EP in a nutshell

(Minka, 2001)

- EP approximates a posterior of the form

$$P(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n p(\mathcal{D}_n | \boldsymbol{\theta}),$$

with a tractable alternative with approximate likelihood terms

$$Q(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n q_n(\boldsymbol{\theta})$$

- $q_n(\boldsymbol{\theta})$ updated iteratively by minimizing a divergence measure

$$\text{KL} \left[P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \overbrace{P(\mathcal{D}_n | \boldsymbol{\theta})}^{\text{exact factor}} \parallel P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \underbrace{q_n(\boldsymbol{\theta})}_{\text{approximation}} \right].$$

- Equivalent to moment-matching.

EP in a nutshell

(Minka, 2001)

- EP approximates a posterior of the form

$$P(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n p(\mathcal{D}_n | \boldsymbol{\theta}),$$

with a tractable alternative with approximate likelihood terms

$$Q(\boldsymbol{\theta} | \mathcal{D}) \propto P(\boldsymbol{\theta}) \prod_n q_n(\boldsymbol{\theta})$$

- $q_n(\boldsymbol{\theta})$ updated iteratively by minimizing a divergence measure

$$\text{KL} \left[P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \overbrace{P(\mathcal{D}_n | \boldsymbol{\theta})}^{\text{exact factor}} \parallel P(\boldsymbol{\theta}) \prod_{i \neq n} q_i(\boldsymbol{\theta}) \times \underbrace{q_n(\boldsymbol{\theta})}_{\text{approximation}} \right].$$

- Equivalent to moment-matching.

EP in the GPPM model

- Approximate posterior for GPPM

$$q(\mathbf{f}, \mathbf{g} | \mathcal{D}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{C}_f) \times \mathcal{N}(\mathbf{g}; \mathbf{0}, \mathbf{C}_g) \prod_{n=1}^N \tilde{t}_n(f_n, g_n),$$

- Σ_{GP} is the joint prior covariance

$$\Sigma_{\text{GP}} = \begin{bmatrix} \mathbf{C}_f & 0 \\ 0 & \mathbf{C}_g \end{bmatrix}.$$

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

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

► Parameters updated by moment-matching.

EP in the GPPM model

- Approximate posterior for GPPM

$$q(\mathbf{f}, \mathbf{g} \mid \mathcal{D}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{GP}}) \prod_{n=1}^N \tilde{t}_n(f_n, g_n),$$

- $\boldsymbol{\Sigma}_{\text{GP}}$ is the joint prior covariance

$$\boldsymbol{\Sigma}_{\text{GP}} = \begin{bmatrix} \mathbf{C}_f & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_g \end{bmatrix}.$$

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

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

► Parameters updated by moment-matching.

EP in the GPPM model

- Approximate posterior for GPPM

$$q(\mathbf{f}, \mathbf{g} \mid \mathcal{D}, \boldsymbol{\theta}) \propto \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{GP}}) \prod_{n=1}^N \tilde{t}_n(f_n, g_n),$$

- $\boldsymbol{\Sigma}_{\text{GP}}$ is the joint prior covariance

$$\boldsymbol{\Sigma}_{\text{GP}} = \begin{bmatrix} \mathbf{C}_f & 0 \\ 0 & \mathbf{C}_g \end{bmatrix}.$$

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

$$\tilde{t}_n(f_n, g_n) = Z_n \mathcal{N}(f_n, g_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \approx \mathcal{N}(y_n; f_n e^{g_n}, \sigma^2)$$

► Parameters updated by moment-matching.

EP in the GPPM model

- Approximate posterior for GPPM

$$q(\mathbf{f}, \mathbf{g} \mid \mathcal{D}, \boldsymbol{\theta}) \propto \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{GP}}) \prod_{n=1}^N \tilde{t}_n(f_n, g_n),$$

- $\boldsymbol{\Sigma}_{\text{GP}}$ is the joint prior covariance

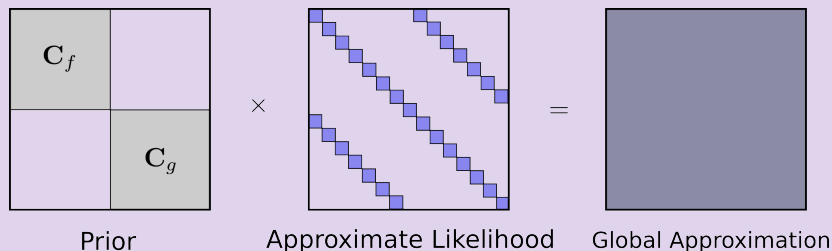
$$\boldsymbol{\Sigma}_{\text{GP}} = \begin{bmatrix} \mathbf{C}_f & 0 \\ 0 & \mathbf{C}_g \end{bmatrix}.$$

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

$$\tilde{t}_n(f_n, g_n) = Z_n \mathcal{N}(f_n, g_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \approx \mathcal{N}(y_n; f_n e^{g_n}, \sigma^2)$$

- ▶ Parameters updated by moment-matching.

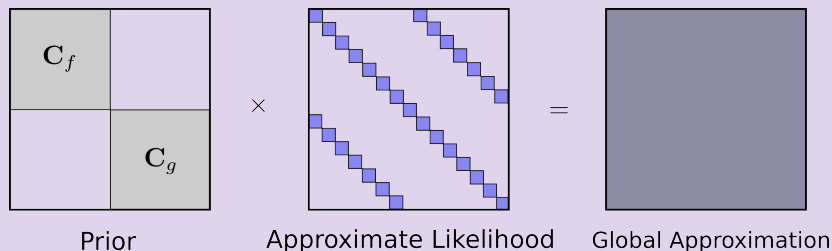
Covariance structure of the approximation



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.

Covariance structure of the 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

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:

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

- “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:

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

- “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:

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

- “zeroth moment”.

EP in the GPPM model

Optimizing Hyperparameters

- Global Approximation

$$\ln Z_{EP} = \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \mu^\top \Sigma^{-1} \mu$$
$$- \frac{1}{2} \ln |\Sigma_{GP}| - \frac{1}{2} \tilde{\mu}^\top \tilde{\Sigma}^{-1} \tilde{\mu} + \sum_{n=1}^N \ln \tilde{Z}_n$$

- GP prior
- Local (likelihood) approximations
- “Zeroth moments”

EP in the GPPM model

Optimizing Hyperparameters

- Global Approximation

$$\ln Z_{EP} = \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

$$- \frac{1}{2} \ln |\boldsymbol{\Sigma}_{GP}| - \frac{1}{2} \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\mu}} + \sum_{n=1}^N \ln \tilde{Z}_n$$

- GP prior
 - Local (likelihood) approximations
 - “Zeroth moments”

EP in the GPPM model

Optimizing Hyperparameters

- Global Approximation

$$\ln Z_{EP} = \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

$$- \frac{1}{2} \ln |\boldsymbol{\Sigma}_{GP}| - \frac{1}{2} \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\mu}} + \sum_{n=1}^N \ln \tilde{Z}_n$$

- GP prior
- Local (likelihood) approximations
- “Zeroth moments”

EP in the GPPM model

Optimizing Hyperparameters

- Global Approximation

$$\ln Z_{EP} = \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

$$- \frac{1}{2} \ln |\boldsymbol{\Sigma}_{GP}| - \frac{1}{2} \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\mu}} + \sum_{n=1}^N \ln \tilde{Z}_n$$

- GP prior
- Local (likelihood) approximations
- “Zeroth moments”

EP in the GPPM model

Making predictions

- Joint predictive distribution on $p(f^*, g^* | \mathbf{x}^*)$ from EP approximation.
- Given $\mathcal{N}(f^*, g^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, 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

$$p(y^* | \mathbf{x}^*, \mathcal{D}) \approx \sum_i \mathcal{N}(y^*; \mu_{f|g_i}^* e^{g_i^*}, v_{f|g_i}^* e^{2g_i^*}).$$

- ▶ Appropriately heavy-tailed
- Linearization
 - ▶ Linearize around the mean.
 - ▶ $p(y^*)$ is Gaussian again.

EP in the GPPM model

Making predictions

- Joint predictive distribution on $p(f^*, g^* | \mathbf{x}^*)$ from EP approximation.
- Given $\mathcal{N}(f^*, g^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, 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

$$p(y^* | \mathbf{x}^*, \mathcal{D}) \approx \sum_i \mathcal{N}(y^*; \mu_{f|g_i}^* e^{g_i^*}, v_{f|g_i}^* e^{2g_i^*}).$$

- ▶ Appropriately heavy-tailed
- Linearization
 - ▶ Linearize around the mean.
 - ▶ $p(y^*)$ is Gaussian again.

EP in the GPMM model

Making predictions

- Joint predictive distribution on $p(f^*, g^* | \mathbf{x}^*)$ from EP approximation.
- Given $\mathcal{N}(f^*, g^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, 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

$$p(y^* | \mathbf{x}^*, \mathcal{D}) \approx \sum_i \mathcal{N}(y^*; \mu_{f|g_i}^* e^{g_i^*}, v_{f|g_i}^* e^{2g_i^*}).$$

- ▶ Appropriately heavy-tailed
- Linearization
 - ▶ Linearize around the mean.
 - ▶ $p(y^*)$ is Gaussian again.

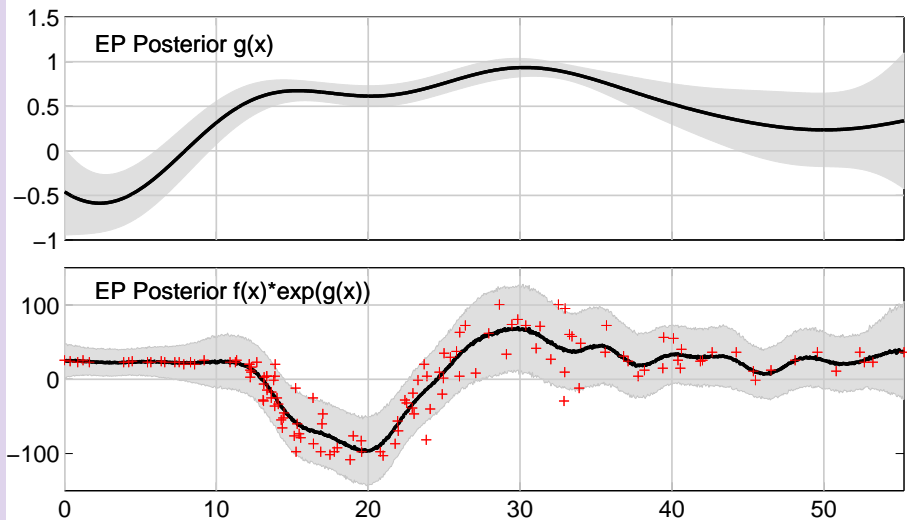
Outline

- 1 Motivation
 - Gaussian Process Regression
 - Predictions and Hyperparameters
 - Nonstationarity
- 2 Gaussian Process Product Model
- 3 Inference
 - Expectation Propagation
 - Hyperparameters
 - Making Predictions
- 4 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

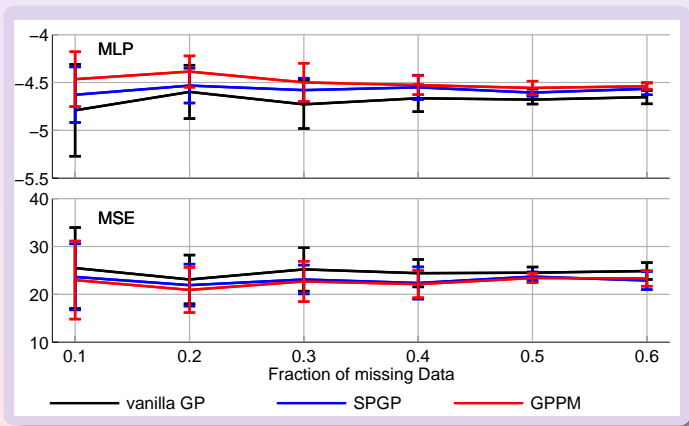
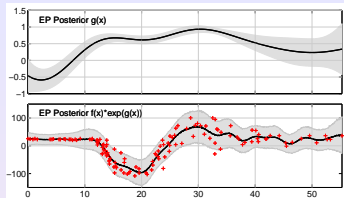
Motorcycle Helmet Data

Classical example of nonstationarity

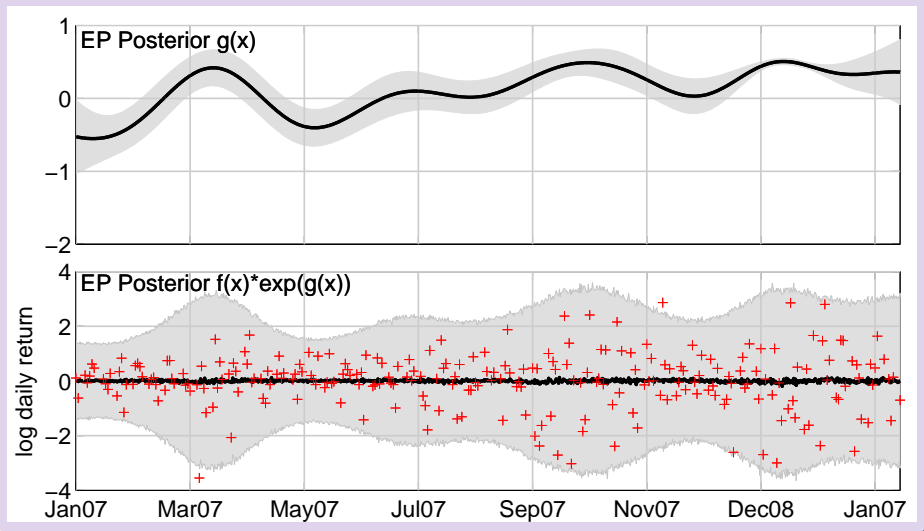


Motorcycle Helmet Data

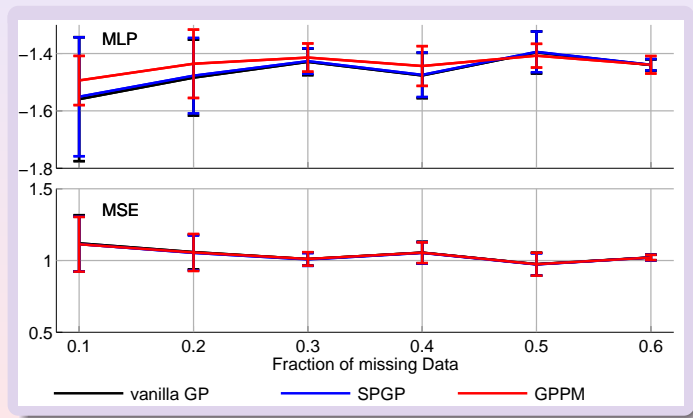
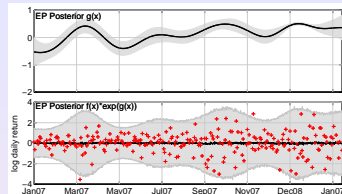
Classical example of nonstationarity



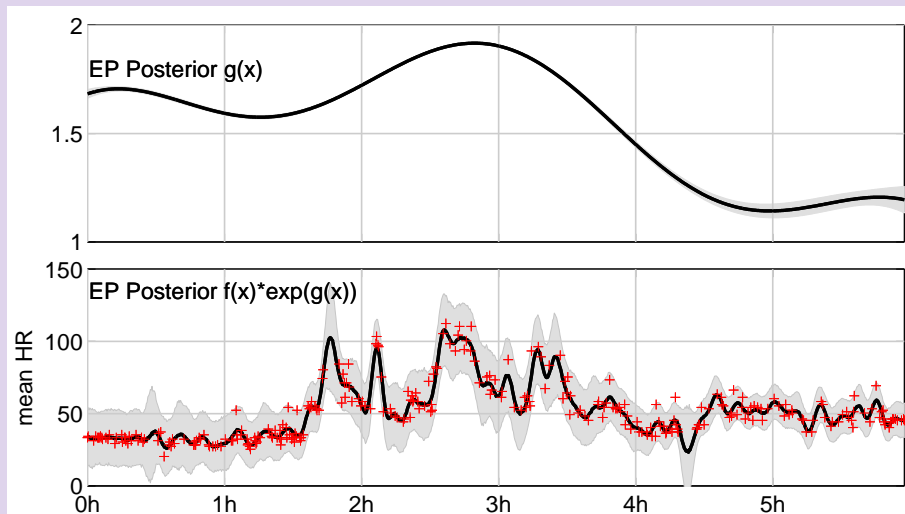
SP500 Log Daily Returns



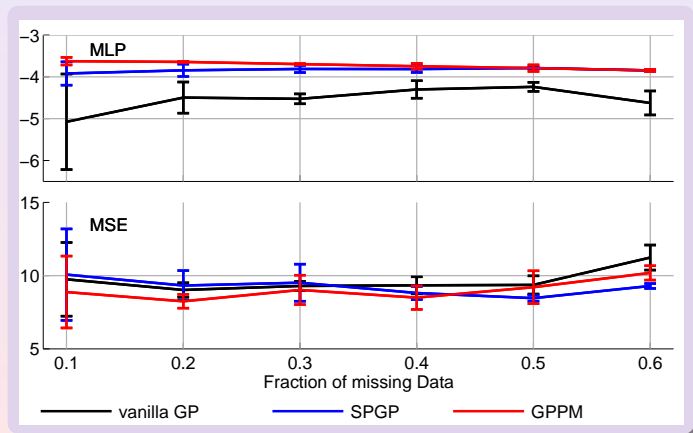
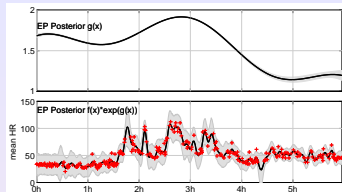
SP500 Log Daily Returns



Heart Rate Data



Heart Rate Data



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

- Goldberg, P., Williams, C., & Bishop, C. (1998). Regression with input-dependent noise: a Gaussian process treatment. *Advances in Neural Processing Systems 10* (pp. 493–499). Cambridge, MA: MIT Press.
- Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Schmidt, A. M., & O'Hagan, A. (2003). Bayesian inference for nonstationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society, Series B*, 65, 745–758.
- Snelson, E., & Ghahramani, Z. (2006). Variable noise and dimensionality reduction for sparse Gaussian processes. *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence*.
- Turner, R., & Sahani, M. (2008). Modeling natural sounds with modulation cascade processes. *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press.

EP in the GPMM model

Making predictions

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

$$\boldsymbol{\mu}^* = \mathbf{K}^\top \left(\boldsymbol{\Sigma}_{GP} + \tilde{\boldsymbol{\Sigma}} \right)^{-1} \tilde{\boldsymbol{\mu}}$$
$$\boldsymbol{\Sigma}^* = \boldsymbol{\kappa} - \mathbf{K}^\top \left(\boldsymbol{\Sigma}_{GP} + \tilde{\boldsymbol{\Sigma}} \right)^{-1} \mathbf{K}$$

- GP Prior
- Local (likelihood) approximations