Fast Laplace Approximation for Gaussian Processes with a Tensor Product Kernel

Perry Groot[a] Markus Peters[b]
Tom Heskes[a] Wolgang Ketter[b]

Radboud University Nijmegen[a]

Erasmus University Rotterdam[b]

Introduction

Gaussian process models

+ rich, principled Bayesian framework
- scalability $\mathcal{O}(N^3)$

Approaches addressing scalability

- approximations / subset selection $\mathcal{O}(M^2 N)$
- exploit additional structure
  - Tensor product kernels
    + Efficient use of Kronecker products on grid data
    - Limited to standard regression, since
      if $\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$
      then $(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y} = \boldsymbol{Q}(\boldsymbol{\Lambda} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Q}^T\boldsymbol{y}$.

Gaussian Processes

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

A Gaussian process (GP) is collection of random variables with the property that the joint distribution of any finite subset is a Gaussian.

A GP specifies a probability distribution over functions $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$ and is fully specified by its mean function $m(\boldsymbol{x})$ and covariance (or kernel) function $k(\boldsymbol{x}, \boldsymbol{x}')$.

Typically $m(\boldsymbol{x}) = \mathbf{0}$, which gives

$$\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_l)\} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}) \text{ with } K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

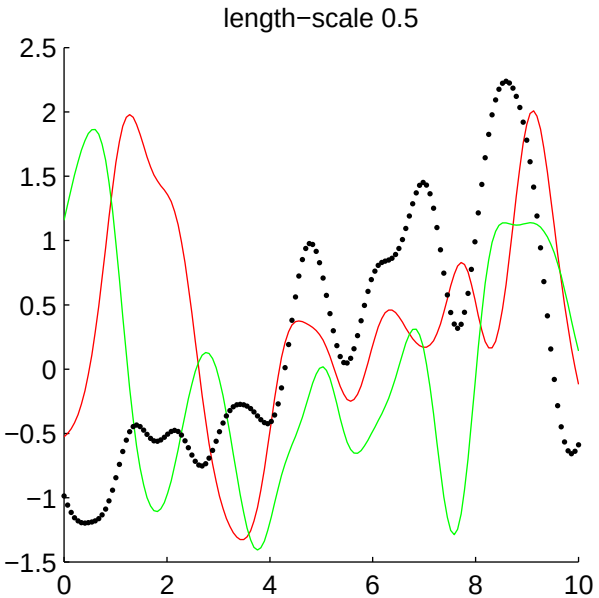
Gaussian Processes - Covariance function

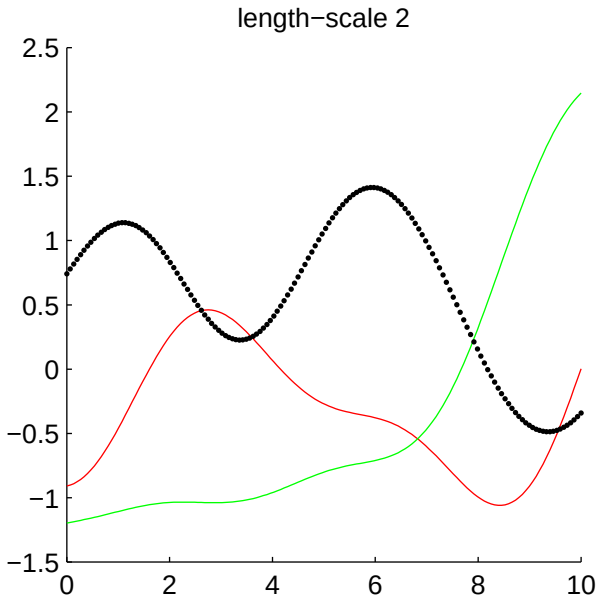
Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Squared exponential (or Gaussian) covariance function:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2\theta^2}\sum_{n=1}^{N}(x_n - x'_n)^2\right)$$

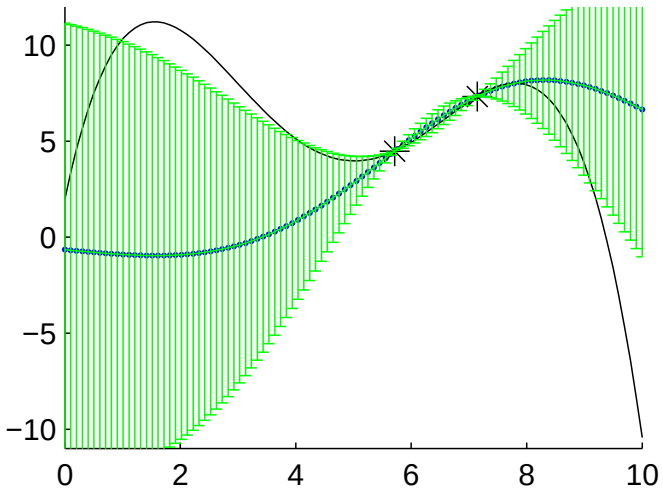
where θ is a length-scale parameter denoting how quickly the functions are to vary.

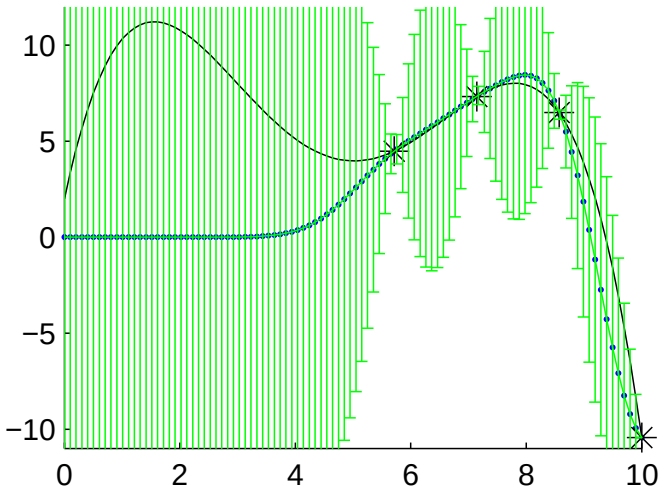


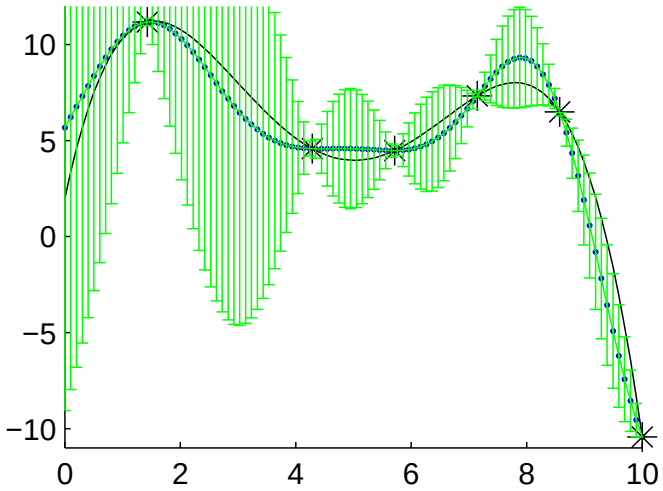


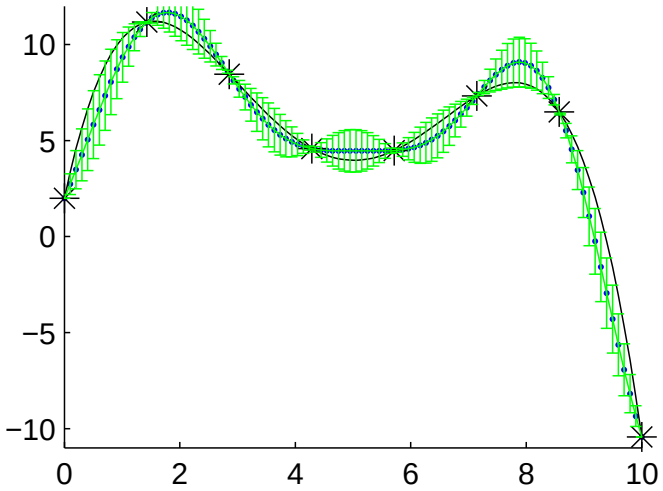
Gaussian Processes - 1D demo

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter









Laplace Approximation

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

For Gaussian process models with non-gaussian likelihood models we need approximations.

Laplace: approximate true posterior $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})$ with a Gaussian $q(\boldsymbol{f})$ centered on the mode of the posterior:

$$q(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|\hat{\mathbf{f}},(\boldsymbol{K}^{-1}+\boldsymbol{W})^{-1}) \tag{1}$$

with $\hat{\mathbf{f}} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\theta})$ and $\boldsymbol{W} = -\nabla\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f})|_{\boldsymbol{f}=\hat{\mathbf{f}}}$.

Kronecker product

Assumptions

- Tensor product kernel function: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_{d=1}^{D} k_d(\boldsymbol{x}_i^d, \boldsymbol{x}_j^d)$.
- data on multi-dimensional grid

This results in a kernel matrix that decomposes into a Kronecker product of matrices of lower dimensions.

$$\boldsymbol{K} = \boldsymbol{K}_1 \otimes \cdots \otimes \boldsymbol{K}_D$$

where

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}$$

Kronecker product

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

The Kronecker product has a convenient algebra

$$(\boldsymbol{A} \otimes \boldsymbol{B})\mathrm{vec}(\boldsymbol{X}) = \mathrm{vec}(\boldsymbol{B}\boldsymbol{X}\boldsymbol{A}^T)$$
$$\boldsymbol{A}\boldsymbol{B} \otimes \boldsymbol{C}\boldsymbol{D} = (\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{B} \otimes \boldsymbol{D})$$
$$(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$$

Operation	**Standard**	**Kronecker**
Storage	$\mathcal{O}(N^2)$	$\mathcal{O}(\sum_{d=1}^{D} N_d^2)$
Matrix vector product	$\mathcal{O}(N^2)$	$\mathcal{O}(N(\sum_{d=1}^{D} N_d))$
Cholesky / SVD	$\mathcal{O}(N^3)$	$\mathcal{O}(\sum_{d=1}^{D} N_d^3)$

MAP estimation

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Let $\boldsymbol{b} = \boldsymbol{W}\boldsymbol{f} + \nabla \log p(\boldsymbol{y}|\boldsymbol{f})$. We need

$$\begin{aligned} \boldsymbol{f}^{new} &= (\boldsymbol{K}^{-1} + \boldsymbol{W})^{-1}\boldsymbol{b} \\ &= \boldsymbol{K}(\boldsymbol{I} - \boldsymbol{W}^{\frac{1}{2}} \underbrace{(\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{\frac{1}{2}})^{-1}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K})\boldsymbol{b}}_{v} \end{aligned}$$

Repeat until convergence:

- iteratively solve $(\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{\frac{1}{2}})\boldsymbol{v} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{b}$
- $\boldsymbol{a} = \boldsymbol{b} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{v}$
- $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{a}$

Marginal Likelihood

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

learn the value of θ, which can be done by minimizing the negative marginal likelihood

$$-\log p(\boldsymbol{y}|\boldsymbol{X},\theta) \approx \frac{1}{2}\hat{\mathbf{f}}^T \boldsymbol{K}^{-1}\hat{\mathbf{f}} - \log p(\boldsymbol{y}|\hat{\mathbf{f}}) + \frac{1}{2}\log|\boldsymbol{B}|$$

with $\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}$.

Reduced-rank approximation

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

We use a reduced-rank approximation:

$$\boldsymbol{K} \approx \boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^T + \boldsymbol{\Lambda}_1$$

with

$$\boldsymbol{Q} = \bigotimes_{d=1}^{D} \boldsymbol{Q}_d, \qquad N \times R \ \texttt{matrix}$$

$$\boldsymbol{S} = \bigotimes_{d=1}^{D} \boldsymbol{S}_d, \qquad R \times R \ \texttt{matrix}$$

$$\boldsymbol{\Lambda}_1 = \mathrm{diag}(\mathrm{diag}(\boldsymbol{K}) - \mathrm{diag}(\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^T))$$

Evaluating Marginal Likelihood

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

$$
\begin{aligned}
|\boldsymbol{B}| &= |\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{\frac{1}{2}}| \\
&\approx |\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\Lambda_1\boldsymbol{W}^{\frac{1}{2}} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^T\boldsymbol{W}^{\frac{1}{2}}| \\
&= |\Lambda_2 + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^T\boldsymbol{W}^{\frac{1}{2}}| \\
&= |\Lambda_2||\boldsymbol{S}||\boldsymbol{S}^{-1} + \boldsymbol{Q}^T\boldsymbol{W}^{\frac{1}{2}}\Lambda_2^{-1}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}| \\
&= |\Lambda_2||\boldsymbol{S}||\boldsymbol{S}^{-1} + \boldsymbol{Q}^T\Lambda_3\boldsymbol{Q}|
\end{aligned}
$$

Gradients

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Need the gradients wrt θ of

$$-\log p(\boldsymbol{y}|\boldsymbol{X},\theta) \approx \frac{1}{2}\hat{\mathbf{f}}^T \boldsymbol{K}^{-1}\hat{\mathbf{f}} - \log p(\boldsymbol{y}|\hat{\mathbf{f}}) + \frac{1}{2}\log|\boldsymbol{B}|$$

which is given by

$$\frac{\partial \log q(\boldsymbol{y}|\boldsymbol{X},\theta)}{\partial \theta_j} = \left.\frac{\partial \log q(\boldsymbol{y}|\boldsymbol{X},\theta)}{\partial \theta_j}\right|_{\text{explicit}} + \sum_{i=1}^{N} \frac{\partial \log q(\boldsymbol{y}|\boldsymbol{X},\theta)}{\partial \hat{f}_i}\frac{\partial \hat{f}_i}{\partial \theta_j}$$

Explicit Derivatives – Kernel Hyperparameters

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Given by

$$\left.\frac{\partial \log q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\partial \theta_j^c}\right|_{\text{explicit}} = \frac{1}{2}\hat{\mathbf{f}}^T \boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\boldsymbol{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\text{tr}\left((\boldsymbol{W}^{-1}+\boldsymbol{K})^{-1}\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\right).$$

Explicit Derivatives – Kernel Hyperparameters

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

$$
\begin{aligned}
&(\boldsymbol{W}^{-1} + \boldsymbol{K})^{-1} \\
&\quad = \boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{\frac{1}{2}})^{-1}\boldsymbol{W}^{\frac{1}{2}} \\
&\quad \approx \boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\Lambda\boldsymbol{W}^{\frac{1}{2}} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^{T}\boldsymbol{W}^{\frac{1}{2}})^{-1}\boldsymbol{W}^{\frac{1}{2}} \\
&\quad = \boldsymbol{W}^{\frac{1}{2}}(\Lambda_2 + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}\boldsymbol{S}\boldsymbol{Q}^{T}\boldsymbol{W}^{\frac{1}{2}})^{-1}\boldsymbol{W}^{\frac{1}{2}} \\
&\quad = \boldsymbol{W}^{\frac{1}{2}}(\Lambda_2^{-1} - \Lambda_2^{-1}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Q}(\boldsymbol{S}^{-1} + \boldsymbol{Q}^{T}\Lambda_3^{-1}\boldsymbol{Q})^{-1}\boldsymbol{Q}^{T}\boldsymbol{W}^{\frac{1}{2}}\Lambda_2^{-1}) \\
&\quad = \Lambda_3 - \Lambda_3\boldsymbol{Q}(\boldsymbol{S}^{-1} + \boldsymbol{Q}^{T}\Lambda_3\boldsymbol{Q})^{-1}\boldsymbol{Q}^{T}\Lambda_3
\end{aligned}
$$

Explicit Derivatives – Kernel Hyperparameters

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

$$\begin{aligned}
&\operatorname{tr}\left(\left(\boldsymbol{W}^{-1}+\boldsymbol{K}\right)^{-1}\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\right) \\
&\quad=\operatorname{tr}\left(\Lambda_3\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\right)-\operatorname{tr}\left(\Lambda_3\boldsymbol{Q}\left(\boldsymbol{S}^{-1}+\boldsymbol{Q}^T\Lambda_3\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}^T\Lambda_3\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\right) \\
&\quad=\operatorname{tr}\left(\Lambda_3\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\right)-\operatorname{tr}\left(\boldsymbol{V}_1^T\frac{\partial \boldsymbol{K}}{\partial \theta_j^c}\boldsymbol{V}_1\right)
\end{aligned}$$

with $\boldsymbol{V}_1=\Lambda_3\boldsymbol{Q}\operatorname{chol}\left[(\boldsymbol{S}^{-1}+\boldsymbol{Q}^T\Lambda_3\boldsymbol{Q})^{-1}\right]$.

Remaining Gradients

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Similar arguments for computing the remaining gradients

Linear Algebra Shortcuts

- $\mathrm{tr}(\boldsymbol{ABC}) = \mathrm{tr}(\boldsymbol{BCA})$
- $\mathrm{tr}(\boldsymbol{AB}^T) = \textsc{SumRows}(\boldsymbol{A} \circ \boldsymbol{B})$
- $\mathrm{diag}(\boldsymbol{ABC}^T) = (\boldsymbol{A} \circ \boldsymbol{C}) \cdot \mathrm{diag}(\boldsymbol{B})$ when $\boldsymbol{B}$ is diagonal

Experiment

Perry Groot, Markus Peters Tom Heskes, Wolgang Ketter

Artificial classification data generated on $\mathcal{X} = [0, 1]^2$ with various grid sizes.