Constraining Gaussian Processes by Variational Fourier Features

Arno Solin
Aalto University

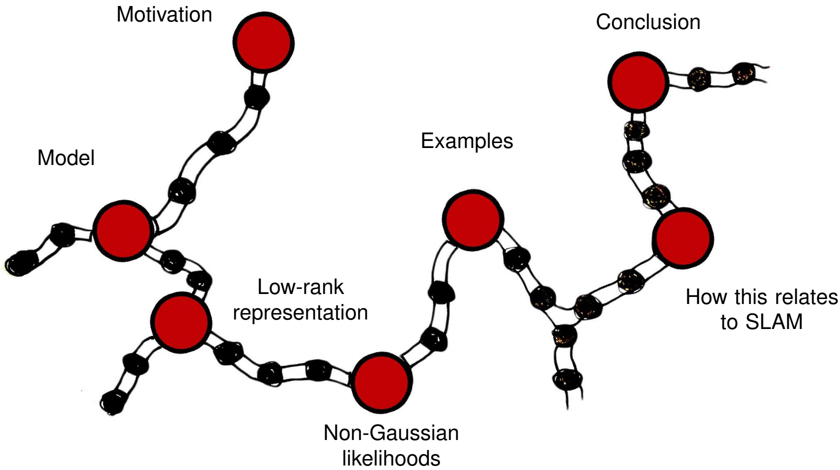
Joint work with

Manon Kok
(and earlier work with Nicolas Durrande,
James Hensman, and Simo Särkkä)

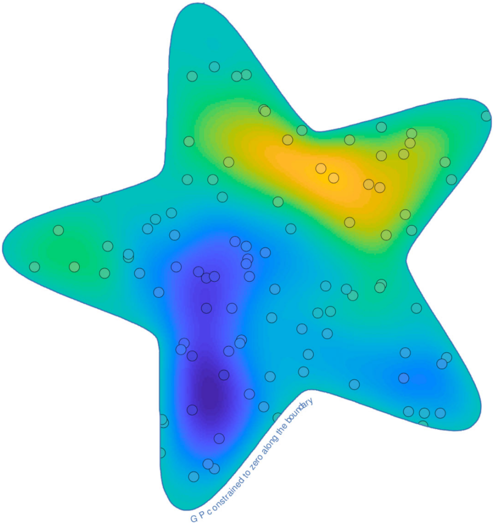
September 12, 2019

@arnosolin arno.solin.fi

Outline

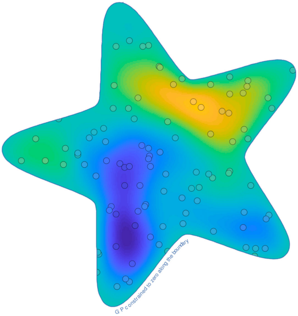

Motivation
Conclusion
Model
Examples
Low-rank
representation
How this relates
to SLAM
Non-Gaussian
likelihoods

The idea



What?

- Gaussian processes (GPs) provide a powerful framework for extrapolation, interpolation, and noise removal in regression and classification
- We constrain GPs to arbitrarily-shaped domains with boundary conditions
- Applications in, *e.g.*, imaging, spatial analysis, robotics, or general ML tasks



Why is this non-trivial?

GPs provide convenient ways for
model specification and inference, but . . .

- **Issue #1:**
  How to represent this prior?
- **Issue #2:**
  Limitations in scaling do large data sets
- **Issue #3:**
  Limitations in dealing with non-Gaussian likelihoods

Hilbert Space Methods for Reduced-Rank GPs

Problem formulation

- Gaussian process (GP) regression problem:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}')),$$
$$y_i = f(\mathbf{x}_i) + \varepsilon_i.$$

- The GP-regression has cubic computational complexity $\mathcal{O}(n^3)$ in the number of measurements.
- This results from the inversion of an $n \times n$ matrix:

$$\mathbb{E}[f(\mathbf{x}_*)] = \kappa(\mathbf{x}_*, \mathbf{x}_{1:n}) \left(\kappa(\mathbf{x}_{1:n}, \mathbf{x}_{1:n}) + \sigma_\mathrm{n}^2 \mathbf{I}\right)^{-1} \mathbf{y}$$
$$\mathbb{V}[f(\mathbf{x}_*)] = \kappa(\mathbf{x}_*, \mathbf{x}_*) - \kappa(\mathbf{x}_*, \mathbf{x}_{1:n}) \left(\kappa(\mathbf{x}_{1:n}, \mathbf{x}_{1:n}) + \sigma_\mathrm{n}^2 \mathbf{I}\right)^{-1} \kappa(\mathbf{x}_{1:n}, \mathbf{x}_*).$$

- Various sparse, reduced-rank, and related approximations have been developed for mitigating this problem.

Covariance operator

- For covariance function $\kappa(\mathbf{x}, \mathbf{x}')$ we can define covariance operator:

$$\mathcal{K}\,\phi = \int \kappa(\cdot, \mathbf{x}')\,\phi(\mathbf{x}')\,\mathrm{d}\mathbf{x}'.$$

- For stationary covariance function $\kappa(\mathbf{x}, \mathbf{x}') \triangleq \kappa(\|\mathbf{r}\|)$; $\mathbf{r} = \mathbf{x} - \mathbf{x}'$ we get

$$S(\boldsymbol{\omega}) = \int \kappa(\mathbf{r})\, e^{-\mathrm{i}\,\boldsymbol{\omega}^\mathsf{T}\mathbf{r}}\,\mathrm{d}\mathbf{r}.$$

- The transfer function corresponding to the operator $\mathcal{K}$ is

$$S(\boldsymbol{\omega}) = \mathscr{F}[\mathcal{K}].$$

- The spectral density $S(\boldsymbol{\omega})$ also gives the approximate eigenvalues of the operator $\mathcal{K}$.

Laplacian operator series

- In isotropic case $S(\omega) \triangleq S(\|\omega\|)$, we can expand

$$S(\|\omega\|) = a_0 + a_1 \|\omega\|^2 + a_2(\|\omega\|^2)^2 + a_3(\|\omega\|^2)^3 + \cdots .$$

- The Fourier transform of the Laplace operator ∇^2 is $-\|\omega\|^2$, *i.e.*,

$$\mathcal{K} = a_0 + a_1(-\nabla^2) + a_2(-\nabla^2)^2 + a_3(-\nabla^2)^3 + \cdots .$$

- Defines a pseudo-differential operator as a series of differential operators.
- Let us now approximate the Laplacian operators with a Hilbert method...

Series expansions of GPs

- Assume a covariance function $\kappa(\mathbf{x}, \mathbf{x}')$ and an inner product, say,

$$\langle f, g \rangle = \int_{\Omega} f(\mathbf{x})\, g(\mathbf{x})\, w(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

- The inner product induces a Hilbert-space of (random) functions.
- If we fix a basis $\{\phi_j(\mathbf{x})\}$, a Gaussian process $f(\mathbf{x})$ can be expanded into a series

$$f(\mathbf{x}) = \sum_{j=1}^{\infty} f_j\, \phi_j(\mathbf{x}),$$

  where f_j are jointly Gaussian.
- If we select ϕ_j to be the eigenfunctions of $\kappa(\mathbf{x}, \mathbf{x}')$ w.r.t. $\langle \cdot, \cdot \rangle$, then this becomes a Karhunen–Loève series.
- In the Karhunen–Loève case the coefficients f_j are independent Gaussian.

Hilbert-space approximation of the Laplacian

- Consider the eigenvalue problem for the Laplacian operators:

$$\begin{cases} -\nabla^2 \phi_j(\mathbf{x}) = \lambda_j^2 \, \phi_j(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \phi_j(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega. \end{cases}$$

- The eigenfunctions $\phi_j(\cdot)$ are orthonormal w.r.t. inner product

$$\langle f, g \rangle = \int_\Omega f(\mathbf{x}) \, g(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

$$\int_\Omega \phi_i(\mathbf{x}) \, \phi_j(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \delta_{ij}.$$

- The negative Laplacian has the formal kernel

$$\ell(\mathbf{x}, \mathbf{x}') = \sum_j \lambda_j^2 \, \phi_j(\mathbf{x}) \, \phi_j(\mathbf{x}')$$

in the sense that

$$-\nabla^2 f(\mathbf{x}) = \int \ell(\mathbf{x}, \mathbf{x}') \, f(\mathbf{x}') \, \mathrm{d}\mathbf{x}'.$$

Approximation of the covariance function

- Recall that we have the expansion

$$\mathcal{K} = a_0 + a_1(-\nabla^2) + a_2(-\nabla^2)^2 + a_3(-\nabla^2)^3 + \cdots .$$

- Substituting the formal kernel gives

$$\begin{aligned}\kappa(\mathbf{x},\mathbf{x}') &\approx a_0 + a_1\,\ell^1(\mathbf{x},\mathbf{x}') + a_2\,\ell^2(\mathbf{x},\mathbf{x}') + a_3\,\ell^3(\mathbf{x},\mathbf{x}') + \cdots \\ &= \sum_j \left[a_0 + a_1\,\lambda_j^2 + a_2\,\lambda_j^4 + a_3\,\lambda_j^6 + \cdots\right]\phi_j(\mathbf{x})\,\phi_j(\mathbf{x}').\end{aligned}$$

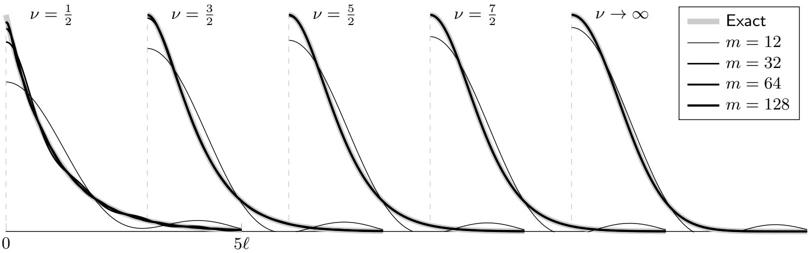
- Evaluating the spectral density series at $\|\boldsymbol{\omega}\|^2 = \lambda_j^2$ gives

$$S(\lambda_j) = a_0 + a_1\lambda_j^2 + a_2\lambda_j^4 + a_3\lambda_j^6 + \cdots .$$

- This leads to the final approximation

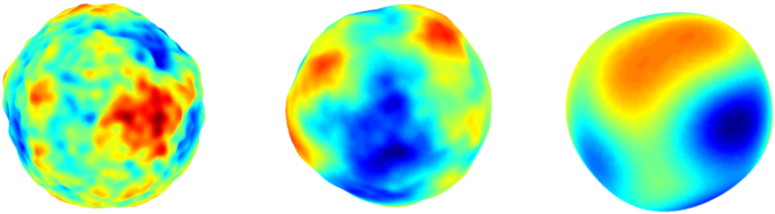
$$\boxed{\kappa(\mathbf{x},\mathbf{x}') \approx \sum_j S(\lambda_j)\,\phi_j(\mathbf{x})\,\phi_j(\mathbf{x}').}$$

Accuracy of the approximation



Approximations to covariance functions of the Matérn class of various degrees of smoothness; $\nu = 1/2$ corresponds to the exponential Ornstein–Uhlenbeck covariance function, and $\nu \to \infty$ to the squared exponential (exponentiated quadratic) covariance function.

Gaussian processes on a sphere



(a) $\nu = \frac{1}{2}$ and $\ell = 0.5$ (b) $\nu = \frac{3}{2}$ and $\ell = 0.5$ (c) $\nu \to \infty$ and $\ell = 0.5$

Easy to apply in simple domains
(hyper-spheres, hyper-cubes, ...)

Reduced-rank method for GP regression

- Recall the GP-regression problem

$$f(\mathbf{x}) \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$$
$$y_i = f(\mathbf{x}_i) + \varepsilon_i.$$

- Let us now approximate

$$f(\mathbf{x}) \approx \sum_{j=1}^{m} f_j \, \phi_j(\mathbf{x}),$$

  where $f_j \sim \mathrm{N}(0, S(\lambda_j))$.
- Via the matrix inversion lemma we then get

$$\mathbb{E}[f(\mathbf{x}_*)] \approx \phi_*^{\mathsf{T}}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \sigma_{\mathrm{n}}^2 \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Phi}^{\mathsf{T}} \mathbf{y},$$
$$\mathbb{V}[f(\mathbf{x}_*)] \approx \sigma_{\mathrm{n}}^2 \phi_*^{\mathsf{T}}(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \sigma_{\mathrm{n}}^2 \boldsymbol{\Lambda}^{-1})^{-1} \phi_*.$$

Computational complexity

- The computation of $\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$ takes $\mathcal{O}(nm^2)$ operations.
- The covariance function parameters do not enter $\mathbf{\Phi}$ and we need to evaluate $\mathbf{\Phi}^\mathsf{T}\mathbf{\Phi}$ only once (nice in parameter estimation).
- The scaling in input dimensionality can be quite bad—but depends on the chosen domain.

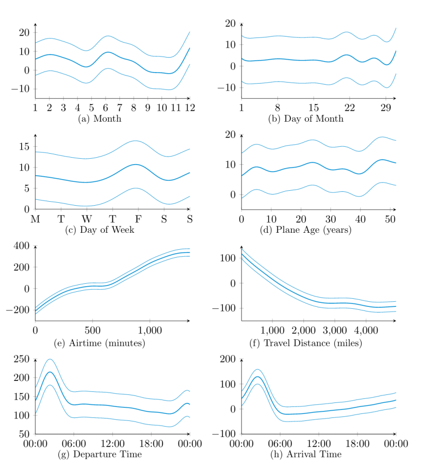
Airline delay example

- Every commercial flight in the US for 2008 ($n \approx 6$ M).
- Inputs, $\mathbf{x}$:
  Age of the aircraft, route distance, airtime, departure time, arrival time, day of the week, day of the month, and month.
- Target, y:
  Delay at landing (in minutes).
- Additive model:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, \sum_{d=1}^{8} \kappa_{\text{se}}(x_d, x'_d))$$

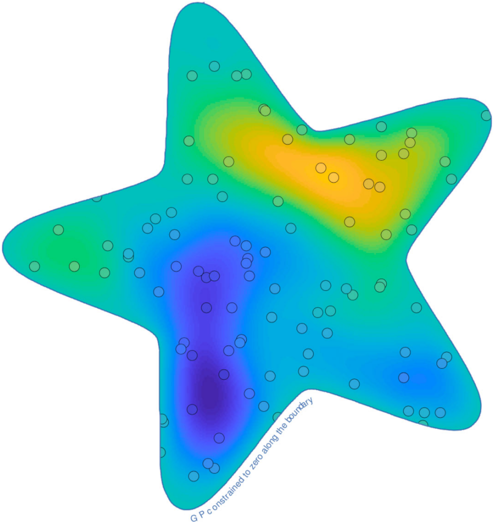
$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma_{\mathrm{n}}^2)$$

Airline delay example

- Every com 

$n \approx 6$ M).
- Inputs, $\mathbf{x}$:
Age of the … eparture time,
arrival time … th, and month.
- Target, y:
Delay at la
- Additive m

Results


GP constrained to zero along the boundary

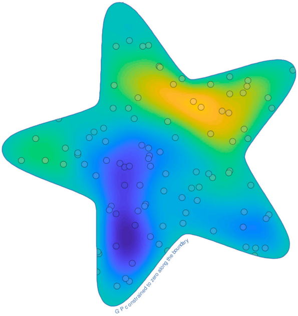
The model

In terms of a GP prior and a likelihood, this can be written as

$$\begin{cases} \quad f(\mathbf{x}) \sim \mathrm{GP}(0, \kappa(\mathbf{x}, \mathbf{x}')), & \mathbf{x} \in \Omega \\ \text{s.t. } f(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega \end{cases}$$

$$\mathbf{y} \mid \mathbf{f} \sim \prod_{i=1}^{n} p(y_i \mid f(\mathbf{x}_i))$$

where $(\mathbf{x}_i, y_i)$ are the n input–output pairs



Why is this non-trivial?

GPs provide convenient ways for
model specification and inference, but . . .

- **Issue #1:**
  How to represent this prior?
- **Issue #2:**
  Limitations in scaling do large data sets
- **Issue #3:**
  Limitations in dealing with non-Gaussian likelihoods

Addressing the three issues

- As a pre-processing step, we solve a Fourier-like generalised harmonic feature representation of the GP prior in the domain of interest
- Both constrains the GP and attains a low-rank representation that is used for speeding up inference
- The method scales as $\mathcal{O}(nm^2)$ in prediction and $\mathcal{O}(m^3)$ in hyperparameter learning (n number of data, m features)
- A variational approach to allow the method to deal with non-Gaussian likelihoods

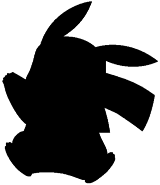
Low-rank representation

- Given a domain $\Omega \subset \mathbb{R}^d$ (d typically 1–3), we project the GP onto the eigenbasis of the Laplace operator, ∇^2, that solves the eigenvalue problem:

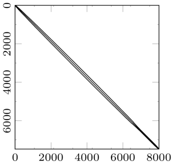
$$\begin{cases} -\nabla^2 \phi_j(\mathbf{x}) = \lambda_j^2 \phi_j(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \phi_j(\mathbf{x}) = 0, & \mathbf{x} \in \partial\Omega. \end{cases}$$

- The approximate eigenvalues and eigenfunctions of the Laplacian in Ω (s.t. the the boundary conditions) can be solved numerically

Domain and discrete Laplacian

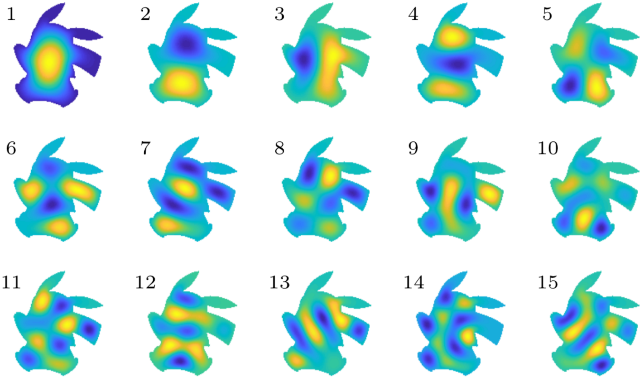


$$+ \; -\nabla^2 \; \approx$$



Finite difference approximation of the operator in a discrete grid of the image.

Harmonic basis functions



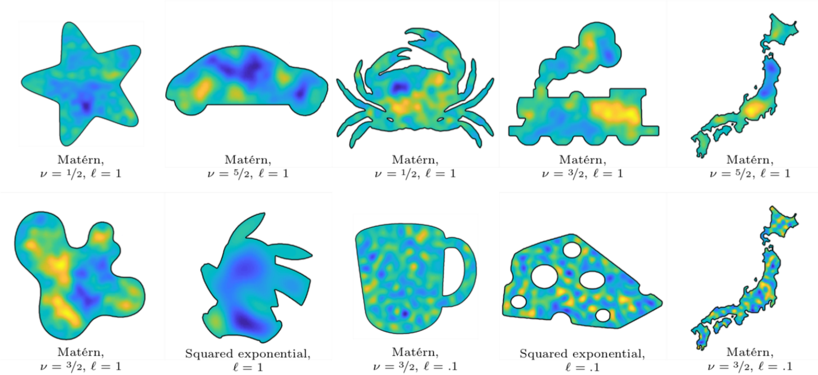
Representation of the GP prior

- We require the covariance function $\kappa(\cdot, \cdot)$ to be stationary
- Leverage the link between stationary covariance functions and the Laplacian for approximating the covariance function by the eigendecomposition and the spectral density function:

$$\kappa(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^{m} S(\lambda_j)\, \phi_j(\mathbf{x})\, \phi_j(\mathbf{x}') = \mathbf{\Phi} \mathbf{\Lambda} \mathbf{\Phi}^\mathsf{T},$$

  where $s(\cdot)$ is the spectral density function of $\kappa(\cdot, \cdot)$
- As $\mathbf{\Phi}$ does not depend on the hyperparameters and $\mathbf{\Lambda}$ is diagonal, we also get a computational boost

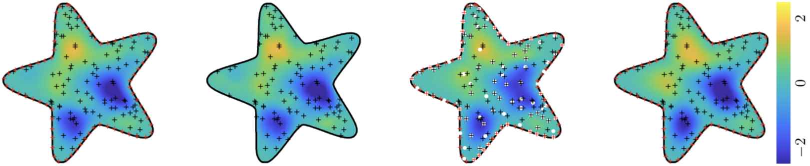
Samples from the GP prior



Non-Gaussian likelihoods

- For non-Gaussian likelihoods, we set up a variational approach and maximize the ELBO
- In practice, we form a Gaussian approximation to the posterior $q(\mathbf{u})$, for the set of m harmonic basis functions
- Optimise the ELBO with respect to the mean and variance of the approximation

Regression example

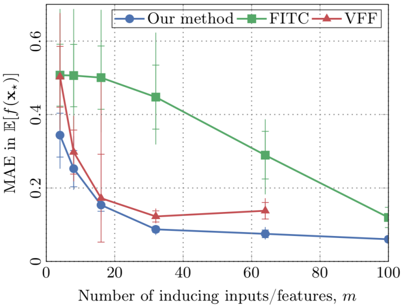


(a) FULL (b) OUR (c) FITC (d) VFF

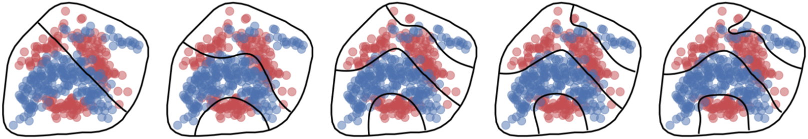
- Alternative approaches: Zero-noise measurements along the boundary for constraining the GP, and applying general-purpose approximations

Regression example

- Naive full GP (baseline)
- Our method
- Fully independent training conditional (FITC)
- Variational Fourier features (VFF)



Banana classification example

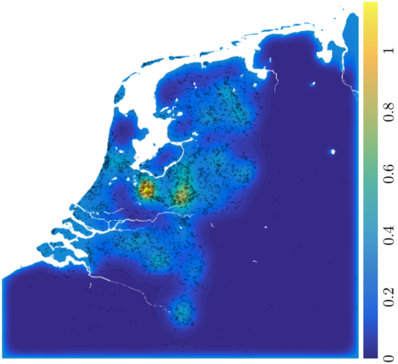


(a) $m = 4$ (b) $m = 8$ (c) $m = 16$ (d) $m = 32$ (e) $m = 64$

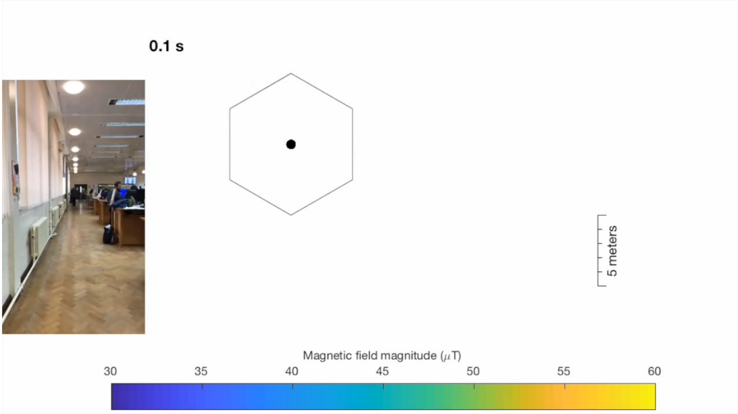
- The outermost decision boundary comes form the prior (know boundary of uncertainty)
- The posterior improves with the number of harmonic basis functions

Modelling tick density in the Netherlands

- 9 months of tick bites from https://tekenradar.nl
- 4,446 data points
- A log-Gaussian Cox process model (Poisson likelihood)
- Modelling the log intensity as a GP with boundary conditions



Simultaneous localisation and mapping (SLAM)

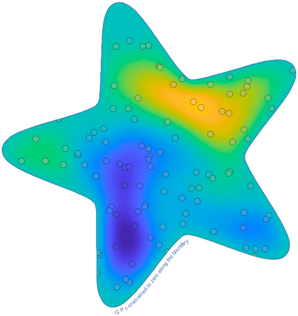


View on YouTube: `https://youtu.be/pbwWLoh6mvI`

Kok and Solin. *Scalable Magnetic Field SLAM in 3D Using Gaussian Process Maps*. FUSION'18.

Recap

- Constraining GPs to arbitrarily-shaped domains with boundary conditions
- Utilizes the link between the stationary covariance functions and the Laplace operator
- Applications in, *e.g.*, imaging, spatial analysis, robotics, or general ML tasks



Bibliography

- A. Solin and M. Kok (2019). Know your boundaries: Constraining Gaussian processes by variational harmonic features. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR 89:2193–2202.
- A. Solin and S. Särkkä (2019). Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*.
- J. Hensman, N. Durrande, and A. Solin (2018). Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 18(151):1–52.
- M. Kok and A. Solin (2018). Scalable magnetic field SLAM in 3D using Gaussian process maps. *Proceedings of the International Conference on Information Fusion (FUSION)*, pages 1353–1360.

- Homepage:
  http://arno.solin.fi
- Twitter:
  @arnosolin