Bayesian autoregression to optimize temporal Matérn kernel Gaussian process hyperparameters

Wouter M. Kouw

International Conference on Probabilistic Numerics 2025

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Problem: GP hyperparameter optimization

Consider a Gaussian process over functions in time and a delta likelihood:

$$p(f \mid t;\, \psi) = \mathcal{GP}\big(f(t) \,\big|\, 0,\, \kappa_\psi(t, t')\big)\,, \qquad p(y_k \mid f, t_k) = \delta\big(y_k - f(t_k)\big)$$

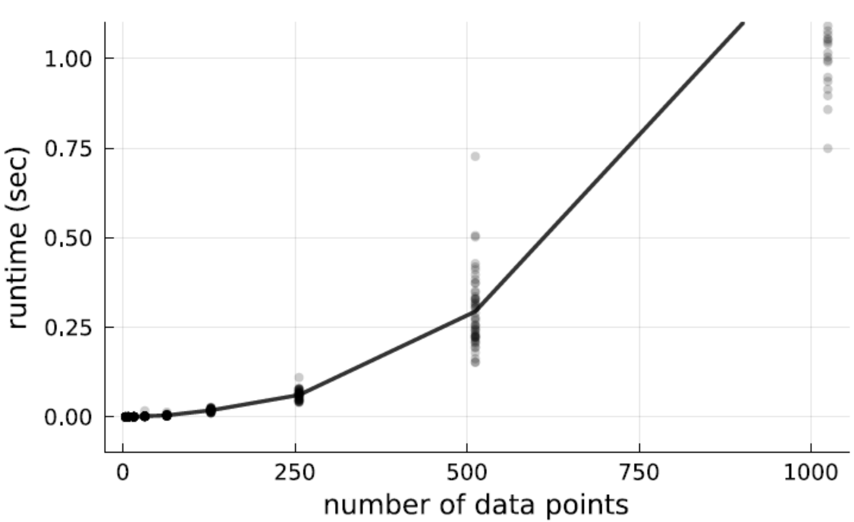
Typical approach to finding kernel hyperparam

$$\psi^* = \underset{\psi \in \Psi}{\arg\max}\, p(\boldsymbol{y} \mid \boldsymbol{t};\, \psi)$$

$$= \underset{\psi \in \Psi}{\arg\max}\ (2\pi)^{-1/2} \big|K_\psi\big|^{-1/}$$

But this requires inverting the kernel covarian

Can this be done faster?



Possible solution

For Matérn kernels, you could convert the GP to an SDE* and try maximum likelihood.

Conversion:

$$F(\omega) = H(\omega)W(\omega) \quad \Longrightarrow \quad S_F(\omega) = |H(\omega)|^2 S_W(\omega) \quad \text{such that} \quad S_F(\omega) = S_\kappa(\omega)$$

Let $S_\kappa(\omega)$ be the power spectral density of the GP governed by the Matérn κ_ψ:

$$S_\kappa(\omega) = \underbrace{\sigma^2 \frac{2\pi^{\frac{1}{2}}\Gamma\left(\nu + \frac{1}{2}\right)}{\Gamma(\nu)} \lambda^{2\nu}}_{:= \varsigma^2} \underbrace{(\lambda^2 + \omega^2)^{-\left(\nu+\frac{1}{2}\right)}}_{= \underbrace{(\lambda + i\omega)^{-\left(\nu+\frac{1}{2}\right)}}_{:= H(i\omega)} (\lambda - i\omega)^{-\left(\nu+\frac{1}{2}\right)}}$$

*Hartikainen & Särkkä (2010)

Possible solution

Unpacking the characteristic polynomial of the transfer function, reveals an order m process:

$$(i\omega)^m F(\omega) + \sum_{n=0}^{m-1} a_n (i\omega)^n F(\omega) = W(\omega)$$

where $a_n = \binom{m}{n} \lambda^{m-n}$ are found through the binomial theorem. Note that $m = \nu + \frac{1}{2}$.

The inverse Fourier transform produces:

$$\frac{d^m f(t)}{dt^m} + \sum_{n=0}^{m-1} a_n \frac{d^n f(t)}{dt^n} = w(t)$$

Example:

$$\nu = \frac{1}{2} \quad \Rightarrow \quad m = 1 \quad \Rightarrow \quad \frac{df(t)}{dt} = -\lambda f(t) + w(t)$$

Possible solution

You could take the SDE, form a state-space model and perform maximum likelihood estimation.

But maximum likelihood is still iterative.

- Can we do without iterations?

State-space models with unknown parameters, states and noise are typically unidentifiable.

- The posterior over parameters will heavily depend on the prior distribution.

I decided to explore a different approach.

Alternate solution

I will use Euler-Maruyama with a higher-order forward finite difference,

$$\frac{d^m f(t)}{dt^m} \approx \frac{1}{\Delta^m} \sum_{n=0}^{m} (-1)^{m-n} \binom{m}{n} f(t + n\Delta)$$

where $\Delta = t_k - t_{k-1}$.

Applying this to each of the derivatives in the SDE:

$$\sum_{n=0}^{m} a_n \frac{d^n f(t)}{dt^n} \approx \sum_{n=0}^{m} \frac{a_n}{\Delta^m} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f_{k+j}$$

where $f_k = f(t_k)$.

The white noise process is discretized to $w_k \sim \mathcal{N}(0, \varsigma^2 \Delta)$.

Alternate solution

We can re-arrange the discretized SDE to a discrete-time autoregressive process:

$$\sum_{n=0}^{m} \frac{a_n}{\Delta^n} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} f_{k+j} = w_k$$

$$\vdots$$

$$f_{k+m} = \sum_{n=0}^{m-1} \theta_n f_{k+n} + \Delta^m w_k$$

where $\theta_n = (-1)^{m-n+1}\binom{m}{n} - \sum_{j=0}^{m-1} a_n \Delta^{m-n} (-1)^{j-n}\binom{j}{n}$ and $\tau = \frac{1}{\Delta^{2m+1}\varsigma^2}$.

Example for $m = 1$:

$$f_{k+1} = (1 - \lambda\Delta) f_k + w_k \quad \text{where } w_k \sim \mathcal{N}(0, \Delta^3 \varsigma^2)$$

Alternate solution

Consider a likelihood function of the form:

autoregressive coefficients

likelihood precision

$$p(y_k|\bar{y}_{k-1},\theta,\tau) = \mathcal{N}(y_k|\theta^\top \bar{y}_{k-1},\tau^{-1})$$

buffer of previous observations

We can then construct a Bayesian filter:

$$p(\theta,\tau\,|y_{1:k}) = \frac{p(y_k|\bar{y}_{k-1},\theta,\tau)}{p(y_k|y_{1:k-1})} p(\theta,\tau\,|y_{1:k-1})$$

Conjugate prior to this autoregressive likelihood is a multivariate Normal – gamma distribution:

$$p(\theta,\tau) = \mathcal{NG}(\theta,\tau\mid\mu_0,\Lambda_0,\alpha_0,\beta_0)$$

Alternate solution

After we've updated our posterior distribution, we can revert back to kernel hyperparameters.

For $m = 1$, this is exact;

$$\lambda = \frac{\Delta}{1-\mu}\ , \qquad \sigma^2 = \frac{\beta}{2(\alpha-1)(1-\mu)\Delta^2}$$

For $m > 1$, we end up with a system of polynomials.

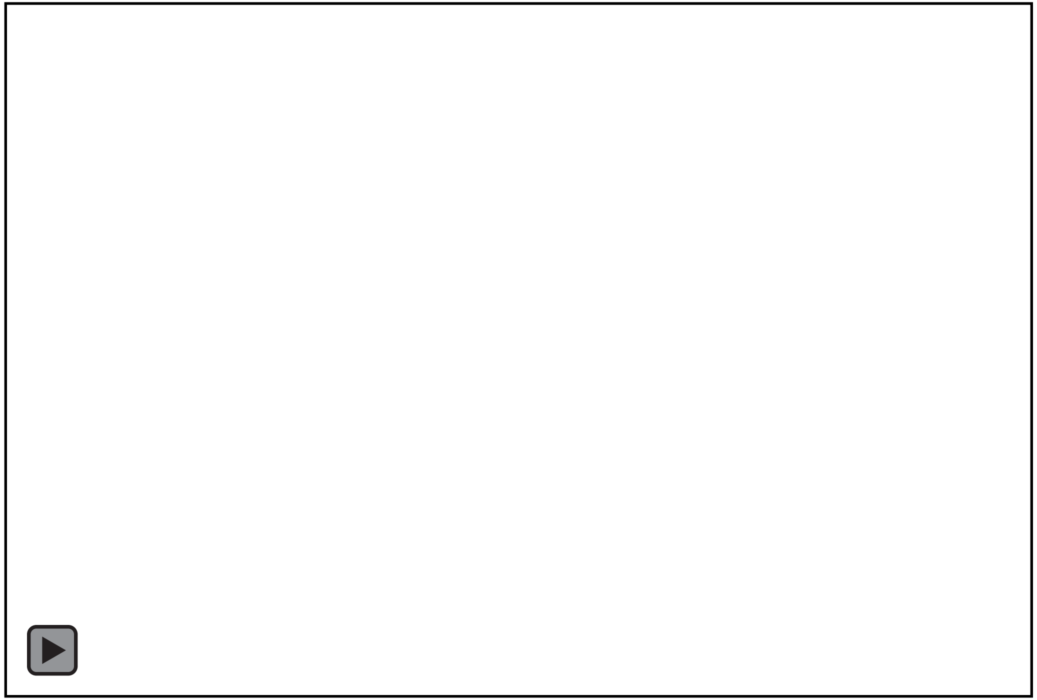
Use a nonlinear least-squares approach, with objectives

$$g_n(\psi) = (\mu_n - \theta_n)^2\ , \qquad g_m(\psi) = \left(\frac{\alpha-1}{\beta} - \tau\right)^2$$

for $n = 0, \dots, m-1$.

Then find the minimizer $\psi^* = \underset{\psi \in \Psi}{\arg\max} \sum_{i=0}^{m} g_i(\psi)$

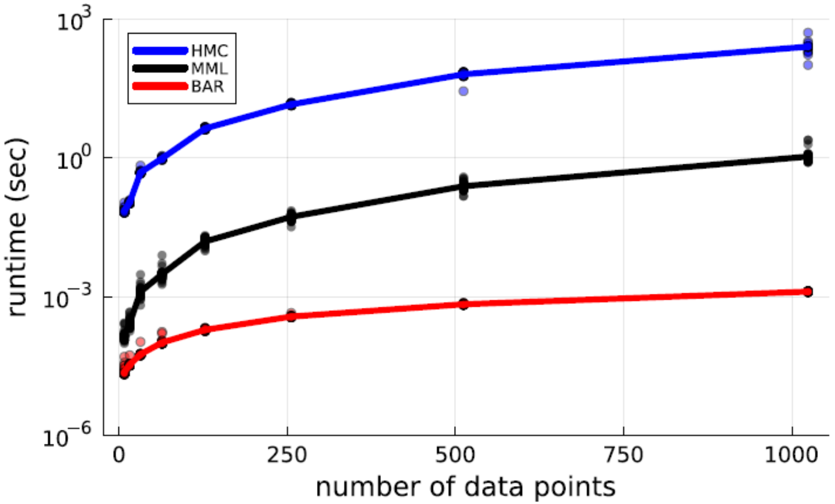
Demo

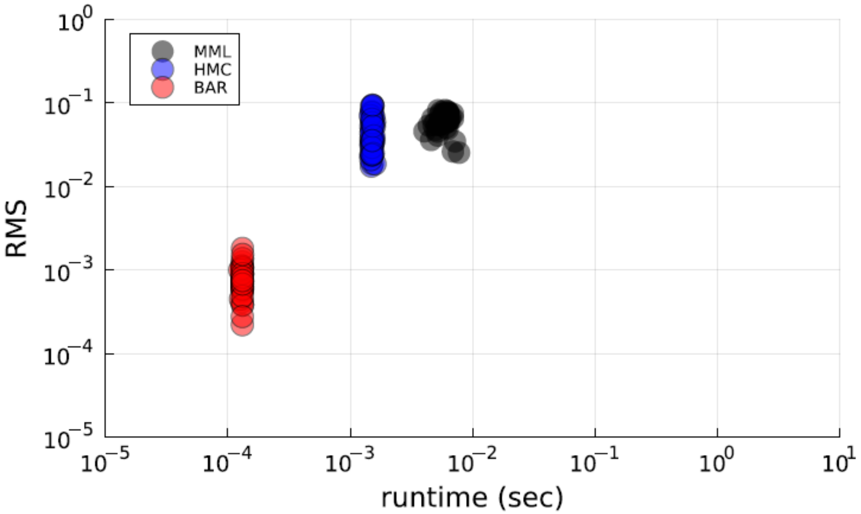


Experiments

50 Simulations: $y \sim \mathcal{GP}\left(\sin(\zeta\bar{t} + \eta)\,, \kappa_\psi(\bar{t})\right)$ for $\zeta \sim U(0,2), \eta \sim U(0,2\pi)$.

Matérn-1/2:

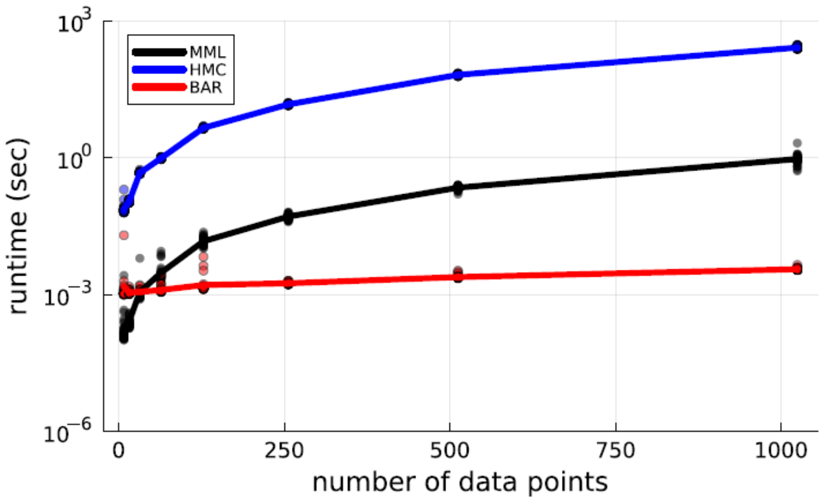


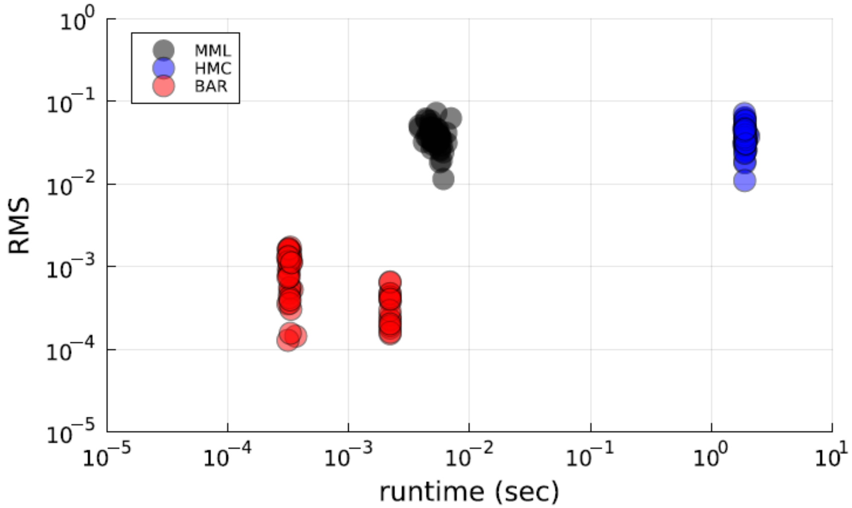


Experiments

50 Simulations: $y \sim \mathcal{GP}\left(\sin(\zeta\bar{t} + \eta)\,, \kappa_\psi(\bar{t})\right)$ for $\zeta \sim U(0,2), \eta \sim U(0,2\pi)$.

Matérn-3/2:





Outlook

Advantages:

- Recursive solution to kernel hyperparameters (you could stop for < N).

Limitations:

- Approximation error by Euler-Maruyama.
- Approximation by nonlinear LS for reversion.

Future work:

- Analysis of asymptotic properties: bias, consistency, stability.
- Generalize to noisy observations.

Thanks for your attention!

paper

rxinfer

TU/e

Extra slides

Matérn-class kernel covariance function:

$$\kappa_{\psi}(t,t') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l} |t-t'| \right)^{\nu} B_{\nu} \left(\frac{\sqrt{2\nu}}{l} |t-t'| \right)$$

Autoregressive parameter posterior has closed-form updates:

$$\Lambda_{k+1} = \Lambda_k + \bar{y}_k \bar{y}_k^{\top}, \qquad \mu_{k+1} = \left(\Lambda_k + \bar{y}_k \bar{y}_k^{\top} \right)^{-1} (\Lambda_k \mu_k + \bar{y}_k y_{k+1}),$$

$$\alpha_{k+1} = \alpha_k + \frac{1}{2}\ , \qquad \beta_{k+1} = \beta_k + \frac{1}{2} (y_{k+1}^2 - \mu_{k+1}^{\top} \Lambda_{k+1} \mu_{k+1} + \mu_k^{\top} \Lambda_k \mu_k)$$