Scalable Log Determinants for Gaussian Process Kernel Learning

David Eriksson[1] Kun Dong[1] Hannes Nickisch[4] David Bindel[2] Andrew Gordon Wilson[3]

Applied Math[1], CS[2], ORIE[3], Philips Research[4]

Cornell University

Gaussian Processes (GPs)

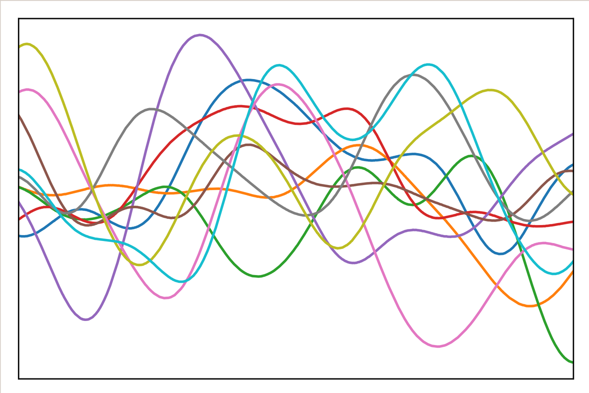
- Multivariate normals are distributions over vectors
- Gaussian processes are distributions over functions

$\mu : \mathbb{R}^d \to \mathbb{R}$ is the mean field; $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the *kernel*

$f \sim GP(\mu, k)$ means

$$\forall X = (x_1, \ldots, x_n), \quad x_i \in \mathbb{R}^d: \quad f_X \sim N(\mu_X, K_{XX}) \quad \text{where}$$

$$f_X \in \mathbb{R}^n; \quad (f_X)_i = f(x_i)$$
$$\mu_X \in \mathbb{R}^n; \quad (\mu_X)_i = \mu(x_i)$$
$$K_{XX} \in \mathbb{R}^{n \times n}; \quad (K_{XX})_{ij} = k(x_i, x_j)$$



Write K_{XX} as K when unambiguous

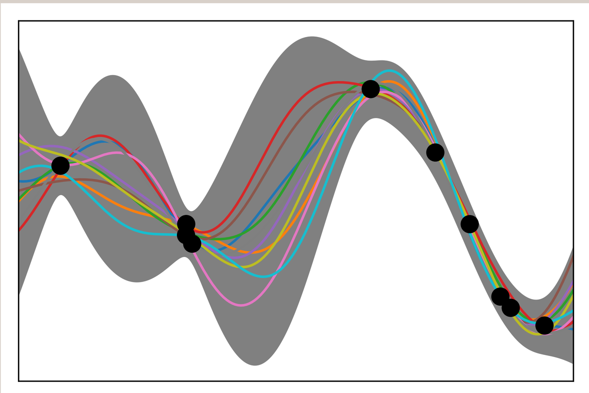
GP Regression

Bayesian framework: prior is $f \sim GP(\mu, k)$
Obtain noisy measurements:

$$y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

Posterior is $GP(\mu', k')$ with

$$\mu'(x) = \mu(x) + K_{xX} c$$
$$k'(x, y) = K_{xy} - K_{xX} \tilde{K}^{-1} K_{Xy}$$



where $\tilde{K}c = y - \mu_X$, $\tilde{K} = K_{XX} + \sigma^2 I$

- Compute c (and hence posterior mean) via Cholesky or CG
- For fast CG, make matvecs with K scale via
  - Low rank approximation (inducing point methods)
  - Interpolation to regular grid + FFT
  - Fast multipole expansions
- What about learning kernel parameters as well?

Kernel learning

- Typically k depends on a vector of *hyperparameters* θ
- Estimate θ from data by maximizing the (log) likelihood

$$\mathscr{L}(\theta|y) = \mathscr{L}_y + \mathscr{L}_{|K|} - \frac{n}{2}\log(2\pi)$$

where (again with $c = \tilde{K}^{-1}(y - \mu_X)$)

$$\mathscr{L}_y = -\frac{1}{2}(y-\mu)^T c, \qquad \frac{\partial \mathscr{L}_y}{\partial \theta_i} = \frac{1}{2} c^T \left(\frac{\partial \tilde{K}}{\partial \theta_i}\right) c$$

$$\mathscr{L}_{|K|} = -\frac{1}{2}\log\det\tilde{K}, \qquad \frac{\partial \mathscr{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\operatorname{tr}\left(\tilde{K}^{-1}\frac{\partial K}{\partial \theta_i}\right),$$

- Can efficiently compute $\mathscr{L}_y$ via iterative method given fast MVMs
- Naively computing $\mathscr{L}_{|K|}$ requires Cholesky factorization

Scaled eigenvalue method

- Approximate eigenvalues λ_i of K_{XX} using the n largest eigenvalues μ_i of K_{YY} on a full grid with m points such that $X \subset Y$:

$$\log|K_{XX} + \sigma^2 I| = \sum_{i=1}^{n} \log(\lambda_i + \sigma^2) \approx \sum_{i=1}^{n} \log\left(\frac{n}{m}\mu_i + \sigma^2\right)$$

- Can handle non-Gaussian likelihoods via the Fielder bound

Stochastic trace estimation

- Our goal is to estimate, for a symmetric positive definite matrix $\tilde{K}$

$$\mathscr{L}_{|K|} = -\frac{1}{2}\operatorname{tr}(\log(\tilde{K})) \quad \text{and} \quad \frac{\partial \mathscr{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\operatorname{tr}\left(\tilde{K}^{-1}\left(\frac{\partial \tilde{K}}{\partial \theta_i}\right)\right)$$

- Stochastic expression for $\mathscr{L}_{|K|}$ and first derivatives:

$$\mathscr{L}_{|K|} = -\frac{1}{2}\mathbb{E}\left[z^T(\log\tilde{K})z\right], \qquad \frac{\partial \mathscr{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\mathbb{E}\left[z^T\tilde{K}^{-1}\frac{\partial K}{\partial \theta_i}z\right],$$

- Common choices of probe vectors z:
  - Hutchinson: $z_i = \pm 1$ with probability 0.5
  - Gaussian: $z_i \sim \mathscr{N}(0, 1)$
- Estimate via sample means with several random probe vectors
- Need to multiply $\log(\tilde{K})$ by probe vectors efficiently

Lanczos

- Function application with fast MVMs $\Longrightarrow$ try Lanczos:
- Lanczos on $\tilde{K}$ computes partial tridiagonalization:

$$\tilde{K}Q_k = Q_k T_k + q_{k+1}e_k^T\beta_k, \qquad Q_k^T Q_k = I$$

$$Q_k \equiv \left[q_1 \ \ldots \ q_k\right], \qquad T_k \equiv \operatorname{tridiag}\begin{pmatrix} \beta_1 & \ldots & \beta_{k-1} \\ \alpha_1 & \alpha_2 & \ldots & \alpha_k \\ \beta_1 & \ldots & \beta_{k-1} \end{pmatrix}$$

- Start from $q_1 = z/\|z\|$ and compute approximations

$$u = \tilde{K}^{-1}z \qquad \approx \|z\| Q_k T_k^{-1} e_1 \qquad \text{(Conjugate gradients)}$$

$$\kappa = z^T\left(\log\tilde{K}\right)z \approx \|z\|^2 e_1^T\left(\log\tilde{T}_k\right)e_1 \qquad \text{(Gauss quadrature)}$$

Chebyshev

- Based on a polynomial approximation of the log
- Minimizes the worst-case error over an interval
- Lanczos is sensitive to the locations of the eigenvalues and tends to yield better accuracy

Fast MVMs

- Our experiments use structured kernel interpolation (SKI)

$$K_{XX} \approx W K_{UU} W^T$$

- W is a sparse n-by-m matrix of interpolation weights
- The points U are referred to as inducing points
- $>$ 1D, rectilinear grid, product covariance $\Longrightarrow$ Kronecker structure
- 1D, regular grid, stationary covariance $\Longrightarrow$ Toeplitz structure
- $>$ 1D, regular grid, stationary covariance $\Longrightarrow$ BTTB structure

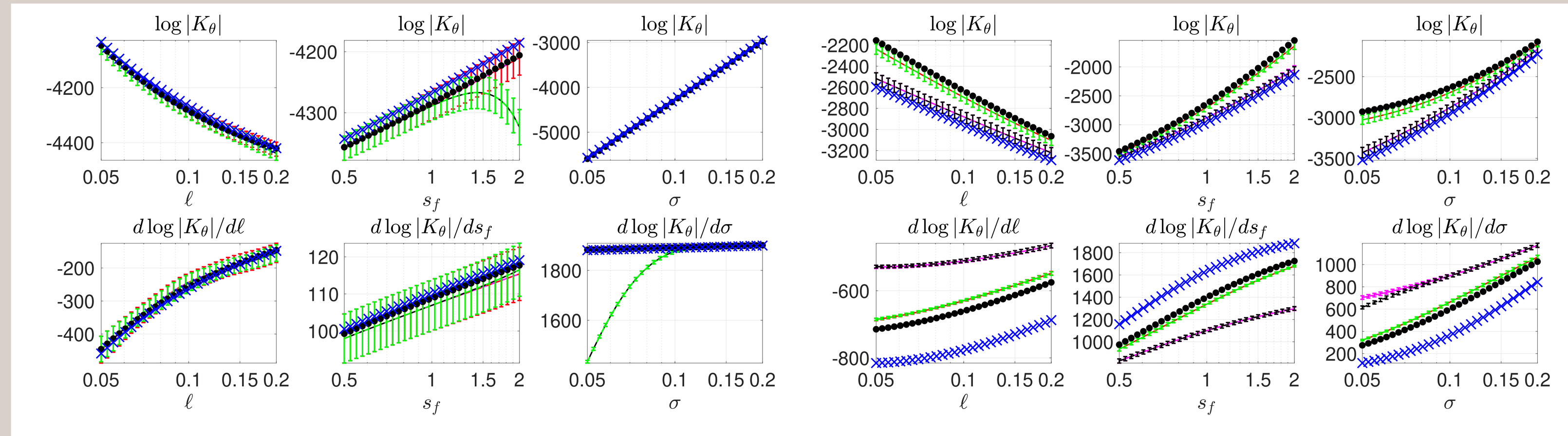
Diagonal correction

- SKI may provide a poor estimate of the diagonal entries
- Modify the SKI approximation to add a diagonal matrix D:

$$K_{XX} \approx W K_{UU} W^T + D, \qquad \operatorname{diag}(K_{XX}) = \operatorname{diag}(W K_{UU} W^T + D)$$

- Not supported by scaled eigenvalues, works with our MVM based approach

Log det accuracy



(a) RBF kernel (b) Matérn 1/2 kernel

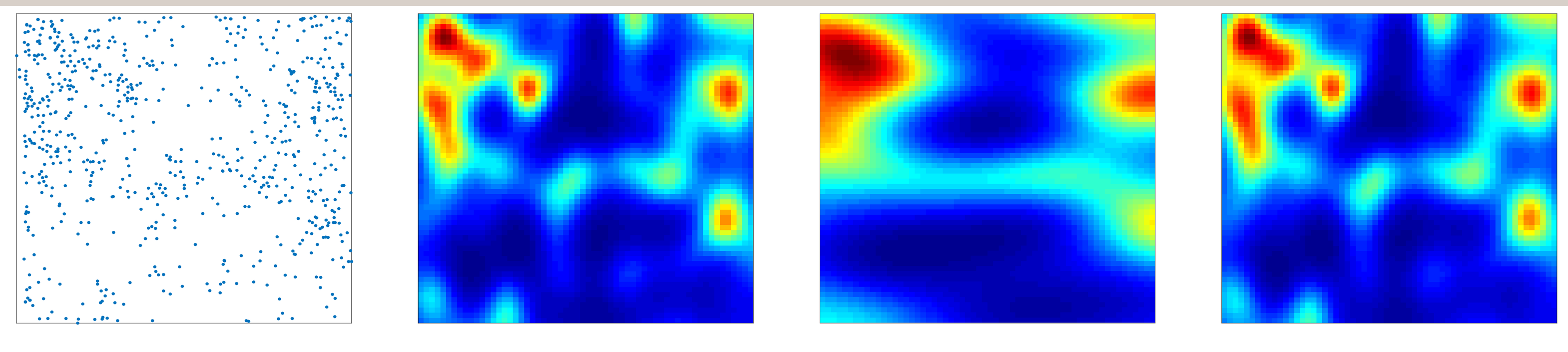
- The data is 1000 points drawn from $\mathscr{N}(0, 2)$.
- The exact values are (•), Lanczos with diagonal replacement is (—), Chebyshev with diagonal replacement is (—), Lanczos without diagonal replacement is (—), Chebyshev without diagonal replacement is (—), Scaled eigenvalues is (×).
- The error bars of Lanczos and Chebyshev were computed from 10 runs

Daily precipitation

Method	n	m	MSE	Time [min]
Lanczos	528k	3M	0.613	14.3
Scaled eigenvalues	528k	3M	0.621	15.9
Exact	12k	-	0.903	11.8

- Precipitation data collected from $\approx$ 5500 US weather stations
- Use induced grid of size $100 \times 100 \times 300$
- Use a subset of 12,000 entries for training with the exact method

Hickory Data Set



(a) Point pattern data (b) Prediction by exact (c) Scaled eigenvalues (d) Lanczos

- Fitted log-Gaussian Cox process model to hickory tree counts in Michigan
- Area discretized using a 60×60 grid
- The scaled eigenvalue method was used in conjunction with the Fiedler bound

Discussion

- New method to efficiently compute the log det and derivatives
- Lanczos outperforms Chebyshev in general
- Our method is flexible and requires only fast MVMs
- Can explore same ideas for computing higher-order derivatives
- Supports diagonal correction and non-Gaussian likelihoods
- Implementations are available at:

`https://github.com/kd383/GPML_SLD`
`https://github.com/jrg365/gpytorch`

https://people.cam.cornell.edu/~dme65/ dme65@cornell.edu