Stochastic Estimators for GP Kernel Learning

David Bindel Kun Dong David Eriksson Andrew Wilson
CS, Applied Math, ORIE
Cornell University

Gaussian Processes (GPs)

Multivariate normals are distributions over vectors;
Gaussian processes are distributions over functions.

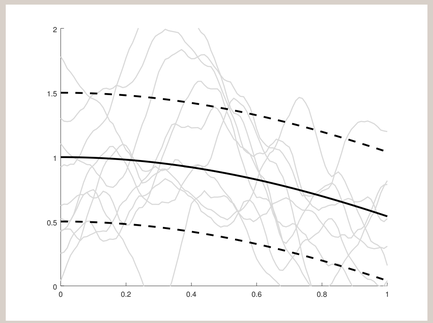
$\mu : \mathbb{R}^d \to \mathbb{R}$ is the mean field; $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the *kernel*.

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

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

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

Write K_{XX} as K when unambiguous.



GP Regression

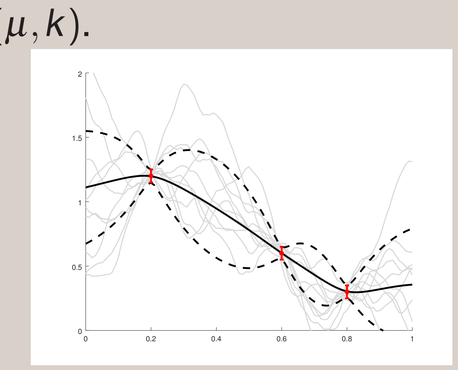
Bayesian framework: prior is $f \sim GP(\mu, k)$.
Obtain noisy measurements:

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

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?

Likelihood and Gradient Estimators

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

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

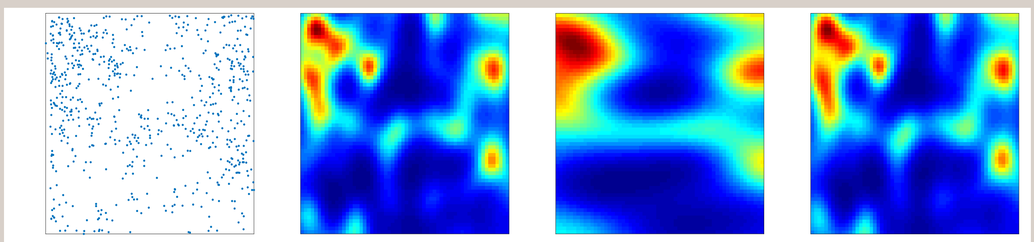
$$\mathscr{L}_{|K|} = -\frac{1}{2}\log\det \tilde{K}, \qquad \frac{\partial \mathscr{L}_{|K|}}{\partial \theta_i} = -\frac{1}{2}\mathrm{tr}\left(\tilde{K}^{-1}\frac{\partial K}{\partial \theta_i}\right),$$

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

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

  for random $z \in \mathbb{R}^n$ with entries independent, mean 0, variance 1.
- Estimate via sample means with several random *probe vectors* (typical choice is Rademacher = Hutchinson estimator).
- Solves and log computations via CG / Lanczos.

Hickory Data Set



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

- Lanczos-based solvers now integrated into GPML
- Better kernel learning vs. cruder approximations to log det
- Example: Fit log-Gaussian Cox process model to 703 hickory trees in Michigan (data from R `spatstat` package)

Re-using Lanczos

Lanczos on $\tilde{K}$ computes partial tridiagonalization:

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

$$Q_K \equiv \begin{bmatrix} q_1 & \ldots & q_k \end{bmatrix}, \qquad T_k \equiv \mathrm{tridiag}\begin{pmatrix} \beta_1 & \ldots & \beta_{k-1} \\ \alpha_1 & \alpha_2 & \ldots & \alpha_k \\ \beta_1 & \ldots & \beta_{k-1} \end{pmatrix}$$

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

$$u = \tilde{K}^{-1} z \approx \|z\| Q_k T_k^{-1} e_1 \qquad \text{(Conjugate gradients)}$$
$$\kappa = z^T \left(\log \tilde{K}\right) z \approx \|z\|^2 e_1^T \left(\log \tilde{T}_k\right) e_1 \qquad \text{(Gauss quadrature)}$$

Open question: does it make sense to precondition both?

Hessian Estimators

- For Hessian of $\mathscr{L}_y$, exploit $\mathbb{E}[zz^T] = I$:

$$\frac{\partial^2 \mathscr{L}_y}{\partial \theta_i \partial \theta_j} = \frac{1}{2} c^T \left(\frac{\partial^2 K}{\partial \theta_i \partial \theta_j} - 2\frac{\partial K}{\partial \theta_i}\tilde{K}^{-1}\frac{\partial K}{\partial \theta_j}\right) c$$
$$= \frac{1}{2}\mathbb{E}\left[c^T \left(\frac{\partial^2 K}{\partial \theta_i \partial \theta_j} - 2\frac{\partial K}{\partial \theta_i} z z^T \tilde{K}^{-1}\frac{\partial K}{\partial \theta_j}\right) c\right]$$

- Can also tackle Hessian of $\mathscr{L}_{|K|}$ with an independent probe $\check{z}$:

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

Re-using computations

For each of m probes z_ℓ, use *one* Lanczos decomposition for

$$u_\ell = \tilde{K}^{-1} z_\ell \qquad \kappa_\ell = z_\ell^T (\log \tilde{K}) z_\ell$$

A few additional matvecs with derivatives:

$$w_{\ell,i} = \frac{\partial K}{\partial \theta_i} z_\ell \qquad w_{\ell,ij} = \frac{\partial^2 K}{\partial \theta_i \partial \theta_j} z_\ell$$

$$d_i = \frac{\partial K}{\partial \theta_i} c \qquad d_{ij} = \frac{\partial^2 K}{\partial \theta_i \partial \theta_j} c$$

The remaining computations are all sums and dot products:

$$\mathscr{L} \approx -\frac{1}{2}\left[(y-\mu_X)^T c + \frac{1}{m}\sum_\ell \kappa_\ell + n\log(2\pi)\right]$$
$$\frac{\partial \mathscr{L}}{\partial \theta_i} \approx -\frac{1}{2}\left[-c^T d_i + \frac{1}{m}\sum_\ell u_\ell^T w_{\ell,i}\right]$$
$$\frac{\partial^2 \mathscr{L}_y}{\partial \theta_i \partial \theta_j} \approx -\frac{1}{2}\left[-c^T d_{ij} + \frac{1}{m}\sum_\ell \left((z_\ell^T d_i)(u_\ell^T d_j) + (z_\ell^T d_j)(u_\ell^T d_i)\right)\right]$$
$$\frac{\partial^2 \mathscr{L}_{|K|}}{\partial \theta_i \partial \theta_j} \approx -\frac{1}{2m}\sum_\ell \left[u_\ell^T w_{\ell,ij} - \frac{1}{m-1}\sum_{\ell' \neq \ell}(u_\ell^T w_{\ell',i})(u_{\ell'}^T w_{\ell,j})\right]$$

Variance Reduction

If lower variance needed, easily available *control variates*.

- Ex: $\check{K} = W_{XU} K_{UU} W_{XU}^T + \Sigma I$, W_{XU} interpolates from U to X points.
- Write $W_{XU} = QR$ (sparse economy).
- Log det and estimate are easy to compute:

$$\log\det \check{K} = \log\det(RK_{UU}R^T + \sigma^2 I) + 2(|X| - |U|)\log\sigma.$$
$$z^T\left(\log \check{K}\right) z = z^T Q \log(RK_{UU}R^T + \sigma^2 I) Q z + (2\log\sigma)(\|z\|^2 - \|Q^T z\|^2)$$

- New estimator for log det of $\tilde{K}$:

$$\log\det \tilde{K} = \mathbb{E}\left[z^T(\log \tilde{K})z - \alpha z^T(\log \check{K})z\right] + \alpha \log\det \check{K}.$$