Stochastic Processes: Lecture 3B

Andrew J. Holbrook

UCLA Biostatistics 270

Spring 2022

Defining GPs

Definition 1

A r.v. X on a Hilbert space H is Gaussian, if for any $h \in H$ the r.v. $\langle X, h \rangle$ follows a Gaussian distribution on $\mathbb{R}$.

Proposition 1

If X is a r.v. on H, then there exist an $m \in H$ and a positive, symmetric, nuclear (trace-class) operator K on H such that:

$$
\begin{aligned}
&E\left(\langle X, h \rangle\right) = \langle m, h \rangle \,, \quad h \in H \,, \\
&E\left(\langle X - m, h \rangle \langle X - m, g \rangle\right) = \langle h, Kg \rangle \,, \quad h, g \in H.
\end{aligned}
$$

m and K are uniquely determined. Conversely, for m and K defined as above, there exists a unique Gaussian distribution on H satisfying these moment conditions.

A definition suited for GP regression

Definition 2
A Gaussian process is a collection of r.v.s, any finite number of which have a joint Gaussian distribution.

A GP $f(x)$ is completely specified by its mean and covariance functions, $m(x)$ and $K(x, x')$. These satisfy

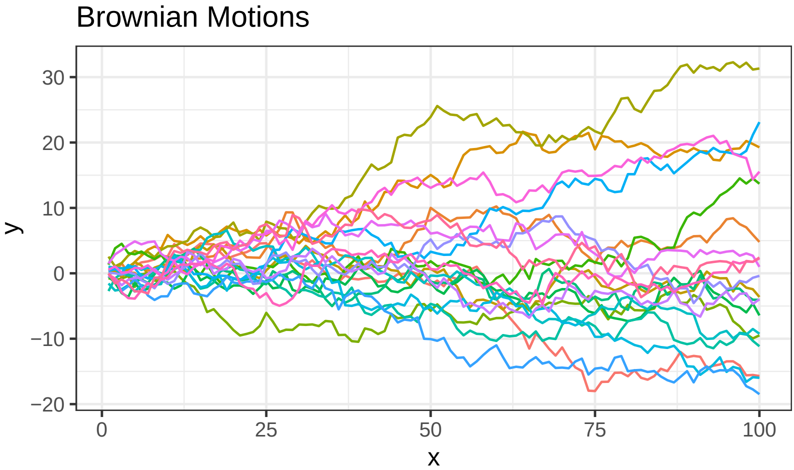
$$\begin{aligned} m(x) &= E(f(x)), \\ K(x, x') &= E\left((f(x) - m(x))\left(f(x') - m(x')\right)\right) . \end{aligned}$$

It is common to use the notation

$$f(x) \sim \mathcal{GP}\left(m(x), K(x, x')\right) .$$

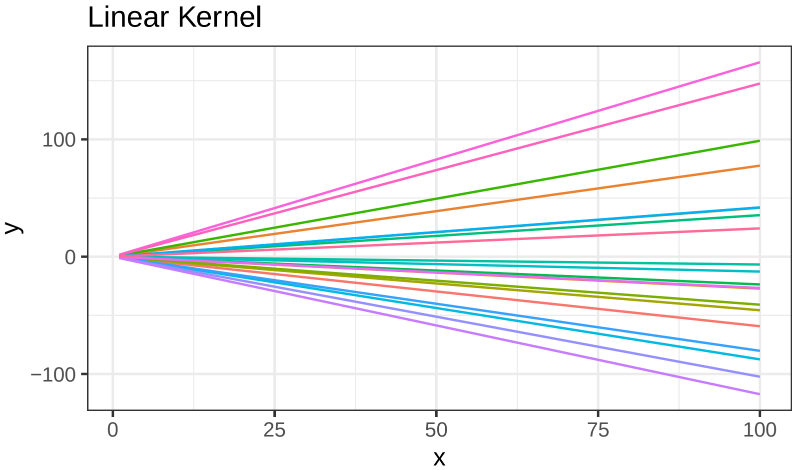
Common Kernels

Brownian motion kernel $K(t, s) = min(t, s)$.



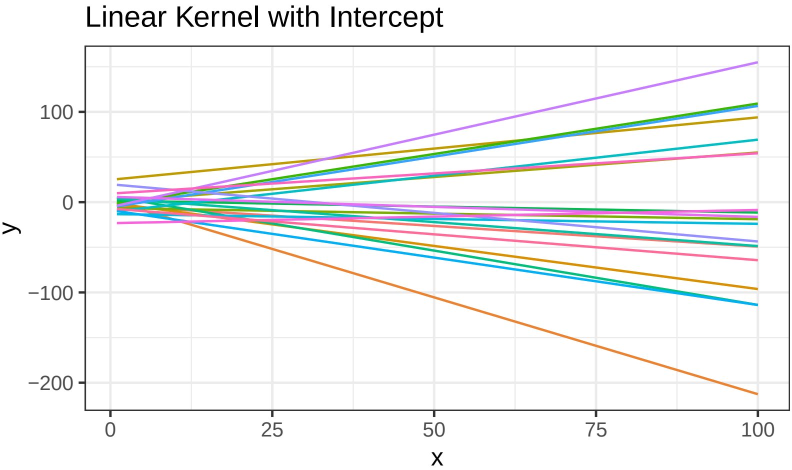
Common Kernels

Linear kernel $K(t,s) = ts$ or $K = \mathrm{x}\mathrm{x}^T$ for vector x.



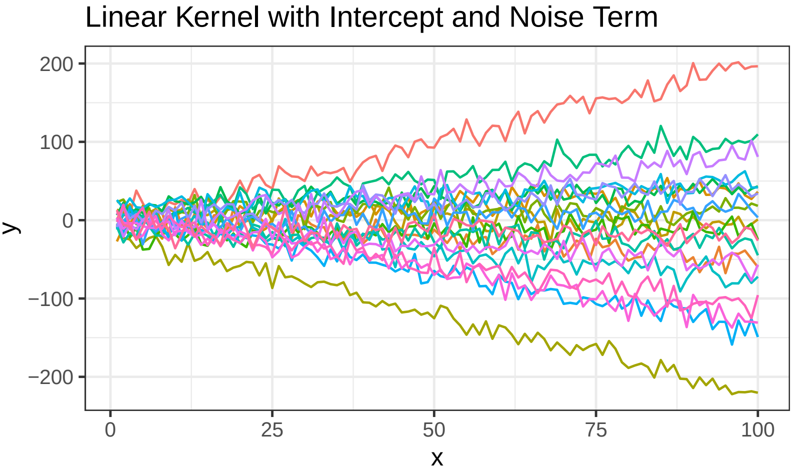
Common Kernels

Linear kernel $K = XX^T$ for X a design matrix $[\alpha\mathbb{1}, \mathrm{x}]$.



Common Kernels

Linear kernel $K = XX^T + \sigma^2 I$ for X a design matrix $[\alpha\mathbb{1}, \mathrm{x}]$.



Common Kernels

Matérn kernel takes $d = d(x, x')$ the distance between x and x':

$$K_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu C_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right).$$

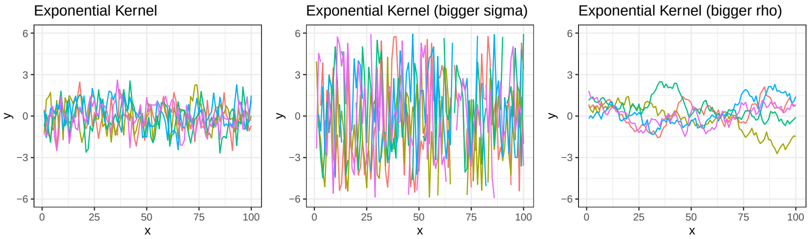
Here, σ^2 is the variance term, C_ν is the modified Bessel function of the second kind, and *rho* is the lengthscale. A GP with covariance K_ν is $\lceil \nu \rceil - 1$ times differentiable.

Examples:

- $\nu = 1/2$: $K_\nu(d) = \sigma^2 e^{-d/\rho}$ (exponential kernel)
- $\nu = 3/2$: $K_\nu(d) = \sigma^2 (1 + \sqrt{3} d/\rho) e^{-\sqrt{3} d/\rho}$
- $\lim_{\nu \to \infty} K_\nu(d) = \sigma^2 e^{-d^2/(2\rho^2)}$
  (squared exponential or radial basis function kernel)

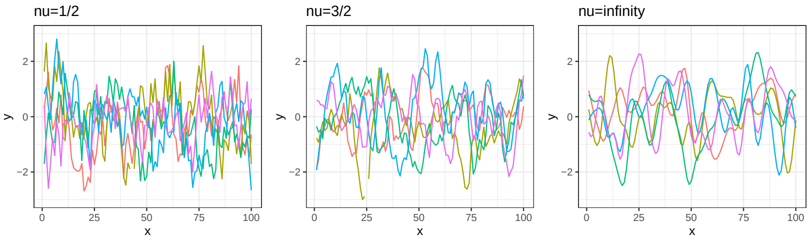
Common Kernels

Exponential kernel $K_{1/2}(d) = \sigma^2 e^{-d/\rho}$.



Common Kernels

Matérn kernels: $\nu = 1/2$, $\nu = 3/2$, $\nu = \infty$.



Building Kernels

Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= ck_1(\mathbf{x}, \mathbf{x}') && (6.13) \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') && (6.14) \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) && (6.15) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) && (6.16) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') && (6.17) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') && (6.18) \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) && (6.19) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}' && (6.20) \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) && (6.21) \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b) && (6.22)
\end{aligned}
$$

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\boldsymbol{\phi}(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$, $k_3(\cdot, \cdot)$ is a valid kernel in $\mathbb{R}^M$, $\mathbf{A}$ is a symmetric positive semidefinite matrix, $\mathbf{x}_a$ and $\mathbf{x}_b$ are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel functions over their respective spaces.

- Bishop (2006). "Pattern Recognition and Machine Learning".

Prediction

Prediction follows from the conditional distribution of a multivariate Gaussian. Write $K(\mathrm{x},\mathrm{x}) := \mathrm{cov}(\mathrm{f}(\mathrm{x}),\mathrm{f}(\mathrm{x}))$ (an abuse of notation) and suppose we have observed

$$\mathrm{f}(\mathrm{x}) \sim N_D(0, K(\mathrm{x},\mathrm{x}))$$

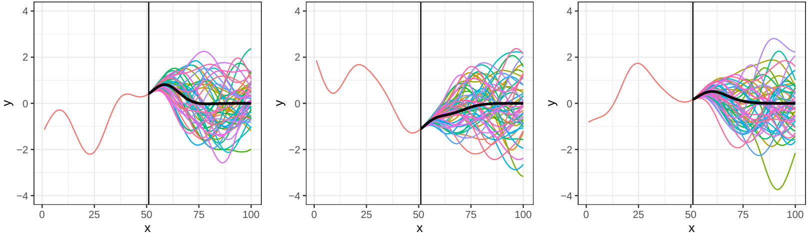
and we would like to predict $\mathrm{f}(\mathrm{x}^*) \sim N_d(0, K(\mathrm{x}^*,\mathrm{x}^*))$. Then we use the joint distribution

$$\begin{pmatrix} \mathrm{f}(\mathrm{x}) \\ \mathrm{f}(\mathrm{x}^*) \end{pmatrix} = N_{D+d}\left(0, \begin{pmatrix} K(\mathrm{x},\mathrm{x}) & K(\mathrm{x},\mathrm{x}^*) \\ K(\mathrm{x}^*,\mathrm{x}) & K(\mathrm{x}^*,\mathrm{x}^*) \end{pmatrix}\right)$$

to conclude that

$$\begin{aligned} \mathrm{f}(\mathrm{x}^*)|\mathrm{f}(\mathrm{x}) \sim N_d\Big(& K(\mathrm{x}^*,\mathrm{x})K(\mathrm{x},\mathrm{x})^{-1}f(\mathrm{x}), \\ & K(\mathrm{x}^*,\mathrm{x}^*) - K(\mathrm{x}^*,\mathrm{x})K(\mathrm{x},\mathrm{x})^{-1}K(\mathrm{x},\mathrm{x}^*)\Big). \end{aligned}$$

Prediction



Bayesian inference

Suppose we've observed N pairs $(\mathrm{f}, \mathrm{X}) = [(f_1, \mathrm{x}_1), \ldots, (f_N, \mathrm{x}_N)]$.
Denote $\boldsymbol{\theta} = (\sigma^2, \rho, \tau^2)$ for the kernel

$$K_\theta(\mathrm{x}_n, \mathrm{x}_{n'}) = \sigma^2 e^{-\frac{1}{2\rho^2}(\mathrm{x}_n - \mathrm{x}_{n'})^2} + \tau^2 \delta_{nn'} \,.$$

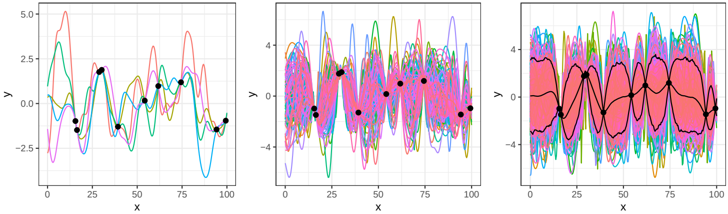
We know $\mathrm{f}|\mathrm{X}, \boldsymbol{\theta} \sim N_N(0, K_\theta(\mathrm{X}))$, where $K_\theta(\mathrm{X}) = [K_\theta(\mathrm{x}_n, \mathrm{x}_{n'})]_{n,n'=1}^N$ is the $N \times N$ covariance matrix. The likelihood is

$$p(\mathrm{f}|\mathrm{X}, \boldsymbol{\theta}) \propto |K_\theta(\mathrm{X})|^{-1/2} e^{-\frac{1}{2}\mathrm{f}^T K_\theta^{-1}(\mathrm{X})\mathrm{f}} \,.$$

Specify priors $p(\sigma^2)$, $p(\rho)$, $p(\tau^2)$ and the posterior becomes

$$p(\boldsymbol{\theta}|\mathrm{f}, \mathrm{X}) \propto |K_\theta(\mathrm{X})|^{-1/2} e^{-\frac{1}{2}\mathrm{f}^T K_\theta^{-1}(\mathrm{X})\mathrm{f}} p(\sigma^2) p(\rho) p(\tau^2) \,.$$

Posterior predictive curves, mean and intervals



Binary classification

Now suppose we've observed N pairs $(y, \mathrm{X}) = [(y_1, \mathrm{x}_1), \ldots, (y_N, \mathrm{x}_N)]$ for y_i binary. We posit Gaussian latent variables $f_1, \ldots, f_n$. Use MCMC to infer

$$
\begin{aligned}
p(\boldsymbol{\theta}, \mathrm{f} | y, \mathrm{X}) \propto & |K_\theta(\mathrm{X})|^{-1/2} e^{-\frac{1}{2}\mathrm{f}^T K_\theta^{-1}(\mathrm{X})\mathrm{f}} \\
& + p(\sigma^2) p(\rho) p(\tau^2) \\
& + \prod_i \left(\frac{e^{f_i}}{1 + e^{f_i}} \right)^{y_i} \left(\frac{1}{1 + e^{f_i}} \right)^{1 - y_i}
\end{aligned}
$$

Note the log of last line simplifies to

$$
\sum_i \left(y_i f_i - \log(1 + e^{f_i}) \right) .
$$

One conditions on posterior samples of $\boldsymbol{\theta}$ and f to get posterior predictive curves $f(x^*)$ for x^* unobserved.

Binary classification

Posterior predictive curves are no longer "pinned" down by observations, since these themselves are inferred latent variabls.