Nonparametric Bayesian Methods

Debdeep Pati
Florida State University

September 30, 2014

Prediction using noisy observations

- It is typical for more realistic modelling situations that we do not have access to function values themselves, but only noisy versions there of $y_i = f(x_i) + \epsilon_i, i = 1, \dots, n$.
- Assuming additive independent identically distributed Gaussian noise with variance σ^2 , the prior on the noisy observations becomes

$$cov(y_p, y_q) = C(x_p, x_q) + \sigma^2 I_{p=q} \implies cov(y) = C(X, X) + \sigma^2 I,$$

- The joint distribution of the observed target values and the function values at the test locations under the prior as

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} C(X,X) + \sigma^2 I & C(X,X_*) \\ C(X_*,X) & C(X_*,X_*) \end{bmatrix}\right)$$

Prediction using noisy observations

- The predictive distribution is

$$f_* \mid X_*, X, y \sim \mathrm{N}(\bar{f}_*, cov(f_*)).$$

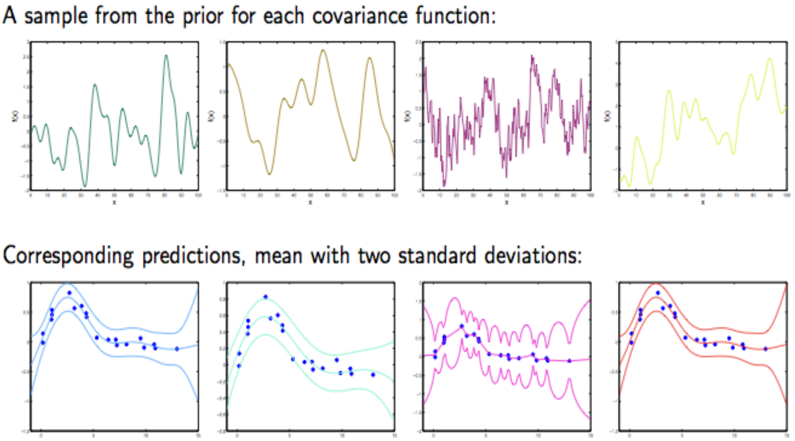
where $\bar{f}_* = E[f_* \mid X, y, X_*] = C(X_*, X)[C(X, X) + \sigma^2]^{-1} y$, and $cov(f_*) = C(X_*, X_*) - C(X_*, X)[C(X, X) + \sigma^2]^{-1} C(X, X_*)$.

- Note first that the mean prediction is a linear combination of observations y; this is sometimes referred to as a linear predictor.
- Another way to look at this equation is to see it as a linear combination of n kernel functions, each one centered on a training point, by writing correspondence with weight-space view compact notation predictive distribution linear predictor representer theorem

$$\bar{f}(x_*) = \sum_{i=1}^{n} \alpha_i C(x_i, x_*), \quad \alpha = (C(X, X) + \sigma^2 I)^{-1} y.$$

Gaussian process predictions using squared exponential cov kernel

Figure: Prediction and predictive intervals



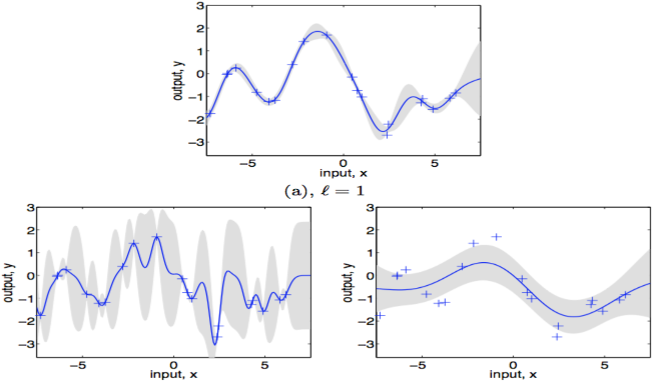
A sample from the prior for each covariance function:

Corresponding predictions, mean with two standard deviations:

Role of hyperparameters

- Typically the covariance functions that we use will have some free parameters.
- For example, the squared-exponential covariance function in one dimension has the following form $C(x_p, x_q) = \sigma_f^2 \exp\{-1/(2l^2)(x_p - x_q)^2\}$.

Gaussian process predictions using squared exponential cov kernel



(a), $\ell = 1$

(b), $\ell = 0.3$

(c), $\ell = 3$

Figure: (a) Data is generated from a GP with hyperparameters $(l, \sigma_f, \sigma_n) = (1, 1, 0.1)$, as shown by the $+$ symbols. Using Gaussian process prediction with these hyperparameters we obtain a 95% confidence region for the underlying function f (shown in grey). Panels (b) and (c) again show the 95% confidence region, but this time for hyperparameter values $(0.3, 1.08, 0.00005)$ and $(3.0, 1.16, 0.89)$ respectively.

Choosing the hyperpriors

- Consider squared-exponential covariance function in one dimension $C(x, x') = \sigma_f^2 \exp\{-A(x - x')^2\}$.
- Conjugate Inverse Gamma hyperprior for σ_f^2, allow heavier tails
- van der Vaart & van Zanten (2008): If $A^d \sim \text{gamma}(a, b)$, optimal rate of convergence adaptively over $C^\alpha[0, 1]^d$ for any $\alpha > 0$. Use Metropolis Hastings algorithm to update A
- Computationally cumbersome, requires matrix evaluation at each stage of the MCMC.
- Use a discrete uniform prior with bounds chosen in such a way that $0.05 < cor(f(x), f(x')) < 0.95$ if $|x - x'| =$ average of the observed intersite distances
- You can save the matrices at the support of the uniform prior before the MCMC.

Series expansion approach

- Mercer's theorem: There exists a sequence of eigenvalues $\lambda_h \downarrow 0$ and an orthonormal system of eigenfunctions ϕ_h, such that

$$C(s,t) = \sum_{h=1}^{\infty} \lambda_h \phi_h(s) \phi_h(t)$$

- Define $\tilde{X}(t) = \sum_{h=1}^{\infty} \lambda_h^{1/2} \; Z_h \; \phi_h(t)$, where Z_h i.i.d. $\mathrm{N}(0,1)$
- $\mathrm{cov}\left(\tilde{X}_s, \tilde{X}_t\right) = \sum_{h=1}^{\infty} \lambda_h \phi_h(s) \phi_h(t) = C(s,t)$
- We can start with a series representation by choosing λ_h and ϕ_h. Different choices lead to splines, neural networks, wavelets, etc

Large spatial datasets

- Large observational and computer-generated datasets:
- Often have spatial and temporal aspects.
- Goal: Make inference on underlying spatial processes from observations at n locations where n is large.

Computational bottleneck

- The posterior predictive involves $(C(X,X)+\sigma^2 I)^{-1}$
- The covariance matrix $C(X,X)$ is large: $n \times n$ for n locations. unstructured: irregular spaced locations. dense: non-negligible correlations.
- Cholesky decomposition of $n \times n$ matrices Generally requires $O(n^3)$ computations and $O(n^2)$ memory.

Options for large n

- Use models that reduce computations and/or storage. Use approximate methods.
- Compactly supported covariance functions.
- Reduced rank covariance functions.
- Leads Statistical and computational efficiency.