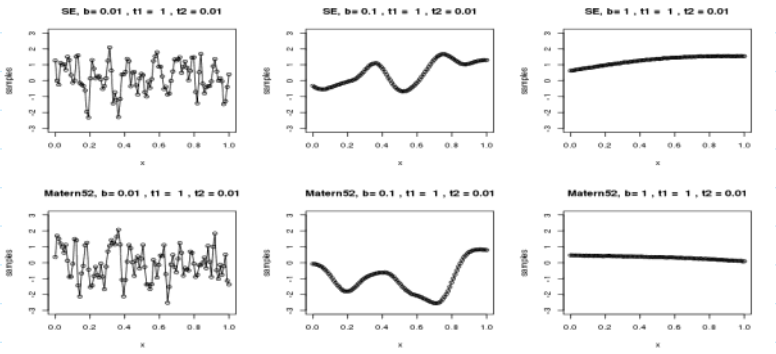
Ch. 3 Gaussian Processes

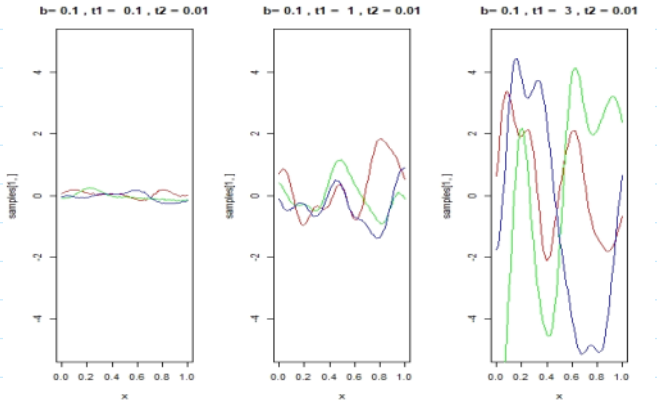
Friday, March 8, 2019 2:06 PM

A) See code

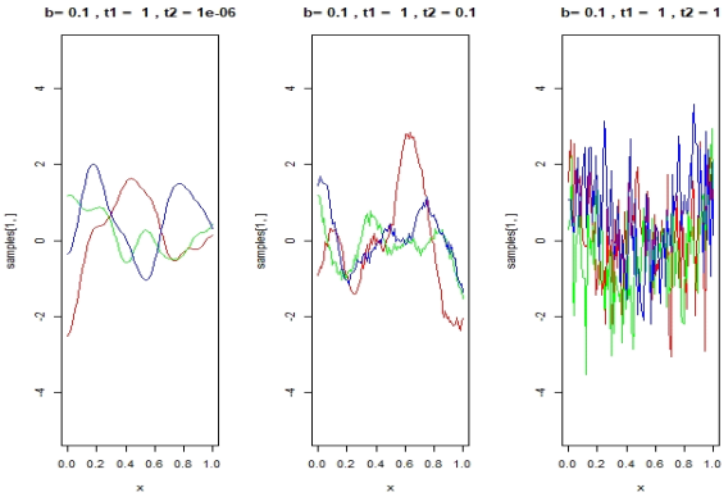
The b value is the bandwidth. It controls how "wiggly" the draw is. A very small value oscillates a lot whereas a very large value produces very smooth and flat graphs.



τ_1 is the amplitude parameter. It controls the amplitude of the oscillations. Smaller values of τ_1 produces very small oscillations and vice versa.



τ_2 is the white noise parameter. If $\tau_2 = 0$, the covariance matrix becomes PSD instead of PD so the cholesky decomposition will fail. In this case, use spectral decomposition



B) Suppose you observe the value of a Gaussian process $f \sim GP(m, C)$ at some points $x_1, \ldots, x_n$. What is the conditional distribution of the value of the process at some new point x^*? Denote the value of the (i,j) element of the covariance matrix as C_{ij}.

The interpretation of the function space view of GPs works as follows. Suppose we have a dependent variable y that can be modeled as $y = f(x) + \varepsilon$ where $f(x)$ has an unknown and potentially infinite number of parameters. The GP approach is a non parametric approach that finds the distribution over all possible functions $f(x)$ that are consistent with the observed data. The GP defines a prior for the distribution of functions that we can specify using a mean function & covariance kernel. The covariance matrix ensures that values that are close together in input space will produce output values that are close together. Furthermore, we only need to be able to define a distribution's values at a finite, but arbitrary set of points $x_{1:n}$. A GP assumes that $p(f(x_1), \ldots, f(x_n))$ is jointly Gaussian with some mean $\mu(x)$ and covariance $C(x)$, $C_{ij} = k(x_i, x_j)$ where k is a positive definite kernel function.

The observed data allows us to derive a posterior together with the prior by utilizing the joint normality of multivariate Gaussians:

$$\begin{bmatrix} f_* \\ f \end{bmatrix} \sim MVN\left(\begin{pmatrix} m \\ m_* \end{pmatrix}, \begin{bmatrix} K(x_*, x_*) & K(x, x_*) \\ K(x_*, x) & K(x, x) \end{bmatrix}\right)$$

Where f denotes training observations $f(x)$ and f_* denotes testing observations $f(x_*)$. To force this joint posterior to contain only those functions that agree with the observed data points we condition the joint Gaussian prior distribution on the observations.

To find: $p(f_* | X_*, X, f)$

We have $p(f_*, f \mid X, X_*) \sim MVN\left(\begin{pmatrix} m \\ m_* \end{pmatrix}, \begin{bmatrix} K(X_*, X_*) & K(x, x_*) \\ K(x_*, x) & K(x, x) \end{bmatrix}\right)$. Therefore, to find $p(f_* | X_*, X, f)$, we can use $p(f_* | f) = \frac{p(f_*, f)}{p(f)}$.

Let $\Sigma = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ and let $\Omega = \Sigma^{-1}$

$$p(f, f_* | X, X_*) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix}^T \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}^{-1} \begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix}\right)$$

$$p(f | X, X_*) = \frac{1}{|2\pi C_{22}|^{1/2}} \exp\left(-\frac{1}{2} f^T C_{22}^{-1} f\right)$$

$$\ln\left(\frac{p(f, f_*)}{p(f)}\right) = \ln(p(f, f_*)) - \ln(p(f))$$

$$= \underbrace{-\frac{1}{2}\ln(2\pi\Sigma)}_{a_1} + \left(-\frac{1}{2}\begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix}^T \Omega \begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix}\right) - \left(\underbrace{\frac{-1}{2}\ln(2\pi C_{22})}_{a_2} - \frac{1}{2} f^T C_{22}^{-1} f\right)$$

$$= a_1 - a_2 - \frac{1}{2}\left[\begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix}^T \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \begin{pmatrix} f_* - m_* \\ f - m \end{pmatrix} - f^T C_{22}^{-1} f\right]$$

$$= a_1 - a_2 - \frac{1}{2}\left[(f_* - m_*)^T \Omega_{11} f_* - m_* + (f - m)^T \Omega_{21}(f_* - m_*) + (f_* - m_*)^T \Omega_{12}(f - m) + (f - m)^T \Omega_{22}(f - m) - f^T C_{22}^{-1} f\right]$$

We are only interested in the pieces that contain f_* so let $(f^T - m)\Omega_{22}(f - m) - f^T C_{22}^{-1} f = a_3$

we are only interested in the pieces that contain f_* so let $(f^T - m)\Omega_{22}(f-m) - f^T C_{22}^{-1} f = a_3$

$$= a_1 - a_2 - \frac{1}{2}\left[(f-m)^T\Omega_{21}(f_*-m^*) + (f_*-m^*)^T\Omega_{12}(f-m) + (f_*-m^*)^T\Omega_{11}(f_*-m^*) + a_3\right]$$

$$= a_1 - a_2 - \frac{1}{2}\left[(f_*-m^*)^T\Omega_{11}(f_*-m^*) - 2(f_*-m^*)^T\Omega_{12}(f-m) + a_3\right]$$

at this point, I will drop all the pieces that don't have to do with f_*

$$\ln(p(f_*|f,X,X_n) \propto -\frac{1}{2}\left[(f_*-m^*)^T\Omega_{11}(f_*-m^*) - 2(f_*-m^*)^T\Omega_{12}(f-m)\right]$$

I will complete the square so that this is in the form $-\frac{1}{2}\left((f_*-v)^T A (f_*-v)\right]$

$$\ln(p(f_*|f,X,X_n) \propto -\frac{1}{2}\left[f_*\Omega_{11}f_* - 2f_*^T\underline{\Omega_{12}(f-m)}\right]$$ I once again drop any term without f_*

$\Omega_{11}V \Rightarrow V = \Omega_{11}^{-1}\Omega_{12}(f-m)$

$$\propto -\frac{1}{2}\left[(f_*-\Omega_{11}^{-1}\Omega_{12}(f-m))^T\Omega_{11}(f_*-\Omega_{11}^{-1}\Omega_{12}(f-m))\right]$$

From a previous excerise, we know

$$\Omega_{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}$$
$$\Omega_{12} = -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

$$\Omega_{11}^{-1}\Omega_{12} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})(-\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$
$$= \Sigma_{12}\Sigma_{22}^{-1}$$

Thus, we have $f_* \sim MVN(C_{12}C_{22}^{-1}(f-m), C_{11} - C_{12}C_{22}C_{21})$ because Ω_{11}^{-1} is the inverse of the new covariance

$$= MVN(K(X_*,X)K(X,X)^{-1}(f-m), K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$$

C) Prove:

Lemma 1. Suppose that the joint dist'n of two vectors y and θ has the following properties:

1.) the conditional distribution for y given θ is multivariate normal, $(y|\theta) \sim N(R\theta, \Sigma)$; and

2.) the marginal dist'n of θ is multivariate normal, $\theta \sim N(m, V)$. Assume that R, Σ, m, and V are all constants.

then the joint dist'n of y and θ is multivariate normal.

Suppose y and θ have the properties given in the lemma. Then the joint density of y and θ is

$$J = \begin{pmatrix} y \\ \theta \end{pmatrix} = \begin{pmatrix} R \\ I \end{pmatrix}\theta + \begin{pmatrix} I \\ 0 \end{pmatrix}\varepsilon \quad \text{where} \quad \theta \sim N(m,V), \; \varepsilon \sim N(0,\Sigma)$$

$$E(y|\theta) = R\theta + I0$$

$$Cov(y|\theta) = 0 + Cov(\varepsilon) = \Sigma$$

$$E(J) = \begin{pmatrix} R \\ I \end{pmatrix}m + \begin{pmatrix} I \\ 0 \end{pmatrix}0 = \begin{pmatrix} Rm \\ m \end{pmatrix}$$

$$
\begin{aligned}
\mathrm{Cov}(\bar{J}) &= \begin{pmatrix} R \\ I \end{pmatrix} V \begin{pmatrix} R \\ I \end{pmatrix}^{\top} + \begin{pmatrix} I \\ 0 \end{pmatrix} (\Sigma) \begin{pmatrix} I \\ 0 \end{pmatrix}^{\top} \\
&= \begin{pmatrix} RV \\ V \end{pmatrix} \begin{pmatrix} R^{\top} & I^{\top} \end{pmatrix} + \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} \begin{pmatrix} I & 0 \end{pmatrix} \\
&= \begin{pmatrix} RVR^{\top} & RV \\ VR^{\top} & V \end{pmatrix} + \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} \Sigma + RVR^{\top} & RV \\ VR^{\top} & V \end{pmatrix}
\end{aligned}
$$