The Laplace Approximation of Gaussian Process Mixture

Zhengdong Lu
Department of Computer Science and Engineering
OGI School of Science and Engineering , OHSU
{zhengdon}@csee.ogi.edu

Various models of Gaussian process mixture have been proposed, mainly to address the non-stationarity in regression [1, 2, 3]. Typically, a Gaussian process mixture consists of M Gaussian processes $F = \{f_1, f_2, \cdots, f_M\}$ with zero mean and covariance function specified by parameters $\Theta = \{\theta_1, \theta_2, \cdots, \theta_M\}$, and a gating network g. For any input x, $g(x) = [g_1(x), g_2(x), \cdots, g_M(x)]$ with

$$g_m(x) \geq 0, m = 1, 2, \cdots, M \text{ and } \sum_{m=1}^{M} g_m(x) = 1.$$

As in [1], the probability model is as follows. Given input $X = \{x_1, x_2, ..., x_N\}$, we define the latent variable $Z = \{z_1, z_2, ..., z_N\}$ to indicate which process each x_i is involved in. The probability of Z, given the gating network g and input X, is

$$P(Z|X, g) = \prod_{i=1}^{N} P(z_i|x_i, g) = \prod_{i=1}^{N} g_{z_i}(x_i). \tag{1}$$

The probability of the output $Y = \{y_1, y_2, ..., y_N\}$ is

$$P(Y|X, g, \Theta) = \sum_{Z} P(Y, Z|X, g, \Theta) = \sum_{Z} P(Y|X, Z, \Theta)P(Z|X, g) \tag{2}$$

where $P(Y|X, Z, \Theta) = \prod_{m=1}^{M} P(\{y_i : z_i = m\}|\{x_i : z_i = m\}; \theta_m)$. Here, $P(\{y_i : z_i = m\}|\{x_i : z_i = m\}; \theta_m)$ is the Gaussian distribution of all the y_i generated by f_m according to Z. The Gaussian process mixture given in equation (1)-(2) is suitable for situation where the domain consists of different regimes that should be described by different types of Gaussian processes. Unfortunately the regression based on equation (1)-(2) is generally intractable due to the exponential number of summations in equation (2). Traditionally this problem is solved by sampling [1]. Our method, instead, considers the Laplace approximation of $P(Y|X, g, \Theta)$, which is a Gaussian process with zero mean and the covariance function as the GP mixture $P(Y|X, g, \Theta)$. Interestingly, this approximation becomes exact when the gating network g gives us a hard partition of domain [3]. Easy to show that the covariance function of Gaussian process mixture is:

$$E(y_i y_j|x_i, x_j, g, \Theta) \quad = \quad \sum_{m=1}^{M} K_m(x_i, x_j)P(z_i = m, z_j = m|x_i, x_j, g) \tag{3}$$

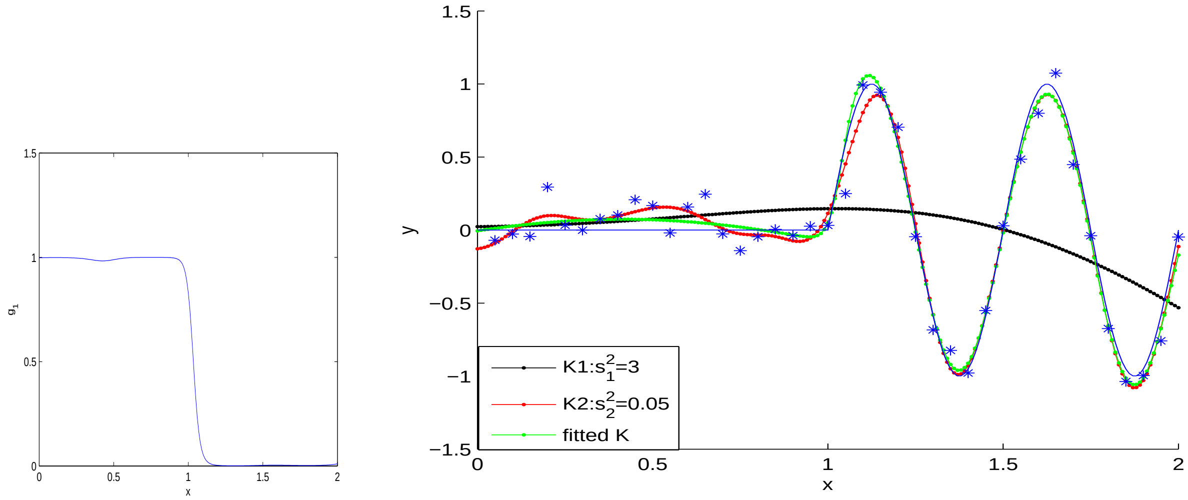
where $K_m(x_i, x_j)$ is the covariance function associated with process f_m. Using $\hat{K}(\cdot, \cdot; g, \Theta)$ to denote the covariance function given in equation (3), our Laplace approximation of the likelihood of Y is $P_L(Y|X, \hat{K}) = \frac{1}{\sqrt{2\pi^N|\hat{K}_X+\sigma^2\mathbf{I}|}}$ $\exp(-\frac{1}{2}y'(\hat{K}_X + \sigma^2\mathbf{I})^{-1}y)$, where σ^2 is the variance of observation noise and $\hat{K}_X$ is the covariance matrix evaluated on X. We find a suitable gating network by maximizing the data likelihood P_L.

Figure 1 and 2 show the regression results of our model on two toy examples. In these two experiments, we considered the mixture of two Gaussian processes with covariance function of process f_m, $(m = 1, 2)$ specified as

$$K_m(x_i, x_j) = \exp(-||x_i - x_j||^2/s_m^2), \ \ m = 1, 2$$

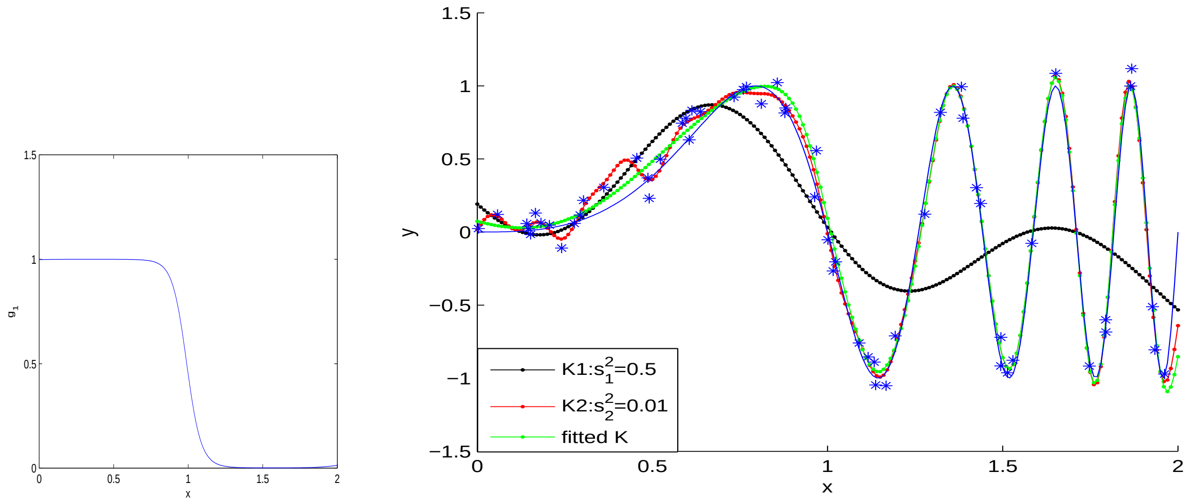
where s_m is the width of the RBF kernel for the m^{th} Gaussian process, which are chosen beforehand. We use the same gating network used in [2] and fit it with gradient descent. We report the regression result given by the probability $P_L(y|x, X, Y, \Theta, \hat{K})$. As shown in the two figures, our model automatically finds the appropriate gating network and yields regression result significantly better than the Gaussian process regression with either f_1 or f_2.

Topic: Learning Algorithms
Preference: oral



(a) Output of g_1 (b) Regression result

Figure 1: The regression result.(b)Blue line: true curve. Blue stars: noisy observation. Black curve: $f_1, s_1^2 = 3$; Red curve: $f_2, s_2^2 = 0.05$; Green curve: Laplace Approximation of mixture of $\{f_1, f_2\}$ with fitted K.



(a) Output of g_1 (b) Regression result

Figure 2: The regression result. (b) Blue line: truth. Blue stars: noisy observation. Black curve: $f_1, s_1^2 = 0.5$; Red curve: $f_2, s_2^2 = 0.01$; Green curve: Laplace Approximation of mixture of $\{f_1, f_2\}$ with fitted K.

References

[1] C. Rasmussen and Z. Ghahramani. Infinite mixtures of gaussian process experts. In *NIPS 14*, 2002.

[2] V. Tresp. Mixtures of gaussian processes. In *NIPS 13*, 2001.

[3] O. Walliams. A Switched Gaussian Process for Estimating Disparity and Segmentation in Binocular Stereo. In *NIPS 19*, 2007.