Variational Inference for Gaussian Process Regression

Daeyoung Lim*
Department of Statistics
Korea University

March 8, 2016

1 Gaussian Process

Gaussian process is a stochastic process specified by a mean function and a covariance function. It could simply be thought of as an infinite-dimensional version of multivariate Gaussian distribution. The multivariate Gaussian random vector consists of infinitely many Gaussian variables. This makes impossible the formation of a mean vector and a covariance matrix since the dimension of the vector and matrix is infinite. We now need sort of a *formula* to consistently compute the elements of the infinite-dimensional mean vector and covariance vector, which are both achieved by constructing functions. A random process that follows Gaussian process is denoted as follows:

$$X(\omega, t) \sim \mathcal{GP}(m(\cdot), \kappa(\cdot)).$$

Without loss of generality, we assume the mean of the process is zero. Then the whole process is completely defined by the covariance function. To construct a covariance matrix of any set of indices, we use the covariance function, K_{ij}:

$$K_{ij} = \mathrm{E}[X_i X_j] = \kappa(x_i, x_j).$$

The covariance function is called "*isotropic*" if it only depends on $|x_i - x_j|$.

2 Mercer's Theorem

The kernel(covariance function), $\kappa(x_i, x_j)$, is a symmetric continuous function

$$\kappa : [a, b] \times [a, b] \to \mathbb{R}$$

where symmetric implies $\kappa(x_i, x_j) = \kappa(x_j, x_i)$. κ is "*positive semidefinite*" as in linear algebra if and only if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j) c_i c_j \geq 0$$

for all $\{(x_i, x_j) : x_i \in [a, b] \text{ and } x_j \in [a, b]\}$ and all choices of $\{(c_i, c_j) : c_i \in \mathbb{R} \text{ and } c_j \in \mathbb{R}\}$. For κ, we can assign a linear operator T_κ on functions defined as follows:

$$[T_\kappa \varphi](x) = \int_a^b \kappa(x, s) \varphi(s)\, ds$$

*Prof. Taeryon Choi

where φ is square-integrable, real-valued functions. (i.e. $\varphi \in L^2[a,b]$) Since T_κ is a linear operator, it falls within the scope of functional analysis that discusses eigenvalues and eigenfunctions of T_κ. Mercer's theorem states that for a positive semidefinite kernel, there exists an orthonormal basis e_k of $L^2[a,b]$ which consists of the eigenfunctions of T_κ whose corresponding eigenvalues λ_k are nonnegative. The spectral decomposition of the kernel is as follows:

$$\kappa(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s)\, e_j(t)$$

whose convergence is absolute and uniform. A more formal statement is given as follows:

Theorem 2.1. *(Mercer's theorem). Let $(\mathcal{X}, \mu)$ be a finite measure space and $\kappa \in L_\infty\left(\mathcal{X}^2, \mu^2\right)$ be a kernel such that $T_\kappa : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ is positive definite. Let $e_i \in L_2(\mathcal{X}, \mu)$ be the normalized eigenfunctions of T_κ associated with the eigenvalues $\lambda_i > 0$. Then:*

1. *the eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ are absolutely summable*
2. $$\kappa\left(x, x'\right) = \sum_{i=1}^{\infty} \lambda_i e_i(x)\, e_i^*\left(x'\right)$$

   holds μ^2 almost everywhere, where the series converges absolutely and uniformly μ^2 almost everywhere.

Replacing the finite measure μ with Lebesgue measure for a stationary covariance function, we obtain

$$\kappa\left(x - x'\right) = \int_{\mathbb{R}^d} e^{2\pi i s^T (x - x')}\, d\mu(s) = \int_{\mathbb{R}^d} e^{2\pi i s^T x} \left(e^{2\pi i s^T x'}\right)^*\, d\mu(s).$$

3 Karhunen-Loève Theorem

Karhunen-Loève(KL) expansion is a version of Fourier series for stochastic processes which differ from each other in that the Fourier series uses sinusoidal basis functions whereas the KL expansion depends on the eigenfunctions of the covariance function. This theorem follows from the Mercer's theorem and the statement goes like this. Let X_t be a zero-mean square-integrable stochastic process indexed over a closed interval $[a,b]$ with a continuous covariance function$\kappa_X(s,t)$. Then, the covariance function is a Mercer kernel with eigenfunctions e_k being an orthonormal basis of $L^2[a,b]$. It allows the following representation:

$$X_t = \sum_{k=1}^{\infty} Z_k e_k(t)$$

where $Z_k = \int_a^b X_t e_k(t)\, dt$ are uncorrelated random variables. Especially when X_t is a Gaussian process, Z_k are independent Gaussian random variables with mean 0.

4 GP regression

Gaussian process is considered to be a dense set of functions. This lends itself directly to the nonparametric setting of Bayesian regression because we do not need to specify any parametric

form of the regression function. The paper considers the stationary squared exponential covariance function,

$$\kappa(h) = \sigma^2 \exp\left(-\frac{1}{2} h^T \Lambda h\right)$$

where Λ is a diagonal matrix whose diagonal entries are $\begin{bmatrix}\lambda_1^2 & \dots & \lambda_d^2\end{bmatrix}$. Now if we consider a random sample $\{s_1, \dots, s_m\}$ from $\mathcal{N}(0, I_d)$, then $\left\{\frac{1}{2\pi}\Lambda^{1/2} s_1, \dots, \frac{1}{2\pi}\Lambda^{1/2} s_m\right\}$ is a random sample from $p_k(s)$. (i.e. $p_k(s)$ is proportional to the power spectral density $S_k(s)$ which is $S_k(s) = \int_{\mathbb{R}^d} e^{-2\pi i s^T h} \kappa(h)\, dh$ and $S_k(s) = \kappa(0)\, p_k(s)$.) The sparse GP approximation is then

$$f(x) \approx \sum_{r=1}^{m} \left[a_r \cos\left\{(s_r \circ x)^T \lambda\right\} + b_r \sin\left\{(s_r \circ x)^T \lambda\right\}\right]$$

where $\lambda = \begin{bmatrix}\lambda_1 & \dots & \lambda_d\end{bmatrix}^T$ and $\circ$ denotes the Hadamard product. The author converts the approximation into a matrix form by construction a coefficient vector and a matrix similar to ordinary design matrix. By doing so, the GP regression reduces to a simple Bayesian regression.

$$\begin{aligned}
\alpha &= \begin{bmatrix}a_1 & \dots & a_m & b_1 & \dots & b_m\end{bmatrix}^T \\
y &= \begin{bmatrix}y_1 & \dots & y_n\end{bmatrix}^T \\
Z &= \begin{bmatrix}Z_1 & \dots & Z_n\end{bmatrix}^T \\
Z_i &= \begin{bmatrix}\cos\left\{(s_1 \circ x_i)^T \lambda\right\} & \dots & \cos\left\{(s_m \circ x_i)^T \lambda\right\} & \sin\left\{(s_1 \circ x_i)^T \lambda\right\} & \dots & \sin\left\{(s_m \circ x_i)^T \lambda\right\}\end{bmatrix}^T \\
\epsilon &= \begin{bmatrix}\epsilon_1 & \dots & \epsilon_n\end{bmatrix}^T
\end{aligned}$$

Now the model formulation converts to

$$y = Z\alpha + \epsilon, \;\; \epsilon \sim \mathcal{N}\left(0, \gamma^2 I_n\right).$$

In other words,

$$y | Z, \alpha, \gamma, \sigma, \lambda \sim \mathcal{N}\left(Z\alpha, \gamma^2 I_n\right).$$

We assign the following priors:

- $\alpha \sim \mathcal{N}\left(0, \frac{\sigma^2}{m} I_{2m}\right)$
- $\lambda \sim \mathcal{N}\left(\mu_\lambda^0, \Sigma_\lambda^0\right)$
- $\sigma \sim$ half-Cauchy (A_σ)
- $\gamma \sim$ half-Cauchy (A_γ)

For variational approximation, we also assign variational distributions for the unknown parameters:

- $q(\alpha) = \mathcal{N}\left(\mu_\alpha^q, \Sigma_\alpha^q\right)$
- $q(\lambda) = \mathcal{N}\left(\mu_\lambda^q, \Sigma_\lambda^q\right)$
- $q(\sigma) = \frac{\exp\left(-C_\sigma^q/\sigma^2\right)}{\mathcal{H}\left(2m-2, C_\sigma^q, A_\sigma^2\right)\sigma^{2m}\left(A_\sigma^2+\sigma^2\right)}$
- $q(\gamma) = \frac{\exp\left(-C_\gamma^q/\gamma^2\right)}{\mathcal{H}\left(2m-2, C_\gamma^q, A_\gamma^2\right)\gamma^{2m}\left(A_\gamma^2+\gamma^2\right)}$

- $\mathcal{H}(p,q,r) = \int_0^\infty x^p \exp\left\{-qx^2 - \log\left(r + x^{-2}\right)\right\}$ (the normalizing constant)

Now that we've arrived at a full package for variational approximation, what is left is computing the lower bound and the updating algorithms for the variational parameters. However, the factor here that makes variational approximation difficult is that the priors are not conditionally conjugate. This calls for nonconjugate variational approximation.

4.1 A few calculations

Before we move on to the next section, there are some points where I had been wondering how to compute things such as the expectation of the variational posterior of σ and γ. So I feel the need to clear up those issues. To do that, we have to take a look at the function $\mathcal{H}(p,q,r)$.

$$\mathcal{H}(p,q,r) = \int_0^\infty x^p \exp\left\{-qx^2 - \log\left(r + x^{-2}\right)\right\} dx, \;\; p \geq 0, \; r > 0$$

If we reparameterize the integration by plugging in $x = 1/\sigma$, which by symmetry also clears up $x = 1/\gamma$, it becomes

$$\begin{aligned} \mathcal{H}(p,q,r) &= \int_\infty^0 \left(\frac{1}{\sigma}\right)^p \exp\left\{-q/\sigma^2\right\} / \left(r + \sigma^2\right) \left(-\frac{1}{\sigma}^2\right) d\sigma \qquad (1) \\ &= \int_0^\infty \frac{\exp\left\{-q/\sigma^2\right\}}{\sigma^{p+2}\left(r + \sigma^2\right)} d\sigma \qquad (2) \end{aligned}$$

Note that the integration bounds are switched in eqn (1) because $\sigma \to \infty$ as $x \to 0$ and $\sigma \to 0$ as $x \to \infty$. Also $dx = -d\sigma/\sigma^2$.

4.1.1 $\mathbb{E}_q[\sigma]$

$$\begin{aligned} \mathbb{E}_q[\sigma] &= \int_0^\infty \sigma \frac{\exp\left\{-C_\sigma^q/\sigma^2\right\}}{\mathcal{H}\left(2m-2, C_\sigma^q, A_\sigma^2\right)\sigma^{2m}\left(A_\sigma^2 + \sigma^2\right)} d\sigma \\ &= \frac{1}{\mathcal{H}\left(2m-2, C_\sigma^q, A_\sigma^2\right)} \int_0^\infty \frac{\exp\left(-C_q^q/\sigma^2\right)}{\sigma^{2m-1}\left(A_\sigma^2 + \sigma^2\right)} d\sigma \\ &= \frac{\mathcal{H}\left(2m-3, C_\sigma^q, A_\sigma^2\right)}{\mathcal{H}\left(2m-2, C_\sigma^q, A_\sigma^2\right)} \end{aligned}$$

4.1.2 $\mathbb{E}_q[\gamma]$

By symmetry,

$$\mathbb{E}_q[\gamma] = \frac{\mathcal{H}\left(n-3, C_\gamma^q, A_\gamma^2\right)}{\mathcal{H}\left(n-2, C_\gamma^q, A_\gamma^2\right)}$$

4.1.3 $\mathbb{E}\left[\sin\left(s^T X\right)\right]$, $\mathbb{E}\left[\cos\left(s^T X\right)\right]$

Let $X \sim \mathcal{N}(\mu, \Sigma)$.

$$\mathbb{E}\left[e^{is^T X}\right] = \mathbb{E}\left[\cos\left(s^T X\right)\right] + i\mathbb{E}\left[\sin\left(s^T X\right)\right] \tag{3}$$

$$= \exp\left\{i\mu^T s - \frac{1}{2}s^T \Sigma s\right\} \text{characteristic function} \tag{4}$$

$$= \exp\left\{i\mu^T s\right\} \exp\left\{-\frac{1}{2}s^T \Sigma s\right\} \tag{5}$$

$$= \left(\cos\left(\mu^T s\right) + i\sin\left(\mu^T s\right)\right) \exp\left\{-\frac{1}{2}s^T \Sigma s\right\} \tag{6}$$

Comparing the real part and imaginary part of eqn (3) and (6),

$$\mathbb{E}\left[\cos\left(s^T X\right)\right] = \cos\left(\mu^T s\right) \exp\left\{-\frac{1}{2}s^T \Sigma s\right\}$$

$$\mathbb{E}\left[\sin\left(s^T X\right)\right] = \sin\left(\mu^T s\right) \exp\left\{-\frac{1}{2}s^T \Sigma s\right\}$$

5 Nonconjugate VB

Mean-field approximation basically assumes that the priors are conditionally conjugate. However, for nonconjugate priors, Knowles and Minka, 2011 suggest a variational message passing algorithm by adding another assumption that the nonconjugate priors belong to one of exponential families. So let's say θ_i is a nonconjugate prior.

$$q_i(\theta_i) = \exp\left\{\eta_i^T T_i(\theta_i) - h_i(\eta_i)\right\}$$

The author uses factor graph to explain nonconjugate variational message passing algorithm but factor graph in the paper is simply collecting parts that contain θ_i from the full joint density $p(y, \theta)$. Now we can define the following:

$$\mathcal{V}_i(\eta_i) = \frac{\partial^2 h_i(\eta_i)}{\partial \eta_i \partial \eta_i^T}$$

$$\eta_i \leftarrow \mathcal{V}_i(\eta_i)^{-1} \sum_{a \in N(\theta_i)} \frac{\partial S_a}{\partial \eta_i}$$

where $S_a = \mathbb{E}_q\left[\log f_a(y, \theta)\right]$. In summary, collect all the parts that have θ_i, take the logarithm and expectation, differentiate with respect to η_i, and multiply it by the inverse of natural gradient.