Advanced Bayesian Computation Weeks 5

Rajarshi Guhaniyogi
Winter 2018

February 9, 2018

Choice of Covariance Functions

- There are certain realistic assumptions often employed on the covariance function $C_{\boldsymbol{\theta}}(t_i, t_j)$.
- Stationary: $C_{\boldsymbol{\theta}}(t_i, t_j) = C_{\boldsymbol{\theta}}(t_i - t_j)$; Isotropic: $C_{\boldsymbol{\theta}}(t_i, t_j) = C_{\boldsymbol{\theta}}(||t_i - t_j||)$.
- The covariance function of a stationary process can be represented as the Fourier transform of a positive finite measure.
- This is the famous Bochner's theorem.
- Let $h = t_i - t_j$, A real-valued function $C_{\boldsymbol{\theta}}(h)$ on $\mathbb{R}^D$ is the covariance function of a stationary real valued random process on $\mathbb{R}^D$ if and only if it can be represented as

$$C_{\boldsymbol{\theta}}(h) = \int cos(2\pi h.s)dH(s),$$

where $H(s)$ is a positive finite measure.

Choice of Covariance Functions

- If $H(s)$ has a density $S(s)$, then $S(s)$ is called the spectral density.
- Note that $C_{\boldsymbol{\theta}}$ and $S(s)$ are Fourier-dual of each other, i.e.

$$C_{\boldsymbol{\theta}}(h) = \int cos(2\pi h.s) dS(s) ds, \;\; S(s) = \int cos(-2\pi h.s) C_{\boldsymbol{\theta}}(h) dh.$$

- Most of the practical applications we take the covariance kernel as an isotropic kernel.
- **Squared Exponential Covariance**: $C_{\boldsymbol{\theta}}(r) = \exp\left(-\frac{r^2}{\phi}\right)$.
- The spectral density is given by $S(s) = (\sqrt{2\pi\phi})^D \exp(-2\pi^2 \phi s^2)$.
- The sample path is infinitely differentiable.
- This is the most popularly used covariance kernel in the machine learning literature.

Choice of Covariance Functions

- **Matern Covariance**:

$$C_{\theta}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\phi} \right) K_{\nu} \left(\frac{\sqrt{2\nu}r}{\phi} \right),$$

  K_{ν} is the modified Bessel function.
- The spectral density is
- Here ν is called the smoothness parameter which determines the smoothness of the sample path.
- As ν is increased, the sample paths are more smooth.
- As $\nu \to \infty$, Matern covarinace kernel converges to the squared exponential covariance kernel.
- If k is the greatest integer less than ν, then the Gaussian process is k times mean square differentiable.

Choice of Covariance Functions

- **Exponential Covariance**:

$$C_{\boldsymbol{\theta}}(r) = \exp\left(-\frac{r}{2\phi}\right).$$

- This is a special case of the Matern covariance kernel with $\nu = 1/2$.
- The sample patha is only continuous, not even differentiable once.
- In the one dimensional case this is the covariance function of the Ornstein-Uhlenbeck (OU) process.
- The OU process [Uhlenbeck process and Ornstein, 1930] was introduced as a mathematical model of the velocity of a particle undergoing Brownian motion.
- Gaussian process with the exponential kernel is not even once mean square differentiable.

Choice of Covariance Functions

- **Rational Quadratic Covariance**:

$$C_{\boldsymbol{\theta}}(r) = \left(1 + \frac{r^2}{2\alpha\phi}\right)^{-\alpha}$$

- This is a scale mixture of squared exponential kernel.
- Sometimes used for a greater flexibility over squared exponential.
- The Gaussian process is infinitely mean squared differentiable for every α.
- As $\alpha \to \infty$, the rational quadratic function takes more and more the shape of a squared exponential covariance function.

Constructing More Complicated Covariance Functions

- Rather than using an isotropic function, one may want to use stationary covariance function.
- Define the distance metric $r(t_i, t_j)^2 = (t_i - t_j)' M (t_i - t_j)$, M is a positive definite matrix.
- Now replace any of the already defined covariance kernels by $C_{\boldsymbol{\theta}}(t_i - t_j) = C_{\boldsymbol{\theta}}(r(t_i, t_j))$.
- Some non-stationary covariance functions are used in some applications.
- For example, dot product covariance function

$$C_{\boldsymbol{\theta}}(t_i, t_j) = \sigma^2 + t_i . t_j$$

- Neural network covariance function is used sometimes.

$$C_{\boldsymbol{\theta}}(t_i, t_j) = \frac{2}{\pi} sin^{-1} \left(\frac{2\tilde{t}_i' \Sigma \tilde{t}_j}{\sqrt{(1 + 2\tilde{t}_i' \Sigma \tilde{t}_i)(1 + 2\tilde{t}_j' \Sigma \tilde{t}_j)}} \right),$$

where $\tilde{t}_i = (1, t_i)'$.

Model and Gaussian process prior

- The non-linear relationship between y and $\boldsymbol{x}$ is given by

$$y = f(\boldsymbol{x}) + \epsilon, \;\; \epsilon \sim N(0, \tau^2)$$

- Need prior distributions on $\{f(\cdot), \tau^2\}$.
- $f(\cdot)$ is assigned a Gaussian process prior distribution $GP(\mu, C_{\boldsymbol{\theta}}(\cdot, \cdot))$
- Lets demonstrate everything with the exponential covariance function. Let $\boldsymbol{\theta} = (\sigma^2, \phi)$,

$$\boxed{C_{\sigma^2,\phi}(t_i, t_j) = \sigma^2 \exp(-||t_i - t_j||/\phi)}.$$

- Note that this is going to create sample paths which are only continuous and not even once differentiable while the function $f(cdot)$ may be more smooth.

Prior

- One may use Matern covariance kernel with a prior on the smoothness parameter ν.
- It was not possible to learn ν.
- Thus it is a common practice to assign the prior distribution with fixing ν.
- For exponential kernel we are fixing $\nu = 2$.
- Prior on σ^2 and τ^2 are assigned inverse-gamma(a,b) prior.
- A Normal prior on μ.
- We need to be careful in assigning prior on ϕ.

Model Fitting

- Suppose we have the data $(y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n)$.
- Thus $y_i = f(\boldsymbol{x}_i) + \epsilon_i$, $\epsilon_i \sim N(0, \tau^2)$.
- Suppose $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)'$ and $\boldsymbol{f} = (f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n))'$.
- Thus $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2 \boldsymbol{I})$.
- From the Gaussian process specification, a priori $\boldsymbol{f} \sim N(\mu \mathbf{1}_n, \boldsymbol{C}_{\sigma^2,\phi})$.
- The posterior distribution

$$p(\phi, \sigma^2, \mu, \tau^2 | \boldsymbol{y}) \propto N(\boldsymbol{y} | \mu \mathbf{1}_n, \boldsymbol{C}_{\sigma^2,\phi} + \tau^2 \boldsymbol{I}) \times N(\mu | \mu_\mu, \sigma_\mu^2) \times IG(\tau^2 | a, b) \times IG(\sigma^2 | a, b) \times p(\phi).$$

Model Fitting

Model fitting proceeds through MCMC steps. We run Gibbs within Metropolis.

- $\mu|-$ follows a normal distribution.
- σ^2, τ^2, ϕ are updated using Metropolis steps.
- In spatial statistics, one uses the classical geostatistical Gaussian process model

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\beta + f(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),\ \ \epsilon(\boldsymbol{s}) \sim N(0, \tau^2).$$

- $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))'$, $\boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1) : \cdots : \boldsymbol{x}(\boldsymbol{s}_1)]'$, $\boldsymbol{f} = (f(\boldsymbol{s}_1), ..., f(\boldsymbol{s}_n))'$.
- Let $\boldsymbol{C}_{\sigma^2,\phi} = ((C_{\sigma^2,\phi}(\boldsymbol{s}_i, \boldsymbol{s}_j)))_{i,j=1}^n$

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{f} + \epsilon,\ \ \epsilon \sim N(\boldsymbol{0}, \tau^2\boldsymbol{I}).$$

Posterior distribution of the function

- Note that

$$p(\boldsymbol{f}|\boldsymbol{y}) \propto N(\boldsymbol{y}|\boldsymbol{f}, \tau^2\boldsymbol{I}) \times N(\boldsymbol{f}|\mu\mathbf{1}_n, \boldsymbol{C}_{\sigma^2,\phi})$$

- Thus $\boldsymbol{f}|\boldsymbol{y}, -$ is a multivariate normal distribution.
- For $\{\mu^{(l)}, \sigma^{2(l)}, \tau^{2(l)}, \phi^{(l)}\}_{l=1}^{L}$ L post burn-in MCMC samples, we draw $\boldsymbol{f}^{(1)}, ..., \boldsymbol{f}^{(L)}$ L MCMC samples of the posterior realization of the function at n data points.
- For inference at an arbitrary point $\boldsymbol{x}$, we calculate the distribution of $f(\boldsymbol{x})|\boldsymbol{y}$.

$$p(f(\boldsymbol{x})|\boldsymbol{y}) = \int [p(f(\boldsymbol{x})|f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n), \sigma^2, \tau^2, \phi, \mu)$$
$$p(f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)|\sigma^2, \tau^2, \phi, \mu, \boldsymbol{y})p(\sigma^2, \tau^2, \phi, \mu|\boldsymbol{y})].$$

- We already know how to draw samples from $p(f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n)|\sigma^2, \tau^2, \phi, \mu, \boldsymbol{y})$ and $p(\sigma^2, \tau^2, \phi, \mu|\boldsymbol{y})$.

Posterior distribution of the function

- $f(\boldsymbol{x})|f(\boldsymbol{x}_1), ..., f(\boldsymbol{x}_n), \sigma^2, \tau^2, \phi, \mu \sim N(\mu_f, \sigma_f^2)$.

$$\sigma_f^2 = C_{\sigma^2,\phi}(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{c}_{\sigma^2,\phi}(\boldsymbol{x})' \boldsymbol{C}_{\sigma^2,\phi}^{-1} \boldsymbol{c}_{\sigma^2,\phi}(\boldsymbol{x}).$$
$$\mu_f = \mu + \boldsymbol{c}_{\sigma^2,\phi}(\boldsymbol{x})' \boldsymbol{C}_{\sigma^2,\phi}^{-1} (\boldsymbol{f} - \mu \boldsymbol{1}_n)$$
$$\boldsymbol{c}_{\sigma^2,\phi}(\boldsymbol{x}) = (C_{\sigma^2,\phi}(\boldsymbol{x}, \boldsymbol{x}_1), ..., C_{\sigma^2,\phi}(\boldsymbol{x}, \boldsymbol{x}_n))'.$$

- For spatial process models, finding posterior distribution of the function is also similar.

Prediction

- Suppose the prediction of response is required at $\boldsymbol{x}$.
- Note that, $y \sim N(f(\boldsymbol{x}), \tau^2)$.
- We have already seen how to draw post burn-in samples $f(\boldsymbol{x})^{(1)}, ..., f(\boldsymbol{x})^{(L)}$ from $f(\boldsymbol{x})|y_1, ..., y_n$.
- Posterior predictive samples $y^{(1)}, ..., y^{(L)}$ are drawn from $y^{(l)} \sim N(f(\boldsymbol{x})^{(l)}, \tau^{2(l)})$.
- In sample prediction can be similarly performed.

Multivariate Gaussian Process model

- Multivariate Gaussian process models are most often used in the geostatistical analysis.
- Let $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), ..., y_m(\boldsymbol{s}))'$, $\boldsymbol{w}(\boldsymbol{s}) = (w_1(\boldsymbol{s}), ..., w_m(\boldsymbol{s}))'$, $\epsilon(\boldsymbol{s}) = (\epsilon_1(\boldsymbol{s}), ..., \epsilon_m(\boldsymbol{s}))'$.
- The multivariate model is given by

$$\boldsymbol{y}(\boldsymbol{s}) = \boldsymbol{B}\boldsymbol{x}(\boldsymbol{s}) + \boldsymbol{w}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),$$

$\boldsymbol{B} = ((\beta_{ij}))_{i,j=1}^{m,p}$, $\epsilon(\boldsymbol{s}) \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$.

- When data observed at $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$,

$$\boldsymbol{y} = vec(\boldsymbol{B}\boldsymbol{X}) + \boldsymbol{w} + \epsilon,$$

$\boldsymbol{w} = (\boldsymbol{w}(\boldsymbol{s}_1)', ..., \boldsymbol{w}(\boldsymbol{s}_n)')'$, $\epsilon = (\epsilon(\boldsymbol{s}_1)', ..., \epsilon(\boldsymbol{s}_n)')'$,
$\boldsymbol{X} = [\boldsymbol{x}(\boldsymbol{s}_1) : \cdots : \boldsymbol{x}(\boldsymbol{s}_n)]$.

- $\epsilon \sim N(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Psi})$.

Modeling Multivariate Gaussian Process

- **Linear Model Coregionalization:** $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{s})$, where $\boldsymbol{v}(\boldsymbol{s}) = (v_1(\boldsymbol{s}), ..., v_m(\boldsymbol{s}))'$, $\boldsymbol{A}$ is an $m \times m$ matrix.
- Is $\boldsymbol{A}$ identifiable?
- Popular specification is $\boldsymbol{A}$ is a lower triangular matrix with diagonal entries all positive.
- $v_1(\boldsymbol{s}), ..., v_m(\boldsymbol{s})$ are assigned independent Gaussian process.
- **Multivariate Matern Kernel:** Specification of Matern kernel for the multivariate Gaussian process so that marginally each component follows a Gaussian process with a univariate Matern kernel.