Gaussian Spatial Processes, Kernels, and Predictors

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

Bayes modeling

This is application of Bayesian thinking to SEL prediction, based the use of a Gaussian process as a "prior distribution" for (the function of $\mathbf{x}$) $\mathrm{E}[y|\mathbf{x}]$. Suppose

$$y = \eta(\mathbf{x}) + \epsilon$$

where

$$\eta(\mathbf{x}) = \mu(\mathbf{x}) + \gamma(\mathbf{x})$$

$\mathrm{E}\epsilon = 0$, $\mathrm{Var}\epsilon = \sigma^2$, the function $\mu(\mathbf{x})$ is known (it could be identically 0) and plays the role of a prior mean for the function (of $\mathbf{x}$)

$$\eta(\mathbf{x}) = \mathrm{E}[y|\mathbf{x}]$$

and (independent of errors ϵ), $\gamma(\mathbf{x})$ is a realization of a mean 0 stationary Gaussian process on $\Re^p$ describing the prior uncertainty for $\eta(\mathbf{x})$.

Gaussian processes and correlation functions

More completely, $\gamma(\mathbf{x})$ has $\mathrm{E}\gamma(\mathbf{x}) = 0$ and $\mathrm{Var}\gamma(\mathbf{x}) = \tau^2$ for all $\mathbf{x}$, and for some appropriate (correlation) function ρ, $\mathrm{Cov}(\gamma(\mathbf{x}), \gamma(\mathbf{z})) = \tau^2 \rho(\mathbf{x} - \mathbf{z})$ for all $\mathbf{x}$ and $\mathbf{z}$ ($\rho(\mathbf{0}) = 1$ and the function of two variables $\rho(\mathbf{x} - \mathbf{z})$ must be positive definite). The Gaussian assumption is that for any finite set of elements $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M$ of $\Re^p$, the vector of values $\gamma(\mathbf{z}_i)$ is multivariate normal.

The simplest standard forms for ρ are product forms, i.e. if ρ_j is a valid 1-D correlation function, then

$$\rho(\mathbf{x} - \mathbf{z}) = \prod_{j=1}^{p} \rho_j(x_j - z_j)$$

is a valid correlation function for a process on $\Re^p$. Standard 1-D correlation functions are $\rho(\Delta) = \exp(-c\Delta^2)$ and $\rho(\Delta) = \exp(-c|\Delta|)$. The first produces "smoother" realizations than the second, and in both cases, the constant c governs how fast realizations vary.

MVN distribution

Then the joint distribution (conditional on the $\mathbf{x}_i$ and assuming that for the training values y_i the ϵ_i are iid independent of the $\gamma(\mathbf{x}_i)$) of the training output values and a value of $\mu(\mathbf{x})$ can be identified and used to find a conditional mean for $\eta(\mathbf{x})$ given the training data. Let

$$\underset{N\times N}{\boldsymbol{\Sigma}} = \left(\tau^2\rho\left(\mathbf{x}_i - \mathbf{x}_j\right)\right)_{\substack{i=1,2,\ldots,N\\ j=1,2,\ldots,N}} \quad \text{and} \quad \underset{N\times 1}{\boldsymbol{\Sigma}\left(\mathbf{x}\right)} = \begin{pmatrix} \tau^2\rho\left(\mathbf{x}-\mathbf{x}_1\right) \\ \vdots \\ \tau^2\rho\left(\mathbf{x}-\mathbf{x}_N\right) \end{pmatrix}$$

For a single value of $\mathbf{x}$,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \\ \eta\left(\mathbf{x}\right) \end{pmatrix} \sim \mathrm{MVN}_{N+1}\left(\begin{pmatrix} \mu\left(\mathbf{x}_1\right) \\ \vdots \\ \mu\left(\mathbf{x}_N\right) \\ \mu\left(\mathbf{x}\right) \end{pmatrix}, \left(\begin{array}{c|c} \left(\boldsymbol{\Sigma}+\sigma^2\mathbf{I}\right) & \boldsymbol{\Sigma}\left(\mathbf{x}\right) \\ \hline \boldsymbol{\Sigma}\left(\mathbf{x}\right)' & \tau^2 \end{array} \right) \right)$$

Conditional mean and a predictor

The conditional mean of $\eta(\mathbf{x})$ given $\mathbf{Y}$ is then

$$\hat{f}(\mathbf{x}) = \mu(\mathbf{x}) + \boldsymbol{\Sigma}(\mathbf{x})' \left(\boldsymbol{\Sigma} + \sigma^2 \mathbf{I}\right)^{-1} \begin{pmatrix} y_1 - \mu(\mathbf{x}_1) \\ \vdots \\ y_N - \mu(\mathbf{x}_N) \end{pmatrix} \qquad (1)$$

Write

$$\underset{N \times 1}{\mathbf{w}} = \left(\boldsymbol{\Sigma} + \sigma^2 \mathbf{I}\right)^{-1} \begin{pmatrix} y_1 - \mu(\mathbf{x}_1) \\ \vdots \\ y_N - \mu(\mathbf{x}_N) \end{pmatrix} \qquad (2)$$

and then (1) implies that

$$\hat{f}(\mathbf{x}) = \mu(\mathbf{x}) + \sum_{i=1}^{N} w_i \tau^2 \rho(\mathbf{x} - \mathbf{x}_i) \qquad (3)$$

Posterior mean and RKHSs

So this development ultimately produces $\mu(\mathbf{x})$ plus a linear combination of the "basis functions" $\tau^2\rho(\mathbf{x}-\mathbf{x}_i)$ as a predictor. Remembering that $\tau^2\rho(\mathbf{x}-\mathbf{z})$ must be positive definite and seeing the ultimate form of the predictor, we are reminded of the RKHS material.

In fact, consider the case where $\mu(\mathbf{x}) \equiv 0$. (If one has some non-zero prior mean for $\eta(\mathbf{x})$, arguably that mean function should be subtracted from the raw training outputs before beginning the development of a predictor. At a minimum, output values should probably be centered before attempting development of a predictor.) Compare (2) and (3) to $\mathbf{b}_\lambda = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{Y}$ and $\hat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^{N} b_{\lambda i}\mathcal{K}(\mathbf{x},\mathbf{x}_i)$ for the $\mu(\mathbf{x}) = 0$ case. So the present "Bayes" Gaussian process development of a predictor under squared error loss based on a covariance function $\tau^2\rho(\mathbf{x}-\mathbf{z})$ and error variance σ^2 is equivalent to a RKHS regularized fit of a function to training data based on a kernel $\mathcal{K}(\mathbf{x},\mathbf{z}) = \tau^2\rho(\mathbf{x}-\mathbf{z})$ and penalty weight $\lambda = \sigma^2$.