DETAILS ON GAUSSIAN PROCESS REGRESSION (GPR) AND SEMI-GPR MODELING

Mohammad Shekaramiz, Todd K. Moon, and Jacob H. Gunther

Electrical and Computer Engineering Department, Utah State University

***Abstract*— This report tends to provide details on how to perform predictions using Gaussian process regression (GPR) modeling. In this case, we represent proofs for prediction using non-parametric GPR modeling for noise-free predictions as well as prediction using semi-parametric GPR for noisy observations.**

1. GAUSSIAN PROCESS REGRESSION

Gaussian processes (GPs) are widely used for modeling a phenomenon based on the observed spatiotemporal data [12]. A GP can be used as a tool for either classification or regression [1, 2, 12]. GPs have been used for decades as a supervised learning tool for regression problems known as Gaussian process regression (GPR) models [1, 2], and are also referred to as kriging, named after the mining engineer D.G. Krige in the geostatistics literature [3–5]. GPR models and kriging methods are applicable to a variety of problems such as the prediction and estimation of temperature, precipitation, missing pixel and un-mixing of pixels in hyperspectral imaging (HSI), human head pose estimation, concentration of carbon dioxide in the atmosphere, etc. [2, 6–11].

In GP modeling, it is assumed that the phenomenon of interest (PoI) can be evaluated via an unknown and probably nonlinear function, which we denote by $f(\cdot)$. The arguments of the function comprise a variable set $\mathbf{u}$ referred to as the input data. For example, $\mathbf{u}$ can be defined as $\mathbf{u} = [u_x, u_y, u_z, t]^T$, where (u_x, u_y, u_z) and t denote the spatial and temporal information about the measurements, respectively. Unlike parametric models such as linear regression, GP is non-parametric. In GP one defines a probability distribution function as a prior over the unknown function $f(\cdot)$, directly. In other words, GP defines a distribution over functions in the function space and the inference is performed directly in this space [2]. This is more general than a parametric model such as Bayesian linear regression, where the prior distribution is defined over the space of parameters. The GP model treats any observation as an outcome of a Gaussian random variable, and all of these random variables are jointly Gaussian. With this setting, any well-defined GP model only needs a mean accompanied with a positive definite covariance function. Under this assumption, GP provides a posterior distribution over the unknown function f once data are observed. Therefore, for any set of N observations with the input data set $\{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$, GP assumes that the distribution $p(f(\mathbf{u}_1), \ldots, f(\mathbf{u}_N))$ is jointly Gaussian with some mean $\boldsymbol{\mu}(U)$ and a covariance matrix $K(U)$, where $U := [\mathbf{u}_1, ..., \mathbf{u}_N]$. The entry in row i and column j of $K(U)$ is denoted by $[K(U)]_{ij} = \kappa(\mathbf{u}_i, \mathbf{u}_j)$, where $\kappa(.,.)$ is a positive definite kernel function. The kernel function specifies

This work is supported by the Grant NASA NNX13 AD 39 A.

the covariance between the pairs of random variables at the corresponding data points. The GP model is defined as follows [13]:

$$\mathbf{f}(U) \sim \mathcal{GP}(\boldsymbol{\mu}(U), K(U)), \tag{1}$$

where $\mathbf{f}(U) := [f(\mathbf{u}_1), \ldots, f(\mathbf{u}_N)]^T$ and $\boldsymbol{\mu}(U) := [\mu(\mathbf{u}_1), \ldots, \mu(\mathbf{u}_N)]^T$.

For the regression purposes, GPR predicts the behavior of the PoI at the unseen data points using the available training data set. The GPR model can also handle noisy observations. Suppose we have access to a set of N noisy observations $\mathbf{y} = \mathbf{f}(U) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 I_N)$ and $y_n = f(\mathbf{u}_n) + \epsilon_n, \forall n = 1, \ldots, N$. The pair $(\mathbf{u}_n, y_n)$ is the nth training data. Using the GPR modeling, the goal then becomes predicting the underlying function f evaluated at some other input data set $U_\star$, i.e., inferring $\mathbf{f}(U_\star)$, where $U_\star := [\mathbf{u}_{\star,1}, \ldots, \mathbf{u}_{\star,M}]^T$. The set $U_\star$ serves as the input test data set. Based on GP modeling, the prior joint distribution between the training and test data can be expressed as [13]

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_\star \end{bmatrix} = \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_{f_\star} \end{bmatrix}, \begin{bmatrix} K(U,U) & K(U_\star, U) \\ K(U, U_\star) & K(U_\star, U_\star) \end{bmatrix} \Big), \tag{2}$$

where $\mathbf{f}_\star$ denotes $\mathbf{f}(U_\star)$ and $[K(U, U_\star)]_{nm} = \kappa(\mathbf{u}_n, \mathbf{u}_{\star,m})$. The predictive distribution over the test data, using the existing rules for conditioning Gaussian distributions, is expressed as follows

$$\mathbf{f}_\star | U, \mathbf{y}, U_\star \sim \mathcal{N}(\boldsymbol{\mu}_{f_\star}, \Sigma_{f_\star}), \tag{3}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{f_\star} &= \boldsymbol{\mu}_y + K(U_\star, U)\big(K(U,U) + \sigma_n^2 I\big)^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) \\ \Sigma_{f_\star} &= K(U_\star, U_\star) - K(U_\star, U)\big(K(U,U) + \sigma_n^2 I\big)^{-1} K(U, U_\star). \end{aligned} \tag{4}$$

Therefore, the point estimate for $\mathbf{f}(U_\star)$ is the mean $\boldsymbol{\mu}_{f_\star}$ and the amount of uncertainty in the estimations is represented by the covariance $\Sigma_{f_\star}$ in (4). Design of the covariance function requires incorporating some prior knowledge about the behavior of the PoI as it determines the amount of correlation between any pair of data points [1]. Some of the most widely used covariance functions are squared exponential kernel ($\kappa_{SE}(u, u') = \exp\{-\frac{(u-u')^2}{2l^2}\}$) and rational quadratic ($\kappa_{RQ}(u, u') = (1 + \frac{(u-u')^2}{2\alpha l^2})^{-\alpha}$), where l and α are hyperparameters [2]. These kernels fall in the category of stationary covariance functions. Once the structure of the covariance function is selected, the corresponding hyperparameters in the model can be chosen either empirically or using some quantified statistical methods. In the empirical approach, the selection of hyperparameters is usually achieved using the empirical features obtained from the observed data such as the smoothness or periodic behavior of the samples.

Remark 1: Although GPs are powerful tools for regression and classification problems, they suffer from high computational complexity as the sample size of the training data set increases. This problem occurs because the estimation of the test data involves inverting the covariance matrix of the training data which grows as more data are collected. Regarding the complexity of GPs, there exist some approaches such as the one for truncated covariance matrices in GPs [14], online sparse matrix GPs (OSMGP) algorithm [8], sparse greedy GP (SGGP) approximation method [15], and reduced rank GP (RRGP) [16]. Furthermore, there exist some studies on estimating the covariance matrix instead of an experimentally designed kernel function. For instance, Xu and Choi provided an approach to estimate and improve the quality of covariance function for anisotropic spatio-temporal GP using mobile sensor networks [17]. The suggested sampling method for such problem is based on minimizing the information-theoretic cost function of the Fisher information [17].

2. PREDICTION USING NON-PARAMETRIC GPR FOR NOISE-FREE OBSERVATIONS

Suppose we observe a training data set $D = \big\{(\mathbf{x}_i, f_i), i = 1, ..., N\big\}$ and $f_i = f(\mathbf{x}_i)$, where $\mathbf{x}_i$ and f_i denote the $i$the set of inputs and the corresponding output, respectively. Given a test set X_* of size $N_* \times D$, the goal is to predict the set of outputs collected into the vector $\mathbf{f}_*$. By definition of the GP, the joint distribution has the following form

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} = \mathcal{N}\Bigg(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}\Bigg),$$

where $K = \kappa(X, X)$ is $N \times N$, $K_* = \kappa(X, X_*)$ is $N \times N_*$, $K_{**} = \kappa(X_*, X_*)$ is $N_* \times N_*$, and $\kappa(.,.)$ is a kernel function. Then, the posterior distribution over $\mathbf{f}_*$ becomes [2, 13]

$$p(\mathbf{f}_*|X_*, X, \mathbf{f}) = \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_{f_*}, \Sigma_{f_*}), \tag{5}$$

where

$$\begin{cases} \boldsymbol{\mu}_{f_*} = \boldsymbol{\mu}_* + K_*^T K^{-1}(\mathbf{f} - \boldsymbol{\mu}(X)),\ \boldsymbol{\mu}_* := \boldsymbol{\mu}(X_*) \\ \Sigma_{f_*} = K_{**} - K_*^T K^{-1} K_*. \end{cases}$$

Below, we provide details on how to derive (5) borrowed from [18].

Remark 2: As a brief review, the inversion of a matrix in block form can be represented as

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}, \tag{6}$$

where

$$\begin{cases} E = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \\ F = -A^{-1}B(D - CA^{-1}B)^{-1} \\ G = -(D - CA^{-1}B)^{-1}CA^{-1} \\ H = (D - CA^{-1}B)^{-1}. \end{cases}$$

Using Remark 2, the logarithm of the posterior distribution over $\mathbf{f}_*$ in (5) is proportional to

$$\log\{p(\mathbf{f}_*|X_*, X, \mathbf{y})\} \propto -\Bigg(\begin{bmatrix} (\mathbf{f} - \boldsymbol{\mu})^T & (\mathbf{f}_* - \boldsymbol{\mu}_*)^T \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} \mathbf{f} - \boldsymbol{\mu} \\ \mathbf{f}_* - \boldsymbol{\mu}_* \end{bmatrix}\Bigg),$$

where

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}^{-1}.$$

Therefore

$$\begin{aligned} \log\{p(\mathbf{f}_*|X_*, X, \mathbf{y})\} &\propto -\Big((\mathbf{f}_* - \boldsymbol{\mu}_*)^T H(\mathbf{f}_* - \boldsymbol{\mu}_*) + (\mathbf{f}_* - \boldsymbol{\mu}_*)^T\big(G(\mathbf{f} - \boldsymbol{\mu}) + F^T(\mathbf{f} - \boldsymbol{\mu})\big)\Big) \\ &\propto -\Big((\mathbf{f}_* - \boldsymbol{\mu}_*)^T H(\mathbf{f}_* - \boldsymbol{\mu}_*) + (\mathbf{f}_* - \boldsymbol{\mu}_*)^T\big((F^T + G)(\mathbf{f} - \boldsymbol{\mu})\big)\Big) \\ &\propto -\Big(\mathbf{f}_*^T H \mathbf{f}_* + \mathbf{f}_*^T\big(-2H\boldsymbol{\mu}_* + (F^T + G)(\mathbf{f} - \boldsymbol{\mu})\big)\Big) \\ &\propto -\Big(\Big(\mathbf{f}_* - H^{-1}\big(H\boldsymbol{\mu}_* - \frac{1}{2}(F^T + G)(\mathbf{f} - \boldsymbol{\mu})\big)\Big)^T H\Big(\star\Big)\Big), \end{aligned} \tag{7}$$

where the term $\star$ in $(A)^T B(\star)$ denotes A. According to (7), the covariance Σ_{f_*} becomes

$$\Sigma_{f_*} = H^{-1} = K_{**} - K_*^T K^{-1} K_*$$

and the mean $\boldsymbol{\mu}_{f_*}$ can be found from

$$\begin{aligned}
\boldsymbol{\mu}_{f_*} &= \boldsymbol{\mu}_* - \frac{1}{2} H^{-1}(-H^T K_*^T K^{-T} - H K_*^T K^{-1})(\mathbf{f} - \boldsymbol{\mu}) \\
&= \boldsymbol{\mu}_* + \frac{1}{2}(H^{-1} H^T K_*^T K^{-T} + K_*^T K^{-1})(\mathbf{f} - \boldsymbol{\mu}) \\
&= \boldsymbol{\mu}_* + \frac{1}{2}(K_*^T K^{-T} + K_*^T K^{-1})(\mathbf{f} - \boldsymbol{\mu}).
\end{aligned}$$

After some simplification, the mean $\boldsymbol{\mu}_{f_*}$ can be represented as

$$\boldsymbol{\mu}_{f_*} = \boldsymbol{\mu}_* + K_*^T K^{-1}(\mathbf{f} - \boldsymbol{\mu}).$$

Therefore, in summary we have

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{f_*}, \Sigma_{f_*}),$$

where

$$\begin{cases} \boldsymbol{\mu}_{f_*} = \boldsymbol{\mu}_* + K_*^T K^{-1}(\mathbf{f} - \boldsymbol{\mu}) \\ \Sigma_{f_*} = K_{**} - K_*^T K^{-1} K_*. \end{cases} \tag{8}$$

3. PREDICTION USING SEMI-PARAMETRIC GPR MODELING FOR NOISY OBSERVATIONS

Consider the following model

$$f(\mathbf{x}) = \boldsymbol{\beta}^T \Phi(\mathbf{x}) + r(\mathbf{x}), \tag{9}$$

where the linear model $\boldsymbol{\beta}^T \Phi(\mathbf{x})$ is used for the mean and the residual $r(\mathbf{x})$ of the process is defined by a GP modeling, which is defined as

$$r(\mathbf{x}) \sim \mathcal{GP}\big(\mathbf{0}, \kappa(\mathbf{x}, \mathbf{x}')\big). \tag{10}$$

In other words, semi-parametric modeling combines the parametric model of the mean and the non-parametric model for the residual of the process. In this case, one can define the following prior for the parameter $\boldsymbol{\beta}$ in (9)

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, B). \tag{11}$$

Then, the posterior distribution over $\mathbf{f}_*$ for GPR with semi-parametric model becomes [18]

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}) = \mathcal{N}\big(\bar{\mathbf{f}}_*, \mathrm{Cov}(\mathbf{f}_*)\big),$$

where

$$\begin{cases} \bar{\mathbf{f}}_* = \Phi_*^T \bar{\boldsymbol{\beta}} + K_*^T K_y^{-1}(\mathbf{y} - \Phi^T \bar{\boldsymbol{\beta}}) \\ \bar{\boldsymbol{\beta}} = (\Phi K_y^{-1} \Phi^T + B^{-1})^{-1}(\Phi K_y^{-1} \mathbf{y} + B^{-1} \mathbf{b}) \\ \mathrm{Cov}(\mathbf{f}_*) = K_{**} - K_*^T K_y^{-1} K_* + R^T (B^{-1} + \Phi K_y^{-1} \Phi^T)^{-1} R \\ R = \Phi_* - \Phi K_y^{-1} K_*. \end{cases} \tag{12}$$

Below the details on how to derive (12) is represented. The set of priors considered here are

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, B), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_N^2 I),$$

where the noise is denoted by $\boldsymbol{\epsilon}$ and is modeled by a zero-mean Gaussian distribution. Therefore, the posterior distribution over $\boldsymbol{\beta}$ is proportional to

$$p(\boldsymbol{\beta}|X, \mathbf{y}) \propto p(\mathbf{y}|X, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{b}, B),$$

and by taking the logarithm from the above equation, we have

$$\begin{aligned} \log\{p(\boldsymbol{\beta}|X, \mathbf{y})\} &\propto -\big((\mathbf{y} - \Phi^T\boldsymbol{\beta})^T K_y^{-1}(\star) + (\boldsymbol{\beta} - \mathbf{b})^T B^{-1}(\star)\big) \\ &\propto -\big(\boldsymbol{\beta}^T(\Phi K_y^{-1}\Phi^T + B^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\Phi K_y^{-1}\mathbf{y} + B^{-1}\mathbf{b})\big). \end{aligned}$$

Therefore, the posterior distribution over $\boldsymbol{\beta}$ becomes

$$p(\boldsymbol{\beta}|X, \mathbf{y}) = \mathcal{N}(\bar{\boldsymbol{\beta}}, \hat{\beta}),$$

where

$$\begin{cases} \bar{\boldsymbol{\beta}} = (\Phi K_y^{-1}\Phi^T + B^{-1})^{-1}(\Phi K_y^{-1}\mathbf{y} + B^{-1}\mathbf{b}) = \hat{\beta}(\Phi K_y^{-1}\mathbf{y} + B^{-1}\mathbf{b}) \\ \hat{\beta} = (\Phi K_y^{-1}\Phi^T + B^{-1})^{-1} \end{cases} \tag{13}$$

yielding to

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}) = \mathcal{N}(\mathbf{f}_*; \boldsymbol{\mu}_*, \Sigma_*),$$

where

$$\begin{cases} \boldsymbol{\mu}_* = \boldsymbol{\mu}(X_*) + K_*^T K_y^{-1}\big(\mathbf{y} - \boldsymbol{\mu}(X)\big) \\ \Sigma_* = K_{**} - K_*^T K_y^{-1} K_*. \end{cases}$$

Therefore,

$$\begin{cases} \boldsymbol{\mu}_* = \Phi_*^T\boldsymbol{\beta} + K_*^T K_y^{-1}(\mathbf{y} - \Phi^T\boldsymbol{\beta}) \\ \Sigma_* = K_{**} - K_*^T K_y^{-1} K_*. \end{cases} \tag{14}$$

Since the above set of equations are dependent on the parameter $\boldsymbol{\beta}$, we integrate out $\boldsymbol{\beta}$ in order to have a non-parametric model, as defined below.

$$\begin{aligned} p(\mathbf{f}_*|X_*, X, \mathbf{y}) &= \int_{\boldsymbol{\beta}} p(\mathbf{f}_*|X_*, X, \mathbf{y}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|X, \mathbf{y}, X_*, \mathbf{y}_*) d\boldsymbol{\beta} \\ &= \int_{\boldsymbol{\beta}} p(\mathbf{f}_*|X_*, X, \mathbf{y}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|X, \mathbf{y}) d\boldsymbol{\beta} \\ &= \int \exp\Big\{ -\big((\mathbf{f}_* - \boldsymbol{\mu}_*)^T\Sigma_*^{-1}(\star) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\hat{\beta}^{-1}(\star)\big)\Big\} d\boldsymbol{\beta} \\ &= \int \exp\Big\{ -\Big(\big(\mathbf{f}_* - \Phi_*^T\boldsymbol{\beta} - K_*^T K_y^{-1}(\mathbf{y} - \Phi^T\boldsymbol{\beta})\big)^T\Sigma_*^{-1}\big(\star\big) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\hat{\beta}^{-1}(\star)\Big)\Big\} d\boldsymbol{\beta} \\ &= \int \exp\{-\Big(\big((\Phi_*^T - K_*^T K_y^{-1}\Phi^T)\boldsymbol{\beta} - (K_*^T K_y^{-1}\mathbf{y} - \mathbf{f}_*)\big)^T\Sigma_*^{-1}\big(\star\big) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\hat{\beta}^{-1}(\star)\Big)\} d\boldsymbol{\beta} \end{aligned}$$

Therefore,

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}) = \Big[\int e^{\Big\{-\Big(\boldsymbol{\beta}^T\big((\Phi_*-\Phi K_y^{-1}K_*)\Sigma_*^{-1}(\star)^T+\hat{\beta}^{-1}\big)\boldsymbol{\beta}-2\boldsymbol{\beta}^T\big((\Phi_*-\Phi K_y^{-1}K_*)\Sigma_*^{-1}(K_*^TK_y^{-1}\mathbf{y}-f_*)+\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\Big)\Big\}} d\boldsymbol{\beta}\Big]\times e^{\Big\{-\big((K_*^TK_y^{-1}\mathbf{y}-\mathbf{f}_*)^T\Sigma_*^{-1}(\star)+\bar{\boldsymbol{\beta}}^T\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\Big\}}. \quad (15)$$

Let's define

$$R := \Phi_* - \Phi K_y^{-1}K_*. \quad (16)$$

Substituting (16) into (15) yields

$$\begin{aligned}
&p(\mathbf{f}_*|X_*, X, \mathbf{y}) = \\
&\Big[\int e^{\Big\{-\Big(\boldsymbol{\beta}^T(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1})\boldsymbol{\beta}-2\boldsymbol{\beta}^T\big(R\Sigma_*^{-1}(K_*^TK_y^{-1}\mathbf{y}-\mathbf{f}_*)+\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\Big)\Big\}} d\boldsymbol{\beta}\Big]\times \\
&\qquad e^{\Big\{-\big((K_*^TK_y^{-1}\mathbf{y}-\mathbf{f}_*)^T\Sigma_*^{-1}(\star)+\bar{\boldsymbol{\beta}}^T\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\Big\}} \\
=&\Big[\int e^{\Big\{-\big(\boldsymbol{\beta}-(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1})^{-1}(R\Sigma_*^{-1}(K_*^TK_y^{-1}\mathbf{y}-\mathbf{f}_*)+\hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big)^T\big(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1}\big)(\star)\Big\}} d\boldsymbol{\beta}\Big]\times \\
&e^{\Big\{\big(R\Sigma_*^{-1}(K_*^TK_y^{-1}\mathbf{y}-\mathbf{f}_*)+\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\big)^T\big(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1}\big)^{-1}(\star)-\big((\mathbf{f}_*-K_*^TK_y^{-1}\mathbf{y})^T\Sigma_*^{-1}(\star)+\bar{\boldsymbol{\beta}}^T\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big)\Big\}}.
\end{aligned}$$

By taking logarithm of the above equation, we then have

$$\begin{aligned}
&\log\{p(\mathbf{f}_*|X_*, X, \mathbf{y})\} \propto \\
&-\Big(\mathbf{f}_*^T\Sigma_*^{-1}\mathbf{f}_* - 2\mathbf{f}_*^T(\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}) - \big(R\Sigma_*^{-1}\mathbf{f}_* - (R\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}-\hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big)^T\times \\
&\qquad\big(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1}\big)^{-1}(\star)\Big),
\end{aligned}$$

which is proportional to

$$\begin{aligned}
\propto -\Big(&\mathbf{f}_*^T\big(\Sigma_*^{-1}-\Sigma_*^{-1}R^T(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1})^{-1}R\Sigma_*^{-1}\big)\mathbf{f}_* \\
&-2\mathbf{f}_*^T\big(\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}-\Sigma_*^{-1}R^T(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1})^{-1}(R\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}-\hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big)\Big)
\end{aligned}$$

Thus

$$\begin{aligned}
&\log\{p(\mathbf{f}_*|X_*, X, \mathbf{y})\} \propto -\Big(\mathbf{f}_*^T\big(\Sigma_*^{-1}-\Sigma_*^{-1}R^T(\hat{\beta}^{-1}+R\Sigma_*^{-1}R^T)^{-1}R\Sigma_*^{-1}\big)\mathbf{f}_* \\
&-2\mathbf{f}_*^T\big(\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}-\Sigma_*^{-1}R^T(R\Sigma_*^{-1}R^T+\hat{\beta}^{-1})^{-1}(R\Sigma_*^{-1}K_*^TK_y^{-1}\mathbf{y}-\hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big)\Big).
\end{aligned}$$

Therefore, the covariance of the posterior distribution on $\mathbf{f}_*$ becomes

$$\begin{aligned}
\Sigma_{f_*} &= \Sigma_*^{-1} - \Sigma_*^{-1}R^T(\hat{\beta}^{-1}+R\Sigma_*^{-1}R^T)^{-1}R\Sigma_*^{-1} \\
&= \Sigma_* + R^T\hat{\beta}R,
\end{aligned} \quad (17)$$

where

$$\begin{cases} \hat{\beta} & = (\Phi K_y^{-1}\Phi^T + B^{-1})^{-1} \\ R & = \Phi_* - \Phi K_y^{-1} K_* \\ \Sigma_* & = K_{**} - K_*^T K_y^{-1} K_* \end{cases}$$

or equivalently,

$$\Sigma_{f_*} = K_{**} - K_*^T K_y^{-1} K_* + R^T (\Phi K_y^{-1}\Phi^T + B^{-1})^{-1} R.$$

Below, we simplify the expected value of $\mathbf{f}_*$.

$$\boldsymbol{\mu}_{\mathbf{f}_*} = \Sigma_{f_*}\big(\Sigma_*^{-1} K_*^T K_y^{-1}\mathbf{y} - \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1}(R\Sigma_*^{-1} K_*^T K_y^{-1}\mathbf{y} - \hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big).$$

Therefore,

$$\boldsymbol{\mu}_{\mathbf{f}_*} = (\Sigma_* + R^T\hat{\beta}R)\big(\Sigma_*^{-1} K_*^T K_y^{-1}\mathbf{y} - \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1}(R\Sigma_*^{-1} K_*^T K_y^{-1}\mathbf{y} - \hat{\beta}^{-1}\bar{\boldsymbol{\beta}})\big)$$

or equivalently

$$\boldsymbol{\mu}_{\mathbf{f}_*} = (\Sigma_* + R^T\hat{\beta}R)\big((\Sigma_*^{-1} - \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1} \times R\Sigma_*^{-1} K_*^T K_y^{-1}\mathbf{y} + \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T \hat{\beta}^{-1})\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big). \tag{18}$$

Notice that

$$(\Sigma_* + R^T\hat{\beta}R)^{-1} = (\Sigma_*^{-1} - \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1} R\Sigma_*^{-1}. \tag{19}$$

Substituting (19) into (18) yields

$$\boldsymbol{\mu}_{\mathbf{f}_*} = (\Sigma_* + R^T\hat{\beta}R)\big((\Sigma_* + R^T\hat{\beta}R)^{-1} K_*^T K_y^{-1}\mathbf{y} + \Sigma_*^{-1} R^T (R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1}\hat{\beta}^{-1}\bar{\boldsymbol{\beta}}\big).$$

Also,

$$(R\Sigma_*^{-1} R^T + \hat{\beta}^{-1})^{-1} = \hat{\beta} - \hat{\beta}R(\Sigma_* + R^T\hat{\beta}R)^{-1} R^T\hat{\beta}.$$

Thus,

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{f}_*} =& K_*^T K_y^{-1}\mathbf{y} + (\Sigma_* + R^T\hat{\beta}R)\Sigma_*^{-1} R^T (I - \hat{\beta}R(\Sigma_* + R^T\hat{\beta}R)^{-1} R^T)\bar{\boldsymbol{\beta}} \\ =& K_*^T K_y^{-1}\mathbf{y} + (\Sigma_* + R^T\hat{\beta}R)\Sigma_*^{-1} R^T\big(I - \hat{\beta}R(I + \Sigma_*^{-1} R^T\hat{\beta}R)^{-1}\Sigma_*^{-1} R^T\big)\bar{\boldsymbol{\beta}}. \end{aligned}$$

By applying matrix inversion lemma, we have

$$I - \hat{\beta}R(I + \Sigma_*^{-1} R^T\hat{\beta}R)^{-1}\Sigma_*^{-1} R^T = I - (I + \hat{\beta}R\Sigma_*^{-1} R^T)^{-1}\hat{\beta}R\Sigma_*^{-1} R^T.$$

Hence

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{f}_*} =& K_*^T K_y^{-1}\mathbf{y} + (\Sigma_* + R^T\hat{\beta}R)\Sigma_*^{-1} R^T\big(I - (I + \hat{\beta}R\Sigma_*^{-1} R^T)^{-1}\hat{\beta}R\Sigma_*^{-1} R^T\big)\bar{\boldsymbol{\beta}} \\ =& K_*^T K_y^{-1}\mathbf{y} + R^T\bar{\boldsymbol{\beta}} + \big(R^T\hat{\beta}R\Sigma_*^{-1} R^T - (\Sigma_* + R^T\hat{\beta}R)\Sigma_*^{-1} R^T (I + \hat{\beta}R\Sigma_*^{-1} R^T)^{-1}\hat{\beta}R\Sigma_*^{-1} R^T\big)\bar{\boldsymbol{\beta}} \\ =& K_*^T K_y^{-1}\mathbf{y} + \Big(R^T(I + \hat{\beta}R\Sigma_*^{-1} R^T) - (I + R^T\hat{\beta}R\Sigma_*^{-1})\big(R^T(I + \hat{\beta}R\Sigma_*^{-1} R^T)^{-1}\hat{\beta}R\Sigma_*^{-1} R^T\big)\Big)\bar{\boldsymbol{\beta}}. \end{aligned}$$

Therefore,

$$\boldsymbol{\mu}_{f_*} = K_*^T K_y^{-1}\mathbf{y} + M\bar{\boldsymbol{\beta}},$$

where

$$M := R^T(I + \hat{\beta}R\Sigma_*^{-1}R^T) - (I + R^T\hat{\beta}R\Sigma_*^{-1})\big(R^T(I + R^T\hat{\beta}R\Sigma_*^{-1})^{-1}\hat{\beta}R\Sigma_*^{-1}R^T\big).$$

Since,

$$R^T(I + \hat{\beta}R\Sigma_*^{-1}R^T) = (I + R^T\hat{\beta}R\Sigma_*^{-1})R^T,$$

we can redefine M as

$$M = R^T(I + \hat{\beta}R\Sigma_*^{-1}R^T)\big(I - (I + \hat{\beta}R\Sigma_*^{-1}R^T)^{-1}\hat{\beta}R\Sigma_*^{-1}R^T\big) \tag{20}$$

and by applying matrix inversion lemma to (20), we will have

$$M = R^T(I + \hat{\beta}R\Sigma_*^{-1}R^T)(I + \hat{\beta}R\Sigma_*^{-1}R^T)^{-1} = R^T.$$

Therefore,

$$\boldsymbol{\mu}_{f_*} = K_*^T K_y^{-1}\mathbf{y} + R^T\bar{\boldsymbol{\beta}},$$

where

$$R = \Phi_* - \Phi K_y^{-1}K_*$$

or equivalently,

$$\mu_{\mathbf{f}_*} = \Phi_*^T\bar{\boldsymbol{\beta}} + K_*^T K_y^{-1}(\mathbf{y} - \Phi^T\bar{\boldsymbol{\beta}}). \tag{21}$$

In summary, we have

$$p(\mathbf{f}_*|X_*, X, \mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{f_*}, \Sigma_{f_*}),$$

where

$$\begin{cases} \boldsymbol{\mu}_{f_*} = \Phi_*^T\bar{\boldsymbol{\beta}} + K_*^T K_y^{-1}(\mathbf{y} - \Phi^T\bar{\boldsymbol{\beta}}) \\ \Sigma_{f_*} = K_{**} - K_*^T K_y^{-1}K_* + R^T(B^{-1} + \Phi K_y^{-1}\Phi^T)^{-1}R \\ \bar{\boldsymbol{\beta}} = (\Phi K_y^{-1}\Phi^T + B^{-1})^{-1}(\Phi K_y^{-1}\mathbf{y} + B^{-1}\mathbf{b}) \\ R = \Phi_* - \Phi K_y^{-1}K_*. \end{cases}$$

4. REFERENCES

[1] J. Q. Shi and T. Choi, *Gaussian Process Regression Analysis for Functional Data*. CRC Press, 2011.

[2] C. Rasmussen and C. Williams, *Gaussian Processes in Machine Learning*. MIT Press, 2006.

[3] P. Goovaerts, *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

[4] E. H. Isaak and R. M. Srivastava, *An Introduction to Applied Statistics*. Oxford University Press, 1989.

[5] J. P. Chiles and P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty*. John Wiley and Sons, 1999.

[6] T. Imbiriba, J. C. M. Bermudez, J.-Y. Tourneret, and C. Richard, "Detection of nonlinear mixtures using Gaussian processes: Application to hyperspectral imaging," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, pp. 7949–7953, May 2014.

[7] T. Wu and Y. Li, "Spatial interpolation of temperature in the United States using residual kriging," *Applied Geography*, vol. 44, pp. 112–120, 2013.

[8] A. Ranganathan, M. H. Yang, and J. Ho, "Online sparse Gaussian process regression and its applications," *IEEE Trans. on Image Proc.*, vol. 20, pp. 391–404, Feb. 2011.

[9] Z. Xing, M. Zhou, A. Castrodad, G. Sapiro, and L. Carin, "Dictionary learning for noisy and incomplete hyperspectral images," *SIAM J. Imag. Sci.*, vol. 5, no. 1, pp. 33–56, 2012.

[10] P. Monestiez, D. Courault, D. Allard, and F. Ruget, "Spatial interpolation of air temperature using evironmental context: Application to a crop model," *Env. and Ecol. Stat.*, vol. 8, pp. 297–309, 2001.

[11] M. R. Holdaway, "Spatial modeling and interpolation of monthly temperature using kriging," *Climate Research*, vol. 6, pp. 215–225, 1996.

[12] M. Shekaramiz, T. K. Moon, and J. H. Gunther, "Exploration and data refinement via multiple mobile sensors based on Gaussian processes," in *51th Asilomar Conf. of Sig., Syst., and Compt.*, pp. 885–889, 2017.

[13] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[14] A. J. Storkey, "Truncated covariance matrices and Toeplitz methods in Gaussian processes," in *Artificial Neural Networks (ICANN)*, 1999.

[15] A. J. Smola and P. L. Bartlett, "Sparse greedy Gaussian process regression," *Advances in Neural Information Processing Systems, Cambridge, Massachussetts, MIT Press*, vol. 13, pp. 619–625, 2001.

[16] J. Quinonero-Candela and C. E. Rasmussen, "Analysis of some methods for reduced rank Gaussian process regression," in *Switching and Learning in Feedback Systems, Springer, Berlin, Heidelberg*, pp. 98–127, 2005.

[17] Y. Xu and J. Choi, "Adaptive sampling for learning Gaussian processes using mobile sensor networks," *Sensors*, vol. 11, no. 3, pp. 3051–3066, 2011.

[18] M. Shekaramiz, *Sparse Signal Recovery Based on Compressive Sensing and Exploration Using Multiple Mobile Sensors*. PhD Dissertation, Utah State University, Digitalcommons, 2018.