Vector-valued Gaussian Processes

Andrew Moore, Grady Wright (Department of Mathematics)

Keywords — Gaussian Processes, Statistics, Velocity Fields

I. Overview of Gaussian Processes

Gaussian processes generalize the multivariate Gaussian distribution and can be used to describe a probability distribution over families of functions.

i. Multivariate Gaussian Distribution

The multivariate Gaussian distribution is used to model *random vectors* (vectors of jointly distributed random variables).

$$
\begin{aligned}
\boldsymbol{x} \in \mathbb{R}^N &\sim \mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\mu} \in \mathbb{R}^N &= (\mu_1, \mu_2, ..., \mu_N)^\top = (\mathbb{E}(x_1), \mathbb{E}(x_2), ..., \mathbb{E}(x_N))^\top \\
\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N} &= \mathbb{E}\left((\boldsymbol{x}-\mu)(\boldsymbol{x}-\mu)^\top\right) = \left[\mathrm{cov}(x_i, x_j)\right]_{i,j=1}^{N} \\
x_i &\sim \mathcal{N}(\mu_i, \boldsymbol{\Sigma}_{ii})
\end{aligned} \tag{1}
$$

ii. Gaussian Processes (GPs)

- *Gaussian process* (GP): an uncountably infinite collection of random variables; any finite sample is a draw from a MV Gaussian distribution.
- GPs are fully specified by a *mean function* m and *covariance (kernel) function* k.
- The kernel function must produce a positive semi-definite matrix when evaluated on a set of input points (or vectors).

We focus on the *squared exponential kernel* $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, defined as:

$$
k(x, x') = \alpha^2 \exp\left(-\frac{1}{2\rho^2}\,\|x - x'\|^2\right). \tag{2}
$$

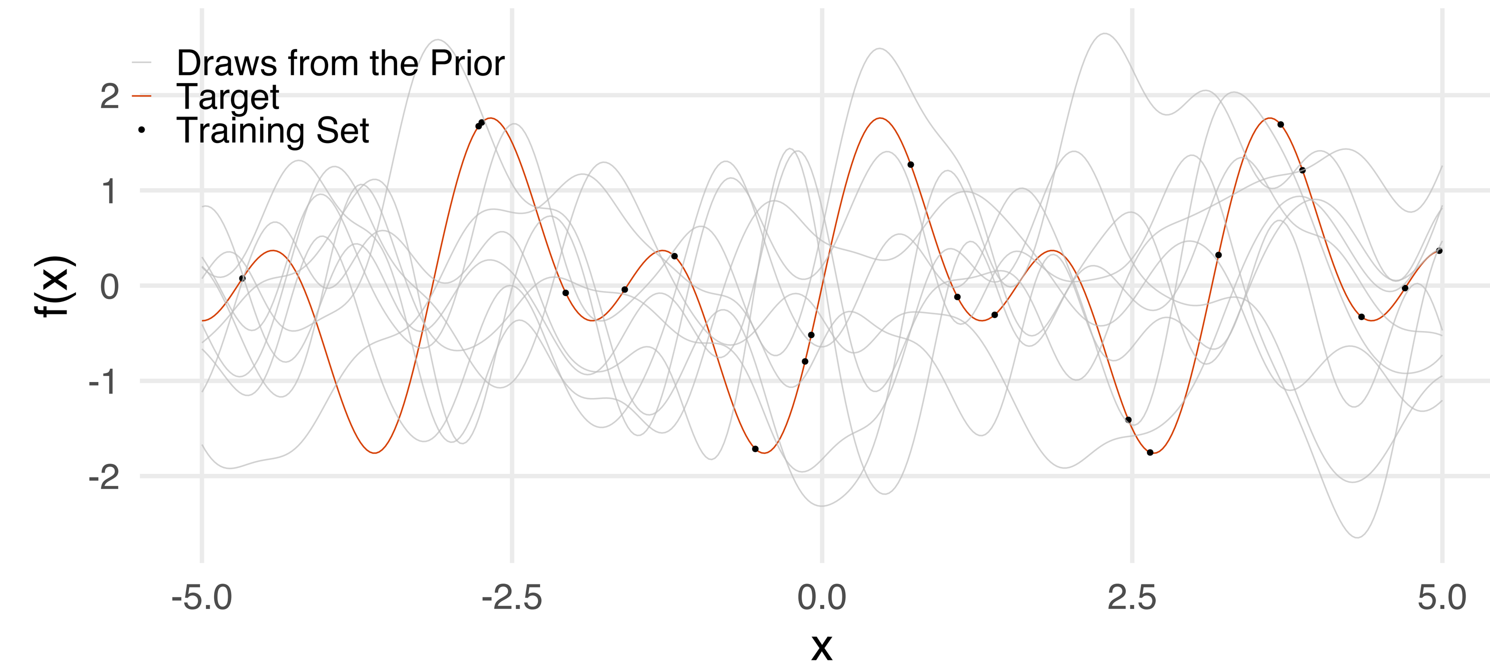
- $\|\cdot\|$ is the Euclidean Norm: $\|\boldsymbol{x}\| = \sqrt{x_1 + x_2 + \cdots + x_N}$
- α and ρ are *hyperparameters* (chosen, or estimated from data)

II. Gaussian Process Regression – Univariate $\mathbf{y}$

Let $S = (\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^d, i \in 1, 2, ..., N\}$ be a researcher's dataset, and let $N = 20$ and $M = 400 - N$. We wish to use S to find an unknown function f that satisfies $\boldsymbol{y} = f(\boldsymbol{x})$, possibly subject to additive noise ε. We can draw samples from the prior distribution:

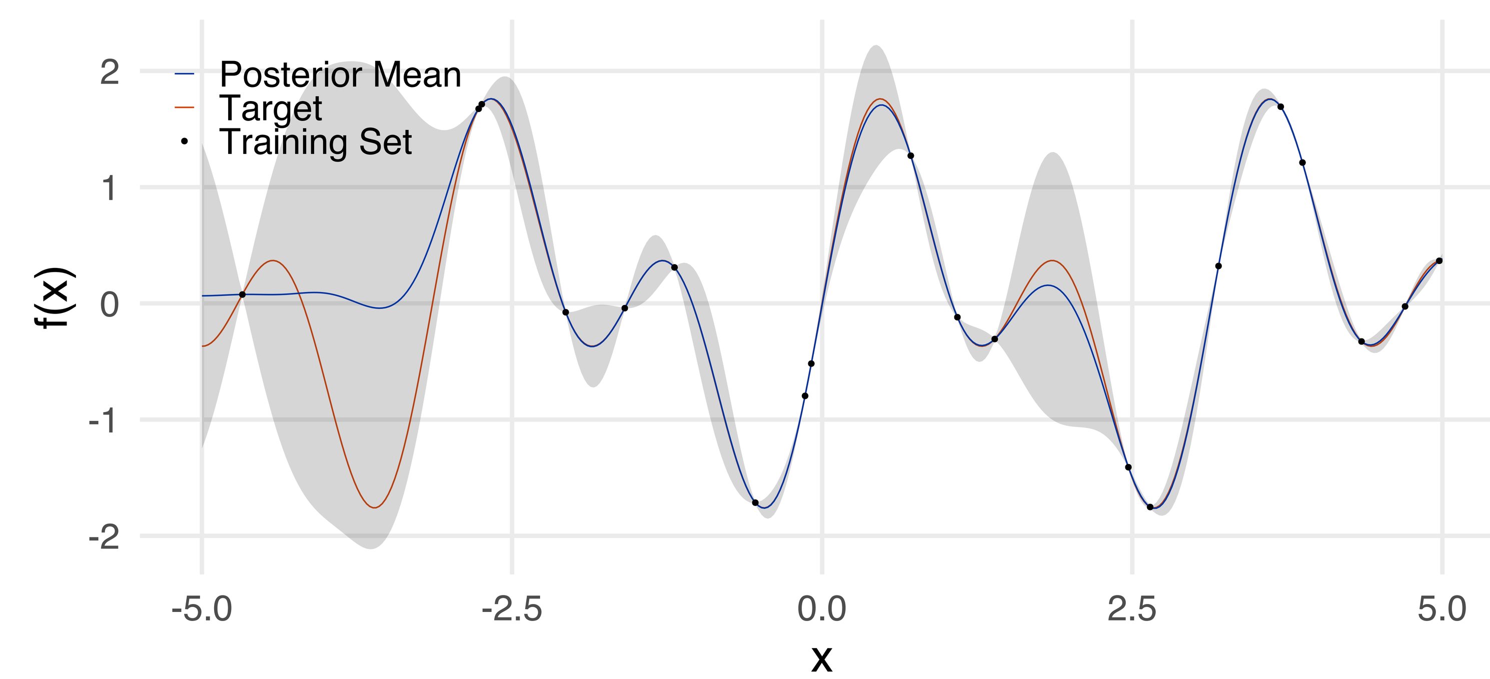
$$
\begin{aligned}
f &\sim \mathcal{GP}(\mathbf{0}, k) \\
\boldsymbol{y}_* &\sim \mathcal{N}_M(\mathbf{0}, k(\boldsymbol{x}_*, \boldsymbol{x}_*)) \\
k(\boldsymbol{x}_*, \boldsymbol{x}_*) \in \mathbb{R}^{M\times M} &= \left[k(\boldsymbol{x}_{*i}, \boldsymbol{x}_{*j})\right]_{i,j=1}^{M}.
\end{aligned} \tag{3}
$$

- Draws from the prior distribution (shown in grey) don't necessarily agree with the data points.
- Kernel choice determines properties of f (e.g., smoothness)



Our prior model for f and the observed data S can be combined to form a *posterior* distribution:

$$
\begin{aligned}
\boldsymbol{y}_* \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}_* &\sim \mathcal{N}_M\left(\hat{\mu}, \hat{\Sigma}\right) \\
\hat{\mu} \in \mathbb{R}^M &= k(\boldsymbol{x}_*, \boldsymbol{x})(k(\boldsymbol{x}, \boldsymbol{x}))^{-1}\boldsymbol{y} \\
\hat{\Sigma} \in \mathbb{R}^{M\times M} &= k(\boldsymbol{x}_*, \boldsymbol{x}_*) - k(\boldsymbol{x}_*, \boldsymbol{x})(k(\boldsymbol{x}, \boldsymbol{x}))^{-1}k(\boldsymbol{x}_*, \boldsymbol{x})^\top \\
k(\boldsymbol{x}, \boldsymbol{x}) \in \mathbb{R}^{N\times N} &= \left[k(x_i, x_j)\right]_{i,j=1}^{N} \\
k(\boldsymbol{x}_*, \boldsymbol{x}) \in \mathbb{R}^{M\times N} &= \left[k(x_{*i}, x_j)\right]_{i,j=1}^{M,N}.
\end{aligned} \tag{4}
$$



III. Multioutput GPR – Vector-valued $\mathbf{y}$

- GPR can be extended to targets with >1 dimensions.
- Velocity fields: $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{N\times 2}$
- Idea: columns of $\boldsymbol{Y}$ might not be independent
- Intrinsic Coregionalization Model (ICM): combines the kernel matrix with a similarity matrix B

$$
\begin{aligned}
&\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{N\times 2},\ \boldsymbol{y} \in \mathbb{R}^{2N} = \mathrm{vec}(Y),\ \boldsymbol{X}_* \in \mathbb{R}^{M\times 2} \\
\boldsymbol{y}_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_* &\sim \mathcal{N}_{2M}\left(\hat{\mu}, \hat{\Sigma}\right) \\
\hat{\mu} \in \mathbb{R}^{2M} &= K_{\boldsymbol{X}_*\boldsymbol{X}}K_{\boldsymbol{X}\boldsymbol{X}}^{-1}\boldsymbol{y} \\
\hat{\Sigma} \in \mathbb{R}^{2M\times 2M} &= K_{\boldsymbol{X}_*\boldsymbol{X}_*} - K_{\boldsymbol{X}_*\boldsymbol{X}}K_{\boldsymbol{X}\boldsymbol{X}}^{-1}K_{\boldsymbol{X}_*\boldsymbol{X}}^\top \\
K_{\boldsymbol{X}\boldsymbol{X}} \in \mathbb{R}^{2N\times 2N} &= B \otimes k(\boldsymbol{X}, \boldsymbol{X}) \\
K_{\boldsymbol{X}_*\boldsymbol{X}} \in \mathbb{R}^{2M\times 2N} &= B \otimes k(\boldsymbol{X}_*, \boldsymbol{X}) \\
K_{\boldsymbol{X}_*\boldsymbol{X}_*} \in \mathbb{R}^{2M\times 2M} &= B \otimes k(\boldsymbol{X}_*, \boldsymbol{X}_*) \\
B \in \mathbb{R}^{2\times 2} &= \mathrm{corr}(\boldsymbol{Y}) = \left(\frac{\langle \boldsymbol{y}_i - \overline{\boldsymbol{y}}_i, \boldsymbol{y}_j - \overline{\boldsymbol{y}}_j\rangle}{\|\boldsymbol{y}_i - \overline{\boldsymbol{y}}_i\|\,\|\boldsymbol{y}_j - \overline{\boldsymbol{y}}_j\|}\right)_{i,j=1}^{2}
\end{aligned} \tag{5}
$$