2. Gaussian Processes.

2.1 Introduction

GPs are stochastic processes identified by a mean function m and a kernel k, $f \sim GP(m, k)$.

$\forall\, x \in M$, $f(x)$ is a random variable (stoch. process).

For a GP, any set $\{f(x_1) \dots f(x_N)\}$ is distributed according to an N-dimensional Gaussian, s.t.

$$\begin{cases} E[f(x_i)] = m(x_i) = m_i \\ Cov[f(x_i), f(x_j)] = k(x_i, x_j) = K_{ij} > 0 \text{ def.} \end{cases}$$

$m: \quad M \to \mathbb{R}$

$k: \quad M \times M \to \mathbb{R}$

2.2 Bayesian Approach to Inverse Problems.

$$y_I = \int_M dx \; C_I(x)\, f(x)$$

- f is promoted to a GP.
- choose a prior for f, $p(f)$, where $f_i = f(x_i)$

  all prior knowledge is encoded in $p(f)$,

  ↳ independent of the data.
- Bayes thm. $\Rightarrow$ posterior distribution

$$\tilde{p}(f) = p(f \mid y) = \frac{p(y \mid f)\, p(f)}{p(y)}$$

$p(y \mid f)$ likelihood

Knowledge about the solution is encoded in $\tilde{p}(f)$. e.g.

central value : $E_{\tilde{p}}[f]$

covariance : $\mathrm{Cov}_{\tilde{p}}[f_i, f_j]$

2.3 Setting the problem

$$x = \{x_i, i = 1 \dots N\} \qquad \rightarrow \qquad f \in \mathbb{R}^N, \quad f_i = f(x_i)$$

$$x^* = \{x_i^*, i = 1 \dots M\} \qquad \qquad f^* \in \mathbb{R}^M, \quad f_i^* = f(x_i^*)$$

Joint prior, depends on a set of hyperparameters θ.

$$p(f, f^* | \theta) = \frac{1}{\sqrt{\det(2\pi K)}} \exp\left\{-\frac{1}{2}\left((f-m)^T, (f^*-m^*)^T\right) K^{-1} \begin{pmatrix} f-m \\ f^*-m^* \end{pmatrix}\right\}$$

where $m_i = m(x_i)$, $m_i^* = m(x_i^*)$

$$K_{ij} = k(x_i, x_j)$$

$$K = \begin{pmatrix} K_{xx} & K_{xx^*} \\ K_{x^*x} & K_{x^*x^*} \end{pmatrix}, \quad (N+M) \times (N+M) \text{ sym., } > 0 \text{ matrix.}$$

e.g. $m(x) = 0$, $\forall x \in M$

$$k(x, x') = \sigma^2 \sqrt{\frac{2\,\ell(x)\,\ell(x')}{\ell(x)^2 + \ell(x')^2}} \exp\left\{-\frac{(x-x')^2}{\ell(x)^2 + \ell(x')^2}\right\}.$$

$\hookrightarrow$ Gibbs kernel

$$\ell(x) = \ell_0 (x + \delta)$$

σ and ℓ_0 are the hyperparameters.

Other choices are possible, depending on the info that we want to encode.

In all cases, the prior is <u>explicit</u>, which is good.

2.4 Data & Theory Predictions

data central values $y = \{ y_I , I = 1 \dots N_{dat} \}$

fluctuations $\varepsilon \sim \mathcal{N}(0, C_y)$

theory prediction for the I^{th} datapoint

$$T_I = \int_M dx \; C_I(x) \, f(x) \simeq \sum_{i=1}^{N} (FK)_{Ii} f_i$$

only f_i involved in the theory prediction, not f_i^*.

f is a GP $\Rightarrow$ T_I are Gaussian variables.

$$\begin{cases} E_p[T_I] = (FK)_{Ii} E_p[f_i] = (FK)_{Ii} m_i \\ Cov_p[T_I, T_J] = (FK)_{Ii} K_{ij} (FK)^T_{jJ} \end{cases}$$

2.5 Posterior Distribution

$$\tilde{p}(f, f^*) = p(f, f^* | y)$$
$$= \int d\theta \; p(f, f^*, \theta | y) \qquad \text{marginalize wrt. } \theta .$$

$$\tilde{p}(f) = \int df^* \, \tilde{p}(f, f^*) \quad \text{and} \quad \tilde{p}(f^*) = \int df \, \tilde{p}(f, f^*)$$

$\tilde{p}(f^*)$ posterior distribution f. values of f that do <u>not</u> enter in the theory prediction.

We have:
$$p(f, f^*, \theta | y) = \underset{(a)}{p(f, f^* | \theta, y)} \; \underset{(b)}{p(\theta | y)}$$

The two factors, (a) and (b), can be computed separately.

(a) $$p(f, f^* | \theta, y) \propto \exp\left\{ -\frac{1}{2} \left((f-m)^T, (f^* - m^*)^T \right) K^{-1} \begin{pmatrix} f - m \\ f^* - m^* \end{pmatrix} \right\} \times$$
$$\times \exp\left\{ -\frac{1}{2} \left((FK) f - y \right)^T C_y^{-1} \left((FK) f - y \right) \right\} .$$

Integrate out f^*

$$\int df^* \, p(f, f^* | \theta, y) \propto \left[\int df^* \exp\left\{ -\frac{1}{2} \begin{pmatrix} f-m \\ f^*-m^* \end{pmatrix}^T \begin{pmatrix} K_{xx} & K_{xx^*} \\ K_{x^*x} & K_{x^*x^*} \end{pmatrix}^{-1} \begin{pmatrix} f-m \\ f^*-m^* \end{pmatrix} \right\} \right] \times$$

$$\times \exp\left\{ -\frac{1}{2} \left((FK)f - y \right)^T C_y^{-1} \left((FK)f - y \right) \right\} .$$

$$= \exp\left\{ -\frac{1}{2} (f-m)^T (K_{xx})^{-1} (f-m) \right\} \exp\left\{ -\frac{1}{2} \left((FK)f - y \right)^T C_y^{-1} \left((FK)f - y \right) \right\} .$$

↳ quadratic form in f.

For <u>linear data</u>, $\tilde{p}(f)$ is a Gaussian

$$\tilde{p}(f | \theta, y) = \mathcal{N}(f; \tilde{m}, \tilde{K}_{xx}) .$$

where
$$\begin{cases} \tilde{m} = m + K_{xx} (FK)^T C_{yT}^{-1} \left[y - (FK)m \right] \\ \tilde{K}_{xx} = K_{xx} - K_{xx} (FK)^T C_{yT}^{-1} (FK) K_{xx} \\ C_{yT} = (FK) K_{xx} (FK)^T + C_y \end{cases}$$

<u>N.B.</u> : $(\tilde{K}_{xx})_{ii} = (\tilde{\Delta} f_i)^2 \leq (K_{xx})^2_{ii} = (\Delta f_i)^2$

since C_{yT} is >0 def.

Posterior dist. yields reduced errors on f:

Integrate out f. Slightly trickier.

$$\tilde{p}(f^* | \theta, y) = \mathcal{N}(f^*; \tilde{m}^*, \tilde{K}_{x^*x^*}) .$$

$$\tilde{m}^* = m^* + K_{x^*x} (FK)^T C_{yT}^{-1} \left[y - (FK)m \right] .$$

↳ updated central value of $E_{\tilde{p}}[f^*]$ because of the correlations introduced by the prior, K_{x^*x}.

Note that the data is <u>independent</u> of f^*.

(b) $p(\theta|y) \propto p(y|\theta)\, p(\theta)$.

$$p(y|\theta) = \frac{1}{\sqrt{\det(2\pi C_{YT})}} \exp\left\{-\frac{1}{2}\left[y-(FK)m\right]^T C_{YT}^{-1}\left[y-(FK)m\right]\right\}.$$

$$y = (FK)f + \varepsilon \qquad \begin{cases} (FK)f \sim \mathcal{N}\left((FK)m, (FK)K_{xx}(FK)^T\right) \\ \varepsilon \sim \mathcal{N}(0, C_Y) \end{cases}$$

Hyperparameters θ appear in C_{YT}, the posterior distr. can only be sampled by MCMC.

If the distribution is sufficiently narrow, θ can be fixed to the mode of the posterior distr.

2.6 Closure Test - 1

Consider synthetic data generated w. a given f_0.

$$y = (FK)f_0 + \eta, \qquad \eta \sim \mathcal{N}(0, C_Y).$$

Study the case of vanishing error: $C_Y = 0$

$$\Rightarrow \quad \tilde{m} = R_{xx}^{(0)} f_0$$

$$R_{xx}^{(0)} = K_{xx}(FK)^T\left[(FK)K_{xx}(FK)^T\right]^{-1}(FK).$$

↳ smearing kernel.

Note the correspondence w. BG solution.

$$\tilde{m} = a_I y_I \qquad \begin{cases} a_I = K_{xx}(FK)^T\left[(FK)K_{xx}(FK)^T\right]^{-1} \\ y_I = (FK)f_0 \end{cases}$$

cfr. w. BG solution

$$w_k(x_0) = \int dx \; c_k(x) \; k(x, x_0) \longrightarrow (FK)_{ki} (K_{xx})_{i i_0}$$

$$\hat{W}_{kJ} = \int dx \; c_k(x) \, k(x, x') \, c_J(x') \longrightarrow (FK) K_{xx} (FK)^T$$

$$a_I = (\hat{W}^{-1})_{IJ} W_J = (w^T)_J (\hat{W}^{-1})_{JI}$$

$\uparrow$ symmetric

Identical sol.n if $k(x,x')$ in the BG metric

$= k(x,x')$ for the GP.

$\tilde{m} \neq f_0$ even in the absence of stat. fluctuations in the data, there is a "reconstruction" error in the space of functions f.

2.7 Bias & Variance

In data space

$$\mathcal{B} = \sum_I (\tilde{T}_I - y_I) = \sum_I (FK)_{Ii} (\tilde{m}_i - f_{0i})$$

$$= \sum (FK) \left[R^{(0)}_{xx} - 1 \right] f_0 = 0$$

data is reproduced exactly!

$$\mathcal{V} = \mathrm{tr}\left[(FK) \, \tilde{K} \, (FK)^T \right]$$

$$= \mathrm{tr}\left[(FK) \left(1 - R^{(0)}_{xx} \right) K_{xx} (FK)^T \right] = 0 \quad !$$

$C_Y = 0 \Rightarrow$ exact reconstruction of data.

2.8 ClosureTest - 2

Adding exp. error : $y_I = (FK)_{Ii} f_{0i} + \eta_I$

$$R_{xx} = K_{xx} (FK)^T \left[(FK) K_{xx} (FK)^T + C_y \right]^{-1} (FK).$$

$$\left\{ \begin{array}{l} \tilde{m} = R_{xx} f_0 + a^T_{xx} \eta \\ \tilde{K}_{xx} = (1 - R_{xx}) K_{xx} (1 - R_{xx})^T + a^T_{xx} C_y a_{xx} \end{array} \right.$$

$$a^T_{xx} = K_{xx} (FK)^T C_{YT}^{-1}$$

$$\Rightarrow R_{xx} = a^T_{xx} (FK).$$

cfr. $\Phi_{\hat{a}}(x) = \sum_I a_I c_I(x)$ for BG.

connection w. BG.

In data space, we can compute bias & variance

$$\left\{ \begin{array}{l} B = (FK) \left[R_{xx} - 1 \right] f_0 + (FK) a^T_{xx} \eta \\ V = (FK)(1 - R_{xx}) K_{xx} (1 - R_{xx})^T (FK)^T + (FK) a^T_{xx} C_y a_{xx} (FK)^T \end{array} \right.$$

A bit of algebra yields, for $C_y \to 0$

$$B = - C_T^{-1} C_y \left((FK) f_0 + \eta \right)$$

$$= - C_T^{-1} C_y\, y$$

$$V = (FK) \left[(1 - R_{xx}) K_{xx} (1 - R_{xx})^T + a^T_{xx} C_y a_{xx} \right] (FK)^T$$

Explicit dependence on the prior, K_{xx}.

All assumptions are exposed!

No minimization needed. Only sample $p(\theta | y)$ by MCMC.