Introduction to Gaussian Processes

Filipe P. Farias
Teleinformatics Engineering Department
Federal University of Ceará
filipepfarias@fisica.ufc.br

Abstract

The Gaussian processes have proven to be a powerfull framework for robust estimation and a flexible model for non-linear *regression*, case which will be the main object of this work, with some implementations of real situations.

1 Applications in disease mapping

1.1 The model

There's several applications using GP and here we'll resume an example for disease mapping presented by [?]. Then let's assume that our phenomenon is ruled by an function f. But, we interested in the distribution of them, considering the approach presented in this work. So, we may say that we evaluated each observation y_i from an unknown function f_i. With this we assume that our observations and our functions are independent and then we can evaluate our joint distribution for the likehood by the product of each one [?].

$$\begin{cases} y_1, y_2, \ldots, y_n \sim \prod_{i=1}^{n} Poisson\left(e_i \exp(f_i)\right) & \text{(1.1a)} \\ f(\mathbf{x})|\theta \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\theta)\right) & \text{(1.1b)} \\ \theta \sim \text{half-t}(\nu, A)^* & \text{(1.1c)} \end{cases}$$

In this case, we used the Poisson distribution for the likelihood because the nature of the process. The phenomenon here is the relative risk of death μ in a region of the country. So, if we consider y the counting of deaths on this region, we can model the phenomenon with a Poisson process which mean in each region is given by the increasing rate of deaths. At this point we have defined e as the standardized expected number of deaths [?], what multiplied by μ reveals, in mean, the rate of deaths in that region. For numerical reasons, we transform $f = \log(\mu)$. Finally, we assume an uncertainty over the parameters of the kernel functions too, then, our hierarchical model stays for the posterior distribution as

$$p(\mathbf{f}|\mathbf{y}, \mathbf{x}) \propto \int p(\mathbf{y}|\mathbf{f}) \mathcal{GP}\left(\mathbf{f}|m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'|\theta)\right) p(\theta) d\theta. \tag{1.2}$$

This function isn't analytically tractable because of the Poisson process, but it is possible its evaluation with approximation methods.

[*] The values ν and A are not arbitrary, but deterministic [?].

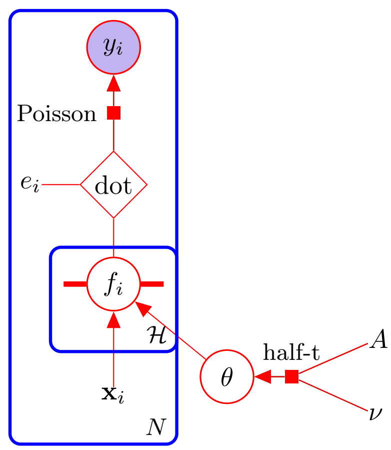

Figure 1: Graphical model for the GP for regression. Colored circles represent observed variables and whited ones represent the unknowns. The thick horizontal bar represents a set of fully connected nodes of the Gaussian field. Note that an observation y_i is conditionally independent of all other nodes given the corresponding latent variable, f_i. Because of the marginalization property of GPs addition of further inputs, $\mathbf{x}$, latent variables, f, and unobserved targets, y_*, does not change the distribution of any other variables.