Introduction to Gaussian Processes - RAW

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

There's several applications using GP and here we'll resume an application for disease mapping. By [Vanhatalo et al., 2010] the problem concerns to use the latent Gaussian field and observe it indirectly from the data, aiming to infer a function for the phenomenon. For this scenario, the data are counts of mortality or morbidity occurrence per area and the pursued function is the relative risk for that area, what is modeled by a Poisson process discussed by [Best et al., 2005].

1.1 The model

The Poisson process is defined if the occurrence, this is the count of deaths y, in a given area i is a Poisson variable and if the counting in one area does not affect the counting in another.

$$y_i \sim Poisson(e_i \mu_i) \tag{1.1}$$

The Poisson variable y_i in each area is defined by its process rate $e_i \mu_i$ [Best et al., 2005,Samat and Percy, 2012], being

$$e_i = \sum_{r=1}^{R} \left(\frac{\sum_{i=1}^{n} y_{i,r}}{\sum_{i=1}^{n} p_{i,r}} \right) p_{i,r} \tag{1.2}$$

the standardized expected number of deaths and μ_i the relative risk of the area. The sum is weighted by social factors, dividing the population p in groups r. In this work, the objective is to infer this relative risk given the data [Lawson, 2013].

Then let's assume $f = \log(\mu)$ [Best et al., 2005]. So, we may say that we evaluated each observation y_i is Poisson distributed and its mean is given by an unknown function $e_i \exp(f_i)$. With this we assume that our observations of the functions are independent and then we can evaluate the joint distribution for likelihood by the product of each one [Vanhatalo, 2010].

We'll model the log risk as a Gaussian process, then latent variables f_i are realizations of the latent function f at the inputs $\mathbf{x}_i$. The GP will be defined by a covariance function $k(\mathbf{x}, \mathbf{x}')$ and a mean function $m(\mathbf{x})$. For disease mapping, we can assume the mean being $m(\mathbf{x}) = \mathbf{0}$, what implies that for a zero function, the risk will be one if there's no spatial variations. The covariance function has a collection of hyperparameters θ.

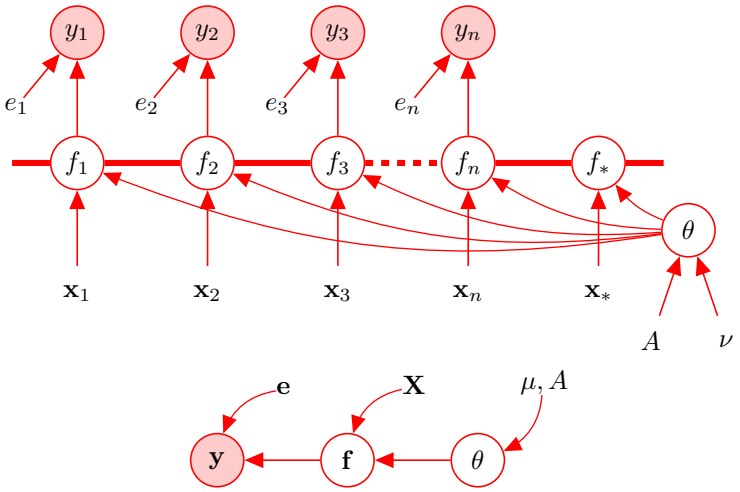

Figure 1: Graphical model for the GP for regression. Colored circles represent observed variables and whited ones represent the unknowns. The thick horizontal bar in f_i node represents a set of fully connected nodes of the Gaussian field. Note that an observation y_i is conditionally independent of 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.

We can represent with a graphical model the dependency of the variables, as in the Figure 1. Then, we can write the model as follows:

$$\begin{cases} y_1, y_2, \ldots, y_n | \mathbf{f} \sim \prod_{i=1}^{n} Poisson\left(e_i \exp(f_i)\right) & \text{(1.3a)} \\ f(\mathbf{x})|\theta \sim \mathcal{GP}\left(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'|\theta)\right) & \text{(1.3b)} \\ \theta \sim \text{half-Student-t}(\nu, A)^* & \text{(1.3c)} \end{cases}$$

We're interested in evaluate the conditional posterior of the latent variable $\mathbf{f}$, i.e. to express the uncertainty about the risk given the knowledge about the number of deaths and the hyperparameters. We can write by the model in the Figure 1, the following conditional

$$p(\mathbf{f}, \mathcal{D}, \theta) = p(\mathbf{y}|\mathbf{f}, \mathbf{e}) p(\mathbf{f}|\mathbf{X}, \theta) p(\theta). \tag{1.4}$$

being $\mathcal{D} = \{\mathbf{X}, \mathbf{y}, \mathbf{p}\}$. The constant $\mathbf{e}$ evaluated in (1.2) depends of $\mathbf{y}$ and $\mathbf{p}$, then we can omit from the probability function. Note that it isn't considered as a random variable even y being a random. This occurs because the evaluation of e_i needs some information about y as in which social group r the death was counted (1.2), what can't be modeled if y was treated as random without rise the complexity of the model. This would lead to a second inference that would be the classification of y. Finally we obtain the posterior for the latent values by marginalizing it, then we have

$$p(\mathbf{f}|\mathcal{D}) = \frac{p(\mathbf{f}, \mathcal{D})}{p(\mathcal{D})} = \frac{\int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) p(\theta) d\theta}{\int \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) p(\theta) d\mathbf{f} d\theta}, \tag{1.5}$$

omitting the constant $\mathbf{e}$ for brevity. The quantity $p(\mathcal{D})$ can be represented by the marginalization of $p(\mathbf{f}, \mathcal{D})$ with respect to $\mathbf{f}$, i.e. $\int p(\mathbf{f}, \mathcal{D}) d\mathbf{f}$. This integral can't be evaluated analytically, then we use approximation methods to do this.

1.2 Conducting the inference

In the last part, we defined the model for the the observations f assuming a Bayesian approach of it. Now we have to obtain the posterior distribution $p(\mathbf{f}|\mathcal{D})$ and conduct the

*The values ν and A are not arbitrary, but deterministic [Vanhatalo et al., 2010].

prediction applying the concepts presented in the Section 4 by finding $\mathbf{f}_*$. Differently of the general Bayesian models, in this application we are not interested in predict the values of y, but the risk $e = \exp(f)$ after observed the deaths y, then we do not derive the predictive distribution. First we have to simulate the distributions of the (1.5).

References

[Best et al., 2005] Best, N., Richardson, S., and Thomson, A. (2005). A comparison of bayesian spatial models for disease mapping. Statistical Methods in Medical Research, 14(1):35–59. PMID: 15690999.

[Lawson, 2013] Lawson, A. (2013). Statistical Methods in Spatial Epidemiology. Wiley Series in Probability and Statistics. Wiley.

[Samat and Percy, 2012] Samat, N. A. and Percy, D. F. (2012). Vector-borne infectious disease mapping with stochastic difference equations: an analysis of dengue disease in malaysia. Journal of Applied Statistics, 39(9):2029–2046.

[Vanhatalo, 2010] Vanhatalo, J. (2010). Speeding up the inference in Gaussian process models. PhD thesis, Aalto University School of Science and Technology, Faculty of Information and Natural Sciences, Department of Biomedical Engineering and Computational Science, Aalto.

[Vanhatalo et al., 2010] Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse gaussian processes. Statistics in medicine, 29:1580–1607.