Rain Gauge Anomaly Detection Using Gaussian Processes

Ciira wa Maina
Nyeri

September 21, 2022

1 Introduction

We aim to explore the use of Gaussian processes to model rainfall potential and exploit this model to detect anomalous rain gauge readings. Faulty readings commonly appear as:

1. Null readings when rainfall occurred
2. Extreme readings when lower rainfall occurred

Here we will focus on the initial case, zero readings when the rainfall was greater than zero.

1.1 Gaussian Processes

Gaussian processes (GPs) can be viewed a priors over functions. For a GP, any finite number of samples from the random process have a joint Gaussian density. The covariance matrix of this joint Gaussian density is computed from a kernel function. A common kernel is the radial basis function (RBF) kernel given by

$$\kappa_{\text{rbf}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right). \tag{1}$$

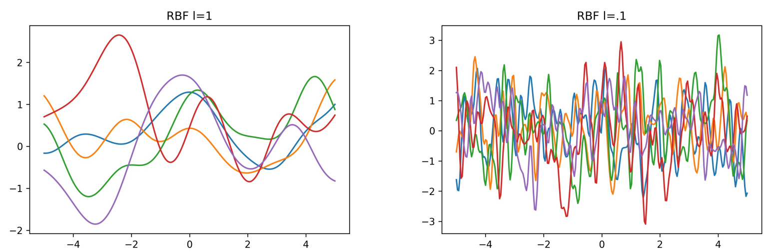
This kernel has two parameters ℓ and σ^2. They determine the nature of functions drawn from the GP.

Sample functions from GPs with an RBF kernel are shown in Figure 1 for two values of ℓ.

1.2 GP Regression and Bayesian Inference

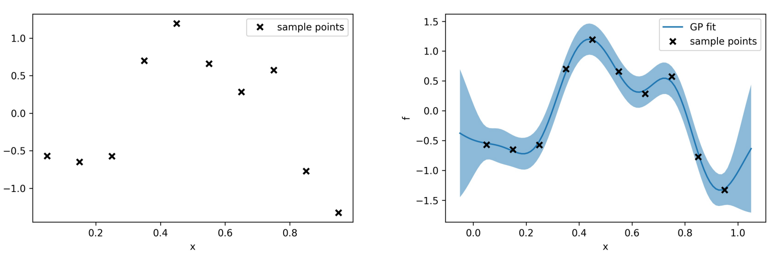
Consider data assumed to follow the following model.

$$y(x) = f(x) + \epsilon$$



(a) $\ell = 1$ (b) $\ell = 0.1$

Figure 1: Functions drawn from a GP as lengthscale varies.



(a) Observations (b) Posterior

Figure 2: GP Regression.

where $f(x) \sim \mathcal{GP}(0, k(x, x'))$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. By combining the prior over f with the likelihood induced by the noise model, we can determine the posterior distribution of f.

$$p(f|y) = \frac{p(y|f)p(f)}{p(y)}$$

With the noise assumed Gaussian, this posterior is tractable and we can draw samples from it. In addition we can determine $p(y)$ which is the marginal likelihood and maximise it to obtain estimates of the kernel parameters.

1.3 Model

The rainfall is observed at L locations, $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_L\}$, over T time intervals with the potential for rain at time t

$$f_t(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}). \tag{2}$$

The rainfall y_{it} observed at location i at time t is given as follows

$$y_{it} = \begin{cases} f_{it} + \epsilon_{it} & \text{w.p} \quad \theta_i \\ \epsilon_{it} & \text{w.p} \quad 1 - \theta_i \end{cases} \tag{3}$$

where $\epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is observation noise. If we introduce a Bernoulli random variable $d_{it} \sim \mathcal{B}(\theta_i)$, we can write the observation model as

$$y_{it} = f_{it} d_{it} + \epsilon_{it} \tag{4}$$

Let $\mathbf{y}_t = [y_{1t}, \ldots, y_{Lt}]^T$ and similarly introduce $\mathbf{f}_t, \mathbf{d}_t, \boldsymbol{\theta}$.

The joint distribution for our problem is given by

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{D} | \boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{t=1}^{T} p(\mathbf{y}_t, \mathbf{f}_t, \mathbf{d_t} | \boldsymbol{\theta}, \boldsymbol{\gamma}) \tag{5}$$

Where $\boldsymbol{\gamma}$ are the hyper-parameters controlling the nature of $\mathbf{K}$. In our case we will use an RBF kernel and therefore $\boldsymbol{\gamma} = \{\sigma^2, \ell\}$.

Let's assume $T = 1$. We have

$$p(\mathbf{y}_t, \mathbf{f}_t, \mathbf{d}_t | \boldsymbol{\theta}, \boldsymbol{\gamma}) = p(\mathbf{y}_t | \mathbf{f}_t, \mathbf{d}_t) p(\mathbf{f}_t | \boldsymbol{\gamma}) p(\mathbf{d}_t | \boldsymbol{\theta}) \tag{6}$$

where

$$p(\mathbf{y}_t | \mathbf{f}_t, \mathbf{d}_t) = \prod_{i=1}^{L} \mathcal{N}(y_{it} | d_{it} f_{it}, \sigma_\epsilon^2),$$

$$p(\mathbf{d}_t | \boldsymbol{\theta}) = \prod_{i=1}^{L} \theta_i^{d_{it}} (1 - \theta_i)^{1 - d_{it}},$$

and

$$p(\mathbf{f}_t | \boldsymbol{\gamma}) = \mathcal{N}(\mathbf{f}_t | \mathbf{0}, \mathbf{K}_t).$$

1.4 Inference

We would like to determine:

- The parameters $\boldsymbol{\gamma}, \boldsymbol{\theta}$
- The posterior $p(\mathbf{f}_t | \mathbf{y}_t)$
- The predictive distribution $p(\mathbf{f}_t^* | \mathbf{y}_t)$

To get the parameters, we would need to maximize the marginal likelihood

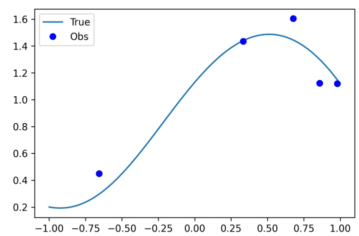
$$\boldsymbol{\gamma}^*, \boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\gamma}} \sum_{\mathbf{d}_t} \int p(\mathbf{y}_t, \mathbf{f}_t, \mathbf{d}_t | \boldsymbol{\theta}, \boldsymbol{\gamma}) d\mathbf{f}_t$$

This is intractable due to the sum over all possible $\mathbf{d}_t$. We proceed via variational inference where we optimize the lower bound on the marginal likelihood.

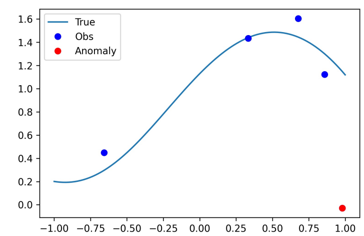
$$\log p(\mathbf{y}_t | \boldsymbol{\theta}, \boldsymbol{\gamma}) \geq \mathbb{E}\{\log p(\mathbf{y}_t, \mathbf{f}_t, \mathbf{d}_t | \boldsymbol{\theta}, \boldsymbol{\gamma})\} - \mathbb{E}\{\log q(\mathbf{f}_t, \mathbf{d}_t)\} \tag{7}$$

where

$$q(\mathbf{f}_t, \mathbf{d}_t) = q(\mathbf{f}_t) \prod_{i=1}^{L} q(d_{it})$$



(a) Data with no anomaly



(b) Fifth observation 'anomalous'

Figure 3: Toy data generation.

is the variational approximation to the true posterior.

We proceed as in [1] to derive the variational approximation to the posterior as well as the hyperparameters. We find that

$$q(\mathbf{f}_t) = \mathcal{N}(\mathbf{f}_t|\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$$

and

$$q(d_{it}) = \eta_i^{d_{it}}(1-\eta_i)^{1-d_{it}}$$

1.5 Experiments

1.5.1 1D Toy Example

We explore $\boldsymbol{\theta}^*$ and the predictive mean in a 1D example where the observations are drawn from a GP with a squared exponential kernel. See Figure 3. We generate sample data and introduce a single anomaly. We then run the VB algorithm and determine $\boldsymbol{\theta}^*$ and the predictive mean. Figure 4, shows a comparison of model fits for an ordinary GP and our model. In this experiment, we get $\boldsymbol{\theta}^* = [1, 1, 1, 1, 0]^T$. Figure 5 shows the evolution of $\boldsymbol{\theta}^*$ during the VB iterations.

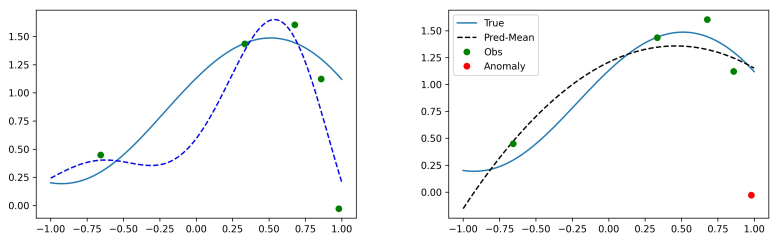
1.5.2 Observations

We note that:

1. The posterior mean appears to 'ignore' the anomalous observation which is promising
2. 'Low' values are assumed anomalous. It appears that the noise variance is over estimated leading to a lack of identifiability.

1.5.3 Next steps

1. Investigate the estimation of noise variance



(a) GP Fit

(b) Anomaly detection model fit

Figure 4: Latent process observed through noise with probability of being switched off.

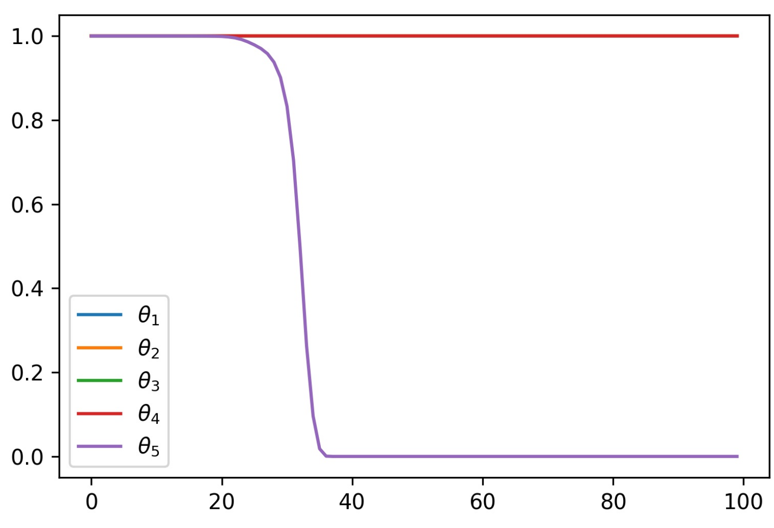


Figure 5: Evolution of $\boldsymbol{\theta}$ during optimization.

2. Verify VB implementation by monitoring the evidence lower bound (ELBO)
3. Experiments with real data

References

[1] Michalis Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. *Advances in neural information processing systems*, 24, 2011.