4M24: Computational Statistics and Machine Learning
High Dimensional MCMC

Samuel McHale sm2431@cam.ac.uk Downing College

May 29, 2023

1 Simulation

a) Gaussian Processes

A Gaussian process ("GP") is an infinite dimensional multivariate Gaussian distribution, with a mean and covariance function. A good choice of covariance function for function approximation is the squared exponential in equation 1, parameterised by a length scale l. The GP can be thought of as a distribution over functions, and we will apply it to find the posterior distribution over an unknown 2D input function given some observations and a GP prior, or, "a bayesian approach to inferring a latent field".

$$\mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \qquad k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right) \tag{1}$$

To work with a GP we must reduce it a high (but finite) dimensional multivariate Gaussian, with a covariance matrix which is the covariance function evaluated across the domain of interest. We will use a GP prior with mean zero:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{0}, C) \qquad C_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

Samples from this GP prior for varying length scales l are shown in figure 1. The small length scale on the left of the figure has a shorter range over which neighbours are correlated, appearing as a bumpier surface, larger length scales are visually smoother.

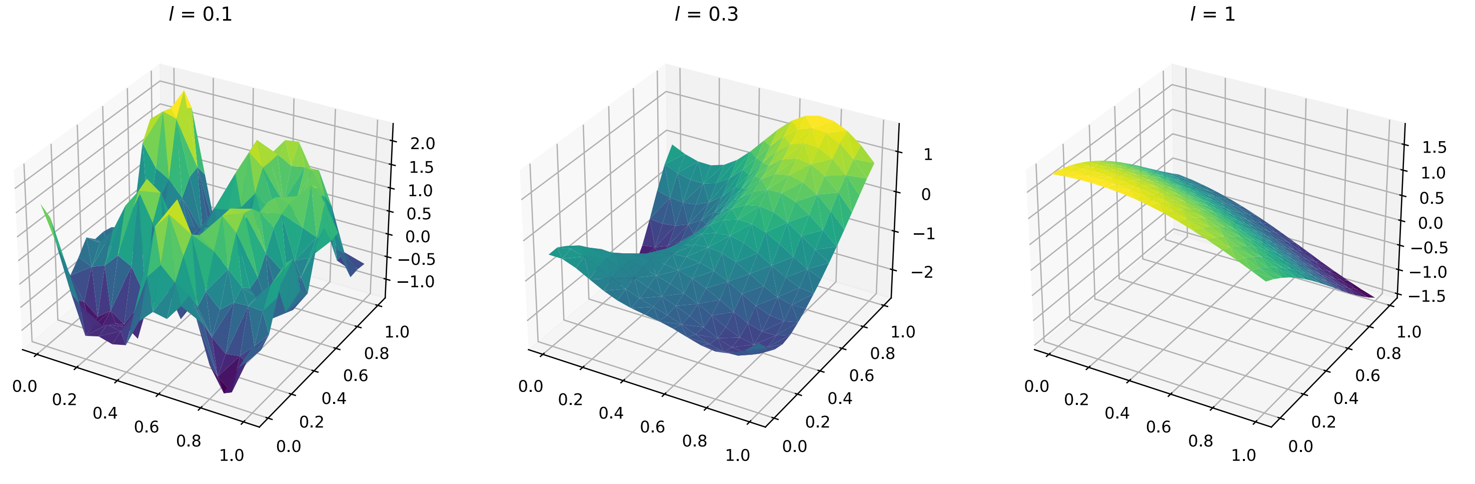


Figure 1: Samples from a Gaussian process prior with varying length scale.

To explore this type of inference we will generate a latent field $\mathbf{u}$ on a $D \times D$ grid, then sub sample 1/4 values at random and add noise to form the observation $\mathbf{v}$. The objective will be to infer $\mathbf{u}$ from $\mathbf{v}$. G is a matrix with only a single one in each column to achieve this sub sampling:

$$\mathbf{v} = G\mathbf{u} + \epsilon \qquad p(\epsilon) = \mathcal{N}(0, 1) \tag{3}$$

Figure 2 shows u with v overlaid.

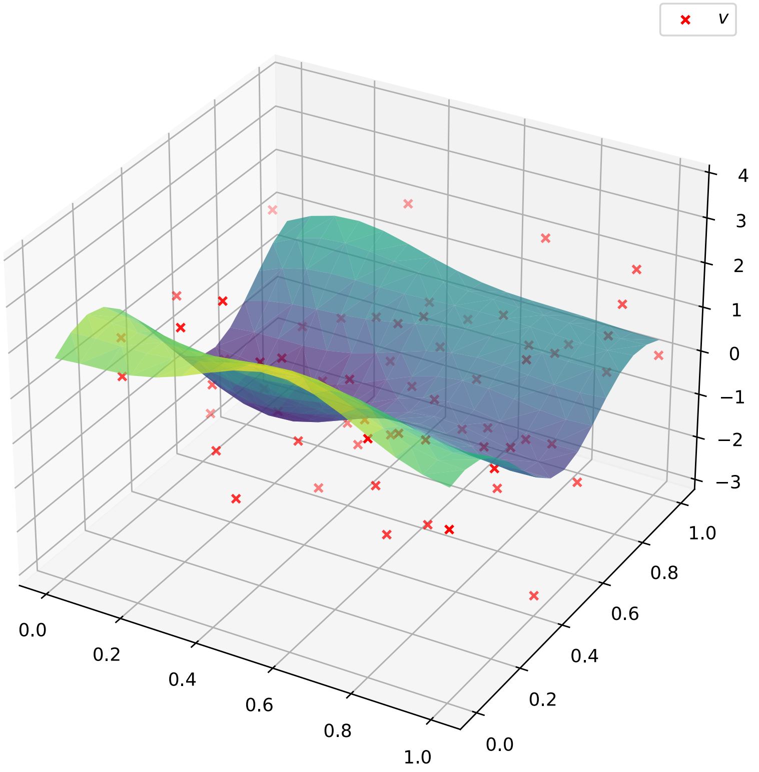


Figure 2: Sub sampled data with noise added $\mathbf{v}$ overlaid on the underlying latent field $\mathbf{u}$.

b) Sampling to Infer the Latent Field

To find the mean of the posterior, we will use the Metropolis Hastings algorithm ("MH") to obtain samples from the posterior and then average them. This avoids needed a closed form solution for the posterior mean, we only need a likelihood proportional to the pdf of the posterior mean, using Bayes rule and from substituting (2) and (3):

$$p(\mathbf{u}|\mathbf{v}) \propto p(\mathbf{v}|\mathbf{u}) \cdot p(\mathbf{u}) = \mathcal{L} \tag{4}$$

$$\log \mathcal{L} = \log p(\mathbf{u}) + \sum_{i}^{M} \log p(v_i|u_i) \tag{5}$$

$$\log \mathcal{L} = \underbrace{-\frac{1}{2}\mathbf{u}^T C^{-1}\mathbf{u}}_{\text{prior}} \underbrace{-\frac{1}{2}(G\mathbf{u}-\mathbf{v})^T(G\mathbf{u}-\mathbf{v})}_{\text{likelihood}} + k \tag{6}$$

MH enables sampling from the target distribution $\pi = \mathcal{L}$, on each iteration a proposed sample $\mathbf{u}' \sim q(\mathbf{u})$ is accepted or rejected. The ratio of the target distribution in the log domain is just the difference, so we can neglect the constant k in (6).

$$\mathbf{u}^{(k+1)} = \begin{cases} \mathbf{u}' & \text{w.p.} \quad \alpha \\ \mathbf{u}^{(k)} & \text{w.p.} \quad 1-\alpha \end{cases} \qquad \alpha = \min\left(\frac{\pi(\mathbf{u}')q(\mathbf{u}^{(k)}|\mathbf{u}')}{\pi(\mathbf{u}^{(k)})q(\mathbf{u}'|\mathbf{u}^{(k)})}, 1\right) \tag{7}$$

For Gaussian Random Walk MH ("GRW") the proposal distribution q is Gaussian

$$q(\mathbf{u}^{(k)}|\mathbf{u}') = \mathcal{N}(\mathbf{u}^{(k)}|\mathbf{u}', 1) \qquad => \qquad \mathbf{u}' = \mathbf{u}^{(k)} + \beta \cdot \boldsymbol{\epsilon} \tag{8}$$

Figure 3 shows the latent field $\mathbf{u}$, observations $\mathbf{v}$, the inferred estimate of $\mathbf{u}$ and the error field using GRW sampling. The estimate is good but not perfect, the blue / purple patches across the centre of the field are slightly out of position and the yellow area at the bottom is a different size / shape. The error field visualises the absolute value of the differences between the two fields.

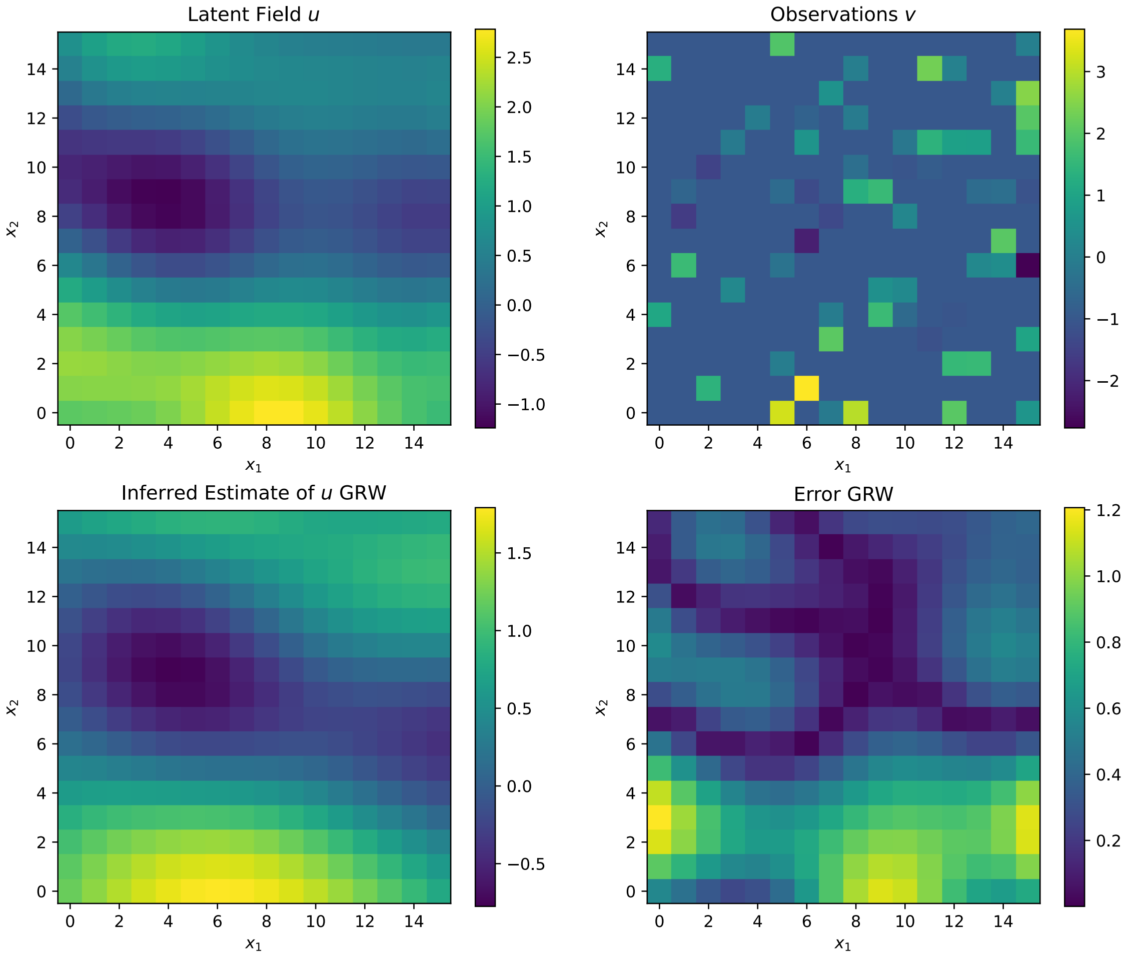


Figure 3: Latent field to be inferred (top left) alongside observations (top right) with inferred field (bottom left) and the error between the actual and inferred latent field (bottom right) with GRW sampling.

For Precondition Crank Nicholson MH ("PCN"), the proposal is

$$\mathbf{u}' = \sqrt{1-\beta^2} \cdot \mathbf{u}^{(k)} + \beta \cdot \boldsymbol{\epsilon} \tag{9}$$

Figure 4 shows the estimate and error found using PCN sampling, they are similar to those obtained using GRW sampling.

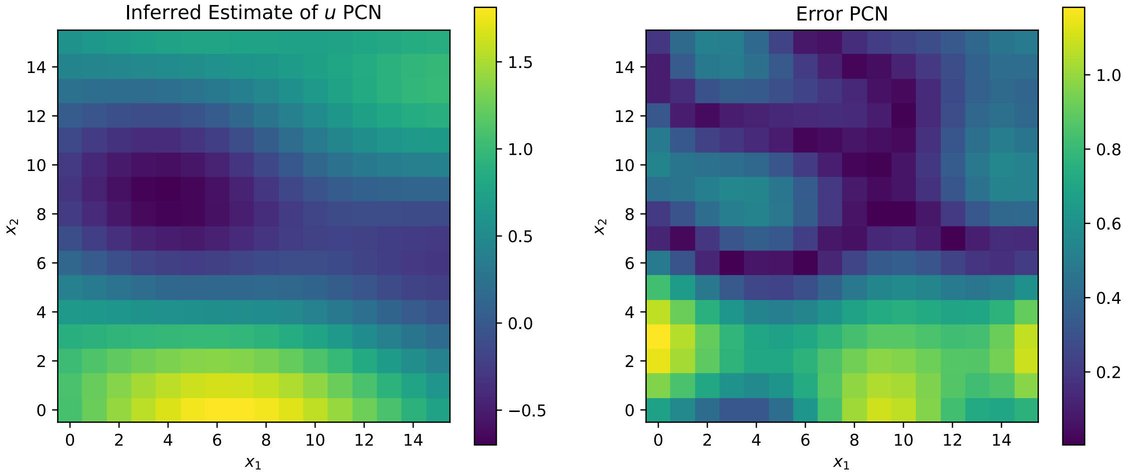


Figure 4: Inferred field and the error between actual and inferred with PCN sampling.

For this Gaussian likelihood problem, a closed form expression for the posterior is relatively easy to derive

$$p(\mathbf{u}|\mathbf{v}) = \mathcal{N}(\mathbf{u}|\boldsymbol{\mu},\Sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{u}^T \underbrace{\Sigma^{-1}} \mathbf{u} - 2\mathbf{u}^T \underbrace{\Sigma^{-1}\boldsymbol{\mu}})\right) \tag{10}$$

From (6):

$$p(\mathbf{u}|\mathbf{v}) \propto p(\mathbf{v}|\mathbf{u})p(\mathbf{u}) \propto \exp\left(-\frac{1}{2}(\mathbf{u}^T C^{-1}\mathbf{u} + (\mathbf{v} - G\mathbf{u})^T(\mathbf{v} - G\mathbf{u}))\right) \tag{11}$$

Omitting constants wrt. $\mathbf{u}$:

$$p(\mathbf{u}|\mathbf{v}) \propto \exp\left(-\frac{1}{2}(\mathbf{u}^T C^{-1}\mathbf{u} - 2\mathbf{u}^T G^T\mathbf{v} + \mathbf{u}^T G^T G\mathbf{u})\right) \tag{12}$$

$$p(\mathbf{u}|\mathbf{v}) \propto \exp\left(-\frac{1}{2}(\mathbf{u}^T \underbrace{(C^{-1} + G^T G)} \mathbf{u} - 2\mathbf{u}^T \underbrace{G^T\mathbf{v}})\right) \tag{13}$$

Then equating $\mathbf{u}$ and $\mathbf{u}^T\mathbf{u}$ terms in (10) and (13)

$$\Sigma = (C^{-1} + G^T G)^{-1} \qquad \mu = \Sigma G^T \mathbf{v} \tag{14}$$

We can use this to compare the performance, figure 5 shows the error in the mean for both algorithms as more samples are generated relative to the exact analytical solution. PCN clearly performs much better exploring the true posterior faster and also achieves a lower error after the burn in period.

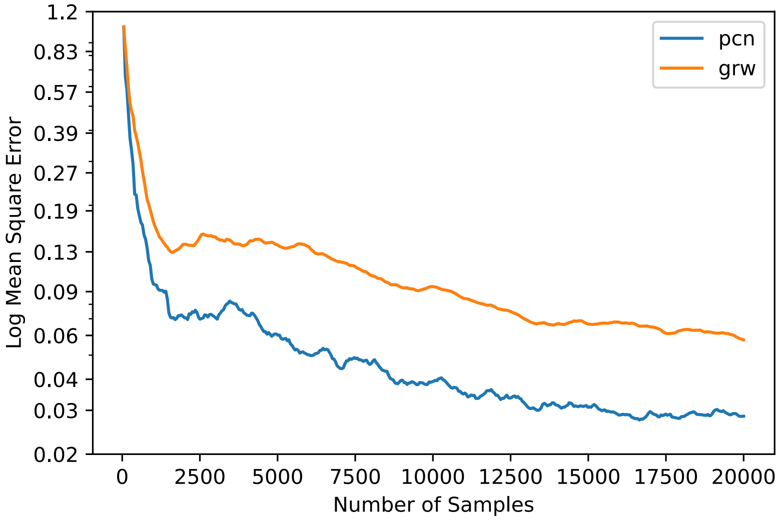
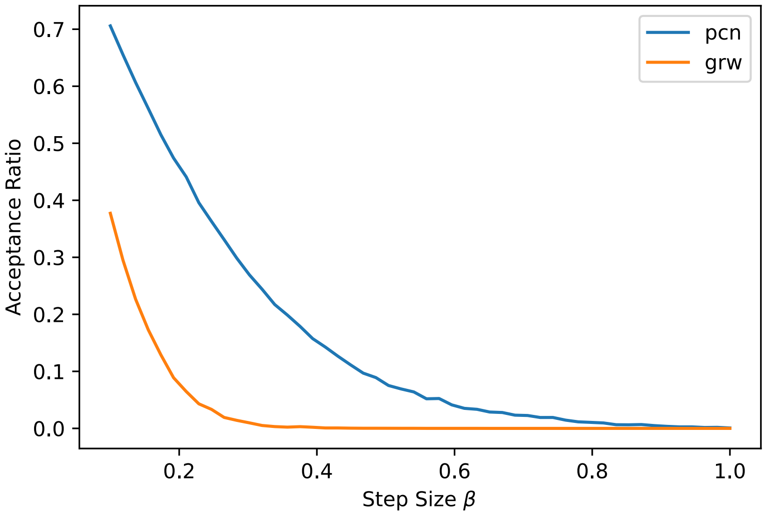


Figure 5: Error of posterior with GRW and PCN wrt. analytical solution.

Both algorithms use a step size parameter β, this controls the distance of the step from one sample to the next. A step size too small will explore the posterior distribution slowly, there will be a longer burn in and more correlation between samples, larger step size means samples are less likely to be accepted. The effect of varying β between 0 and 1 is shown in figure 6, PCN's acceptance ratio is more resistant to increasing β, remaining higher throughout.

The relationship between acceptance ratio and step size matters because a larger step size will explore the posterior more in a fixed number of samples. This also gives the samples less correlation. However, this is a trade off with the acceptance ratio, as a small acceptance ratio will take longer to find each sample as more samples are calculated and rejected.



s

Figure 6: Relationship between acceptance ratio and step size with GRW and PCN sampling.

With a constant size dataset of observations, increasing the number of data points in the inference grid N affects the acceptance ratio as shown in figure 7. PCN actually has constant acceptance ratio as the dimensions grow, because it is properly defined in an infinite dimensional Hilbert space unlike GRW.

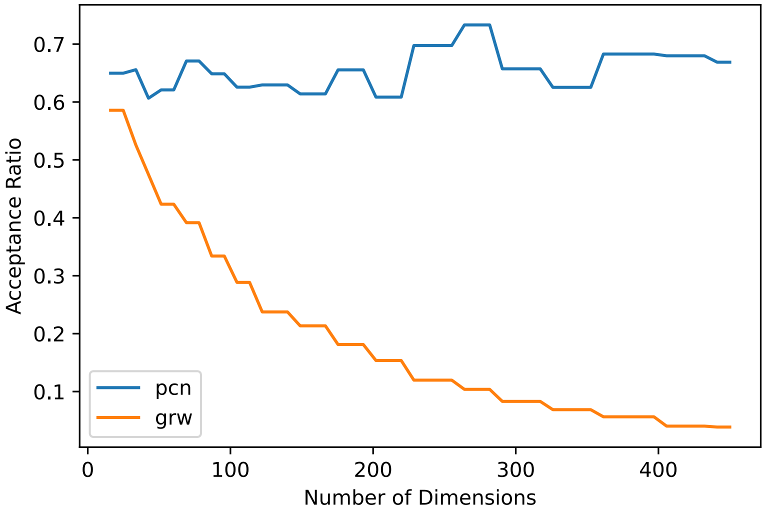


Figure 7: Relationship between acceptance ratio and dimensions N for constant M.

c) Probit Likelihood

We now consider a probit latent field $\mathbf{t}$, generated from $\mathbf{u}$ as follows

$$t_i = \begin{cases} 1 & v_i \leq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad p(t_i = 1|\mathbf{u}) = \Phi([G\mathbf{u}]_i) \tag{15}$$

The probit likelihood is

$$p(\mathbf{t}|\mathbf{u}) = \prod_i^M p(t_i = 1|\mathbf{u})^{t_i} \cdot (1 - p(t_i = 1|\mathbf{u}))^{(1-t_i)} \tag{16}$$

$$\log p(\mathbf{t}|\mathbf{u}) = \mathbf{t} \cdot \log \Phi(G\mathbf{u}) + (1 - \mathbf{t}) \cdot \log(1 - \Phi(G\mathbf{u})) \tag{17}$$

The predictive distribution is

$$p(t^* = 1|\mathbf{t}) = \int p(t^* = 1|\mathbf{u})p(\mathbf{u}|\mathbf{t})d\mathbf{u} \approx \frac{1}{N}\sum_{n}^{N} p(t^* = 1|\mathbf{u}^{(n)}) \tag{18}$$

$$= \frac{1}{N}\sum_{n}^{N} \Phi(\mathbf{u}^{(n)}) \qquad \mathbf{u}^{(n)} \sim p(\mathbf{u}|\mathbf{t}) \tag{19}$$

The inferred predictive distribution (using PCN sampling) from (19) in figure 8 shows higher confidence in regions with a lot of evidence of the latent field, these succeed in approximately aligning with the true latent field. Lower confidence is shown in areas with conflicting or little evidence. This confidence level is an advantage of bayesian approaches, the decision rule can now be tailored to different applications. For example if there is a higher penalty for misclassifying a 0 as a 1, we can threshold 1 assignment at a higher probability value > 0.8 or we could add another "unknown" output classification for low confidence areas close to 0.5.

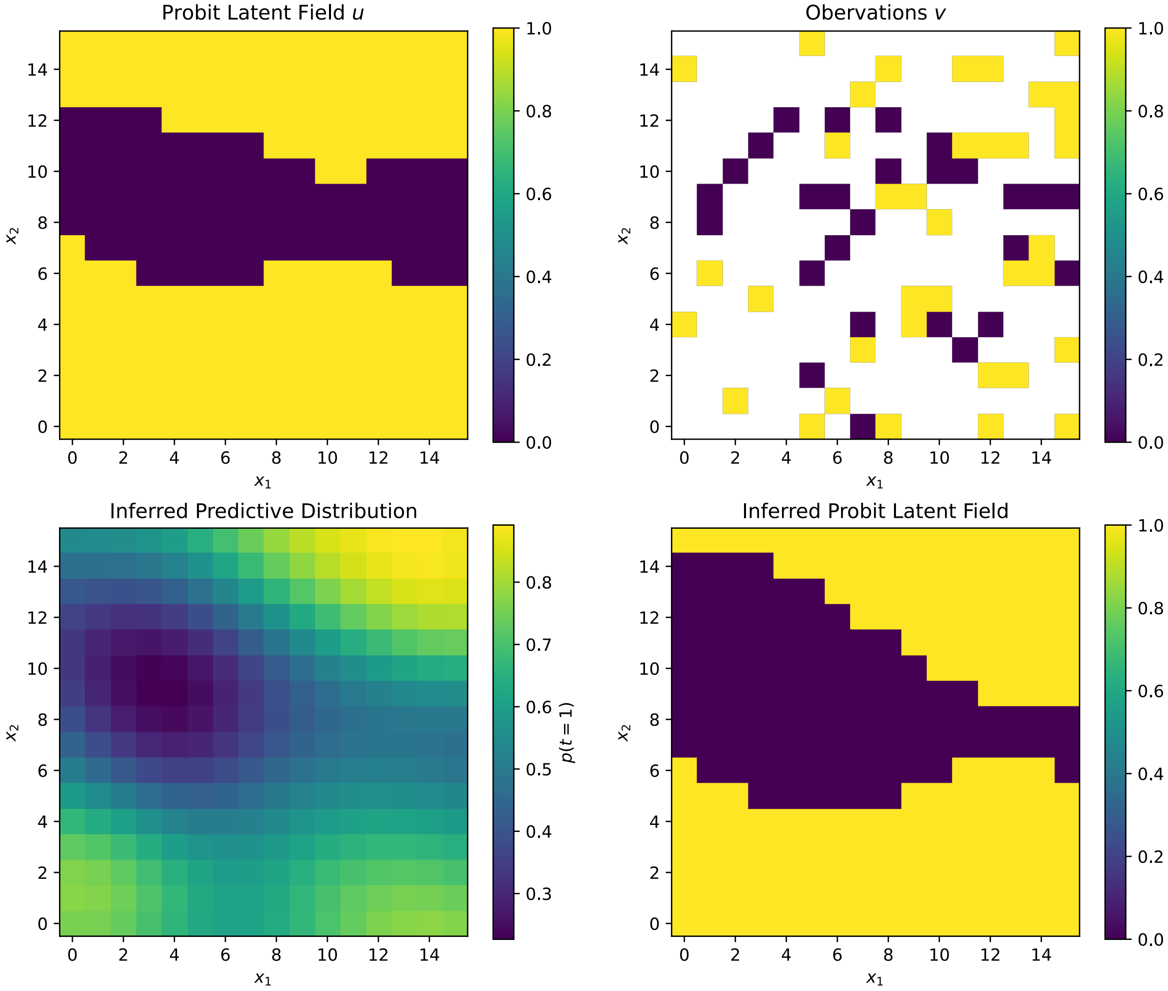


Figure 8: Probit latent field (top left), observations from the probit latent field (top right), the inferred predictive distribution (bottom left) and the hard assignments based on the inferred preditive field (bottom right).

d) Inferring Length Scale

To infer the length scale we define an error as the mean number of misclassifications

$$\text{error} = \text{mean}(\text{abs}(\mathbf{u}_{\text{true}} - \mathbf{u}_{\text{inferred}})) \tag{20}$$

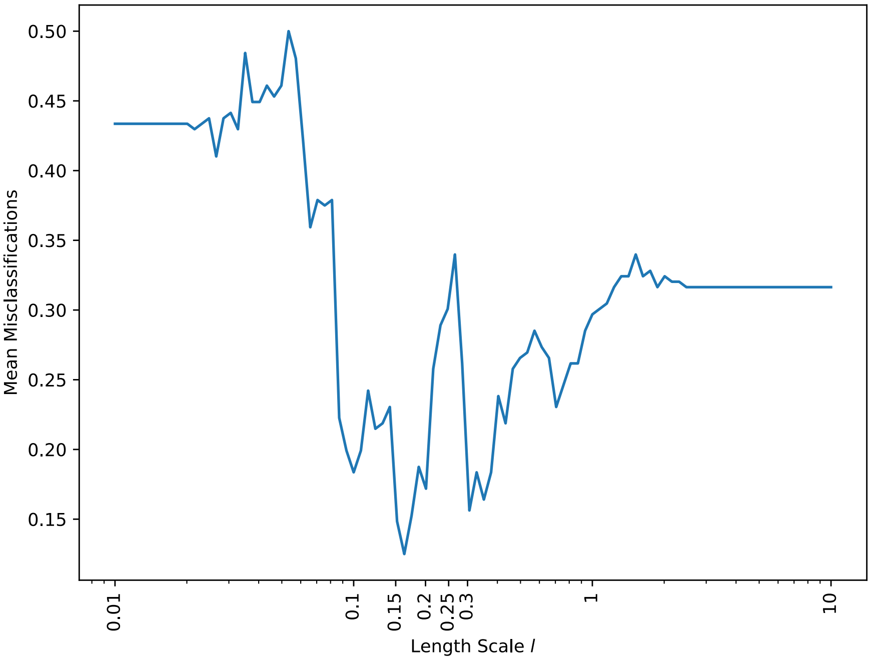


Figure 9: Logarithmic plot of length scale against error, with minimum around 0.1 to 0.3

The inference in figure 9 yields an approximate value, a noisy underestimate in the range of 0.1 to 0.3, of the length scale from which the data was generated, 0.3. This is likely because adding noise reduces correlation between neighbours, which corresponds to a smaller length scale for a GP as shown in part a).

2 Spatial

In this section we will apply this inference to a real world problem, using bike theft count data covering Lewisham, we subsample to give observed data **c** (shown in figure 10) and attempt to infer the expected counts for the rest of Lewisham.

e) Poisson Likelihood

The poisson distribution is a suitable model for bike theft counts, we exponentiate the latent field to give the strictly positive poisson rate θ for the distribution. The likelihood is derived as follows.

$$p(\mathbf{c}|\boldsymbol{\theta}) = \prod_{i=1}^{M} \frac{e^{-\theta_i}\theta_i^{c_i}}{c_i!} \qquad \theta_i = \exp([G\mathbf{u}]_i) \tag{21}$$

$$\log p(\mathbf{c}|\boldsymbol{\theta}) = \sum_{i}^{M} -\theta_i + c_i \cdot \log\theta_i - \log c_i! \tag{22}$$

Grouping constants wrt. **u** as k:

$$\log p(\mathbf{c}|\mathbf{u}) = \sum_{i}^{M} -\exp([G\mathbf{u}]_i) + c_i \cdot [G\mathbf{u}]_i + k \tag{23}$$

f) Spatial Inference

Deriving the expected counts by marginalising over the posterior

$$p(c^* = k|\mathbf{c}) = \int p(c^* = k|\mathbf{u})p(\mathbf{u}|\mathbf{c})d\mathbf{u} \approx \frac{1}{n}\sum_{j}^{n} p(c^* = k|\mathbf{u}^{(j)}) \tag{24}$$

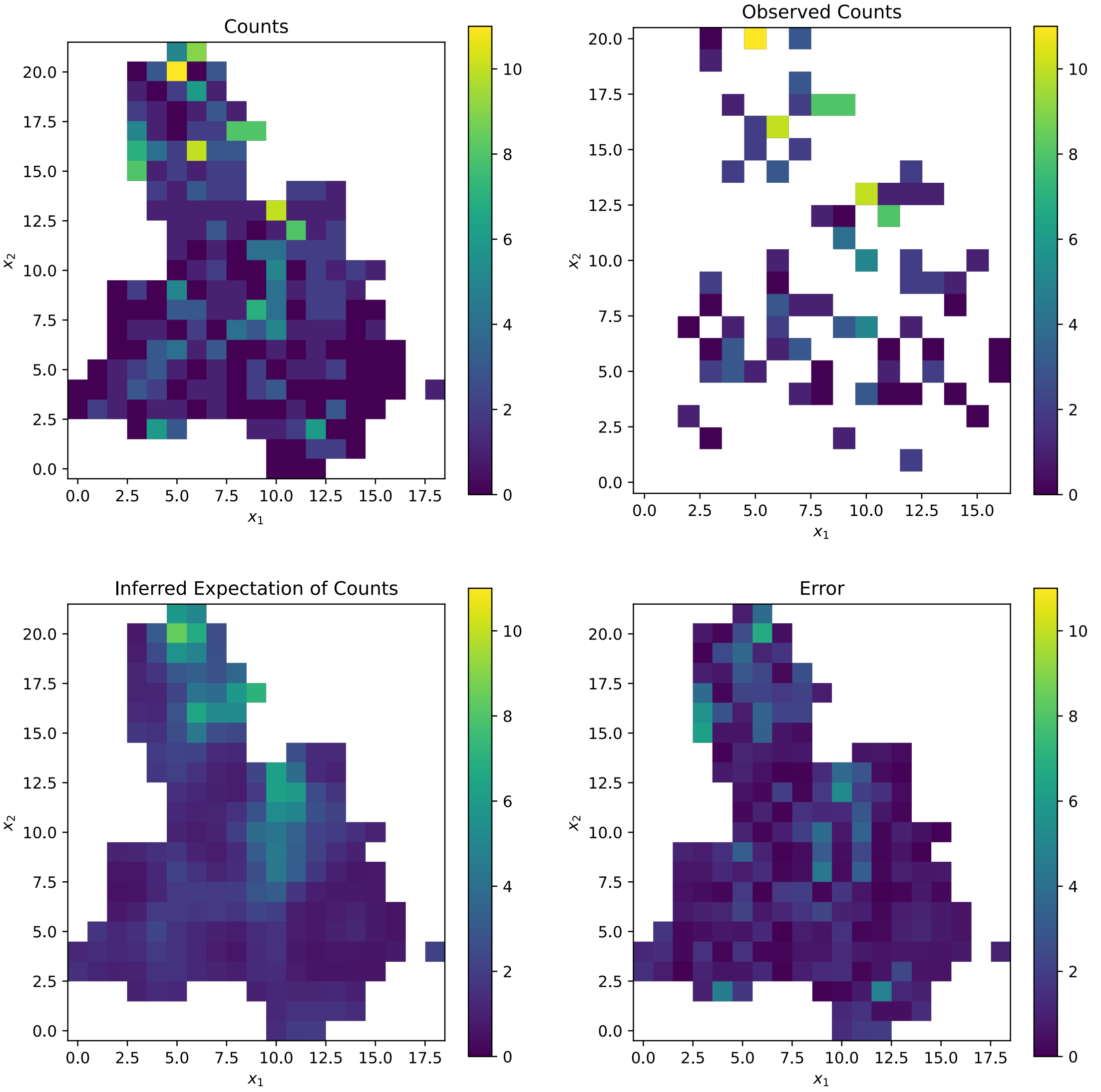
$$\mathop{\mathbb{E}}_{p(c^*|\mathbf{c})}[c^*] = \sum_{k}^{\infty} k \cdot p(c^* = k|\mathbf{c}) \tag{25}$$

Combining (24) with (25) and using the fact that the poisson rate θ is the expectation of the poisson distribution:

$$\approx \frac{1}{n}\sum_{j}^{n}\sum_{k}^{\infty} k \cdot p(c^* = k|\mathbf{u}^{(j)}) = \frac{1}{n}\sum_{j}^{n} \mathop{\mathbb{E}}_{p(c^*|\mathbf{u}^{(j)})}[c^*] = \frac{1}{n}\sum_{j}^{n} \theta^{*(j)} \tag{26}$$

And substituting θ from (21). The inferred expected counts are shown in figure 10 with length scale 1.2.

$$\mathop{\mathbb{E}}_{p(c^*|\mathbf{c})}[c^*] = \frac{1}{n}\sum_{j}^{n} \exp(\mathbf{u}^{*(j)}) \tag{27}$$


Counts

Observed Counts

Inferred Expectation of Counts

Error

Figure 10: True counts, observed counts, inferred expected counts and error for bike theft in Lewisham with $l = 1.2$

Length Scale Analysis

In this scenario we have access to noisy data covering the whole domain of Lewisham and we have subsampled it to form our observations. One option for setting the length scale is to minimise the error wrt. the original data, as done to estimate the length scale in probit classification, however this has two drawbacks:

1. In a another scenario, we might not have data for the whole domain, it would be better to use a general approach to setting l which doesn't rely on the original data. This is a practical issue.
2. The original data is still noisy, the error will be lowest when the length scale is small so that the noise (which may not be zero) is treated as zero and incorporated into the latent field. This is a principle issue, this turns out to just mishandle the noise assumption.

This approach is shown in figure 11, the error plateaus once the length scale is small enough < 0.6, above this value the larger length scale smoothes out the noise, increasing the error wrt. the noisy original data. This is also observed in figure 12, the extremely small length scale is noisy, modelling the noise and the large length scale is overly smoothed.

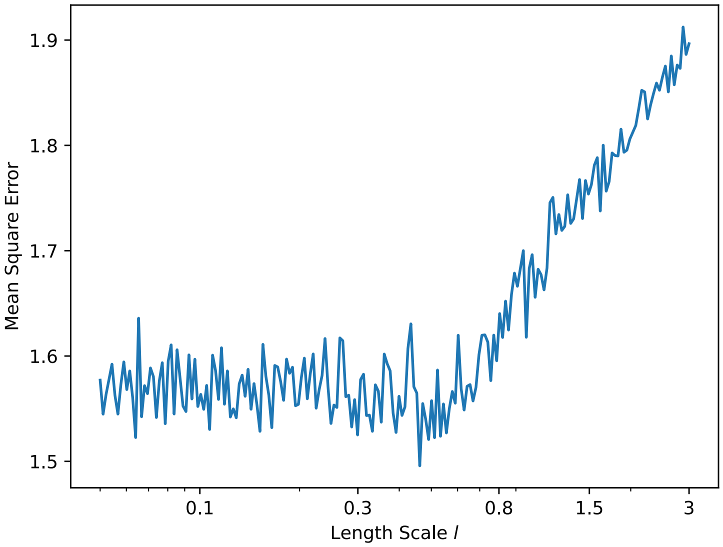


Figure 11: Relationship between error and length scale l.

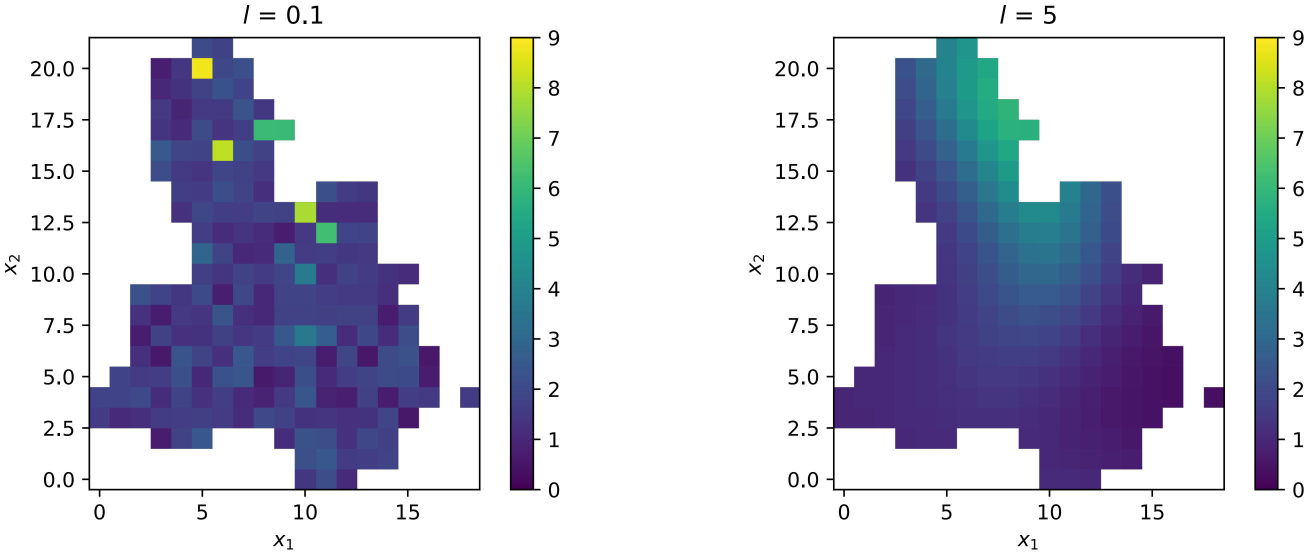


Figure 12: Inferred counts for extreme length scales.

Instead, the model evidence is a better way to quantify the quality of a model, it rewards the quality of the fit and penalises the complexity of the model.

$$p(\mathcal{D}|\mathcal{M}) = p(\mathbf{v}|\mathcal{M}) = \int p(\mathbf{v}|\mathbf{u}, \mathcal{M})p(\mathbf{u}|\mathcal{M})d\mathbf{u} \tag{28}$$

Ideally a closed form expression for the model evidence could be derived but this is likely to be difficult or impossible for this poisson likelihood and GP prior. Another approach to computing it is using a Monte Carlo estimate, although if the function space is very large importance sampling my be necessary. This model is very complex and minimising error didn't seem appropriate as mentioned, so I opted to use my intuition to determine the appropriate noise level. I chose a length scale of $\mathbf{l = 1.2}$ that produces a smooth field while still allowing local variation in the expected count.

3 Conclusion

Gaussian Proccesses have been shown to be a probabilistic, general and effective way of inferring simulated latent fields from observation data. High dimensional MCMC Metropolis Hastings sampling methods are good ways to find the posterior distribution especially in cases where the likelihood is not Gaussian. The PCN-MH proposal mechanism is properly defined in infinite dimensions and it was empirically demonstrated to perform better in high dimensions than another proposal GRW-MH. These methods worked well with a probit likelihood too and managed to roughly infer the length scale used in generation. This was then applied to real spatial data with a poisson likelihood and again performed well. Options for choosing the length scale for this problem were also investigated.