Assignment 2: GMRF:swithNon-Gaussian Data

Wanru Cheng & Kuan Teh Wan

December 2024

Introduction

This assignment explores the use of Gaussian Markov Random Fields (GMRFs) to model non-Gaussian data, specifically Poisson-distributed accident counts in British coal mines. The GMRF model allows us to capture spatial and temporal dependencies in the data, leading to more accurate inference and prediction.

Data

The data for this analysis consists of the number of accidents per year in British coal mines from 1750 to 1980. The data is modeled as a Poisson process with a latent spatial effect. The latent spatial effect is assumed to follow a Gaussian Markov Random Field (GMRF) with a conditional autoregressive (CAR) or simultaneous autoregressive (SAR) structure.

Goal

The primary goal of this assignment is to:

- Estimate the parameters of the GMRF model, including the precision matrix Q and the fixed effect coefficients β.
- Reconstruct the latent field of log-intensities, capturing the spatial and temporal patterns in the accident data.
- Evaluate the performance of the model by assessing the fit to the data and the predictive accuracy.
- Analyze the impact of different model specifications (e.g., different prior distributions, different spatial structures) on the results.

By achieving these goals, we aim to gain a deeper understanding of the factors influencing accident rates in British coal mines and to develop a robust statistical model for future analysis and prediction.

Theory

Q1:

The probability mass function (PMF) of the Poisson distribution is given by:

$$p(y_i|z_i) = \frac{e^{z_i y_i} exp(e^{-z_i})}{y_i!}$$

and the observation log-density:

$$\log p(y_i|z_i) = y_i z_i - e^{z_i} - \log y_i!$$

then we have its first derivative wrt. z_i as

$$\frac{d}{dz_i} \log p(y_i|z_i) = y_i - e^{z_i}$$

also the second derivative

$$\frac{d^2}{d^2 z_i} \log p(y_i|z_i) = -e^{z_i}$$

Q2:

Given that the proportionality of posterior of $\tilde{x}$:

$$p(\tilde{x}|y,\theta) = \frac{p(y|\tilde{x},\theta)p(\tilde{x}|\theta)}{p(y)}$$

then we have the log-posterior of $\tilde{x}$

$$\log p(\tilde{x}|y,\theta) = \log p(y|\tilde{x},\theta) + \log p(\tilde{x}|\theta) - \log p(y)$$

Q3:

Then to further get its first and second derivative(wrt. $\tilde{x}$), we have to expand its likelihood function as well as prior distribution.

Likelihood:

$$\log p(y|\tilde{x},\theta) = \sum_{i=1}^{n} y_i(Ax) - exp(Ax) - log(y_i!)$$

Prior:

$$\log p(\tilde{x}|\theta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\tilde{x}^T Q \tilde{x}$$

1st derivative:

$$\frac{d}{dx} \log p(\tilde{x}|y,\theta) = A^T y_i - A^T exp(Ax) - Qx$$

2nd derivative:

$$\frac{d^2}{d^2x} \log p(\tilde{x}|y,\theta) = -A^T exp(Ax)A - Q$$

Q4:

The approximate logliklihood of $P(\theta|y)$ is:

$$log\, P(\theta|y) \approx \; log\, P(y|\hat{x}^{(0)},\theta) + log\, P(\hat{x}^{(0)}|\theta) - log\, P_G(\hat{x}^{(0)}|y,\theta)$$

$$= log\, P(\theta|y) \approx \; log\, P(y|\hat{x}^{(0)},\theta) + log\, P(\hat{x}^{(0)}|\theta) - log\,(|Q_{x|y}^{\frac{1}{2}}|)$$

Q5:

$$\mathcal{C} = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \end{bmatrix}$$

$$\mathcal{G} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Q6:

If the distance between the points are h instead of 1, the derivative of the basis function will be scaled by 1/h, the integral of the basis function will be scaled by h, thus the C and G matrics are as follows:

$$\mathcal{C}_{\langle} = h * \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \end{bmatrix}$$

$$\mathcal{G}_{\langle} = 1/h * \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Results and Conclusion

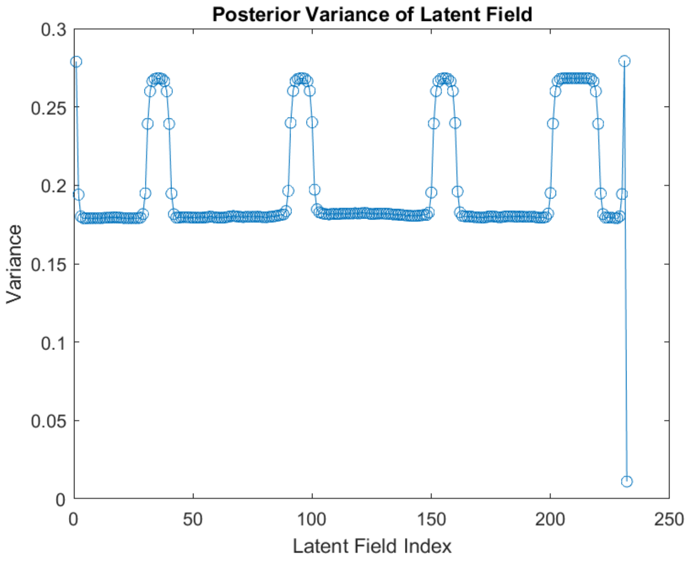

Figure 1: Caption

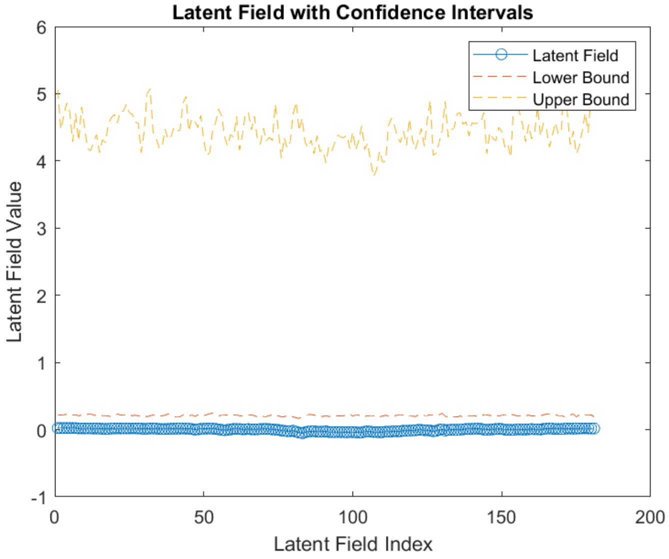

Figure 2: Caption

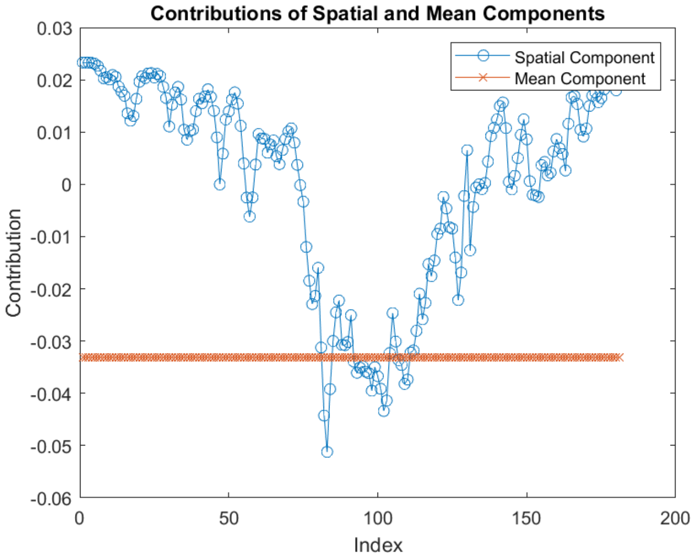

Figure 3: Reconstruction of Latent Field

Spatial Component:

The spatial component demonstrated significant fluctuations, indicating that spatial variations play a crucial role in shaping the reconstructed field. The amplitude of these fluctuations was substantial, suggesting that spatial dependencies are highly influential in determining the underlying process. While the specific spatial patterns were not immediately interpretable without further context, the dominance of the spatial component underscores the importance of incorporating spatial relationships into the model.

Mean Component:

In contrast, the mean component exhibited relatively little variation, remaining largely constant with a slight downward trend. This suggests that the mean component primarily captures a global average or baseline level in the data, with spatial variations having a more dominant influence on the overall field. The smaller magnitude of the mean component compared to the spatial component further emphasizes the importance of spatial factors in explaining the observed data.

Overall:

These observations highlight the critical role of spatial dependencies in the reconstructed field. The dominance of the spatial component suggests that models that fail to account for spatial relationships may provide inaccurate representations of the underlying process. Further analysis of the spatial component could reveal specific spatial patterns and structures, providing valuable insights into the underlying processes generating the data.