4M24 Coursework 2024/2025

High-Dimensional MCMC

Computational Statistics & Machine Learning

Coursework 4M24: High-Dimensional MCMC

- 25% of overall grade
- ~10 hours
- Python 'Skeleton' code provided
- Deadline start of Lent term: **Wednesday 22nd January 2025**
- Coursework sheet with questions
- Coursework files on Moodle

Deliverables

- Maximum 10 page report (including any figures & appendix)
- Answers to questions (a)-(f)
- No need to include code

Coursework Files

coursework.pdf

- Coursework description with questions (a)-(f)

functions.py

- Plotting functions provided
- MCMC algorithms with gaps – fill in TODO

simulation.py

- Questions (a)-(d)

spatial.py

- Questions (e)-(f)

data.csv

- Bike theft count data, (x,y) locations and corresponding number of bike thefts

Part I : Simulation

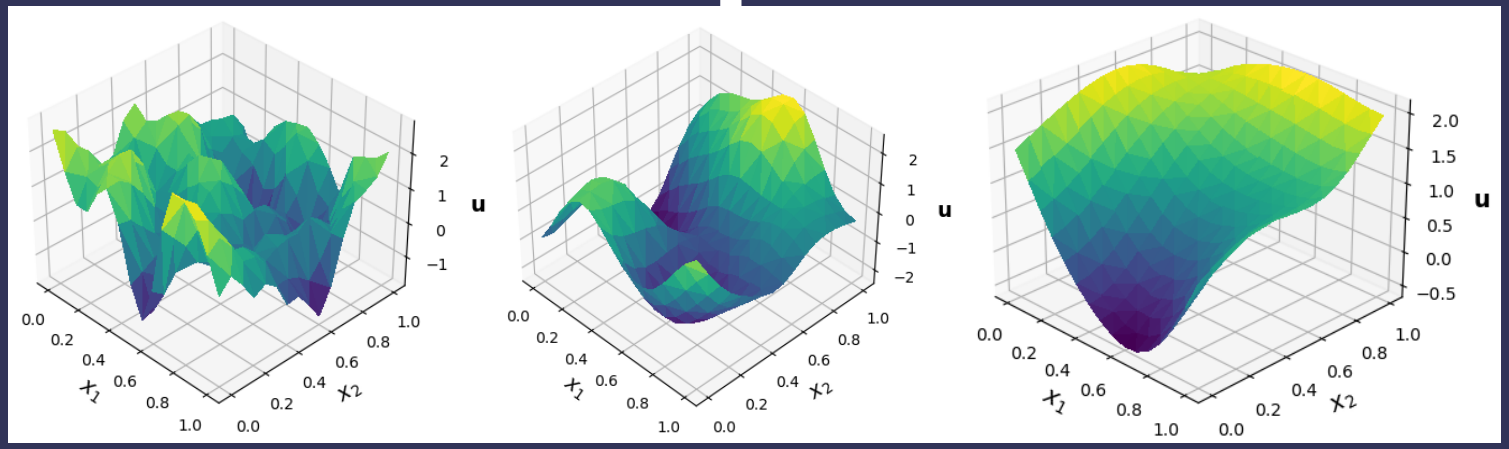
We pick the subsample locations at random

$$G = \begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots \\ \vdots & \vdots & & \vdots & & \vdots & \end{bmatrix} \qquad \epsilon \sim N(0, I)$$

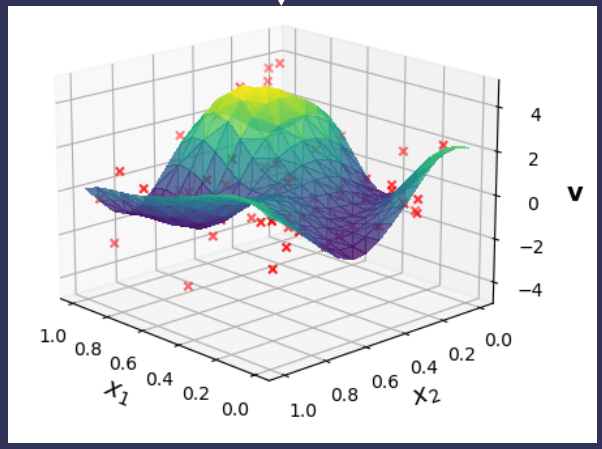
$$u(x) \sim GP(0, k(x, x'))$$

Subsample & Observe

$$v = Gu + \epsilon$$



Simulate from Gaussian Process



Generated Data

Part I : Simulation

Prior	$p(\boldsymbol{u}) = N(0, \boldsymbol{K})$
Likelihood	$p(\boldsymbol{v} \mid \boldsymbol{u}) = N(\boldsymbol{G}\boldsymbol{u}, \mathrm{I})$
Posterior	$p(\boldsymbol{u} \mid \boldsymbol{v})$

Question (b)

- Now try to infer the original (high-dimensional) field $u(\boldsymbol{x})$ that generated the (lower-dimensional) observed data
- You will sample from this posterior using:
  - Gaussian Random Walk Metropolis-Hastings (GRW-MH)
  - Preconditioned Crank-Nicolson (pCN)

MCMC Algorithms

GRM-MC $u' = u + \beta\zeta, \quad \zeta \sim N(0,K)$

pCN $u' = \sqrt{1-\beta^2}u + \beta\zeta, \quad \zeta \sim N(0,K)$

Part I : Simulation

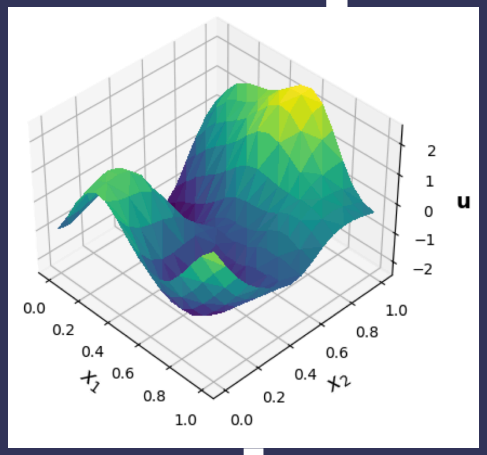
Questions (c), (d) – Probit Classification

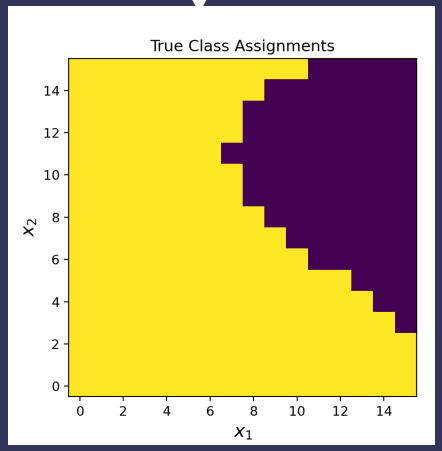
$$t_{true} = \begin{cases} 0 & u_i < 0 \\ 1 & u_i \geq 0 \end{cases}$$

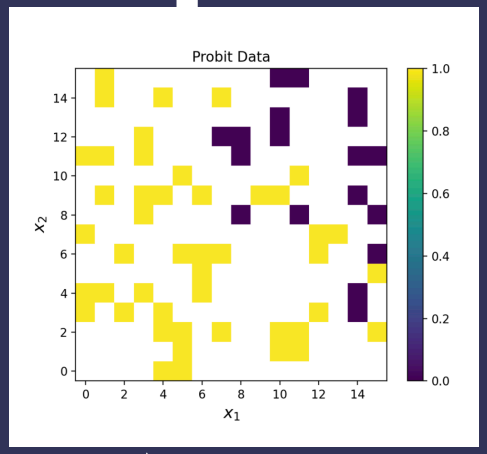
$$t_i = \begin{cases} 0 & v_i < 0 \\ 1 & v_i \geq 0 \end{cases}$$

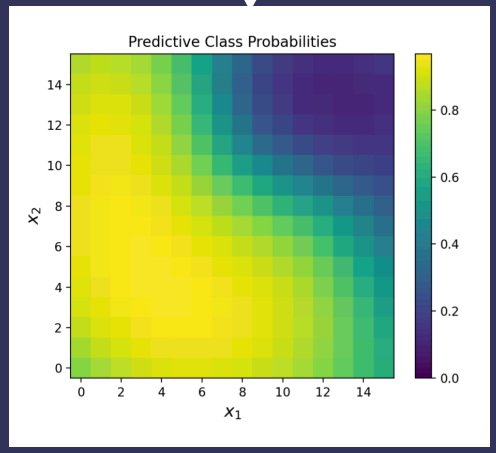
Threshold

Sample $p(\boldsymbol{u} \mid \boldsymbol{t}) \rightarrow$ find $p(t^* = 1 \mid \boldsymbol{t})$





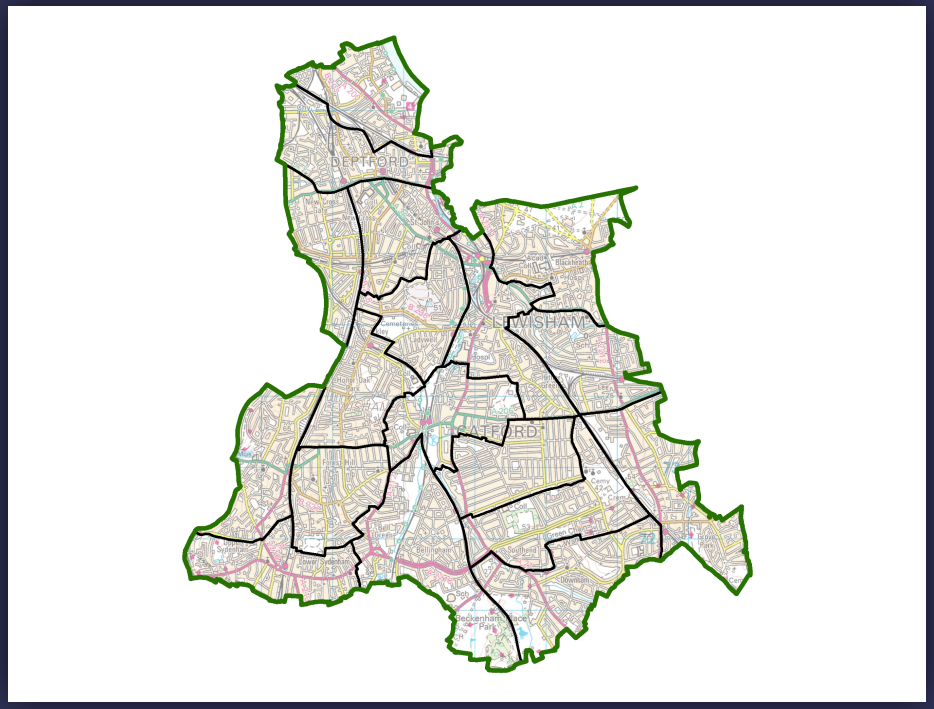




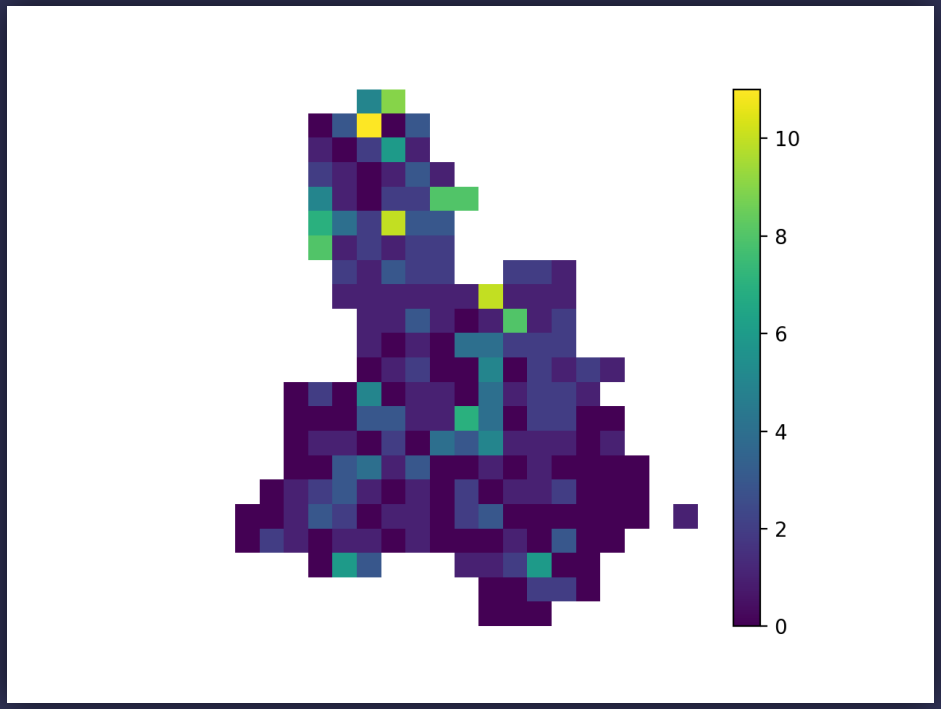
Subsample, Observe with Noise & Threshold

Part II : Spatial

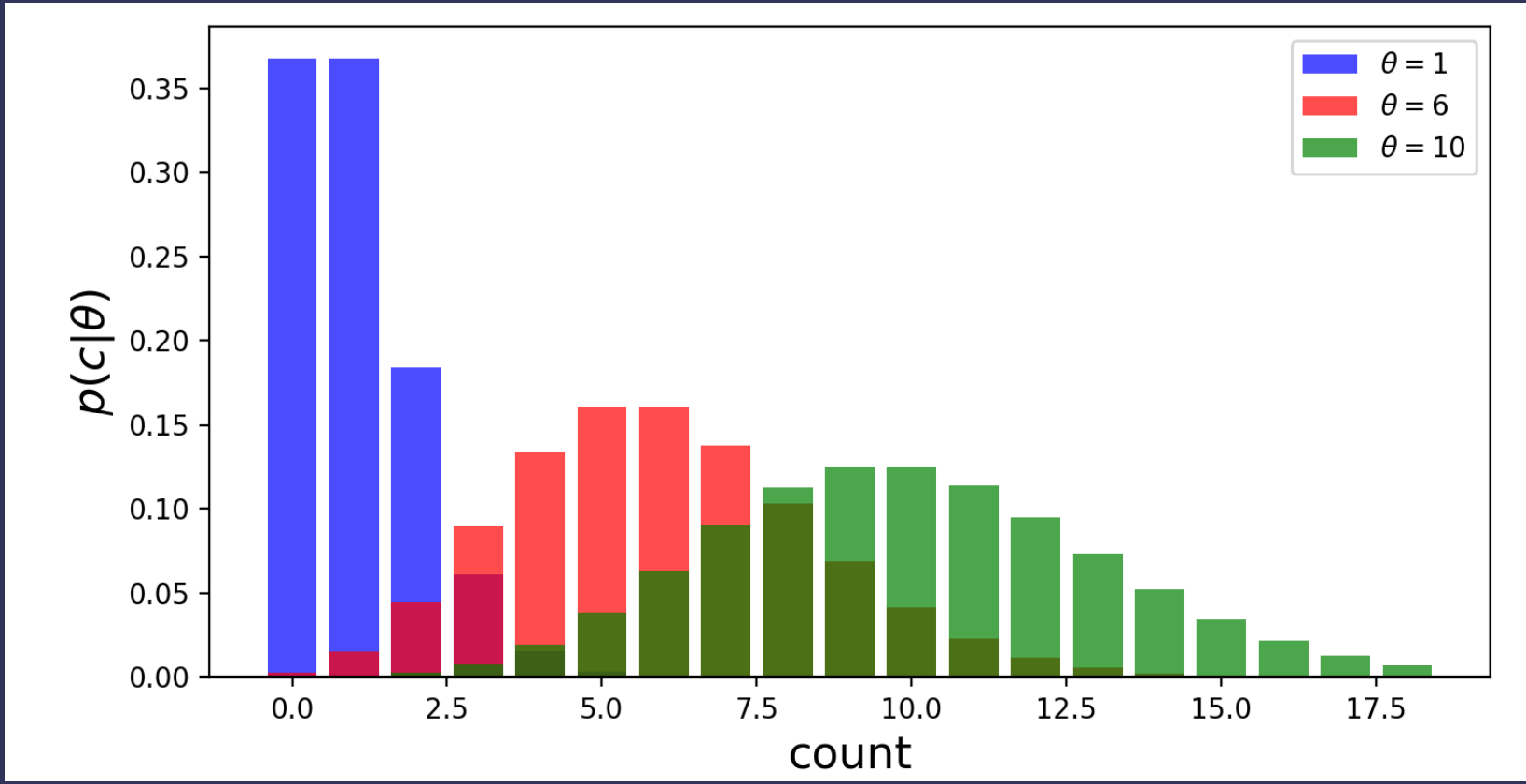
Lewisham Borough



Bike Theft Counts



Part II : Spatial



Part II : Spatial

$$G = \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots \\ \vdots & \vdots & & \vdots & & \vdots & \end{bmatrix}}_{N} \Big\} M$$

Prior $p(\boldsymbol{u}) = N(0, \boldsymbol{K})$

Mapping $\theta_i = e^{[Gu]_i}$

Likelihood $p(\boldsymbol{c}|\boldsymbol{\theta}) = \prod_{i=1}^{M} f(c_i|\theta_i)$ $\qquad f(c_i|\theta_i) = \dfrac{e^{-\theta_i}\theta_i^{c_i}}{c_i!}$

Posterior $p(\boldsymbol{u}|\boldsymbol{c})$

- Want to infer the bike theft counts, c^*, at *all* data locations, using posterior samples given subsampled data
- Transform posterior samples at location i, $\left\{u^{*(j)}\right\}_{j=1}^{n}$ to rate samples $\left\{\theta^{*(j)}\right\}_{j=1}^{n}$ $(\theta^* = e^{u^*})$
- Use rate samples at each location to infer $\mathbb{E}[c^*]$, i.e. the expected/mean counts at each location
- Compare these counts to the true values

Problems?

- Ask on Moodle discussion page
- Check Jupyter Notebooks (Lecture_11.ipynb)
- Wikipedia/Online (Cholesky decomposition, log-likelihoods , pCN etc.)
  - https://makarandtapaswi.wordpress.com/2011/07/08/cholesky-decomposition-for-matrix-inversion/
  - https://en.wikipedia.org/wiki/Poisson_distribution
  - https://en.wikipedia.org/wiki/Preconditioned_Crank-Nicolson_algorithm
- Email me: ag933@cam.ac.uk

Good Luck!