4M24 Coursework 2022/2023

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 18th January 2023**
- 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

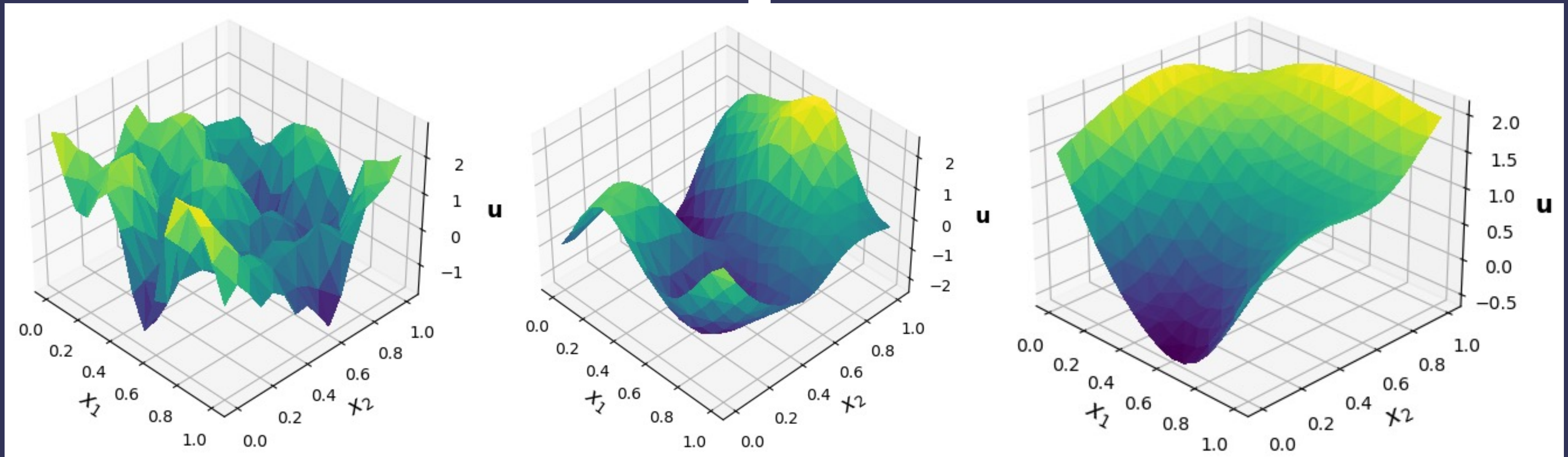
Part I : Simulation

We pick the subsample locations at random

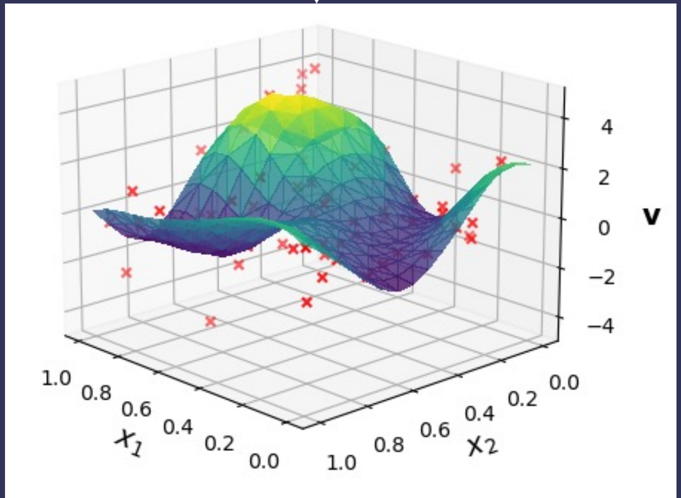
$$G = \begin{bmatrix} 1 & 0 & \dots & 0 & \dots & 0 & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 & \dots \\ \vdots & \vdots & & \vdots & & \vdots & \end{bmatrix} \qquad \epsilon \sim N(0, I)$$

Subsample & Observe $\quad v = Gu + \epsilon$

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



Simulate from Gaussian Process



Generated Data

Part I : Simulation

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

Question (b)

- Now try to infer the original (high-dimensional) field $u(\boldsymbol{x})$ that generated the (lower-dimensional) observed data
- This is done by sampling from the posterior using both Metropolis-Hastings (GRW) and preconditioned Crank-Nicholson (pCN)

Part I : Simulation

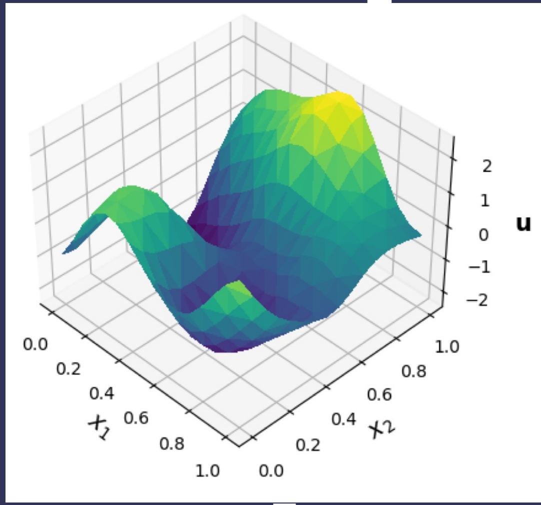
Questions (c), (d) – Probit Classification

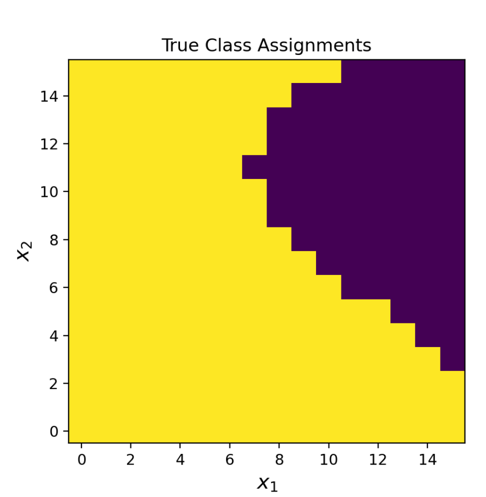
$$t_{true} = \begin{cases} 0 & u_i < 0 \\ 1 & u_i \geq 0 \end{cases}$$

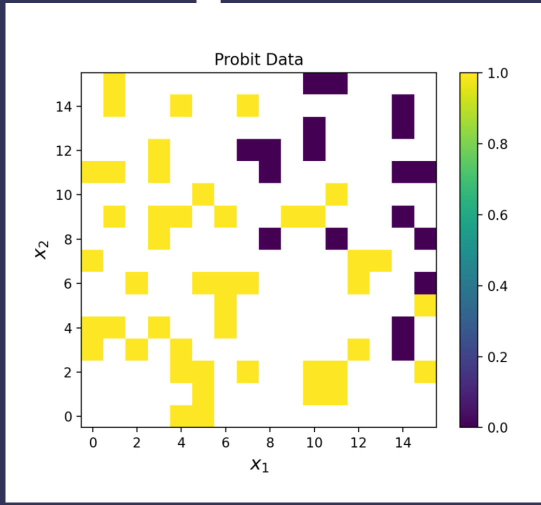
$$t_i = \begin{cases} 0 & {\color{red}v_i} < 0 \\ 1 & {\color{red}v_i} \geq 0 \end{cases}$$

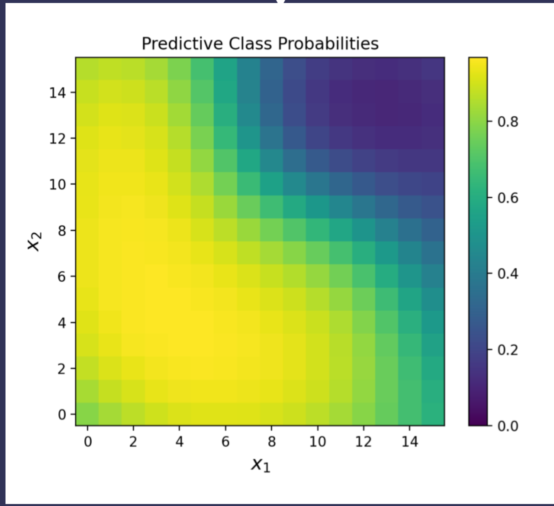
Threshold

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





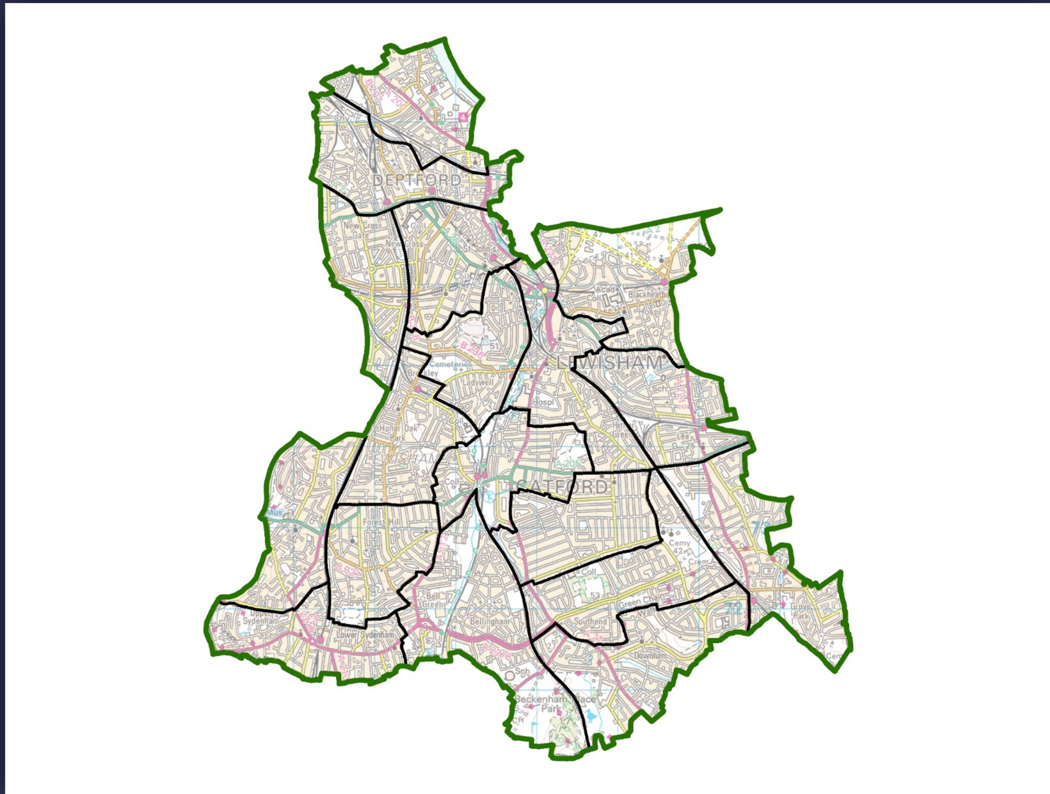




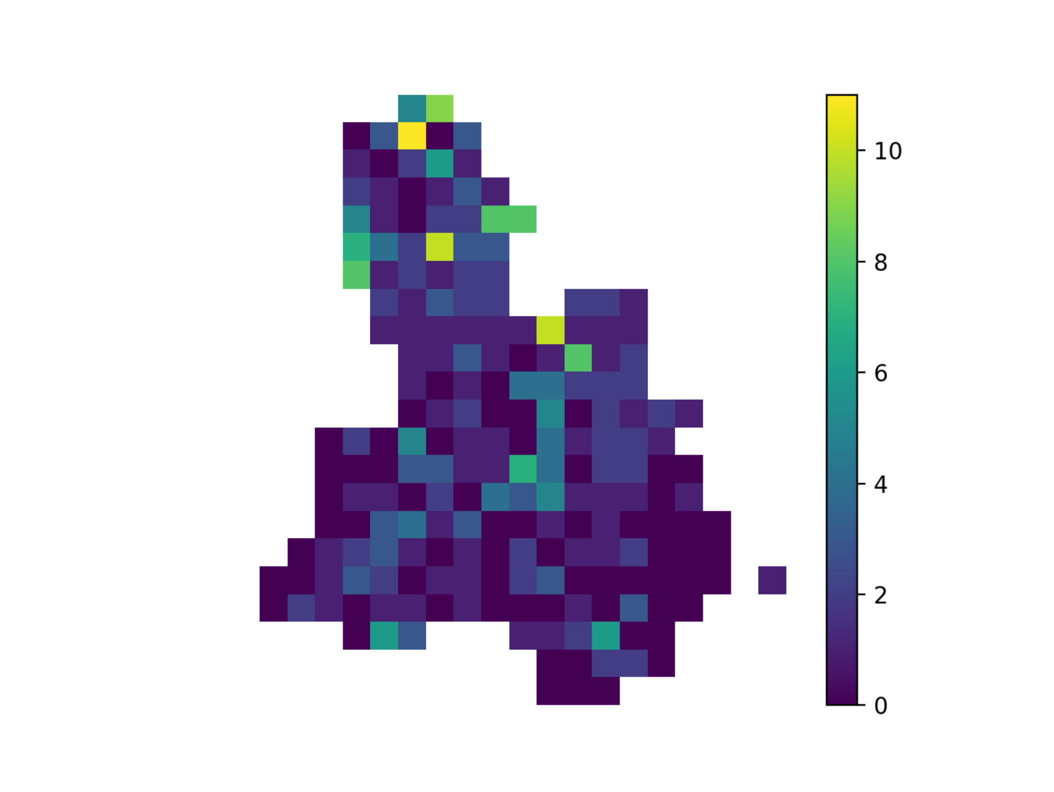
Subsample, Observe with Noise & Threshold

Part II : Spatial

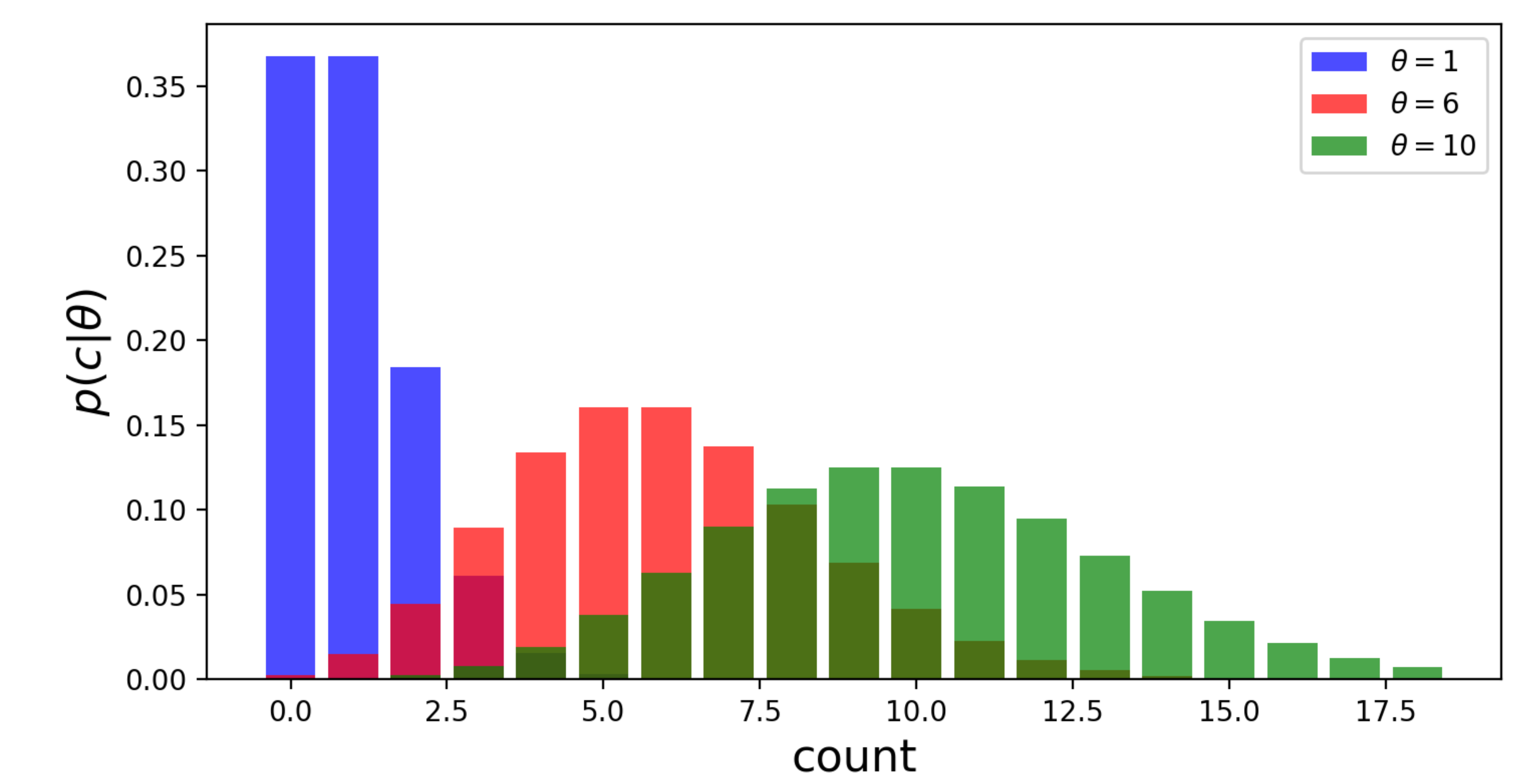
Lewisham Borough



Bike Theft Counts



Part II : Spatial



Part II : Spatial

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

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

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

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

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

$$f(c_i|\theta_i) = \frac{e^{-\theta_i}\theta_i^{c_i}}{c_i!}$$

- 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 $\mathbb{E}[\mathrm{c}^*]$, 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!