Log-Gaussian Cox Process for London crime data

Jan Povala

with

Louis Ellam
Dr Seppo Virtanen
Prof Mark Girolami

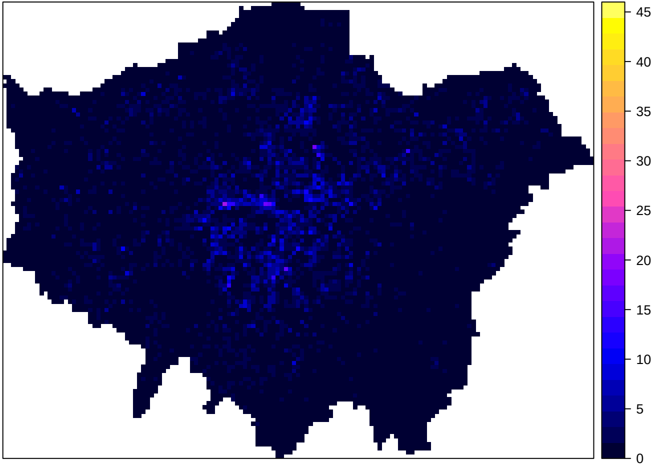
July 24, 2018

Outline

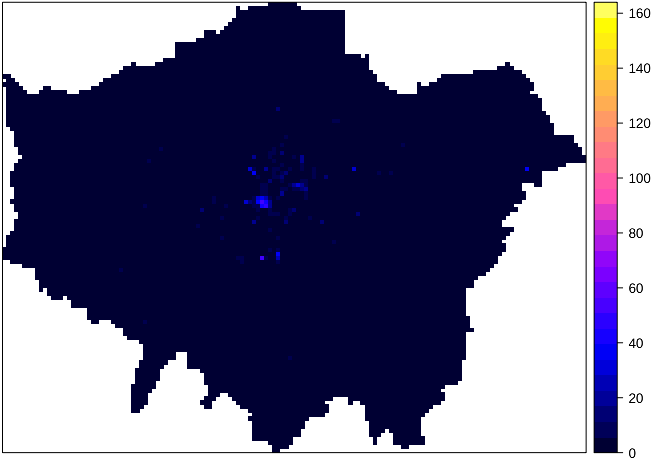
Aims and Objectives

- Modelling of crime and short-term forecasting.
- Two stages:
  1. *Inference* - what is the underlying process that generated the observations?
  2. *Prediction* - use the inferred process's properties to forecast future values.

Burglary



Theft from the person



Outline

Cox Process

Cox process is a natural choice for an environmentally driven point process (Diggle et al., 2013).

Definition

Cox process $\mathsf{Y}(\boldsymbol{x})$ is defined by two postulates:

1. $\Lambda(\boldsymbol{x})$ is a nonnegative-valued stochastic process;
2. conditional on the realisation $\lambda(\boldsymbol{x})$ of the process $\Lambda(\boldsymbol{x})$, the point process $\mathsf{Y}(\boldsymbol{x})$ is an inhomogeneous Poisson process with intensity $\lambda(\boldsymbol{x})$.

Log-Gaussian Cox Process

- Cox process with intensity driven by a fixed component $Z_{\boldsymbol{x}}^{\top}\boldsymbol{\beta}$ and a latent function $f(\boldsymbol{x})$:

$$\Lambda(\boldsymbol{x}) = \exp\left(Z_{\boldsymbol{x}}^{\top}\boldsymbol{\beta} + f(\boldsymbol{x})\right),$$

  where $f(\boldsymbol{x}) \sim \mathcal{GP}(0, k_\theta(\cdot,\cdot))$, $Z_{\boldsymbol{x}}$ are socio-economic indicators, and $\boldsymbol{\beta}$ are their coefficients.
- Discretised version of the model:

$$\mathsf{y}_i \sim \text{Poisson}\left(\exp\left[Z_{\boldsymbol{x}_i}^{\top}\boldsymbol{\beta} + f(\boldsymbol{x}_i)\right]\right).$$

Inference

We would like to infer the posterior distributions of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\mathbf{f}$:

$$p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) p(\mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta})}{p(\mathbf{y})},$$

where

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) p(\mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{\beta} d\mathbf{f},$$

which is intractable.

Solutions

1. Laplace approximation
2. **Markov Chain Monte Carlo sampling**
3. ...

Markov Chain Monte Carlo (MCMC)

- Sampling from the joint posterior distribution:

$$p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta})p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\beta}),$$

  using Hamiltonian Monte Carlo (HMC).
- Challenges:
  - $\boldsymbol{\theta}$, $\mathbf{f}$, and $\boldsymbol{\beta}$ are strongly correlated.
  - High dimensionality of $\mathbf{f}$ - every iteration requires the inverse and the determinant of $\boldsymbol{K}$.
  - Choosing the mass matrix in the HMC algorithm.

Computation

Flaxman et al. (2015), Saatçi (2012)

- The calculations above require $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ space.
- Cheaper linear algebra available if separable kernel functions are assumed, *e.g.* in $D = 2$ dimensions:

$$k((x_1, x_2), (x_1', x_2')) = k_1(x_1, x_1')k_2(x_2, x_2')$$

  implies that $\boldsymbol{K} = \boldsymbol{K}_1 \otimes \boldsymbol{K}_2$.
- Applying the above properties, the inference can be performed using $\mathcal{O}\left(Dn^{\frac{D+1}{D}}\right)$ operations and $\mathcal{O}\left(Dn^{\frac{2}{D}}\right)$ space.

Outline

Experiment

Model

- Factorisable covariance function (product of two Matérns).
- Uninformative prior for $\boldsymbol{\theta}$.
- $\mathcal{N}(\mathbf{0}, 10\boldsymbol{I})$ prior for $\boldsymbol{\beta}$.

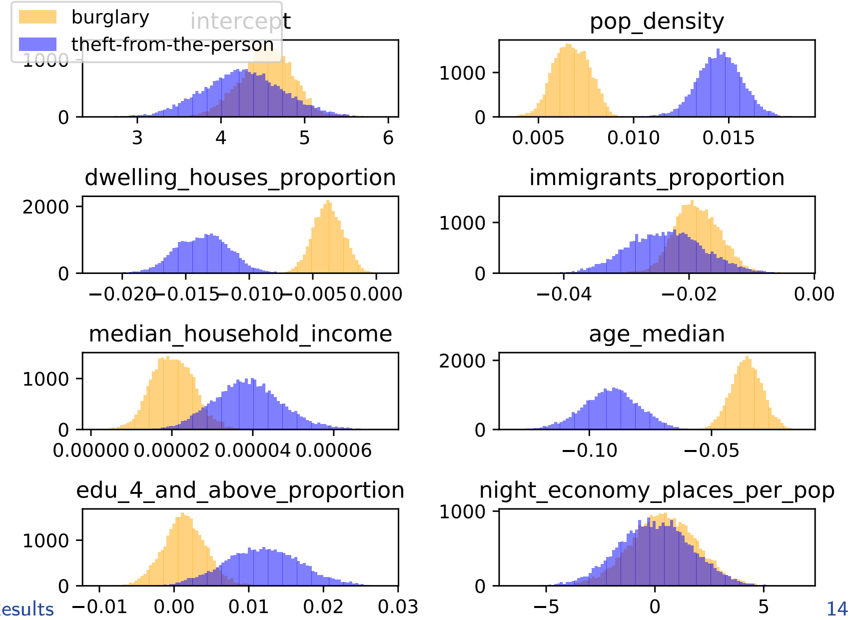
Dataset

- *Burglary*, *Theft from the person* data for 2016.
- Grid: 59x46, one cell is an area of 1km by 1km.
- Missing locations are treated with a special noise model.

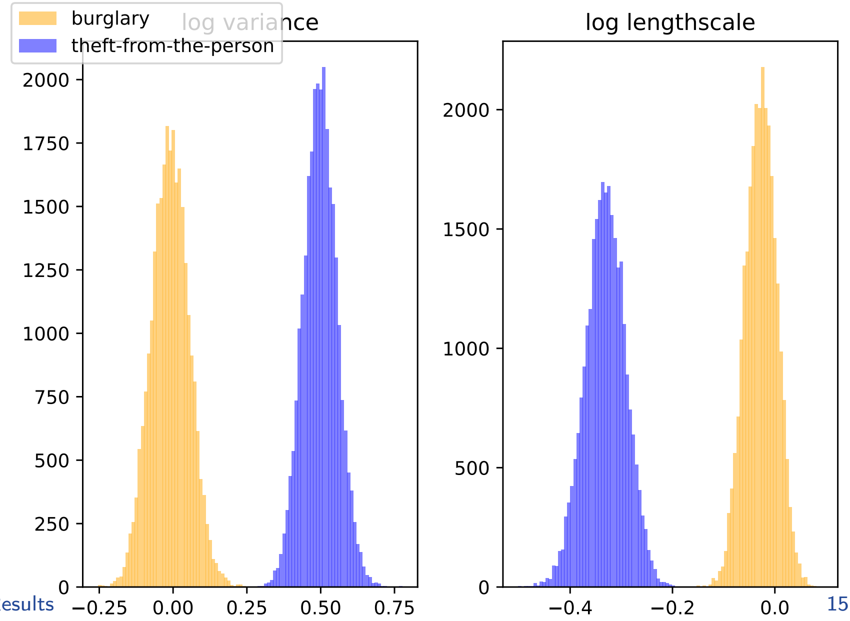
Inferred random variables

- Coefficients ($\boldsymbol{\beta}$) for various socio-economic indicators.
- Two hyperparameters $\boldsymbol{\theta}$: lengthscale(ℓ), marginal variance (σ^2).
- Latent field $\mathbf{f}$.

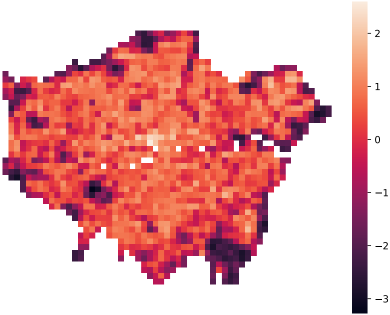
Socio-economic indicators



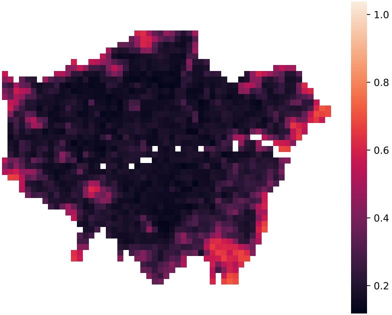
Hyperparameters



Latent field - Burglary

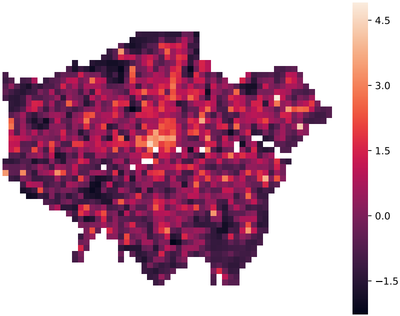


(a) mean

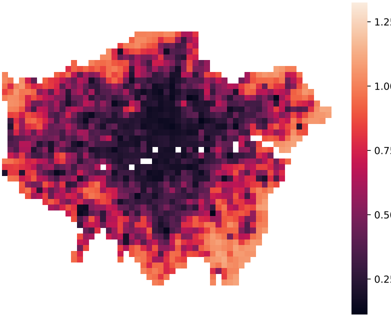


(b) standard deviation

Latent field - Theft from the person



(c) mean



(d) standard deviation

Model Fit - RMSE

We compare our model with inferences made using Poisson regression (GLM) using the root mean square error metric:

Burglary

MCMC	6.59224
GLM	30.39759

Theft from the person

MCMC	4.71420
GLM	69.61551

Discussion

- Effects missing in the GLM model are spatially correlated. This could imply two possibilities:
  - Model is missing a covariate that is spatially correlated.
  - The true process driving criminal activity is spatially correlated.
- Socio-economic indicators from the census data are 'static' and might struggle to explain more 'dynamic' crime types, e.g. *burglary* vs. *violence against person*.

Outline

Next steps

- Benchmark against INLA (Lindgren, Rue, and Lindström, 2011).
- Looking at a possibility to extend it into spatio-temporal case.

Bibliography I

Diggle, Peter J. et al. (2013). “Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm”. en. In: *Statistical Science* 28.4, pp. 542–563. ISSN: 0883-4237. DOI: 10.1214/13-STS441. URL: http://projecteuclid.org/euclid.ss/1386078878.

Flaxman, Seth et al. (2015). “Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. Vol. 37. ICML’15. Lille, France: JMLR.org, pp. 607–616.

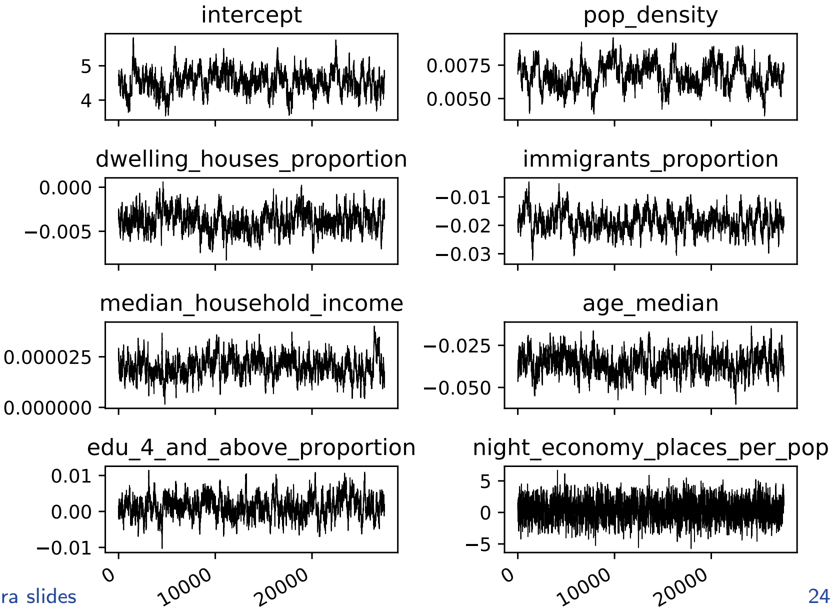
Bibliography II

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, pp. 423–498. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2011.00777.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2011.00777.x/abstract.

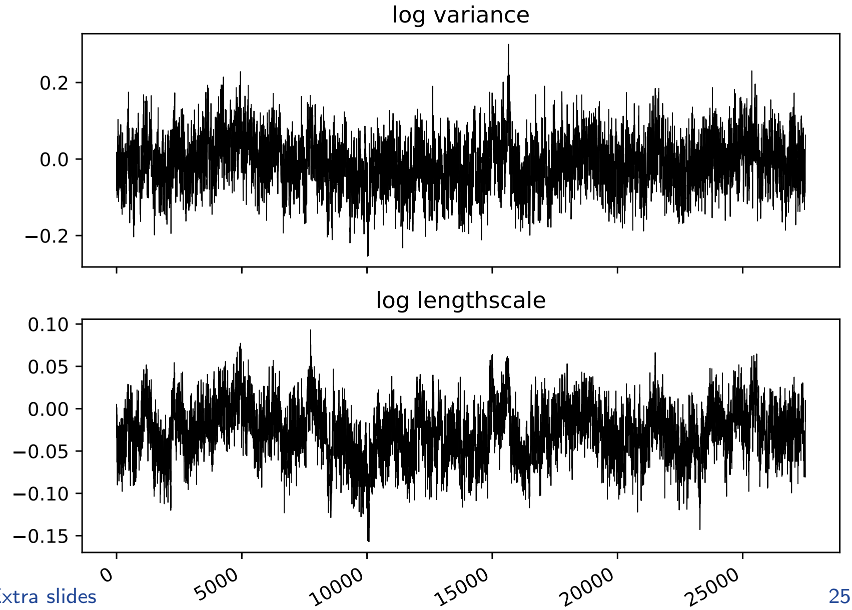
Saatçi, Yunus (2012). "Scalable inference for structured Gaussian process models". PhD Thesis. Citeseer.

Wilson, Andrew Gordon et al. (2014). "Fast Kernel Learning for Multidimensional Pattern Extrapolation". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Cambridge, MA, USA: MIT Press, pp. 3626–3634. URL: http://dl.acm.org/citation.cfm?id=2969033.2969231.

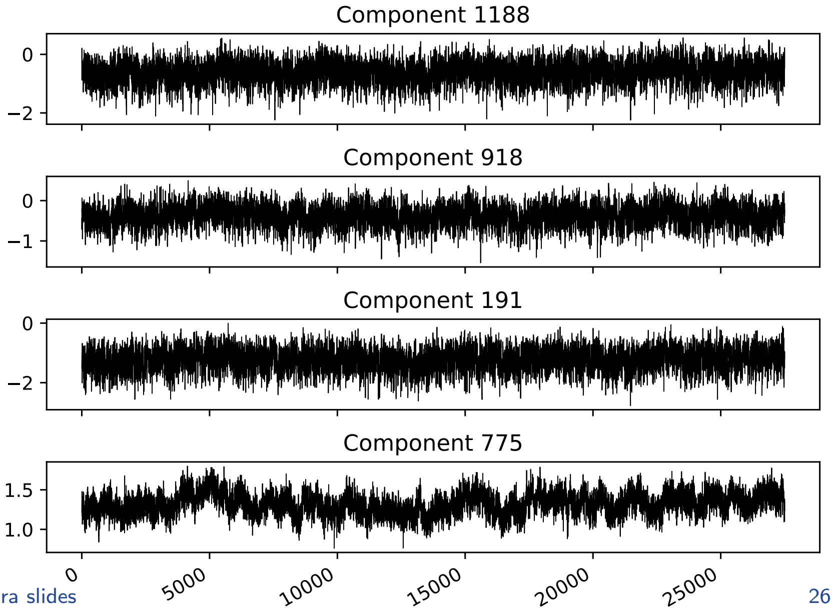
β traceplots



θ traceplots



f traceplots



Laplace Approximation

Flaxman et al. (2015)

- For simplicity, we assume non-parametric model (no fixed term), and treat $\boldsymbol{\theta}$ as a point estimate got by maximising marginal likelihood.
- Approximate the posterior distribution of the latent surface by:

$$p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \approx \mathcal{N}\left(\hat{\mathbf{f}}, -\left(\nabla\nabla\Psi(\mathbf{f})|_{\hat{\mathbf{f}}}\right)^{-1}\right),$$

  where $\Psi(\mathbf{f}) := \log p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \overset{\text{const}}{=} \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta}) + \log p(\mathbf{f}|\boldsymbol{\theta})$ is unnormalised log posterior, and $\hat{\mathbf{f}}$ is the mode of the distribution.
- Newton's method to find $\hat{\mathbf{f}}$.

Matérn Covariance Function

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{\ell} \right)$$

We fix $\nu = 2.5$ as it is difficult to jointly estimate ℓ and ν due to identifiability issues.

Kronecker Algebra

Saatçi (2012)

- Matrix-vector multiplication $(\otimes_d \mathbf{A}_d)\,\boldsymbol{b}$ in $\mathcal{O}(n)$ time and space.
- Matrix inverse: $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$
- Let $\boldsymbol{K}_d = \boldsymbol{Q}_d \boldsymbol{\Lambda}_d \boldsymbol{Q}_d^\top$ be the eigendecomposition of $\boldsymbol{K}_d$. Then, the eigendecomposition of $\boldsymbol{K} = \otimes_d \boldsymbol{K}_d$ is given by $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} = \otimes_d \boldsymbol{Q}_d$, and $\boldsymbol{\Lambda} = \otimes_d \boldsymbol{\Lambda}_d$. The number of steps required is $\mathcal{O}\left(Dn^{\frac{3}{D}}\right)$.

Incomplete grids

Wilson et al. (2014)

We have that $\mathsf{y}_i \sim \text{Poisson}(\exp(\mathsf{f}_i))$. For the points of the grid that are not in the domain, we let $\mathsf{y}_i \sim \mathcal{N}(\mathsf{f}_i, \epsilon^{-1})$ and $\epsilon \to 0$. Hence,

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i \in \mathcal{D}} \frac{\left(e^{\mathsf{f}_i}\right)^{\mathsf{y}_i} e^{-e^{\mathbf{f}_i}}}{\mathsf{y}_i!} \prod_{i \notin \mathcal{D}} \frac{1}{\sqrt{2\pi\epsilon^{-1}}} e^{\frac{-\epsilon(\mathbf{y}_i - \mathbf{f}_i)^2}{2}}$$

The log-likelihood is thus:

$$\sum_{i \in \mathcal{D}} [\mathsf{y}_i \mathsf{f}_i - \exp(f_i) + \text{const}] - \frac{1}{2} \sum_{i \notin \mathcal{D}} \epsilon(\mathsf{y}_i - \mathsf{f}_i)^2$$

We now take the gradient of the log-likelihood as

$$\nabla \log p(\mathbf{y}|\mathbf{f})_i = \begin{cases} \mathsf{y}_i - \exp(\mathsf{f}_i), & \text{if } i \in \mathcal{D} \\ \epsilon(\mathsf{y}_i - \mathsf{f}_i), & \text{if } i \notin \mathcal{D} \end{cases}$$

and the hessian of the log-likelihood as

$$\nabla\nabla \log p(\mathbf{y}|\mathbf{f})_{ii} = \begin{cases} -\exp(\mathsf{f}_i), & \text{if } i \in \mathcal{D} \\ -\epsilon & \text{if } i \notin \mathcal{D} \end{cases}.$$