Log-Gaussian Cox Process for London crime data

Jan Povala

May 3, 2018

Outline

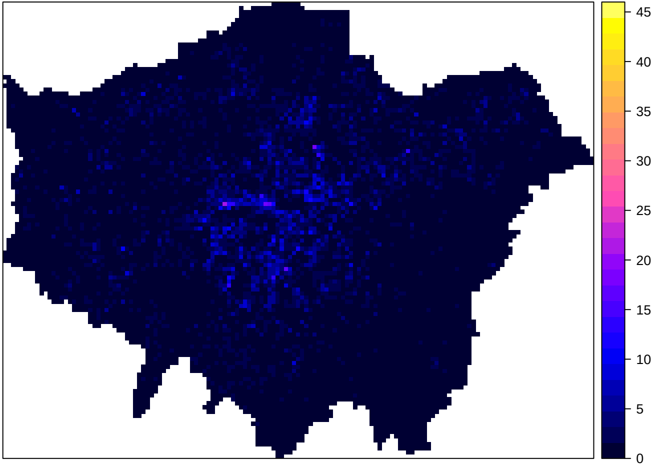
Motivation

Methodology

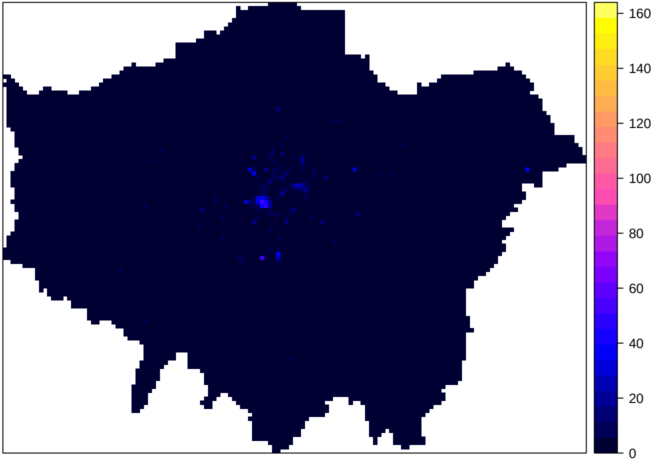
Results

Next steps

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 is defined by two postulates:

1. $\Lambda = \left\{\Lambda(\boldsymbol{x}) : \boldsymbol{x} \in \mathbf{R}^2\right\}$ is a nonnegative-valued stochastic process;
2. conditional on the realisation $\Lambda(\boldsymbol{x}) = \lambda(\boldsymbol{x}) : \boldsymbol{x} \in \mathbf{R}^2$, the point process is an inhomogeneous Poisson process with intensity $\lambda(x)$.

Log-Gaussian Cox Process

- Cox process with intensity driven by a Gaussian Process $f(\boldsymbol{x})$:

$$\Lambda(\boldsymbol{x}) = \exp\left(f(\boldsymbol{x})\right).$$

- Tractability of multivariate Normal distribution carries over to the associated Cox process.
- A common approach for analysis is to introduce a grid over the domain X.

Field inference - Laplace Approximation

Flaxman et al. (2015)

- Approximate the posterior distribution of the Gaussian Process by:

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

  where $\Psi(\mathbf{f}) := \log p(\mathbf{f}|\mathbf{y}, X) \overset{\text{const}}{=} \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|X)$ is unormalised log posterior.
- Newton's method to find $\hat{\mathbf{f}}$.

Field inference - Newton Optimisation

Flaxman et al. (2015)

- The Newton optimisation step:

$$\mathbf{f}^{\text{new}} \leftarrow \mathbf{f}^{\text{old}} - (\nabla\nabla\Psi)^{-1}\,\nabla\Psi$$

- $\nabla\nabla\Psi$ and $\nabla\Psi$ require inverting the covariance matrix of the GP:

$$\begin{aligned}\nabla\Psi(\mathbf{f}) =& \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f}\\ \nabla\nabla\Psi(\mathbf{f}) =& -\boldsymbol{W} - \boldsymbol{K}^{-1},\end{aligned}$$

where $W := -\nabla\nabla \log p(\mathbf{y}|\mathbf{f})$.

Hyperparameters - Marginal Likelihood

Flaxman et al. (2015)

- Not the whole story: $p(\mathbf{f}|X)$ should be $p(\mathbf{f}|X, \boldsymbol{\theta})$, where $\boldsymbol{K}$ depends on $\boldsymbol{\theta}$.
- Marginal log-likelihood:

$$\begin{aligned}\log p(\mathbf{y}|X, \boldsymbol{\theta}) &= \log \int \exp\left[\Psi(\mathbf{f})\right] d\mathbf{f} \\ &\approx \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|\end{aligned}$$

Computation

Flaxman et al. (2015)

The computatios require matrix inverse, matrix determinant, and matrix-vector multiplications:

- Conjugate gradient for inverting K, exploiting Kronecker structure
- Determinant approximation due to Fiedler (1971):

$$\begin{aligned}\log |\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}| &= \log \left(|\boldsymbol{K} + \boldsymbol{W}^{-1}||\boldsymbol{W}| \right) \\ &\leq \log \left\{ \prod_i \left(e_i + W_{ii}^{-1} \right) \prod_i W_{ii} \right\} \\ &= \sum_i \log \left(1 + e_i W_{ii} \right),\end{aligned}$$

  where $e_1, \ldots, e_n$ are sorted eigenvalues of $\boldsymbol{K}$.
- Matrix-vector multiplication are efficient due to Kronecker structure.

Separable kernels, Kronecker methods

Flaxman et al. (2015)

- Separable kernel functions:

$$k((x_1, y_1), (x_2, y_2)) = k_1(x_1, x_2)k_2(y_1, y_2)$$

- On a regular grid, we get:

$$\boldsymbol{K} = \boldsymbol{K}_1 \otimes \boldsymbol{K}_2$$

Outline

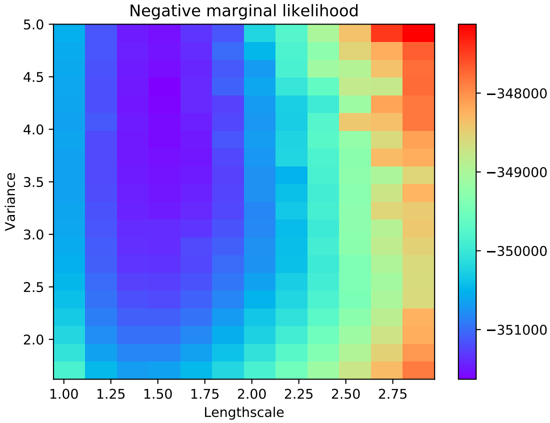
Experiment

Spatial model with isotropic Matérn covariance function:

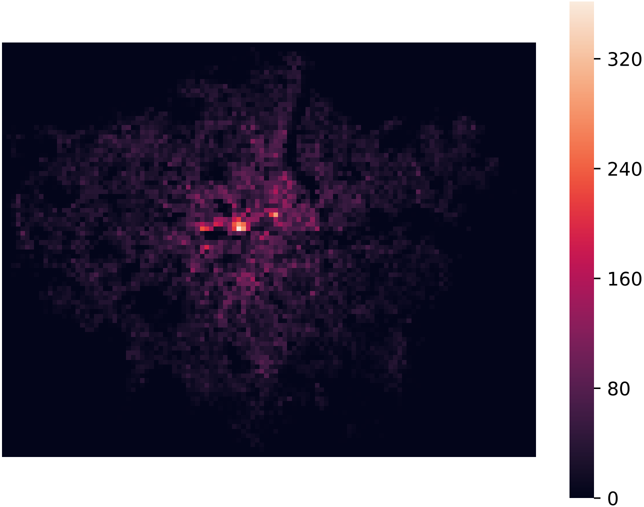
- Dataset used: 2016 data
- Crime types: Burglary, Theft from the person
- Grid: 117x91, one cell is an area of 500m by 500m.
- Two parameters inferred: lengthscale(ℓ), marginal variance (σ^2)

Burglary - inferred parameters

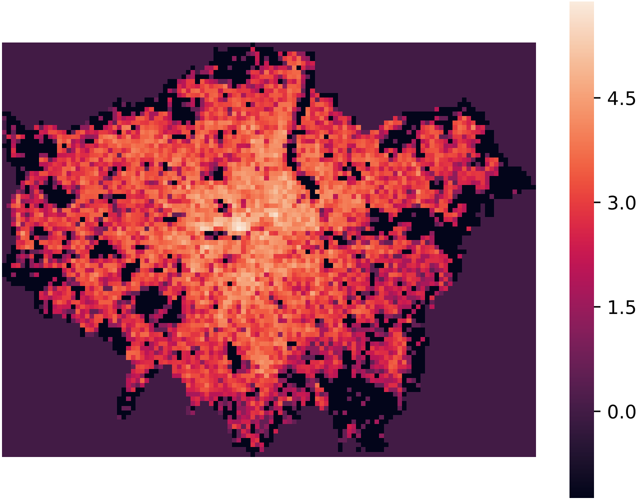
Inferred parameters: $\ell = 1.45$, and $\sigma^2 = 4.32$



Burglary - counts

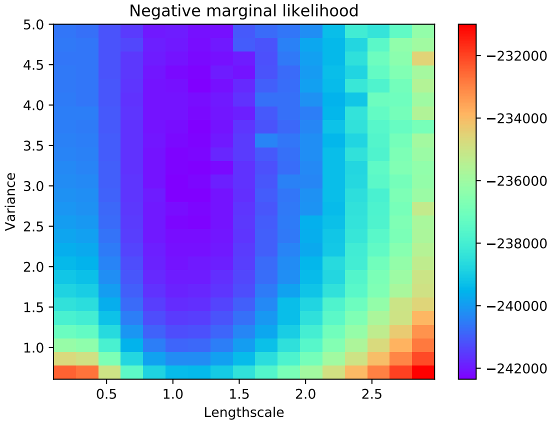


Burglary - latent field

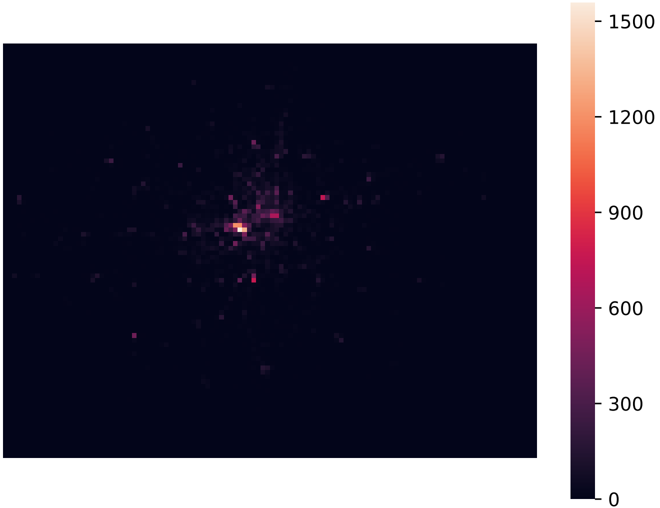


Theft from the person - inferred parameters

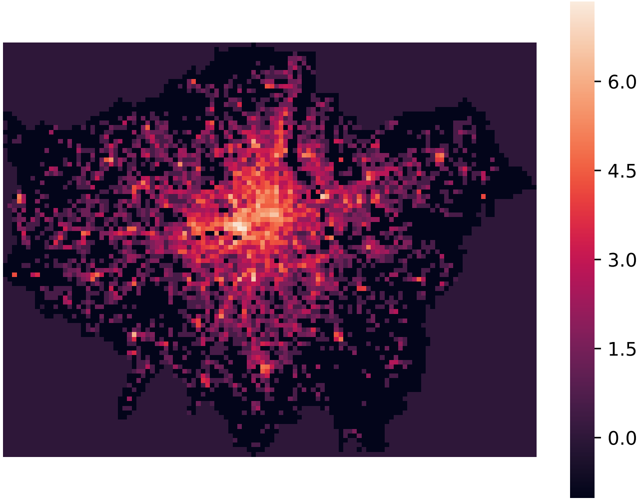
Inferred parameters: $\ell = 1.11$, and $\sigma^2 = 2.80$



Theft from the person - counts



Theft from the person - latent field



Outline

Time component, and predictions

Possible options are:

- A kernel with period of 12 months for seasonal variation (Flaxman, 2014):

$$k_P\left(t, t'\right) = \exp\left(-\frac{2\sin^2\left(\frac{(t-t')\pi}{12}\right)}{\ell^2}\right)$$

- Spectral mixture kernel with Q components (Flaxman et al., 2015):

$$k(\tau) = \sum_{q=1}^{Q} w_q \exp\left(-2\pi^2\tau^2 v_q\right)\cos\left(2\pi\tau\mu_q\right)$$

Stochastic PDEs

- Finite Element Method to solve SPDEs as described in Lindgren, Rue, and Lindström (2011).
- Sigrist, Künsch, and Stahel (2015) solve transport-diffusion SPDE using spectral methods on a grid.

More on this from Seppo.

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.

Fiedler, Miroslav (1971). "Bounds for the Determinant of the Sum of Hermitian Matrices". In: *Proceedings of the American Mathematical Society* 30.1, p. 27. ISSN: 00029939. DOI: 10.2307/2038212. URL: http://www.jstor.org/stable/2038212?origin=crossref.

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

Flaxman, Seth R. (2014). *A General Approach to Prediction and Forecasting Crime Rates with Gaussian Processes*. Tech. rep. Heinz College Technical Report, 2014. URL https://www. ml. cmu. edu/research/dap-papers/dap_flaxman. pdf.

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.

Bibliography III

Sigrist, Fabio, Hans R. Künsch, and Werner A. Stahel (2015). "Stochastic partial differential equation based modelling of large space-time data sets". en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77.1, pp. 3–33. ISSN: 13697412. DOI: 10.1111/rssb.12061. URL: http://doi.wiley.com/10.1111/rssb.12061.

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.

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 use $\nu = 2.5$.

Inference Newton step details

Kronecker Algebra

Incomplete grids

Wilson et al. (2014)

We have that $\mathsf{y}_i \sim \mathrm{Poisson}(\mathsf{f}_i)$. 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:

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

We now take the gradient of the $\log$ of the 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

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