# Multi-task learning of point processes

Virginia Aglietti
*Supervisors*: Dr Theo Damoulas & Professor David Firth
*Collaborator*: Dr Edwin Bonilla (UNSW)

University of Warwick

30th April 2018

# Outline


Research overview

Log Gaussian Cox Process Networks
- Related work
- The LGCPN model
- Scalable Variational Inference
- Experiments
- Future directions

Long term PhD directions

References

# Research overview

**Focus:** Bayesian multi-task learning of point processes.

**Motivation:** Many social processes (e.g. crime, traffic accidents, diseases etc.) can be seen as point processes characterized by:

- Cross-correlation;
- Spatial and temporal correlation;
- Correlation structure that changes in time and space (non-stationary) in a dependent manner (non-separable);

**Goal:** Exploit the current data rich environment to develop scalable algorithms capable of **jointly modelling** social processes taking into account their complexities.

# Research overview

**Focus:** Bayesian multi-task learning of point processes.

**Motivation:** Many social processes (e.g. crime, traffic accidents, diseases etc.) can be seen as point processes characterized by:

- Cross-correlation;
- Spatial and temporal correlation;
- Correlation structure that changes in time and space (non-stationary) in a dependent manner (non-separable);

**Goal:** Exploit the current data rich environment to develop scalable algorithms capable of **jointly modelling** social processes taking into account their complexities.

LGCPN

# The LGCP model

The Log-Gaussian Cox Process (LGCP) [Møller et al., 1998] is an **inhomogeneous Poisson process** with a **stochastic intensity** function:

$$y_A|\lambda(\mathbf{x}) \sim \text{Poisson}\left(\int_{\mathbf{x}\in A} \lambda(\mathbf{x})d\mathbf{x}\right),$$

where:

$$\lambda(\mathbf{x}) = \exp\{f(\mathbf{x})\}, \quad f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$$

The **multivariate** LGCP (MLGCP) [Diggle et al., 2013] considers $P$ types of points with an intensity given by:

$$\lambda_p(\mathbf{x}) = \exp(\beta + f_0(\mathbf{x}) + f_p(\mathbf{x}))$$

where $f_0(\mathbf{x}_n)$ indicates a GP common to all tasks while $f_p(\mathbf{x}_n)$ denotes a GP specific to the point process of type $p$. Inference proceeds via a MALA algorithm.

# The LMC/ICM model

The linear model of coregionalisation (LCM) is an established technique in geostatistics. The outputs are expressed as linear combinations of $Q$ independent Gaussian processes:

$$y_p(\mathbf{x}) = f_p(\mathbf{x}), \quad f_p(\mathbf{x}) = \sum_{q=1}^{Q} w_{pq} u_q(\mathbf{x}).$$

The covariance $\mathrm{Cov}[f_p(\mathbf{x}), f_{p'}(\mathbf{x}')]$ is given by $K(\mathbf{x}, \mathbf{x}') = \sum_{q=1}^{Q} B_q k_q(\mathbf{x}, \mathbf{x}')$ where $B_q$ is known as a coregionalisation matrix. A Bayesian treatment has been proposed by Schmidt & Gelfand (2003). The weight parameters have an inverse Wishart prior and inference proceeds with MCMC.

# Advantages of LGCPN

**Drawbacks of existing work:**

- Computational cost and mixing properties of the MCMC chain;
- Limited flexibility.

We attempt to adress some of the drawbacks with LGCPN. LGCPN has the following advantages:

- **Fully Bayesian** treatment allowing uncertainty quantification;
- **Scalable inference** framework solving convergence and mixing issues;
- **Inferred** mixing weights;
- **Flexible** model in capturing cross-correlations.

# The LGCPN model

**Latent functions:**

- $Q$ **uncorrelated** $\mathcal{GP}$:

$$p(\mathbf{F}|\theta) = \prod_{q=1}^{Q} p(\mathbf{F}_{\bullet q}|\theta_q)$$
$$= \prod_{q=1}^{Q} \mathcal{N}(\mathbf{F}_{\bullet q}|\mathbf{0}, \mathbf{K}_{xx}^{q}),$$

where $\theta_q$ are the hyper-parameters for the $q$-th latent function.

**Mixing weights:**

- **Independent weights** across tasks and latent functions:

$$p(\mathbf{W}) = \prod_{p=1}^{P} \prod_{q=1}^{Q} \mathcal{N}(w_{pq}; 0, \sigma_{pq}^{2}$$

- **Coupled weights**, independent across latent processes:

$$p(\mathbf{W}|\theta_w) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{W}_{\bullet q}; \mathbf{0}, \mathbf{K}_{w}^{q}$$

where $\theta_w$ denotes the hyper-parameters.

# The LGCPN model

**Likelihood function:** $P$ LGCP with a log intensity given by the linear combination of the latent functions and the task-specific offset $\phi_p$ :

$$p(\mathbf{Y}|\mathbf{F},\mathbf{W}) = \prod_{n=1}^{N}\prod_{p=1}^{P} \text{Poisson}\left(y_{np}; \exp\left(\sum_{q=1}^{Q} w_{pq} f_q(x_n) + \phi_p\right)\right)$$

**Computational grid:** We introduce an approximation by considering a computational grid on the spatial extend and representing each grid cell with its centroid. We approximate the integral in the LGCP likelihood function by replacing the intensity's infinite dimensional distribution with a finite approximation.

# Plate diagram for LGCPN

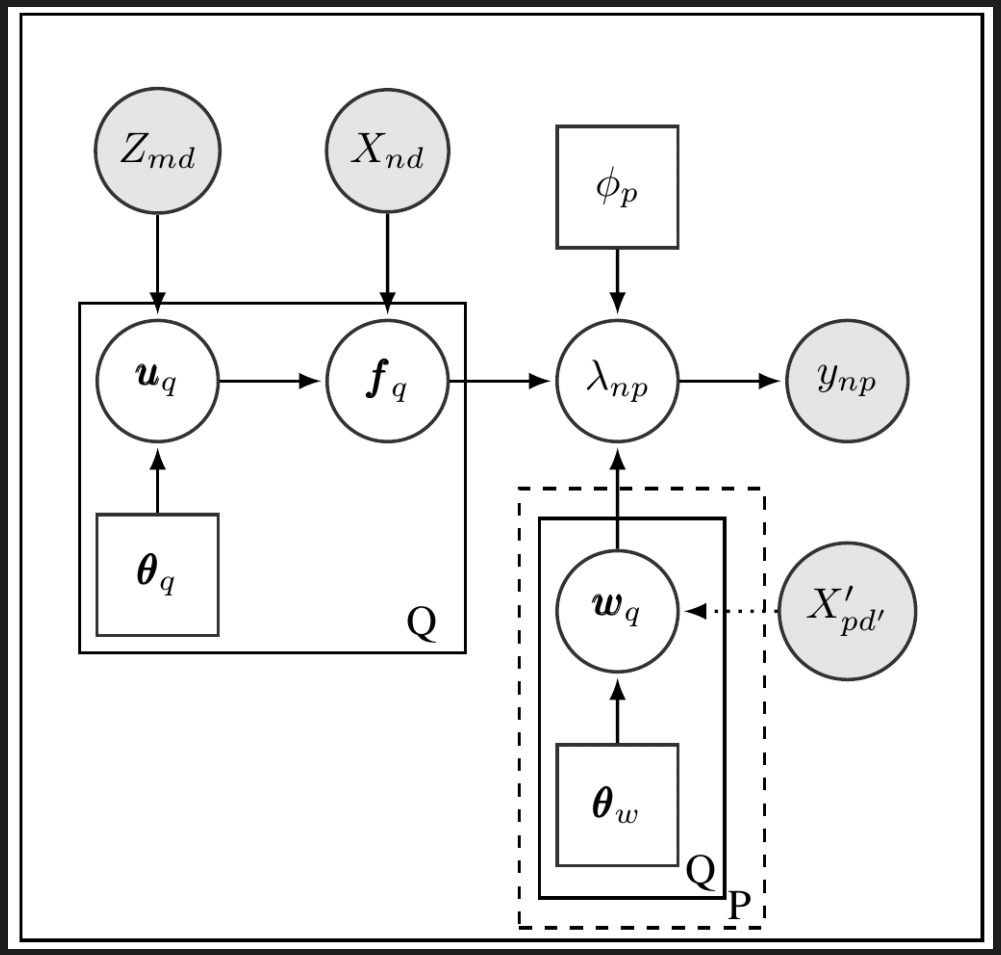


Figure 1: LGCPN- Plate diagram. $X'_{pd'}$ represents the inputs for the GP prior on $\mathbf{W}$. When placing a Normal prior on each $w_{pq}$, we introduce the additional factorization across $P$ (dashed plate).

# Scalable Variational Inference

We introduce an **augmented prior over the latent functions** [Titsias, 2009 and Bonilla et al., 2016] defined by:

$$p(\mathbf{U}|\theta) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{0}, \mathbf{K}_{zz}^{q}) \quad \text{and} \quad p(\mathbf{F}|\mathbf{U}, \theta) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{F}_{\bullet q}; \tilde{\mu}_q, \widetilde{\mathbf{K}}_q),$$

where $\tilde{\mu}_q = \mathbf{K}_{xz}^{q}(\mathbf{K}_{zz}^{q})^{-1}\mathbf{U}_{\bullet q}$ and $\widetilde{\mathbf{K}}_q = \mathbf{K}_{xx}^{q} - \mathbf{K}_{xz}^{q}(\mathbf{K}_{zz}^{q})^{-1}\mathbf{K}_{zx}^{q}$.

$\mathbf{U}_{\bullet q}$ denotes the inducing process for $\mathbf{F}_{\bullet q}$ computed in the $M \times D$ matrix $\mathbf{Z}_q$ of *inducing inputs* ($M \ll N$).

# Scalable Variational Inference

$$\underbrace{p(\mathbf{F}, \mathbf{U}, \mathbf{W}|\mathcal{D})}_{\text{Posterior distribution}} \rightarrow \underbrace{q(\mathbf{F}, \mathbf{U}, \mathbf{W}|\nu)}_{\text{Variational distribution}} = p(\mathbf{F}|\mathbf{U}) \underbrace{q(\mathbf{U}|\nu_u)}_{(1)} \underbrace{q(\mathbf{W}|\nu_w)}_{(2)}$$

(1) $q(\mathbf{U}|\nu_u) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{m}_q, \mathbf{S}_q)$

(2) $q(\mathbf{W}|\nu_w) = \prod_{q=1}^{Q} \mathcal{N}(\mathbf{W}_{\bullet q}; \omega_q, \Omega_q)$

where $\nu_u = \{\mathbf{m}_q, \mathbf{S}_q\}$ are the variational parameters.

**Minimization of the KL divergence** $\rightarrow$ Maximisation of the evidence lower bound $(\mathrm{ELBO})$:

$$\mathcal{L}_{\text{elbo}}(\nu) = \mathcal{L}_{\text{kl}}(\nu) + \mathcal{L}_{\text{ell}}(\nu), \tag{1}$$

$$\mathcal{L}_{\text{kl}}(\nu) = -\mathrm{KL}(q(\mathbf{F}, \mathbf{U}, \mathbf{W}|\nu) \| p(\mathbf{F}, \mathbf{U}, \mathbf{W})), \tag{2}$$

$$\mathcal{L}_{\text{ell}}(\nu) = \mathbb{E}_{q(\mathbf{F}, \mathbf{U}, \mathbf{W}|\nu)}[\log p(\mathbf{Y}|\mathbf{F}, \mathbf{W})]. \tag{3}$$

# Computational complexity

**The time complexity of the algorithm is of order $\mathcal{O}(M^3)$.**

- Eq. 2 can be written as $-\mathrm{KL}(q(\mathbf{U}|\boldsymbol{\nu}_u)\|p(\mathbf{U})) - \mathrm{KL}(q(\mathbf{W}|\boldsymbol{\nu}_w)\|p(\mathbf{W}))$, thus including distributions over M-dimensional variables and P-dimensional variables (M $\gg$ P). The computational complexity of the KL term is thus independent of N.

- Eq. 3 decomposes as a sum of expectations over N:

$$\mathcal{L}_{\text{ell}}(\boldsymbol{\nu}) = \sum_{n=1}^{N}\sum_{p=1}^{P} \mathbb{E}_{q(\mathbf{F}_{n\bullet})q(\mathbf{W}_{p\bullet})}[\log p(y_{np}|\mathbf{F}_{n\bullet}, \mathbf{W}_{p\bullet}, \phi_p)].$$

  This enables the use of stochastic optimization techniques thus also making this term independent of N.

# Predictive intensity

This framework allows to have a closed form solution for all the **moments of the predictive intensity** by exploiting the MGF of the product of two normal random variables. Given $X \sim \mathcal{N}(\mu_X, \mu_X)$ and $Y \sim \mathcal{N}(\mu_Y, \mu_Y)$ , then $Z = XY$ has $\mathrm{MGF}_Z(t)$ defined as:

$$\mathrm{MGF}_Z(t) = \frac{\exp\left[\frac{t\mu_X\mu_Y + 1/2(\mu_Y^2\sigma_X^2 + \mu_X^2\sigma_Y^2)t^2}{1 - t^2\sigma_X^2\sigma_Y^2}\right]}{\sqrt{1 - t^2\sigma_X^2\sigma_Y^2}} \tag{4}$$

The moment generating function for $V = \sum_{q=1}^{Q} w_{pq} f_q(\mathbf{x})$ is the product of $Q$ moment generating functions of the form given in Eq. 5. Thus the $n$-th moment for $\lambda(\mathbf{x}^*)$:

$$\mathbb{E}\left[\left[\exp\left(\sum_{q=1}^{Q} w_{pq} f_q(\mathbf{x}^*)\right)\right]^n\right] = \mathrm{MGF}_V(n)$$

# Transfer experiment - Synthetic data

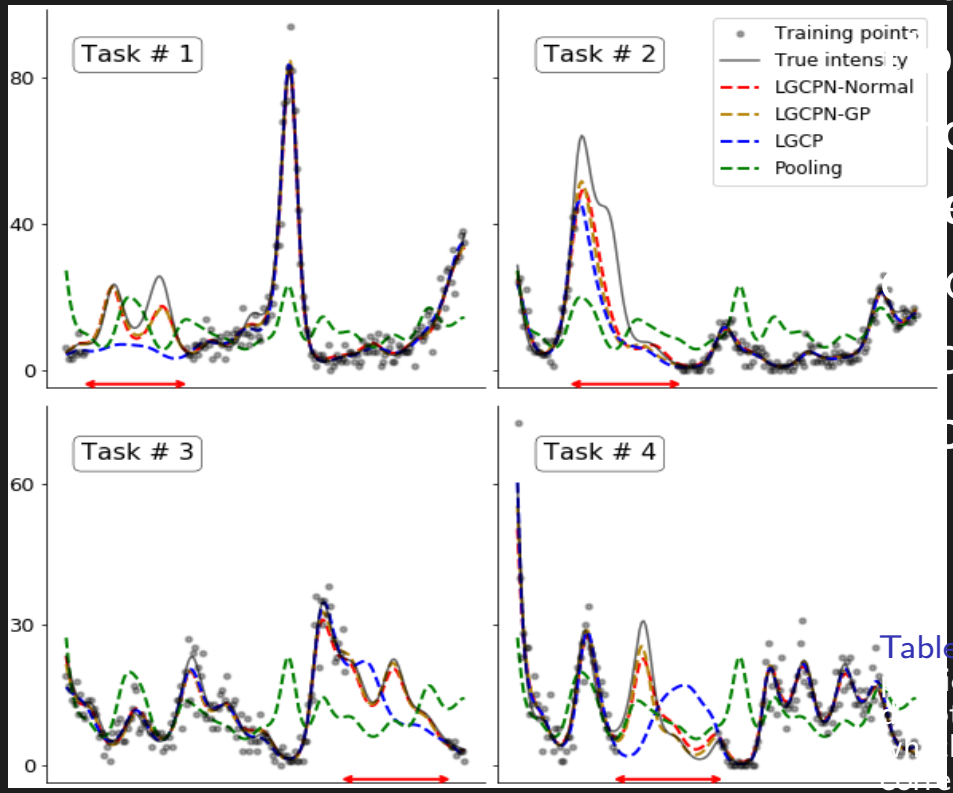


Figure 2: Four tasks with missing data in the marked intervals.

We consider **two latent functions** at 200 evenly spaced points and combine them via **task-specific mixing weights** to produce the final intensities for the correlated tasks. We then **remove 50 contiguous samples** from each task.

Table 1: Results on the synthetic dataset when making predictions on the missing intervals. Our model is denoted by either LGCPN-N or LGCPN-GP according to whether the prior over the weights is uncorrelated or correlated, respectively.

| Method \Task # | RMSE 1 | RMSE 2 | RMSE 3 | RMSE 4 | NLPL (PER CELL) 1 | NLPL (PER CELL) 2 | NLPL (PER CELL) 3 | NLPL (PER CELL) 4 |
|---|---|---|---|---|---|---|---|---|
| LGCPN-N | 4.45 | 13.41 | 4.61 | 5.06 | **3.13** | **7.78** | **3.04** | **3.19** |
| LGCPN-GP | **4.40** | 15.13 | **4.46** | **4.60** | 4.12 | 9.79 | 3.30 | 3.21 |
| LGCP | 11.05 | 17.63 | 7.74 | 14.02 | 26.43 | 31.33 | 21.09 | 32.35 |
| POOLING | 9.31 | 25.90 | 10.35 | 8.02 | 6.24 | 18.46 | 6.62 | 4.58 |
| ICM | 5.07 | **8.58** | 4.62 | 7.01 | 4.72 | 10.66 | 3.47 | 5.36 |

# Transfer experiment - Bovine tuberculosis (BTB) events

The BTB dataset (Fig. 3) consists of locations of BTB incidents in Cornwall, UK, in the period 1989–2002. We consider the four most common BTB genotypes (GT: 9, 12, 15 and 20).

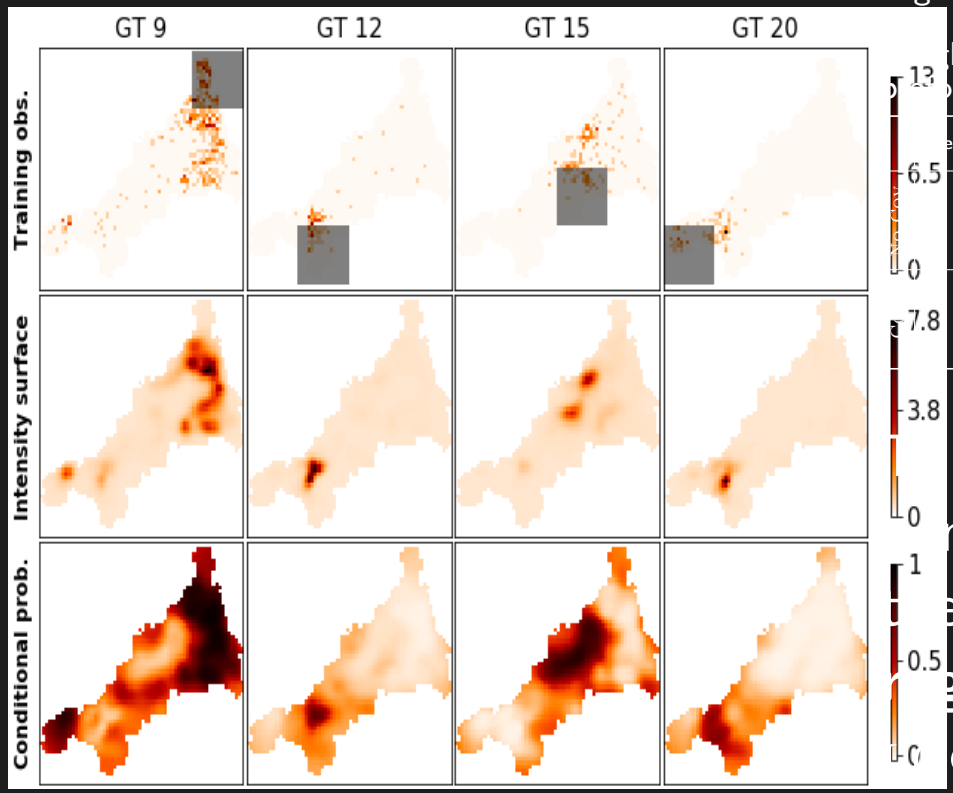


Figure 3: *Row 1:* Events of BTB in Cornwall, UK (period 1989-2002). Shaded areas represent missing data folds. Counts are on a 64×64 regular grid.

Table 2: Root mean squared error (RMSE) and mean negative log predictive likelihood on BTB with missing for our model (LGCPN) and the standard le-task) LGCP. The last three rows gives rmances obtained when considering covariates.

| ethod \btb type # | RMSE gt 9 | RMSE GT 12 | RMSE GT 15 | RMSE GT 20 | NLPL (PER CELL) GT 9 | NLPL (PER CELL) GT 12 | NLPL (PER CELL) GT 15 | NLPL (PER CELL) GT 20 |
|---|---|---|---|---|---|---|---|---|
| LGCPN-N | **0.35** | 0.13 | **0.14** | 0.14 | 0.36 | 0.08 | 0.15 | 0.09 |
| LGCPN-GP | **0.35** | **0.12** | **0.14** | **0.13** | **0.35** | **0.07** | **0.14** | **0.08** |
| LGCP | 23.15 | 28.61 | 18.45 | 54.65 | 10.09 | 11.46 | 7.79 | 22.72 |
| LGCPN-N | 0.37 | 0.13 | 0.15 | 0.14 | 0.38 | **0.08** | **0.14** | 0.10 |
| LGCPN-GP | **0.35** | **0.12** | **0.14** | **0.13** | **0.36** | **0.08** | 0.15 | **0.08** |
| LGCP | 14.03 | 16.92 | 9.90 | 32.00 | 6.37 | 6.91 | 4.32 | 13.51 |

e transfer appears less promi-
nt in this dataset due to
e presence of **spatially seg-**
**gated tasks**. We note how
CPN avoids negative transfer.

# Transfer experiment - Crime events

The dataset CRIME includes latitude and longitude locations of 7 main types of crime reported to the New York City Police Department (NYPD) in 2016.

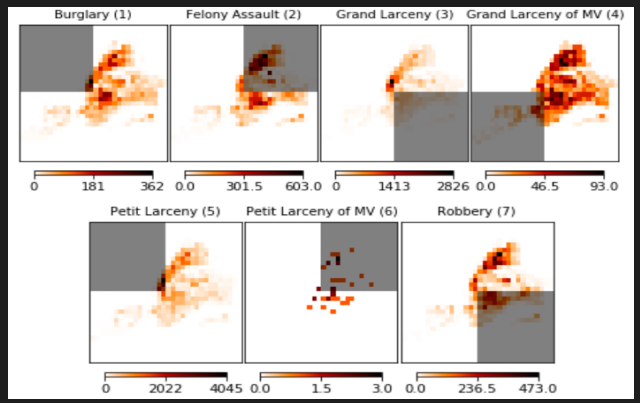


Figure 4: Counts of crime events on a $32 \times 32$ regular grid. Shaded areas represent missing data folds.

Table 3: CRIME dataset. Performances of alternative methods on the missing regions. 1. Burglary 2. Fellony Assault 3. Grand Larceny 4. Grand Larceny of MV 5. Petit Larceny 6. Petit Larceny of MV 6. Robbery.

| Method \Crime type # | RMSE 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| LGCPN-N | **9.30** | 23.39 | 78.96 | 7.90 | 89.94 | **0.20** | 12.54 |
| LGCPN-GP | 9.47 | **20.67** | 52.49 | 7.42 | **71.35** | **0.20** | **10.53** |
| LGCP | 26.27 | 50.14 | 114.89 | 10.62 | 209.83 | 2.61 | 38.53 |
| ICM | 23.26 | 35.17 | **51.96** | **6.91** | 101.12 | 0.27 | 22.34 |
| | NLPL (PER CELL) 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| LGCPN-N | **1.94** | 5.45 | 9.38 | **1.71** | 8.49 | **0.10** | **1.95** |
| LGCPN-GP | 2.26 | **5.18** | **5.83** | 1.93 | **8.26** | 0.18 | 2.23 |
| LGCP | 31.82 | 66.80 | 120.55 | 13.92 | 277.87 | 2.25 | 47.30 |
| ICM | 4.44 | 9.23 | 6.83 | 1.79 | 15.84 | 0.18 | 4.62 |

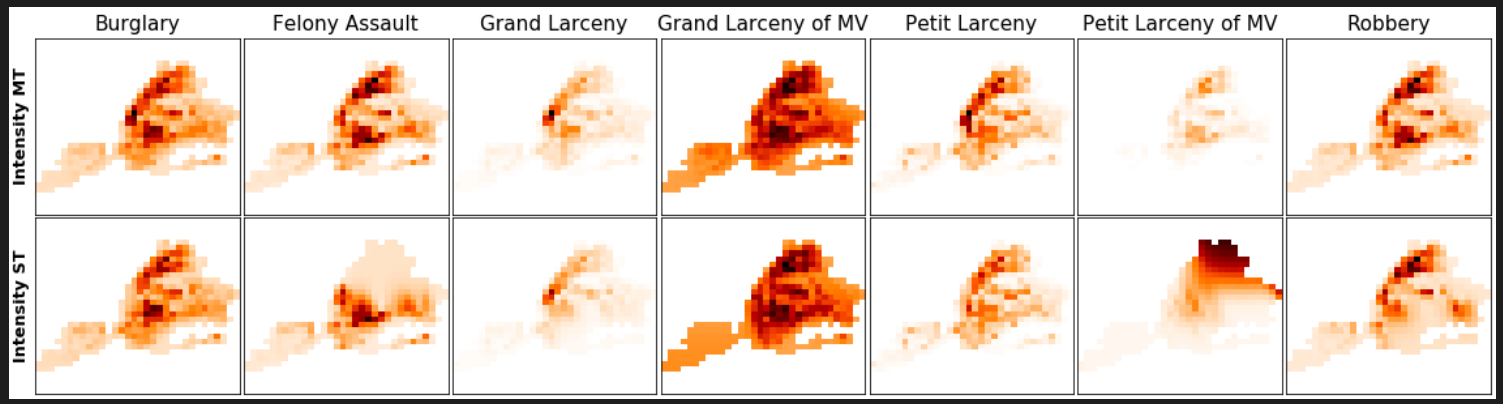


Figure 5: Estimated intensity surface when introducing missing data regions as in Fig. 4. *Row 1:* LGCPN *Row 2:* LGCP

# Short term directions

- **Validation** of LGCPN with alternative metrics (e.g. spatial k-fold);
- Comparison of **transfer capabilities** with respect to MLGCP;
- Impose additional **structure on the mixing weights** so as to increase transfer and improve interpretability;
- Develop a **spatio temporal** LGCPN framework to allow for real time monitoring of crime or desease events;
- Look at other **approximation techniques**.

# Long term directions

- Develop a **continuous formulation** for LGCPN thus getting rid of the approximation introduced by the regual grid;
- Incorporate **heteroscedacity** and **non-stationarity**.
- Adress **non-separability** between the spatial and temporal dimensions of the processes.
- Go **beyond GPs**: explore Student-$t$ processes or Gamma processes.
- Extend the framework to **other point processes**. Examples are negative binomial processes or Hawkes processes.

Additional **long term directions** include:

- Link the modelling side to the **policy intervention** issue: develop a probabilistic framework for analysing and assessing the impact of policies on spatio-temporal point processes thereby potentially driving the policy making process.
- ...

**Additional projects:**

- CUSP London: Assessing the impact of traffic policies on the number of accidents happening in London.
- ATI: Participation to the "Clean Air" project which aims at developing algorithms for air quality monitoring and forecasting in London.

**Thanks for your attention.**

# References I



David M Blei, Alp Kucukelbir, and Jon D McAuliffe.
Variational inference: A review for statisticians.
*Journal of the American Statistical Association*, (just-accepted), 2017.


Edwin V Bonilla, Karl Krauth, and Amir Dezfouli.
Generic inference in latent Gaussian process models.
*arXiv preprint arXiv:1609.00577*, 2016.


Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor.
Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm.
*Statistical Science*, pages 542–563, 2013.

# References II



C. Lloyd, T. Gunter, M. A. Osborne, and S. J. Roberts.
Variational Inference for Gaussian Process Modulated Poisson Processes.
In *International Conference on Machine Learning*, 2015.


Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen.
Log gaussian cox processes.
*Scandinavian journal of statistics*, 25(3):451–482, 1998.


Alexandra M Schmidt and Alan E Gelfand.
A bayesian coregionalization approach for multivariate pollutant data.
*Journal of Geophysical Research: Atmospheres*, 108(D24), 2003.

# References III


Benjamin Taylor, Tilman Davies, Barry Rowlingson, and Peter Diggle.
Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in r.

*Journal of Statistical Software*, 63:1–48, 2015.

Michalis K Titsias.
Variational learning of inducing variables in sparse gaussian processes.
5:567–574, 2009.

# Supplementary material

# Gaussian Processes

**Definition:** a Gaussian process ($\mathcal{GP}$) is a collection of random variables, any finite number of which have Gaussian distributions.

A $\mathcal{GP}$ is fully specified by a mean function $m(x)$ and covariance function $k(x, x')$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k_\theta(\boldsymbol{x}, \boldsymbol{x'})).$$

where $\theta$ represents the kernel hyperparameters.

The entropy term for $\mathbf{U}$ is given by:

$$
\begin{aligned}
\mathcal{L}^{u}_{\text{ent}}(\nu_u) &= -\mathbb{E}_{q(\mathbf{U}|\nu_u)}[\log q(\mathbf{U}|\nu_u)] \\
&= -\int q(\mathbf{U}|\nu_u) \log q(\mathbf{U}|\nu_u) d\mathbf{U} \\
&= -\sum_{q=1}^{Q} \int q(\mathbf{U}_{\bullet q}|\nu_u) \log q(\mathbf{U}_{\bullet q}|\nu_u) d\mathbf{U}_{\bullet q} \\
&= -\sum_{q=1}^{Q} \int \mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{m}_q, \mathbf{S}_q) \log \mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{m}_q, \mathbf{S}_q) d\mathbf{U}_{\bullet q} \\
&= -\sum_{q=1}^{Q} \left[ \mathcal{N}(\mathbf{m}_q; \mathbf{m}_q, \mathbf{S}_q) - \frac{1}{2} \operatorname{tr} (\mathbf{S}_q)^{-1} \mathbf{S}_q \right] \\
&= \frac{1}{2} \sum_{q=1}^{Q} \left[ M \log 2\pi + \log |\mathbf{S}_q| + M \right].
\end{aligned}
$$

The cross-entropy term for $\mathbf{U}$ is given by:

$$\begin{aligned}
\mathcal{L}^{u}_{\text{cross}}(\boldsymbol{\nu}_u) &= \mathbb{E}_{q(\mathbf{U}|\boldsymbol{\nu}_u)}[\log p(\mathbf{U})] \\
&= \int q(\mathbf{U}|\boldsymbol{\nu}_u) \log p(\mathbf{U}) d\mathbf{U} \\
&= \sum_{q=1}^{Q} \int q(\mathbf{U}_{\bullet q}|\boldsymbol{\nu}_u) \log p(\mathbf{U}_{\bullet q}) d\mathbf{U}_{\bullet q} \\
&= \sum_{q=1}^{Q} [\mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{m}_q, \mathbf{S}_q) \log \mathcal{N}(\mathbf{U}_{\bullet q}; \mathbf{0}, \mathbf{K}^q_{zz})] \\
&= \sum_{q=1}^{Q} \left[ \log \mathcal{N}(\mathbf{m}_q; \mathbf{0}, \mathbf{K}^q_{zz}) - \frac{1}{2} \operatorname{tr} (\mathbf{K}^q_{zz})^{-1} \mathbf{S}_q \right].
\end{aligned}$$

When placing a correlated prior on the mixing weights, the entropy term for $\boldsymbol{W}$ is given by:

$$\begin{aligned}
\mathcal{L}^{w}_{\text{ent}}(\boldsymbol{\nu}_w) &= -\int q(\mathbf{W}|\boldsymbol{\nu}_w)\log q(\mathbf{W}|\boldsymbol{\nu}_w)d\mathbf{W} \\
&= -\sum_{q=1}^{Q}\int \mathcal{N}(\mathbf{W}_{\bullet q};\boldsymbol{\omega}_q,\boldsymbol{\Omega}_q)\log\mathcal{N}(\mathbf{W}_{\bullet q};\boldsymbol{\omega}_q,\boldsymbol{\Omega}_q)d\mathbf{W}_{\bullet q} \\
&= -\sum_{q=1}^{Q}\left[\mathcal{N}(\boldsymbol{\omega}_q;\boldsymbol{\omega}_q,\boldsymbol{\Omega}_q) - \frac{1}{2}\,\mathrm{tr}\,(\boldsymbol{\Omega}_q)^{-1}\boldsymbol{\Omega}_q\right] \\
&= \frac{1}{2}\sum_{q=1}^{Q}\left[P\log 2\pi + \log|\boldsymbol{\Omega}_q| + P\right].
\end{aligned}$$

The cross-entropy term for $\boldsymbol{W}$ is given by:

$$
\begin{aligned}
\mathcal{L}^{w}_{\text{cross}}(\boldsymbol{\nu}_w) &= \mathbb{E}_{q(\mathbf{W}|\boldsymbol{\nu}_w)}[\log p(\mathbf{W})] \\
&= \int q(\mathbf{W}|\boldsymbol{\nu}_w) \log p(\mathbf{W}) d\mathbf{W} \\
&= \sum_{q=1}^{Q} \int q(\mathbf{W}_{\bullet q}|\boldsymbol{\nu}_w) \log p(\mathbf{W}_{\bullet q}) d\mathbf{W}_{\bullet q} \\
&= \sum_{q=1}^{Q} \int \mathcal{N}(\mathbf{W}_{\bullet q}; \boldsymbol{\omega}_q, \boldsymbol{\Omega}_q) \log \mathcal{N}(\mathbf{W}_{\bullet q}; \mathbf{0}, \mathbf{K}^q_w)) d\mathbf{W}_{\bullet q} \\
&= \sum_{q=1}^{Q} \left[ \log \mathcal{N}(\boldsymbol{\omega}_q; \mathbf{0}, \mathbf{K}^q_w) - \frac{1}{2} \operatorname{tr} (\mathbf{K}^q_w)^{-1} \boldsymbol{\Omega}_q \right].
\end{aligned}
$$

When placing an independent prior and approximate posterior over $\mathbf{W}$, the terms $\mathcal{L}^w_{\text{ent}}$ and $\mathcal{L}^w_{\text{cross}}$ get further simplified in:

$$\begin{aligned}
\mathcal{L}^w_{\text{ent}}(\boldsymbol{\nu}_w) &= -\int q(\mathbf{W}|\boldsymbol{\nu}_w) \log q(\mathbf{W}|\boldsymbol{\nu}_w) d\mathbf{W} \\
&= -\sum_{q=1}^{Q}\sum_{p=1}^{P}\int \mathcal{N}(w_{pq}; \omega_{pq}, \Omega_{pq}) \log \mathcal{N}(w_{pq}; \omega_{pq}, \Omega_{pq}) dw_{pq} \\
&= \frac{1}{2}\sum_{q=1}^{Q}\sum_{p=1}^{P}\left[\log 2\pi + \log \Omega_{pq} + 1\right],
\end{aligned}$$

$$
\begin{aligned}
\mathcal{L}^{w}_{\text{cross}}(\boldsymbol{\nu}_w) &= \int q(\mathbf{W}|\boldsymbol{\nu}_w) \log p(\mathbf{W}) d\mathbf{W} \\
&= \sum_{q=1}^{Q} \sum_{p=1}^{P} \int q(w_{pq}|\boldsymbol{\nu}_w) \log p(w_{pq}) dw_{pq} \\
&= \sum_{q=1}^{Q} \sum_{p=1}^{P} \int \mathcal{N}(w_{pq}; \omega_{pq}, \Omega_{pq}) \log \mathcal{N}(w_{pq}; 0, \sigma^2_{pq}) dw_{pq} \\
&= \sum_{q=1}^{Q} \sum_{p=1}^{P} \left[ \log \mathcal{N}(\omega_{pq}; \mathbf{0}, \Omega_{pq}) - \frac{\Omega_{pq}}{2\sigma^2_{pq}} \right],
\end{aligned}
$$

Consider now the LGCPN model in which $\lambda(\mathbf{x}) = \exp(\sum_{q=1}^{Q} wf(\mathbf{x}))$. In this case, the likelihood for the continuous case can be written as:

$$p(Y|\lambda) = \exp\left[-\sum_{p=1}^{P}\int_{\tau}\lambda_p(\mathbf{x})dx\right]\prod_{p=1}^{P}\prod_{n_p}^{N_p}\lambda_p(\mathbf{x}_{n_p}) \tag{5}$$

Notice that $n_p$ indicates the location of the $n$-th event for the $p$-th task. The expected log likelihood is defined as:

$$\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}\left[-\sum_{p}^{P}\int_{\tau}\lambda_p(\mathbf{x})d\mathbf{x} + \sum_{p=1}^{P}\sum_{n_p}^{N_p}\log\lambda_p(\mathbf{x}_{n_p})\right] \tag{6}$$

Replacing the expression for the intensity we get:

$$-\sum_{p=1}^{P}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\exp(\sum_{q=1}^{Q}w_p f(\mathbf{x}))q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F}d\mathbf{x}+ \tag{7}$$

$$+\sum_{p=1}^{P}\sum_{n_p}^{N_p}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\log\left[\exp(\sum_{q=1}^{Q}w_p f(\mathbf{x}_{n_p}))\right]q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F} \tag{8}$$

$$=-\sum_{p=1}^{P}\int_{\tau}\mathbb{E}\left[\exp(\sum_{q=1}^{Q}w_p f(\mathbf{x}))\right]d\mathbf{x} \tag{9}$$

$$+\sum_{p=1}^{P}\sum_{n_p}^{N_p}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\sum_{q=1}^{Q}w_p f(\mathbf{x}_{n_p})q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F}d\mathbf{x} \tag{10}$$

$$=-\sum_{p=1}^{P}\int_{\tau}\mathbb{E}\left[\exp(\sum_{q=1}^{Q}wf(\mathbf{x}))\right]d\mathbf{x}+\sum_{p=1}^{P}\sum_{n_p}^{N_p}\mathbb{E}(\sum_{q=1}^{Q}w_p f(\mathbf{x}_{n_p})) \tag{11}$$

We know that the expected value of $\exp(\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p}))$ is equal to:

$$\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\lambda_p(\mathbf{x})) = \tag{12}$$

$$= \exp(\phi_p) \prod_{q=1}^{Q} \frac{1}{\sqrt{1 - \Omega_{pq}^2 \Sigma_{nn}^q}} * \tag{13}$$

$$* \exp\left( -\frac{1}{2\Omega_{pq}^2} \left( \omega_{pq}^2 + \frac{(\Omega_{pq}^2 \mu_q(\mathbf{x}) + \omega_{pq})^2}{\Omega_{pq}^2 \Sigma_{nn}^q - 1} \right) \right) \tag{14}$$

$\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}(\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p}))$ is equal to:

$$\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}(\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p})) = \tag{15}$$

$$= \mathbb{E}_{q(\mathbf{W})}\left(\mathbb{E}_{q(\mathbf{F})}\left[\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p})|\mathbf{W}\right]\right) \tag{16}$$

$$= \mathbb{E}_{q(\mathbf{W})}\left[\sum_{q=1}^{Q} w_p \mu_q(\mathbf{x}_{n_p})\right] \tag{17}$$

$$= \sum_{q=1}^{Q} \omega_p \mu_q(\mathbf{x}_{n_p})) \tag{18}$$

We are left with the following integral:

$$-\sum_{p=1}^{P}\left[\prod_{q=1}^{Q}\frac{1}{\sqrt{1-\Omega_{pq}^{2}\Sigma_{nn}^{q}}}\exp\left(-\frac{\omega_{pq}^{2}}{2\Omega_{pq}^{2}}\right)\int_{\tau}\exp\left(\frac{(\Omega_{pq}^{2}\mu_{q}(\mathbf{x})+\omega_{pq})^{2}}{\Omega_{pq}^{2}\Sigma_{nn}^{q}-1}\right)d\mathbf{x}\right.$$

where the posterior mean for $q(\mathbf{F})$ computed in $\mathbf{x}$ is defined as $\mu_q(\mathbf{x}) = k_{xz}^{q}(K_{zz})^{-1}m_q$.
What we need to solve is the following integral:

$$\int_{\tau}\exp\left(\frac{(\Omega_{pq}^{2}k_{xz}^{q}(K_{zz})^{-1}m_{q}+\omega_{pq})^{2}}{\Omega_{pq}^{2}\Sigma_{nn}^{q}-1}\right)d\mathbf{x} \qquad (19)$$

$$\mathrm{KL}(q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)\|p(\mathbf{F},\mathbf{U},\mathbf{W}|\mathcal{D})) = \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log\frac{q(\mathbf{F})}{p(\mathbf{F}|y)}\right]$$

$$= \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log q(\mathbf{F}) - \log p(\mathbf{F}|y)\right]$$

$$= \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log q(\mathbf{F})\right] - \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log p(\mathbf{F},y)\right] + \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log p(y)\right]$$

$$= -\left[\mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log p(\mathbf{F},y)\right] - \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log q(\mathbf{F})\right]\right] + \mathbb{E}_{q(\mathbf{F},\mathbf{U},\mathbf{W}|\nu)}\left[\log p\right.$$

$$= -\mathrm{ELBO} + \log p(y)$$

The second term in this equation doesn't depend on $q(\cdot)$ thereby minimization of the KL divergence is equivalent to the maximisation of the ELBO.

$$
\begin{aligned}
\log p(y) &= \log \int p(\mathbf{F}, y) d\mathbf{F} \\
&= \log \int p(\mathbf{F}, y) \frac{q(\mathbf{F})}{q(\mathbf{F})} d\mathbf{F} \\
&= \log \left[ \mathbb{E}_{q(\mathbf{F})} \frac{p(\mathbf{F}, y)}{q(\mathbf{F})} \right] \\
&\geq \mathbb{E}_{q(\mathbf{F})} \left[ \log \frac{p(\mathbf{F}, y)}{q(\mathbf{F})} \right] \\
&= \mathbb{E}_{q(\mathbf{F})} \left[ \log p(\mathbf{F}, y) \right] - \mathbb{E}_{q(\mathbf{F})} \left[ \log q(\mathbf{F}) \right] \\
&= \text{ELBO}
\end{aligned}
$$

Here we derive a closed form expression for the expected value of the predictive intensities, say $\mathbb{E}_{\boldsymbol{\lambda}|\mathcal{D}}(\boldsymbol{\lambda})$, which is defined as:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\lambda}|\mathcal{D}}(\boldsymbol{\lambda}) &= \mathbb{E}_{q(\mathbf{W},\mathbf{F})}(\exp(\mathbf{W}\mathbf{F}^T + \phi)) \\ &= \exp(\phi)\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\exp(\mathbf{W}\mathbf{F}^T)\end{aligned}$$

where $\phi$ represents the offsets to the log mean and $q(\mathbf{W})$ and $q(\mathbf{F})$ denote the variational distributions over the latent functions and the weights respectively. For each task $p$ and test point $n$ the expeted value of the predicted intensity is defined as:

$$\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\lambda_{np}) = \exp(\phi_p)\mathbb{E}_{q(\mathbf{W})}\left[\mathbb{E}_{q(\mathbf{F})}(\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet}^T)|\mathbf{W})\right]$$

Given the defined $q(\mathbf{F})$, the inner expectation can be written as:

$$\mathbb{E}_{q(\mathbf{F})}(\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet}^{T})|\mathbf{W}) = \prod_{q=1}^{Q} \mathbb{E}_{q(\mathbf{F})}(\exp(w_{pq}f_q(\mathbf{x}_n)|\mathbf{W}) \quad (20)$$

Denote by $\boldsymbol{\mu}_q$ and $\Sigma^q$ the mean and covariance parameters of $q(\mathbf{F})$. Each term on the right hand side of Eq. 20 represents the expected value of a log-Normal distribution with mean equal to $\mu_q(\mathbf{x}_n) + \phi_p$ and variance equal to $\Sigma_{nn}^q$. We can thus rewrite:

$$\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\lambda_{np}) = \exp(\phi_p)\mathbb{E}_{q(\mathbf{W})}\left[\sum_{q=1}^{Q} \exp(w_{pq}\mu_q(\mathbf{x}_n) + \frac{1}{2}w_{pq}^2\Sigma_{nn}^q)\right]$$

Given the factorisation of $q(\mathbf{W})$ across latent functions, we can rewrite this last equation as:

$$\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\lambda_{np}) = \exp(\phi_p) \prod_{q=1}^{Q} \int \exp(w_{pq}\mu_q(\mathbf{x}_n) + \frac{1}{2} w_{pq}^2 \Sigma_{nn}^q) p(\mathbf{W}_{\bullet q}) d\mathbf{W}_{\bullet q}$$

Solving the integrals we find:

$$\begin{aligned}\mathbb{E}_{q(\mathbf{W})q(\mathbf{F})}(\lambda_{np}) &= \\ &= \exp(\phi_p) \prod_{q=1}^{Q} \frac{1}{\sqrt{1 - \Omega_{pq}^2 \Sigma_{nn}^q}} * \\ &\quad * \exp\left( -\frac{1}{2\Omega_{pq}^2} \left( \omega_{pq}^2 + \frac{(\Omega_{pq}^2 \mu_q(\mathbf{x}_n) + \omega_{pq})^2}{\Omega_{pq}^2 \Sigma_{nn}^q - 1} \right) \right)\end{aligned}$$