Efficient Inference in Multi-task Cox Process Models Supplementary Material

Virginia Aglietti
University of Warwick
The Alan Turing Institute

Theodoros Damoulas
University of Warwick
The Alan Turing Institute

Edwin V. Bonilla
CSIRO's Data61
UNSW

1 Derivation of the KL-divergence Term

The KL-divergence terms composing the ELBO can be written as $\mathcal{L}_{\text{kl}}(\boldsymbol{\nu}) = \mathcal{L}^u_{\text{ent}}(\boldsymbol{\nu}_u) + \mathcal{L}^u_{\text{cross}}(\boldsymbol{\nu}_u) + \mathcal{L}^w_{\text{ent}}(\boldsymbol{\nu}_w) + \mathcal{L}^w_{\text{cross}}(\boldsymbol{\nu}_w)$ where each term is given by:

$$\mathcal{L}^u_{\text{cross}}(\boldsymbol{\nu}_u) = \sum_{q=1}^{Q} \left[\log \mathcal{N}(\mathbf{m}_q; \mathbf{0}, \mathbf{K}^q_{zz}) - \frac{1}{2} \text{ tr } (\mathbf{K}^q_{zz})^{-1} \mathbf{S}_q \right] \tag{1}$$

$$\mathcal{L}^u_{\text{ent}}(\boldsymbol{\nu}_u) = \frac{1}{2} \sum_{q=1}^{Q} \left[M \log 2\pi + \log |\mathbf{S}_q| + M \right] \tag{2}$$

$$\mathcal{L}^w_{\text{cross}}(\boldsymbol{\nu}_w) = \sum_{q=1}^{Q} \left[\log \mathcal{N}(\boldsymbol{\omega}_q; \mathbf{0}, \mathbf{K}^q_{w}) - \frac{1}{2} \text{ tr } (\mathbf{K}^q_{w})^{-1} \boldsymbol{\Omega}_q \right] \tag{3}$$

$$\mathcal{L}^w_{\text{ent}}(\boldsymbol{\nu}_w) = \frac{1}{2} \sum_{q=1}^{Q} \left[\ P \log 2\pi + \log |\boldsymbol{\Omega}_q| + P \right], \tag{4}$$

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 simplified further, reducing the computational cost significantly when a large number of tasks is considered. Here we derive the expressions for Eqs. (2)–(4).

The cross-entropy term for $\mathbf{U}$ (Eq. (1)) is given by:

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

The entropy term for $\mathbf{U}$ (Eq. (2)) is given by:

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

When placing a coupled prior on the mixing weights, the cross-entropy term for $\boldsymbol{W}$ (Eq. (3)) is given by:

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

The entropy term for $\boldsymbol{W}$ (Eq. (4)) is given by:

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

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:

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

and in:

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

where Ω_{pq} represents the p-th diagonal term of $\boldsymbol{\Omega}_q$.

2 Closed form evaluation of the ELL term

The MCPM model formulation allows to derive a closed form expression for the moments of the intensity function. Here we provide details about the derivations and obtain an expression for the first moment of $\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet})$ which has been used in the closed form evaluation of $\mathcal{L}_{\text{ell}}$.

In order to compute the moments of λ we can exploit the moment generating function (MGF) of the product of two normal random variables. Denote by X and Y two independent and normally distributed random variables. The variable $Z = XY$ has $\text{MGF}_Z(t)$ defined as:

$$\text{MGF}_Z(t) = \frac{\exp\left[\frac{t\mu_X\mu_Y + 1/2(\mu_Y^2\sigma_X^2 + \mu_X^2\sigma_Y^2)t^2}{1 - t^2\sigma_X^2\sigma_Y^2}\right]}{\sqrt{1 - t^2\sigma_X^2\sigma_Y^2}}. \qquad (5)$$

Now define $V = \sum_{q=1}^{Q} X_qY_q$ where $X_q \perp\!\!\!\perp Y_q, \forall q$, $X_q \perp\!\!\!\perp X_{q'}, \forall q, q'$ and $Y_q \perp\!\!\!\perp Y_{q'}, \forall q, q'$. Given these assumptions, the MGF for V is defined as the product of Q MGF of the form given in Eq. (5). We have $\text{MGF}_V(t) = \prod_{q=1}^{Q}\text{MGF}_{Z_q}(t)$. This implies that:

$$\mathbb{E}(\lambda_p) = \mathbb{E}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet})\right] = \text{MGF}_V(1) \qquad (6)$$

where $X_q = \omega_{pq}$ and $Y_q = f_{nq}$.

Exploiting Eq. (6) we can derive a closed form expression for $\mathcal{L}_{\text{ell}}$:

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{F}),q(\mathbf{W})}\left[\log(p(\mathbf{Y}|\mathbf{F},\mathbf{W}))\right] =\\ &= -\sum_{n=1}^{N}\sum_{p=1}^{P}\mathbb{E}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet} + \phi_p) + y_{np}\log(\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet} + \phi_p)\right.\\ &\left.+\log\Gamma(y_{np}+1)\right]\\ &= \sum_{n=1}^{N}\sum_{p=1}^{P}\mathbb{E}\left[-\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet} + \phi_p)y_{np}\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet} + y_{np}\phi_p\right.\\ &\left.+\log\Gamma(y_{np}+1)\right]\\ &= -\sum_{n=1}^{N}\sum_{p=1}^{P}\mathbb{E}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet} + \phi_p)\right] + \sum_{n=1}^{N}\sum_{p=1}^{P}\left[y_{np}\mathbb{E}(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet}) +\right.\\ &\left.+y_{np}\phi_p + \log\Gamma(y_{np}+1)\right]\\ &= -\sum_{n=1}^{N}\sum_{p=1}^{P}\exp(\phi_p)MGF_V(1) +\\ &\sum_{n=1}^{N}\sum_{p=1}^{P}\sum_{q=1}^{Q}(y_{np}\omega_{pq}\mu_q(\mathbf{x}_n) + y_{np}\phi_p + \log\Gamma(y_{np}+1))\end{aligned} \qquad (7)$$

Given the moments of $q(\mathbf{W}_{p\bullet})$ and $q(\mathbf{F}_{n\bullet})$ we can write:

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{F}_{n\bullet})q(\mathbf{W}_{p\bullet})}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet})\right]\\ &= \prod_{q=1}^{Q}\frac{\exp\left[\frac{\omega_{pq}\mu_{nq} + 1/2(\mu_{nq}^2\Omega_{pq} + \omega_{pq}^2\Sigma_{nn}^{q})}{1 - \Omega_{pq}\Sigma_{nn}^{q}}\right]}{\sqrt{1 - \Omega_{pq}\Sigma_{nn}^{q}}}\end{aligned} \qquad (8)$$

Defining $\delta_X = \mu_X/\sigma_X$ in Eq. (5) we can rewrite $\text{MGF}_Z(t)$ as:

$$\mathrm{MGF}_Z(t) = \frac{\exp\left[\frac{t\mu_X\mu_Y + 1/2(\mu_Y^2\sigma_X^2 + \mu_X^2\sigma_Y^2)t^2}{1 - t^2\frac{\mu_X^2\sigma_Y^2}{\delta_X^2}}\right]}{\sqrt{1 - t^2\frac{\mu_X^2\sigma_Y^2}{\delta_X^2}}} \tag{9}$$

As δ_X increases, MGF$_Z(t)$ converges to the form:

$$\mathrm{MGF}_Z(t) = \exp\left[t\mu_X\mu_Y + 1/2(\mu_Y^2\sigma_X^2 + \mu_X^2\sigma_Y^2)t^2\right] \tag{10}$$

which is the MGF of a Gaussian distribution with mean and variance given by $\mu_X\mu_Y$ and $\mu_Y^2\sigma_X^2 + \mu_X^2\sigma_Y^2$ respectively (Seijas-Macías and Oliveira, 2012). This implies that, for increasing values of δ_{X_q} the sum of the products of Gaussians tends to a Gaussian distribution.

3 Relationship to existing literature

As mentioned in §2, when $\mathbf{\Omega}_q \to 0$, $\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet}$ converges to a Gaussian distribution. Depending on the number of latent GPs included in the model (Q) and the moments of $q(\mathbf{W}_{p\bullet})$, MCPM will thus converge either to an ICM (or LCM) or to a MLGCP or to an LGCP. When $Q \neq P$, we have $\log\lambda_p(\mathbf{x}^{(n)}) = \sum_{q=1}^{Q}\omega_{pq}\mathbf{F}_{n\bullet}$ for each n and p. We can thus write:

$$\begin{aligned}
&\lim_{\mathbf{K}_w \to 0}\mathrm{Cov}(\log\lambda_p(\mathbf{x}), \log\lambda_{p'}(\mathbf{x}')) \\
&= \sum_{q=1}^{Q}\sum_{q'=1}^{Q}\omega_{pq}\omega_{p'q'}\mathrm{Cov}(\mathbf{F}_{\bullet q}, \mathbf{F}_{\bullet q'}) \\
&= \sum_{q=1}^{Q}\underbrace{\omega_{pq}\omega_{p'q'}}_{\boldsymbol{B}_q}\widetilde{\mathbf{K}}^q_{\mathbf{xx}'}
\end{aligned}$$

where we have exploited the independence assumption between $\mathbf{F}_{\bullet q}$ and $\mathbf{F}_{\bullet q'}$ for $q \neq q'$.

When $Q = P + 1$ and $\mathbf{W}_{P\times(P+1)} = [\boldsymbol{I}_P \;\; \mathbb{1}_P]$, the intensity for each task will be determined by the $(P+1)$-th common GP and by the p-th task specific GP. We thus recover the MLGCP formulation.

Finally, when $Q = P$ and $\mathbf{W}_{P\times P} = \boldsymbol{I}_P$, the intensity for each task will be determined only by the p-th task specific GP. We thus recover the LGCP formulation.

We summarise these results in the following lemma:

LEMMA 1 MCPM *generalizes* ICM, MLGCP *and* LGCP. *As Cov*$(w_{pq}, w_{p'q'}) \to 0, \forall p, q, p', q'$, *for* $Q \neq P$ *we have* $\hat{\lambda}_{\text{MCPM}} \to \hat{\lambda}_{\text{ICM}}$ *(or a* $\hat{\lambda}_{\text{MCPM}} \to \hat{\lambda}_{\text{LCM}}$ *depending on the assumed covariance functions for the latent* GP*s) where the intensity parameters are jointly determined by the moments of* $\mathbf{F}$ *and* $\mathbf{W}$*:*

$$\begin{aligned}
&\lim_{\substack{Cov(w_{pq}, w_{p'q'}) \to 0 \\ \forall p,q,p',q'}}\mathrm{Cov}(\log\lambda_p(\mathbf{x}), \log\lambda_{p'}(\mathbf{x}')) \\
&\qquad = \sum_{q=1}^{Q}\underbrace{\gamma_{pq}\gamma_{p'q'}}_{\boldsymbol{B}_{q_{(p,p')}}}\widetilde{\mathbf{K}}^q_{\mathbf{xx}'}
\end{aligned}$$

where $\boldsymbol{B}_q \in \mathbb{R}^{P\times P}$ *is known as coregionalisation matrix. For* $Q = P + 1$ *and* $\mathbf{W}_{P\times(P+1)} = [\boldsymbol{I}_P \;\; \mathbb{1}_P]$ *we have* $\hat{\lambda}_{\text{MCPM}} \to \hat{\lambda}_{\text{MLGCP}}$*. Finally, for* $Q = P$ *and* $\mathbf{W}_{P\times P} = \boldsymbol{I}_P$ *we have* $\hat{\lambda}_{\text{MCPM}} \to \hat{\lambda}_{\text{LGCP}}$*.*

When having task descriptors $\mathbf{h}$, we can view the log intensity as a function of the joint space of input features and task descriptors *i.e.* $\log\lambda(\mathbf{x}, \mathbf{h})$. It is possible to show that under our independence prior assumption between weights ($\mathbf{W}$) and latent functions ($\mathbf{F}$), the prior covariance over the log intensities (evaluated at inputs $\mathbf{x}$ and $\mathbf{x}'$ and tasks p and p') is given by:

$$\mathbb{C}\mathrm{ov}[\log\lambda_p(\mathbf{x}), \lambda_{p'}(\mathbf{x}')] \quad = \sum_{q=1}^{Q}\kappa_w^q(\mathbf{h}^p, \mathbf{h}^{p'})\kappa_f^q(\mathbf{x}, \mathbf{x}')$$

where $\mathbf{h}^p$ denotes the p-th task descriptors. At the observed data $\{\mathbf{X}, \mathbf{H}\}$, assuming a regular grid, the MCPM prior covariance over the log intensities is $\mathbb{C}\mathrm{ov}[\log\boldsymbol{\lambda}(\mathbf{X}), \log\boldsymbol{\lambda}(\mathbf{X})] = \sum_{q=1}^{Q}\mathbf{K}_w^q \otimes \mathbf{K}_f^q$. This is effectively the LCM prior with $\mathbf{K}_w^q$ denoting the coregionalization matrix. Importantly, the two methods differ substantially in terms of inference. While in LCM a point estimate of $\mathbf{K}_w^q$ is generally obtained, MCPM proceeds by optimizing the hyperparameters for $\mathbf{K}_w^q$ and doing full posterior estimation for both $\mathbf{W}$ and $\mathbf{F}$. In addition, by adopting a process view on $\mathbf{W}$, we increase the model flexibility and accuracy by capturing additional correlations across tasks while being able to generalize over unseen task descriptors. Last but not least, by considering our priors and approximate posteriors over $\mathbf{W}$ and $\mathbf{F}$ separately, instead of a single joint prior over the log intensities, we can exploit state-of-the art inducing variable approximations (Titsias, 2009) over each $\mathbf{W}_{\bullet q}$ and $\mathbf{F}_{\bullet q}$ separately, instead of dealing with a sum of Q Kronecker products for which there is not an efficient decomposition when $Q > 2$ (Rakitsch et al., 2013).

4 Continuous MCPM formulation

Following a common approach, in this work we introduce a computational grid on the spatial extend and consider the cells' centroids as inputs of MCPM. Here we discuss the continuous formulation of our model. The likelihood function for the continuous MCPM model can be written as:

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

where we assume all events to be distinct and we denote as n_p the location of the n-th event for the p-th task. This implies an expected log likelihood term defined as:

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

Replacing the expression for MCPM intensity in the previous equation we get:

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}\log(p(Y|\lambda)) \\
&= -\sum_{p=1}^{P}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet})q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F}d\mathbf{x}+ \\
&+\sum_{p=1}^{P}\sum_{n_p}^{N_p}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\log\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet})\right]q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F} \\
&= -\sum_{p=1}^{P}\int_{\tau}\mathbb{E}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet})\right]d\mathbf{x} \\
&+\sum_{p=1}^{P}\sum_{n_p}^{N_p}\int_{\tau}\int_{\mathbf{F}}\int_{\mathbf{W}}\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet}q(\mathbf{W})q(\mathbf{F})d\mathbf{W}d\mathbf{F}d\mathbf{x} \\
&= -\sum_{p=1}^{P}\int_{\tau}\mathbb{E}\left[\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet})\right]d\mathbf{x} + \sum_{p=1}^{P}\sum_{n_p}^{N_p}\mathbb{E}(\mathbf{W}_{p\bullet}\mathbf{F}_{n_p\bullet})
\end{aligned}$$

for a bounded region τ. The expected value of $\exp(\mathbf{W}_{p\bullet}\mathbf{F}_{n\bullet})$ can be computed as in Eq. 8 while $\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}(\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p}))$ is equal to:

$$\mathbb{E}_{q(\mathbf{F})q(\mathbf{W})}(\sum_{q=1}^{Q} w_p f(\mathbf{x}_{n_p})) = \sum_{q=1}^{Q}\omega_p\mu_q(\mathbf{x}_{n_p})) \qquad (11)$$

We are thus left with an intractabe integral of the form:

$$\begin{aligned}
-\sum_{p=1}^{P}\Bigg[\prod_{q=1}^{Q}\frac{1}{\sqrt{1-\Omega_{pq}^2\Sigma_{nn}^q}}\exp\left(-\frac{\omega_{pq}^2}{2\Omega_{pq}^2}\right)\ldots \\
\ldots\int_{\tau}\exp\left(\frac{(\Omega_{pq}^2\mu_q(\mathbf{x})+\omega_{pq})^2}{\Omega_{pq}^2\Sigma_{nn}^q-1}\right)d\mathbf{x}\Bigg]
\end{aligned}$$

where the posterior mean for $q(\mathbf{F})$ computed in $\mathbf{x}$ is defined as $\mu_q(\mathbf{x}) = k_{xz}^q(K_{zz})^{-1}m_q$.

This integral could be approximated using a series expansion but this would result in a computationally difficult problem.

5 Pseudo-algorithm

Algorithm 1 illustrates the MCPM algorithm:

Algorithm 1 LGCPN

```
Inputs: Observational dataset $\mathcal{D} = \{\mathbf{x}_p^{(i)} \in \tau, \forall p = 1, ..., P\}_{i=1}^{I}$ for bounded region $\tau$ where $I$ denotes the number of events. Number of latent GPs $Q$. Number of mini-batches b of size $B$.
Output: Optimized hyper-parameters, posterior moments of $\boldsymbol{\lambda}$

 Discretize event locations $\mathcal{D}$ in $Y \in \mathbb{R}^{N\times P}$ given the grid size.
Initialize: $i \leftarrow 0$, $\boldsymbol{\eta}^{(0)} = (\boldsymbol{\theta}, \boldsymbol{\theta}_w, \boldsymbol{\phi}, \boldsymbol{\nu}_u, \boldsymbol{\nu}_w)$
repeat
    $\{X_{train} \in \mathbb{R}^{B\times D}, Y_{train} \in \mathbb{R}^{B\times P}\} \rightarrow$ get-next-MiniBatch($\mathcal{D}$)
    for j=0 to b do
        $\max_{\boldsymbol{\mu}} \mathcal{L}_{\text{elbo}}(\boldsymbol{\eta}^{(i)})$ (Eqs. (2)–(4) and (7))
        $\boldsymbol{\eta}^{(i)} \leftarrow \boldsymbol{\eta}^{(i-1)} - \rho\nabla_{\boldsymbol{\eta}}\mathcal{L}_{\text{elbo}}(\boldsymbol{\eta}^{(i-1)})$
        $i = i + 1$
    end for
until convergence criterion is met.
 $\boldsymbol{\eta}^* \leftarrow \boldsymbol{\eta}^{(i-1)}$
 $\mathbb{E}[\boldsymbol{\lambda}(\mathbf{x})^t] = \exp(t\boldsymbol{\phi}^*)\text{MGF}_{\mathbf{WF}|\boldsymbol{\eta}^*}(t)$
```

6 Plate diagram

Here we provide the plate diagram for MCPM with independent prior on the mixing weights:

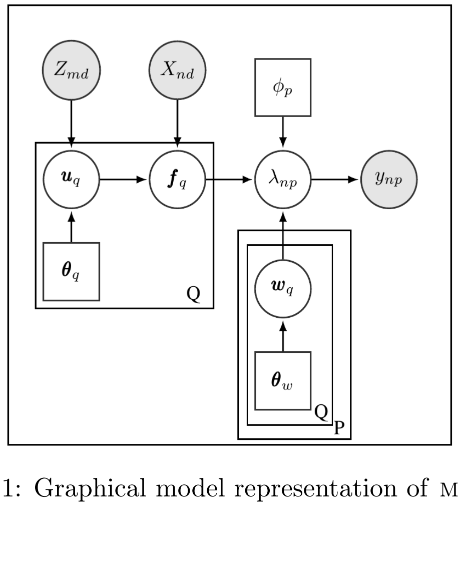

Figure 1: Graphical model representation of MCPM-N.

7 Algorithmic efficiency

Evaluating $\mathcal{L}_{\text{ell}}$ in closed form, we are able to significanlty speed up the algorithm by getting rid of the

Monte Carlo evaluations, see Fig. 2 and Fig. 3.

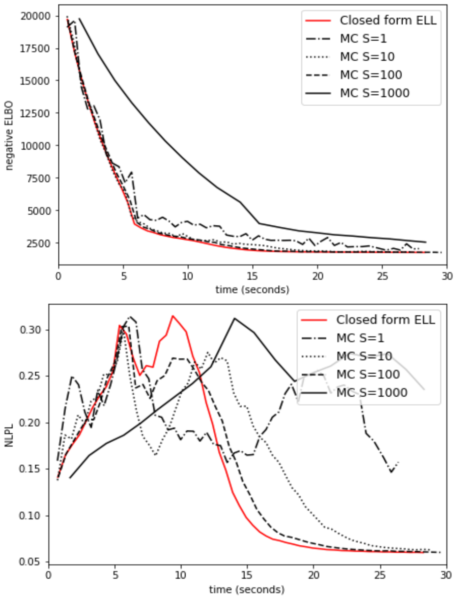

Figure 2: Synthetic data. MC estimate of ELL vs. Closed form evaluation of ELL. *Left:* Negative ELBO values over time. *Right:* NLPL values for one task over time. S denotes the number of samples used in the MC evaluation.

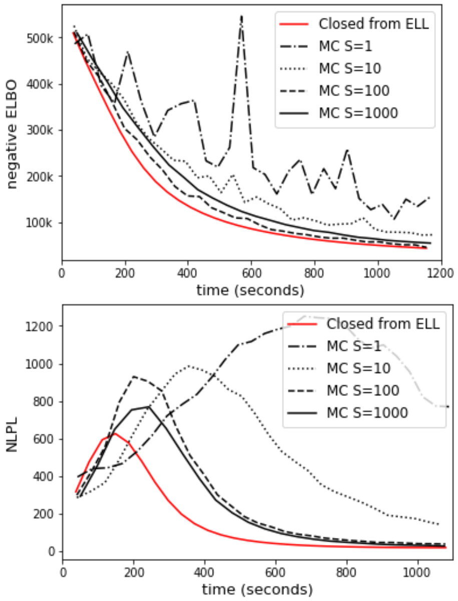

Figure 3: CRIME data. MC estimate of ELL vs. Closed form evaluation of ELL. *Left:* Negative ELBO values over time. *Right:* NLPL values for one task over time. S denotes the number of samples used in the MC evaluation.

8 Additional experimental results

Synthetic experiments Here we report additional performance metrics for the two synthetic experiments included in the text. Tab. 1 gives the coverage numbers for the first synthetic experiment while Tab. 2 and Tab. 3 display the RMSE and coverage performances for the second synthetic dataset. Fig. 4 gives the predicted counts distributions for the second synthetic dataset.

Table 1: S1 dataset. In-sample/Out-of-sample 90% CI coverage for the predicted event counts distributions.

	Empirical Coverage (EC)			
	1	**2**	**3**	**4**
MCPM-N	0.80/0.12	**0.99**/0.58	0.92/0.57	**0.94**/**0.83**
MCPM-GP	**0.95**/**0.19**	0.72/**0.67**	**1.00**/**0.78**	0.92/0.75
ICM	0.75/0.03	0.66/0.60	0.62/0.50	0.93/0.42

Table 2: S2 dataset. RMSE performance when making predictions on the interval $[80, 100]$.

	RMSE									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MCPM-N	**1.10**	**1.15**	**0.89**	0.17	0.95	0.99	1.10	0.63	1.50	0.55
MCPM-GP	1.15	1.43	0.91	**0.13**	0.94	**0.97**	1.19	**0.58**	**1.43**	0.70
MTPP	1.20	1.70	1.12	0.17	**0.91**	1.05	**1.05**	1.11	1.61	**0.49**

Table 3: S2 dataset. In-sample/Out-of-sample 90% CI coverage for the predicted event counts distributions.

	Empirical Coverage									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MCPM-N	**1.00**/**1.00**	**1.00**/**1.00**	0.95/0.99	0.66/**1.00**	**1.00**/0.86	0.97/**1.00**	**0.99**/**1.00**	0.88/**1.00**	0.92/0.95	**1.00**/**1.00**
MCPM-GP	0.99/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**	**0.99**/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**	**1.00**/**1.00**
MTPP	0.77/0.77	0.82/0.73	0.86/**1.00**	0.93/**1.00**	0.75/0.83	0.96/0.84	0.78/0.54	0.99/**1.00**	0.66/0.88	0.74/0.95

8.1 Crime data experiments

Here we report the RMSE performances for MCPM and competing models on the CRIME dataset (Tab. 4). In Fig. 5 and Fig. 6 we give the estimated intensities and conditional probabilities for the CRIME complete data experiment. Finally, in Fig. 7 we show the conditional probabilities for the missing data experiment.

8.2 BTB data experiments

In Fig. 8 we show the estimated conditional probabilities on the origin color scale used by Diggle et al. (2013). In Fig. 9 we give the estimated intensity surfaces for the complete data experiment. Finally, in Fig. 10 we show the estimated intensity surfaces for the missing data experiment.

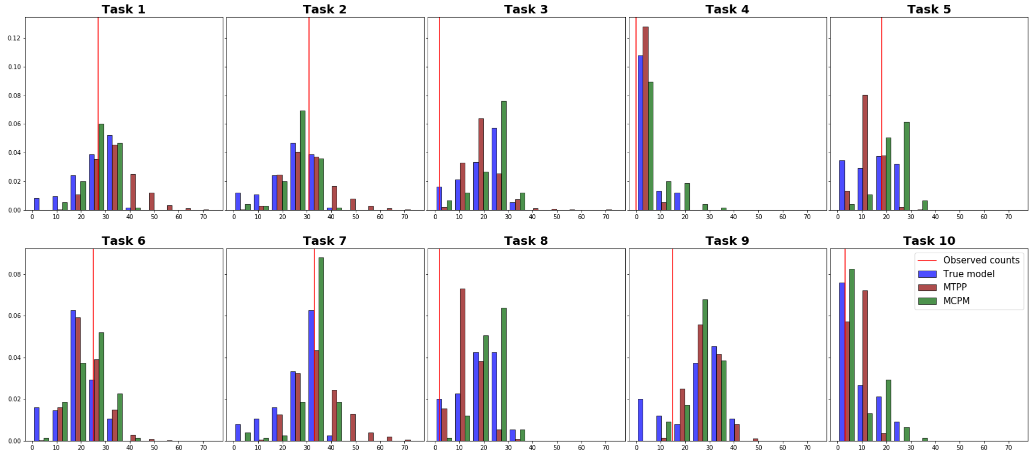

Figure 4: Predicted empirical distributions of event counts in $[80, 100]$ for the second synthetic dataset.

Table 4: CRIME dataset. Performance on the missing regions. Standard errors in parentheses.

	Standardized RMSE						
	1	**2**	**3**	**4**	**5**	**6**	**7**
MCPM	1.74 (0.42)	2.91 (1.06)	**3.00** (1.22)	**2.75** (0.82)	3.57 (1.99)	**11.70** (2.32)	**1.54** (0.29)
MCPM-GP	**1.71** (0.39)	**1.91** (0.33)	3.40 (1.80)	2.96 (1.03)	**2.00** (0.47)	12.18 (2.76)	1.62 (0.33)
LGCP	5.16 (1.81)	4.68 (0.99)	8.93 (5.22)	3.09 (0.50)	7.69 (3.68)	36.96 (5.43)	5.19 (1.21)
ICM	3.36 (1.04)	3.64 (0.83)	3.70 (1.89)	2.97 (1.22)	3.05 (0.97)	12.36 (1.99)	2.82 (0.62)

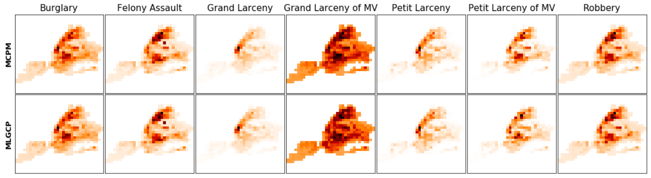

Figure 5: CRIME dataset. Estimated intensity surface with MCPM (*first row*) and MLGCP (*second row*). The color scale used for each crime is given in Fig. (5).

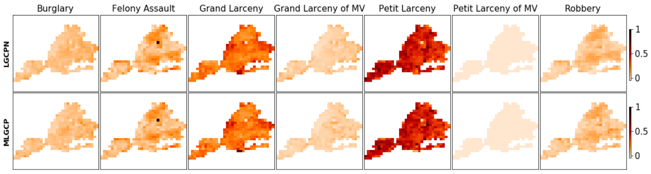

Figure 6: CRIME dataset. Estimated conditional probabilities in the complete data setting. *Row 1:* MCPM *Row 2:* MLGCP.

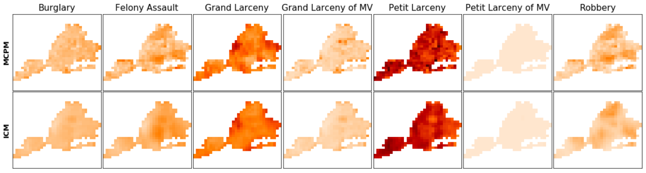

Figure 7: CRIME dataset. Estimated conditional probabilities when introducing missing data regions. *Row 1:* MCPM *Row 2:* LGCP.

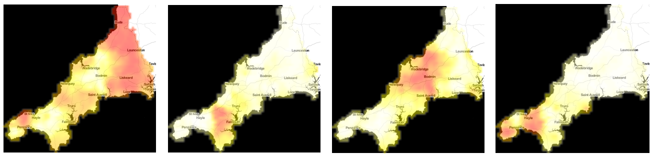

Figure 8: MLGCP- BTB dataset. Estimated conditional probabilities plotted on the color scale used by Diggle et al. (2013) and Taylor et al. (2015). The first plots corresponds to GT 9, the second to GT 12, the third to GT 15 and the fourth to GT 20.

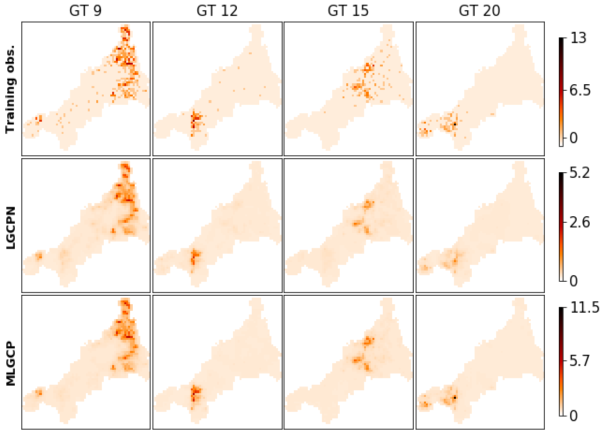

Figure 9: Estimated intensity surfaces in the complete data setting. *First row:* Training data. *Second row:* MCPM *Third row:* MLGCP

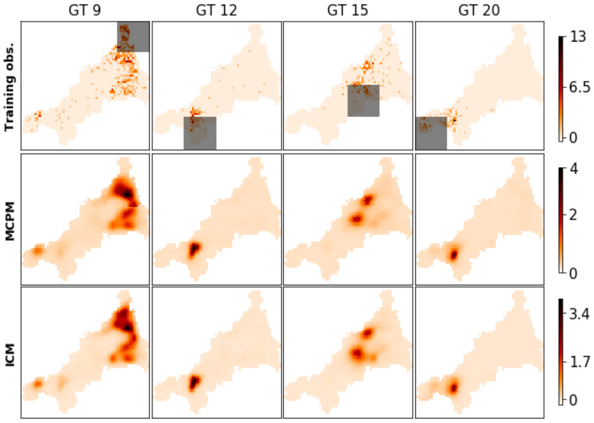

Figure 10: Estimated intensity surfaces in the missing data (shaded regions) setting. *First row:* Training data. *Second row:* MCPM *Third row:* ICM

References

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Rakitsch, B., Lippert, C., Borgwardt, K., and Stegle, O. (2013). It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. *Neural Information Processing Systems.*

Seijas-Macías, A. and Oliveira, A. (2012). An approach to distribution of the product of two normal variables. *Discussiones Mathematicae Probability and Statistics*, 32(1-2):87–99.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in r. *Journal of Statistical Software*, 63:1–48.

Titsias, M. K. (2009). Variational learning of inducing variables in sparse gaussian processes. *Artificial Intelligence and Statistics*, 5:567–574.