A. Supplementary Material

Accompanying the submission *Fast Bayesian Intensity Estimation for the Permanental Process*.

A.1. Exact Expected Log Loss

We evaluate our estimated $\hat{\lambda}$ using the expectation under the true $\mathrm{PP}(\lambda)$ of the log likelihood under $\mathrm{PP}(\hat{\lambda})$, where PP is the Poisson process. Adams et al. (2009) approximate this quantity using Monte Carlo, employing numerical integration for (1). It turns out that for the computational cost of one such numerical integration, we may compute the expected loss using standard results for Lévy processes (Cont & Tankov, 2004). An elementary self contained argument runs as follows:

$$\begin{aligned}\mathbb{E}_{X\sim \mathrm{PP}(\lambda)}\left[\log p_{X\sim \mathrm{PP}(\hat{\lambda})}(X)\right] &= \mathbb{E}_{\mathrm{card}(X)}\left[\mathbb{E}_{X\sim \mathrm{PP}(\lambda)|\,\mathrm{card}(X)}\left[\log p_{X\sim \mathrm{PP}(\hat{\lambda})}(X)\right]\right]\\ &= \mathbb{E}_{\mathrm{card}(X)}\left[\mathrm{card}(X)\left(\log\hat{\Lambda}(\Omega)+H(\lambda,\hat{\lambda})\right)-\hat{\Lambda}(\Omega)\right]\\ &= \Gamma(\Omega)\left(\log\hat{\Lambda}(\Omega)+H(\lambda,\hat{\lambda})\right)-\hat{\Lambda}(\Omega)\\ &= \int_{\boldsymbol{x}\in\Omega}\left(\lambda(\boldsymbol{x})\log\hat{\lambda}(\boldsymbol{x})-\hat{\lambda}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x},\end{aligned}$$

where Ω is the sampling domain, $H(\lambda,\hat{\lambda}) \coloneqq \int_{\boldsymbol{x}\in\Omega}\frac{\lambda(\boldsymbol{x})}{\Lambda(\Omega)}\log\frac{\hat{\lambda}(\boldsymbol{x})}{\hat{\Lambda}(\Omega)}\,\mathrm{d}\boldsymbol{x}$ is the cross-entropy between the probability density functions proportional to λ and $\hat{\lambda}$ and we recall $\Lambda(S) \coloneqq \int_{x\in S}\lambda(x)\,\mathrm{d}x$. The first line is the tower law of expectation. To see the second line, note that we may sample $X\sim\mathrm{PP}(\lambda)$ by first sampling $\mathrm{card}(X)\sim\mathrm{Poisson}(\Lambda(\Omega))$, and then drawing each element of X according to the probability density proportional to λ. The third line uses the Poisson expectation $\mathbb{E}_{\mathrm{card}(X)}\left[\mathrm{card}(X)\right]=\Gamma(\Omega)$ and the fourth some simple algebra.

As an aside, we may therefore write the Kullback-Leibler divergence in a form resembling that for probability distributions:

$$\begin{aligned}D_{\mathrm{KL}}\left(\mathrm{PP}(f)\middle\|\,\mathrm{PP}(g)\right) &= \mathbb{E}_{X\sim \mathrm{PP}(\lambda)}\left[\log p_{X\sim \mathrm{PP}(\lambda)}(X)-\log p_{X\sim \mathrm{PP}(\hat{\lambda})}(X)\right]\\ &= \int_{\boldsymbol{x}\in\Omega}\left(f(\boldsymbol{x})\log\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}+g(\boldsymbol{x})-f(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

A.2. Bayesian Decision Theory for the Expected Log Loss

To determine the intensity function which maximises the expected log likelihood we define the loss

$$\ell(\lambda,\lambda') \coloneqq \mathbb{E}_{n_i\sim N(B_i),i=1,2,\ldots,m|\lambda}\log p\left(N(B_i)=n_i, i=1,2,\ldots,m|\lambda'\right)$$

where $N(B_i)$ is the random variable representing the number of points in the set $B_i\subseteq\Omega$, Ω is the domain of the process and we recall $\Lambda(S)\coloneqq\int_{x\in S}\lambda(x)\,\mathrm{d}x$. It is well known that (Baddeley, 2007)

$$p(N(B_i)=n_i, i=1,2,\ldots,m|\lambda)=\prod_i\frac{\Lambda(B_i)^{n_i}}{n_i!}\exp(-\Lambda(\Omega)).$$

Bayesian decision theory considers the expected loss

$$L(\lambda')\coloneqq\mathbb{E}_{\lambda|D}\left[\ell(\lambda,\lambda')\right],$$

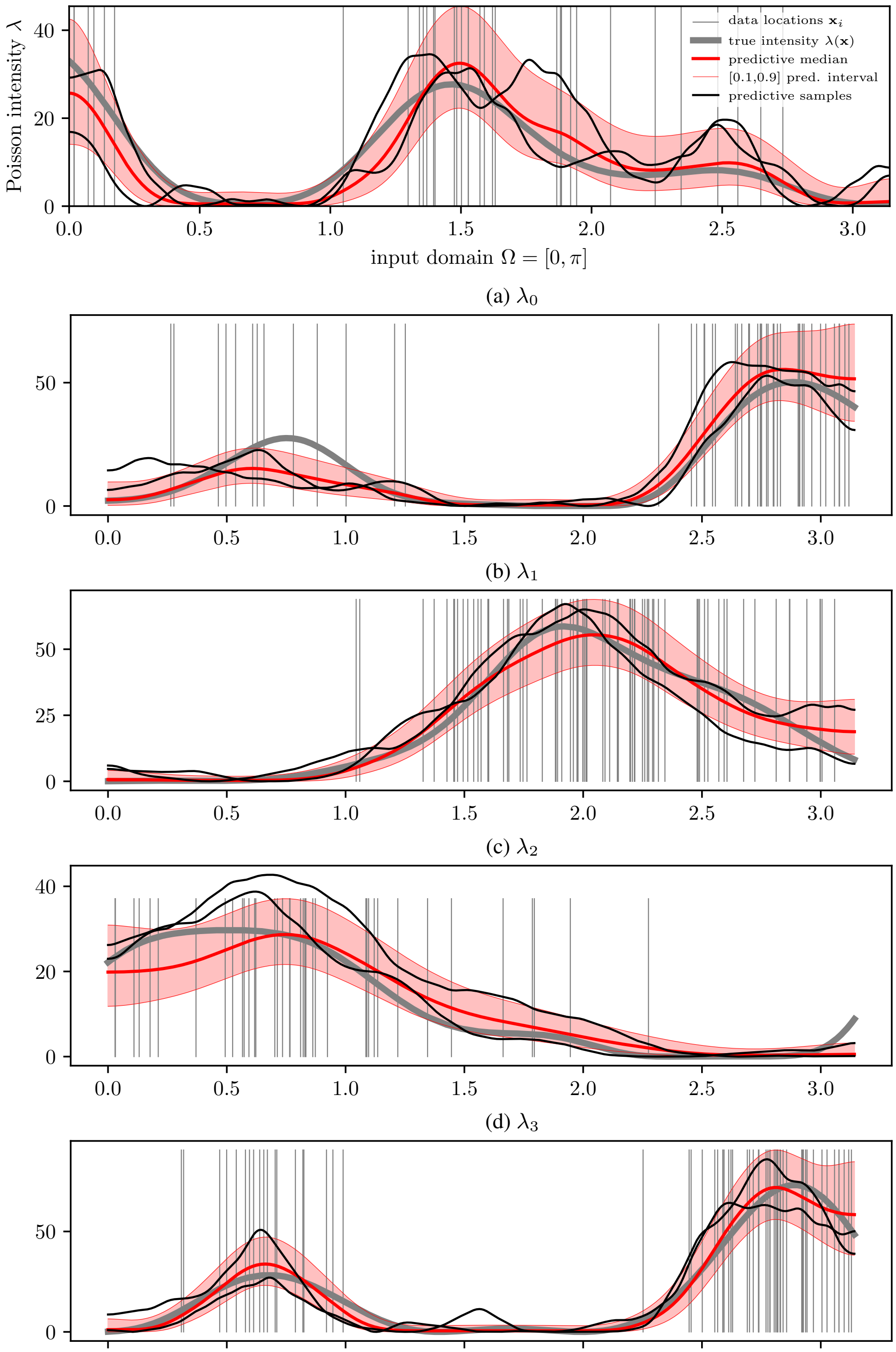
where the expectation is with respect to the posterior predictive distribution given the data D. Combining these expressions and assuming without loss of generality that $\Omega=\bigcup_i B_i$ yields

$$L(\lambda')=\mathbb{E}_{\lambda|D}\left[\mathbb{E}_{n_i\sim N(B_i),i=1,2,\ldots,m|\lambda}\left[\sum_i\left(n_i\log\Lambda'(B_i)-\log(n_i!)-\Lambda'(B_i)\right)\right]\right].$$

The optimal choice is $\Lambda^*\coloneqq\mathrm{argmax}_{\lambda'}L(\lambda')$, so by stationarity

$$\begin{aligned}\lambda^*(B_i) &= \mathbb{E}_{\lambda|D}\left[\mathbb{E}_{n_i\sim N(B_i)|\lambda}\left[n_i\right]\right]\\ &= \mathbb{E}_{\lambda|D}\left[\Lambda(B_i)\right],\end{aligned}$$

and so $\lambda^*=\mathbb{E}_{\lambda|D}[\lambda]$, the expectation of the posterior predictive distribution.



(a) λ_0

(b) λ_1

(c) λ_2

(d) λ_3

(e) λ_4

Figure 6. Predictive distributions for the test problems of subsection 6.2.

A.3. Standard Laplace Approximations for the GP

Following *e.g.* (Rasmussen & Williams, 2006), assume that we are given an independent and identically distributed sample $\{(\boldsymbol{x}_i, y_i)\}_{1\le i\le m}$, and the goal is to estimate $p(y|\boldsymbol{x})$. Let the true joint in $\boldsymbol{f} = (f(\boldsymbol{x}_i))_i, \boldsymbol{y} = (y(\boldsymbol{x}_i))_i$ be

$$\begin{aligned}\log p(\boldsymbol{y}, \boldsymbol{f}|X, k) &= \log p(\boldsymbol{y}|\boldsymbol{f}) + \log p(\boldsymbol{f}|X, k)\\ &= \log p(\boldsymbol{y}|\boldsymbol{f}) - \frac{1}{2}\boldsymbol{f}^\top K^{-1}\boldsymbol{f} - \frac{1}{2}\log|K| - \frac{m}{2}\log 2\pi,\end{aligned}$$

where $K = (k(\boldsymbol{x}_i), \boldsymbol{x}_j)_{ij}$ and $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m)$. The Laplace approximation fits a normal to the posterior,

$$\begin{aligned}\log p(\boldsymbol{f}|\boldsymbol{y}, X) &\approx \log \mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}}, Q)\\ &= -\frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^\top Q^{-1}(\boldsymbol{f} - \hat{\boldsymbol{f}}) - \frac{1}{2}\log|Q| - \frac{m}{2}\log 2\pi\\ &:= \log q(\boldsymbol{f}|\boldsymbol{y}, X).\end{aligned}$$

$\hat{\boldsymbol{f}}$ and Q come from a second order approximation of the log posterior at its mode, *i.e.*

$$\begin{aligned}\hat{\boldsymbol{f}} &= \operatorname*{argmax}_{\boldsymbol{f}} p(\boldsymbol{y}|\boldsymbol{f}, X)\\ &= \operatorname*{argmax}_{\boldsymbol{f}} p(\boldsymbol{y}, \boldsymbol{f}|X)\\ Q^{-1} &= -\left.\frac{\partial^2}{\partial\boldsymbol{f}\partial\boldsymbol{f}^\top}\log p(\boldsymbol{y}, \boldsymbol{f}|X)\right|_{\boldsymbol{f}=\hat{\boldsymbol{f}}}\\ &= K^{-1} + W\\ W_{ii} &= -\left.\frac{\partial^2}{\partial f_i^2}\log p(y_i|f_i)\right|_{f_i=\hat{f}_i}\end{aligned}$$

Taylor expanding $\log p(\boldsymbol{y}, \boldsymbol{f}|X)$ at $\boldsymbol{f} = \hat{\boldsymbol{f}}$,

$$\begin{aligned}\log p(\boldsymbol{y}, \boldsymbol{f}|X) &\approx \log p(\boldsymbol{y}, \hat{\boldsymbol{f}}|X) - \frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^\top Q^{-1}(\boldsymbol{f} - \hat{\boldsymbol{f}})\\ &= \log p(\boldsymbol{y}|\boldsymbol{f} = \hat{\boldsymbol{f}}) - \frac{1}{2}\hat{\boldsymbol{f}}^\top K^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|K| - \frac{m}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^\top Q^{-1}(\boldsymbol{f} - \hat{\boldsymbol{f}})\\ &:= \log q(\boldsymbol{y}, \boldsymbol{f}|X)\end{aligned} \tag{17}$$

Now

$$\log\int \exp(-\frac{1}{2}\boldsymbol{x}^\top H^{-1}\boldsymbol{x})d\boldsymbol{x} = \frac{m}{2}\log 2\pi + \frac{1}{2}\log|H|$$

So we get the approximate marginal likelihood

$$\begin{aligned}\log Z &:= \log p(\boldsymbol{y}|X)\\ &\approx \log\int q(\boldsymbol{y}, \boldsymbol{f}|X)d\boldsymbol{f}\\ &= \log p(\boldsymbol{y}|\boldsymbol{f} = \hat{\boldsymbol{f}}) - \frac{1}{2}\hat{\boldsymbol{f}}^\top K^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|K| - \frac{1}{2}\log|K^{-1} + W|\\ &= \log p(\boldsymbol{y}|\boldsymbol{f} = \hat{\boldsymbol{f}}) - \frac{1}{2}\hat{\boldsymbol{f}}^\top K^{-1}\hat{\boldsymbol{f}} - \frac{1}{2}\log|I + KW|\end{aligned} \tag{18}$$

This is a standard textbook approach (Rasmussen & Williams, 2006), but we can get the same approximation via

$$\log p(\boldsymbol{y}|X) \approx \log q(\boldsymbol{y}, \hat{\boldsymbol{f}}|X) - \log q(\hat{\boldsymbol{f}}|\boldsymbol{y}, X), \tag{19}$$

since the right hand side is true for all $\boldsymbol{f}$, not just $\hat{\boldsymbol{f}}$. Hence we need only subtract the approximate log likelihoods as above. By evaluating at $\hat{\boldsymbol{f}}$, the second r.h.s. term in (17), vanishes immediately.