Journal of Classification (2021) 38:188–211
https://doi.org/10.1007/s00357-020-09363-4

Model-based Clustering of Count Processes

Tin Lok James Ng[1] · Thomas Brendan Murphy[2]

Published online: 2 July 2020

Abstract
A model-based clustering method based on Gaussian Cox process is proposed to address the problem of clustering of count process data. The model allows for nonparametric estimation of intensity functions of Poisson processes, while simultaneous clustering count process observations. A logistic Gaussian process transformation is imposed on the intensity functions to enforce smoothness. Maximum likelihood parameter estimation is carried out via the EM algorithm, while model selection is addressed using a cross-validated likelihood approach. The proposed model and methodology are applied to two datasets.

Keywords Count process · Clustering · Gaussian process · Gaussian Cox process · Mixture models

1 Introduction

Model-based clustering techniques (Fraley and Raftery 2002; Bouveyron and Brunet-Saumard 2014; McNicholas 2016; Bouveyron et al. 2019) have been widely used in many applications where sample observations consist of multivariate data taking values in Euclidean space. In this approach, it is often assumed that the observations arise from a finite mixture distribution that is a mixture of two or more components, where each component is a probability density function and each component has an associated probability. Recently, clustering techniques for functional data (Abraham et al. 2003; Giacofci et al. 2013; Jacques and Preda 2014) have been proposed where each observation is a stochastic process taking values in an infinite dimensional space.

In some applications, one observes a collection of count processes as data, where each count process consists of event times observed over a fixed time interval. Some examples include the occurrence of natural disasters in various locations, observed times of failure of

✉ Tin Lok James Ng
jamesng@uow.edu.au

Thomas Brendan Murphy
brendan.murphy@ucd.ie

[1] School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

[2] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

multiple electronic components, and temporal sequence of action potentials generated by neurons. Compared with clustering finite dimensional observations and functional observations, clustering techniques for count process data are underdeveloped. In this paper, we propose a mixture Poisson processes model for count process data observed in a fixed time interval $[0, T)$, where each mixture component g is a non-homogeneous Poisson process governed by an intensity function $\lambda_g(t)$, $t \in [0, T)$.

The proposed model is an extension of the Poisson mixture model proposed in Côme and Latifa (2014), which was used to analyze the Paris bike-sharing system. In Côme and Latifa (2014), event time points for bikes arriving and departing from each station are obtained, and the observations are aggregated at 1-h intervals to produce the counts. A generative model based on Poisson mixtures was proposed and an EM algorithm was used to estimate the intensities and clustering of bike stations. Our approach incorporates a Gaussian process prior on the intensity functions $\{\lambda_g(t)\}_{g=1}^{G}$, where G is the number of clusters or intensity functions. The Gaussian process prior can enhance the smoothness of the resulting intensity estimates. We adopt the EM algorithm to cluster the count processes and estimate the cluster-specific intensity functions. Standard errors of the estimated intensity functions are estimated using a jackknife approach, and a cross-validated likelihood approach (Smyth 2000) is used to perform model selection.

The rest of this article is organized as follows. In Section 2, we briefly review the definitions of Poisson processes and Gaussian processes, and the relevant literature on Gaussian process modulated Poisson processes. Section 3 presents the mixture of Poisson processes model and develops an EM algorithm for intensity function estimation and clustering; standard error estimates for the estimated intensity functions and model selection are also considered. Two datasets are analyzed in Section 5 by the proposed methodology. The paper concludes in Section 6 with a discussion of the proposed modeling approach.

2 Preliminaries

Definitions of Poisson processes and Gaussian processes are stated in this section in order to motivate our proposed model.

2.1 Poisson Process

Non-homogeneous Poisson process (NHPP) is a popular model for count process data. A temporal NHPP defined on an interval $[0, T) \subset \mathbb{R}$ is associated with a non-negative and locally integrable intensity function $\lambda(t)$ for $t \in [0, T)$. That is, for any bounded region $B \subset [0, T)$, the volume $\Lambda(B) = \int_B \lambda(s)ds$ is finite. Furthermore, let $N(B)$ be the number of events in B, we have:

1. $N(B)$ follows a Poisson distribution with rate $\Lambda(B)$.
2. Given $N(B)$, the location of events within B are i.i.d. with density $\lambda(t)/\Lambda(B)$.

2.2 Gaussian Process

A random scalar function $g(s) : \mathbb{S} \to \mathbb{R}$ is said to have a Gaussian process prior, if for any finite collection of points $\{s_n\}_{n=1}^{N} \in \mathbb{S}$, the function values $\{g(s_n)\}_n^N$ follow a multivariate

Gaussian distribution. The Gaussian process prior can be defined by a mean function $m(\cdot) : \mathbb{S} \rightarrow \mathbb{R}$ and a covariance function $k(\cdot, \cdot) : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}$; the mean and covariance functions further depend on hyperparameters ϕ.

The idea of Gaussian process modeling is to place a prior directly on the space of functions without parameterizing the random function. A Gaussian process approach provides a convenient way to incorporate dependence structure for points in the space $\mathbb{S}$ where the dependence between two points is typically determined by their distance and orientation. Rasmussen and Williams (2006) contains a detailed review of theory and applications of Gaussian process.

One attractive feature of Gaussian processes is the variety of covariance functions one can choose from, which leads to different level of smoothness of the underlying random function to be modeled. Therefore, prior knowledge and specifications about the shape of the underlying function can be incorporated by selecting different covariance functions. Gaussian process modeling has been successfully applied to regression (Williams and Rasmussen 1996), classification (Kim and Ghahramani 2006), density estimation (Murray et al. 2009), and point process modeling (Adams et al. 2009) problems.

2.3 Gaussian Cox Process

The combination of a Poisson process and a Gaussian process prior is known as a Gaussian Cox process. It provides an attractive modeling framework to infer the underlying intensity function since one only needs to specify the form of the Gaussian process mean and covariance functions. This approach has been adopted in various applications, including neuroscience (Cunningham et al. 2008), finance (Basu and Dassios 2002), and forestry (Heikkinen and Arjas 1999).

3 Model and Methodology

3.1 Mixture of Poisson Processes

We assume that there are G normalized intensity functions $\lambda = \{\lambda_g\}_{g=1}^{G}$, where $\lambda_g : [0, T) \rightarrow \mathbb{R}^+$, $\int_0^T \lambda_g(s)ds = 1$, and we let τ_g be the probability that a point process has intensity function λ_g, for $g = 1, 2, \ldots, G$. We employ the logistic density transform to the intensity functions $\{\lambda_g\}_{g=1}^{G}$:

$$\lambda_g(t) = \frac{\exp(f_g(t))}{\int_0^T \exp(f_g(s))ds}. \tag{1}$$

Logistic Gaussian process priors have been proposed for Bayesian nonparametric density estimation where theoretical properties have been investigated (Leonard 1978; Lenk 1988, 1991; Tokdar and Ghosh 2007).

A zero-mean Gaussian process prior $GP(0, k(s, s'))$ is assigned to the random functions $f = \{f_g\}_{g=1}^{G}$, where $k(s, s')$ is the covariance function. The covariance function defines the nearness or similarity of input points, and it is a basic assumption that input points that are close are likely to have similar values of intensity. It is common to assume that the covariance function is stationary, that is, it is a function of the $x - x'$. In this paper, we assume the one-dimensional squared-exponential covariance function, which has the following form:

$$k(s, s') = \sigma^2 \exp\left[-\frac{1}{2l^2}(s - s')^2\right], \tag{2}$$

for two input points s and s'. The hyperparameters $\phi = (l, \sigma)$ determine the properties of the covariance function. In particular, l determines the smoothness of the covariance function, and as l increases, the covariance between two input points increases. The incorporation of a Gaussian process prior increases the amount of smoothness in the estimated intensity functions.

Assuming that the intensity function is a transformation of random realization from a Gaussian process provides a convenient way to specify prior beliefs about the intensity function without choosing a particular functional form. Unfortunately, the likelihood involves an integral of an infinite-dimensional random function which is computationally intractable. Various inference methods for Gaussian process have been proposed including Markov Chain Monte Carlo (MCMC) (Adams et al. 2009) and variational Bayesian approach (Lloyd et al. 2015).

To make inference more tractable, we further discretize the interval $[0, T)$ into m equal length sub-intervals, and let $\{t_k\}_{k=1}^m$ be the mid-points of the sub-intervals $[T(k-1)/m, Tk/m)$, $k = 1, \cdots, m$. We assume that the random functions $\{f_g\}_{g=1}^G$; hence, the intensity functions $\{\lambda_g\}_{g=1}^G$ are piecewise constant on each interval. Hence, we have for any $t \in [0, T)$:

$$f_g(t) = \sum_{k=1}^{m} f_g(t_k) I\left\{t \in [T(k-1)/m, Tk/m)\right\},$$

where $I(\cdot)$ is the indicator function. To simplify the notation, for each $g = 1, \cdots, G$, we let $\mathbf{f}_g = (f_{g,k})_{k=1}^m$ denote the vector of function evaluations at the m mid-points $\{t_k\}_{k=1}^m$; that is, $f_{g,k} = f_g(t_k)$ for $k = 1, \cdots, m$. The Gaussian process prior (1) results in a Gaussian distribution for $\mathbf{f}_g$:

$$p(\mathbf{f}_g | \{t_k\}_{k=1}^m, \theta) = \mathbb{N}(\mathbf{f}_g | 0, K) \tag{3}$$

where K is an $m \times m$ covariance matrix that depends on the mid-points $\{t_k\}_{k=1}^m$.

We note that when the length scale parameter $l \to 0$ in the squared exponential covariance function (2), the covariance matrix K converges to $\sigma^2 I_m$, and the prior distribution for $\mathbf{f}_g$ tends to:

$$p(\mathbf{f}_g | \{t_k\}_{k=1}^m, \theta) = \mathbb{N}(\mathbf{f}_g | 0, \sigma^2 I_m), \tag{4}$$

where I_m is the $m \times m$ identity matrix. In this limiting case, each $f_{g,k}$ has a normal prior with variance σ^2, but $f_{g,k}$ and $f_{g,k'}$ are independent for $k \neq k'$.

This discretization leads to efficient inference of intensity functions since the resulting computation is independent of the size of data. In contrast, the cost of each MCMC step in Adams et al. (2009) scales cubically in the size of data.

After the discretization, for any $t \in [0, T)$, the intensity function evaluated at t can be expressed as:

$$\begin{aligned} \lambda_g(t) &= \frac{\exp(f_g(t))}{\frac{1}{m}\sum_{k=1}^m \exp(f_g(t_k))} \\ &= \frac{\exp(f_g(t))}{\frac{1}{m}\sum_{k=1}^m \exp(f_{g,k})}. \end{aligned} \tag{5}$$

To take into account the variation in overall intensities between individual count process data, we introduce scaling factors $\alpha = \{\alpha_i\}_{i=1}^{N}$ so that the count process i with associated normalized intensity λ_g has overall intensity $\alpha_i \lambda_g$.

The choice of hyperparameters ϕ represents the prior belief of smoothness of the intensity function. We adopt a cross-validated likelihood approach along with a grid search (Section 3.5) to determine ϕ. The number of sub-intervals m for the discretization could also be determined using the cross-validated likelihood approach, although doing so would significantly increase the model search space and computational time. A practical approach is to start with a small number of sub-intervals m, and gradually increase m until the desired level of resolution is obtained. The effect of choosing different m on the estimated intensity functions and classification accuracy of the count process observations will be examined in Section 4.

It is possible to simulate N count processes from the model where each count process i consists of a vector of event times $\{x_{i,j}\}_{j=1}^{n_i}$ and n_i is the number of events. We first obtain G random functions $\{f_g\}_{g=1}^{G}$ evaluated at the m mid-points by drawing G random variables from the multivariate Gaussian distribution given in (3). The logistic transformation (5) allows us to obtain the G intensity functions $\{\lambda_g\}_{g=1}^{G}$. We then draw from a multinomial distribution $z_i \sim \mathbb{M}(1, \{\tau_g\}_{g=1}^{G})$ to determine the intensity function λ_g from which the observations of count process i are simulated. We then obtain the overall intensity function $\alpha_i \lambda_{z_i}$ associated with the count process i. Various techniques for non-homogeneous process simulation can then be applied to draw observations $\{x_{i,j}\}_{j=1}^{n_i}$ in time interval $[0, T)$. The generative process of the model is summarized in Algorithm 1.

Algorithm 1 Mixture of poisson processes simulation
Input: $\{\alpha_i\}_{i=1}^{N}$, $\{\tau_g\}_{g=1}^{G}$, θ, T, m,
Output: $\{\{x_{i,j}\}_{j=1}^{n_i}\}_{i=1}^{N}$

```
for $g = 1, \cdots, G$ do
    Draw random function $f_g$ according to (3)
    Obtain intensity function $\lambda_g$ according to (5)
end for
for $i = 1, \cdots, N$ do
    Draw cluster label for count process $z_i \sim \mathbb{M}(1, \{\tau_g\}_{g=1}^{G})$
    Draw observations $\{x_{ij}\}_{j=1}^{n_i}$ in time interval $[0, T)$ according to a non homogeneous Poisson process with intensity $\alpha_i \lambda_{z_i}$.
end for
```

3.2 Consistency of Penalized MLE Under Mixture of Poisson Processes

The incorporation of a Gaussian process prior can be understood as a roughness penalty on the intensity functions, where intensity functions that are less smooth receive a larger penalty. The penalized maximum likelihood approach (MLE) is commonly adopted in mixture modelling (Ciuperca et al. 2003; Fraley and Raftery 2007; Chen et al. 2008). For finite normal mixtures, the penalized MLE approach is used to tackle the likelihood degeneracy problem (Fraley and Raftery 2007; Chen et al. 2008) where penalties are imposed on the variances of mixture components. In our case, the Gaussian process prior penalty ensures that the resulting estimated intensity functions are smooth.

The study of consistency properties of the MLE of finite mixture models has attracted significant research interests (Cheng and Liu 2001; Chen et al. 2016; Chen 2017). In this section, we derive the consistency property of the penalized MLE of mixture of Poisson processes model. We let $x = \{x_i\}_{i=1}^{N}$ be the observed event times of N count processes, where $x_i = \{x_{i,j}\}_{j=1}^{n_i}$ are the event times of process i. We can alternatively represent x_i as $y_i = (y_{i,1}, \dots, y_{i,m})^T$ where $y_{i,k}$ denotes the number of events in the kth interval for the ith count process. These two representations are equivalent when m is assumed to be fixed. Let $\mathcal{F} = \{\mathbf{f} \in \mathbb{R}^m\}$ be the parameter space of each mixture component. The penalized density of y_i is given by:

$$h(y_i; \alpha_i, P_{\mathcal{F}}) = \sum_{g=1}^{G} \tau_g h(y_i; \alpha_i, \mathbf{f}_g), \tag{6}$$

where

$$h(y_i; \alpha_i, \mathbf{f}_g) = \left(\exp(-\alpha_i) \prod_{k=1}^{m} \frac{(\alpha_i \lambda_k^{(g)})^{y_{i,k}}}{y_{i,k}!} \right) p(\mathbf{f}_g) \tag{7}$$

is the penalized density for the gth mixture component, and $P_{\mathcal{F}}$ is the mixing distribution. To suppress notation, we write P for $P_{\mathcal{F}}$ when no confusion arises. We define $\mathbb{P}$ to be the space which consists of all mixing distributions, and for two mixing distributions P_1 and P_2 in $\mathbb{P}$, we define the distance $D(\cdot, \cdot)$

$$D(P_1, P_2) = \int_{\mathcal{F}} |P_1(\mathbf{f}) - P_2(\mathbf{f})| \exp(-|\mathbf{f}|) d\mathbf{f}$$

where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^m$. We say that $P \to P_0$ if $D(P, P_0) \to 0$. Suppose $P^* \in \mathbb{P}$ is the true mixing distribution and $\hat{P}$ is an estimator, then $\hat{P}$ is strongly consistent when $D(\hat{P}, P^*) \to 0$ almost surely.

A necessary condition for the consistency of $\hat{P}$ is the identifiability of the mixture model. The following proposition on identifiability is a direct consequence of Theorem 4.2 of Sapatinas (1995) where the identifiability of arbitrary multivariate power-series mixtures is proved.

Proposition 3.1 *The mixture model defined in* Eqs. 6 *and* 7 *is identifiable. That is, let $H(y_i; \alpha_i, P)$ be the cumulative distribution function of $h(y_i; \alpha_i, P)$. If for any α_i, $H(y_i; \alpha_i, P) = H(y_i; \alpha_i, P^*)$ for all y_i, then $D(P, P^*) = 0$.*

In particular, the mixture model (6) can be formulated in the form of Equation 4.1 of Sapatinas (1995), and since a multivariate Poisson mixture model has infinite divisible univariate marginals, the mixture model (6) is identifiable by Theorem 4.2 of Sapatinas (1995).

Consistency of MLE is established in the following result where the proof is deferred to Appendix A.

Proposition 3.2 *Let $\hat{P}$ be the MLE of mixture of Poisson processes model, then $\hat{P}$ is strongly consistent.*

3.3 EM Algorithm

Given observed event times of N count processes, we want to group the count procesess into G clusters while simultaneously estimating the intensity functions of the G Poisson process. As we are unable to observe the true cluster assignment for each count process, we adopt the EM algorithm (Dempster et al. 1977) for maximum likelihood estimation.

Let $\mathbf{f} = (\mathbf{f}_1, \cdots, \mathbf{f}_G)^T$ be the vector of multivariate Gaussian random variable with distribution defined in Eq. 3, the (penalized) likelihood function for the observations with G clusters can be written as:

$$L(\theta; x) = \prod_{i=1}^{n} \left[\sum_{g=1}^{G} \tau_g \exp(-\alpha_i) \prod_{j=1}^{n_i} \lambda_g(x_{i,j}) \right] \left[\prod_{g=1}^{G} p(\mathbf{f}_g) \right] \tag{8}$$

where $p(\mathbf{f}_g)$ is given in Eq. 3, and $\theta = (\alpha, \tau, \mathbf{f})$ are the model parameters. In non-Bayesian setting, the Gaussian process prior imposed on random functions $\{f_g\}_{g=1}^{G}$ can be understood as penalty terms. The maximum penalized likelihood estimates of $\{f_g\}_{g=1}^{G}$ are equivalent to *maximum a posterior* estimates under the Bayesian framework.

We let $z = \{z_i\}_{i=1}^{n}$ be the latent assignment of observations (count processes) to clusters, where $z_i = (z_{i1}, z_{i2}, \cdots, z_{iG})^T$ and

$$z_{ig} = \begin{cases} 1 & \text{if count process } i \text{ has intensity function } \alpha_i \lambda_g \\ 0 & \text{otherwise} \end{cases} .$$

By introducing the latent variables z, we can write the complete data likelihood function as:

$$L(\theta; x, z) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\tau_g \exp(-\alpha_i) \prod_{j=1}^{n_i} (\alpha_i \lambda_g(x_{i,j})) \right]^{z_{ig}} \left[\prod_{g=1}^{G} p(\mathbf{f}_g) \right] \tag{9}$$

Taking the logarithm of Eq. 9 gives the complete data log-likelihood function:

$$\begin{aligned} \log L(\theta; x, z) = & \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[\log(\tau_g) - \alpha_i + \sum_{j=1}^{n_i} \left\{ \log(\alpha_i) + \log(\lambda_g(x_{ij})) \right\} \right] \\ & + \sum_{g=1}^{G} \log(p(\mathbf{f}_g)) \end{aligned} \tag{10}$$

During the E-step, we need to compute the probability of cluster assignment for each count process $\hat{\pi}_i^g \equiv E(Z_{ig} | x_i, \hat{\theta}^{(t)})$ conditional on current parameter estimates $\hat{\theta}^{(t)} = (\hat{\alpha}^{(t)}, \hat{\tau}^{(t)}, \hat{\mathbf{f}}^{(t)})$. We have:

$$\begin{aligned} \hat{\pi}_i^g &= p\{Z_{ig} = 1 | x_i, \hat{\theta}^{(t)}\} \\ &= \frac{\hat{\tau}_g^{(t)} \exp(-\hat{\alpha}_i^{(t)}) \prod_{j=1}^{n_i} (\hat{\alpha}_i^{(t)} \hat{\lambda}_g^{(t)}(x_{i,j}))}{\sum_{h=1}^{G} \hat{\tau}_h^{(t)} \exp(-\hat{\alpha}_i^{(t)}) \prod_{j=1}^{n_i} (\hat{\alpha}_i^{(t)} \hat{\lambda}_h^{(t)}(x_{i,j}))} \end{aligned} \tag{11}$$

During the M-step, we need to maximize the conditional expectation of the complete data log-likelihood given in Eq. 9, $E_{Z|x,\hat{\theta}^{(t)}}(\log L(\theta; x, Z))$ with respect to the conditional distribution of Z given the model parameters $\hat{\theta}^{(t)}$. The updates for $\alpha = \{\alpha_i\}_{i=1}^N$ and $\tau = \{\tau\}_{g=1}^G$ can be derived analytically and are given below.

$$\hat{\alpha}_i = n_i \tag{12}$$

$$\hat{\tau}_g = \frac{\sum_{i=1}^n \hat{\pi}_i^g}{n} \tag{13}$$

The update formula for scaling factor α_i does not depend on $\hat{\pi}_i^g$; hence, it only needs to be computed once.

The optimization of the conditional expected complete data log-likelihood with respect to $\{\mathbf{f}_g\}_{g=1}^G$ does not result in explicit update rule for $\{\mathbf{f}_g\}_{g=1}^G$. As a result, we employ Newton's method to obtain estimates for $\{\mathbf{f}_g\}_{g=1}^G$. The EM algorithm is summarized in Algorithm 2, and its full derivation is given in Appendix B. With the discretization of the time domain, the computational complexity of the proposed algorithm is independent of the number of events in each count process observations. Since both the grid centers and hyperparameters are fixed, the inverse of the covariance matrix can be pre-computed.

Algorithm 2 EM Algorithm
Input: G, x, θ
Output: Estimated parameters $(\hat{\alpha}, \hat{\tau}, \hat{\lambda})$, a Posterior cluster assignment probabilities $\{\{\hat{\pi}_i^g\}_{g=1}^G\}_{i=1}^N$

```
for $i = 1, \cdots, N$ do
    Set $\hat{\alpha}_i = n_i$
end for
Initialise $\{\hat{\tau}_g\}_{g=1}^G$
do
    E Step
    for $i = 1, \cdots, N$ do
        for $g = 1, \cdots, G$ do
            Compute a Posterior probability $\hat{\pi}_i^g$ according to Eq. 11
        end for
    end for

    M Step
    for $g = 1, \cdots, G$ do
        Compute $\hat{\tau}_g$ according to Eq. 13
    end for
    for $g = 1, \cdots, G$ do
        Compute $\hat{\mathbf{f}}_g$ by Newton's method
        Compute $\hat{\lambda}_g$ by Eq. 5
    end for
while Not Converged
```

3.4 Uncertainty Estimates of Intensity Function

It is often desirable to obtain standard errors and construct confidence intervals for the model parameters (α, τ, λ). The primary interest is on the intensity functions $\lambda = \{\lambda_g\}_{g=1}^G$; we propose the jackknife technique to achieve this (e.g., O'Hagan et al. (2019)).

For each intensity function λ_g for $g = 1, 2, \ldots, G$ and $i = 1, 2, \ldots, N$, we obtain the estimates $\hat{\lambda}_g^{(i)}$ for using the EM algorithm with the ith count process removed from the data. To solve the label switching problem that typically occurs in mixture modeling, we then re-order the estimates $\{\hat{\lambda}_g^{(i)}\}_{g=1}^G$ for $i = 1, \cdots, N$ so that the squared distance $\sum_{g=1}^G \sum_{k=1}^m (\hat{\lambda}_g(t_k) - \hat{\lambda}_g^{(i)}(t_k))^2$ is minimized.

For $g = 1, \cdots, G$, and for $k = 1, \cdots, m$, we let

$$\bar{\lambda}_g(t_k) = \frac{1}{N} \sum_{i=1}^N \hat{\lambda}_g^{(i)}(t_k)$$

be the estimator of intensity function g at time t_k based on all sub-samples. The estimated variance of the intensity function estimator using jackknife technique is given below.

$$\mathrm{Var}_{jack}(\hat{\lambda}_g(t_k)) = \frac{N-1}{N} \sum_{i=1}^N (\hat{\lambda}_g^{(i)}(t_k) - \bar{\lambda}_g(t_k))^2 \tag{14}$$

for $g = 1, \cdots, G$ and $k = 1, \cdots, m$. As a result, the confidence interval for each estimated intensity function $\hat{\lambda}_g$ is given by:

$$\left(\hat{\lambda}_g(t_k) - c\sqrt{\mathrm{Var}_{jack}(\hat{\lambda}_g(t_k))}, \hat{\lambda}_g(t_k) + c\sqrt{\mathrm{Var}_{jack}(\hat{\lambda}_g(t_k))}\right)$$

for $k = 1, \cdots, m$ and for c is chosen depending on the desired confidence level, and is usually set to 2 in practice. The lower bound of the confidence interval is set to 0 if it is a negative number.

While the jackknife method can be computationally expensive, it can be performed completely in parallel. For much larger N, the jackknife method may be replaced by the leave-k-out cross validation to reduce computational burden where k is some small integer.

3.5 Model Selection

We want to determine the number of Poisson processes in the mixture model G as well as the hyperparameters ϕ for the Gaussian process prior. The application of standard information criteria in performing model selection is problematic due to the difficulty in defining the dimension of the model or the number of parameters.

Smyth (2000) proposed using cross-validated likelihood approach to choose the number of components in a mixture model. In this paper, we extend this method along with grid search to determine both the number of Poisson processes G and the hyperparameters ϕ. Consider the general case of choosing the optimal model $\mathcal{M}_k$ from a set of models $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_{k_{max}}\}$ where k_{max} is the total number of models considered. Let $\theta^{(k)}$ be the parameters corresponding to the model $\mathcal{M}_k$, and let $l(\theta^{(k)}; D)$ denote its log-likelihood model with parameters $\theta^{(k)}$ evaluated on dataset D. Assuming dataset D is observed, and for a fixed model $\mathcal{M}_k$, the method works by repetitively partitioning the data set into two sets, one of which is to train the model $\mathcal{M}_k$ and obtain estimates of parameters $\hat{\theta}^{(k)}$

by maximizing the log-likelihood and the other is for testing the model by evaluating its log-likelihood.

Assuming that the data D is partitioned into J sets, $\{S_1, S_2, \ldots, S_J\}$. For jth partition, let $D \setminus S_j$ be the dataset used for training the model and let:

$$\hat{\theta}^{(k)}(D \setminus S_j) := \underset{\theta^{(k)}}{\operatorname{argmax}}\, l(\theta^{(k)}; D \setminus S_j)$$

be the estimated parameters. We then evaluate the log-likelihood of the test set S_j with the estimated parameters $\hat{\theta}^{(k)}(D \setminus S_j)$ to obtain $l(\hat{\theta}^{(k)}(D \setminus S_j); S_j)$. The cross-validated estimate of the test log-likelihood is defined as:

$$l_k^{cv} = \frac{1}{J}\sum_{j=1}^{J} l(\hat{\theta}^{(k)}(D \setminus S_j); S_j) \tag{15}$$

In the case of mixture of Poisson processes model, $\mathcal{M}_k$ represents a model with $G^{(k)}$ Poisson processes and hyperparameters $\phi^{(k)}$. We determine the number of Poisson processes G and the hyperparameters ϕ by choosing the model M_k with the highest value of cross-validated likelihood l_k^{cv}. The choice of J for the cross validation has attracted much research interests (Kohavi 1995; Bengio and Grandvalet 2004; Zhang and Yang 2015). Experimental results in the literature shows that moderate values of J tend to reduce the variance of the test log-likelihood (Kohavi 1995). Performing cross validation with several random splits also helps reduce the variance. However, a too large J relative to the size of data results in only a low number of sample combinations, thus limiting the number of iterations that are different.

3.6 Clusters Versus Mixture Components

Baudry et al. (2010) argued the difference between clusters and mixture components and proposed a method based on entropy criterion to check if mixture components are modeling distinct clusters. Let $\hat{\pi}_i^g$ be the estimated a posterior probability of count process i belongs to cluster g, the entropy of a particular mixture model with G components is given by:

$$\mathrm{Ent}(G) = -\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{\pi}_i^g \log(\hat{\pi}_i^g)$$

A greedy algorithm is then used to combine the mixture components where at each stage, the two mixture components to be merged are chosen so as to minimize the resulting entropy. The decrease in the entropy at each step of the procedure may help guide the choice of the number of clusters.

4 Simulation

We evaluate the performance of the proposed model and the EM algorithm developed in Section 3 using simulation studies. For each combination of the number of sub-intervals

Table 1 L_2 distance between estimated and true intensity functions and classification accuracy of the EM algorithm under various parameter settings

m	$\boldsymbol{\sigma}$	$\boldsymbol{l}$	$\boldsymbol{L_2}$ distance	Classification accuracy
20	1	0	0.101	1
20	1	0.005	0.077	1
20	1	0.01	0.083	1
20	1	0.02	0.082	1
20	2	0	0.275	0.920
20	2	0.005	0.219	0.926
20	2	0.01	0.177	0.945
20	2	0.02	0.202	0.938
20	5	0	0.384	0.786
20	5	0.005	0.332	0.859
20	5	0.01	0.376	0.830
20	5	0.02	0.421	0.819
40	1	0	0.182	1
40	1	0.005	0.153	1
40	1	0.01	0.136	1
40	1	0.02	0.163	1
40	2	0	0.321	0.923
40	2	0.005	0.206	0.968
40	2	0.01	0.283	0.947
40	2	0.02	0.295	0.920
40	5	0	0.628	0.802
40	5	0.005	0.408	0.869
40	5	0.01	0.517	0.840
40	5	0.02	0.581	0.802

m and the hyperparameters (σ, l) specified in Table 1, we simulate 100 count process observations according to a mixture of three point processes with intensity functions:

$$\lambda_1(s) = 10\mathcal{B}(s; 2, 2)$$
$$\lambda_2(s) = 10\mathcal{B}(s; 3, 10)$$
$$\lambda_3(s) = 10\mathcal{B}(s; 0.5, 0.5)$$

where $\mathcal{B}(.; a, b)$ denotes the density function of the Beta distribution with parameters a and b. The mixing proportion and the scaling factors are set as $\tau = (1/3, 1/3, 1/3)$ and $\alpha = (50, \ldots, 50)$, respectively. The EM algorithm is then applied to estimate the intensity functions and to obtain clustering structure. The procedure described above is repeated 500 times in order to obtain an estimate of the L_2 distance between the estimated and true intensity functions and its classification accuracy.

The entries in Table 1 with length scale parameter $l = 0$ correspond to the limiting case described in Section 3.1 where different points in the time domain are assumed to be independent. We see from Table 1 that imposing covariance structure on the points improves model fit as well as classification accuracy. As the true intensity functions are relatively smooth, the accuracies in estimation and classification are similar with $m = 20$ and $m = 40$.

The true intensity functions, the average of the estimated intensity functions resulting from 500 replications, and the estimated 95% confidence interval are shown in Fig. 1. We can see that the intensity functions can be consistently recovered. A higher level of uncertainty is observed at locations where the true intensity function has large gradient.

The experiments were performed using a 12-core Intel i7-8700 computer with a clock speed of 3.2 GHz. The average computing times for running a single EM algorithm are 1.26 s for $m = 20$, and 2.74 s for $m = 40$.

5 Examples

5.1 Washington Bike-Sharing Scheme Data

We apply the mixture of Poisson processes model to analyze the Washington bike-sharing scheme data.[1] Information including start date and time, end date and time, start station and end station of a trip are publicly available. For each station in the bike-sharing scheme, we consider the end time of a trip as an event. We fit the model to the data over the period of 18 March 2016 through 24 March 2016 with 362 active stations and a total of 62,234 events.

We partition the time interval into 84 sub-intervals, which corresponds to the aggregation of events for each 2-h interval. We find choosing 84 sub-intervals ensures a good trade-off between resolution of details and fluctuation of the intensity functions. To determine the number of clusters K and hyperparameters $\phi = (l, \sigma)$ for the covariance function, we apply the cross-validated log-likelihood approach described in Section 3.5. In particular, various combination of values were chosen for K, l and σ and the cross-validated log-likelihood is estimated for each of the combinations. The four-cluster model consistently scores the highest for various combinations of hyperparameters l and σ.

The results obtained from fitting a four-cluster model with hyperparameters $l = 0.01$ and $\sigma = 1$ are presented here. Each bike station is plotted on the map in Fig. 2 where its color represents the maximum a posteriori cluster membership. The estimated intensity functions along with the estimated 95% confidence interval are given in Fig. 3. The confidence intervals are esimated using the jackknife approach as described in Section 3.4. We summarize the four estimated intensity functions using a text descriptor (Table 2).

It can be seen from Fig. 3 that the overall level of activities tends to be substantially higher on weekdays compared with weekends. The estimated intensity function for cluster 3 has two peaks on each working day which implies that these stations are popular destinations both in the morning and afternoon. We observe that the bike stations in cluster 4 tend to be concentrated in the central area of Washington, and the corresponding estimated intensity function has a strong peak in the morning. It is very likely that these stations are located near business districts and riders arrive at work in the morning. On the other hand, cluster 1 has most stations located further from the city center, and the corresponding estimated intensity function has a strong peak in the afternoon. Hence, it is plausible that riders arrive home from work in the afternoon.

Figure 2 shows some spatial effect in the clustering of bike stations whereby stations that are close in distance tend to be in the same cluster. A potential extension of the model to take into account the spatial effect is to allow the cluster assignment probabilities depending on covariates of count processes.

[1] Historical data available at https://www.capitalbikeshare.com/

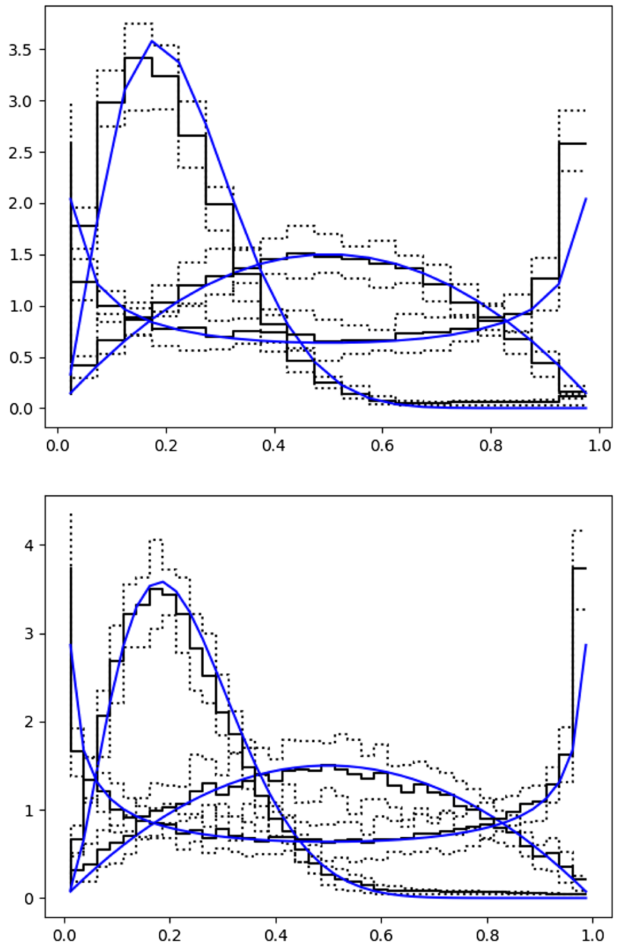

Fig. 1 Estimated intensity functions. Blue: true intensity functions. Black solid: mean estimated intensity functions. Black dot: estimated 95% confidence intervals. Top: number of sub-intervals = 20. Bottom: number of sub-intervals = 40

The predicted versus actual number of events and the quantile–quantile plot of the standardized residuals for each bike station at various time intervals are shown in Figs. 4 and 5. We see that the proposed model provides an adequate fit to the data although the observed quantiles tend to slightly deviate from the theoretical ones at the tails.

We apply the method proposed by Baudry et al. (2010) and presented in Section 3.6 to check if there are multiple components modeling a cluster. Both the plots of entropy versus the number of mixture components and entropy versus cumulative count of merged observations (Fig. 6) show that the reduction in entropy is insignificant at each merging step. Hence, it appears that the four mixture components are modeling four distinct clusters.

5.2 Reality Mining Data

The Reality Mining Dataset was collected in 2004 (Eagle and Pentland 2006). The goal of this experiment was to explore the capabilities of the smart phones that enabled social

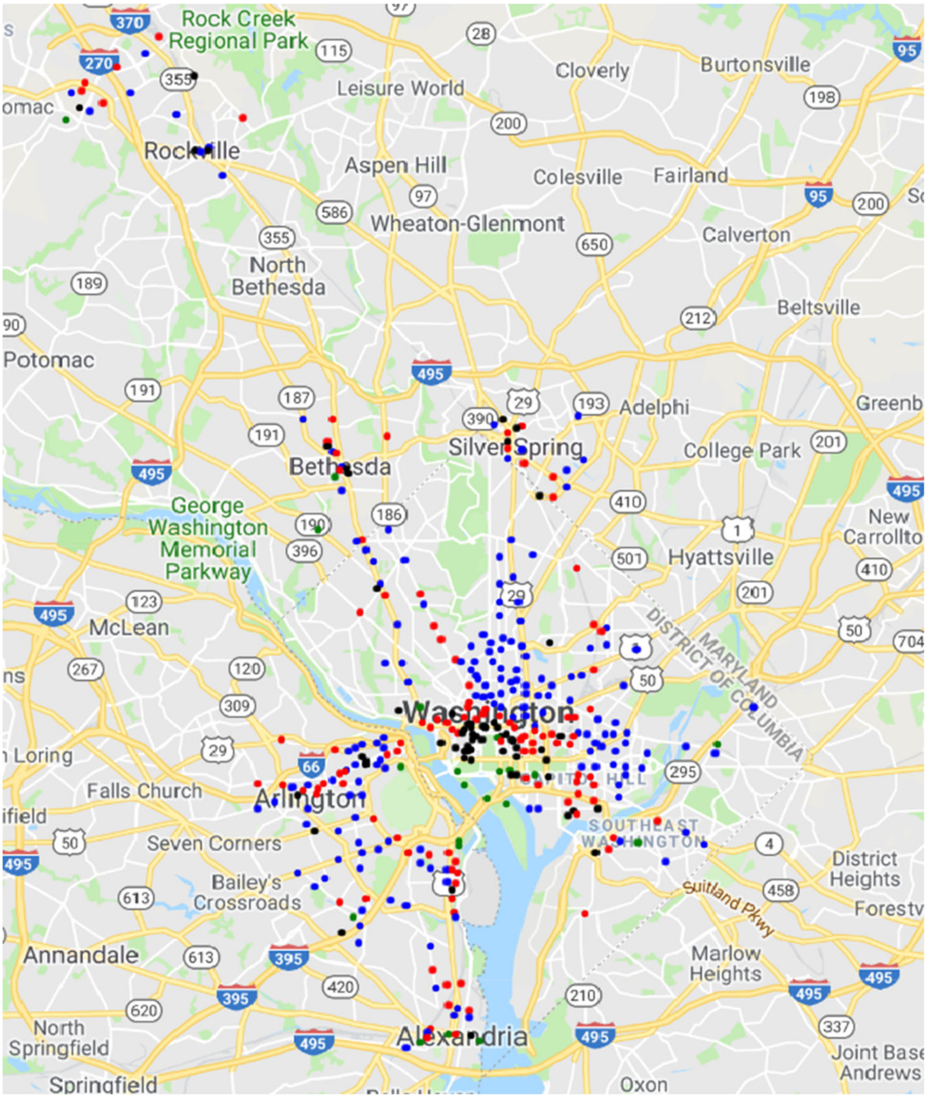

Fig. 2 A map of the bike stations colored by maximum a posteriori cluster membership. Blue: cluster 1. Green: cluster 2. Red: cluster 3. Black: cluster 4

scientists to investigate human interactions. In particular, 75 students or faculty in the MIT Media Laboratory and 25 students at the MIT Slogan business school were the subjects of the experiment where the times of phone calls and SMS messages between subjects were recorded. For each subject in the study, we treat the time when an outgoing call was made as an event time.

To mitigate the effect of subjects dropping out of study, we focus on the core of 86 people who have made outgoing calls during the pe.riod between 24 September 2004 and 8 January 2005.[2]

[2] Data obtained from http://sociograph.blogspot.ie/2011/04/communication-networks-part-2-mit.html

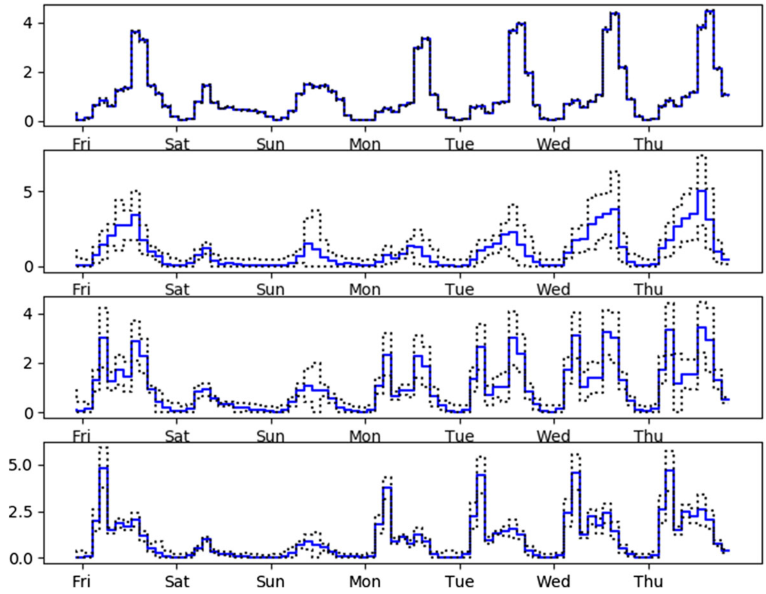

Fig. 3 Estimated intensity functions and confidence intervals for the bike-sharing dataset, where the x-axis shows the days of the week (start from 12 am for each day) and the y-axis shows the intensities

We partition the time interval into 106 sub-intervals which represents an aggregation of events to 1-day intervals. As in the case of bike-sharing scheme dataset, we determine the number of clusters K and hyperparameters $\phi = (l, \sigma)$ by applying the cross-validated log-likelihood method with grid search. The five-cluster model is selected by the method and the results from fitting a five-cluster model with hyperparameters $l = 0.01$ and $\sigma = 1$ are shown in Fig. 7.

A text descriptor is provided in Table 3 to summarize the five estimated intensity functions. The five intensity functions reveal contrasting behavior of the participants where participants in cluster 1 tend to be consistently active while participants in cluster 5 make very few outgoing calls from early November. We observe that participants in cluster 3

Table 2 Descriptions of the estimated intensity functions for the Washington bike-sharing dataset

Color	Cluster name	Description
Blue	Sharp peak in the afternoon	Low level of activities in the morning with a sharp peak in the afternoon
Green	No peak	Increasing level of activities, without sharp peak
Red	Two Peaks	One peak in the morning, and another in the afternoon
Black	Early Peak	One sharp peak in the morning

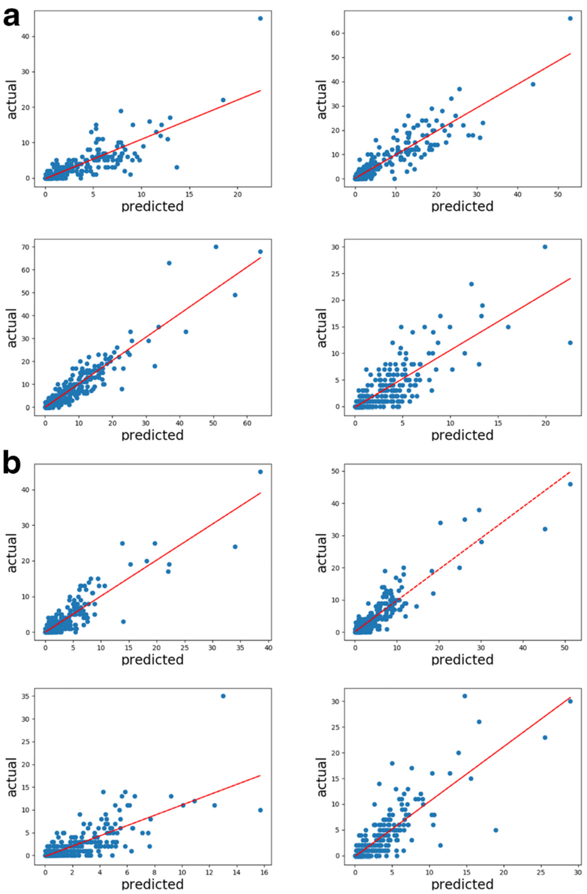

Fig. 4 Top two rows: predicted versus actual number of events for each bike station in every 10th sub-interval. Bottom two rows: quantile–quantile plots of standardised residuals for each bike station in every 10th sub-interval

tend to make increasing number of phone calls as time passes while the reverse is true for participants in cluster 4.

The predicted versus actual number of events and the quantile-quantile plot of the standardized residuals for each bike station at various time intervals are shown in Figs. 8 and 9. Both the plots of entropy versus number of components and entropy versus cumulative count of merged observations (Fig. 10) show that the reduction in entropy is insignificant at each merging step.

6 Discussion

In this paper, we have proposed a general framework to model and cluster count processes data for which penalized likelihood estimation using the EM algorithm is practical to implement. The approach has a number of advantages over existing methods in clustering point processes data. First, the incorporation of a Gaussian process prior enhances smoothness of

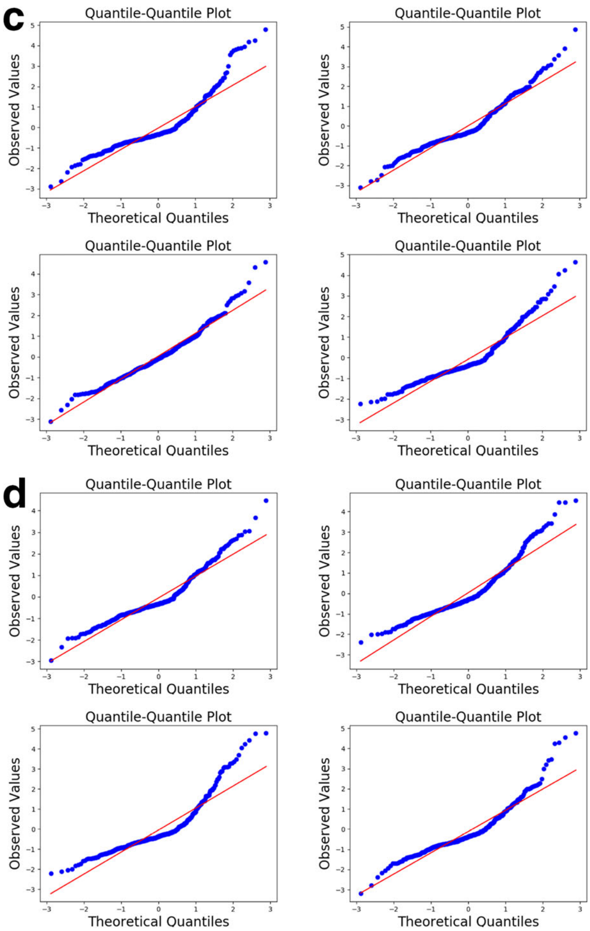

Fig. 5 Bottom two rows: quantile–quantile plots of standardised residuals for each bike station in every 10th sub-interval

the estimated intensity functions. The discretizations of the time domain and Gaussian processes lead to a computationally efficient inference algorithm. Furthermore, the jackknife method provides a sound framework to estimate the uncertainties in intensity functions, while the selection of hyperparameters and number of clusters is performed using a cross-validated log-likelihood approach. We have applied the proposed method to two real-world datasets and obtained interesting and interpretable results.

The proposed modeling framework can be generalized and extended. The length of each sub-interval is assumed to be equal under the current framework. Allowing the length of the sub-intervals to vary would be more appropriate when the underlying intensity has varying levels of smoothness. When covariates of count process observations are available, they may be incorporated in the cluster assignment probabilities. Furthermore, the development of an efficient Markov Chain Monte Carlo algorithm for the developed model is also of interest. In applications such as natural disaster modeling, it is reasonable to assume that the arrival of an event affects future arrivals for some period of time. In such scenarios, alternative point processes such as Hawkes process may be more appropriate.

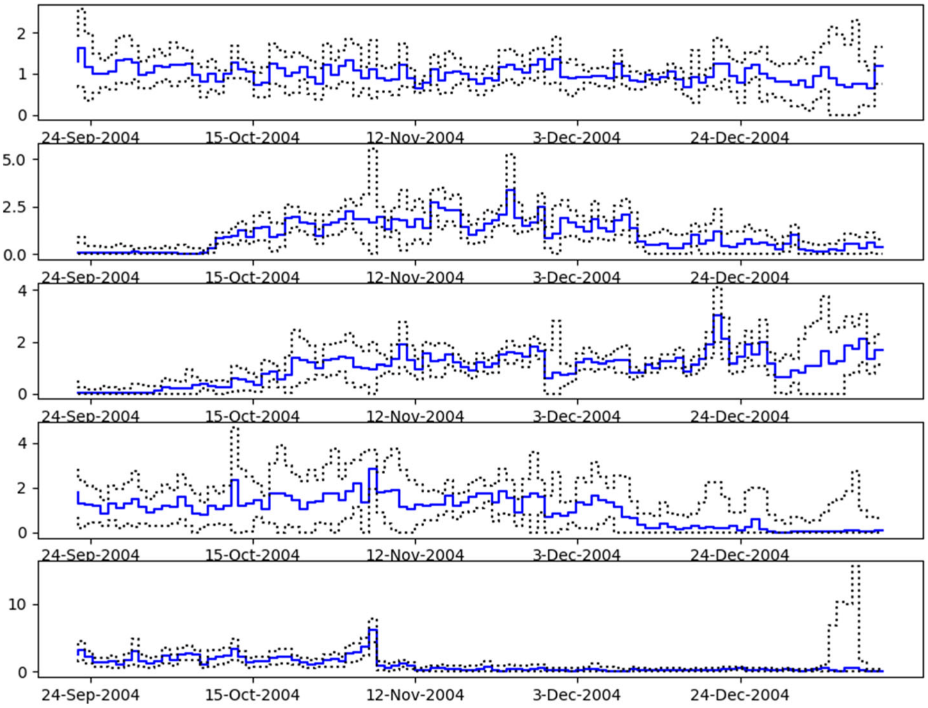

Fig. 6 Entropy plots for the bike-sharing data. Top: entropy versus number of components. Bottom: entropy versus cumulative count of merged observations

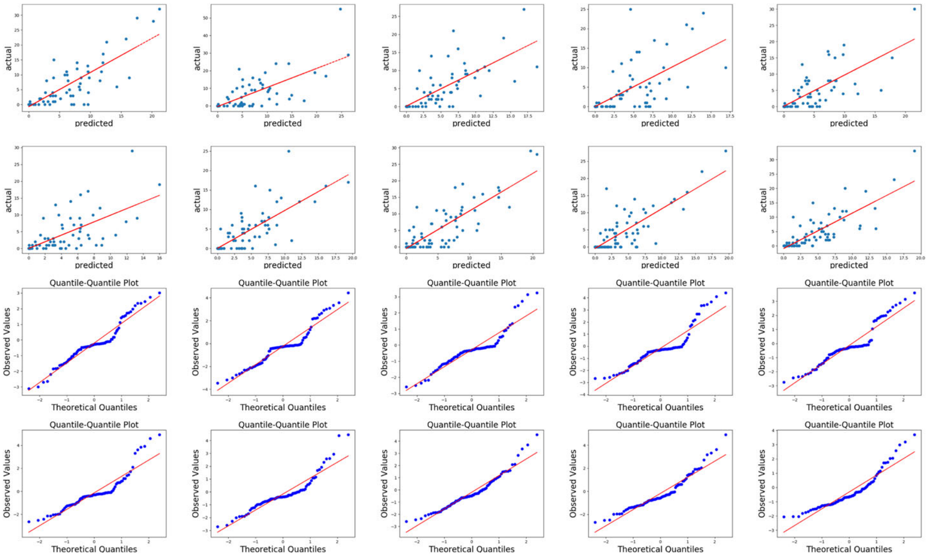

Fig. 7 Estimated intensity functions and confidence intervals for the Reality Mining dataset

Table 3 Descriptions of the estimated intensity functions for the Reality Mining data set

No.	Cluster name	Description
1	Stable	Level of activities very stable
	Rising and declining	Level of activities rising and declining
3	Rising and stable	Level of activities gradually increasing, and remain stable from October
4	Stable and declining	Level of activities remain stable until December, and die off during Christmas
5	Die off	Activities die off at early November

Appendix A: Consistency of Penalized MLE

Sufficient conditions for $\hat{P}$ to be strongly consistent are given in Kiefer and Wolfowitz (1956) and are stated below.

- (C1) Identifiability: Let $H(y_i; \alpha_i, P)$ be the cumulative distribution function of $h(y_i; \alpha_i, P)$. If for any α_i, $H(y_i; \alpha_i, P) = H(y_i; \alpha_i, P^*)$ for all y_i, then $D(P, P^*) = 0$.

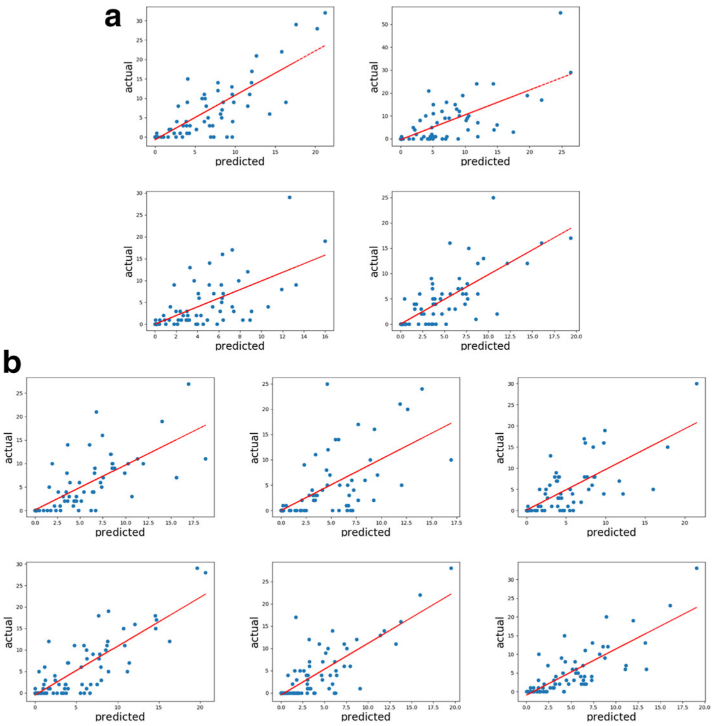

Fig. 8 Top two rows: predicted versus actual number of events for each participant in every 10th sub-interval

- (C2) Continuity: The component parameter space $\mathcal{F}$ is a closed set. For all y_i, α_i and any given P_0,

$$\lim_{P \to P_0} h(y_i; \alpha_i, P) = h(y_i; \alpha_i, P_0).$$

- (C3) Finite Kullback–Leibler Information: For any $P \neq P^*$, there exists an $\epsilon > 0$ such that

$$E^*[\log(h(Y_i; \alpha_i, B_\epsilon(P))/h(Y_i; \alpha_i, P^*))]^+ < \infty,$$

where $B_\epsilon(P)$ is the open ball of radius $\epsilon > 0$ centered at P with respect to the metric D, and

$$h(Y_i; \alpha_i, B_\epsilon(P)) = \sup_{\tilde{P} \in B_\epsilon(P)} h(Y_i; \alpha_i, \tilde{P}).$$

- (C4) Compactness: The definition of the mixture density $h(y_i; \alpha_i, P)$ in P can be continuously extended to a compact space $\overline{\mathbb{P}}$.

Identifiability of the mixture model is established in Proposition 3.1. In the case of mixture of Poisson processes, the component parameter space $\mathcal{F}$ is clearly closed.

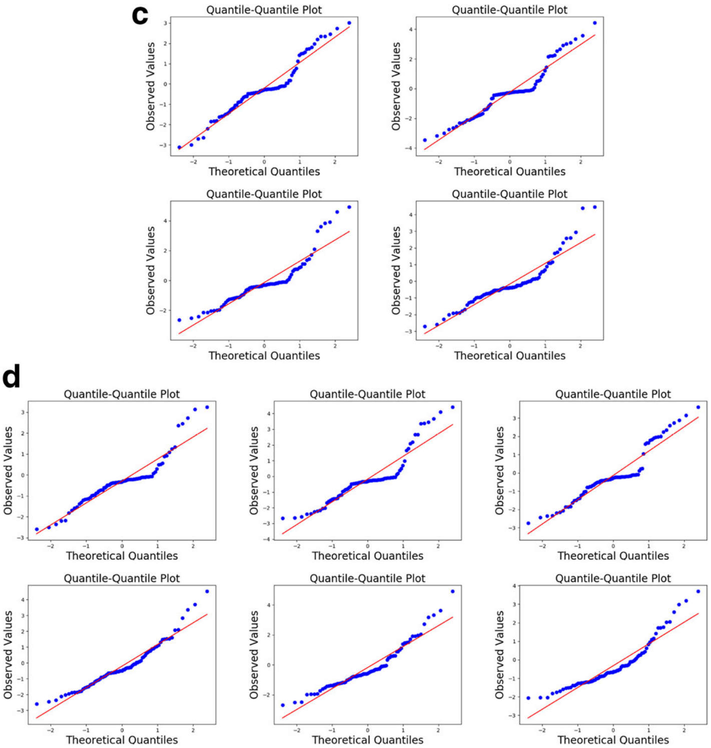

Fig. 9 Bottom two rows: quantile–quantile plots of standardized residuals for each participant in every 10th sub-interval

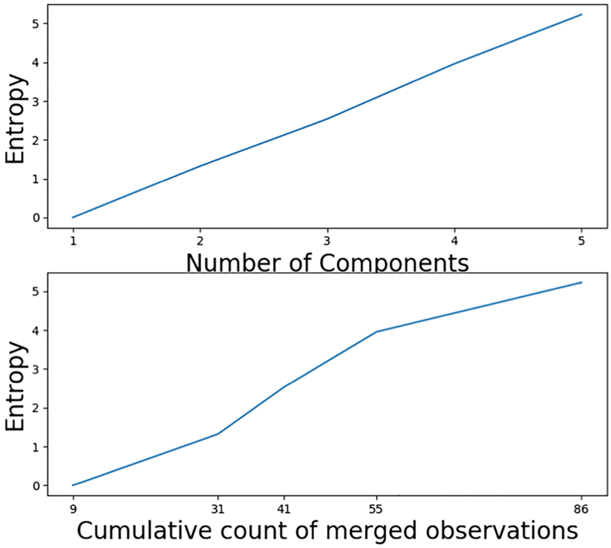

Fig. 10 Entropy plots for Reality Mining data. Top: entropy versus number of components. Bottom: entropy versus cumulative count of merged observations

To compactify the space of mixing distributions $\mathbb{P}$, we notice that the distance between any two mixing distributions is bounded above by:

$$\int_{\mathcal{F}} \exp(-|\mathbf{f}|)d\mathbf{f} < \infty.$$

We can extend $\mathbb{P}$ to $\overline{\mathbb{P}}$ by including all sub-distributions ρP for any $\rho \in [0, 1)$ as in Chen (2017).

We now show the continuity of $h(y_i; \alpha_i, P)$ on $\overline{\mathbb{P}}$. Recall that $D(P_m, P) \to 0$ if and only if $P_m \to P$ in distribution if only if $\int u(\mathbf{f})dP_m(\mathbf{f}) \to \int u(\mathbf{f})dP_0(\mathbf{f})$ for all bounded and continuous function $u(\cdot)$.

For any given $\mathbf{f}_0 \in \mathcal{F}$ and $\alpha_i > 0$, we clearly have $\lim_{\mathbf{f}\to\mathbf{f}_0} h(y_i; \alpha_i, \mathbf{f}) = h(y_i; \alpha_i, \mathbf{f}_0)$, and $\lim_{|\mathbf{f}|\to\infty} h(y_i; \alpha_i, \mathbf{f}) = 0$. Hence, $h(y_i; \alpha_i, \mathbf{f})$ is continuous and bounded on $\mathcal{F}$. Suppose that $P_m \to P_0 \in \overline{\mathbb{P}}$ in distribution, we have that:

$$h(y_i; \alpha_i, P_m) = \int h(y_i; \alpha_i, \mathbf{f})dP_m(\mathbf{f}) \to \int h(y_i; \alpha_i, \mathbf{f})dP_0(\mathbf{f}) = h(y_i; \alpha_i, P_0)$$

which shows that $h(y_i; \alpha_i, P)$ is continuous in $\mathbb{P}$.

To prove finite Kullback–Leibler information, we need the extra sufficient condition that all scaling factors $\{\alpha_i\}_i$ are bounded above, that is, $\alpha_i < \alpha^{(M)} < \infty$ almost surely $\forall i$. Let $\mathbf{f}_0$ be a support point of P^*. There must be a positive constant δ such that:

$$h(y_i; \alpha_i, P^*) \geq \delta \left(\exp(-\alpha_i) \prod_{k=1}^{m} \frac{(\alpha_i \lambda_k)^{y_k}}{y_k!} \right) p(\mathbf{f}_0).$$

uniformly for all $\alpha_i < \alpha^{(M)}$. Therefore, we have that:

$$E^*(\log h(y_i; \alpha_i, P^*)) \geq \sum_{k=1}^{m} \log(\alpha_i \lambda_k) E^*(Y_{i,k}) - \sum_{k=1}^{m} E^*(\log Y_{i,k}!) - \alpha_i + \log(\delta) + \log p(\mathbf{f}_0).$$

Since $\alpha_i < \alpha^{(M)}$ almost surely, both $E^*(Y_{i,k})$ and $E^*(\log(Y_{i,k}!))$ are finite for all k. Hence, $E^*(\log h(y_i; \alpha_i, P^*)) > -\infty$. Since $h(y_i; \alpha_i, P^*)$ is bounded from above, $E^*(\log h(y_i; \alpha_i, P^*)) < \infty$. Therefore, $E^*|\log h(y_i; \alpha_i, P^*)| < \infty$.

Since $h(Y_i; \alpha_i, B_{\epsilon(P)}) < c < \infty$ for any P, ϵ and α_i,

$$E^*[\log(h(Y_i; \alpha_i, B_\epsilon(P))/h(Y_i; \alpha_i, P^*))]^+ < \log(c) - E^*(\log h(Y_i; \alpha_i, P^*)) < \infty.$$

Therefore, we have shown that the four conditions above are satisfied and the penalized MLE under mixture of Poisson process is strongly consistent.

Appendix B: Derivation of EM Algorithm

Recall the complete data log-likelihood defined in Eq. 10. The E-step requires the computation of the expected value of the complete data log-likelihood function with respect to the conditional distribution of Z given x and current estimates of parameters $\hat{\theta}^{(t)}$. This conditional expression can be expressed as:

$$\begin{aligned} E_{Z|x,\hat{\theta}^{(t)}}(\log L(\theta; x, Z)) = & \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{\pi}_i^g \left[\log(\tau_g) - \alpha_i + \sum_{j=1}^{n_i} \log(\alpha_i) + \log(\lambda_g(x_{ij})) \right] \\ & + \sum_{g=1}^{G} \log(p(\mathbf{f}_g)) \end{aligned} \tag{16}$$

The M-step finds the parameters θ that maximize the expression above. It is straightforward to show (by differentiation) that the values of α and τ that maximize the expression above are given by Eqs. 12 and 13 respectively.

To optimize Eq. 16 with respect to $\mathbf{f}_g$, we write $\mathbf{f}_g = (f_{g,1}, \cdots, f_{g,m})^T$ and retain terms that involve $\mathbf{f}_g$ for $g = 1, \cdots, G$ to obtain:

$$\begin{aligned} Q \equiv & \sum_{i=1}^{n} \hat{\pi}_i^g \sum_{j=1}^{n_i} \log(\lambda_g(x_{ij})) + \log(p(\mathbf{f}_g)) \\ = & \sum_{i=1}^{n} \pi_i^g \sum_{j=1}^{n_i} \log\left[\frac{\exp(f_g(x_{i,j}))}{\sum_{k=1}^{m}\exp(f_{g,k})}\right] - \frac{1}{2}\mathbf{f}_g^T K^{-1}\mathbf{f}_g + const \\ = & \sum_{i=1}^{n} \hat{\pi}_i^g \sum_{j=1}^{n_i} f_g(x_{i,j}) - \sum_{i=1}^{n} \hat{\pi}_i^g n_i \log\left[\sum_{k=1}^{m}\exp(f_{g,k})\right] - \frac{1}{2}\mathbf{f}_g^T K^{-1}\mathbf{f}_g + const \\ = & \mathbf{b}^T\mathbf{f}_g - \sum_{i=1}^{n} \hat{\pi}_i^g n_i \log\left[\sum_{k=1}^{m}\exp(f_{g,k})\right] - \frac{1}{2}\mathbf{f}_g^T K^{-1}\mathbf{f}_g + const \end{aligned}$$

where $\mathbf{b} = (b_1, \cdots, b_m)^T$ with y_k defined below.

$$b_k = \sum_{i=1}^{n} \hat{\pi}_i^g \sum_{j=1}^{n_i} I(x_{ij} \in [s_{k-1}, s_k))$$

where $[s_{k-1}, s_k)$ is the kth interval of the m equally spaced intervals of $[0, T)$.

For $g = 1, \cdots, G$, we now optimize Q with respect to $\mathbf{f}_g$, and as analytical expression does not exist, we use Newton's method. For each g, the Jacobian $J_g(Q)$ of Q can be written as:

$$J_g(Q) = \mathbf{y} - \mathbf{u}_g \sum_{i=1}^{n} \hat{\pi}_i^g n_i - \frac{1}{2} \mathbf{f}_g^T (K^{-1} + K^{-T})$$

where $\mathbf{u}_g = (u_{g,1}, \cdots, u_{g,m})^T$ with

$$u_{g,l} = \frac{\exp(f_{g,l})}{\sum_{k=1}^{m} \exp(f_{g,k})}$$

for $l = 1, \cdots, m$. The Hessian $H_g(Q)$ can be written as:

$$H_g(Q) = -(\mathrm{diag}(\mathbf{u}_g) - \mathbf{u}_g \mathbf{u}_g^T) \sum_{i=1}^{n} \pi_i^g n_i - \frac{1}{2} (K^{-1} + K^{-T})$$

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Statist.*, *30*, 581–595.

Adams, R.P., Murray, I., MacKay, D.J.C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th annual international conference on machine learning* (pp. 9–16).

Basu, S., & Dassios, A. (2002). A Cox process with log-normal intensity. *Insurance sMath. Econom.*, *31*, 297–302.

Baudry, J.-P., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.*, *19*, 332–353.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: a review. *Comput. Statist. Data Anal.*, *71*, 52–78.

Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E. (2019). *Model-based clustering and classification for data science: with applications in R*. Cambridge: Cambridge University Press.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statist Sci.*, *32*, 47–63.

Chen, J., Li, S., Tan, X. (2016). Consistency of the penalized MLE for two-parameter gamma mixture models. *Sci China Math.*, *59*, 2301–2318.

Chen, J., Tan, X., Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statist Sinica*, *18*, 443–465.

Cheng, R.C.H., & Liu, W.B. (2001). The consistency of estimators in finite mixture models. *Scand J. Statist.*, *28*, 603–616.

Ciuperca, G., Ridolfi, A., Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scand J. Statist.*, *30*, 45–59.

Côme, E., & Latifa, O. (2014). Model-based count series clustering for Bike-sharing system usage mining, a case study with the Vélib' system of Paris, ACM Transactions on Intelligent Systems and Technology (TIST) - Special Section on Urban Computing, 5.

Cunningham, J.P., Shenoy, K.V., Sahani, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th international conference on machine learning* (pp. 192–199).

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, *39*, 1–38. with discussion.

Eagle, N., & Pentland, A. (2006). Reality mining: sensing complex social systems. *Pers Ubiquitous Comput.*, *10*, 255–268.

Fraley, C., & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, *97*, 611–631.

Fraley, C., & Raftery, A.E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*, 155–181.

Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, *69*, 31–40.

Heikkinen, J., & Arjas, E. (1999). Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, *55*, 738–745.

Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. *Comput Statist. Data Anal.*, *71*, 92–106.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, *27*, 887–906.

Kim, H., & Ghahramani, Z. (2006). Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 1948–1959.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on artificial intelligence - vol. 2, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., IJCAI'95, pp. 1137–1143.

Lenk, P.J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, *83*, 509–516.

Lenk, P.J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, *78*, 531–543.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B*, *40*, 113–146. with discussion.

Lloyd, C., Gunter, T., Osborne, M.A., Roberts, S.J. (2015). Variational inference for Gaussian process modulated Poisson processes. In *Proceedings of the 32nd international conference on international conference on machine learning* -, (Vol. 37 pp. 1814–1822).

McNicholas, P.D. (2016). *Mixture model-based classification*. Boca Raton: CRC Press.

Murray, I., MacKay, D., Adams, R.P. (2009). The Gaussian process density sampler. In Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.) *Advances in neural information processing systems 21* (pp. 9–16): Curran Associates, Inc.

O'Hagan, A., Murphy, T.B., Scrucca, L., Gormley, I.C. (2019). Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Comput Stat.*, *34*, 1779–1813.

Rasmussen, C.E., & Williams, C.K.I. (2006). *Gaussian processes for machine learning, adaptive computation and machine learning*. MIT Press: Cambridge.

Sapatinas, T. (1995). Identifiability of mixtures of power-series distributions and related characterizations. *Ann. Inst. Statist. Math.*, *47*, 447–459.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Stat Comp.*, *10*, 63–72.

Tokdar, S.T., & Ghosh, J.K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *J. Statist. Plann. Inference*, *137*, 34–42.

Williams, C.K.I., & Rasmussen, C.E. (1996). Gaussian processes for regression. In *Advances in neural information processing systems 8* (pp. 514–520). Cambridge: MIT Press.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econom.*, *187*, 95–112.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.