

# Comparative evaluation of various frequentist and Bayesian non-homogeneous Poisson counting models

Marco Grzegorzczak<sup>1</sup> · Mahdi Shafiee Kamalabad<sup>1</sup>

Received: 5 August 2015 / Accepted: 13 September 2016 / Published online: 5 October 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

**Abstract** In this paper a comparative evaluation study on popular non-homogeneous Poisson models for count data is performed. For the study the standard homogeneous Poisson model (HOM) and three non-homogeneous variants, namely a Poisson changepoint model (CPS), a Poisson free mixture model (MIX), and a Poisson hidden Markov model (HMM) are implemented in both conceptual frameworks: a frequentist and a Bayesian framework. This yields eight models in total, and the goal of the presented study is to shed some light onto their relative merits and shortcomings. The first major objective is to cross-compare the performances of the four models (HOM, CPS, MIX and HMM) independently for both modelling frameworks (Bayesian and frequentist). Subsequently, a pairwise comparison between the four Bayesian and the four frequentist models is performed to elucidate to which extent the results of the two paradigms ('Bayesian vs. frequentist') differ. The evaluation study is performed on various synthetic Poisson data sets as well as on real-world taxi pick-up counts, extracted from the recently published New York City Taxi database.

**Keywords** Non-homogeneous count data · Mixture model · Changepoint model · Hidden Markov model · Bayesian paradigm · Frequentist paradigm · New York City Taxi data

---

**Electronic supplementary material** The online version of this article (doi:[10.1007/s00180-016-0686-y](https://doi.org/10.1007/s00180-016-0686-y)) contains supplementary material, which is available to authorized users.

---

✉ Marco Grzegorzczak  
m.a.grzegorzczak@rug.nl  
Mahdi Shafiee Kamalabad  
m.shafiee.kamalabad@rug.nl

<sup>1</sup> Johann Bernoulli Institute (JBI), Rijksuniversiteit Groningen, 9747 AG Groningen, The Netherlands

## 1 Introduction

The Poisson distribution is one of the most popular statistical standard tools for analysing (homogeneous) count data, i.e. integer-valued samples. For modelling non-homogeneous count data, e.g. time series where the number of counts depends on time and hence systematically differs over time, various extensions of the standard Poisson model have been proposed and applied in the literature. More appropriate non-homogeneous Poisson models can be easily obtained by embedding the standard Poisson model into other statistical frameworks, such as changepoint models (CPS), finite mixture models (MIX), or hidden Markov models (HMM). The three aforementioned modelling approaches have become very popular statistical tools throughout the years for the following three reasons: (i) First, each of the three modelling approaches is of a generic nature so that it can be combined with a huge variety of statistical distributions and models to extend their flexibilities. (ii) Second, the statistical methodology behind those generic models is rather simple, described in lots of textbooks on Statistics and the model inference is feasible. (iii) Third, the three approaches can be easily formulated and implemented in both conceptual frameworks: the standard ‘frequentist’ framework and the Bayesian framework.

Despite this popularity, the performances of the resulting non-homogeneous models have never been systematically compared with each other in the statistical literature. This paper tries to fill this gap and presents a comparative evaluation study on non-homogeneous Poisson count data, for which those three well-known statistical models (changepoint models, mixture models and hidden Markov models) are implemented in both conceptual frameworks: the frequentist framework and the Bayesian framework.

More precisely, for the evaluation study the standard homogeneous Poisson model (HOM) and three non-homogeneous variants thereof, namely a Poisson changepoint model (CPS), a Poisson free mixture model (MIX), and a Poisson hidden Markov model (HMM) are implemented in a frequentist as well as in a Bayesian framework. The goal of the presented study is to systematically cross-compare the performances. Thereby the focus is not only on cross-comparing the generic modelling approach for non-homogeneity (CPS, MIX and HMM), but also on comparing the frequentist model instantiations with the Bayesian model instantiations. The study is performed on various synthetic data sets as well as on real-world taxi pick-up counts, extracted from the recently published New York City Taxi (NYCT) database. In all presented applications it is assumed that the Poisson parameter does *not* depend on any external covariates so that the changes are time-effects only. That is, the non-stationarity is implemented intrinsically by temporal changepoints, at which the Poisson process spontaneously changes its values.

Within this introductory text no literature references have been given, since detailed descriptions of all those generic statistical concepts, mentioned so far, can be found in many standard textbooks on Statistics, and therefore, in principle, will be familiar for most of the readers. However, in Sect. 2, where the models are described and mathematically formulated, explicit literature references will be provided for all models under comparison.

## 2 Methodology

### 2.1 Mathematical notations

Let  $\mathbf{D}$  denote a  $n$ -by- $T$  data matrix, whose columns refer to equidistant time points,  $t \in \{1, \dots, T\}$ , and whose rows refer to independent counts,  $i \in \{1, \dots, n\}$ , which were observed at the corresponding time points. The element  $d_{i,t}$  in the  $i$ th row and  $t$ th column of  $\mathbf{D}$  is the  $i$ th count, which was observed at the  $t$ th time point. Let  $\mathbf{D}_{\cdot,t} := (d_{1,t}, \dots, d_{n,t})^\top$  denote the  $t$ th column of  $\mathbf{D}$ , where “ $\top$ ” denotes vector transposition.  $\mathbf{D}_{\cdot,t}$  is then the vector of the  $n$  observed counts for time point  $t$ .

Assume that the time points  $1, \dots, T$  are linked to  $K$  Poisson distributions with parameters  $\theta_1, \dots, \theta_K$ . The  $T$  time points can then be assigned to  $K$  components, which represent the  $K$  Poisson distributions. More formally, let the allocation vector  $\mathbf{V} = (v_1, \dots, v_T)^\top$  define an allocation of the time points to components, where component  $k$  represents a Poisson distribution with parameter  $\theta_k$ .  $v_t = k$  means that time point  $t$  is allocated to the  $k$ th component and that the observations at  $t$  stem from a Poisson distribution with parameter  $\theta_k$  ( $t = 1, \dots, T$  and  $k = 1, \dots, K$ ). Note that the  $n$  independent counts within each column are always allocated to the same component, while the  $T$  columns (time points) are allocated to different components. Define  $\mathbf{D}^{[k]}$  to be the sub-matrix, containing only the columns of  $\mathbf{D}$  that are allocated to component  $k$ .

The probability density function (pdf) of a Poisson distribution with parameter  $\theta > 0$  is:

$$p(x|\theta) = \frac{\theta^x \cdot \exp\{-\theta\}}{x!} \tag{1}$$

for  $x \in \mathbb{N}_0$ . Assuming that all counts, allocated to  $k$ , are realisations of independently and identically distributed (iid) Poisson variables with parameter  $\theta_k$ , the joint pdf is given by:

$$p(\mathbf{D}^{[k]}|\theta_k) = \prod_{t=1}^T \mathcal{I}_{\{v_t=k\}}(t) \cdot p(\mathbf{D}_{\cdot,t}|\theta_k) \tag{2}$$

where  $\mathcal{I}_{\{v_t=k\}}(t)$  indicates whether time point  $t$  is allocated to component  $k$ , and

$$p(\mathbf{D}_{\cdot,t}|\theta_k) = \prod_{i=1}^n p(d_{i,t}|\theta_k) = \frac{(\theta_k)^{\{\sum_{i=1}^n d_{i,t}\}} \cdot \exp\{-n \cdot \theta_k\}}{d_{1,t}! \cdot \dots \cdot d_{n,t}!} \tag{3}$$

is the joint pdf of the counts in the  $t$ th column of  $\mathbf{D}$ . Given the allocation vector  $\mathbf{V}$ , which allocates the data into  $K$  sub-matrices  $\mathbf{D}^{[1]}, \dots, \mathbf{D}^{[K]}$ , and independent component-specific Poisson distributions with parameters  $\theta_1, \dots, \theta_K$ , the joint pdf of  $\mathbf{D}$  is:

$$p(\mathbf{D}|\mathbf{V}, \boldsymbol{\theta}) = \prod_{k=1}^K p(\mathbf{D}^{[k]}|\theta_k) \tag{4}$$

where  $\boldsymbol{\theta} := (\theta_1, \dots, \theta_K)^\top$ , and  $p(\mathbf{D}^{[k]}|\theta_k)$  was defined in Eq. (2).

Now assume that the allocation vector  $\mathbf{V}$  is known and fixed, while the component-specific Poisson parameters are unknown and have to be inferred from the data  $\mathbf{D}$ .

Following the frequentist paradigm, the parameters can be estimated by the Maximum Likelihood (ML) approach. The ML estimators which maximise the log-likelihood

$$l(\boldsymbol{\theta}|\mathbf{V}, \mathbf{D}) := \log\{p(\mathbf{D}|\mathbf{V}, \boldsymbol{\theta})\} \quad (5)$$

are given by  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_K)^\top$ , where  $\hat{\theta}_k$  is the empirical mean of all counts in  $\mathbf{D}^{[k]}$ . Assuming that  $T_k$  time points are allocated to component  $k$ , the matrix  $\mathbf{D}^{[k]}$  contains  $T_k \cdot n$  counts and

$$\hat{\theta}_k = \frac{1}{n \cdot T_k} \sum_{t=1}^{T_k} \mathcal{I}_{\{v_t=k\}}(t) \sum_{i=1}^n d_{i,t} \quad (6)$$

In a Bayesian setting the Poisson parameters in  $\boldsymbol{\theta}$  are assumed to be random variables as well, and prior distributions are imposed on them. The standard conjugate prior for a Poisson model with parameter  $\theta_k > 0$  is the Gamma distribution:

$$p(\theta_k|a, b) = \frac{b^a}{\Gamma(a)} \cdot (\theta_k)^{a-1} \exp\{-\theta_k \cdot b\} \quad (7)$$

where  $a$  is the shape and  $b$  is the rate parameter. Due to standard conjugacy arguments, for each component  $k$  the posterior distribution is a Gamma distribution with parameters  $\tilde{a} = a + \xi^{[k]}$  and  $\tilde{b} = b + n \cdot T_k$

$$p(\theta_k|\mathbf{D}^{[k]}) = \frac{(b + n \cdot T_k)^{a+\xi^{[k]}}}{\Gamma(a + \xi^{[k]})} \cdot (\theta_k)^{a+\xi^{[k]}-1} \exp\{-\theta_k \cdot (b + n \cdot T_k)\} \quad (8)$$

where  $\xi^{[k]}$  is the sum of all  $n \cdot T_k$  elements of the  $n$ -by- $T_k$  (sub-)matrix  $\mathbf{D}^{[k]}$ . The marginal likelihood can be computed in closed-form:

$$\begin{aligned} p(\mathbf{D}^{[k]}|a, b) &= \int_0^\infty p(\mathbf{D}^{[k]}|\theta_k) p(\theta_k|a, b) d\theta_k \\ &= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{\prod_{t=1}^{T_k} \prod_{i=1}^n (d_{i,t}^{[k]})!} \cdot \frac{\Gamma(a + \xi^{[k]})}{(T_k \cdot n + b)^{\xi^{[k]}+a}} \end{aligned} \quad (9)$$

where  $d_{i,t}^{[k]}$  is the element in the  $i$ th row and  $t$ th column of  $\mathbf{D}^{[k]}$ .

Imposing independent Gamma priors on each component  $k \in \{1, \dots, K\}$  induced by the allocation vector  $\mathbf{V}$ , the marginal likelihood for the complete data matrix  $\mathbf{D}$  is:

$$p(\mathbf{D}|\mathbf{V}) = \prod_{k=1}^K p(\mathbf{D}^{[k]}|a, b) \quad (10)$$

where the dependence on the fixed hyperparameters  $a$  and  $b$  on the left hand side of the last equation was suppressed.

So far it has been assumed that the allocation vector  $\mathbf{V}$  is known and fixed, although  $\mathbf{V}$  will be unknown for many real-world applications so that  $\mathbf{V}$  also has to be inferred from the data  $\mathbf{D}$ . The next section is therefore on the allocation vector inference.

## 2.2 Allocation vector inference

The standard frequentist Poisson or Bayesian Poisson–Gamma model assumes that the data are homogeneous so that all time points  $t = 1, \dots, T$  always belong to the same component; i.e.  $K = 1$  and  $\mathbf{V} = (1, \dots, 1)^\top$ . These models are referred to as the homogeneous (HOM) models. The HOM model is not adequate if the number of counts varies over time, and non-homogeneous Poisson models, which infer the underlying allocation, have to be used instead. Prominent approaches to model non-homogeneity include: multiple changepoint processes (CPS), finite mixture models (MIX), and hidden Markov models (HMM). CPS impose a set of changepoints which divide the time series  $1, \dots, T$  into disjunct segments. Although this is a very natural choice for temporal data, the disadvantage is that the allocation space is restricted, as data points in different segments cannot be allocated to the same component; i.e. a component once left cannot be revisited. E.g. for  $T = 6$  the true allocation  $\mathbf{V} = (1, 1, 2, 2, 2, 1)^\top$  cannot be modelled and the best CPS model approximation might be:  $\mathbf{V}^{CPS} = (1, 1, 2, 2, 2, 3)^\top$ . The MIX model, on the other hand, is more flexible, as it allows for a free allocation of the time points so that  $\mathbf{V}$  is part of the configuration space. But MIX does not take the temporal ordering of the data points into account. It treats the  $T$  time points as interchangeable units. This implies in the example above that all allocation vectors, which allocate  $T_1 = 3$  time points to component  $k = 1$  and  $T_2 = 3$  time points to component  $k = 2$ , are always equally supported a priori; including unlikely allocations, such as:  $\mathbf{V}^* = (1, 2, 1, 2, 1, 2)^\top$ .

A compromise between CPS and MIX is the hidden Markov model (HMM). HMM allows for an unrestricted allocation vector configuration space, but unlike MIX it does not ignore the order of the time points. A homogeneous first-order HMM imposes a (homogeneous) Markovian dependency among the components  $v_1, \dots, v_T$  of  $\mathbf{V}$  so that the value of  $v_t$  depends on the value of the preceding time point  $v_{t-1}$ , and the homogeneous state-transition probabilities can be such that neighbouring points are likely to be allocated to the same component, while components once left can be revisited. The aforementioned Poisson models, can be implemented in a frequentist as well as in a Bayesian framework, yielding 8 non-homogeneous Poisson models in total, see Table 1 for an overview.

## 2.3 The frequentist framework

The learning algorithms for the non-homogeneous frequentist models learn the best-fitting model for each number of components  $K$ , and the goodness of fit increases in  $K$ . Restricting  $K$  to be in between 1 and  $K_{MAX}$ , for each approach (CPS, MIX and HMM) the best fitting model with  $K$  components, symbolically  $\mathcal{M}_K$ , can be learnt from the data  $\mathbf{D}$ . The Bayesian Information criterion (BIC), proposed by Schwarz (1978), is a well-known model selection criterion and balances between the goodness of fit and

**Table 1** Overview to the eight (non-)homogeneous Poisson models under comparison

	Frequentist version (FREQ)	Bayesian version (BAYES)
Homogeneous model (HOM) with 1 parameter	See Sect. 2.3.1. Well-known standard model from frequentist textbooks with closed-form solution	See Sect. 2.4.1. Well-known standard model from Bayesian textbooks with closed-form solution
Changepoint model (CPS) with $K$ parameters	See Sect. 2.3.2. For each $K$ the best changepoint set can be determined with the Segment Neighbourhood Search Algorithm. BIC is used for model selection	See Sect. 2.4.2. Model averaging via MCMC, based on changepoint birth, death, and reallocation moves
Finite mixture model (MIX) with $2K - 1$ parameters	See Sect. 2.3.3. For each $K$ , the ML estimators of the incomplete model can be inferred with the EM algorithm. BIC is used for model selection	See Sect. 2.4.3. Model averaging via MCMC, based on the moves of the allocation sampler
Hidden Markov model (HMM) with $K^2 + 2K - 1$ parameters	See Sect. 2.3.4. For each $K$ , the ML estimators of the incomplete model can be inferred with the EM algorithm. BIC is used for model selection	See Sect. 2.4.4. Model averaging via MCMC, based on the moves of the allocation sampler and four additional moves

Detailed explanations are given in the main text

model sparsity. According to the BIC, among a set of models  $\{\mathcal{M}_1, \dots, \mathcal{M}_{K_{MAX}}\}$ , the one with the lowest BIC value is considered the most appropriate one with the best trade-off (fit vs. sparsity). Given the  $n$ -by- $T$  data set matrix  $\mathbf{D}$ , and models  $\mathcal{M}_K$  with  $K$  components and  $q_{\mathcal{M}_K}$  parameters ( $K = 1, \dots, K_{MAX}$ ), the BIC of  $\mathcal{M}_K$  is defined as

$$BIC(\mathcal{M}_K) = -2 \cdot \log\{p(\mathbf{D}|\mathcal{M}_K)\} + q_{\mathcal{M}_K} \cdot \log(n \cdot T) \quad (11)$$

where  $n \cdot T$  is the number of data points in  $\mathbf{D}$ , and for  $K = 1$  each of the three non-homogeneous model becomes the homogeneous model  $\mathcal{M}_1$  (see Sect. 2.3.1).

### 2.3.1 The homogeneous frequentist Poisson model (FREQ-HOM)

The homogeneous model  $\mathcal{M}_1$  assumes that the counts stay constant over time, i.e. that there is only one single component,  $K = 1$ , and that the allocation vectors assign all data points to this component, i.e.  $\mathbf{V} = \mathbf{1} = (1, \dots, 1)^\top$ . Hence,  $\mathbf{D}^{[1]} = \mathbf{D}$  and according to Eq. (6), the maximum likelihood (ML) estimator of the single ( $q_{\mathcal{M}_1} = 1$ ) Poisson parameter  $\theta := \theta_1$  is the empirical mean of all  $T \cdot n$  data points in  $\mathbf{D}$ .

### 2.3.2 The frequentist changepoint Poisson model (FREQ-CPS)

A changepoint model uses a changepoint set of  $K - 1$  changepoints,  $C = \{c_1, \dots, c_{K-1}\}$ , where  $1 < c_1 < \dots < c_{K-1} < T$ , to divide the time points  $1, \dots, T$  into  $K$  disjunct segments. Time point  $t$  is assigned to component  $k$  if  $c_{k-1} < t \leq c_k$ , where  $c_0 = 0$  and  $c_K = T$  are pseudo changepoints. This means for the  $t$ th element,  $v_t$ , of the allocation vector,  $\mathbf{V}_C$ , implied by  $C$ :  $v_t = k$  if  $c_{k-1} < t \leq c_k$ . A changepoint set  $C$  with  $K - 1$  changepoints implies a segmentation  $\mathbf{D}^{[1]}, \dots, \mathbf{D}^{[K]}$  of the data matrix  $\mathbf{D}$ , and the ML estimators  $\hat{\theta}_k$  for the segment-specific Poisson parameters  $\theta_k$  can be computed with Eq. (6). The model fit can be quantified by plugging the ML estimators  $\hat{\theta}$  into the log-likelihood in Eq. (5):

$$l(\hat{\theta}_C | \mathbf{V}_C, \mathbf{D}) := \log\{p(\mathbf{D} | \mathbf{V}_C, \hat{\theta}_C)\} \quad (12)$$

where  $\mathbf{V}_C$  is the allocation vector implied by  $C$ , and  $\hat{\theta}_C$  is the vector of ML estimators. The best fitting set of  $K - 1$  changepoints,  $C^K$ , i.e. the set maximising Eq. (12), can be found recursively by the segment neighbourhood search algorithm. This algorithm, proposed by Auger and Lawrence (1989), employs dynamic programming to find the best fitting changepoint set with  $K - 1$  changepoints for each  $K$  ( $2 \leq K \leq K_{MAX}$ ). The algorithm is outlined in Sect. 1 of the supplementary material. The best changepoint model  $\mathcal{M}_{\hat{K}}$  minimises the BIC in Eq. (11), and the output of the algorithm is the corresponding allocation vector  $\hat{\mathbf{V}}_{CPS}$  and the segment-specific ML-estimators  $\hat{\theta}_{CPS} := \hat{\theta}_{C^{\hat{K}}}$ .

### 2.3.3 The frequentist finite mixture Poisson model (FREQ-MIX)

In a frequentist finite mixture model with  $K$  components the time points  $1, \dots, T$  are treated as interchangeable units from a mixture of  $K$  independent Poisson distributions

with parameters  $\theta_1, \dots, \theta_K$  and mixture weights  $\pi_1, \dots, \pi_K$ , where  $\pi_k \geq 0$  for all  $k$ , and  $\sum_{k=1}^K \pi_k = 1$ . The columns  $\mathbf{D}_{:,t}$  of the data matrix  $\mathbf{D}$  are then considered as a sample from this Poisson mixture distribution with pdf:

$$p(\mathbf{D}_{:,t}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^K \pi_k \cdot p(\mathbf{D}_{:,t}|\theta_k) \quad (13)$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$  is the vector of Poisson parameters,  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^\top$  is the vector of mixture weights, and  $p(\mathbf{D}_{:,t}|\theta_k)$  can be computed with Eq. (3). The maximisation of Eq. (13) in the parameters  $(\boldsymbol{\theta}, \boldsymbol{\pi})$  is analytically not feasible so that the ML estimates have to be determined numerically. For mixture distributions this can be done with the Expectation Maximisation (EM) algorithm (Dempster et al. 1977). The mathematical details of the EM algorithm are provided in Sect. 2 of the supplementary material. The best mixture model  $\mathcal{M}_{\hat{K}}$  minimises the BIC in Eq. (11), where  $q_K = K + (K - 1)$  is the number of Poisson and (free) mixture weight parameters.<sup>1</sup> The output of the EM-algorithm is the best number of components  $\hat{K}_{MIX}$ , the corresponding  $T$ -by- $\hat{K}_{MIX}$  allocation probability matrix  $\hat{\mathbf{A}}_{MIX}$ , whose elements  $\Delta_{t,k}$  are the probabilities that time point  $t$  belongs to component  $k$ , and the vector of ML estimators  $\hat{\boldsymbol{\theta}}_{MIX}$ .

### 2.3.4 The frequentist Hidden Markov Poisson model (FREQ-HMM)

The key assumption of a hidden Markov model (HMM) with  $K$  components ('states') is that the (unobserved) elements  $v_1, \dots, v_T$  of the allocation vector  $\mathbf{V}$  follow a (homogeneous) first-order Markovian dependency. That is,  $\{v_t\}_{t=1, \dots, T}$  is considered a homogeneous Markov chain of order  $\tau = 1$  with the state space  $S = \{1, \dots, K\}$ , the initial distribution  $\Pi = (\pi_1, \dots, \pi_K)$ , where  $\pi_k \geq 0$  is the probability that  $v_1$  is equal to  $k$ , and the  $K$ -by- $K$  transition (probability) matrix  $\mathbf{A}$ , whose elements  $a_{i,j} \geq 0$  are the transition probabilities for a transition from state  $i$  to state  $j$ :  $a_{i,j} = P(v_{t+1} = j | v_t = i)$  for all  $t \in \{1, \dots, T - 1\}$ .<sup>2</sup> Assume that there are  $K$  state-dependent Poisson distributions so that each state  $k \in \{1, \dots, K\}$  corresponds to a Poisson distribution with parameter  $\theta_k$ . The data matrix  $\mathbf{D}$  is then interpreted as a sequence of its  $T$  columns,  $\mathbf{D}_{:,1}, \dots, \mathbf{D}_{:,T}$ , and  $v_t = k$  means that column  $\mathbf{D}_{:,t}$  is a vector of  $n$  realisations of the  $k$ th Poisson distribution with parameter  $\theta_k$ . Mathematically, this means:

$$p(\mathbf{D}_{:,t}|v_t = k) = p(\mathbf{D}_{:,t}|\theta_k) \quad (14)$$

where  $p(\mathbf{D}_{:,t}|\theta_k)$  was defined in Eq. (3). The Hidden Markov model is now fully specified and has the unknown parameters  $\Pi$ ,  $\mathbf{A}$  and  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)^\top$ . For given parameters  $(\Pi, \mathbf{A}, \boldsymbol{\theta})$ , the distribution of the unknown ('hidden') state sequence  $v_1, \dots, v_T$  can be inferred recursively with the forward and backward algorithm. And by combining the forward and backward algorithms with the EM algorithm,

<sup>1</sup> The  $K$  mixture weights fulfil:  $\sum_{k=1}^K \pi_k = 1$ .

<sup>2</sup> It holds:  $\sum_{k=1}^K \pi_k = 1$ , and  $\sum_{j=1}^K a_{i,j} = 1$  for all  $i$ .



the best HMM model  $\mathcal{M}_{\hat{K}}$ , which minimises the BIC in Eq. (11), can be numerically determined. The details of the inference procedure are provided in Sect. 3 of the supplementary material. For a HMM model with  $K$  components the total number of parameters is  $q_K = \mathcal{K} + (\mathcal{K} - 1) + (\mathcal{K}^2 - \mathcal{K})$ , i.e. the sum of the Poisson parameters, the free initial probability parameters and the free transition probability parameters.<sup>3</sup> The output of the EM algorithm, as described in Sect. 3 of the supplementary paper, is the best number of components  $\hat{K}_{HMM}$ , the corresponding  $T$ -by- $\hat{K}_{HMM}$  allocation probability matrix  $\hat{\Delta}_{HMM}$ , whose elements  $\Delta_{t,k}$  are the probabilities that time point  $t$  belongs to component  $k$ , and the ML-estimators  $\hat{\theta}_{HMM}$ .

### 2.4 The Bayesian framework

The Bayesian models employ a Poisson-Gamma model, for which the marginal likelihood  $p(\mathbf{D}|\mathbf{V})$  can be computed with Eq. (10). While the homogeneous model, described in Sect. 2.4.1, keeps  $K = 1$  fixed, the three non-homogeneous models have to infer  $K$  and the unknown allocation vector  $\mathbf{V}$ . In a Bayesian framework this means that prior distributions have to be imposed on  $\mathbf{V}$  and  $K$ . The three non-homogeneous models, described below, assume that the joint prior distribution can be factorized,  $p(\mathbf{V}, K) = p(\mathbf{V}|K) \cdot p(K)$ , and impose on  $K$  a truncated Poisson distribution with parameter  $\lambda$  and the truncation  $1 \leq K \leq K_{MAX}$  so that  $p(K) \propto \lambda^K \cdot \exp\{-\lambda\} \cdot (K!)^{-1}$ .

Subsequently, the prior on  $\mathbf{V}$  is specified conditional on  $K$ . The marginal likelihood  $p(\mathbf{D}|\mathbf{V})$  and the two prior distributions  $p(K)$  and  $p(\mathbf{V}|K)$  together fully specify the Bayesian model, and Markov Chain Monte Carlo (MCMC) simulations are used to generate samples  $(\mathbf{V}^{(1)}, K^{(1)}), \dots, (\mathbf{V}^{(R)}, K^{(R)})$  from the posterior distribution:

$$p(\mathbf{V}, K|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{V}) \cdot p(\mathbf{V}|K)p(K) \tag{15}$$

The Bayesian models, described below, differ only by the conditional prior  $p(\mathbf{V}|K)$ .

#### 2.4.1 The homogeneous Bayesian Poisson–Gamma model (BAYES–HOM)

The homogeneous Bayesian model assumes that the counts do not vary over time, so that  $K = 1$  and  $\mathbf{V} = (1, \dots, 1)^T =: \mathbf{1}$  and  $\mathbf{D}^{[1]} = \mathbf{D}$ . According to Eqs. (9–10), the marginal likelihood of the BAYES–HOM model is then given by

$$\begin{aligned} p(\mathbf{D}^{[1]}|\mathbf{V} = \mathbf{1}) &= \int_0^\infty p(\mathbf{D}|\theta)p(\theta|a, b)d\theta \\ &= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{\prod_{t=1}^T \prod_{i=1}^n (d_{i,t})!} \cdot \frac{\Gamma(a + \xi)}{(T \cdot n + b)^{\xi+a}} \end{aligned}$$

where  $d_{i,t}$  is the element in the  $i$ th row and  $t$ th column of  $\mathbf{D}$ , and  $\xi$  is the sum of all  $n \cdot T$  elements of  $\mathbf{D}$ .

<sup>3</sup> Note that:  $\sum_{k=1}^K \pi_k = 1$ , and  $\sum_{l=1}^K a_{k,l} = 1$  for  $k = 1, \dots, K$ .

### 2.4.2 The Bayesian changepoint Poisson–Gamma model (BAYES–CPS)

There are various possibilities to implement a Bayesian changepoint model, and here the classical one from Green (1995) is used. The prior on  $K$  is a truncated Poisson distribution, and each  $K$  is identified with  $K - 1$  changepoints  $c_1, \dots, c_{K-1}$  on the discrete set  $\{1, \dots, T - 1\}$ , where  $v_t = k$  if  $c_{k-1} < t \leq c_k$ , and  $c_0 := 1$  and  $c_K := T$  are pseudo changepoints. Conditional on  $K$ , the changepoints are assumed to be distributed like the even-numbered order statistics of  $L := 2(K - 1) + 1$  points uniformly and independently distributed on  $\{1, \dots, T - 1\}$ . This implies that changepoints cannot be located at neighbouring time points and induces the prior distribution:

$$P(\mathbf{V}|K) = \frac{1}{\binom{T-1}{2(K-1)+1}} \prod_{k=0}^{K-1} (c_{k+1} - c_k - 1) \quad (16)$$

The BAYES–CPS model is now fully specified and  $K$  and  $\mathbf{V}$  can be sampled from the posterior distribution  $p(\mathbf{V}, K | \mathbf{D})$ , defined in Eq. (15), with a Metropolis–Hastings MCMC sampling scheme, based on changepoint birth, death and re-allocation moves (Green 1995).

Given the current state at the  $r$ th MCMC iteration:  $(\mathbf{V}^{(r)}, K^{(r)})$ , where  $\mathbf{V}^{(r)}$  can be identified with the changepoint set:  $C^{(r)} = \{c_1, \dots, c_{K^{(r)}-1}\}$ , one of the three move types is randomly selected (e.g. each with probability  $1/3$ ) and performed. The three move types (i–iii) can be briefly described as follows:

- (i) In the changepoint reallocation move one changepoint  $c_j$  from the current changepoint set  $C^{(r)}$  is randomly selected, and the replacement changepoint is randomly drawn from the set  $\{c_{j-1} + 2, \dots, c_{j+1} - 2\}$ . The new set  $C^*$  gives the new candidate allocation vector  $\mathbf{V}^*$ ; the number of components stays unchanged:  $K^* = K^{(r)}$ .
- (ii) The changepoint birth move randomly draws the location of one single new changepoint from the set of all valid new changepoint locations:

$$B^\dagger := \left\{ c \in \{1, \dots, T - 1\} : |c - c_j| > 1 \forall j \in \{1, \dots, K^{(r)} - 1\} \right\} \quad (17)$$

Adding the new changepoint to  $C^{(r)}$  yields  $K^* = K^{(r)} + 1$ , and the new set  $C^*$ , which yields the new allocation vector  $\mathbf{V}^*$ .

- (iii) The changepoint death move is complementary to the birth move. It randomly selects one of the changepoints from  $C^{(r)}$  and proposes to delete it. This gives the new changepoint set  $C^*$  which yields the new candidate allocation vector  $\mathbf{V}^*$ , and  $K^* = K^{(r)} - 1$ .

For all three moves the Metropolis–Hastings acceptance probability for the new candidate state  $(\mathbf{V}^*, K^*)$  is given by  $A = \min\{1, R\}$ , with

$$R = \frac{p(\mathbf{D}|\mathbf{V}^*)}{p(\mathbf{D}|\mathbf{V}^{(r)})} \cdot \frac{p(\mathbf{V}^*|K^*)p(K^*)}{p(\mathbf{V}^{(r)}|K^{(r)})p(K^{(r)})} \cdot Q \quad (18)$$

where  $Q$  is the Hastings ratio, which can be computed straightforwardly for each of the three move types (see, e.g., Green (1995)). If the move is accepted, set  $\mathbf{V}^{(r+1)} = \mathbf{V}^*$  and  $K^{(r+1)} = K^*$ , or otherwise leave the state unchanged:  $\mathbf{V}^{(r+1)} = \mathbf{V}^{(r)}$  and  $K^{(r+1)} = K^{(r)}$ .

### 2.4.3 The Bayesian finite mixture Poisson–Gamma model (BAYES–MIX)

Here, the Bayesian finite mixture model instantiation and the Metropolis Hastings MCMC sampling scheme proposed by Nobile and Fearnside (2007) is employed. The prior on  $K$  is a truncated Poisson distribution, and conditional on  $K$ , a categorical distribution (with  $K$  categories) and probability parameters  $\mathbf{p} = (p_1, \dots, p_K)^\top$  is used as prior for the allocation variables  $v_1, \dots, v_T \in \{1, \dots, K\}$ . That is,  $\sum_{k=1}^K p_k = 1$  and  $p(v_t = k) = p_k$ . The probability of the allocation vector  $\mathbf{V} = (v_1, \dots, v_T)^\top$  is then given by:

$$p(\mathbf{V}|\mathbf{p}) = \prod_{k=1}^K (p_k)^{n_k} \tag{19}$$

where  $n_k = |\{t \in \{1, \dots, T\} : v_t = k\}|$  is the number of time points that are allocated to component  $k$  by  $\mathbf{V}$ . Imposing a conjugate Dirichlet distribution with parameters  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$  on  $\mathbf{p}$  and marginalizing over  $\mathbf{p}$ , yields the closed-form solution:

$$p(\mathbf{V}|K) = \int p(\mathbf{V}|\mathbf{p})p(\mathbf{p}|\boldsymbol{\alpha})d\mathbf{p} = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\sum_{k=1}^K (n_k + \alpha_k))} \prod_{k=1}^K \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{20}$$

The BAYES–MIX model is now fully specified, and the posterior distribution is invariant to permutations of the components’ labels if:  $\alpha_k = \alpha$ . A Metropolis-Hastings MCMC sampling scheme, proposed by Nobile and Fearnside (2007) and referred to as the “allocation sampler”, can be used to generate a sample from the posterior distribution in Eq. (15). The allocation sampler consists of a simple Gibbs move and five more involved Metropolis-Hastings moves. Given the current state at the  $r$ th iteration:  $(\mathbf{V}^{(r)}, K^{(r)})$  the Gibbs move keeps the number of components fixed,  $K^{(r+1)} = K^{(r)}$ , and just re-samples the value of one single allocation variable  $v_t^{(r)}$  from its full conditional distribution. This yields a new allocation vector  $\mathbf{V}^{(r+1)}$  with a re-sampled  $t$ th component  $v_t^{(r+1)}$ . As this Gibbs move has two disadvantages, Nobile and Fearnside (2007) propose to use five additional Metropolis Hastings MCMC moves. (i) As the Gibbs move yields only very small steps in the allocation vector configuration space, Nobile and Fearnside (2007) propose three additional Metropolis Hastings MCMC moves, referred to as the M1, M2 and M3 move, which also keep  $K^{(r)}$  fixed but allow for re-allocations of larger sub-sets of the allocation variables  $v_1^{(r)}, \dots, v_T^{(r)}$ . (ii) As neither the Gibbs move nor the M1-M3 moves can change the number of components, Nobile and Fearnside (2007) also propose a pair of moves, referred to as the Ejection-and Absorption move, which generate a new or delete an existing component, so that  $K^{(r+1)} = K^{(r)} + 1$  or  $K^{(r+1)} = K^{(r)} - 1$ , respectively. The technical details of the moves can be found in Nobile and Fearnside (2007).

#### 2.4.4 The Bayesian hidden Markov Poisson–Gamma model (BAYES–HMM)

The focus is on a Bayesian hidden Markov model instantiation, which was recently proposed in Grzegorzcyk (2016) in the context of non-homogeneous dynamic Bayesian network models. The prior on  $K$  follows a truncated Poisson distribution, and for each  $K$  a HMM model with  $K$  states is used to model the allocation vector  $\mathbf{V}$ . To this end,  $\mathbf{V}$  is identified with the temporally ordered sequence of its components:  $v_1, \dots, v_T$ , and it is assumed that the latter sequence describes a homogeneous first order Markov chain with a uniform initial distribution and a  $K$ -by- $K$  transition matrix  $\mathbf{A}$ .

Let  $a_{l,k}$  be the element in the  $l$ th row and  $k$ th column of the transition matrix  $\mathbf{A}$ .  $a_{l,k}$  is then the probability for a transition from component  $l$  to component  $k$ , and  $\sum_{k=1}^K a_{l,k} = 1$ . For a homogeneous Markov chain this means:  $a_{l,k} = P(v_t = k | v_{t-1} = l, \mathbf{A}, K)$  for all  $t$ , and hence:

$$\begin{aligned} p(\mathbf{V}|\mathbf{A}, K) &= p(v_1, \dots, v_T|\mathbf{A}, K) = p(v_1|K) \prod_{t=2}^T p(v_t|v_{t-1}, \mathbf{A}, K) \\ &= \frac{1}{K} \prod_{k=1}^K \prod_{l=1}^K (a_{l,k})^{n_{l,k}} \end{aligned} \quad (21)$$

where  $n_{l,k} = |\{t \in \{2, \dots, T\} : v_t = k \wedge v_{t-1} = l\}|$  is the number of transitions from  $l$  to  $k$  in the sequence  $v_1, \dots, v_T$ .

Each row  $\mathbf{A}_{l,\cdot}$  of the transition matrix  $\mathbf{A}$  defines the probability vector of a categorical random variable (with  $K$  categories), and on each vector  $\mathbf{A}_{l,\cdot}$  an independent Dirichlet prior with parameter vector  $\boldsymbol{\alpha}_l = (\alpha_{l,1}, \dots, \alpha_{l,K})^\top$  can be imposed:

$$p(\mathbf{A}_{l,\cdot}|\boldsymbol{\alpha}_l) = \frac{\prod_{k=1}^K \Gamma(\alpha_{l,k})}{\Gamma(\sum_{k=1}^K \alpha_{l,k})} \prod_{k=1}^K (a_{l,k})^{\alpha_{l,k}-1} \quad (22)$$

Marginalizing over the transition matrix  $\mathbf{A}$  in Eq. (21), i.e. marginalizing over the row vectors  $\mathbf{A}_{1,\cdot}, \dots, \mathbf{A}_{K,\cdot}$ , where each row vector  $\mathbf{A}_{l,\cdot}$  has an independent Dirichlet prior, defined in Eq. (22), gives the marginal distribution:

$$p(\mathbf{V}|K) = \int_{\mathbf{A}_{1,\cdot}} \dots \int_{\mathbf{A}_{K,\cdot}} p(\mathbf{V}|\mathbf{A}, K) \left\{ \prod_{l=1}^K p(\mathbf{A}_{l,\cdot}|\boldsymbol{\alpha}_l) \right\} d\mathbf{A}_{1,\cdot} \dots d\mathbf{A}_{K,\cdot} \quad (23)$$

Inserting Eq. (21) into Eq. (23) yields:

$$P(\mathbf{V}|K) = \frac{1}{K} \prod_{l=1}^K \left( \int_{\mathbf{A}_{l,\cdot}} P(\mathbf{A}_{l,\cdot}|\boldsymbol{\alpha}_l) \prod_{k=1}^K (a_{l,k})^{n_{l,k}} d\mathbf{A}_{l,\cdot} \right) \quad (24)$$

The inner integrals in Eq. (23) correspond to Multinomial-Dirichlet distributions, which can be computed in closed form:

$$P(\mathbf{V}|K) = \frac{1}{K} \prod_{l=1}^K \frac{\Gamma(\sum_{k=1}^K \alpha_{l,k})}{\Gamma(\sum_{k=1}^K n_{l,k} + \alpha_{l,k})} \prod_{k=1}^K \frac{\Gamma(n_{l,k} + \alpha_{l,k})}{\Gamma(\alpha_{l,k})} \tag{25}$$

The BAYES–HMM model is now fully specified, and with  $\alpha_{l,k} = \alpha$  in Eq. (22) the marginal distribution  $P(\mathbf{V}|K)$  in Eq. (25) is invariant to permutations of the states’ labels.

In principle, the allocation sampler from Nobile and Fearnside (2007) from Sect. 2.4.3 can also be used to generate a sample from the posterior distribution in Eq. (15). However, the allocation sampler moves have been developed for finite mixture models, where data points are treated as interchangeable units without any order. Hence, the allocation sampler moves are sub-optimal when a Markovian dependency structure among temporal data points is given. In Grzegorzcyk (2016) it has been shown that the performance of the allocation sampler can be significantly improved in terms of convergence and mixing by including two new pairs of complementary Metropolis–Hastings moves. These two pairs of moves, referred to as the ‘inclusion and exclusion moves’ and the ‘birth and death moves’ in Grzegorzcyk (2016), exploit the temporal structure of the data points. A detailed description of these moves can be found in Grzegorzcyk (2016).

### 3 Validation

Table 1 gives an overview to the models from Sect. 2, and Table 2 shows the outputs of those models. The outputs range from a scalar ML estimate (FREQ–HOM) to an MCMC sample of allocation vectors (e.g. BAYES–HMM). For each model the output inferred from  $\mathbf{D}$  can be used to estimate the probability of a new validation data set  $\mathbf{D}$ . Assume that in addition to the  $n$ -by- $T$  data matrix  $\mathbf{D}$  from Sect. 2.1, another  $\tilde{n}$ -by- $T$  data matrix  $\tilde{\mathbf{D}}$  is given and that the time points  $1, \dots, T$  in  $\mathbf{D}$  and  $\tilde{\mathbf{D}}$  can be mapped onto each other.

Each non-homogeneous **Bayesian model** with  $K$  components and allocation vector  $\mathbf{V}$  inferred from  $\mathbf{D}$  can then be used to subdivide the new data matrix  $\tilde{\mathbf{D}}$  into submatrices  $\tilde{\mathbf{D}}^{[1]}, \dots, \tilde{\mathbf{D}}^{[K]}$ , and the predictive probability for the  $k$ th sub-matrix  $\tilde{\mathbf{D}}^{[k]}$  is:

$$\begin{aligned} p(\tilde{\mathbf{D}}^{[k]}|\mathbf{D}^{[k]}) &= \int_0^\infty p(\tilde{\mathbf{D}}^{[k]}|\theta_k)p(\theta_k|\mathbf{D}^{[k]})d\theta_k \\ &= \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})} \cdot \frac{1}{\prod_{t=1}^{T_k} \prod_{i=1}^{\tilde{n}} (\tilde{a}_{i,t}^{[k]})!} \cdot \frac{\Gamma(\tilde{a} + \tilde{\xi}^{[k]})}{(T_k \cdot \tilde{n} + \tilde{b})^{\tilde{\xi}^{[k]} + \tilde{a}}} \end{aligned} \tag{26}$$

where  $T_k$  is the number of columns allocated to  $k$ ,  $\tilde{a} = a + \xi^{[k]}$  and  $\tilde{b} = b + n \cdot T_k$  are the posterior parameters, defined above Eq. (8),  $\tilde{\xi}^{[k]}$  is the sum of all  $\tilde{n} \cdot T_k$  elements of the  $\tilde{n}$ -by- $T_k$  sub-matrix  $\tilde{\mathbf{D}}^{[k]}$ , and  $\tilde{a}_{i,t}^{[k]}$  is the element in the  $i$ th row and  $t$ th column of  $\tilde{\mathbf{D}}^{[k]}$ .

**Table 2** Overview of the outputs of the eight (non-)homogeneous Poisson models

	Frequentist version (FREQ)	Bayesian version (BAYES)
Homogeneous model (HOM) with 1 parameter	$K = 1, V = (1, \dots, 1)^T$ scalar ML-estimator $\hat{\theta}$ estimated from the $n$ -by- $T$ values in $D$	$K = 1, V = (1, \dots, 1)^T$ 1-dimensional posterior distribution $p(\theta D)$ inferred from the $n$ -by- $T$ values in $D$
Changepoint model (CPS) with $\hat{K}$ parameters	$\hat{K}_{CPS}$ , one concrete allocation vector instantiation $\hat{V}_{CPS}$ , and a vector of the $\hat{K}_{CPS}$ component-specific ML-estimators $\hat{\theta}_{CPS}$	A sample $\{K^{(r)}, V^{(r)}\}_{r=1, \dots, R}$ , and for each $r$ a set of $K^{(r)}$ posterior distributions $p(\theta_{k,r}   D^{[k,r]})$ where $\theta_{k,r}$ and $D^{[k,r]}$ refer to the $k$ th component of the $r$ th sample
Finite mixture model (MIX) with $2\hat{K} - 1$ parameters	$\hat{K}_{MIX}$ , a vector of the $\hat{K}_{MIX}$ component-specific ML estimators $\hat{\theta}_{MIX}$ , and a $T$ -by- $K$ matrix $\hat{\Delta}_{MIX}$ , whose elements $\hat{\Delta}_{t,k}$ are the estimated allocation probabilities $p(v_t = k D)$	A sample $\{K^{(r)}, V^{(r)}\}_{r=1, \dots, R}$ , and for each $r$ a set of $K^{(r)}$ posterior distributions $p(\theta_{k,r}   D^{[k,r]})$ where $\theta_{k,r}$ and $D^{[k,r]}$ refer to the $k$ th component of the $r$ th sample
Hidden Markov model (HMM) with $\hat{K}^2 + 2\hat{K} - 1$ parameters	$\hat{K}$ , a vector of the $\hat{K}_{HMM}$ component-specific ML estimators $\hat{\theta}_{HMM}$ , and a $T$ -by- $K$ matrix $\hat{\Delta}_{HMM}$ whose elements $\Delta_{t,k}$ are the estimated allocation probabilities $p(v_t = k D)$	A sample $\{K^{(r)}, V^{(r)}\}_{r=1, \dots, R}$ , and for each $r$ a set of $K^{(r)}$ posterior distributions $p(\theta_{k,r}   D^{[k,r]})$ where $\theta_{k,r}$ and $D^{[k,r]}$ refer to the $k$ th component of the $r$ th sample

Detailed explanations are given in the main text

The (logarithmic) predictive probability of  $\tilde{\mathbf{D}}$  conditional on  $K$  and  $\mathbf{V}$  is then given by:

$$\log\{p(\tilde{\mathbf{D}}|\mathbf{D}, \mathbf{V}, K)\} = \sum_{k=1}^K \log\{p(\tilde{\mathbf{D}}^{[k]}|\mathbf{D}^{[k]})\} \tag{27}$$

For the homogeneous Bayesian model with  $K = 1$ ,  $\mathbf{D}^{[1]} = \mathbf{D}$ , and  $\tilde{\mathbf{D}}^{[1]} = \tilde{\mathbf{D}}$ , the predictive probability of  $\tilde{\mathbf{D}}$  can be computed analytically. For each non-homogeneous Bayesian model  $\mathcal{M}$  an MCMC simulation generates a sample  $\{\mathbf{V}^{(r)}, K^{(r)}\}_{r=1, \dots, R}$  from the posterior distribution  $p(K, \mathbf{V}|\mathbf{D})$  in Eq. (15), and the predictive probability of model  $\mathcal{M}$  can be approximated by:

$$\log\{p(\tilde{\mathbf{D}}|\mathbf{D}, \mathcal{M})\} \approx \frac{1}{R} \sum_{r=1}^R \log\{p(\tilde{\mathbf{D}}|\mathbf{D}, \mathbf{V}^{(r)}, K^{(r)})\} \tag{28}$$

For the **frequentist models** it can be proceeded similarly: After data matrix  $\mathbf{D}$  has been used to learn a model and its ML-estimates, the probability of the new data matrix  $\tilde{\mathbf{D}}$ , given the model and the ML estimates learnt from  $\mathbf{D}$ , is a measure which corresponds to a Bayesian predictive probability. The homogeneous model and the changepoint model both output concrete values for  $\hat{K}$  and  $\hat{\mathbf{V}}$ , and:

$$\log\{p(\tilde{\mathbf{D}}|\hat{K}, \hat{\mathbf{V}}, \hat{\boldsymbol{\theta}})\} = \sum_{k=1}^{\hat{K}} \log\{p(\tilde{\mathbf{D}}^{[k, \hat{\mathbf{V}}]}|\hat{\boldsymbol{\theta}}_k)\} \tag{29}$$

where  $\hat{K}$ ,  $\hat{\mathbf{V}}$ , and  $\hat{\boldsymbol{\theta}}$  are those values inferred from the training data  $\mathbf{D}$ ,  $\tilde{\mathbf{D}}^{[k, \hat{\mathbf{V}}]}$  is the  $k$ th submatrix of the validation data  $\tilde{\mathbf{D}}$  implied by  $\hat{\mathbf{V}}$ , and  $p(\tilde{\mathbf{D}}^{[k, \hat{\mathbf{V}}]}|\hat{\boldsymbol{\theta}}_k)$  can be computed with Eq. (2).<sup>4</sup> **FREQ-MIX** and **FREQ-HMM** both infer the number of components  $\hat{K}$  and the Poisson parameters  $\hat{\boldsymbol{\theta}}$  but no concrete allocation vector. They infer a  $\hat{K}$ -by- $T$  matrix  $\hat{\mathbf{\Delta}}$ , whose elements  $\hat{\Delta}_{k,t}$  are the probabilities that time point  $t$  is allocated to component  $k$ , symbolically  $\hat{\Delta}_{k,t} = \hat{p}(v_t = k|\mathbf{D})$ . The probability of the new data set  $\tilde{\mathbf{D}}$  is then given by:

$$\log\{p(\tilde{\mathbf{D}}|\hat{K}, \hat{\mathbf{\Delta}}, \hat{\boldsymbol{\theta}})\} = \sum_{t=1}^T \log\left\{\sum_{k=1}^{\hat{K}} \hat{\Delta}_{k,t} \cdot p(\tilde{\mathbf{D}}_{\cdot,t}|\hat{\boldsymbol{\theta}}_k)\right\} \tag{30}$$

## 4 Data

### 4.1 Synthetic data

Synthetic count data matrices are generated as follows: let  $\mathbf{V}^* = (v_1^*, \dots, v_T^*)^T$  be the true allocation vector, which allocates each time point  $t \in \{1, \dots, T\}$  to a component

<sup>4</sup> For the homogeneous model it holds:  $\hat{K} = 1$  and  $\tilde{\mathbf{D}}^{[1, \hat{\mathbf{V}}]} = \tilde{\mathbf{D}}$ .

$k \in \{1, \dots, K^*\}$ , where  $v_t^* = k$  means that  $t$  is allocated to  $k$ . Given  $\mathbf{V}^*$ ,  $n$ -by- $T$  data set matrices  $\mathbf{D}^*$  can be obtained by sampling each matrix element  $d_{i,t}^*$  independently from a Poisson distribution with parameter  $\theta_{v_t^*}$  ( $i = 1, \dots, n$  and  $t = 1, \dots, T$ ).

The focus of the study is on different allocation vectors  $\mathbf{V}^*$  with different component-specific Poisson parameters  $\theta_1, \dots, \theta_{K^*}$ . Let  $\mathbf{P}^* = (p_1, \dots, p_T)$  denote a row vector whose element  $p_t$  is the Poisson parameter for time point  $t$ . That is,  $p_t = \lambda$  means that  $\mathbf{V}^*$  allocates time point  $t$  to a component with Poisson parameter  $\theta_{v_t^*} = \lambda$ . The row vector  $\mathbf{P}^*$  will be referred to as the vector of Poisson parameters.

Let  $\mathbf{s}_m$  denote a row vector of length  $m$ , whose elements are all equal to  $s \in \mathbb{N}$ ,  $\mathbf{s}_m = (s, \dots, s)$ . The situation, where an allocation vector  $\mathbf{V}^*$  allocates  $T = 4 \cdot m$  time points to  $K^* = 4$  equidistant coherent segments of length  $m$ , with the four component-specific Poisson parameters  $\theta_1 = 1, \theta_2 = 5, \theta_3 = 3$ , and  $\theta_4 = 8$ , can then be defined compactly:

$$\mathbf{P}^* = (\underbrace{1, \dots, 1}_{m\text{-times}}, \underbrace{5, \dots, 5}_{m\text{-times}}, \underbrace{3, \dots, 3}_{m\text{-times}}, \underbrace{8, \dots, 8}_{m\text{-times}}) =: (\mathbf{1}_m, \mathbf{5}_m, \mathbf{3}_m, \mathbf{8}_m)$$

For the situation where the allocation vector follows a free mixture model, e.g., by allocating  $T = 2 \cdot m$  time points to  $K^* = 2$  components with Poisson parameters  $\theta_1 = 1$  and  $\theta_2 = 5$ , let  $\mathbf{P}^* = \mathbf{MIX}(\mathbf{1}_m, \mathbf{5}_m)$  denote that  $\mathbf{P}^*$  is a row vector whose elements are a random permutation of the elements of the vector  $(\mathbf{1}_m, \mathbf{5}_m)$ .

With regard to the real-world Taxi data, described in Sect. 4.2, each data matrix  $\mathbf{D}$  is built with  $T = 96$  columns (time points) and  $n \in \{1, 2, 4, 8, 16\}$  rows (independent samples per time point). An overview to the allocation schemes (vectors of Poisson parameters), employed in the comparative evaluation study, is given in Table 4 of the supplementary material. For each of the four allocation scenarios (HOM, CPS, MIX, and HMM) two different vectors of Poisson parameters are considered. Data matrices are built with a varying no. of rows  $n \in \{1, 2, 4, 8, 16\}$  and  $T = 96$  columns. For each of the resulting  $4 \cdot 2 \cdot 5 = 40$  combinations, 25 independent data matrix instantiations are generated, i.e. 1000 data matrices in total. Subsequently, for each of those 1000 data matrix instantiations a  $\tilde{n}$ -by- $T$  validation data matrix with  $\tilde{n} = 30$  and  $T = 96$  is sampled the same way (using the same vector of Poisson parameters).<sup>5</sup>

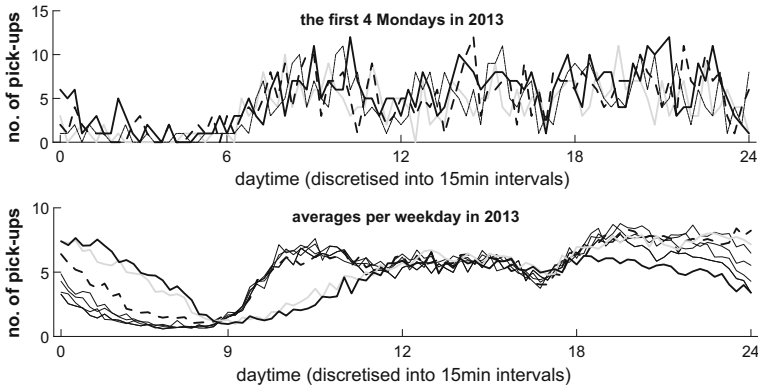
## 4.2 The New York City Taxi (NYCT) data from 2013

Through a ‘Freedom of Information Law’ request from the ‘New York City Taxi and Limousine Commission’ a dataset, covering information of about 700 million taxi trips in New York City (USA) from the calendar years 2010–2013, was published and stored by the University of Illinois (Donovan and Work 2015). In the NYCT database, for each trip various details are provided; e.g. (i) the number of transported passengers, (ii) the pick-up and drop-off dates and daytimes, (iii) the GPS coordinates, where the passenger(s) were picked up and dropped off.<sup>6</sup> In this paper the focus is on the pick-up

<sup>5</sup> Note that  $T$  and  $\tilde{n}$  have been set in accordance with the NYCT data, described in Sect. 4.2.

<sup>6</sup> The NYCT data can be downloaded from: <http://dx.doi.org/10.13012/J8PN93H8>





**Fig. 1** New York City Taxi pick-up time series. To shed some light onto the variability of the daily profiles, the *upper panel* shows the time series of the first four Mondays in 2013. The *lower panel* shows the seven weekday averages in 2013. Three weekdays with slightly deviating profiles have been highlighted: Sunday (**bold black**), Saturday (*grey*), and Friday (*dotted black*)

dates and daytimes of about 170 million taxi rides in the most recent year 2013, so that only a fractional amount of the data is used. Each pick-up is interpreted as a ‘taxi call’, so that it can be analysed how the number of taxi calls varies over the daytime. The data preparation can be summarised as follows: For each of about 170 million taxi rides from 2013 the pick-up date and daytime are extracted and down-sampled by a factor of 1000 (by randomly selecting 0.1% of the extracted samples), before all entries corresponding to US holidays are withdrawn.<sup>7</sup> Subsequently, there remain 169,596 date-and-time entries, which subdivide onto the 7 weekdays as indicated in Table 5 of the supplementary material. Discretising the daytimes into  $T = 96$  equidistant time intervals,<sup>8</sup> each covering 15 min of the 24-h day, and binning the pick-up times of each individual day into the  $T = 96$  time intervals, gives a 355-by-96 data matrix  $\mathbf{D}$ , whose elements  $d_{i,t}$  are the number of taxi pick-ups (or taxi calls) on the  $i$ th day in time interval  $t$ . Since the seven weekdays might show different patterns, the data set matrix  $\mathbf{D}$  is subdivided into seven  $n_w$ -by- $T$  sub-matrices  $\mathbf{D}_w$  ( $w = 1, \dots, 7$ ), where  $w$  indicates the weekday, and  $n_w \in \{46, 50, 51, 52\}$  varies with the weekdays (see Table 5 of the supplementary material). Figure 1 shows the number of Taxi calls for the first four Mondays in 2013 and the weekday averages.

In the study the weekdays are analysed separately, as they are likely to show different patterns. For each weekday  $n \in \{1, 2, 4, 8, 16\}$  rows (days) are randomly selected from  $\mathbf{D}_w$ , before  $\tilde{n} = 30$  of the remaining  $n_w - n$  rows are randomly selected to build a validation data matrix. Repeating this procedure 5-times independently yields 150 data matrix pairs  $\mathbf{D}_{w,n,u}$  and  $\tilde{\mathbf{D}}_{w,\tilde{n},u}$ , where  $w \in \{1, \dots, 7\}$  indicates the weekday,  $n \in \{1, 2, 4, 8, 16\}$  and  $\tilde{n} = 30$  indicate the number of rows of  $\mathbf{D}_{w,n,u}$  and  $\tilde{\mathbf{D}}_{w,\tilde{n},u}$ ,

<sup>7</sup> The following US holidays in 2013 are excluded: Jan 1 (New Year’s Day), Jan 21 (Martin Luther King), Feb 18 (Presidents’ Day), May 27 (Memorial Day), Jul 4 (Independence Day), Sep 2 (Labor Day), Oct 14 (Columbus Day), Nov 11 (Veterans Day), Nov 28 (Thanksgiving Day) and Dec 25 (Christmas Day).

<sup>8</sup> The time information is provided in seconds in the format: hh-mm-ss, ranging from 00-00-00 (midnight) to 23-59-59 (last second of the day).

and  $u \in \{1, \dots, 5\}$  indicates the replicate. Each  $\mathbf{D}_{w,n,u}$  is a  $n$ -by-96 matrix and each  $\tilde{\mathbf{D}}_{w,n,u}$  is a 30-by-96 matrix.<sup>9</sup>

## 5 Simulation details

For all models the maximal number of components is set to  $K_{MAX} = 10$ . In the Gamma priors, see Eq. (7), both hyperparameters  $a$  and  $b$  are set to 1 so as to obtain rather uninformative priors. In terms of equivalent sample sizes this setting corresponds to one ( $b = 1$ ) additional pseudo observation with one single taxi call ( $a = 1$ ) for each component. The hyperparameter of the truncated Poisson prior on the number of components of the non-homogeneous Bayesian models is set to  $\lambda = 1$ , meaning that a priori only one single component is expected ( $K = 1$ ). Furthermore, all hyperparameters of the Dirichlet priors of the BAYES-MIX and the BAYES-HMM model are set to 1. That is, it was set  $\boldsymbol{\alpha} = \mathbf{1}$  above Eq. (20) and  $\boldsymbol{\alpha}_l = \mathbf{1}$  ( $l = 1, \dots, K$ ) in Eq. (22). In terms of equivalent samples sizes this can be interpreted as one pseudo count per mixture component (BAYES-MIX) or transition (BAYES-HMM), respectively. The two homogeneous models (FREQ-HOM and BAYES-HOM) as well as the frequentist changepoint model (FREQ-CPS) always output deterministic solutions. The EM-algorithm, which is used for inferring the FREQ-MIX and the FREQ-HMM model, can get stuck in local optima. Therefore, the EM algorithm is run 10 times independently for each data set with different randomly sampled initialisations of the Poisson parameters.  $\epsilon = 0.001$  is used for the stop-criterion (see Tables 1, 2 in the supplementary paper). For each  $K$  the output with the highest maximal likelihood value was selected, while the other EM algorithm outputs were withdrawn.<sup>10</sup> (The maximal likelihood value was typically reached several times, suggesting that running the EM algorithm 10 times is sufficient for the analysed data.) The non-homogeneous Bayesian models are inferred with MCMC simulations, and a pre-study was performed to determine the required number of MCMC iterations. This pre-study was based on eight data sets with  $n = 16$ , one from each of the 8 allocation scenarios shown in Table 4 of the supplementary material. On each of these data sets 5 independent MCMC simulations with different allocation vector initialisations were performed. Trace-plot diagnostics of the quantity:  $\log(\text{Likelihood}) + \log(\text{Prior})$ , which is proportional to the log posterior probability, as well as scatter plots of the pairwise co-allocation probabilities,  $\hat{p}(v_{t_1} = v_{t_2} | \mathbf{D})$  for  $t_1, t_2 \in \{1, \dots, T\}$ , indicated that the following MCMC simulation setting is sufficient: The burn-in phase is set to 25,000 MCMC iterations, before  $R = 250$  equidistant samples are taken from the subsequent 25,000 MCMC iterations (sampling phase).

<sup>9</sup> Note that the same number of validation samples ( $\tilde{n} = 30$ ) is sampled for each  $n$  to ensure that the predictive probabilities  $p(\tilde{\mathbf{D}}_{w,\tilde{n},u} | \mathbf{D}_{w,n,u})$  are comparable for different  $n$ .

<sup>10</sup> Note that the mixture weights and the transition probabilities were always initialised uniformly, i.e.  $\pi_k = 1/K$  (FREQ-MIX) and  $a_{i,j} = 1/K$  (FREQ-HMM).

## 6 Comparative evaluation study

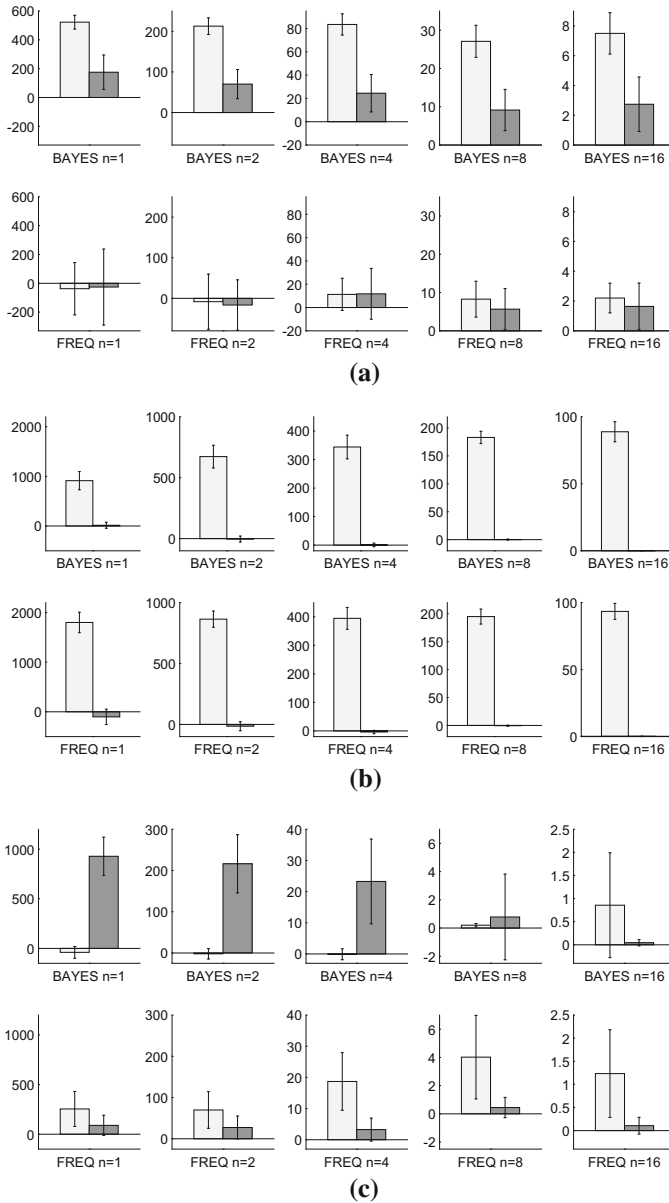
First, the synthetic data from Sect. 4.1 are analysed with the eight models listed in Tables 1, 2. The first finding is that the homogeneous models (FREQ–HOM and BAYES–HOM) yield substantially lower predictive probabilities than the non-homogeneous models for the non-homogeneous data. This is not unexpected, as the homogeneous models can per se not deal with non-homogeneity (e.g. changepoint-segmented data). For clarity of the plots, the results of the homogeneous models are therefore left out whenever their inclusion would have led to substantially different scales.

Figures 2, 3 and 4 show histograms of the average log predictive probability differences with separate histograms for the Bayesian and the frequentist models. Here, the four models (HOM, CPS, MIX and HMM) are compared independently within the Bayesian and within the frequentist framework without comparing the two paradigms (Bayesian vs. frequentist). In each histogram the models being most consistent with the data (i.e. being most consistent with the data generation process), are used as ‘reference’ models.<sup>11</sup> In a complementary study the four Bayesian models and the four frequentist models are compared in a pairwise manner. In Figs. 5 and 6 for each of the four models (HOM, CPS, MIX and HMM) the average log predictive probability differences (‘Bayesian results minus frequentist results’) are plotted against the average log predictive probability of the Bayesian and the frequentist results. The curves (‘differences vs. means’) are known as ‘Tukey mean-difference’ or ‘Bland-Altman’ plots, see Cleveland (1994) or Bland and Altman (1995).

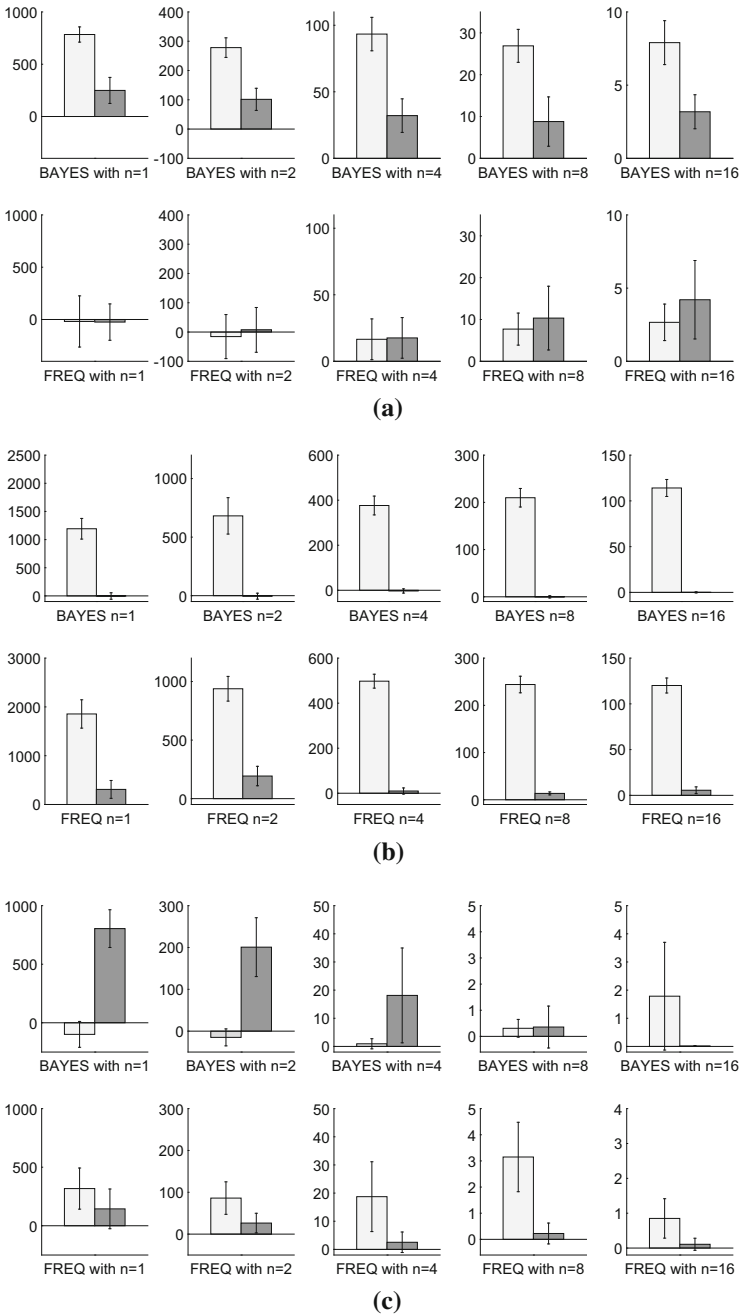
### 6.1 Global trends, Figs. 2–6

Before studying the individual results in more detail, two global trends become obvious. Figures 2, 3 and 4 show that the predictive probability differences between the non-homogeneous models get consistently lower as the number of samples  $n$  per time point  $t$  increases. The only exception appears for the mixture data [panels (b) in Figs. 2, 3], as the changepoint models (BAYES–CPS and FREQ–CPS) can per se not deal with mixture allocations, even when the sample size  $n$  is large. That is, for sufficiently informative data each non-homogeneous model can approximate all kinds of non-homogeneity, unless there is a clear mismatch between the dependency structure in the data and the inference model, as observed for the CPS models on mixture data. The second global finding from Figs. 5 and 6 is that the pairwise differences between the Bayesian and the frequentist models consistently converge towards zero as the number of samples  $n$  increases. That is, asymptotically for all four models the Bayesian variant and the frequentist variant perform equally well for all data scenarios.

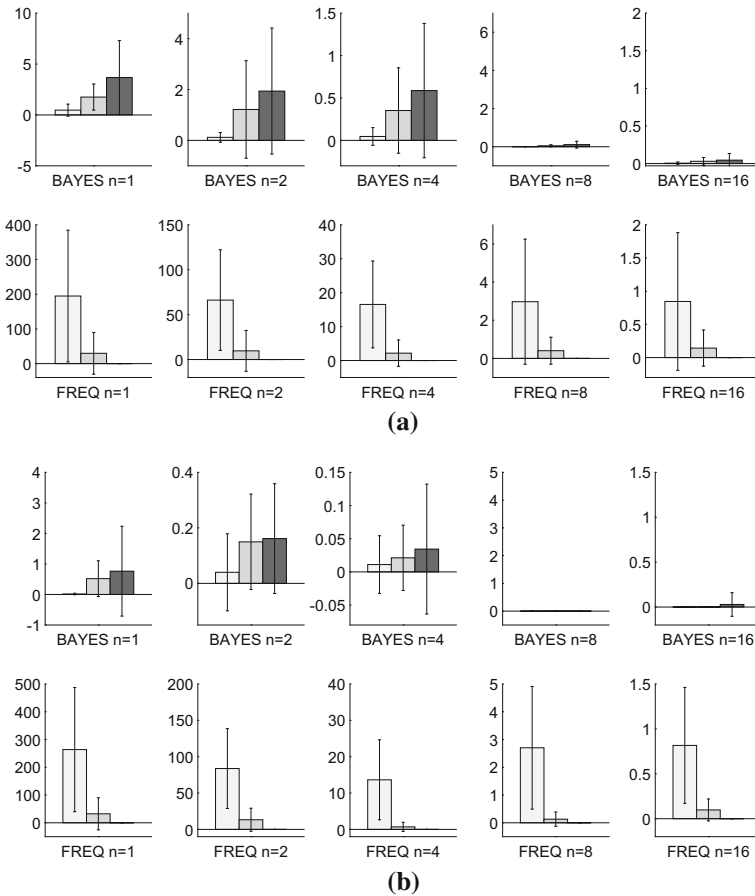
<sup>11</sup> For example the changepoint models (FREQ–CPS and BAYES–CPS) are used as references for the two changepoint-segmented data scenarios:  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m)$  and  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m, 5_m, 6_m)$ .



**Fig. 2** Cross-method comparison on synthetic data—part 1/3. **a–c** Histograms of the average log predictive probability differences for three non-homogeneous allocation scenarios with *error bars* representing standard deviations. In each **(a–c)** the *upper row* refers to the Bayesian models while the *bottom row* refers to the frequentist models. In each panel the differences between the reference (=most consistent with the data) model and the other two non-homogeneous models are shown. The homogeneous models led to substantially lower predictive probabilities and the results are therefore not shown. From *left to right* the sample size  $n$  increases and the scale of the y-axis changes. **a** Changepoint data  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m)$ . *Left* CPS–MIX, *right* CPS–HMM. **b** Mixture data  $\mathbf{P}^* = MIX(1_m, 5_m)$ . *Left* MIX–CPS, *right* MIX–HMM. **c** Hidden Markov data  $\mathbf{P}^* = (1_m, 5_m, 1_m, 5_m)$ . *Left* HMM–CPS, *right* HMM–MIX



**Fig. 3** Cross-method comparison on synthetic data—part 2/3. **a–c** Histograms of the average log predictive probability differences for three more non-homogeneous allocation scenarios. See caption of Fig. 2 for further details. **a** Changepoint data  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m, 5_m, 6_m)$ . *Left* CPS–MIX, *right* CPS–HMM. **b** Mixture data  $\mathbf{P}^* = MIX(1_m, 2_m, 4_m, 8_m)$ . *Left* MIX–CPS, *right* MIX–HMM. **c** Hidden Markov data  $\mathbf{P}^* = (1_m, 5_m, 1_m, 5_m, 1_m, 5_m)$ . *Left* HMM–CPS, *right* HMM–MIX

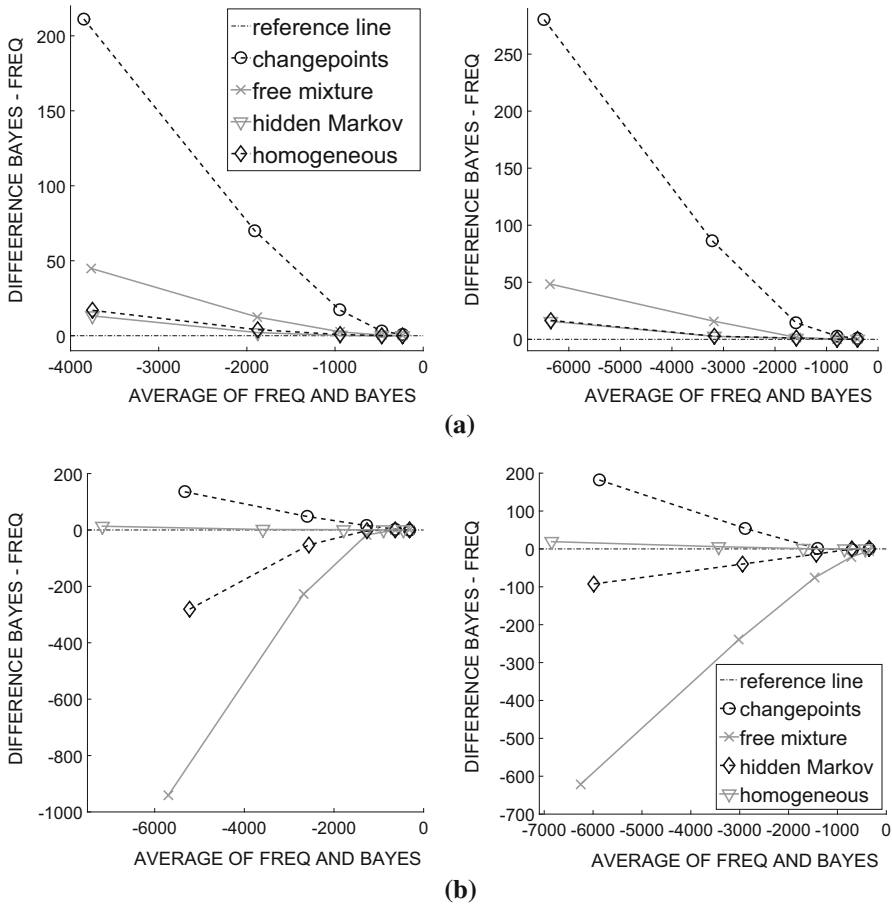


**Fig. 4** Cross-method comparison on synthetic data—part 3/3. **a, b** histograms of the average log predictive probability differences for the two homogeneous data scenarios. In both panels the *upper (lower) row* refers to the Bayesian (frequentist) models and the sample size  $n$  increases from *left to right*. In each panel the differences between the homogeneous (HOM) model and the three non-homogeneous models (CPS, MIX and HMM) are shown. Note that the FREQ–HMM model results never differed from the FREQ–HOM results and that the scales of the y-axis differ. **a** Homogeneous data  $\mathbf{P}^* = (\mathbf{1}_m)$ . *Left* HOM–CPS, *centre* HOM–MIX, *right* (HOM–HMM). **b** Homogeneous data  $\mathbf{P}^* = (\mathbf{5}_m)$ . *Left* HOM–CPS, *centre* HOM–MIX, *right* HOM–HMM

## 6.2 Specific trends, Figs. 2–4

### 6.2.1 Homogeneous data, Fig. 4

The differences in the log predictive probabilities are relatively low, except for the frequentist changepoint model (FREQ–CPS). That is, for homogeneous data only FREQ–CPS overfits the data for low sample sizes  $n$ , while the other non-homogenous models are never inferior to the homogeneous reference models. A further analysis (results not shown) reveals that FREQ–CPS yields low predictive probabilities, as it tends to impose too many changepoints. For low sample sizes  $n$ , single columns (or coherent sequences of columns) can—by chance—have exceptional large values.

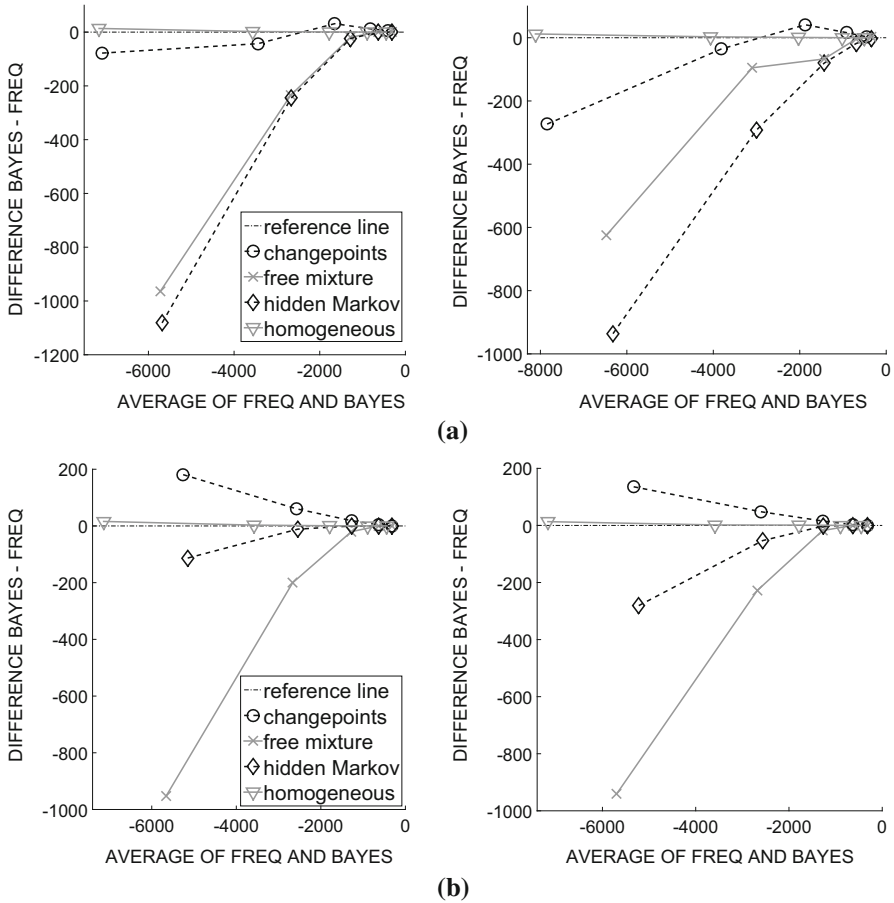


**Fig. 5** Tukey mean-difference plots to compare the performances of the frequentist and the Bayesian models on synthetic data—part 1/2. The *two panels* show the average log predictive probability differences for the homogeneous data (a) and for the changepoint segmented data (b). In the four plots for each of the four models (HOM, CPS, MIX and HMM) the log predictive probability differences (Bayesian–frequentist) have been plotted against the average log predictive probabilities (of Bayesian and frequentist). The *five symbols* on each *line* correspond to the values obtained for the five sample sizes  $n \in \{1, 2, 4, 8, 16\}$ . **a** Homogeneous data. *Left*  $\mathbf{P}^* = (1_m)$ , *right*  $\mathbf{P}^* = (5_m)$ . **b** Changepoint data. *Left*  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m)$ , *right*  $\mathbf{P}^* = (1_m, 2_m, 3_m, 4_m, 5_m, 6_m)$

Unlike the relatively robust Bayesian variant (BAYES–CPS), the frequentist changepoint model (FREQ–CPS) separates (or ‘cuts out’) those columns by setting two surrounding changepoints. The Bayesian changepoint variant appears to have a more effective penalty against over-fitting and does not allow for changepoints at neighbouring positions so that single columns cannot be ‘cut out’.

6.2.2 Changepoint-segmented data, panels (a) in Figs. 2 and 3

For all sample sizes  $n$  the Bayesian changepoint model (BAYES–CPS) performs significantly better than the Bayesian mixture model (BAYES–MIX) and the Bayesian



**Fig. 6** Tukey mean-difference plots to compare the performances of the frequentist and the Bayesian models on synthetic data—part 2/2. The two panels show the average log predictive probability differences plotted against the average log predictive probabilities for the mixture data (a) and for the hidden Markov model data (b); for further details see caption of Fig. 5. a Mixture data. Left  $P^* = \text{MIX}(1_m, 5_m)$ , right  $P^* = \text{MIX}(1_m, 2_m, 4_m, 8_m)$ . b Hidden Markov data. Left  $P^* = (1_m, 5_m, 1_m, 5_m)$ , right  $P^* = (1_m, 5_m, 1_m, 5_m, 1_m, 5_m)$ .

hidden Markov model (BAYES–HMM). The differences to the reference model (BAYES–CPS) show that BAYES–MIX performs consistently worse than BAYES–HMM. The reason becomes obvious from Fig. 8 in Sect. 7: BAYES–HMM approximates the underlying allocation better than BAYES–MIX, as BAYES–HMM—unlike BAYES–MIX—does not ignore the temporal order of the data points. For the frequentist models, the trend on changepoint-segmented data is slightly different: For small  $n \leq 2$  there is no difference in the performance of the non-homogeneous models. Only for  $n \geq 4$  the changepoint model (FREQ–CPS) performs better than its competitors. Thereby the mixture model (FREQ–MIX) performs better than the hidden Markov model (FREQ–HMM) for  $n \geq 4$ . Figure 9 in Sect. 7 suggests that this can be



explained as follows: *FREQ-MIX* possesses fewer parameters than *FREQ-HMM* (see Table 1) so that its BIC-penalty is lower (see Fig. 9). Consequently, *FREQ-MIX* can approximate the underlying segmentation better than *FREQ-HMM*. For low  $n \leq 2$  there is no difference between *FREQ-CPS* and the other models, as the frequentist changepoint model (*FREQ-CPS*) tends to overfit the data, as discussed above (see homogeneous data) and demonstrated in Sect. 7 (see Fig. 10).

### 6.2.3 Free-mixture data, panels (b) in Figs. 2 and 3

The Bayesian and the frequentist models show very similar trends. The changepoint models (*CPS*) are substantially outperformed by the free mixture reference models (*MIX*), while the hidden Markov models (*HMM*) are competitive to the mixture models (*MIX*). Only for small  $n \leq 2$  *FREQ-HMM* appears to be slightly inferior to *FREQ-MIX*. Figure 9 in Sect. 7 suggests that this is due to the higher BIC-penalty of the *FREQ-HMM* model. However, for the scenario  $\text{MIX}(\mathbf{1}_m, \mathbf{5}_m)$  and  $n = 1$  the increased BIC-penalty turns out to be advantageous for *FREQ-HMM*. Unlike *FREQ-HMM*, *FREQ-MIX* tends to overfit the data with  $n = 1$  by re-allocating outliers (columns with large values) to additional components.

### 6.2.4 Hidden-Markov data, panels (c) in Figs. 2 and 3

Among the Bayesian models, the mixture model (*BAYES-MIX*) is clearly outperformed by the hidden Markov model (*BAYES-HMM*) for low sample sizes  $n \leq 4$ . For larger sample sizes  $n \geq 8$  the differences decrease. The Bayesian changepoint model (*BAYES-CPS*) is competitive to *BAYES-HMM*, as it approximates the underlying dependency structure by additional changepoints; see Fig. 8 in Sect. 7.<sup>12</sup> For the frequentist models a complementary trend can be observed: The changepoint model (*FREQ-CPS*) is consistently inferior to the reference model (*FREQ-HMM*), while the mixture model (*FREQ-MIX*) is competitive for all  $n$ . Again *FREQ-CPS* tends to overfit the data (by cutting out columns with large realisations by surrounding changepoints), see Fig. 8 in Sect. 7. The disadvantage of *FREQ-MIX*, to ignore the temporal order of the data points, appears to be compensated by its relatively low BIC-penalty (see Fig. 9 in Sect. 7).

## 6.3 Bayesian versus frequentist

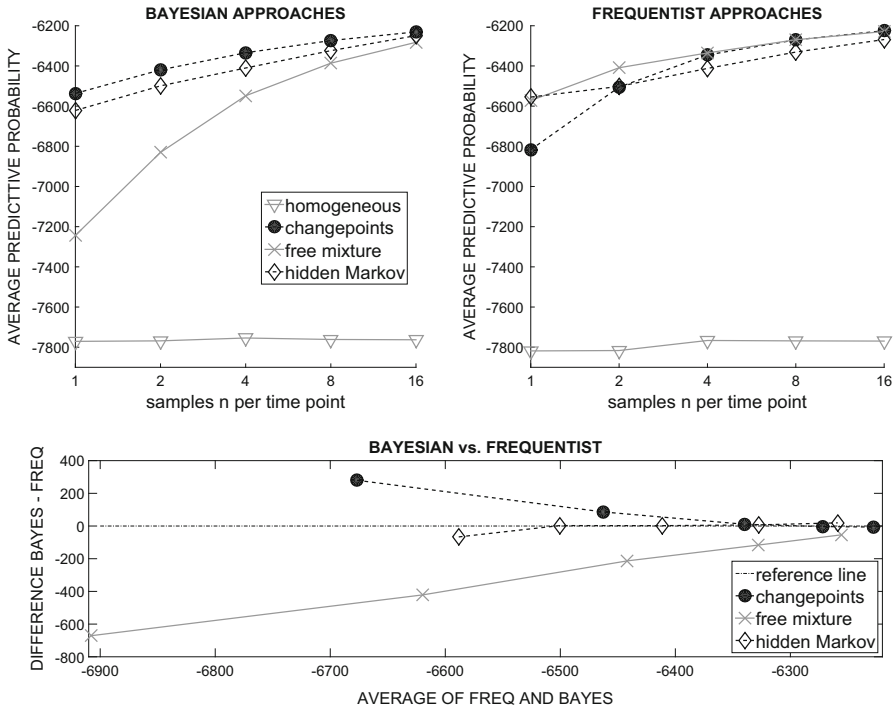
The Tukey-mean-difference plots of the pairwise predictive probability differences between the four Bayesian and the four frequentist models in Figs. 5 and 6 show that both paradigms yield nearly identical results for large sample sizes ( $n \geq 8$ ), while significant differences can be observed for small sample sizes  $n$ . Most remarkably are the following two trends: (i) Except for the mixture data [panel (a) in Fig. 6],

<sup>12</sup> Note that the selected Poisson means ( $\theta = 1$  and  $\theta = 5$ ) yield components with very dissimilar values. This makes it easy for the changepoint model to distinguish them and to approximate the non-stationarity by setting an increased number of changepoints, e.g. 3 changepoints for  $(\mathbf{1}_m, \mathbf{5}_m, \mathbf{1}_m, \mathbf{5}_m)$ .

for low sample sizes  $n$  the Bayesian changepoint model (BAYES–CPS) is superior to the frequentist changepoint model (FREQ–CPS). (ii) Except for the homogeneous data [panel (a) in Fig. 5], the frequentist hidden Markov model (FREQ–HMM) and especially the frequentist mixture model (FREQ–MIX) are superior to their Bayesian counterparts (BAYES–HMM and BAYES–MIX). The reason for the superiority of the Bayesian changepoint model (BAYES–CPS) is that the frequentist variant (FREQ–CPS) has a clear tendency towards over-fitting for uninformative data (for low  $n$ ); see Figs. 8 and 10 in Sect. 7 for more details. Unlike the Bayesian changepoint-model instantiation, FREQ–CPS infers only one single allocation vector (changepoint set) without any model-averaging. The low number of parameters of FREQ–CPS (see Table 1) yields a relatively low BIC-penalty. Single columns of the data matrix, which by chance have larger values than the other columns, can be ‘cut out’ so that the FREQ–CPS model is very susceptible to over-fitting. On the other hand, the superiority of the frequentist mixture (FREQ–MIX) and the frequentist hidden Markov model (FREQ–HMM) over its Bayesian counterparts can be explained by the Multinomial-Dirichlet prior on the allocation vector. Both Bayesian models (BAYES–MIX and BAYES–HMM) employ Multinomial-Dirichlet priors for the allocation vectors, which can yield very strong prior penalties for non-homogeneous allocation vectors. As shown in Fig. 9 in Sect. 7, BAYES–MIX is strongly penalized for all forms of non-homogeneity and BAYES–HMM is strongly penalized for mixture allocation vectors. This bottleneck of the Multinomial-Dirichlet prior for allocation vectors has already been analysed and discussed in Grzegorzcyk et al. (2010) and renders the Bayesian model variants inappropriate for small samples sizes  $n$ , i.e. for uninformative data, where the effect of the likelihood is small compared to the effect of the Multinomial-Dirichlet prior.

#### 6.4 The New York City Taxi (NYCT) data

The results for the NYCT data are shown in Fig. 7. The top plots shows the average log predictive probabilities for the Bayesian models (left) and the frequentist models (right) for different sample sizes  $n$ . The lower panel provides Tukey mean-difference plots to visualise the pairwise differences between the Bayesian and the frequentist models. The upper plots show that the homogeneous models (FREQ–HOM and BAYES–HOM) perform show the worst performance on the NYCT data. This is not unexpected, as Fig. 1 shows that the Taxi pick-up data are clearly non-stationary. Among the Bayesian models, the changepoint-model (BAYES–CPS) performs best for all sample sizes  $n$ , and asymptotically (i.e. as  $n$  increases) the non-homogeneous Bayesian models perform equally well. Among the frequentist models the mixture model (FREQ–MIX) shows the best performance. For  $n = 1$  FREQ–MIX and FREQ–HMM perform approximately equally well, while FREQ–CPS performs significantly worse. For  $n = 2$  the FREQ–MIX model performs better than both competitors. And for larger  $n$  ( $n \geq 4$ ) FREQ–MIX and FREQ–CPS perform equally well, while FREQ–HMM performs slightly worse. The Tukey mean-difference plot in the bottom of Fig. 7 shows that the Bayesian and frequentist models asymptotically perform equally well. For the lower samples sizes  $n$  the trends are consistent with the earlier observations for the synthetic data. The Bayesian changepoint model (BAYES–CPS) is superior to its

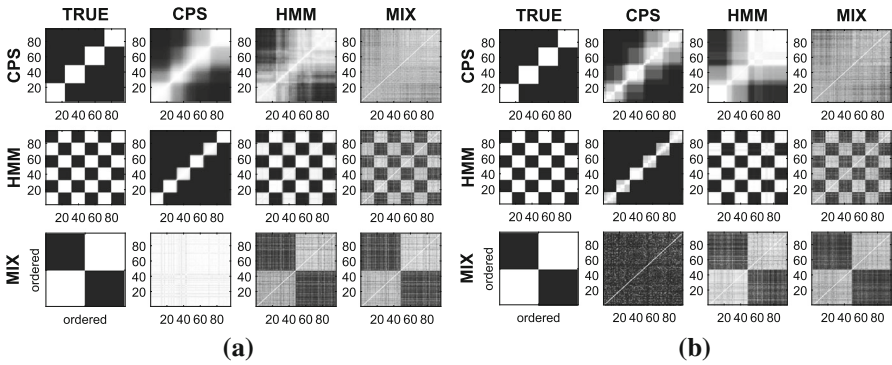


**Fig. 7** Results for the New York City Taxi data. In the upper plots the average log predictive probabilities (averaged across 35 data sets; i.e. 5 randomly sampled data instantiations per weekday) of the Bayesian models (*upper left*) and the frequentist models (*upper right*) have been plotted against the number of samples  $n$  per time point  $t$ . In the *lower plot* for each of the three non-homogeneous models (CPS, MIX and HMM) the average log predictive probability differences (BAYES–FREQ) have been plotted against the average log predictive probability of FREQ and BAYES. The five symbols on each line correspond to the values obtained for the sample sizes  $n \in \{1, 2, 4, 8, 16\}$ . In the lower plot the Bayesian (frequentist) model is superior when the *curve/symbol* is *above (below)* the reference line

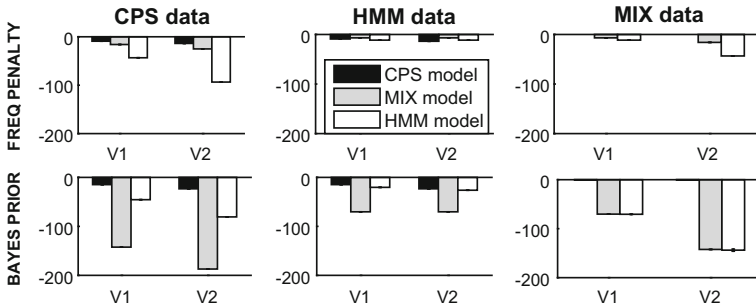
frequentist counterpart (FREQ–CPS), while the opposite trend can be observed for the mixture and the hidden Markov model. The  $p$  values of two-sided one-sample  $t$  tests for the predictive probability differences between the best Bayesian model (BAYES–CPS) and the best frequentist model (FREQ–CPS) are computed to determine whether the performances differ significantly for any  $n$ . Given the relatively small  $t$  test sample size of  $n_d = 7$  weekdays,<sup>13</sup> the five  $p$  values (for  $n = 1, 2, 4, 8, 16$ ) are higher than the standard level  $\alpha = 0.05$ , indicating that the best Bayesian and the best frequentist model are performing approximately equally well on the NYCT data.<sup>14</sup>

<sup>13</sup> That is one (average) predictive probability difference per weekday; the differences for the 5 data replicates per weekday are averaged, as they are very similar to each other.

<sup>14</sup>  $p$  values:  $p = 0.30$  ( $n = 1$ ),  $0.53$  ( $n = 2$ ),  $p = 0.96$  ( $n = 4$ ),  $p = 0.45$  ( $n = 8$ ), and  $p = 0.72$  ( $n = 16$ ).



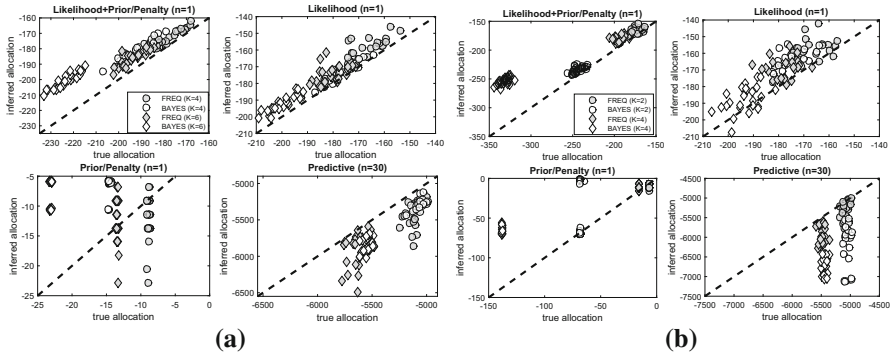
**Fig. 8** Heatmap representations of the inferred connectivity structures for the non-homogeneous models. **a** Refers to the Bayesian models, **b** to the frequentist models. Both panels are arranged as 3-by-4 matrices with rows corresponding to the true allocation vectors: CPS:  $(1_m, 2_m, 3_m, 4_m)$  (top), HMM:  $(1_m, 5_m, 1_m, 5_m, 1_m, 5_m)$  (centre), and MIX: MIX  $(1_m, 5_m)$  (bottom). The first columns show the true connectivity structures, and rows 2–4 correspond to the three non-homogeneous models: CPS, HMM, and MIX. The heatmaps give the inferred probabilities  $p(v_s = v_t | \mathbf{D})$  of two points  $s$  and  $t$  belonging to the same component. The probabilities are represented by a grey shading, where white corresponds to 1, and black corresponds to 0. The axes refer to the  $T = 96$  time points. All connectivity probabilities  $p(v_s = v_t)$  are averaged over 25 data instantiations with  $n = 1$  observation per time point. The time points of the MIX data in the last rows have been ordered w.r.t. the two mixture components. **a** Heatmaps of Bayesian model variants. **b** Heatmaps of frequentist model variants



**Fig. 9** Comparison of penalty terms for the non-homogeneous models. The plot is arranged as a 2-by-3 matrix, and the rows refer to the frequentist (top) and the Bayesian (bottom) models. The columns refer to three different allocation scenarios (CPS data, HMM data, and MIX data) and for each scenario two variants (V1 and V2) are distinguished. The segmentation schemes correspond to those used in the comparative evaluation study in Sect. 6, see Table 2 in the supplementary material for an overview. The bars give the penalties (BIC or prior probability) of the three models (CPS, MIX and HMM) for the true underlying allocation. As the CPS models cannot infer the true allocation of mixture data, the bars are not shown. For the CPS models it is assumed that they approximate HMM data by additional changepoints, e.g. (HMM, V1):  $(1_m, 5_m, 1_m, 5_m)$  is approximated by setting 3 changepoints

### 7 Further model diagnostics

This section provides additional diagnostic plots for the synthetic data, analysed in Sect. 6. The goal is to shed more light onto the relative merits and shortcomings of the models under comparison and to derive some conclusions of general validity. Since



**Fig. 10** Comparing the best scoring models with the ‘true’ models. **a** Diagnostics for the changepoint models (CPS–FREQ and CPS–BAYES) on changepoint-segmented data:  $(1_m, 2_m, 3_m, 4_m)$  (circles) and  $(1_m, 2_m, 3_m, 4_m, 5_m, 6_m)$  (diamonds). **b** Diagnostics for the mixture models (MIX–FREQ and MIX–BAYES) on mixture data: MIX  $(1_m, 5_m)$  (circles) and MIX  $(1_m, 2_m, 4_m, 8_m)$  (diamonds). In both panels there are 4 scatter plots, in which features of the ‘best’ models (BAYES: highest posterior, FREQ: best BIC) are plotted against the corresponding features of the ‘true’ models, using the true allocations. Symbols that refer to Bayesian models are white-colored, the frequentist symbols are grey-colored. Upper left scores versus scores (BAYES:  $\log(\text{likelihood} + \text{prior})$ , FREQ: BIC value); upper right log-likelihood versus log-likelihood; lower left penalty versus penalty (BAYES:  $\log(\text{prior})$ , FREQ: BIC-penalty); and lower right predictive probability versus predictive probability. See supplementary material for a similar plot for time series with  $n = 16$  data points per time point. **a** Changepoint models (FREQ and BAYES) on changepoint data with  $n = 1$ . **b** Mixture models (FREQ and BAYES) on mixture data with  $n = 1$

the predictive probabilities differed most significantly for sparse data, the focus of this section is on time series with only  $n = 1$  observation per time point. The first analysis investigates to which extent the non-homogeneous models are capable of inferring the true underlying allocation vectors. To this end, for three of the allocation scenarios, namely  $(1_m, 2_m, 3_m, 4_m)$ ,  $(1_m, 5_m, 1_m, 5_m, 1_m, 5_m)$ , and MIX  $(1_m, 5_m)$ , the average probabilities,  $p(v_s = v_t | \mathbf{D})$ , that two time points  $s$  and  $t$  are allocated to the same component, are computed. The two panels of Fig. 8 show heatmap representations of those connectivity probabilities for the Bayesian and for the frequentist model variants. The heatmaps show the following trends: **(1st rows, CPS data)**: The MIX models fail to infer the true allocation; the time points are not sufficiently separated and the heatmaps appear unstructured. The HMM models perform better and their heatmaps show that the first time points and the last time points both build connected segments; only the centre changepoint is improperly inferred. The heatmaps of the CPS models are very similar to the true heatmap. Although the centre changepoint is a little bit diffuse, it can be seen that the data consist of 3–4 connected segments. **(2nd rows, HMM data)**: The HMM models reconstruct the true allocation almost perfectly, while the CPS models segment the data into too many segments ( $K = 6$ ). This is the reason why changepoint-based models are inappropriate for HMM data. As discussed earlier, CPS models cannot re-visit components once left so that HMM allocations are not in their allocation vector configuration spaces. CPS-models have to approximate HMM allocations by setting additional changepoints; this is the reason for their suboptimal performances on HMM data. The mixture models, in principle, infer the right trends. But it can be seen from the heatmaps that the separations between

the components are weaker than those of the HMM models. (**3rd rows, MIX data**): The CPS models fail to infer the segmentation of the mixture data. The HMM and the MIX models perform approximately equally well and correctly divide the time points into  $K = 2$  components, though the inferred separations appear to be slightly too weak. (**BAYESIAN vs. FREQ**): The heatmaps of the Bayesian and the frequentist model variants are very similar, except for the heatmaps of the changepoint-models on mixture data (see bottom rows, 2nd columns in Fig. 8). While CPS-BAYES does not separate the time points, the frequentist counterpart (CPS-FREQ) shows the opposite behaviour: CPS-FREQ separates the time points into (too) many short segments.

Overall, the findings are in agreement with the results from Sect. 6: The heatmaps in Fig. 8 confirm that the HMM models can properly infer HMM data and mixture data while their performances on changepoint-segmented data are suboptimal (the changepoints are not properly inferred). The CPS models and the MIX models completely fail for certain allocation scenarios: The CPS model cannot deal with mixture data and can only approximate the segmentation of HMM data by setting too many changepoints. The mixture models, which do not exploit the temporal order of the data points, are inappropriate for changepoint-segmented data. Another finding is the difference between the two changepoint models: CPS-BAYES and CPS-FREQ. For mixture data CPS-BAYES infers ‘undercomplex’ allocations with too few changepoints (mostly  $K = 1$ ), while CPS-FREQ infers ‘overcomplex’ allocations with too many changepoints (even the maximum of  $K = 10$  changepoints is reached). The latter finding suggests that the frequentist changepoint model has a tendency towards overfitting the data by setting too many changepoints.

A comparison of the penalty terms for the true allocations is given in Fig. 9. The top row of Fig. 9 shows the BIC-penalties of the frequentist models, the bottom row shows the Bayesian (log) prior probability penalties. (**CPS data and HMM data**) The penalties of the frequentist and the Bayesian models are comparable except for the MIX model. For both types of data the penalties of the Bayesian mixture model are substantially higher than the penalties of its frequentist counterpart. (**MIX data**) The CPS model cannot infer the true mixture allocations. The Bayesian HMM model and the Bayesian MIX model are penalized significantly stronger than their frequentist counterparts. (**MIX-BAYES vs. HMM-BAYES**) The bottom row of Fig. 9 shows that the penalties of HMM-BAYES and MIX-BAYES are nearly identical for mixture data, while MIX-BAYES has substantially higher penalties for CPS data and MIX data. That is, the Bayesian mixture model (MIX-BAYES) is ‘over-penalized’ for non-mixture data.

The last diagnostic compares the performances of the Bayesian and the frequentist models with respect to over-fitting issues. To this end, certain ‘features’ of the ‘best’ inferred models (FREQ-models minimising the BIC score; BAYES-models with the highest posterior score) are compared with the corresponding ‘features’ of the true models, i.e. models which are based on the true allocation vectors. The ‘features’ are: (i) the scores, (ii) the (marginal) likelihood values, (iii) the prior penalty terms, and (iv) the predictive probabilities for new data. Figure 10 gives scatter plots in which the features of the best inferred models are plotted against the features of the true models. The scatter plots can be interpreted as follows: Symbols are above (below) the diagonal when the feature of the best inferred model is higher (lower) than the feature

of the true model. (Figure 10a, **changepoint-divided data**): The best inferred models yield higher scores and higher likelihoods than the true models. That is, both models variants (BAYES and FREQ) fit the data better than the true models. But the scatter plot of the penalty terms show that the Bayesian CPS model consistently infers ‘under-penalized’ models (i.e. models with too few changepoints) while the frequentist CPS model also infers ‘over-penalized’ models (i.e. models with too many changepoints). The scatter plot of the predictive probabilities shows the implication. The inferred models are inferior to the true models (all symbols are below the diagonal), but the predictive probabilities of the ‘under-complex’ CPS–BAYES model are better than those of the ‘over-complex’ CPS–FREQ model. Thereby the most ‘over-complex’ CPS–FREQ models yield the lowest predictive probabilities. This clearly shows that the frequentist changepoint model is more susceptible to over-fitting than its Bayesian counterpart. (Figure 10b, **mixture data**): The upper panels show that MIX–FREQ again consistently overfits the data, while the MIX–BAYES model sometimes yields lower likelihoods than the true model (see diamond symbols in the upper right panel). The scatter plot of the penalty term shows that MIX–BAYES infers ‘undercomplex’ models (with too few mixture components) for scenario (V2, mixture with  $K = 4$  components) and sometimes for scenario (V1, mixture with  $K = 2$  components). This suggests that MIX–BAYES overpenalizes the complexity of the allocation vectors, so that ‘undercomplex’ models are inferred. This explains why MIX–FREQ is superior to the over-penalized (and thus ‘undercomplex’) MIX–BAYES model (see scatter plot of the predictive probabilities). Figure 1 in the supplementary material shows the same diagnostics for time series with  $n = 16$  data points per time point. The results show that those issues of under- and over-penalisation diminish/disappear as the data get more informative.

## Summary

The additional diagnostics, shown in Figs. 8, 9 and 10, confirm four of the empirical findings from Sect. 6.

1. In Sect. 6 the CPS-models showed suboptimal performances for HMM and MIX data. Figure 8 shows that CPS models are inferior to HMM models for both types of data because the true allocations are not part of their allocation vector configuration spaces. Only for changepoint segmented data, where the HMM models do not properly infer the true allocation, the CPS models are superior to the HMM models.
2. In Sect. 6 the MIX-models showed suboptimal performances for CPS and HMM data. Figure 8 shows that the MIX models cannot properly infer CPS and HMM allocations, while the HMM models show moderate performances for all types of data.
3. In Sect. 6 the Bayesian mixture model (MIX–BAYES) was found to be inferior to its frequentist counterpart (MIX–FREQ). As seen from Figs. 9 and 10, the Bayesian mixture variant is over-penalised. This renders the frequentist HMM model preferable to the Bayesian mixture model.
4. In Sect. 6 it was also found that the Bayesian CPS model is superior to its frequentist counterpart (CPS–FREQ). As seen from Fig. 10a, the frequentist changepoint

model tends to overfit the data. This renders the Bayesian changepoint model preferable to the frequentist CPS model.

## 8 Conclusions

In this paper the results of a comparative evaluation study on eight (non-)homogeneous models for (Poisson) count data were presented. The study was performed on various synthetic data sets and on taxi pick-up counts, extracted from the recently published New York City Taxi (NYCT) database, described in Sect. 4. For the study the standard homogeneous Poisson model (HOM) and three non-homogeneous Poisson models, namely a changepoint model (CPS), a free mixture model (MIX) and a hidden Markov model (HMM), were implemented following the frequentist paradigm (FREQ) and the Bayesian paradigm (BAYES); see Tables 1 and 2 in Sect. 2 for an overview. The empirical findings from Sects. 6 and 7 suggest the following conclusions:

Asymptotically, i.e. for sufficiently informative data (here: quantified in terms of the sample size  $n$  per time point  $t$ ), there is no difference between the paradigms. The Bayesian and the frequentist models perform equally well. For less informative data (here: for small  $n$ ) there are significant differences, as described in more detail below. While the homogeneous model variants (FREQ–HOM and BAYES–HOM) cannot deal with non-homogeneity, the non-homogeneous models, except for the frequentist changepoint model (FREQ–CPS), do not overfit homogeneous data. Thus, it can be recommended applying non-homogeneous approaches, even if the data might be homogeneous. Moreover, if the data is informative enough, in both frameworks (Bayesian and frequentist) all three non-homogeneous models can approximate all kinds of non-homogeneity, unless there is a clear mismatch between the model and the underlying data. E.g. in Sects. 6 and 7 it was found that the changepoint models (FREQ–CPS and BAYES–CPS) perform badly for mixture data. The hidden Markov models (FREQ–HMM and BAYES–HMM) appear to be superior to the mixture models (FREQ–MIX and BAYES–MIX), since they are competitive on free-mixture data, and superior on hidden Markov and changepoint-segmented data.<sup>15</sup> In a pairwise comparison of the four Bayesian and the four frequentist models it was found for less informative data (here: small  $n$ ) that the Bayesian changepoint model (BAYES–CPS) is superior to its frequentist counterpart (FREQ–CPS), while the opposite trend could be observed for the mixture model and the hidden Markov model. The superiority of the Bayesian changepoint model (BAYES–CPS) over the frequentist variant (FREQ–CPS) is due to the fact the frequentist model variant is very susceptible to over-fitting (see Fig. 10a in Sect. 7). The inferiority of the Bayesian free mixture model (BAYES–MIX) and the Bayesian hidden Markov (BAYES–HMM) model to their frequentist counterparts is caused by the allocation vector priors. Both Bayesian models employ the Multinomial-Dirichlet prior, which is known to impose a very strong penalty on non-homogeneous allocations (see Fig. 9 in Sect. 7), rendering the Bayesian variants inappropriate for less informative data sets (here: for small samples

<sup>15</sup> Though still worse than the changepoint models (FREQ–CPS and BAYES–CPS), which can be seen as reference models for changepoint-segmented data.



sizes  $n$ ) than the two frequentist variants **FREQ–MIX** and **FREQ–HMM** (see, e.g., Fig. 10b).

For the real-world New York City Taxi (NYCT) data very similar trends could be observed. For sufficiently informative data (here: for large  $n$ ) all non-homogeneous models led to approximately identical results, and for uninformative data (here: for small  $n$ ) it was found that the Bayesian changepoint model (**BAYES–CPS**) performs better than its frequentist counterpart, while the frequentist mixture model (**FREQ–MIX**) performs better than its Bayesian counterpart (see Fig. 7). It was also found that the performances of the best Bayesian model (**BAYES–CPS**) and the best frequentist model (**FREQ–MIX**) do not differ significantly for any  $n$ . Finally, it should be noted that potential ‘overdispersion’ problems (i.e. potential violations of the Poisson model assumption) were not taken into account within the presented study. Unlike for the synthetic data, where all data points were actually sampled from Poisson distributions so that over-dispersion problems could not arise, overdispersion could have been present for the real-world NYCT data application. Therefore, the (undispersed) models, considered here, might have been suboptimal for the NYCT data and better results could perhaps have been obtained by taking the potential over-dispersion properly into account; e.g. by replacing the Poisson distribution by the more flexible negative binomial distribution or by applying more advanced Poisson model approaches.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Auger I, Lawrence C (1989) Algorithms for the optimal identification of segment neighborhoods. *Bull Math Biol* 51:39–54
- Bland J, Altman D (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 346:1085–1087
- Cleveland W (1994) *The elements of graphing data*, 2nd edn. Hobart Press, summit
- Dempster AP, Laird NM, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–38
- Donovan B, Work D (2015) Using coarse GPS data to quantify city-scale transportation system resilience to extreme events. In: *Proceedings of the transportation research board 94th annual meeting*. Washington, to appear
- Green P (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732
- Grzegorzczak M (2016) A non-homogeneous dynamic Bayesian network with a hidden Markov model dependency structure among the temporal data points. *Mach Learn* 102:155–207
- Grzegorzczak M, Husmeier D, Rahnenführer R (2010) Modelling non-stationary gene regulatory processes. *Adv Bioinform* 2010, Article ID 749848
- Nobile A, Fearnside A (2007) Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Stat Comput* 17:147–162
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464