HEC MONTRÉAL

École affiliée à l'Université de Montréal

Prediction problems using maximum entropy models

par
Lotfi Khribi

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(option Méthodes quantitatives de gestion)

Avril 2017

HEC MONTRÉAL

École affiliée à l'Université de Montréal

Cette thèse intitulée :

Prediction problems using maximum entropy models

Présentée par :

Lotfi Khribi

a été évaluée par un jury composé des personnes suivantes :

M. François Bellavance
HEC Montréal
Président-rapporteur

M. Marc Fredette
HEC Montréal
Codirecteur de recherche

Mme Brenda MacGibbon
UQAM
Codirectrice de recherche

M. Denis Larocque
HEC Montréal
Membre du jury

M. Lajmi Lakhal Chaieb
Université Laval
Examinateur externe

M. Daniel Parent
HEC Montréal
Représentant du directeur de HEC Montréal

RÉSUMÉ

Dans cette thèse, nous étudierons les problèmes liés aux problèmes de prédiction. En particulier, nous considérons la prédiction des événements récurrents. Pour cela, nous développons différents modèles de prédiction lorsque ces événements peuvent être modélisés en utilisant des processus de Poisson homogènes et non-homogènes. Parmi ces modèles, on s'intéresse à ceux qui utilisent des effets aléatoires car ils possèdent des caractéristiques intéressantes. Nous proposons des modèles prédictifs utilisant des techniques bayésiennes empiriques et le principe du maximum d'entropie afin de modéliser les effets aléatoires avec des distributions régissant la distribution inconnue des paramètres. Nous montrerons que pour la prédiction des événements récurrents notre premier modèle qui utilise comme loi a priori la loi d'entropie maximum à deux moments qui correspond à la loi normale tronquée se compare favorablement au modèle classique de la négative binomiale qui utilise comme loi a priori la loi gamma. On présente ensuite une extension de l'approche développée dans notre premier modèle. En effet, en raison de la condition sur l'utilisation de la loi a priori d'entropie maximum à deux moments, notre premier modèle est contraint à considérer seulement les cas où le coefficient de variation était inférieur ou égal à 1. Ici, nous enlevons cette restriction par l'utilisation des lois d'entropie maximum avec un nombre de moments d'ordre plus élevé et nous l'appliquons dans la prédiction des événements récurrents tout en utilisant des processus de Poisson homogènes et non-homogènes. Nous évaluons la performance de nos modèles par des études de simulation approfondies et par quelques ensembles de données.

Mots clés : Modèles de Poisson mixtes, Événements récurrents, Principe du maximum d'entropie, Processus de Poisson homogène et non–homogène, Méthode des moments, Maximum de vraisemblance.

Méthodes de recherche : Méthode mixte, modélisation mathématique, recherche quantitative.

ABSTRACT

In this thesis, we will study issues related to prediction problems. In particular, we consider the prediction of recurrent events. For this, we develop different prediction models when these events can be modeled using homogeneous or nonhomogeneous Poisson processes. Amongst these models, we are interested in those using random effects because they possess interesting features. We propose a predictive model using empirical Bayes techniques and the maximum entropy principle in order to model the random effects for the unknown parameters. We will show that for the prediction of recurrent events, our first model using as a prior the two moments maximum entropy distribution, which is equivalent to the truncated normal distribution, compared very favorably to the negative binomial model that uses as a prior the gamma distribution. We also present an extension of the approach developed in our first model: because of the two moment condition on our maximum entropy priors, we were restricted to considering only cases where the coefficient of variation was less than or equal to 1. We remove this restriction by the use of higher moment maximum entropy priors in the prediction of recurrent events using homogeneous and nonhomogeneous Poisson processes. We assess the performance of such models through extensive simulation studies and some real data sets.

Keywords: Recurrent events; The maximum entropy principle; Mixed-Poisson; Homogeneous and Nonhomogeneous Poisson process; Moment matching; Maximum likelihood.
Research methods: Mixed methods; mathematical modeling; quantitative research.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

À ma mère qui souhaitait un médecin dans la famille ;
ce sera bien un docteur mais sans ordonnance.

REMERCIEMENTS

Je tiens en premier lieu à remercier chaleureusement mes deux directeurs de thèse, Marc Fredette et Brenda MacGibbon, pour leurs encadrements, leurs patiences, leurs confiances et leurs grande disponibilités tout au long de ces années de recherche. J'ai pris un grand plaisir à travailler avec eux. Merci Marc et Brenda!

Tous mes remerciements, également, aux membres du jury pour leur temps consacré à lire et à évaluer ma thèse.

Je tiens à exprimer ma profonde gratitude à ma mère Chadlia qui m'a grandement soutenu et encouragé à réaliser mon rêve d'accomplir mes études doctorales.

Le plus grand remerciement vont à ma chère épouse Imen Nakhli pour avoir su me faire croire que tout est possible même dans les moments les plus critiques, à mon adorable fille Yasmine, à mon fils le capitaine Youssef et à ma fille Rayhane l'éponge pour leur patiences, leur soutiens et leur encouragements durant mes études doctorales.

Je remercie également mes deux GRANDE familles, les familles Khribi et Nakhli, pour vos encouragements tout au long de ce long projet.

Je suis reconnaissant envers mes amis proches Michel, Zeineb, Laila et Severien , qui m'ont été pour moi d'un soutien inestimable.

Un grand merci à mes amis Mounir, Nidhal, Mohamed-Ali et Walid pour leur supports, leur encouragements et leur bonnes paroles.

Merci aussi à mon Frère Mohamed Jabir pour ses conseils et pour son soutien informatique.

Je remercie bien évidemment les différents organismes qui m'ont appuyé financièrement tout au long du parcours doctoral : Le syndicat des chargé(e)s de cours de l'UQAM (SCCUQ) avec la bourse de perfectionnement longue durée et le HEC Montréal.

Je dédie finalement ce travail à l'âme de mon cher père Mekki et à mes chers grands parents en particulier à ma grand-mère Hallouma qui ont sacrifié leurs vies pour moi, et qui ont été mes repères.

INTRODUCTION

Dans cette thèse, nous étudierons les problèmes liés aux problèmes de prédiction. En particulier, nous considérons la prédiction des événements récurrents, événements qui se produisent à répétition au fil du temps. La prédiction des événements récurrents a été discutée dans plusieurs contextes comme par exemple dans les réclamations de garantie des véhicules (Fredette et Lawless, 2007) ou les réclamations dans le domaine des assurances (England et Verrall, 2002), où les prédictions sont utilisées, par exemple, pour la planification fiscale.

Les modèles de Poisson mixtes ont été trouvés utiles lorsque les événements récurrents présentent une surdispersion avec l'utilisation du modèle de Poisson, c'est-à-dire quand la variance est plus grande que la moyenne. Cette surdispersion est souvent attribuée à l'hétérogénéité inobservée entre les événements. Cette surdispersion peut être comptabilisée en utilisant des intensités différentes pour chaque processus, et ainsi utiliser une distribution a priori pour ces intensités inconnues. Nous avons l'intention d'utiliser des techniques bayésiennes empiriques et le principe du maximum d'entropie afin de modéliser cette distribution a priori.

Tout d'abord, la notion d'entropie apparait dans de nombreux domaines, comme par exemple en physique, où l'entropie est une grandeur thermodynamique associée à un système de particules. En théorie de l'information, où l'entropie quantifie le manque d'information, la maximisation de l'entropie se comprend comme une "diminution de l'information" ou une "augmentation du désordre". L'entropie d'une distribution de probabilité finie π pour une variable aléatoire contenue X est une fonction définie par :

$$H = -\int_x \pi(x) \ln(\pi(x)) dx.$$

Cette fonction a été introduite en premier par Shannon (Shannon (1948)). Il est possible de fournir à l'entropie de Shannon une interprétation en terme d'information moyenne. En effet, dans le cas discret, l'information au sens de Shannon apportée par une réalisation x du variable aléatoire discrète X est $-\ln(p_x)$. L'information moyenne est donc donnée par $-p_x \ln(p_x)$, qui est l'expression

de Shannon dans le cas discret. À la base de l'utilisation de cette entropie de Shannon, la réalisation d'un événement rare apporte plus d'information sur le phénomène que la réalisation d'un événement fréquent. Le principe de maximum d'entropie (PME) choisit en premier lieu les quantités statistiques que l'on juge essentielles pour résumer l'information apportée par un jeu de données. La loi de probabilité décrivant le phénomène aléatoire, n'apparaît qu'après et doit vérifier des contraintes mettant en jeu ces quantités statistiques essentielles.

La recherche d'un modèle mixte par application du PME suppose en premier lieu de modéliser l'incertitude par une loi qui maximise l'entropie de Shannon parmi les lois qui vérifient les contraintes imposées par un jeu de données. Donc le PME choisit en premier lieu les quantités statistiques que l'on juge essentielles pour résumer l'information apportée par ce jeu de données. Ce qui fait que la loi de probabilité décrivant le phénomène aléatoire, n'apparaît qu'après et doit vérifier des contraintes mettant en jeu ces quantités statistiques essentielles.

Pour cela et afin de modéliser la distribution a priori qui tient compte d'une possible hétérogénéité inobservée entre les événements en utilisant le PME, nous développons dans cette thèse différents modèles de prédiction lorsque ces événements récurrents peuvent être modélisés en utilisant des processus de Poisson homogènes et non-homogènes.

Dans le premier chapitre, nous proposons un modèle prédictif permettant de prédire les événements récurrents en utilisant des processus de Poisson homogènes lorsque la distribution régissant la distribution des paramètres est inconnue. Nous avons l'intention d'utiliser des techniques bayésiennes empiriques et le PME afin de modéliser l'information a priori. Cette approche a également été motivée par le succès de l'utilisation de la loi a priori gamma pour ce type de problème. Le choix de la loi a priori gamma dans les modèles Poisson mixtes a été motivé par ses belles propriétés mathématiques lorsqu'elle est utilisé avec les processus de Poisson. Ici, nous proposons d'appliquer la méthode des moments pour estimer les paramètres de la loi a priori d'entropie maximum, c'est-à-dire maximiser l'entropie soumise à seulement deux contraintes que les deux premiers moments soient égaux aux moments empiriques et ainsi obtenir comme solution la loi normale tronquée (tronquée au-dessous de zéro). Nous avons montré que pour la prédiction des événements récurrents notre modèle à effets aléatoires qui utilise comme loi a priori la loi normale tronquée se compare favorablement au modèle classique de la négative binomiale utilisant la loi a priori gamma pour les effets aléatoires. Le chapitre 2 présente une extension de l'approche développée dans le premier chapitre. En effet, en raison de la condition sur l'utilisation de la loi a priori d'entropie maximum à deux moments, nous avons été contraint à considérer seulement les cas où le coefficient de variation était inférieur ou égal à 1. Dans ce chapitre, nous enlevons cette restriction par l'utilisation des lois d'entropie maximum avec un nombre de moments d'ordre plus élevé et nous l'appliquons dans la

prédiction des événements récurrents tout en utilisant des processus homogènes de Poisson. Nous évaluons la performance de nos modèles par des études de simulation approfondies et par quelques ensembles de données. Dans le troisième chapitre, l'hypothèse de l'homogénéité dans le temps est relaxée et nous proposons un nouveau modèle prédictif à effets aléatoires permettant la prédiction des événements récurrents toute en utilisant des processus de Poisson non-homogènes. De plus, l'hétérogénéité possible entre les unités a été modélisée en utilisant des lois d'entropie maximum avec un nombre de moments d'ordre plus élevé au lieu de la loi a priori gamma. Nous appliquons notre modèle sur un ensemble de données réel provenant d'un programme de fidélisation et nous comparons son adéquation au modèle classique qui utilise comme loi a priori la loi gamma.

Chapter 1

The Poisson Maximum Entropy Model for Homogeneous Poisson Processes

Abstract

Our main interest is parameter estimation using maximum entropy methods in the prediction of future events for a Homogeneous Poisson process (HPP) when the distribution governing the distribution of the parameters is unknown. We intend to use empirical Bayes techniques and the maximum entropy principle to model the prior information. This approach has also been motivated by the success of the gamma prior for this problem, since it is well known that the gamma maximizes Shannon entropy under appropriately chosen constraints. However, as an alternative, we propose here to apply one of the often used methods to estimate the parameters of the maximum entropy prior. It consists of moment matching, that is, maximizing the entropy subject to the constraint that the first two moments equal the empirical ones and we obtain the truncated normal distribution (truncated below at the origin) as a solution. We also use maximum likelihood estimation (MLE) methods to estimate the parameters of the truncated normal distribution for this case. These two solutions, the gamma and the truncated normal, which maximize the entropy under different constraints are tested as to their effectiveness for prediction of future events for homogeneous Poisson processes by measuring their coverage probabilities, the suitably normalized lengths of their pre-

diction intervals and their goodness-of-fit measured by the Kullback-Leibler criterion and a discrepancy measure. The estimators obtained by these methods are compared in an extensive simulation study to each other as well as to the estimators obtained using the completely noninformative Jeffreys' prior and the usual frequency methods. We also consider the problem of choosing between the two maximum entropy methods proposed here, that is, the gamma prior and the truncated normal prior, estimated both by matching of the first two moments and, by maximum likelihood, when faced with data and we advocate the use of the sample skewness and kurtosis. The methods are also illustrated on two examples: one concerning the occurrence of mammary tumors in laboratory animals taking part in a carcinogenicity experiment and the other, a warranty data set from the automobile industry.

Keywords: Recurrent events; mixed-Poisson; Jeffreys' prior; moment matching; maximum likelihood estimation, skewness, kurtosis.

1.1 Introduction

This paper investigates the prediction of recurrent events, events which occur repeatedly over time. A considerable amount of such data is seen in a number of different subject areas, for example, in marketing (Wang et al. (2007)), in finance (Zellner and Tobias (2001) and Ximing (2003) and in engineering and reliability contexts (Fredette and Lawless (2007)). Mixed Poisson models have been found useful where recurrent events display extra-Poisson variation, that is, where the variance is usually larger than the mean. Such overdispersion is often attributed to unobserved heterogeneity between events. This overdispersion may be accounted for by using different rates for each process and then using a prior distribution on these unknown rates.

Let $\mathbf{N}(s,t)$ be the random variable representing the number of events occurring for a subject in the time interval $[s,t]$. For convenience we write $\mathbf{N}(t)$ for $\mathbf{N}(0,t)$. We will only consider continuous time processes where two events cannot occur simultaneously. Many different types of such processes are discussed in the literature (see Cook and Lawless (2007)), but the Poisson process (PP) is a popular one used by statisticians to model such recurrent events. The intensity function satisfies:

$$\lambda(t|H(t)) = \lim_{\triangle t \to 0} \frac{P[\mathbf{N}(t, t+\triangle t) = 1|H(t)]}{\triangle t},$$

where $H(t)$ denotes the history of the process up to time t.

For the Homogeneous Poisson processes (HPPs), considered here, the rates are the unknown parameters. Suppose that we have n subjects and $\mathbf{N}_i(t)$ denotes the number of events occurring for a subject i up to time t. The model is defined by:

$$\begin{aligned} \mathbf{N}_i(t)|\lambda_i &\sim PP(\lambda_i), \\ \lambda_i &\sim \pi(\lambda_i), \end{aligned} \tag{1.1}$$

where $i = 1, ..., n$, the processes are independent, and, the rates λ_i have an unknown distribution whose density is denoted by $\pi(\lambda_i)$. We also suppose here that each process is observed up to a fixed time t_{1i} and that we are interested in finding a point predictor or a prediction interval for the $\mathbf{N}_i(t_{1i}, t_{2i})|N(t_1)$. Throughout this article, $(\lambda_1, \lambda_2 ..., \lambda_n)$, $(\mathbf{N}_1(t_{11}), ..., \mathbf{N}_n(t_{1n}))$ and $(\mathbf{N}_1(t_{11}, t_{21}), ..., \mathbf{N}_n(t_{1n}, t_{2n}))$ will be denoted by $\underline{\lambda}$, $\underline{\mathbf{N}}(t_1)$ and $\underline{\mathbf{N}}(t_1, t_2)$ respectively.

The entropy of a probability distribution $\pi(\lambda)$, first introduced by Shannon (1948), is a measure of the amount of information contained in the distribution which can be written as follows:

$$H = -\int_{\lambda} \pi(\lambda) \ln(\pi(\lambda)) d\lambda. \tag{1.2}$$

The larger the entropy, the less information is provided by the distribution. It has been largely popularized by Jaynes (1957, 1968, 1982), Good (1963), Zellner (1977) and Skilling (1989) among others.

Many authors have worked on the problem of maximizing H subject to various side conditions (e.g., Zellner and Highfield (1988) and Mohammad-Djafari (1991), Jaynes (1982), Lisman and Van Zuylen (1972), Rao et al. (1973). As shown in Mohammad-Djafari (1991), the maximum entropy prior allows us to determine the optimal probability distribution $\pi(\lambda)$ from a finite set of equations involving expectations of some known functions. Usually, these known functions $\phi_k(\lambda)$ are either the powers of λ or its logarithm. Wragg and Dowson (1970) gave a procedure for fitting maximum entropy distributions subject to the constraint that the first K moments of this distribution are equal to the empirical moments.

The priors that we compare for the prediction of recurrent events for a homogeneous Poisson process here have at most two parameters. We use the maximum entropy gamma prior where the gamma prior is found by solving the equations involving expectations of 1, λ and $\log(\lambda)$ as well as the prior from the Poisson maximum entropy model (MaxEnt) where the prior is found by solving the equations involving expectations of 1, λ and λ^2. This yields a truncated normal distribution (truncated below at the origin). Henceforth it will be simply be called the truncated normal prior or the MaxEnt prior.

We focus on comparing by simulation the performance of these priors to the Jeffreys' prior (Jeffreys, 1946, 1961) and the usual frequency methods for HPP. In order to achieve a fairer comparison with the maximum entropy gamma prior where the usual parameter estimators are the maximum likelihood ones (MLE), we also propose to use maximum likelihood methods for estimating the two parameters in the truncated normal distribution prior. It will be shown that for HPP's, this differs from the usual MLE method for exponential family distributions when the constraints are the first two moments and the MLE and moment matching methods give the same estimates (Mohammad-Djafari and Idier(1991)). We also consider the gamma model with the parameters estimated by matching the first two moments and compare all four methods with the Jeffreys' prior and the frequency methods.

The remainder of this paper is organized as follows. In Section 2, we introduce our Poisson-MaxEnt model and show that for matching on the first two theoretical moments, the MaxEnt prior distribution corresponds to the truncated normal prior(truncated below at the origin). Section 3 develops the two methods of estimation for the parameters of this model: moment matching (MM) and the MLE parameter estimation method for the truncated normal. In Section 4, the performances of the proposed matching moment approach and its comparison with the use of the gamma conjugate prior using MLE for the parameters, the noninformative Jeffreys' prior, the prior obtained using moment matching for the gamma, the prior obtained using the MLE method of parameter estimation for the truncated normal and the usual frequentist methods are studied through Monte Carlo simulations. The performance is evaluated by the Kullback-Leibler divergence, a discrepancy measure and the coverage probability and the suitably normalized length of the prediction intervals. This simulation study indicates that the MaxEnt prior with a HPP is preferable to the noninformative Jeffreys' prior for the prediction of recurrent events studied here. Moreover, we also show that the MaxEnt model is an interesting alternative in some cases to the classical negative binomial

(NB) model obtained with the conjugate gamma prior. We also indicate the conditions under which the maximum entropy matching moment and the MLE method for the truncated normal entropy methods perform better than the classical negative binomial model and vice versa. We also show how the empirical skewness and kurtosis estimates can be used to choose between the gamma and the truncated normal priors when faced with real data. These methods are also illustrated on two real examples, one of them concerning the occurrence of mammary tumors in laboratory animals taking part in a carcinogenicity experiment and the other one, a warranty data set from the automobile industry. A general discussion and possible extensions of these methods with concluding remarks are presented in Section 5.

1.2 A Poisson Maximum Entropy model

1.2.1 Maximum Entropy Prior

The goal is to find the density function $\pi(\lambda)$ that maximizes the entropy H given by (1.2), subject to

$$\int_{\mathbf{R}^+} \phi_k(\lambda)\pi(\lambda)d\lambda \quad = \quad \hat{\mu}_k \quad k = 0, 1, 2. \tag{1.3}$$

where $\phi_0(\lambda) = 1$, $\phi_1(\lambda) = \lambda$ and $\phi_2(\lambda) = \lambda^2$ and $\hat{\boldsymbol{\mu}}_{\mathbf{0}} = 1$ and $\hat{\boldsymbol{\mu}}_{\mathbf{1}}$ and $\hat{\mu}_2$ the two non-central empirical moments of the distribution respectively.

Following the same procedure for fitting maximum entropy distributions used by Zellner and Highfield (1988) to find the function $\pi(\lambda)$ that maximizes the entropy by solving the nonlinear problem (1.3), we apply the Lagrange multiplication method (Weinstock, 1952) which yields the following maximum entropy distribution:

$$\pi(\lambda|\alpha_0, \alpha_1, \alpha_2) = \exp\left(-\alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2\right),$$

with normalization constant defined by:

$$\begin{aligned} e^{\alpha_0} &= \int_{\mathbb{R}^+} \exp\left(-\alpha_1\lambda - \alpha_2\lambda^2\right)d\lambda \\ &= \frac{\sqrt{\pi}}{\sqrt{\alpha_2}}\Phi\left(-\frac{\alpha_1}{\sqrt{2\alpha_2}}\right)e^{\frac{\alpha_1^2}{4\alpha_2}}, \end{aligned}$$

where $\Phi(.)$ is the distribution function of the standard normal distribution.

This maximum entropy density over $[0, \infty)$ is not always of the above form. Indeed, Wragg and Dowson (1970) have shown that when fitting the first two moments for distributions over $[0, \infty)$ this maximum entropy approach is not appropriate if the centred moments satisfy $(\boldsymbol{\mu_2'}) > 2(\boldsymbol{\mu_1})^2$ or equivalently the coefficient of variation > 1, because all densities of the form $\pi(\lambda) = exp(-\alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2)$ with $\alpha_2 > 0$ have the property that $\boldsymbol{\mu_2'} \leq 2\boldsymbol{\mu_1}^2$.

However, any law whose density is proportional to $\exp(-\alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2)$ for $\lambda \geq 0$ takes the form of a left truncated normal law with mean μ and variance σ^2 and with $\alpha_1 = -\frac{\mu}{\sigma^2}$ and $\alpha_2 = \frac{1}{2\sigma^2}$.

Two main reasons have lead us to consider situations where the first two moments are known and thus to use the MaxEnt prior with two parameters: first, this is because the NB model which will serve as one of the models for comparison also uses the maximum entropy gamma prior with two parameters; secondly, for the simulation study in Section 4, the different distributions used to generate the unknown parameters $\underline{\lambda}$ are also distributions with at most two parameters.

1.2.2 Model specifications

If we let the λ_i's be truncated normal random variables, then model (1) becomes an empirical Bayes model given by:

$$\begin{aligned} \mathbf{N}_i(t)|\lambda_i &\sim PP(\lambda_i), \\ \pi(\lambda_i; \mu, \sigma^2) &= \frac{e^{\frac{-(\lambda_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}(1 - \Phi(\frac{-\mu}{\sigma}))}. \end{aligned} \tag{1.4}$$

For the Poisson-Maximum Entropy model (1.4), the joint posterior distribution of all the unknown parameters $\underline{\lambda}|\underline{N}(t)$ is given by

$$\begin{aligned} \pi(\underline{\lambda}|\underline{N}(t_1); \mu, \sigma^2) &= \frac{P\big[\underline{\mathbf{N}}(t_1) = \underline{N}(t_1)|\underline{\lambda}\big]\pi(\underline{\lambda}; \mu, \sigma^2)}{\int_\lambda P\big[\underline{\mathbf{N}}(t_1) = \underline{N}(t_1)|\underline{\lambda}\big]\pi(\underline{\lambda}; \mu, \sigma^2)d\lambda} \\ &= \prod_{i=1}^{n} \frac{\lambda_i^{N_i(t_{1i})} e^{-\frac{(\lambda_i-(\mu-\sigma^2 t_{1i}))^2}{2\sigma^2}}}{\int_{\lambda_i} \lambda_i^{N_i(t_{1i})} e^{-\frac{(\lambda_i-(\mu-\sigma^2 t_{1i}))^2}{2\sigma^2}} d\lambda_i}. \end{aligned} \tag{1.5}$$

Given $\pi(\underline{\lambda}|\underline{N}(t_1);\mu,\sigma^2)$, the parameters of interest $\underline{\lambda}|\underline{N}(t_1)$ can be estimated from this posterior distribution. Unfortunately, direct mathematical derivation of $\pi(\underline{\lambda}|\underline{N}(t_1);\mu,\sigma^2)$ usually involves a high-dimensional integration to obtain the normalizing constant which is a product of $\int_{\lambda_i} \lambda_i^{N_i(t_{1i})} e^{-\frac{(\lambda_i-(\mu-\sigma^2 t_{1i}))^2}{2\sigma^2}} d\lambda_i$ and although recursion formulas are avaiblable, this derivation is not mathematically tractable. In this case, the posterior distributions will not have a known closed form, but rather complicated high dimensional densities, which makes direct inference almost impossible. The problem of estimating the mean and variance of the posterior distribution can be solved by generating a large number of samples from the posterior distribution using Markov chain Monte Carlo (MCMC) implemented in WinBUGS (Spiegelhalter et al.(2003)), and from these samples, we can obtain appropriate parameter estimates of $(\underline{\lambda}|\underline{N}(t_1);\mu,\sigma^2)$.

1.3 Estimating unknown Poisson Maximum Entropy parameters

The most popular prior distribution used with the Poisson is the gamma distribution which is the maximum entropy prior and also the conjugate one and also maximizes Shannon entropy. The parameters are usually estimated by the maximum likelihood method. The usual maximum entropy estimation method, denoted by(MaxEnt) here, uses matching moments (MM). Thus, as mentioned in Section 2, we also introduce the use of maximum likelihood methods for estimating the parameters of an HPP. This is called Pseudo-MaxEnt here. Thus we have two different empirical Bayes methods for estimating the parameters of the Poisson-Maximum Entropy model to compare with the maximum entropy gamma model where its parameters are estimated both by MLE and MM methods.

The objective of both estimation approaches in the Poisson-Maximum Entropy case is to choose the probability distribution $\pi(\lambda_i;\mu,\sigma^2) = \frac{e^{\frac{-(\lambda_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}(1-\Phi(\frac{-\mu}{\sigma}))}$ for the unknown parameter λ_i which best represents the observed data $\underline{N}(t)$.

1.3.1 Maximum Entropy Method

Henceforth the maximum entropy method which uses the matching moments (MM) estimation method for the parameters in the empirical Bayes Poisson-Maximum Entropy model (1.4) will be called the MaxEnt method. We can show that the expectation and variance of a left truncated normal distribution with truncation below at 0 for an unknown parameter λ_i are defined by $\mathbb{E}[\lambda_i \mid \lambda_i \geq 0] = \mu + g(\mu,\sigma)\sigma$ and $\mathbb{V}ar[\lambda_i \mid \lambda_i \geq 0] = \sigma^2\Big[1 - g(\mu,\sigma)\Big(g(\mu,\sigma) + \frac{\mu}{\sigma}\Big)\Big]$, where

$g(\mu,\sigma) = \frac{\phi(-\frac{\mu}{\sigma})}{(1-\Phi(\frac{-\mu}{\sigma}))}$ and $\phi(.)$ is the density function of the standard normal distribution.

Letting $\mathbf{R}_i = \frac{\mathbf{N}_i(t_{1i})}{t_{1i}}$, the first two central moments of λ_i are:

$$\mathbb{E}[\mathbf{R}_i] = \mathbb{E}[\lambda_i] = \mu + g(\mu,\sigma)\sigma$$

and

$$\mathbb{V}ar[\mathbf{R}_i] = \mathbb{V}ar[\lambda_i] + \mathbb{E}[\lambda_i](t_{1i})^{-1}.$$

Note that because the variance of $\mathbf{R}_i$ is different for each observation, $(t_{1i})^{-1}$ will be replaced with its average value when considering the vector $\underline{\mathbf{N}}(t_1)$. This approach has been used before to deal with the same kind of data (e.g. Gaver and O'Muircheartaigh, 1987). Our own simulations suggest that this approach provides accurate estimates even when the t_{1i}'s vary greatly from one process to another.

Thus we have to solve the following non-linear system of equations:

$$\begin{aligned} \mu + g(\mu,\sigma)\sigma &= \overline{R}, \\ \sigma^2\Big[1 - g(\mu,\sigma)\Big(g(\mu,\sigma) + \frac{\mu}{\sigma}\Big)\Big] &= s_R^2 - \overline{R}\overline{t_{1i}^{-1}}, \end{aligned} \tag{1.6}$$

where $\overline{t_{1i}^{-1}}$ is the sample average of $\frac{1}{(t_{1i})}$'s, while $\overline{R}$ and s_R^2 are respectively the sample mean and variance of the R_i's. The lack of a closed-form solution means that the MaxEnt estimators must be found using an iterative approach. For this we use the "nleqlsv" R package.

It is important to mention here that we cannot apply directly the methods presented in Mohammad-Djafari and Idier(1991) because our situation is more complex. Indeed, our method must take into account that the process $\mathbf{N}_i(t)$ depends on time t_{1i} since these counts are generated from a Poisson distribution with rates $\lambda_i t_{1i}$ and therefore the first two central moments of λ_i are given by their empirical estimates $\overline{R}$ and $s_R^2 - \overline{R}\overline{t_{1i}^{-1}}$ which differ from those given by Mohammad-Djafari and Idier (1991) for other problems.

1.3.2 The Pseudo-Maximum Entropy Method

We introduce the pseudo-maximum entropy (Pseudo-MaxEnt) method using MLE instead of MM for the empirical Bayes MaxEnt model (1.4). Thus to obtain the Pseudo-MaxEnt estimators of the parameters we construct the marginal likelihood L of the empirical Bayesian Poisson-Maximum Entropy model (1.4)

$$\begin{aligned} L(\mu,\sigma^2|\underline{N}(t_1)) &= \int_{\lambda} P[\underline{\mathbf{N}}(t_1)=\underline{N}(t_1)|\lambda]\pi(\underline{\lambda};\mu,\sigma^2)d\underline{\lambda} \\ &= \prod_{i=1}^{n} \frac{t_{1i}^{N_i(t_{1i})}}{N_i(t_{1i})!}\frac{1}{(1-\Phi(\frac{-\mu}{\sigma}))\sqrt{2\pi\sigma^2}} I_i(N_i) \end{aligned}$$

with

$$I_i(N_i) = \int_0^{\infty} \lambda_i^{N_i(t_{1i})} e^{\frac{-(\lambda_i-\mu)^2}{2\sigma^2} - t_{1i}\lambda_i} d\lambda_i. \tag{1.7}$$

The log-likelihood is

$$l(\mu,\sigma^2|\underline{N}(t_1)) \propto \sum_{i=1}^{n} \Big[-0.5ln(\sigma^2) - ln\Big(1-\Phi(\frac{-\mu}{\sigma})\Big) + ln\Big(I_i(N_i)\Big)\Big]. \tag{1.8}$$

Using Lebesgue's Dominated Convergence Theorem, Talvila (2001) gave necessary and sufficient conditions to interchange the order of differentiation and integration which are verified for $I(N_i)$ here. Interchanging differentiation and integration, the first derivatives of (1.7) with respect to μ and σ^2 are given by the following recurrent equations

$$\begin{aligned} \frac{\partial I_i(N_i)}{\partial \mu} &= \frac{1}{\sigma^2}\Big[I_i(N_i+1) - \mu I_i(N_i)\Big]. \\ \frac{\partial I_i(N_i)}{\partial \sigma^2} &= \frac{1}{2\,\sigma^4}\left[I_i(N_i+2) - 2\,\mu\, I_i(N_i+1) + \mu^2\, I_i(N_i)\right]. \end{aligned}$$

Using the derivatives above, the first partial derivatives of (1.8) are

$$\begin{aligned} \frac{\partial l(\mu,\sigma^2|\underline{N}(t_1))}{\partial \mu} &= -\frac{n}{\sigma}\frac{\phi(-\frac{\mu}{\sigma})}{(1-\Phi(\frac{-\mu}{\sigma}))} + \sum_{i=1}^{n}\Big[\frac{\frac{\partial I_i(N_i)}{\partial \mu}}{I_i(N_i)}\Big], \\ \frac{\partial l(\mu,\sigma^2|\underline{N}(t_1))}{\partial \sigma^2} &= -\frac{n}{2\,\sigma^2} + \frac{n\,\mu\,\phi(-\mu/\sigma)}{2\,\sigma^3\,(1-\Phi(-\mu/\sigma))} + \sum_{i=1}^{n}\frac{\partial I_i(N_i)}{\partial \sigma^2}\Big/ I_i(N_i). \end{aligned} \tag{1.9}$$

Using the second derivatives of (1.9), the Hessian matrix of (1.8) is

$$H = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix},$$

where

$$\begin{aligned}
d_{11} &= \frac{n}{\sigma^2}\left(\phi'(-\mu/\sigma)\,(1-\Phi(-\mu/\sigma)) + \phi^2(-\mu/\sigma)\right) \Big/ (1-\Phi(-\mu/\sigma))^2 + \\
&\qquad + \sum_{i=1}^{n}\left(\frac{\partial^2 I_i(N_i)}{\partial\mu^2} \Big/ I_i(N_i) - \left(\frac{\partial I_i(N_i)}{\partial\mu}\right)^2 \Big/ (I_i(N_i))^2\right), \\
d_{12} &= \frac{n\,\mu}{2\,\sigma^4}\left(\frac{(\sigma/\mu)\,\phi(-\mu/\sigma) - \phi'(-\mu/\sigma)}{1-\Phi(-\mu/\sigma)} - \frac{\phi^2(-\mu/\sigma)}{(1-\Phi(-\mu/\sigma))^2}\right) + \\
&\qquad + \sum_{i=1}^{n}\left(\frac{\partial^2 I_i(N_i)}{\partial\mu\,\partial\sigma^2} \Big/ I_i(N_i) - \left(\frac{\partial I_i(N_i)}{\partial\mu}\,\frac{\partial I_i(N_i)}{\partial\sigma^2}\right) \Big/ (I_i(N_i))^2\right), \\
d_{22} &= \frac{n}{2\,\sigma^4} + \frac{n\,\mu^2}{4\,\sigma^6}\left(\frac{\phi'(-\mu/\sigma) - (3\,\sigma/\mu)\,\phi(-\mu/\sigma)}{1-\Phi(-\mu/\sigma)} + \frac{\phi^2(-\mu/\sigma)}{(1-\Phi(-\mu/\sigma))^2}\right) + \\
&\qquad + \sum_{i=1}^{n}\left(\frac{\partial^2 I_i(N_i)}{\partial\sigma^4} \Big/ I_i(N_i) - \left(\frac{\partial I_i(N_i)}{\partial\sigma^2}\right)^2 \Big/ (I_i(N_i))^2\right).
\end{aligned}$$

The analytic solutions to equations (1.9) are extremely difficult to obtain; thus it is natural to use the Newton-Raphson algorithm or any of its variants to estimate the parameters μ and σ^2 . We have chosen to use the Quasi-Newton algorithm (also known as a variable metric algorithm), published in Broyden et al. (1970) for the solution.

In order to know how the maximum likelihood estimation method performs in practice for the Pseudo-MaxEnt model, we conducted an additional simulation study to measure the performance of this method. For this, we used different values of μ and σ^2 and calculated the bias, the mean squared error(MSE), the mean, the standard deviation, the minimum and the maximum for both estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$. We also did an analogous study using the score function for the gamma to estimate μ and σ^2 using the same efficiency scores and found that the results were very comparable.

As the existence and the uniqueness of the maximum likelihood solution for the estimator of the truncated normal mean and variance is an important question we designed another simulation study to compare the sign changes of the truncated normal and the gamma score functions. Fixing μ and varying σ^2 we studied the behavior of the score function based on the first partial derivative

of the likelihood function with respect to μ from equation (1.9) and in the gamma score function. We were interested in the sign changes in the score function and classified them into 3 categories: more than one sign change, one sign change (from positive to negative) and no sign changes. We performed an analogous simulation study by fixing σ^2 and varying μ. Although the gamma score function never had more than one solution, the percentage of multiple crossings for the truncated normal score function was very low (usually less than 3%) except for small μ and small σ^2 where the results were somewhat more problematic. The percentage of times when the gamma and truncated normal score functions were never zero were very comparable, although the gamma percentages were usually slightly lower. The results of these studies were encouraging for the use of the method of solving equations(1.9) here.

1.4 Simulation study and Data applications

1.4.1 Simulation Study

The performance of the maximum entropy methods with the MM and MLE methods for prediction in homogeneous Poisson processes was evaluated in the simulation study described below, based on the following criteria: the Kullback-Leibler criterion, a discrepancy measure, point prediction and finally coverage probability and length of the prediction interval.

The effects of both the model and the estimation method were studied. Different pairs of parameter constraints were arbitrarily chosen to reflect different orders of heterogeneity and magnitude. There are six pairs of constraints of interest in our simulation study for the unknown parameters $\underline{\lambda}$ (see Table 1.1).

Pairs	$\mathbf{E}[\lambda_i]$	$\mathbf{Var}[\lambda_i]$	Coefficient of Variation (CV)	Estimated Truncation
1	1	0.1	0.32	0.08%
2	1	0.3	0.56	3.80%
3	1	0.8	0.89	13.00%
4	5	2.5	0.32	0.08%
5	5	7.5	0.56	3.80%
6	5	20	0.89	13.00%

Table 1.1 Different pairs of parameters using in the simulation study.

In this section, we present the results for the pairs 2 and 4 of the constraints presented in Table 1.1. These two pairs are chosen in order to reflect different orders of heterogeneity and magnitude.

The results for the other four pairs are presented in the Supplemental Material.

Our aim here is to compare the effectiveness of the gamma and the truncated normal priors, as well as the Jeffreys prior and the usual frequentist methods for prediction in homogeneous Poisson processes in different situations. Therefore, in our simulation study, we have chosen several different priors as well as the gamma and the truncated normal to represent the unknown prior distribution for the Poisson process: the binomial, the uniform, the Weibull, the lognormal and the Inverse Gaussian. The idea behind the use of different distributions is that we want to study which model presented in this paper is more robust to the real distribution of the random effects.

As mentioned in the previous section, the posterior distribution is a complex function and difficult to compute. Our choice, therefore is to perform a simulation study to compute its parameters such as the mean, the variance etc. The study can be described as follows. For each set of pairs in Table 1.1, we generate $b = 2000$ samples from $n = 20$ HPP's and assumed that these processes are observed up to the times $t_1 = \{5, 5.5, ..., 9.5, 10, ..., 14.0, 14.5\}$ and that we want to predict the number of occurrences for each process up to the times $t_2 = \{12.5, ..., 12.5, 17.5, ..., 17.5\}$ respectively. The idea behind this choice is to represent different values of $(t_{2i} - t_{1i})$. More precisely, in this simulation study we start by generating a vector of the unknown parameter $\underline{\lambda}$ from the prior $\pi_0(\lambda)$ and then for each sample j we generate the counts $\underline{N}_j^*(t_1)$ and $\underline{N}_j^*(t_1, t_2)$ having a Poisson distribution with a vector of rates $\underline{\lambda} t_1$ and $\underline{\lambda}(t_2 - t_1)$ respectively. Note that we use both continuous and discrete distributions for π_0 such as the gamma, truncated normal(truncated below at the origin), Weibull, lognormal, inverse Gaussian and binomial to generate the unknown parameters $\underline{\lambda}$.

The MaxEnt and the Pseudo-MaxEnt methods of estimation are compared to four others: two non-Bayesian plug-in methods assuming Poisson processes with identical and different fixed rates and both of them estimated by the MLE method, the method based on the noninformative Jeffreys' prior and the method based on the maximum entropy gamma prior where the parameters are estimated using both the MLE and MM methods. Table 1.2 presents these methods and gives the point predictor $\widehat{N}_i(t_{1i}, t_{2i})$ and the predictive density $\widetilde{f}_p(N_i(t_{1i}, t_{2i}|\underline{N}(t_{1i})))$ for each estimation method used and with the data described by the model generated as indicated in the previous paragraph. Note that the posterior distribution of $\underline{\lambda}$ for the maximum entropy methods was obtained using an MCMC of $100,000$ realizations after a burn-in of 1000 iterations.

Model	Notation	$\widehat{N}_i(t_{1i}, t_{2i})$	$\tilde{f}_p(\mathrm{Ni}(t_{1i}, t_{2i})\|\underline{N}(t_1))$
$\mathrm{N_i(t)} \sim \mathcal{PP}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda})$	$(t_{2i} - t_{1i})\left(\frac{\sum_{i=1}^{n} \mathrm{Ni}(t_{1i})}{\sum_{i=1}^{n} t_{1i}}\right)$	$\frac{e^{(t_{2i}-t_{1i})\left(\frac{\sum_{i=1}^{n} \mathrm{Ni}(t_{1i})}{\sum_{i=1}^{n} t_{1i}}\right)}}{\mathrm{Ni}(t_{1i}, t_{2i})!}\ x$ $\left[(t_{2i} - t_{1i})\left(\frac{\sum_{i=1}^{n} \mathrm{Ni}(t_{1i})}{\sum_{i=1}^{n} t_{1i}}\right)\right]^{\mathrm{Ni}(t_{1i}, t_{2i})}$
$\mathrm{N_i(t)} \sim \mathcal{PP}(\hat{\lambda}_i)$	$\mathcal{P}(\hat{\lambda}_i)$	$(t_{2i} - t_{1i})\left(\frac{\mathrm{Ni}(t_{1i})}{t_{1i}}\right)$	$\frac{e^{((t_{2i}-t_{1i})\left(\frac{\mathrm{Ni}(t_{1i})}{t_{1i}}\right)}}{\mathrm{Ni}(t_{1i}, t_{2i})!}\ x$ $\left[(t_{2i} - t_{1i})\left(\frac{\mathrm{Ni}(t_{1i})}{t_{1i}}\right)\right]^{\mathrm{Ni}(t_{1i}, t_{2i})}$
$\mathrm{N_i(t)} \sim \mathcal{PP}(\lambda_i)$ $\pi_0(\lambda_i) = 1/\sqrt{\lambda_i}$	*Jeffreys*	$(t_{2i} - t_{1i})\left(\frac{0.5 + \mathrm{Ni}(t_{1i})}{0 + t_{1i}}\right)$	$\frac{\Gamma(0.5 + \mathrm{Ni}(t_{1i}) + \mathrm{Ni}(t_{1i}, t_{2i}))}{\Gamma(0.5 + \mathrm{Ni}(t_{1i}))\ \mathrm{Ni}(t_{1i}, t_{2i})!}\ x$ $\left(\frac{t_{2i} - t_{1i}}{t_{1i}}\right)^{\mathrm{Ni}(t_{1i}, t_{2i})} \left(\frac{t_{1i}}{t_{2i}}\right)^{0.5+\mathrm{Ni}(t_{1i})}$
$\mathrm{N_i(t)}\,\lvert\,\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim$ $\mathrm{Gamma}(\hat{a}_{mle}, \hat{b}_{mle})$	$\mathcal{G}$ $(\hat{a}_{mle}, \hat{b}_{mle})$	$(t_{2i} - t_{1i})\left(\frac{\hat{a}_{mle} + \mathrm{Ni}(t_{1i})}{\hat{b}_{mle} + t_{1i}}\right)$	$\frac{\Gamma(\hat{a}_{mle} + \mathrm{Ni}(t_{1i}) + \mathrm{Ni}(t_{1i}, t_{2i}))}{\Gamma(\hat{a}_{mle} + \mathrm{Ni}(t_{1i}))\ \mathrm{Ni}(t_{1i}, t_{2i})!} x$ $\left(\frac{t_{2i} - t_{1i}}{\hat{b}_{mle} + t_{1i}}\right)^{\mathrm{Ni}(t_{1i}, t_{2i})} \left(\frac{\hat{b}_{mle} + t_{1i}}{\hat{b}_{mle} + t_{2i}}\right)^{\hat{a}_{mle}+\mathrm{Ni}(t_{1i})}$
$\mathrm{N_i(t)}\,\lvert\,\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim$ $\mathrm{Gamma}(\hat{a}_{mm}, \hat{b}_{mm})$	$\mathcal{G}$ $(\hat{a}_{mm}, \hat{b}_{mm})$	$(t_{2i} - t_{1i})\left(\frac{\hat{a}_{mm} + \mathrm{Ni}(t_{1i})}{\hat{b}_{mm} + t_{1i}}\right)$	$\frac{\Gamma(\hat{a}_{mm} + \mathrm{Ni}(t_{1i}) + \mathrm{Ni}(t_{1i}, t_{2i}))}{\Gamma(\hat{a}_{mm} + \mathrm{Ni}(t_{1i}))\ \mathrm{Ni}(t_{1i}, t_{2i})!} x$ $\left(\frac{t_{2i} - t_{1i}}{\hat{b}_{mm} + t_{1i}}\right)^{\mathrm{Ni}(t_{1i}, t_{2i})} \left(\frac{\hat{b}_{mm} + t_{1i}}{\hat{b}_{mm} + t_{2i}}\right)^{\hat{a}_{mm}+\mathrm{Ni}(t_{1i})}$
$\mathrm{N_i(t)}\,\lvert\,\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim TruncNormal$ $(\hat{\mu}_{mle}, \hat{\sigma}^2_{mle})$	*Pseudo* $-$ *MaxEnt* $(\hat{\mu}_{mle}, \hat{\sigma}^2_{mle})$	$(t_{2i} - t_{1i})\ \lambda_i^*$	$\frac{e^{(t_{2i}-t_{1i})\lambda_i^*}}{\mathrm{Ni}(t_{1i}, t_{2i})!}[(t_{2i} - t_{1i})\ \lambda_i^*]^{\mathrm{Ni}(t_{1i}, t_{2i})}$ where λ_i^* is the posterior mean simulated from $\pi(\lambda\|N(t_1); \hat{\mu}_{mle}, \hat{\sigma}^2_{mle})$
$\mathrm{N_i(t)}\,\lvert\,\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim TruncNormal$ $(\hat{\mu}_{mm}, \hat{\sigma}^2_{mm})$	*MaxEnt* $(\hat{\mu}_{mm}, \hat{\sigma}^2_{mm})$	$(t_{2i} - t_{1i})\ \lambda_i^*$	$\frac{e^{(t_{2i}-t_{1i})\lambda_i^*}}{\mathrm{Ni}(t_{1i}, t_{2i})!}[(t_{2i} - t_{1i})\ \lambda_i^*]^{\mathrm{Ni}(t_{1i}, t_{2i})}$ where λ_i^* is the posterior mean simulated from $\pi(\lambda\|N(t_1); \hat{\mu}_{mm}, \hat{\sigma}^2_{mm})$

Table 1.2 Point predictor and predictive density for each method.

Kullback-Leibler criterion

In attempting to judge the goodness-of-fit of a given predictive approach, we require some overall measure of the divergence between each derived parameter-free predictive density $\widetilde{f}_p(\underline{N}(t_1, t_2)|\underline{N}(t_1)))$ obtained from each of the methods mentioned in Table 1.2 and the true one denoted by $f_{\pi_0}(\underline{N}(t_1, t_2)|\underline{N}(t_1); \underline{\lambda})$. An appropriate measure for this is the average Kullback-Leibler (1951) divergence, DIV_{KL}, between these two predictive densities defined by

$$DIV_{KL}\left(f_{\pi_0}; \widetilde{f}_p\right) = \mathbb{E}\Big[\int \log\Big\{\frac{f_{\pi_0}(\underline{N}(t_1, t_2)|\underline{N}(t_1); \underline{\lambda})}{\widetilde{f}_p(\underline{N}(t_1, t_2)|\underline{N}(t_1))}\Big\}\pi_0(\underline{\lambda})d\underline{\lambda}\Big].$$

where the expectation is taken over all possible values of the $(\underline{N}(t_1), \underline{N}(t_1, t_2))$ and π_0 is the true distribution of $\underline{\lambda}$ used in the simulation.

If we have two contenders, say $\widetilde{f}_{p_1}(\underline{N}(t_1, t_2)|\underline{N}(t_1))$ and $\widetilde{f}_{p_2}(\underline{N}(t_1, t_2)|\underline{N}(t_1))$, for the role of estimate then we define the average Kullback-Leibler difference in distance, (D_{KL}):

$$\begin{aligned} D_{KL}\Big(\widetilde{f}_{p_1}; \widetilde{f}_{p_2}\Big) &= \big\{DIV_{KL}(f_{\pi_0}; \widetilde{f}_{p_1}) - DIV_{KL}(f_{\pi_0}; \widetilde{f}_{p_2})\big\} \\ &= \mathbb{E}\Big[\int \log\Big\{\frac{\widetilde{f}_{p_2}(\underline{N}(t_1, t_2)|\underline{N}(t_1))}{\widetilde{f}_{p_1}(\underline{N}(t_1, t_2)|\underline{N}(t_1))}\Big\}\pi_0(\underline{\lambda})d\underline{\lambda}\Big]. \end{aligned}$$

If this difference is negative, then $\widetilde{f}_{p_1}$ is said to be a better estimate than $\widetilde{f}_{p_2}$. We use the simulation study just described to estimate D_{KL}. In fact, the average KL difference in divergence will be estimated by simulating $b = 2000$ samples of $n = 20$ HPPs using

$$\hat{D}_{KL}\Big(\tilde{f}_{p_2}, \tilde{f}_{p_1}\Big) = \sum_{j=1}^{b} \log\Big\{\frac{\tilde{f}_{p_2}(N_j^*(t_1, t_2)|N_j^*(t_1))}{\tilde{f}_{p_1}(N_j^*(t_1, t_2)|N_j^*(t_1))}\Big\},$$

where $N_j^*(t_1)$ and $N_j^*(t_1, t_2)$ are the counts generated with the Poisson distribution having rates $\lambda_{j,i}t_{1i}$ and $\lambda_{j,i}(t_{2i} - t_{1i})$ respectively for the jth sample, $j = 1, ..., b$ and for the ith process, $i = 1, ..., n$ and each parameter vector $\underline{\lambda}$ is generated from the prior π_0.

We note that in Table 1.3 the MaxEnt method is used as the reference in calculating the difference in KL divergence, which will be called the average KL distance, and thus plays the role of $\widetilde{f}_{p_1}$. Thus a negative value of the average KL distance $\hat{D}_{KL}$ means that the approach considered perform better than the MaxEnt method.

The results given in Table 1.3 indicate that when the amount of truncation and the coefficient of variation are not very large, the MaxEnt and the Pseudo-MaxEnt methods of estimation perform better than the method based on a noninformative Jeffreys' prior, whatever the distribution used to generate the unknown parameters $\underline{\lambda}$. Also, these methods have

slightly better results than the method based on the maximum entropy gamma prior when the unknown parameters are generated with truncated normal, Weibull, uniform or binomial priors. These maximum entropy methods are slightly less efficient than the method based on the maximum entropy gamma prior for the cases where the unknown parameters $\underline{\lambda}$ are generated using the gamma, lognormal, or inverse Gaussian distributions. We see from the simulation results that, although the maximum entropy methods do not always provide the predictive density closest to the true one, it is always very close. This is not the case for the method based on the Jeffreys' prior. From this simulation study, we also note that both the non-Bayesian plug-in methods assuming Poisson processes with identical and different fixed rates are usually the furthest from the true predictive density. The MaxEnt method is the closest one if all other distances are positive.

Moments for λ_i	Random effects	$P(\hat{\lambda})$	$P(\hat{\lambda}_i)$	Jeffreys	Gamma $(\hat{a}_{mm},\hat{b}_{mm})$	Gamma $(\hat{a}_{mle},\hat{b}_{mle})$	$Pseudo-MaxEnt$ $(\hat{\mu}_{mle},\hat{\sigma}^2_{mle})$
$E[\lambda_i]=1$ $Var[\lambda_i]=0.3$	Truncated Normal	13.616	2.773	0.669	0.214	0.130	-0.010
	Uniform	14.977	2.607	0.624	0.374	0.231	-0.003
	Weibull	12.351	2.942	0.767	0.018	0.017	0.001
	Gamma	11.360	3.007	0.792	-0.101	-0.108	0.062
	LNormal	10.375	3.195	0.922	-0.253	-0.254	0.091
	InvGauss	10.170	3.264	0.859	-0.247	-0.248	0.099
$E[\lambda_i]=5$ $Var[\lambda_i]=2.5$	Truncated Normal	22.709	2.258	0.650	0.113	0.093	0.016
	Binomial	22.249	2.275	0.563	0.107	0.083	0.032
	Uniform	22.142	2.356	0.540	0.007	0.014	0.013
	Weibull	22.689	2.395	0.605	0.073	0.075	0.017
	Gamma	20.976	2.304	0.582	-0.064	-0.044	0.044
	LNormal	19.786	2.359	0.597	-0.119	-0.107	0.056
	InvGauss	20.248	2.200	0.486	-0.136	-0.115	0.066

Table 1.3 Comparison of the average KL distance with the MaxEnt method (with MM) as reference method. This method is the closest one if all distances are positive.

Point prediction

To compare the adequacy of each point prediction method for $\widehat{\underline{N}}(t_1, t_2)$ obtained from one of the seven methods used in this simulation study, we analyzed the following discrepancy

$$D = \sqrt{\frac{\sum_{i=1}^{n} \left(N_i(t_{1i}, t_{2i}) - \hat{N}_i(t_{1i}, t_{2i}) \right)^2}{n}}, \tag{1.10}$$

where $\hat{N}_i(t_{1i}, t_{2i})$ is the point predictor provided by the method chosen (*cf.* Table 1.2). The value of D represents, for a given sample of n processes, the root mean square error between the real value of $N_i(t_{1i}, t_{2i})$ and its prediction.

The results from our simulation study are presented in Table 1.4. Each cell contains the average value obtained from (1.10) over the 2000 samples. The last column "True" contains results assuming the full knowledge of λ_i i.e $\hat{N}_i(t_{1i}, t_{2i}) = (t_{2i} - t_{1i})\mathbb{E}[\lambda_i | N_i(t_{1i})]$. We give the true value as a means of showing that the error of the these point predictions is mainly due to the fact that the random variables are predicted and not because the true parameters are unknown. It should be noted that the "True" value is calculated assuming we know the original model and its parameters. Thus it represents an absolute minimum value for the average discrepancy. We note that the smallest average discrepancy excluding the "True" column in this table for a given distribution of $\underline{\lambda}$ is written in bold font.

The first thing we notice in Table 1.4 is that both the MaxEnt and the pseudo-MaxEnt priors and the methods based on the maximum entropy gamma prior outperform the method based on the noninformative Jeffreys' prior and the two plug-in methods assuming HPPs with either identical or different rates. We expected this result for the two plug-in methods since all others methods treat $\underline{\lambda}$ as a random vector. Neither the MaxEnt nor the Pseudo-MaxEnt methods nor the method based on the maximum entropy gamma prior performs equally well for all the different distributions of the random effects. The MaxEnt or the Pseudo-MaxEnt methods give slightly better results than the maximum entropy gamma method when the unknown parameters $\underline{\lambda}$ are generated with the truncated normal, Weibull, uniform and bi-

Moments for λ_i	Random effects	$\mathbb{P}(\hat{\lambda})$	$\mathbb{P}(\hat{\lambda}_i)$	Jeffreys	Gamma $(\hat{a}_{mle}, \hat{b}_{mle})$	Gamma $(\hat{a}_{mm}, \hat{b}_{mm})$	*MaxEnt* $(\hat{\mu}_{mm}, \hat{\sigma}^2_{mm})$	*PseudoMaxEnt* $(\hat{\mu}_{mle}, \hat{\sigma}^2_{mle})$	True
$E[\lambda_i]=1$ $Var[\lambda_i]=0.3$	Truncated Normal	3.744	2.943	2.962	2.785	2.785	**2.771**	2.773	2.750
	Uniform	3.790	2.913	2.933	2.775	2.774	**2.741**	2.742	2.728
	Weibull	3.726	2.953	2.976	2.792	2.791	**2.783**	2.789	2.749
	Gamma	3.704	2.927	2.948	**2.761**	2.765	2.776	2.784	2.734
	LNormal	3.687	2.931	2.959	2.757	**2.755**	2.789	2.806	2.723
	InvGauss	3.657	2.910	2.928	2.747	**2.744**	2.783	2.797	2.724
$E[\lambda_i]=5$ $Var[\lambda_i]=2.5$	Truncated Normal	9.728	6.511	6.524	6.246	6.244	**6.216**	6.224	6.185
	Binomial	9.650	6.546	6.560	6.303	6.300	**6.276**	6.287	6.251
	Uniform	9.797	6.581	6.585	6.352	6.350	**6.329**	**6.329**	6.293
	Weibull	9.767	6.595	6.610	6.330	6.328	6.301	**6.300**	6.264
	Gamma	9.718	6.543	6.551	6.292	**6.287**	6.301	6.316	6.251
	LNormal	9.641	6.585	6.597	6.319	**6.313**	6.348	6.376	6.273
	InvGauss	9.720	6.549	6.565	6.323	**6.316**	6.356	6.380	6.276

Table 1.4 Comparison of the discrepancies using different distributions for the unknown parameters $\underline{\lambda}$ with "True" representing the full knowledge of $\underline{\lambda}$.

nomial distributions. These are the same distribution where the maximum entropy methods had smaller average KL differences in divergence. However these methods are less efficient than the maximum entropy gamma method in the case where the distributions are gamma, lognormal or inverse Gaussian. Nevertheless, even when our methods do not provide the smallest average discrepancy, their values are always close to the smallest one.

We can say that our MaxEnt and pseudo-MaxEnt methods, compared to other methods, seem to predict well when the amount of truncation and coefficient of variation are not very large. It is interesting to see that when we increase the expectation of λ_i to 5, both the MaxEnt method and the method based on the maximum entropy gamma prior with matching moments (MM) are often slightly more efficient than the same methods using maximum likelihood estimation.

Coverage probability and length of the prediction interval criteria

Although we may be interested in the goodness-of-fit of the method, one of the main preoccupations for recurrent event processes is prediction. If predictive densities and point prediction are of interest, the effectiveness of a method in finding adequate prediction intervals should also be assessed by checking their ability to have the desired coverage probability and a reasonable length. Here we seek to construct prediction intervals for a future random variable $\mathbf{N_i}(t_{1i}, t_{2i})$, given observed data $N(t_1)$.

To estimate the actual coverage probability of the prediction intervals, we used the simulated 2000 samples of the $n = 20$ HPP's. For each process, we calculated a one-sided 95% prediction interval of the form $[0, U(N(t_1))]$ for each of the seven methods considered. Since the quantity to predict is discrete, we randomized each interval. For each interval of the form $[0, U(N(t_1))]$ the length was normalized by dividing by $(t_{2i} - t_{1i})$ in order to render the interval lengths comparable, so that the average length becomes a meaningful measure. The results of these simulations are presented in Table 1.5. Note that each cell of this table contains the proportion of the $2000 \times 20 = 40000$ counts that were included in the corresponding 95% prediction interval and the number in parentheses corresponds to the average length of these intervals. Note that since the predictive densities used ignore the uncertainty between the estimated and the true parameters, we expect to obtain prediction intervals with coverage proportions below the nominal level of 95%. This problem can be corrected by calibrating the prediction intervals (Fredette and Lawless, 2007).

For the empirical Bayes methods, we notice how close to 0.95 the coverage probabilities are. The coverage probabilities given by these methods are never smaller than 94%. Also, it is interesting to note that MaxEnt and pseudo-MaxEnt methods, the methods based on the maximum entropy gamma prior and the method based on Jeffreys' prior yield similar coverage probabilities whatever the distribution used. However, the plug-in method assuming Poisson processes with different rates seems to use shorter intervals but fails to reach the desired coverage probability compared to the empirical Bayes methods. But with coverage

Moments for λ_i	Random effects	$\mathbb{P}(\hat{\lambda})$	$\mathbb{P}(\hat{\lambda}_i)$	Jeffreys	Gamma $(\hat{a}_{mle}, \hat{b}_{mle})$	Gamma $(\hat{a}_{mm}, \hat{b}_{mm})$	$MaxEnt$ $(\hat{\mu}_{mm}, \hat{\sigma}^2_{mm})$	$PseudoMaxEnt$ $(\hat{\mu}_{mle}, \hat{\sigma}^2_{mle})$
$E[\lambda_i]=1$ $Var[\lambda_i]=0.3$	Truncated Normal	0.858 (1.8184)	0.891 (1.7662)	0.949 (2.1173)	0.946 (1.9835)	0.946 (1.9774)	0.947 (1.9747)	**0.947 (1.9688)**
	Uniform	0.848 (1.8125)	0.888 (1.7736)	0.950 (2.1225)	0.947 (1.9903)	0.947 (1.9811)	0.948 (1.9755)	**0.950 (1.9599)**
	Weibull	0.862 (1.8150)	0.887 (1.7721)	0.949 (2.1253)	0.945 (1.9829)	0.945 (1.9811)	0.946 (1.9815)	**0.946 (1.9602)**
	Gamma	0.866 (1.8072)	0.887 (1.7642)	0.949 (2.1220)	0.946 (1.9665)	0.947 (1.9696)	0.945 (1.9715)	**0.945 (1.9548)**
	LNormal	0.875 (1.8309)	0.886 (1.7885)	0.948 (2.1481)	0.946 (1.9877)	0.947 (1.9917)	0.945 (1.9909)	**0.944 (1.9911)**
	InvGauss	0.873 (1.8190)	0.887 (1.7903)	0.950 (2.1483)	0.948 (1.9792)	0.948 (1.9861)	0.947 (1.9977)	**0.946 (1.9649)**
$E[\lambda_i]=5$ $Var[\lambda_i]=2.5$	Truncated Normal	0.824 (7.1259)	0.898 (7.1069)	0.952 (7.7308)	0.950 (7.5527)	0.950 (7.5527)	0.950 (7.5362)	**0.949 (7.5104)**
	Binomial	0.819 (7.1259)	0.895 (7.1069)	0.950 (7.7308)	0.949 (7.5527)	0.948 (7.5527)	**0.948 (7.5362)**	**0.948 (7.5104)**
	Uniform	0.804 (7.0999)	0.894 (7.0833)	0.949 (7.7066)	0.947 (7.5307)	0.947 (7.5328)	0.948 (7.5175)	**0.946 (7.5178)**
	Weibull	0.817 (7.1305)	0.896 (7.0891)	0.951 (7.7116)	0.950 (7.5431)	0.950 (7.5376)	0.949 (7.5215)	**0.949 (7.5210)**
	Gamma	0.829 (7.1179)	0.899 (7.1128)	0.951 (7.7378)	0.949 (7.5447)	0.949 (7.5497)	0.949 (7.5367)	**0.947 (7.5363)**
	LNormal	0.839 (7.1384)	0.897 (7.0997)	0.952 (7.7238)	0.950 (7.5375)	0.950 (7.5440)	0.949 (7.5324)	**0.947 (7.5283)**
	InvGauss	0.834 (7.0823)	0.897 (7.0609)	0.950 (7.6843)	0.946 (7.4884)	0.946 (7.4975)	0.945 (7.4888)	**0.943 (7.4762)**

Table 1.5 Coverage proportions and length of 95% prediction intervals.

probability close to 0.95 and shorter prediction intervals, the Pseudo-MaxEnt method seems to be the most parsimonious method.

From these simulations, we can draw three conclusions regarding the MaxEnt and pseudo-MaxEnt methods: first, they are efficient methods for point or interval prediction; secondly, they are very robust to the rate homogeneity/heterogeneity assumption when the amount of truncation and the empirical coefficient of variation are not very large; thirdly, the predictive model based on the maximum entropy methods performs better than the method based on the noninformative Jeffreys' prior and also performs slightly better than the method based on the maximum entropy gamma prior when the unknown parameters $\underline{\lambda}$ are generated with the truncated normal, Weibull or uniform distributions. However, these proposed methods are usually slightly less efficient than or equivalent to the maximum entropy gamma prior

method in the other cases.

To explain why the MaxEnt and pseudo-MaxEnt methods performed better for some distributions but not for others we examined the skewness and kurtosis for all these distributions. Using one of the pairs of parameter constraints used in this simulation study ($\mathbb{E}[\lambda_i] = 5$ and $\mathbb{V}ar[\lambda_i] = 2.5$), we see in Table 1.6 that the value of skewness and kurtosis are divided into two groups according to the distribution used. The first set of distributions that contains the truncated normal, Weibull, uniform and binomial and where the maximum entropy methods performed better than other methods have kurtosis values less than 3 and very small skewness values. For the second group which has kurtosis values greater than 3 and larger skewness values containing the gamma, lognormal and inverse Gaussian distributions, the maximum entropy gamma method performed better than the other methods.

Another restriction on the use of the maximum entropy methods is that if the coefficient

Moments for λ_i	Random effects	Skewness	Kurtosis
$\mathbb{E}[\lambda_i] = 5$ and $\mathbb{V}ar[\lambda_i] = 2.5$	TruncNormal	0.030	2.94
	Binomial	0.000	2.80
	Uniform	0.000	1.80
	Weibull	0.024	2.71
	Gamma	0.632	3.60
	LNormal	0.980	4.76
	InvGauss	0.949	4.50

Table 1.6 Skewness and kurtosis indicators for different unknown parameters distributions and parameter combination defined with $\mathbb{E}[\lambda_i] = 5$ and $\mathbb{V}ar[\lambda_i] = 2.5$.

of variation > 1 then the system of equations (1.6) has no solution so another method must be found (Wragg and Dowson, 1970). In this case, we would propose to use the maximum entropy gamma prior method as it is the most commonly used method for HPP's and based on our simulation results it performs well in these situations.

1.4.2 Data applications

In this section, we apply the maximum entropy methods to two examples. The first one concerns the occurrence of mammary tumors in laboratory animals taking part in a carcinogenicity experiment and the second one, a warranty data set from the automobile industry.

Mammary tumors in a carcinogenicity study

We consider the data presented in Gail et al. (1980) on the time to development of mammary tumors for $n = 48$ female rats. Rats were exposed to a carcinogen and 60 days later were randomized to receive either a treatment (23 animals) or to be part of a control group (25 animals). The data show the days on which new tumors occurred for each animal over a time period of 122 days. For this data set, we limited our interest to the treatment group which consists of 23 observations in order to have a sample size comparable to the sample size (n = 20) used in our simulation study. The main objective of this study here is the prediction of tumor occurrence for the 23 female rats in the treatment group using the maximum entropy methods and the comparison of the results obtained with those using the other methods studied.

Let $N_i(t)$ be the number of distinct tumors occurring up to time t for the ith subject $(i = 1, ..., n)$; time is elapsed time since the start of the study. To ensure that the MaxEnt and pseudo-MaxEnt methods and the other methods are adaptable to these data, we treat the occurrence times as continuous, as Gail et al.(1980) did. We find point predictors for $N_i(t_{1i}, t_{2i})$ for different values of t_{1i} increasing towards t_{2i} using the methods described here. Table 1.7 presents the average distance between the actual value of $N_i(t_{1i}, t_{2i})$ and its prediction, for $t_{2i} = 122$, using the different values of $t_{1i} = 10, 20, ..., 120$. It should be noted that for this data the estimated coefficient of variation $(\widehat{\sigma}/\widehat{\mu})$ is always less than 1 and the smallest average discrepancy appears in bold font in Table 1.7. From this table, we can see the disadvantage of using the method based on the Jeffreys' prior and that the maximum entropy methods have a better performance. The Pseudo-MaxEnt gives the best prediction with few exceptions, followed closely by the empirical Bayes method using the maximum entropy gamma prior with MM and MLE estimation methods. We also note that most of the

time, the estimators using MM for gamma and MaxEnt priors were less efficient than those obtained by MLE methods.

t_{1i}(in days)	Jeffreys	Gamma $(\widehat{a}_{mm}, \widehat{b}_{mm})$	Gamma $(\widehat{a}_{mle}, \widehat{b}_{mle})$	MaxEnt $(\widehat{\mu}_{mm}, \widehat{\sigma}^2_{mm})$	PseudoMaxEnt $(\widehat{\mu}_{mle}, \widehat{\sigma}^2_{mle})$
10	9.4324	2.1150	2.0796	2.1257	**2.0782**
20	4.5697	1.7352	1.7326	1.7356	**1.7306**
30	3.2929	2.8665	**1.5762**	**1.5762**	1.5768
40	2.8895	2.5810	1.5695	1.5695	**1.5431**
50	2.3072	1.3563	1.3547	1.3565	**1.3534**
60	1.6611	1.1715	1.1721	**1.1717**	1.1801
70	1.3013	1.1178	1.1237	**1.1170**	1.1196
80	1.2540	0.9183	0.9088	0.9070	**0.8939**
90	1.0108	0.8161	0.8140	0.8168	**0.8101**
100	0.7710	0.6606	0.6609	**0.6600**	0.6614
110	0.4619	0.3789	0.3779	0.3798	**0.3775**
120	0.2280	0.1699	**0.1695**	0.1699	**0.1695**

Table 1.7 Absolute error discrepancy of point predictors with different values of t_{1i} for the mammary tumors in a carcinogenicity data set (Treatment group).

Automobile Warranty Claims Study

Now we apply the same methods to a warranty data set from the automobile industry to predict the eventual number of warranty claims using the data already observed. This data set which describes warranty claims contains warranty information on 42188 cars which were sold over a period of 171 weeks. Each car had a three year or 36000 mile warranty, whichever came first. The following times were recorded for each car: production time, time of sale and, claim time(s) if any. This prediction of the number of claims is important for many reasons such as comparisons across production years or when interest centers on extrapolating the data to forecast the total cost of repairs.

Let $N_i(-\infty, t)$ be the total number of warranty claims for the i^{th} car t days after it was sold. We note that some claims could occur between the production day and the day of sale and that's why $N_i(-\infty, 0)$ is not necessarily equal to 0. Since $N_i(-\infty, 0)$ and $N_i(0, t)$ often

have to be modeled differently (Fredette and Lawless, 2007), we will focus here, without loss of generality, on the prediction of $\mathbf{N}_i(0, 365)$, the total number of warranty claims for each car i the first year after its sale. The range of the number of claims for each car was 0 to 22 claims and the total number of claims was 33438. We note here that in order to make this data set adaptable to a homogeneous Poisson process and therefore suitable for our methods, we prefer to use cars only during the first year of their three year warranty since the rate of occurrence of claims is almost constant throughout this time period. After one year, the rate of occurrence usually decreases because of mileage dropout. Table 1.8 shows the distribution of total claims amongst all the cars.

Number of claims	**Number of cars**
0	26.693
1	$7,911$
2	$3,421$
3	$1,773$
4	939
5	555
6	380
7	188
8	112
9	84
10+	129
$33,438$	$42,188$

Table 1.8 Frequency Distribution of all Warranty Claims.

We can see that 63% of the cars never had a warranty claim, 19% of the cars had only one claim and only 18% of the cars had 2 or more claims before the end of the warranty. Figure 1.1 gives a histogram of the occurrence times of claims during the year where each car is potentially under warranty. It appears that, except maybe for the first 50 days, the rate of occurrence of claims appears homogenous over the warranty claims time period.

In order to study this warranty data set from the automobile industry using the methods described here for point prediction, we will find predictors and the actual coverage probability

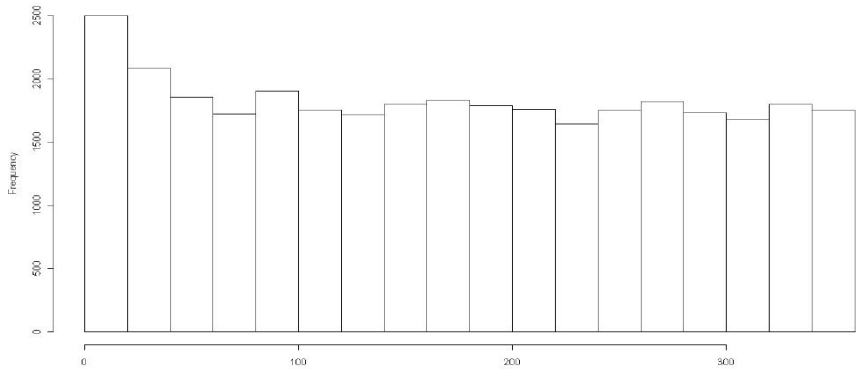

Figure 1.1 Histogram of the occurrence times.

of the prediction intervals for $N_i(t_{1i}, t_{2i})$ with different values of t_{1i} converging towards t_{2i}.

Table 1.9 presents the average distance between the actual value of $N_i(t_{1i}, t_{2i})$ and its prediction, where $t_{2i} = 365$ and with the different values of $t_{1i} = (45, 85, ..., 365)$. From this table, we can see the disadvantage of using the method based on the Jeffreys' prior. In fact using either of the MaxEnt and pseudo-MaxEnt methods usually gives the best prediction, followed closely by the empirical Bayes method using the maximum entropy gamma prior with the MLE method. We note that the MM estimators for the maximum entropy gamma method produced results less efficient than the MaxEnt and Pseudo-MaxEnt methods.

t_{1i}(in days)	Jeffreys	Gamma $(\widehat{a}_{mm}, \widehat{b}_{mm})$	Gamma $(\widehat{a}_{mle}, \widehat{b}_{mle})$	MaxEnt $(\widehat{\mu}_{mm}, \widehat{\sigma}^2_{mm})$	PseudoMaxEnt $(\widehat{\mu}_{mle}, \widehat{\sigma}^2_{mle})$
45	9.5166	6.6803	4.8853	4.8840	4.8802
85	4.7006	3.4990	2.5209	2.5198	2.5205
125	2.9445	2.3051	1.7251	1.7243	1.7237
165	2.0343	1.6734	1.3182	1.3176	1.3181
205	1.4567	1.2505	1.0467	1.0462	1.0466
245	1.0570	0.9482	0.8384	0.8380	0.8376
285	0.7340	0.6835	0.6352	0.6350	0.6349
325	0.4533	0.4382	0.4249	0.4249	0.4250
365	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1.9 Absolute error of point predictors with different values of t_{1i} for the automobile warranty claims data sets.

As both maximum entropy priors, the gamma and the truncated normal introduced here perform well for prediction in homogeneous Poisson processes but in different cases one performs better than the other we are faced with the question of which one to choose. The first test is to calculate the empirical centred moments of the data and determine whether $\boldsymbol{\mu_2^{'}} \leq 2\boldsymbol{\mu_1}^2$. If this answer is negative, then we must choose the gamma. If positive, then Table 1.6 in Section 4 indicates to us that the empirical skewness and kurtosis can be calculated and compared to the skewness and kurtosis calculated based on the gamma prior or truncated normal prior. We propose that this could be used as a decision aid. We illustrate this concept with the 2 data sets.

	Skewness	Kurtosis
Empirical Hypothetical	**0.4566984**	**2.136324**
Hypothetical truncated normal	**1.653976**	**3.002978**
Hypothetical gamma	2.005476	9.032898

Table 1.10 Theoretical and empirical skewness and kurtosis values for the mammary tumors in a carcinogenicity data set (Treatment group).

	Skewness	Kurtosis
Empirical Hypothetical	**1.388075**	**3.624457**
Hypothetical truncated normal	**1.646746**	**3.001122**
Hypothetical gamma	1.99911	8.994664

Table 1.11 Theoretical and empirical skewness and kurtosis values for the automobile warranty claims data sets.

The results in Tables 1.10 and 1.11 allow us say that for both data sets it is advantageous to use the truncated normal prior rather than the gamma prior since both hypothetical skewness and kurtosis values using the truncated normal prior for this prior are much closer to the empirical ones. This is especially true for the warranty data.

1.5 Concluding Remarks and Extensions

For recurrent event data modeled by homogeneous Poisson processes, we propose maximum entropy methods matching for the first and second moments for prediction of recurrent events and we show that the resulting truncated normal prior when it exists is an interesting alternative in many cases to the method using the maximum entropy gamma prior resulting

in the NB posterior distribution. The existence of the truncated normal prior as the solution requires that the empirical $\boldsymbol{\mu_2'} \leq 2\boldsymbol{\mu_1}^2$. Provided this solution exists, we use both matching moments and maximum likelihood to estimate the parameters for the truncated normal prior. Moreover, it has been shown that the truncated normal prior using either matching moments or maximum likelihood estimation methods is preferable to a totally noninformative Jeffreys' prior for the prediction of recurrent events studied here. A possible reason for this may be that maximum entropy methods allow us to make use of information that is available and find a prior that is as noninformative as possible using this minimal level of information. It is also superior to the usual frequency methods studied here.

In the comparison between the truncated normal prior and the gamma prior for prediction of future events in the homogeneous Poisson processes, the first criterion to check is whether the empirical $\boldsymbol{\mu_2'} \leq 2\boldsymbol{\mu_1}^2$ so that we can be more certain that the solution as a truncated normal prior exists. It should also be noted that the MaxEnt and the Pseudo-MaxEnt methods in the simulation study seem to predict well for the set of pairs of constraints on the unknown parameter vector $\underline{\lambda}$ when the amount of truncation and the coefficient of variation for the λ_i are not very large. Although both priors perform well, in some cases one method appears superior to the other and vice versa. To help in this choice when analysing data, we propose to compare the empirical skewness and kurtosis to the hypothetical skewness and kurtosis obtained by assuming that the data were generated by either the truncated normal prior or the gamma prior.

Both methods of estimation, the MaxEnt and Pseudo-MaxEnt, produce very good results in terms of the comparison criteria we used. Although the method of Pseudo-MaxEnt does not offer a major improvement over the MaxEnt (matching moments) method, it is a promising technique which requires further study.

Finally, we conjecture that the MaxEnt and Pseudo-MaxEnt methods could provide more flexibility for modeling various recurrent events distributions with heavy tails and various shapes by matching higher moments. This will be the subject of our future research for homogeneous Poisson processes and also for nonhomogeneous Poisson processes.

Acknowledgements

The research of the first author was partially supported by a fellowship from Université du Québec à Montréal (UQAM). The research of the second and third authors was partially supported by NSERC of Canada. All three authors want to thank the referee whose comments and suggestions greatly improved this manuscript.

1.6 References

• Broyden, C.G. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute for Mathematics and Applications,* **6**, 222–231.

• Cook, R.J. and Lawless,J.F (2007). *The Statistical Analysis of Recurrent Events.* Springer, New York.

• Fredette, M. and Lawless, J.F. (2007). Finite horizon prediction of recurrent events with application to forecast of warranty claims. *Technometrics,* **49**, 66–80.

• Gail, M. H., Santner, T. J. and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics,* **36**, 255–266.

• Gaver, D. P. and O'Muircheartaigh, I. G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics,* **29**, 1–15.

• Good I.J.(1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics,* **34**, 911–934.

• Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical Review,* **106**, 620–630, and **108**, 171–190.

• Jaynes, E.T. (1968). Prior probabilities. *IEEE Transactions Systems Science and Cybernetics,* **SSC-4**, 227–241.

• Jaynes, E.T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE,* **70**, 939–952.

• Jaynes, Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society,* A **186**, 453–461.

• Jeffreys, H. (1961). *Theory of Probability,* 3rd edition, Oxford University Press.

• Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics,* **22**, 79–86.

• Lisman, J.H.C. and Van Zuylen, M.C.A. (1972). Note on the generation of most probable

frequency distributions, *Statistica Neerlandica,* **26**, 19–23.

• Mohammad-Djafari, A. (1991). A Matlab program to calculate the maximum entropy distributions. Appeared in *Maximum Entropy and Bayesian Methods,* Series *Fundamental Theories of Physics*, **50**, 221–233.

• Mohammad-Djafari, A. (1992). Maximum likelihood estimation of the lagrange parameters of the maximum entropy distributions. Appeared in *Maximum Entropy and Bayesian Methods,* Series *Fundamental Theories of Physics*, **50**, 131–139.

• Rao, C.R., Kagan, A.M. and Y.V. Linnik (1973). *Characterization Problems in Mathematical Statistics,* John Wiley and Sons, New York.

• Shannon, C.E.(1948). The mathematical theory of communication. *Bell System Technical Journal,* **27**, 379–423.

• Skilling, J. (1989). Classic Maximum Entropy, In *Maximum Entropy and Bayesian Methods,* J. Skilling, Ed. Kluwer Academic, Norwell, MA. 45–52.

• Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual. Version 1.4* (http://www.mrc-bsu.cam.ac.uk/bugs.). Technical Report, Medical Research Council Biostatistics Unit. Cambridge.

• Talvila, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign, *American Mathematical Monthly,* **108**, 544–548.

• Ximing, W., (2003). Calculation of maximum entropy densities with application to income distribution, *Journal of Econometrics,* **115**, 347–354.

• Wang, H.M, Kalwani, M.U., and Akura, T. (2007). A Bayesian multivariate Poisson regression model of cross-category store brand purchasing behaviour, *Journal of Retailing and Consumer Services,* **14**, 369–382.

• Weinstock, R. (1952). *Calculus of Variations - With Applications to Physics and Engineering,* McGraw-Hill, New York.

• Wragg, A. and Dowson, D.C. (1970). Fitting continuous probability density functions over $(0, \infty)$ using information theory ideas. *IEEE Transactions on Information Theory,* **IT-16**, 226–230.

• Zellner, A. (1977). Maximal Data Information Prior Distributions, *In New Developments in the Applications of Bayesian Methods.* Ed. A. Aykac and C. Brumat, Amsterdam: North-Holland, 211-232.

• Zellner, A. and R. Highfield, R. (1988). Calculation of maximum entropy distributions and

approximation of marginal posterior distributions, *Journal of Econometrics,* **37**, 195–209.

• Zellner, A. and Tobias, J., (2001). Further results on Bayesian method of moments analysis of the multiple regression model. *International Economic Review,* **42**, 121–139.

Chapter 2

Choosing between higher moment maximum entropy models and its application to homogeneous point processes

Abstract

Using random effects in the modeling of homogeneous Poisson processes (HPP) has proved effective (Cook and Lawless (2007) and Gongjun et al. (2015)). We (Khribi et al. (2015)) compared the truncated normal prior (truncated at 0) to the usual gamma prior in the prediction for homogeneous Poisson processes and we concluded that the truncated normal prior (which is equivalent to the 2-moment maximum entropy prior) compared very favorably to the gamma one. This method was called the general Poisson-MaxEnt model. Unfortunately, because of the 2-moment condition on our maximum entropy prior, we were restricted to considering only cases where the coefficient of variation was less than or equal to 1 (Wragg and Dowson (1970)). Here we remove this restriction by the use of the k-moment maximum entropy prior ($k > 2$). The effectiveness of the general Poisson-MaxEnt model with this k-moment prior for prediction in HPP was measured by two goodness-of-fit criteria: Kullback-Leibler divergence and a discrepancy measure. The estimators obtained by these methods are compared to the estimators obtained with the two moment maximum entropy prior and the gamma prior

used by (Khribi et al. (2015)). The likelihood ratio test is used in order to determine when to stop adding higher order moments. We also illustrated on two examples: one concerning the occurrence of mammary tumors in laboratory animals taking part in a carcinogenicity experiment and the other, a warranty data set from the automobile industry.

Keywords: Recurrent events; mixed-Poisson; the maximum entropy principle; moment matching; maximum likelihood estimation; discrepancy measure; Kullback-Leibler divergence; likelihood ratio test; mean square prediction error.

2.1 Introduction

In the study of the prediction problem for homogeneous Poisson processes (HPP), the recurrent events often display extra-Poisson variation. This occurs in various fields such as biomedicine (Gongjun et al. (2015)), marketing (Brijs et al. (2004)) and reliability (Fredette and Lawless (2007)). Such variation is usually handled in an empirical Bayesian fashion and the gamma prior is the most common choice. In Khribi et al. (2015) we compared the performance of the two moment maximum entropy prior to other common ones such as the gamma prior in this prediction problem. We also compared two different estimation methods: the commonly used matching moments used by Aroian (1948), Wragg and Dowson (1970), Zellner and Highfield (1987), Mohammad-Djafari (1991, 1992), Ximing Wu (2003) and Khribi et al. (2015) as well as maximum likelihood (ML), as suggested by Mohammad-Djafari (1992), to estimate the moments in the two moment problem. The maximum likelihood estimation method did as well as the moment matching method and often outperformed it. Unfortunately, the Wragg and Dowson (1970) result implies that only cases where the coefficient of variation is less than one can be considered with this entropy prior. This has led us to consider higher moment maximum entropy priors here for this problem of prediction of recurrent events.

To the best of our knowledge, the maximum entropy prior with more than two moments has not previously been used in the prediction of recurrent events. Our aim here is to use the general Poisson-maximum entropy (Poisson-MaxEnt) model for this problem, where the heterogeneity between events is taken into account by the use of the prior obtained by maxi-

mizing the entropy. Given the excellent performance of the MLE method for the two moment problem and its ease of computation compared to matching moments especially when the number of moments is greater than 2, we chose to use MLE here. This will be called the MLE-MaxEnt method. The maximum entropy density with k-moments ($k > 2$) is of exceptional interest because it covers a wide variety of possibilities. For example, these densities have the possibility of bimodality, an important property which has been discussed in the literature (e.g., Eisenberger(1964) and Broadbent (1966)). We compare the k-moment maximum entropy prior distribution for the Poisson process with the two moment maximum entropy prior and the gamma prior which is popular among the conjugate priors and results in the negative binomial (NB) posterior distribution. It should be noted that the gamma distribution can also be considered as a maximum entropy distribution under different constraints (Mohammad-Djafari (1991)).

The remainder of this paper is organized as follows. In Section 2, we describe the maximum entropy principle and we recall the definition of a homogeneous Poisson process (HPP) and then we introduce the Poisson-MaxEnt model. Section 3 discusses the maximum likelihood approach to estimate the vector of parameters of this general Poisson-MaxEnt model. In Section 4, the performance of the k-moment maximum entropy priors for different values of k proposed here and their comparison with the use of the gamma conjugate prior are studied through Monte Carlo simulations. In order to test our methods we used many different priors to generate the original random effects including the k-moment maximum entropy priors, the gamma, the generalized gamma, the Weibull, the lognormal, the uniform and the inverse gaussian.

The performance of the k-moment maximum entropy priors is evaluated by the Kullback-Leibler criterion (Kullback-Leibler (1951)) and a discrepancy measure equal to the root mean square prediction error between the predicted value obtained using a specific prediction model and the estimator obtained here using our methods. Discrepancy is a general term usually measuring differences between an empirical value and the theoretical one. It is used in many different types of applications (for example, Chen et al. (2008) and Dick and Pillichshammer

(2014)).

The method is also illustrated using two real examples modelled by homogeneous Poisson processes, one of them concerning the occurrence of mammary tumors in laboratory animals taking part in a carcinogenicity experiment and the other one, a warranty data set from the automobile industry. Here we propose several different prediction models using k-moment priors with different values for k. We study the performance of such prediction models using the absolute error discrepancy equal to the absolute difference between point predictors calculated with different models for the random effects. We also assess the adequacy of these prediction models for these data sets on the basis of the likelihood ratio test. A general discussion with concluding remarks is presented in Section 5.

2.2 The Maximum Entropy Principle and The Homogeneous Poisson Process

Here we describe the maximum entropy principle and we introduce the homogeneous Poisson process (HPP) and the Poisson-MaxEnt model.

2.2.1 The Maximum Entropy principle

As we saw in our study (Khribi et al. (2015)), the entropy of a probability density $\pi(\lambda)$ is a measure of the amount of information contained in the density and was first defined by Shannon (1948) as

$$H = -\int_{\lambda} \pi(\lambda) \ln(\pi(\lambda)) d\lambda.$$

The goal is to maximize H subject to certain side conditions. The usual choice to determine $\pi(\lambda)$ is to use a finite set of expectations $\mu_j = \mathbb{E}[\phi_j(\lambda)]$ of known functions $\phi_j(\lambda), j = 0, ..., k$ and to match these empirical moments. This is called the matching moment (MM) estimation method. These known functions $\phi_j(\lambda)$ are often the arithmetic non-central moments of the form $\phi_j(\lambda) = \lambda^j, j = 0, ..., k$. In this simple case, using the arithmetic non-central moments, maximizing the likelihood yields the same estimates as the matching moment method (Mohammad-Djafari (1992)).

To find the function $\pi(\lambda)$ that maximizes the entropy of this nonlinear problem using matching moments, we form the Lagrangian

$$L \;=\; \sum_{j=0}^{k} \alpha_j \Big(\int_{\mathbf{R}^+} \lambda^j \pi(\lambda) d\lambda - \mu_j \Big).$$

where $\underline{\boldsymbol{\alpha}}$ is a vector of Lagrange multipliers. Applying the Lagrange's multiplication method (Weinstock (1952)), the following k-moment maximum entropy prior distribution is defined by:

$$\pi(\lambda|\underline{\boldsymbol{\alpha}}) \;=\; A \exp\,(-\sum_{j=1}^{k} \alpha_j \lambda^j), \tag{2.1}$$

where $\underline{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, ..., \alpha_k)$ and with normalization constant defined by:

$$A \;=\; \frac{1}{\int_{\mathbb{R}^+} \exp\,(-\sum_{j=1}^{k} \alpha_j \lambda^j) d\lambda}.$$

Clearly the k-moment maximum entropy distribution given k non-central moments that maximizes the entropy is of interest for modeling data. For example, to specify a mixture consisting of a linear combination of two normal distributions, we need five independent parameters. This case has been treated in detail by several authors (e.g., Cohen (1967) and Gridgeman (1970)), and it can clearly give rise to a bimodal probability distribution. However, unless we can readily identify the existence of two component distributions, it would seem beneficial if we use a model requiring fewer than five parameters, if available; the maximum entropy prior given four non-central moments is just such a model.

2.2.2 Homogeneous Poisson Process

We let $N(s,t)$ represent the number of events occurring for a subject in the time interval $(s,t]$ with $N(t)$ representing $N(0,t)$. To model such recurrent events, many different types of processes are discussed in the literature (see Cook and Lawless (2007)) where the Poisson process (PP) is one of the most popular ones. We will consider here only continuous time processes where two events cannot occur simultaneously. The number of events can be defined

through the intensity function,

$$\lambda(t|H(t)) = \lim_{\triangle t \to 0} \frac{P[N(t, t+\triangle t) = 1|H(t)]}{\triangle t},$$

where $H(t)$ denotes the history of the process up to time t.

Here the intensity function we use to model these events is the one corresponding to a HPP where the rates are the unknown parameters. Suppose that we have n subjects and that $N_i(t)$ denotes the number of events occurring for a subject i up to time t. When these processes are time-homogeneous, the model is defined by:

$$\begin{aligned} N_i(t)|\lambda_i &\sim PP(\lambda_i), \\ \lambda_i &\sim \pi(\lambda_i), \end{aligned} \tag{2.2}$$

where the processes are independent and $i = 1, ..., n$. We also suppose here that each process is observed up to a fixed time t_{1i} and that the interest is to find a point predictor or a prediction interval for $N_i(t_{1i}, t_{2i})|N(t_{1i})$'s for $i = 1, ..., n$. Throughout this article, $(\lambda_1, \lambda_2 ..., \lambda_n)$, $(N_1(t_{11}), ..., N_n(t_{1n}))$ and $(N_1(t_{11}, t_{21}), ..., N_n(t_{1n}, t_{2n}))$ will be denoted by $\underline{\lambda}$, $\underline{N}(t_1)$ and $\underline{N}(t_1, t_2)$ respectively.

2.2.3 Model specification of the Poisson-MaxEnt model

The general Poisson-MaxEnt model with k-moments that we develop is then given by:

$$\begin{aligned} N_i(t)|\lambda_i &\sim HPP(\lambda_i), \\ \pi(\lambda_i; \underline{\boldsymbol{\alpha}}) &= A \exp\left(-\sum_{j=1}^{k} (\alpha_j \lambda_i^j)\right), \end{aligned} \tag{2.3}$$

where $A = \frac{1}{\int_{\mathbb{R}^+} \exp\left(-\sum_{j=1}^{k} \alpha_j \lambda^j\right) d\lambda}$ is a normalization constant and $\underline{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, ..., \alpha_k)$ is a vector of parameters. Clearly, α_k must be positive.

For the general Poisson-Maximum Entropy model (2.3), the joint posterior distribution of all the unknown parameters $\underline{\boldsymbol{\lambda}}|\underline{N}(t_1)$ is given by

$$\begin{aligned}\pi(\underline{\boldsymbol{\lambda}}|\underline{N}(t_1);\underline{\boldsymbol{\alpha}}) &= \frac{P[\underline{\mathbf{N}}(t_1)=\underline{N}(t_1)|\underline{\boldsymbol{\lambda}}]\pi(\underline{\boldsymbol{\lambda}};\underline{\boldsymbol{\alpha}})}{\int_{\underline{\boldsymbol{\lambda}}} P[\underline{\mathbf{N}}(t_1)=\underline{N}(t_1)|\underline{\boldsymbol{\lambda}}]\pi(\underline{\boldsymbol{\lambda}};\underline{\boldsymbol{\alpha}})d\underline{\boldsymbol{\lambda}}} \\ &= \prod_{i=1}^{n} \frac{\lambda_i^{N_i(t_{1i})}\exp\Big(-\lambda_i(\alpha_1+t_{1i})-\sum_{j=2}^{k}\alpha_j\lambda_i^j\Big)}{\int_{\boldsymbol{\lambda_i}}\lambda_i^{N_i(t_{1i})}\exp\Big(-\lambda_i(\alpha_1+t_{1i})-\sum_{j=2}^{k}\alpha_j\lambda_i^j\Big)d\lambda_i}. \end{aligned} \tag{2.4}$$

Hence, using this conditional density, the density function for $N_i(t_{1i},t_{2i})|N_i(t_{1i})$ is then given by

$$\begin{aligned}P[N_i(t_{1i},t_{2i})=n|N_i(t_{1i});\underline{\boldsymbol{\alpha}}] &= \frac{(t_{2i}-t_{1i})^n}{n!\int_{\lambda_i}\lambda_i^{N_i(t_{1i})}\exp\Big(-\lambda_i(\alpha_1+t_{1i})-\sum_{j=2}^{k}\alpha_j\lambda_i^j\Big)d\lambda_i} \\ &\times\int_{\lambda_i}\lambda_i^{(N_i(t_{1i})+n)}\exp\Big(-\lambda_i(\alpha_1+t_{2i})-\sum_{j=2}^{k}\alpha_j\lambda_i^j\Big)d\lambda_i. \end{aligned} \tag{2.5}$$

We note that the posterior distribution (2.4) will not have a known closed form, but includes rather complicated high dimensional densities, which makes direct inference almost impossible because direct mathematical derivation of these posterior distributions involves high-dimensional integration which is not mathematically tractable to obtain the normalizing constant. For this reason, we generate from this posterior distribution a large number of samples using Markov chain Monte Carlo (MCMC) implemented in WinBUGS (Spiegelhalter et al.(2003)), and from these samples, we can obtain appropriate parameter estimates such as the posterior mean of $\underline{\boldsymbol{\lambda}}|(\underline{N}(t_1);\underline{\boldsymbol{\alpha}})$, where $\underline{\boldsymbol{\alpha}}$ is estimated by the methods described in the next section.

2.3 Estimating unknown Poisson-Maximum Entropy parameters

The objective of both estimation approaches, matching moments and maximum likelihood estimation, in the general Poisson-Maximum Entropy model is to choose the probability distribution $\pi(\lambda_i;\underline{\boldsymbol{\alpha}}) = A\exp\left(-\sum_{j=1}^{k}(\alpha_j\lambda_i^j)\right)$ for the unknown vector of parameters $\underline{\boldsymbol{\alpha}}$ which best represents the observed data $\underline{N}(t)$.

In this section, we discuss ways to estimate the vector of the k parameters $\underline{\boldsymbol{\alpha}}$ of the general Poisson-MaxEnt model. In the study (Khribi et al. (2015)) for the $k = 2$ problem, the parameters of the maximum entropy distribution with only two moments in the Poisson-MaxEnt model were estimated by two methods, the usual maximum entropy estimation method which uses matching moments (MM) and the maximum likelihood method, referred to as the MLE-MaxEnt here. However, the MLE-MaxEnt method seemed more promising than the MaxEnt method. Here, we favour the MLE-MaxEnt method because it is computationally less complex than MM method when $k > 2$. For completeness the matching moment (MM) method for the MaxEnt-Poisson model will be described in appendix A.

2.3.1 The MLE-Maximum Entropy Method for the Poisson-MaxEnt Model

Here we introduce the MLE-maximum entropy (MLE-MaxEnt) method using MLE for estimating the parameters of the empirical Bayes MaxEnt model (2.3). To obtain them, we construct the marginal likelihood L of the empirical Bayes general Poisson-Maximum Entropy model (2.3)

$$\begin{aligned} L(\underline{\boldsymbol{\alpha}}|\underline{N}(t_1)) &= \int_{\underline{\lambda}} P[\underline{\mathbf{N}}(t_1) = \underline{N}(t_1)|\underline{\boldsymbol{\lambda}}]\pi(\underline{\lambda};\underline{\boldsymbol{\alpha}})d\underline{\lambda} \\ &= \prod_{i=1}^{n} \frac{t_{1i}^{N_i(t_{1i})}}{N_i(t_{1i})!} I_i(N_i), \end{aligned}$$

with

$$I_i(N_i) = \int_0^{\infty} \lambda_i^{N_i(t_{1i})} \exp\Big[-\Big(\sum_{j=1}^{k} \alpha_j \lambda_i^j + t_{1i}\lambda_i \Big)\Big] d\lambda_i. \tag{2.6}$$

Ignoring the terms that do not depend on $\underline{\boldsymbol{\alpha}}$, the log-likelihood is given by

$$l(\underline{\boldsymbol{\alpha}}|\underline{N}(t_1)) \propto \sum_{i=1}^{n} \Big[ln\Big(I_i(N_i) \Big)\Big]. \tag{2.7}$$

One method to find the maximum of (2.7) is to take the partial derivatives and set them equal to 0 (Zellner and Highfield (1987)). Using Lebesgue's Dominated Convergence The-

orem, Talvila (2001) gave necessary and sufficient conditions to interchange the order of differentiation and integration which are verified for $I(N_i)$ here. Interchanging differentiation and integration, the first derivatives of (2.7) with respect to $\alpha_1, \alpha_2, ..., \alpha_k$ are given by the following equations

$$\frac{\partial l(N_i)}{\partial \alpha_j} = \sum_{i=1}^{n} \left[\frac{-\int_0^\infty \lambda_i^{(N_i(t_{1i})+j)} \exp\left[-\left(\sum_{j=1}^{k} \alpha_j \lambda_i^j + t_{1i}\lambda_i \right) \right]}{\int_0^\infty \lambda_i^{(N_i(t_{1i}))} \exp\left[-\left(\sum_{j=1}^{k} \alpha_j \lambda_i^j + t_{1i}\lambda_i \right) \right]} \right] = 0, j = 1, 2, ..., k \quad (2.8)$$

The analytic solutions to the equations in (2.8) are difficult to obtain; thus it is natural to use a numerical method to estimate directly the vector of parameters $\underline{\boldsymbol{\alpha}}$ that maximize the log-likelihood (2.7).

We have chosen MATLAB "fminsearchbnd", a nonlinear optimization method which is derivative-free and allows bounds on the variables for this MLE problem. Under our model (2.3) for HPP, matching moments and the MLE-maximum entropy methods for Poisson-MaxEnt model will not yield the same estimates. This differs from the simple case without Poisson processes considered by Mohammad-Djafari (Mohammad-Djafari (1992)).

2.4 Simulation studies and data applications

2.4.1 Simulation Studies

Through extensive simulation studies presented in this section, we will study and compare the performance of our general Poisson-MaxEnt model with this k-moment prior ($k > 2$) to the models using the two moment maximum entropy prior or the gamma prior where the parameters were estimated using the MLE method. For comparison, we use the following goodness-of-fit criteria: Kullback-Liebler distance and a discrepancy measure for point predictions equal to the root mean square prediction error.

Throughout this study, we know that the advantage of using the general Poisson-MaxEnt model with this k-moment prior ($k > 2$) for prediction in HPP is that it can be used regard-

less of the values of the coefficient of variation. This allows us to reflect different orders of heterogeneity in the data. Among all results obtained with different values of the coefficient of variation, we only present here the results for two of them in order to be concise: the first one represents a case where the value of the coefficient of variation is less than 1 and the second where it is greater than 1. The latter case is used in order to show the benefit of using the general Poisson-MaxEnt model with this k-moment prior when $k > 2$ and thus removing the restriction of Wragg and Dowson (1970) that we have in this situation when $k \leq 2$.

In order to study which models among those presented in this paper are more robust to the real distribution of $\underline{\boldsymbol{\lambda}}$, we have chosen several different priors to represent the unknown prior distribution for the HPP, such as the maximum entropy distributions given 2, 4 and 6 non-central moments, the gamma, the Weibull, the lognormal, the inverse Gaussian, the generalized gamma and the continuous uniform distribution, to generate the unknown parameters $\underline{\lambda}$.

Moreover for each value of the coefficient of variation used for these simulation studies, we generate $b = 2,000$ samples from $n = 20$ HPP's and assume that these processes are observed up to the times $t_1 = \{5, 5.5, ..., 9.5, 10, ..., 14.0, 14.5\}$ and we want to predict $N_i(t_{1i}, t_{2i})$ for each process i up to the times $t_2 = \{12.5, ..., 12.5, 17.5, ..., 17.5\}$. The idea behind this choice is to represent different values of $(t_{2i} - t_{1i})$.

Kullback-Liebler divergence

One method which allows us to compare the performance of these models is to use their predictive distributions for $\mathbf{N}_i(t_{1i}, t_{2i})$. Such a comparison can be done by evaluating how close each predictive density $\widetilde{f}_p(y|x)$ is to the true density $f(y|x;\theta)$ where θ is a vector of unknown parameters. In the literature, to judge the goodness-of-fit of a given predictive method (e.g. Vidoni (1995) and Komaki (1996)), a common approach has been to assess this relative closeness with the average Kullback-Liebler (KL) divergence (Kullback and Leibler

(1951)) which is defined by

$$D_{KL}\Big[\widetilde{f}_p(y|x); f(y|x;\theta)\Big] = \mathbb{E}\Big[\log\Big\{\frac{f(\mathbf{Y}|\mathbf{X};\theta)}{\widetilde{f}_p(\mathbf{Y}|\mathbf{X})}\Big\}\Big] = \int f(y|x;\theta)\log\Big(\frac{f(y|x;\theta)}{\widetilde{f}_p(y|x)}\Big)dy,$$

where X and Y represent an actual and a future random variable respectively. We note also that this divergence is positive unless $\widetilde{f}_p(y|x)$ always coincides with $f(y|x;\theta)$.
If the real distribution of $\mathbf{N}_i(t_{1i}, t_{2i})$ is known, we can compare the distance between these predictive densities and the real density of $\mathbf{N}_i(t_{1i}, t_{2i})$. This should give us an indication of the ability of these methods to provide adequate prediction of $\mathbf{N}_i(t_{1i}, t_{2i})|\mathbf{N}_i(t_{1i})$.
We measure which predictive density considered is closer to the true one, $f(N(t_1, t_2)|N(t_1);\theta)$, as follows. If we have two contenders, for example, $\widetilde{f}_{MLE-MaxEnt}(N(t_1, t_2)|N(t_1))$ and $\widetilde{f}_p(N(t_1, t_2)|N(t_1))$, for the role of estimates of the true one, $f(N(t_1, t_2)|N(t_1);\theta)$, then $\widetilde{f}_{MLE-MaxEnt}(N(t_1, t_2)|N(t_1))$ is closer in terms of KL divergence than $\widetilde{f}_p(N(t_1, t_2)|N(t_1))$ if
$D_{KL}\Big[\widetilde{f}_p(N(t_1, t_2)|N(t_1)); f(N(t_1, t_2)|N(t_1);\theta)\Big] - D_{KL}\Big[\widetilde{f}_{MLE\ \ MaxEnt}(N(t_1, t_2)|N(t_1)); f(N(t_1, t_2)|N(t_1);\theta)\Big]$
$= D_{KL}\Big[\widetilde{f}_{MLE\ \ MaxEnt}(N(t_1, t_2)|N(t_1)), \widetilde{f}_p(N(t_1, t_2)|N(t_1))\Big]$ is positive.

Here, the average KL divergence will be estimated by simulating $b = 2,000$ samples of $n = 20$ HPP and will be defined by

$$\begin{aligned}\hat{D}_{KL}&\Big[\tilde{f}_{MLE-MaxEnt}(N(t_1, t_2)|N(t_1)), \tilde{f}_p(N(t_1, t_2)|N(t_1))\Big] \\ &= \frac{1}{b}\sum_{j=1}^{b}\log\Big\{\frac{\tilde{f}_{MLE-MaxEnt}(N_j^*(t_1, t_2)|N_j^*(t_1))}{\tilde{f}_p(N_j^*(t_1, t_2)|N_j^*(t_1))}\Big\},\end{aligned} \tag{2.9}$$

where $N_j^*(t_1)$ and $N_j^*(t_1, t_2)$ are the counts generated for the jth sample, $j = 1, ..., b$ and and $\tilde{f}_p(N_j^*(t_1, t_2)|N_j^*(t_1))$ is the predictive density obtained from the other model to which we are comparing our estimator.

The results of these simulations are presented in Table 2.1. We use the priors in the first column of Table 2.1 to generate the "true" random effects and we estimate, using MLE, these random effects from our four chosen models: the gamma, the 2-moment, the 4-moment and the 6-moment maximum entropy prior. Each cell of this table contains the value of the average KL divergence given by (9) between the predictive density $\tilde{f}_{MLE-MaxEnt}(N_j^*(t_1, t_2)|N_j^*(t_1))$

Moments for λ_i	Random effects	$\text{Log}[\frac{MLE\ 6Moments}{Gamma}]$	$\text{Log}[\frac{MLE\ 6Moments}{MLE\ 2Moments}]$	$\text{Log}[\frac{MLE\ 6Moments}{MLE\ 4Moments}]$
$E[\lambda_i]=1$ $Var[\lambda_i]=0.3$ c.v.=0.56	Gamma	**-1.0**	0.8	0.3
	MaxEnt2MM	0.7	**-0.9**	0.2
	MaxEnt4MM	5.9	6.3	**-0.5**
	MaxEnt6MM	8.2	7.5	4.1
	Ggamma	7.4	6.2	0.9
	Weibull	3.2	2.4	0.7
	LNormal	1.9	1.1	0.4
	InvGauss	2.2	1.7	0.8
	Uniform	7.6	6.4	2.8
$E[\lambda_i]=1$ $Var[\lambda_i]=1.5$ c.v.=1.2	Gamma	**-0.3**	31.3	9.7
	MaxEnt2MM	13.1	11.8	6.2
	MaxEnt4MM	13.7	27.8	**-0.0**
	MaxEnt6MM	16.2	29.7	12.7
	Ggamma	12.9	25.7	7.9
	Weibull	15.3	26.8	10.7
	LNormal	13.4	28.2	8.9
	InvGauss	12.9	24.7	7.4

Table 2.1 Comparison of the average KL distance with the general Poisson-MaxEnt model with the 6-moment prior as reference model with different values of the coefficient of variation (c.v.). To render the table more readable, the values of the KL distances have been multiplied by 1000.

for the general Poisson-MaxEnt model with the 6-moment reference prior and the other predictive densities $\tilde{f}_p(N_j^*(t_1,t_2)|N_j^*(t_1))$ for the other models. We note that a negative value on a line in this table for a given distribution of $\underline{\boldsymbol{\lambda}}$ is written in bold font and it indicates that the predictive model in that column performs better than our reference model in terms of KL divergence. Therefore, the absence of negative values on a given line indicates that our reference method is the most suitable for this distribution of $\underline{\boldsymbol{\lambda}}$. It is also noted that the higher this value is for the other models, the better the performance of our reference model compared to the other models.

We note first that the table indicates that the general Poisson-MaxEnt model with this 6-moment prior (as a reference model) performs well compared to the other models: the values are always positive except for some cases where the true distribution of $\underline{\boldsymbol{\lambda}}$ corresponds perfectly to the method used (Gamma, MaxEnt2MM or MaxEnt4MM). Indeed, when the value of the coefficient of variation is ≤ 1 and the random effects are neither generated by the gamma or a MaxEnt prior, we note that the Poisson-MaxEnt model with the 2-moment

prior and the NB model (gamma prior estimated with the MLE method) are similar in terms of performance where their predictive densities are closest to the true predictive density $f(N(t_1, t_2)|N(t_1))$. Our reference model performed clearly better than the Poisson-MaxEnt models with the 2-moment and slightly outperformed the 4-moment prior. However, when the value of the coefficient of variation is > 1, the NB model performs much better than the Poisson-MaxEnt model with 2 moments. Nevertheless, our reference model performs is still clearly better than the NB model and the Poisson-MaxEnt model with 2 or 4 moments except when the random effects are generated by the gamma or the 4-moment prior. Except the few exceptions mentioned above where the true distribution of $\underline{\boldsymbol{\lambda}}$ corresponds perfectly to the method used, we see that our reference model always provide the closest predictive density. Moreover, whatever the value of the coefficient of variation used for these simulation studies, our reference model has a better performance compared to the other models and thus exhibits a robustness to the type of distribution of $\underline{\lambda}$.

Finally we note that when the value of the coefficient of variation is greater than 1 the Poisson-MaxEnt model with the maximum entropy 2-moment prior gives a **positive** value of (2.9) (=**11.8**) for the KL divergence in spite of the fact that theoretically the coefficient of variation must be ≤ 1 in order for this prior to be defined (Wragg and Dowson (1970)). The use of this approach is therefore not recommended in this case; however, the results are presented here as well for illustrative purposes.

Discrepancy measure

We also compare the adequacy of each point prediction method for $N_i(t_{1i}, t_{2i})$ obtained from one of the four models for the random effects used in this simulation study. We used the following discrepancy measure, defined here as follows:

$$D = \sqrt{\frac{\sum_{i=1}^{n} \left(N_i(t_{1i}, t_{2i}) - \hat{N}_i(t_{1i}, t_{2i}) \right)^2}{n}}, \tag{2.10}$$

where $\hat{N}_i(t_{1i}, t_{2i})$ is the point predictor provided by the model chosen and estimated using maximum likelihood estimation. The value of D represents, for a given sample of n processes,

the root mean square prediction error between the true value of $N_i(t_{1i}, t_{2i})$ and its predictor. The above discrepancy measure is called the root mean square prediction error. Discrepancy is a term often used to describe a method which compares an empirical value to a theoretical one.

The results of these simulations are presented in Table 2.2. We use the priors in the first column of Table 2.2 to generate the "true" random effects and we estimate, using MLE, these random effects from our four chosen models: the gamma, the 2-moment, the 4-moment and the 6-moment maximum entropy prior.

We begin by generating the 20 λ_i's ($i = 1, ..., 20$) using one of the models in the first column of Table 2.2. From this we can obtain $N_i(t_{1i}) = Poisson(t_{1i}\lambda_i)$, $i = 1, ..., 20$ and $N_i(t_{1i}, t_{2i}) = Poisson((t_{2i} - t_{1i})\lambda_i)$, $i = 1, ..., 20$. We then estimate $\lambda_i|N_i(t_{1i})$ by the method suggested in Section 2.3, that is, the Markov Chain Monte Carlo method of Spiegelhalter et al. (2003) implemented in WinBUGS. The predictor of $N_i(t_{1i}, t_{2i})$ will equal $(t_{2i} - t_{1i})$ times the posterior mean of $\lambda_i|N_i(t_{1i})$ and it will be denoted by $\hat{N}_i(t_{1i}, t_{2i})$. These values are then put into equation (2.10) to obtain the discrepancy. Table 2.2 consists of the value over 1 of the ratio of two discrepancy measures where the denominator is calculated using the true random effects distribution and the numerator is calculated using one of our four chosen models. We note that in Table 2.2 the smallest value in this table for a given distribution of $\underline{\boldsymbol{\lambda}}$ is written in bold font and it corresponds to the most suitable model. A value close to 0 means that the point predictor $\hat{N}_i(t_{1i}, t_{2i})$ provided by the model chosen is very close to the true value and thus that model performs very well. For example, a value of **1.11** in the first line of Table 2.2 means that a prediction based on this model (the NB model) would be on average **1.11**% off from the best possible prediction measured by our discrepancy measure when all the true parameters are known.

From the results in this table, the first thing we can say is that the general Poisson-MaxEnt model with the 6-moment prior distribution has the best performance in term of our discrepancy measure when predicting $N_i(t_{1i}, t_{2i})$. Even when our model does not provide the smallest value , its value is always very close to the smallest one. It is also robust to the type of distribution used to generate the random effects. That means, the real distribution of the rates $\underline{\lambda}$

Moments for λ_i	Random effets	Gamma $(\widehat{a}_{mle}, \widehat{b}_{mle})$	*MLE 2Moments*	*MLE 4Moments*	*MLE 6Moments*
$E[\lambda_i]=1$ $Var[\lambda_i]=0.3$ c.v.=0.56	**Gamma**	**1.11**	1.29	1.26	1.15
	MaxEnt2MM	0.90	**0.72**	0.86	0.79
	MaxEnt4MM	0.47	0.50	**0.22**	0.32
	MaxEnt6MM	0.61	0.68	0.36	**0.11**
	Ggamma	0.72	0.79	0.54	**0.32**
	Weibull	0.32	0.47	0.29	**0.22**
	LNormal	1.04	1.18	0.97	**0.89**
	InvGauss	0.75	0.93	0.68	**0.61**
	Uniform	0.47	0.72	0.40	**0.25**
$E[\lambda_i]=1$ $Var[\lambda_i]=1.5$ c.v.=1.2	**Gamma**	**0.38**	1.41	0.54	0.40
	MaxEnt2MM	0.58	1.45	0.51	**0.33**
	MaxEnt4MM	0.73	1.20	**0.22**	**0.22**
	MaxEnt6MM	0.98	1.41	0.54	**0.40**
	Ggamma	0.98	1.24	0.62	**0.40**
	Weibull	1.09	1.49	0.69	**0.44**
	LNormal	0.90	1.12	0.51	**0.33**
	InvGauss	0.94	1.20	0.47	**0.33**

Table 2.2 Comparison using our discrepancy measures, the root mean square prediction error, for the gamma and the general Poisson-MaxEnt model with $k = 2, 4$ or 6 moments versus the best possible prediction assuming full knowledge of λ_i that is, where the λ_i are generated by one of the models listed in the column of random effects. We note that the smallest percentage of error prediction in this table for a given distribution of $\underline{\lambda}$ is written in bold font.

seems to have a small influence on our method. Our model give us the smallest value (value written in bold font) whatever the distribution used to generate the unknown parameters $\underline{\lambda}$ with the exception of the cases where the random effects were generated by the gamma or MaxEnt priors and this corresponds to the same pattern observed using the KL divergence.

When the value of the coefficient of variation is greater than 1, the Poisson-MaxEnt model with the 2-moment prior give us the **largest** value of the ratio (=**1.45**) although we used the 2-moment maximum entropy distribution to generate the random effects for the unknown parameters $\underline{\lambda}$. Again we refer to the Wragg and Dowson result concerning all densities of the form $\pi(\lambda) = Ae^{-(\alpha_1\lambda+\alpha_2\lambda^2)}$ require that the coefficient of variation ≤ 1 and although we do not recommend this approach in this case, we have added the results for illustrative purposes.

It appears from these simulations that the general Poisson-MaxEnt model with the 6-moment prior distribution is more effective than the other models in finding point predictors of $N_i(t_{1i}, t_{2i})$, the number of events of a HPP.

Finally, we conclude that although that the NB model can be used in the prediction problem for HPP whatever the value of coefficient of variation, so can the maximum entropy prior with $k > 2$ moments. However, these simulation studies indicate that the general Poisson-MaxEnt model with higher moments gives us a better performance than the NB model in the case where the value of coefficient of variation is ≤ 1. When $k = 2$, the Poisson-MaxEnt with 2 moments and the NB models are similar in term of performance. On the other hand, when $k > 2$ and the value of the coefficient of variation is greater than 1, then the general Poisson-MaxEnt model with the k-moment prior ($k = 4$ or 6) truly outperforms the classical NB model.

2.4.2 Data applications

In this section, we apply the general Poisson-MaxEnt model with the k-moment prior using the MLE-MaxEnt estimation method to two examples. The first one concerns the time to occurrence of mammary tumors in laboratory rats in a carcinogenicity experiment using data from Gail et al. (1980) and the second one, a warranty data set from the automobile industry using data from Kalbfleisch et al.(1991). These two examples have been previously analysed using 2-moment priors in Khribi et al. (2015). For these two examples, we propose a suitable prediction model and study the performance of such model through the discrepancy measure given by (10). But first, we propose an approach to determine an adequate value for k in practical applications.

Likelihood Ratio Tests

The likelihood ratio test (LRT) is a hypothesis test that helps us to determine when to stop adding higher order moments. Using the log-likelihood functions for two models, let us say the null model with the k-moment maximum entropy prior with $k = 2$ or 4 and the alternative model, the model with the $(k+2)$-moment maximum entropy prior, then the test statistic is the ratio of the log-likelihood of the null model to the alternative model:

$$\Gamma \quad = \quad -2\log\Big(\frac{l(\underline{\boldsymbol{\alpha}}_1|\underline{N}(t_1))}{l(\underline{\boldsymbol{\alpha}}_2|\underline{N}(t_1))}\Big) \tag{2.11}$$

where $l(\underline{\boldsymbol{\alpha}}_1|\underline{N}(t_1))$ and $l(\underline{\boldsymbol{\alpha}}_2|\underline{N}(t_1))$ are the log-likelihood of the null and alternative models respectively.
This is a statistical test for nested models which rejects the null hypothesis with a given significance level based on the chi-square distribution (Wilks (1938)). Through successive testing using the LRT, we determine the number of moments necessary for the k-moment prior.

Mammary tumors in a carcinogenicity study

We consider the data presented in Gail et al. (1980) on the time to development of mammary tumors for $n = 48$ female rats. Rats were exposed to a carcinogen and 60 days later were randomized to receive either a treatment or no treatment. The data consisted of the days on which new tumors occurred for each animal over a time period of 122 days. We limited our interest to the 23 observations in the treatment group in order to have a sample size comparable to the sample size (n = 20) used in our simulation study. We studied these using the MLE-MaxEnt estimation method for the general Poisson-MaxEnt model with the k-moment prior and the comparison of the results obtained with models which are defined with the two moment maximum entropy prior and the gamma prior using the MLE method.

Let $N_i(t)$ (t = time elapsed since the start of the study) be the number of distinct tumors occurring up to time t for the ith subject ($i = 1, ..., n$) and $n = 23$. To ensure that the general Poisson-MaxEnt model and the NB model are adaptable to these data, we treat the occurrence times as continuous, as Gail et al.(1980) did. We find point predictors for $N_i(t_{1i}, t_{2i})|N_i(t_{1i})$ at times $t_{1i} = 10, 20, ..., 120$ and $t_{2i} = 122$ for all $i = 1, ..., 23$ of this 122 day long longitudinal process. We first determine the value of k by comparing the model based on a $(k+2)$-maximum entropy prior for the random effects versus the one based on the k-moment one using the likelihood ratio test (Wilks (1938)). Then we study the performance of our model through our discrepancy measure.

For each time t_{1i}, Table 2.3 presents the likelihood ratio test (LRT) results where the last three columns indicate respectively the p-values of the k-moment maximum entropy prior

with k=2, 4 and 6 as the null models versus the $(k+2)$-moment maximum entropy prior as the alternative models. Note that the last column shows us the number of moments suggested for our model.

Based on the results in Table 2.3 with a significance level equal to 5%, we can say that the LRT always rejects the general Poisson-MaxEnt model with the 8-moment prior. Therefore, it supports our model with the 6-moment maximum entropy prior as the adequate prediction model when the maximum entropy prior is used. This means adding other moments does not allow us to reject our 6-moment predictive model; thus, the LRT provides a stopping rule. However, we should note that there is one exception, when $t_{1i} = 120$ the LRT suggest $k = 2$.

t_{1i}	p-value of LRT (MaxEnt_2MM_vs_4MM)	p-value of LRT (MaxEnt_4MM_vs_6MM)	p-value of LRT (MaxEnt_6MM_vs_8MM)	Number of Moments Suggested
10	< 0.01%	**< 0.01%**	78.90%	**6**
20	< 0.01%	**< 0.01%**	87.05%	**6**
30	< 0.01%	**< 0.01%**	91.30%	**6**
40	< 0.01%	**< 0.01%**	98.74%	**6**
50	< 0.01%	**< 0.01%**	100%	**6**
60	0.04%	**0.09%**	100%	**6**
70	0.22%	**0.13%**	100%	**6**
80	0.79%	**0.57%**	100%	**6**
90	0.91%	**1.33%**	100%	**6**
100	1.52%	**3.98%**	100%	**6**
110	3.46%	**4.08%**	100%	**6**
120	**5.05%**	9.63%	100%	**2**

Table 2.3 The likelihood ratio test for the mammary tumors in a carcinogenicity data set.

Table 2.4 presents the absolute error difference between the real value of $N_i(t_{1i}, t_{2i})$ and its predictor $\hat{N}_i(t_{1i}, t_{2i})|\hat{N}_i(t_{1i})$ defined in Section 4.1.2. For example, a value of **1.86** in the first line of Table 2.4 means that a prediction based on this model (the general Poisson-MaxEnt model with the 6-moment prior) would be on average **1.86** from the real value of $N_i(t_{1i}, t_{2i})$. The likelihood ratio test appears to be an appropriate stopping rule, since its

corresponding average discrepancy values (values in bold font) are always very close to the smallest absolute error discrepancy. Furthermore, the most interesting result in Table 2.4 is that Poisson MaxEnt model suggested by the LRT always outperforms the NB model.

t_{1i}(in days)	Gamma $(\widehat{a}_{mle},\widehat{b}_{mle})$	*MLE 2Moments*	*MLE 4Moments*	*MLE 6Moments*	*MLE 8Moments*
10	2.08	2.08	1.91	**1.86**	1.85
20	1.73	1.73	1.62	**1.57**	1.57
30	1.58	1.58	1.46	**1.42**	1.42
40	1.57	1.54	1.41	**1.38**	1.38
50	1.35	1.35	1.27	**1.25**	1.25
60	1.17	1.18	1.13	**1.09**	1.09
70	1.12	1.12	0.95	**0.91**	0.91
80	0.91	0.89	0.86	**0.83**	0.82
90	0.81	0.81	0.77	**0.74**	0.74
100	0.66	0.66	0.58	**0.56**	0.56
110	0.38	0.38	0.32	**0.29**	0.29
120	0.17	**0.17**	0.16	0.16	0.16

Table 2.4 Absolute error discrepancy of point predictors with different values of t_{1i} for the mammary tumors in a carcinogenicity data set with MLE-MaxEnt estimation method.

Automobile Warranty Claims Study

Now we apply the same methods to a warranty data set from the automobile industry presented in Kalbfleisch et al.(1991) and Fredette and Lawless (2007)) to predict the eventual number of warranty claims using the data already observed. This data set which describes warranty claims contains warranty information on 42 188 cars which were sold over a period of 171 weeks.

Here $N_i(t)$ represents the number of claims at time t since the day of sale. We are interested in predicting $\mathbf{N}_i(365)$, the total number of warranty claims for each car i during the first year after its sale. The range of the number of claims for each car was 0 to 22 claims and the total number of claims was 33 438. Table 2.5 shows the distribution of total claims

amongst all the cars.

Number of claims	Number of cars
0	26.693
1	7,911
2	3,421
3	1,773
4	939
5	555
6	380
7	188
8	112
9	84
10+	129
33,438	42,188

Table 2.5 Frequency Distribution of Warranty Claims during the first year after the day of sale (Khribi et al. (2015)).

Figure 2.1 (Khribi et al. (2015)) gives a histogram of the occurrence times of claims during the year where each car is potentially under warranty. Except possibly for the first 50 days, the rate of occurrence of claims appears homogenous over the warranty claims time period.

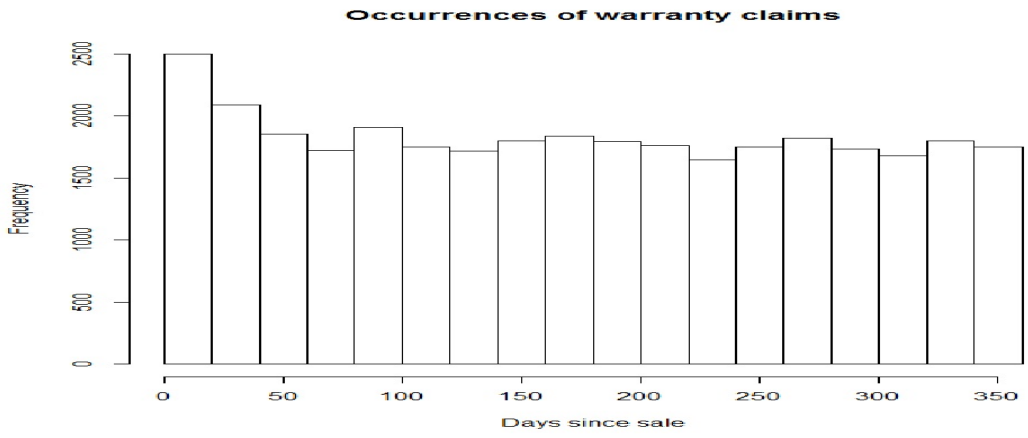

Figure 2.1 Histogram of the occurrence times (Khribi et al. (2015)).

Here, we calculate point predictors for $N_i(t_{1i}, t_{2i})$ with different values of t_{1i} converging towards $t_{2i} = 365$ with these predictive models. For every time t_{1i}, Table 2.6 present the LRT

results where the last three columns indicate respectively the p-values of the k-moment maximum entropy prior with k=2, 4 and 6 as the null models versus the $(k+2)$-moment maximum entropy prior as the alternative models. Note that the last column shows us the number of moments required for our model. Moreover, Table 2.7 presents the absolute error discrepancy measure as t_{2i} converges towards 365, using the different values of $t_{1i} = (45, 85, ..., 365)$.

Based on the results in Table 2.6 and with a significance level equal to 5%, we can say that the LRT supports the model with the 6-moment maximum entropy prior as adequate for prediction and thus we do not need to add other moments. This validates the results we see in Table 2.7, where the average absolute error discrepancy for the general Poisson-MaxEnt model with the 6-moment maximum entropy prior (values in bold font) are always very close to the lowest one. The only exception occurs when t_{1i} equals 365, where the LRT supports the MaxEnt prior with only 2-moments.

$\mathbf{t_{1i}}$	p-value of LRT (MaxEnt_2MM_vs_4MM)	p-value of LRT (MaxEnt_4MM_vs_6MM)	p-value of LRT (MaxEnt_6MM_vs_8MM)	Number of Moments Suggested
45	< 0.01%	**< 0.01%**	72.18%	**6**
85	< 0.01%	**0.02%**	83.73%	**6**
125	< 0.01%	**0.10%**	90.84%	**6**
165	0.04%	**0.19%**	97.61%	**6**
205	0.11%	**0.68%**	99.91%	**6**
245	0.85%	**3.02%**	100%	**6**
285	2.48%	**4.30%**	100%	**6**
325	3.07%	**4.08%**	100%	**6**
365	**44.85%**	93.32%	100%	**2**

Table 2.6 The likelihood ratio test for the automobile warranty claims data sets.

Finally, we note from these two examples that when $k = 2$, the Poisson-MaxEnt with 2-moments and the NB models are similar in terms of performance. On the other hand, when $k > 2$, the general Poisson-MaxEnt predictive model suggested by the LRT is clearly more adequate than the classical NB model using the conjugate gamma prior.

2.5 Concluding Remarks and Extensions

In this paper we have outlined a model, the general Poisson-MaxEnt model with the k-moment prior for prediction problem for HPP. This model is tested as to its effectiveness for prediction measured by different goodness-of-fit criteria. We note that the use of this prior

t_{1i}(in days)	Gamma $(\widehat{a}_{mle}, \widehat{b}_{mle})$	*MLE 2Moments*	*MLE 4Moments*	*MLE 6Moments*	*MLE 8Moments*
45	4.89	4.88	4.70	**4.61**	4.61
85	2.52	2.52	2.41	**2.34**	2.34
125	1.73	1.72	1.64	**1.57**	1.57
165	1.32	1.32	1.25	**1.19**	1.19
205	1.05	1.05	0.94	**0.90**	0.90
245	0.84	0.84	0.76	**0.73**	0.73
285	0.64	0.63	0.58	**0.55**	0.55
325	0.42	0.43	0.40	**0.38**	0.38
365	0.00	**0.00**	0.00	0.00	0.00

Table 2.7 Absolute error discrepancy of point predictors with different values of t_{1i} for the automobile warranty claims data sets with MLE-MaxEnt estimation method.

with more than two moments allows us to remove the restriction of Wragg and Dawson that the coefficient of variation must be less than one.

We used maximum likelihood estimation method to estimate the parameters in the general Poisson-MaxEnt model because it is computationally less complex than the matching moments procedure when $k > 2$. The use of the k-moment maximum entropy prior produced very good results in terms of the comparison criteria (KL divergence and a discrepancy measure) we used here with different values of the coefficient of variation in our simulation studies. Finally, we have shown using two data sets the effectiveness of the general Poisson-MaxEnt model for prediction problems. The LRT was used in the analysis of the two data sets as a stopping rule for adding more moments.

We know that the classical NB model obtained with the conjugate gamma prior is the usual choice for prediction problems. This model can be used whatever the value of coefficient of variation. Through these simulation studies and the two data sets, we have seen that the Poisson-MaxEnt model with $k > 2$ has generally a better performance than the NB model whether the value of coefficient of variation is smaller or greater than one.

In our future research, it will be very interesting to use these methods allowing prediction of recurrent events using flexible nonhomogeneous Poisson processes with priors that have

heavy tails and various shapes and with possible heterogeneity amongst the individual units modeled with higher moment maximum entropy priors.

Appendix A: Moment Matching Estimation for the Poisson-MaxEnt Model

Here we illustrate how to use the moment matching (MM) estimation for the Poisson-MaxEnt method.

We consider the maximum entropy method for the empirical Bayes general Poisson-MaxEnt model with moment matching (MM). This involves matching the first k unconditional moments with $k > 2$ with their corresponding empirical non-central moments.
Letting $R_i = \frac{N_i(t_{1i})}{t_{1i}}$, we can show that the first k unconditional moments for an unknown parameter λ_i truncated at 0 are defined by:
$E[\lambda_i] = E[R_i] = \boldsymbol{\mu_1}$; $E[\lambda_i^2] = E[R_i^2] - \frac{1}{t_{1i}}\boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ and for $j > 2$, we have the following recurrence formula

$$\mu_j \;=\; E[R_i^j] - \sum_{s=1}^{j-1} t_{1i}^{s-j}\mu_s\Big[\frac{1}{s!}\sum_{m=0}^{s}(-1)^{s-m}\binom{s}{m}m^{j-1}\Big] \qquad j = 3, 4, ..., k.$$

Then we have to solve the following nonlinear system of equations

$$\int_{\mathbf{R}^+} \lambda_i^j \pi(\lambda|\underline{\boldsymbol{\alpha}})d\lambda_i \;=\; \hat{\boldsymbol{\mu}}_{\mathbf{j}} \qquad j = 1, 2, ..., k. \tag{2.12}$$

with $\pi(\lambda|\underline{\boldsymbol{\alpha}}) = A\exp\left(-\sum_{j=1}^{k}(\alpha_j\lambda_i^j)\right)$ with $A = \frac{1}{\int_{\mathbb{R}+}\exp\left(-\sum_{j=1}^{k}\alpha_j\lambda^j\right)d\lambda}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{1}} = \overline{R}$; $\hat{\boldsymbol{\mu}}_{\mathbf{2}} = \overline{R^2} - \overline{t_{1i}^{-1}}\hat{\boldsymbol{\mu}}_{\mathbf{1}}$; and the following recurrence formula for $\hat{\boldsymbol{\mu}}_{\mathbf{j}}$

$$\hat{\boldsymbol{\mu}}_{\mathbf{j}} \;=\; \overline{R^j} - \sum_{s=1}^{j-1}\overline{t_{1i}^{s-j}}\hat{\mu}_s\Big[\frac{1}{s!}\sum_{m=0}^{s}(-1)^{s-m}\binom{s}{m}m^{j-1}\Big] \qquad j = 3, 4, ..., k.$$

where $\overline{t_{1i}^{-j}}$ is the sample average of $\frac{1}{t_{1i}^j}$, while $\overline{R^j}$ is the sample average of the R_i^j's.

One way to solve this problem is to transform the nonlinear system of equations (2.12) into an unconstrained optimization problem and then use a numerical integration and the

"fminsearchbnd" MATLAB function described earlier to obtain an exact density

$$\pi(\lambda|\underline{\boldsymbol{\alpha}}) = A\exp\left(-\sum_{j=1}^{k}(\alpha_j\lambda_i^j)\right) \quad for \quad k > 2. \tag{2.13}$$

Here we note that Mohammad-Djafari (1991), has implemented his numerical method in MATLAB which allows us to estimate the vector of parameters in the maximum entropy distribution.

Another way is to specify the k-moment prior completely. For that, we substitute (2.13) into (2.12) and solve this highly nonlinear set of equations for the $\underline{\boldsymbol{\alpha}}$ in terms of the k known empirical moments.

For given $\hat{\boldsymbol{\mu}}_{\mathbf{1}}$, $\hat{\boldsymbol{\mu}}_{\mathbf{2}}$,...,$\hat{\boldsymbol{\mu}}_{\mathbf{k}}$ the corresponding values of α_1, α_2,...,α_k are obtained by solving the set nonlinear equations

$$\int_{\mathbf{R}^+} \lambda_i^j e^{-\left[\sum_{j=1}^{k}\alpha_j\lambda_i^j\right]} \;=\; \hat{\boldsymbol{\mu}}_{\mathbf{j}} \int_{\mathbf{R}^+} e^{-\left[\sum_{j=1}^{k}\alpha_j\lambda_i^j\right]} \qquad j = 1, 2, ..., k. \tag{2.14}$$

We see that this is not an easy problem as it involves, among other things, the integration of an exponential function in which the exponent is of degree k and also that no general analytic solution exist for this highly nonlinear set of equations. That's why we adopt a numerical method that should lead to good approximate solutions for the vector $\underline{\boldsymbol{\alpha}}$.
Let $\underline{\boldsymbol{\alpha}}^0 = (\alpha_1^0, \alpha_2^0, ..., \alpha_k^0)$ is a vector of initial values of $\underline{\boldsymbol{\alpha}}$ and let ϵ the vector defined by the equations

$$\epsilon_j \;=\; \alpha_j - \alpha_j^0 \qquad j = 1, 2, ..., k. \tag{2.15}$$

By linearizing (2.14) we see that the ϵ_j approximately satisfy k simultaneous equations of the form

$$(W_{i+j} - \hat{\boldsymbol{\mu}}_{\mathbf{i}} W_i)\epsilon_i \;=\; W_i - \hat{\boldsymbol{\mu}}_{\mathbf{i}} C_0 \qquad i, j = 1, 2, ..., k \tag{2.16}$$

where

$$W_t \;=\; \int_{\mathbf{R}^+} \lambda_i^t e^{-\left[\sum_{j=1}^{k} \alpha_j \lambda_i^j\right]} d\lambda_i \qquad t = 1, 2, ..., 2k.$$

Thus, given a sufficiently good initial approximation $(\alpha_1^0, \alpha_2^0, ..., \alpha_k^0)$, the system (2.16) is solved for $\underline{\boldsymbol{\alpha}} = \underline{\boldsymbol{\alpha}}^0 + \boldsymbol{\epsilon}$, which becomes our new vector of trial $\underline{\boldsymbol{\alpha}}$, and iterations continue until $\boldsymbol{\epsilon}$ becomes appropriately small.

We note that for $t < k$ a numerical calculation with MATLAB has to be performed, but for $t \geq k$ recurrence relations of the form

$$W_t \;=\; \frac{1}{k\alpha_k}\Big[(t+1-k)W_{(t-k)} - \sum_{j=1}^{k-1} \alpha_j W_{(t+j-k)}\Big]$$

can be used.

2.6 References

• Aroian, L.A. (1948). The fourth degree exponential distribution function, *Annals of Mathematical Statistics,* **19**, 589–592.

• Brijs, T., Karlis, D., Swinnen, G., Vanhoof, K., Wets, W. and Manchanda, P., (2004). A multivariate Poisson mixture model for marketing applications *Statistica Neerlandica,* **58**, No. **3**, 322–348.

• Broadbent, D. E. (1966). A difficulity in assessing bimodality in certain distributions, *British Journal of Mathematical and Statistical Psychology,* **19**, 125–126.

• Chen, F.; Curran, P.; Bollen, K.A.; Kirby, J.and Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models, *Sociological Methods & Research*, **36**, No.**4**, 462–494.

• Cohen, A.C. (1967). Estimation in mixture of two normal distributions, *Technometrics* **9**, 15–28.

• Cook, R.J. and Lawless,J.F. (2007). *The Statistical Analysis of Recurrent Events*, New York, Springer.

• Dick, J. and Pillichshammer, F.(2014). *Discrepancy Theory and Quasi-Monte Carlo Inte-*

gration, New York, Springer.

• Eisenberger,I. (1964). Genesis of bimodal distributions, *Technometrics*, **6**, 357–363.

• Fredette, M. and Lawless, J.F. (2007). Finite horizon prediction of recurrent events with application to forecast of warranty claims. *Technometrics,* **49**, 66–80.

• Gail, M. H., Santner, T. J. and Brown, C. C. (1980), An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics,* **36**, 255–266.

• Gongjun, X., Sy Han, C., Chiung-Yu, H., Mei-Cheng, W. and Jun, Y. (2015). Joint scale-change models for recurrent events and failure time. *Journal of the American Statistical Association,* **111**, 1–38.

• Gridgeman, N. T. (1970). A comparison of two methods of analysis of mixtures of normal distributions, *Technometrics*, **12**, 823–833.

• Kalbfleisch, J. D., Lawless, J. F. and Robinson, J. (1991), Methods for the analysis and prediction of warranty claims, *Technometrics*, **33**, 273–285.

• Khribi, L., Fredette, M. and MacGibbon, B. (2016). The Poisson maximum entropy model for homogeneous Poisson processes. *Communications in Statistics - Simulation and Computation,* **45**, No. **09**, 3435–3456.

• Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika*, **83**, 299–313.

• Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics,* **22**, 79–86.

• Lawless, J.F. and Fredette, Marc (2005). Frequentist prediction intervals and predictive distributions, *Biometrika*, **92**, 529–542.

• Mohammad-Djafari, A. (1991). A Matlab program to calculate the maximum entropy distributions. Appeared in *Maximum Entropy and Bayesian Methods,* Series *Fundamental Theories of Physics*, **50**, 221–233.

• Mohammad-Djafari, A. (1992). Maximum likelihood estimation of the lagrange parameters of the maximum entropy distributions. Appeared in *Maximum Entropy and Bayesian Methods,* Series *Fundamental Theories of Physics*, **50**, 131–139.

• Shannon, C.E.(1948). The mathematical theory of communication. *Bell System Technical Journal,* **27**, 379–423.

• Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual. Version 1.4* (http://www.mrc-bsu.cam.ac.uk/bugs.). Technical Report, Medical Research Coun-

cil Biostatistics Unit. Cambridge.

• Talvila, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign, *American Mathematical Monthly,* **108**, 544–548.

• Vidoni, P. (1995). A note on modified estimative prediction limits and distributions, *Biometrika,* **85**, 949–953.

• Weinstock, R. (1952). *Calculus of Variations - With Applications to Physics and Engineering*, New York, McGraw-Hill. • Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics*, **9**, No.**1**, 60–62.

• Wragg, A. and Dawson, D.C. (1970). Fitting continuous probability density functions over $(0, \infty)$ using information theory ideas, *IEEE Transactions on Information Theory,* **IT-16**, 226–230.

• Wu, X., (2003). Calculation of maximum entropy densities with application to income distribution, *Journal of Econometrics* **115**, No.**2**, 347–354.

• Zellner, A. and R. Highfield, R. (1987). Calculation of maximum entropy distributions and approximation of marginal posterior distributions, *Journal of Econometrics,* **37**, 195–209.

Chapter 3

A Nonhomogeneous Poisson process predictive model using maximum entropy prior random effects with application to predict purchases

Abstract

Using a higher order maximum entropy prior for the random effects in the prediction of future events for a homogeneous Poisson process compared favorably to the usual gamma prior (Khribi et al., 2016). In this paper the time homogeneity assumption is relaxed and we propose a predictive model for recurrent events using flexible nonhomogeneous Poisson processes (NHPP). In addition, possible heterogeneity amongst the individual units is modeled using higher moment maximum entropy priors instead of the gamma prior. We assess the performance of such a model with a real data set from a loyalty program and compared its adequacy to the negative binomial model using the conjugate gamma prior.

Keywords: Recurrent events; mixed-Poisson; Nonhomogeneous Poisson process; maximum entropy principle; maximum likelihood; forecasting.

3.1 Introduction

This paper deals with prediction problems, specially problems when recurrent events are concerned. In many settings, it is important for example to predict the numbers of events that will occur in future time periods. The motivation for our work lies in the prediction of individual activity level based on the data already observed or other events that occur for individual units or subjects in a population. The database at hand was obtained from a loyalty program at a major commercial airline.

The prediction of recurrent events has been discussed in specific contexts such as warranty claims (e.g. Khribi et al., 2016 and Fredette and Lawless, 2007), insurance claims (e.g. England and Verrall, 2002) where predictions are used, for example, for fiscal planning or for taxation purposes and often in software reliability (e.g. Amin et al., 2013) where decisions about the time to stop testing and releasing software are influenced by predictions of the number of new bugs that would be found if testing were to continue. In these situations, the recurrent events often display extra-Poisson variation usually handled by an empirical Bayesian model using a gamma prior. To relax the time homogeneity assumption, we develop a model allowing finite-horizon prediction of recurrent events using flexible nonhomogeneous Poisson processes with higher moment maximum entropy priors random effects as a predictive model. Such a model has not previously been used in the prediction of recurrent events.

According to our two previous articles (Khribi et al., 2015) and (Khribi et al., 2016), we have seen the importance of using the higher moment maximum entropy prior instead of the gamma prior when we deal with problems when recurrent events are concerned. For this reason, the main idea of our article is to compare the performance of the proposed predictive model using the higher moment maximum entropy prior for the random effects with the negative binomial (NB) model used for example by (Fredette and Lawless, 2007) to forecast automobile warranty claims which uses a gamma prior for the random effects. The choice of a gamma prior was motivated by its nice mathematical properties when used with Poisson processes.

The remainder of this paper is organized as follows. In Section 2, we outline the assumptions relative to the use of mixed nonhomogeneous Poisson models for predicting recurrent events. Then, we describe the maximum entropy principle before providing details on the development and estimation of our model. In Section 3, the performance of the proposed approach is studied in a particular data setting from a major airline company and its comparison with the NB model using the conjugate gamma prior. We finally conclude with a discussion of our results, limitations, and avenues for future research in Section 4.

3.2 Prediction of Recurrent Events with Mixed Poisson Models

Here we present the nonhomogeneous Poisson processes, then we recall the maximum entropy principle used to take account the possible heterogeneity amongst the individuals and presented for example in our studies (Khribi et al., 2015 and Khribi et al., 2016), finally we define the Poisson-maximum entropy model.

3.2.1 Mixed Nonhomogeneous Poisson Processes

The motivation for our work lies in the prediction of a real data set from a loyalty program or other events that occur for individual units or subjects in a population. That is, there is a finite population of units $i = 1, \ldots, n$ and we wish to predict the total number of events, for an individual or the whole population over a specified time period $(0, T]$ on the basis of events that have already occurred up to given times $t_i \leq T$ for the units in the population. In practice, this interval $(0, T]$ would typically refer to a calendar or fiscal year time period, and the various t_i's would usually take the same value for all units.

Let t represent the number of days elapsed since the beginning of the calendar year, and let $N_i(u, v)$ denote the number of events in the age interval $u < t \leq v$. The objective is then to predict $N_i(0, T)$ the total number of events associated with unit i up until time T, where T could possibly represent the end of the year. Of course, as an added benefit, it may eventually be useful to predict the total number of events during the whole year by predicting

$$N_+ (0, T) = \sum_{i=1}^{n} N_i (0, T) . \tag{3.1}$$

Of course, $N_i(0,T)$ will ultimately be known for each i once the calendar year is over. However, in several situations, it might be useful to predict this value on the basis of previous experience with this unit but also on the basis of events that have already occurred during that calendar year. Because $N_i(0,t_i)$ is known for each $i = 1,\ldots,n$ where $0 \le t_i \le T$, prediction of $N_i(0,T)$ is equivalent to predicting $N_i(t_i,T)$.

For convenience, we consider continuous time processes where two events cannot occur simultaneously. From this point on, we also write $N(t)$ for $N(0,t)$. Different types of recurrent events processes are discussed in the literature on point processes (Grandell, 1997). These are all characterized by an event intensity function

$$\lambda(t|\, H(t)) = \lim_{\Delta t \to 0} \frac{P[N(t,t+\Delta t) = 1|\, H(t)]}{\Delta t} \tag{3.2}$$

where $H(t)$ denotes the history of the process up to time t. Poisson processes are Markovian because (3.2) depends only on t. The intensity, or rate, function is then simply denoted by $\lambda(t)$, and

$$N(t) \sim PP(\lambda(t))$$

means that $N(t)$ is a nonhomogeneous Poisson process (NHPP) with rate function $\lambda(t)$.

It is well known that in a Poisson process, the total number of events over any interval has a Poisson distribution, and that the number of events $N(s_1,t_1)$ and $N(s_2,t_2)$ in two nonoverlapping time intervals $(s_1,t_1]$ and $(s_2,t_2]$ are independent. These two properties make Poisson processes easy to use with prediction problems involving recurrent events. However, in populations with heterogeneous units, it is generally necessary to extend the models by including unit-specific random effects. Such models are termed random-effects, or mixed, Poisson processes (e.g., Lawless, 1987; Grandell, 1997).

We model the rate function for a single process with parametric forms $\lambda(t;\alpha,\beta) = \alpha f(t;\beta)$, where α is a scalar and β is a vector of low dimension. This parameterization is convenient because $f(t;\beta)$ and α measure different aspects of a NHPP; the function $f(t;\beta)$ describes the shape of the rate function, and α represents the overall event frequency. In the finite-horizon problems, it is convenient to choose α so that $\mathrm{E}[N(0,T)] = \alpha$, in which case $\int_0^T f(t;\beta)dt = 1$.

That is, $f(t;\beta)$ has the form of a probability density function over $(0,T]$.

3.2.2 The Maximum Entropy principle

As we saw in our studies (Khribi et al., 2015 and Khribi et al., 2016), the entropy of a probability density $\pi(\alpha)$ is a measure of the amount of information contained in the density which was first defined by Shannon (1948) as

$$H = -\int_{\alpha} \pi(\alpha)\ln(\pi(\alpha))d\alpha.$$

The goal is to maximize H subject to certain side conditions. The usual choice to determine $\pi(\alpha)$ is to use a finite set of expectations $\mu_j = \mathbb{E}[\phi_j(\alpha)]$ of known functions $\phi_j(\alpha), j = 0, ..., k$. This is called the matching moment (MM) estimation method. These known functions $\phi_j(\alpha)$ are often the arithmetic non-central moments of the form $\phi_j(\alpha) = \alpha^j, j = 0, ..., k$. In this simple case using the arithmetic non-central moments maximizing the likelihood yields the same estimates as the matching moment method (Mohammad-Djafari (1992)).

To find the function $\pi(\alpha)$ that maximizes the entropy of this nonlinear problem using matching moments we form the Lagrangian

$$L = \sum_{j=0}^{k} \gamma_j \Big(\int_{\mathbf{R}^+} \alpha^j \pi(\alpha)d\alpha - \mu_j \Big).$$

where $\boldsymbol{\gamma}$ is a vector of Lagrange multipliers. Applying the Lagrange's multiplication method (Weinstock, 1952). The following k moment maximum entropy prior distribution is defined by:

$$\pi(\alpha|\boldsymbol{\gamma}) = A\exp\,(-\sum_{j=1}^{k}\gamma_j\alpha^j), \tag{3.3}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_k)$ and with normalization constant defined by:

$$A = \frac{1}{\int_{\mathbb{R}^+} \exp\left(-\sum_{j=1}^{k} \gamma_j \alpha^j\right) d\alpha}.$$

3.2.3 Model specification of the general Poisson-maximum entropy model

To consider scenarios in which heterogeneity is observed among the processes for different units, we incorporate unobservable *iid* random effects in our model by using the k moment maximum entropy distribution given k non-central moments. The general Poisson-MaxEnt model considered in this article is

$$\begin{aligned} N_i(t)|\alpha_i &\sim PP(\alpha_i f(t;\beta)), \\ \pi(\alpha_i;\boldsymbol{\gamma}) &= A \exp\left(-\sum_{j=1}^{k} \gamma_j \alpha_i^j\right), \end{aligned} \tag{3.4}$$

where $i = 1, \ldots, n$ and $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)$ and with normalization constant defined by:

$$A = \frac{1}{\int_{\alpha_i} \exp\left(-\sum_{j=1}^{k} \gamma_j \alpha_i^j\right) d\alpha_i}.$$

We will propose later an efficient criterion which allows us to determine the number of moments necessary for the k-moment priors in the model (3.4). And for our particular data set studied here it will be seen further that the model (3.4) performs very well when the number of moments k is equal to 4.

3.2.4 Prediction

We seek to construct prediction intervals for a future random variable Y, given observed data $X = x$. Such intervals are of the form $(L(x), U(x))$, and we attempt to find intervals where $P[L(X) \leq Y \leq U(X)]$ equals some specified fixed value $1 - \zeta$, in which case $(L(x), U(x))$ is called a $1 - \zeta$ prediction interval (e.g., Lawless and Fredette, 2005) and $1 - \zeta$ is called its coverage probability.

In the context discussed in this article, we wish to use the information regarding the n processes that are available at a certain given time to make predictive statements about the remaining number of events that would be observed. As it is the focus of our article, only the prediction of a single count $N_i(t_i, T)$ is discussed here. It is easy to make the extension to predict the sum of all counts (1).

For each process, the information available to make our prediction consists of the total number of events, $N_i(t_i)$, and the set of occurrence times, $\tau_i(t_i) = \{\tau_{i1}, \ldots, \tau_{iN_i(t_i)}\}$. The conditional distribution $\pi(\boldsymbol{\alpha}|(\mathbf{N}(t), \tau); \boldsymbol{\gamma}, \boldsymbol{\beta})$ of the random effects is defined by:

$$\begin{aligned}\pi(\boldsymbol{\alpha}|(\mathbf{N}(t), \tau); \boldsymbol{\gamma}, \boldsymbol{\beta}) &= \frac{\mathbf{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\mathbf{N}(t), \tau))\pi(\boldsymbol{\alpha}; \boldsymbol{\gamma})}{\int_{\boldsymbol{\alpha}} \mathbf{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\mathbf{N}(t_1), \tau))\pi(\boldsymbol{\alpha}; \boldsymbol{\gamma}) d\boldsymbol{\alpha}} \\ &= \prod_{i=1}^{n} \frac{\alpha_i^{N_i(t_i)} \exp\Big(- \alpha_i(\gamma_1 + F(t_i; \boldsymbol{\beta})) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big)}{\int_{\boldsymbol{\alpha_i}} \alpha_i^{N_i(t_i)} \exp\Big(- \alpha_i(\gamma_1 + F(t_i; \boldsymbol{\beta})) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big) d\alpha_i} \end{aligned} \quad (3.5)$$

where

$$\mathbf{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}|(\mathbf{N}(t), \tau)) = \prod_{i=1}^{n} \left(\prod_{j=1}^{N_i(t_i)} \alpha_i f(\tau_{ij}; \boldsymbol{\beta}) \right) e^{-\alpha_i F(t_i; \boldsymbol{\beta})}$$

and $\mathbf{N}(t) = (N_1(t), \ldots, N_n(t))$.
Then for each process i the $\alpha_i|(\mathbf{N}(t), \tau)$ has a distribution defined by:

$$\pi(\alpha_i|(\mathbf{N}(t), \tau); \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\alpha_i^{N_i(t_i)} \exp\Big(- \alpha_i(\gamma_1 + F(t_i; \beta)) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big)}{\int_{\alpha_i} \alpha_i^{N_i(t_i)} \exp\Big(- \alpha_i(\gamma_1 + F(t_i; \beta)) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big) d\alpha_i}$$

where $F(t; \beta) = \int_0^t f(u; \beta) du$.

Hence, using this conditional density, the density function for $N_i(t_i, T)|N_i(t_i; \boldsymbol{\gamma}, \boldsymbol{\beta})$ is given by

$$\begin{aligned} P[N_i(t_i, T) = n|N_i(t_i); \boldsymbol{\gamma}, \boldsymbol{\beta}] &= \frac{(F(T; \boldsymbol{\beta}) - F(t_i; \boldsymbol{\beta}))^n}{n! \int_{\alpha_i} \alpha_i^{N_i(t_i)} \exp\Big(- \alpha_i(\gamma_1 + F(t_i; \boldsymbol{\beta})) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big) d\alpha_i} \\ \times \int_{\alpha_i} \alpha_i^{(N_i(t_i)+n)} \exp\Big(- \alpha_i(\gamma_1 + F(T; \boldsymbol{\beta})) - \sum_{j=2}^{k} \gamma_j \alpha_i^j \Big) d\alpha_i. \end{aligned} \quad (3.6)$$

Note that the occurrence times do not appear in this distribution; only knowledge of $N_i(t_i)$ is needed to determine this conditional distribution. However, the occurrence times will enter into the estimation of the model parameters β.

3.2.5 Discrepancy measure

In order to compare the adequacy of the point prediction method for $\mathbf{N}_i(t_i, T)$ obtained from our model, we calculate the prediction error between the real value of $\mathbf{N}_i(t_i, T)$ and its predictor $\hat{\mathbf{N}}_i(t_i, T)$.

For this, we used a discrepancy measure. Discrepancy is a general term usually measuring differences between an empirical value and its expected one. It is used in many different types of applications (Chen et al., 2008). Our discrepancy measure is defined as follow:

$$D = \sqrt{\frac{\sum_{i=1}^{n} \left(N_i(t_i, T) - \hat{N}_i(t_i, T) \right)^2}{n}}. \tag{3.7}$$

where the point predictor $\hat{N}_i(t_i, T)$ is defined by $\hat{N}_i(t_i, T) = \mathbb{E}[N_i(t_i, T)|\mathbf{N}(t_1); \boldsymbol{\gamma}, \boldsymbol{\beta}] = (F(T; \boldsymbol{\beta}) - F(t_i; \boldsymbol{\beta}))\mathbb{E}[\alpha_i|\mathbf{N}(t_1); \boldsymbol{\gamma}, \boldsymbol{\beta}]$ with $\mathbb{E}[\alpha_i|\mathbf{N}(t); \boldsymbol{\gamma}, \boldsymbol{\beta}]$ is the posterior mean of $\boldsymbol{\alpha_i}|(\mathbf{N}(t); \boldsymbol{\gamma}, \boldsymbol{\beta})$ given by (3.5) and where all the unknown parameters are replaced by their estimations (see Section 2.6).

As we saw in our studies (Khribi et al., 2015 and Khribi et al., 2016), the posterior distributions (3.5) will not have a known closed form, but it is a rather complicated high dimensional density, which makes direct inference almost impossible. For this reason, we can generate from this posterior distribution a large number of samples using Markov chain Monte Carlo (MCMC) implemented in WinBUGS (Spiegelhalter et al., 2003), and from these samples, we can obtain appropriate parameters estimate like the posterior mean of $\boldsymbol{\alpha}|(\mathbf{N}(t_1); \boldsymbol{\gamma}, \boldsymbol{\beta})$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are estimated by the method described in the next section.

3.2.6 Estimating unknown Poisson-Maximum Entropy parameters

In this section, we will discuss ways to estimate the vector of the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the general Poisson-MaxEnt model (3.4). These estimates will then substitute for the real parameters in the point prediction and prediction intervals previously mentioned. In the study (Khribi et al., 2015) where the predictive model for the prediction of recurrent events uses homogeneous Poisson processes, the parameters of the maximum entropy prior distribution were estimated by two methods, the usual maximum entropy estimation method which uses matching moments (MM) and the maximum likelihood method, referred to as the Pseudo-MaxEnt. Because we have seen in Khribi et al., (2016) that the MLE-MaxEnt method is computationally less complex than MM method when $k > 2$, we use in this study the MLE-MaxEnt method to estimate the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

The MLE-Maximum Entropy Method for the Poisson-MaxEnt Model

For the empirical Bayes MaxEnt model (3.4), we introduce the MLE-maximum entropy (MLE-MaxEnt) method using MLE for estimating the vector of the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. We start by construct the marginal likelihood L of the empirical Bayes general Poisson-Maximum Entropy model (3.4)

$$
\begin{aligned}
L(\boldsymbol{\gamma},\boldsymbol{\beta}|(\mathbf{N}(t),\boldsymbol{\tau})) &= \int_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha},\boldsymbol{\beta}|(\mathbf{N}(t),\boldsymbol{\tau}))\pi(\boldsymbol{\alpha};\boldsymbol{\gamma})d\boldsymbol{\alpha} \\
&= \int_{\boldsymbol{\alpha}} \left[\prod_{i=1}^{n}\left(\prod_{j=1}^{N_i(t_i)} \alpha_i f(\tau_{ij};\boldsymbol{\beta})\right)e^{-\alpha_i F(t_i;\boldsymbol{\beta})}\right]\left(\prod_{i=1}^{n}\frac{e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\right)d\boldsymbol{\alpha} \\
&= \prod_{i=1}^{n}\left[\frac{\left(\prod_{j=1}^{N_i(t_i)} f(\tau_{ij};\boldsymbol{\beta})\right)}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\int_{\boldsymbol{\alpha_i}} \alpha_i^{N_i(t_i)} e^{\left(-\alpha_i(\gamma_1+F(t_i;\boldsymbol{\beta}))-\sum_{j=2}^{k}\gamma_j\alpha_i^j\right)}d\alpha_i\right] \\
&= \prod_{i=1}^{n}\left[\frac{\left(\prod_{j=1}^{N_i(t_i)} f(\tau_{ij};\boldsymbol{\beta})I_i(N_i(t_i))\right)}{\int_{\alpha_i} e^{(-\sum_{j=1}^{k}\gamma_j\alpha_i^j)}d\alpha_i}\right] \qquad (3.8)
\end{aligned}
$$

with

$$I_i(N_i(t_i)) = \int_{\boldsymbol{\alpha_i}} \alpha_i^{N_i(t_i)} e^{\left(-\alpha_i(\gamma_1 + F(t_i;\boldsymbol{\beta})) - \sum_{j=2}^{k} \gamma_j \alpha_i^j\right)} d\alpha_i. \tag{3.9}$$

The log-likelihood is given by

$$l(\boldsymbol{\gamma}, \boldsymbol{\beta} | (\mathbf{N}(t), \boldsymbol{\tau})) = \sum_{i=1}^{n} \left[\log\left(\int_{\alpha_i} e^{(-\sum_{j=1}^{k} \gamma_j \alpha_i^j)} d\alpha_i \right) + \sum_{j=1}^{N_i(t_i)} \log\left(f(\tau_{ij}; \boldsymbol{\beta}) \right) + \log\left(I_i(N_i(t_i)) \right) \right] \tag{3.10}$$

Using Lebesgue's Dominated Convergence Theorem, (Talvila, 2001) gave necessary and sufficient conditions to interchange the order of differentiation and integration for (3.10) which are verified here. We can find the estimate of the vector of the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ by solving the score equations,

$$\frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta} | (\mathbf{N}(t), \boldsymbol{\tau}))}{\partial \gamma_j} = \sum_{i=1}^{n} \left[\frac{\int_{\alpha_i} \alpha_i^j e^{-\sum_{j=1}^{k} \gamma_j \alpha_i^j} d\alpha_i}{\int_{\alpha_i} e^{-\sum_{j=1}^{k} \gamma_j \alpha_i^j} d\alpha_i} - \frac{I_i(N_i(t_i) + j)}{I_i(N_i(t_i))} \right] = 0,$$

$$\frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\beta} | (\mathbf{N}(t), \boldsymbol{\tau}))}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{\frac{\partial f(\tau_{ij}; \boldsymbol{\beta})}{\partial \beta}}{f(\tau_{ij}; \boldsymbol{\beta})} - \frac{\partial F(t_i; \boldsymbol{\beta})}{\partial \beta} \frac{I_i(N_i(t_i) + 1)}{I_i(N_i(t_i))} \right] = 0,$$

The analytic solutions to these score equations are impossible to obtain; we thus use a numerical method to estimate directly the vector of the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ that maximize the log-likelihood (3.10). As we did in Khribi et al. (2016), we have chosen MATLAB "fminsearchbnd", a nonlinear optimization method which is derivative-free and allows bounds on the variables.

3.2.7 Plug-in Prediction Intervals

A prediction interval for $N_i(t_i, T)$ is an interval $[L(N(t), \tau(t)), U(N(t), \tau(t))]$ such that

$$P[L(N(t), \tau(t)) \leq N_i(t_i, T) \leq U(N(t), \tau(t)); \gamma, \beta] = 1 - \zeta.$$

Such an interval is called an exact $1-\zeta$ prediction interval for $N_i(t_i, T)$. In most settings (including the one considered in this paper), one cannot find exact prediction intervals when the parameters $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are unknown. This is analogous to the non-existence of exact confidence intervals for parameters in most statistical models. The alternative is to find an interval with an approximate coverage probability of $1-\zeta$. This can be accomplished in one way by finding an interval $[L, U]$ such that

$$P[L \leq N_i(t_i, T) \leq U; \hat{\gamma}, \hat{\beta}] = 1 - \zeta, \tag{3.11}$$

where only $N_i(t_i, T)$ is treated as a random variable, and where $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ are the MLE estimates obtained from the likelihood function based on the observed data and defined by (3.8).

The interval (3.11) is called a "plug-in" $1-\zeta$ prediction interval. Essentially, our method assumes that (3.6) is the true distribution and that the true parameter values are in fact $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ and thus ignores completely the uncertainty in $(\hat{\gamma}, \hat{\beta})$ relative to $(\boldsymbol{\gamma}, \boldsymbol{\beta})$. When the observed data set is very large, so that $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ can be assumed close to $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, then the coverage probability of this interval will be close to 1-ζ. However, in the case were the observed data set is not very large, our method can be improved by "calibrating" the plug-in intervals as was done by Fredette and Lawless (2007). We note that the calibration procedure still provides an approximate coverage probability for the prediction interval (3.11).

Plug-in prediction intervals with an approximate coverage probability of $1-\zeta$ can easily be obtained from the $\zeta/2$ and the $1-\zeta/2$ quantiles based on the predictive probability function $P[N_i(t_i, T) = n | N_i(t_i); \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}]$ given by (3.6).

3.3 Predicting the number of flights taken by frequent flyers

The context of this research is frequent flyer status within a specific airline loyalty program. Frequent flyer programs involve the systematic collection of detailed information regarding members' flying activities, thus allowing prediction of individual activity level based on the data already observed. The database at hand was obtained from the loyalty program of a major American commercial airline. It includes information on individual top-tier frequent

flyers for a period of 3 years starting January 1st, 2004 and provides, for each frequent flyer, a unique identifier along with the various dates that flights have been flown. To qualify for top-tier "Gold" membership, each frequent flyer had to fly at least 20 times over the first calendar year—that is, between January 1st, 2004 and December 31st, 2004 inclusively.

The quantities we wish to predict are $N_i(366, 731)$ for each frequent flyer i—that is, the number of flights taken by each frequent flyer between the first and last day of the second calendar year. The dataset contains such data for 5,000 frequent flyers. In the second year, each of them had actually flown between 0 and 158 flights. Table 3.1 gives the distribution of total number of flights in year 2 for those who had qualified for top-tier membership at the end of year 1.

Number of flights	Total
0	94
1∼10	721
11∼19	1112
20+	3073

Table 3.1 Distribution of the number of flights taken over year 2 by frequent flyers who had qualified for top-tier status by the end of year 1.

During a given year, the managers want to estimate, for the new year in progress, the eventual number of flights to be flown by each frequent flyer according to past data. Here we show how the methods in Section 2 can be used to predict the number of flights to be flown by each member, or the total flights to be flown for a group of, or all, frequent flyers.

Figure **??** provides a plot of the number of flights each day for the first 3 years considered, for those with top-tier frequent flyer membership after the first year. This graph clearly shows the seasonal (that is, nonhomogeneous) character of the flying habits of top-tier frequent flyers. It also shows that a number of flights flown daily diminishes with time as members from this top-tier cohort leave the program and/or the company, or diminish flying habits.

We now propose to use model (3.4) to predict the total number of flights by a given top-tier frequent flyer over a calendar year. Fredette and Lawless (2007) proposed a similar prediction model for forecasting automobile warranty claims, but instead of using a higher order maximum entropy prior for the random effects he uses the gamma prior and Laguerre

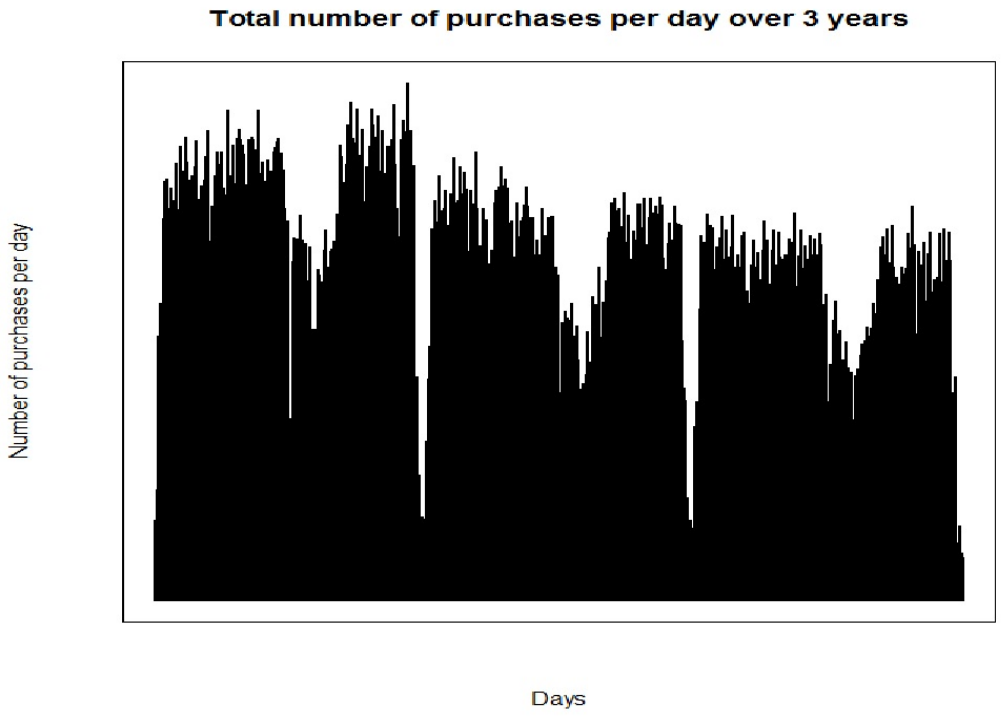

Figure 3.1 Total number of purchases per day over 3 years

polynomials function for $f(t;\beta)$. The choice of a suitable parametric form for $f(t;\beta)$ in (3.4) is crucial, because our predictions necessarily involve extrapolation into the future. Ideally, the shape of this function would be the same every year to reflect the periodicity of flying habits of frequent flyers. In addition, we would like to allow for a potential reduction of the amplitude of this function to reflect the fact that the number of flights usually diminishes over time.

We thus consider the function

$$f(t;\beta) = p(t-366;\beta_1,\beta_2) \times \exp\{C(d_t;\beta_3,\ldots,\beta_{K+3})\},$$

where:

- d_t is the number of the day of the year. For example, $d_1 = d_{366+1} = d_{366+365+1} = 1$ (the first year was a leap year). This will allow the function to retain the same shape year after year.

- At the beginning of the second year, we incorporate a decreasing proportion $p(.;\beta_1,\beta_2)$ to reflect the fact that some customers are likely to leave the program over time. Because of the obvious relationship between this phenomenon and a survival problem, we opted for a survival function $p(t-366,\beta_1,\beta_2) = S(t;\beta_1,\beta_2)$ such that $S(0;\beta_1,\beta_2) = 1$ and decreases thereafter. We used the Weibull survival function $S(t;\beta_1,\beta_2) = \exp\{(-t\beta_1)^{\beta_2}\}$ which is probably, along with the log-normal survival function, the most popular distribution for survival problems.

- $C(t;\beta)$ is a cubic spline. Cubic splines are continuous piecewise cubic polynomials used in curve fitting. They have been found to have nice properties with good ability to fit sharply curving shapes (Harrel, 2001). In order to use a cubic spline, we first have to determine an appropriate number of knots. Between each of these knots, the continuous function $C(t;\beta)$ is a cubic polynomial. Based on the data available after the first year, we found out that it was sufficient here to use $K = 4$ knots. In order to have approximately the same number of recurrent events between consecutive knots, the knots are the 20%, 40%, 60%, and 80% quantiles of all the occurrence times observed that 1$^{\text{st}}$ year (i.e., 70, 140, 220, and 300 days). The explicit form of this piecewise cubic polynomial is given by:

$$\begin{aligned} C(t;\beta_3,\ldots,\beta_9) =& \beta_3 t + \beta_4 t^2 + \beta_5 t^3 + \beta_6 (t-70)_+^3 + \\ & \beta_7 (t-140)_+^3 + \beta_8 (t-220)_+^3 + \beta_9 (t-300)_+^3 \end{aligned}$$

  where $(.)_+$ is the positive part of what is inside the parenthesis.

As Figure **??** shows, the use of splines in this case does allow for our model to follow rather well the bimodal distribution of flying behaviour among the top-tier frequent flyers over the first year of data, used to estimate our model.

Our approach improves in three ways the one proposed by Fredette and Lawless (2007) for the problem at hand.

- The use of the higher moment maximum entropy prior instead of the gamma prior as a random effects when we have a possible heterogeneity amongst the individual units

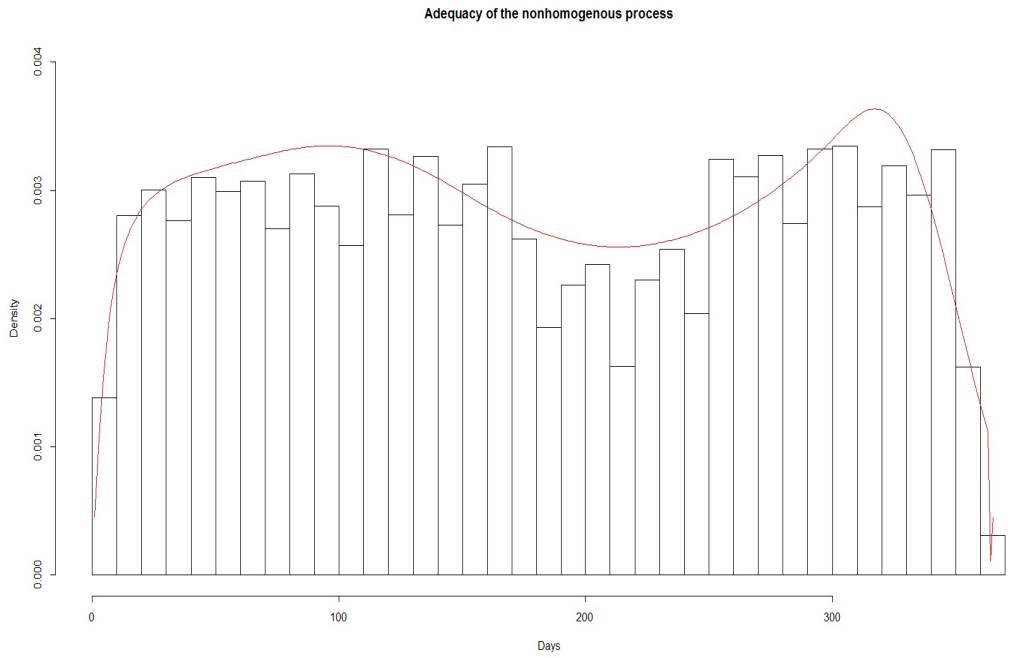

Figure 3.2 Adequacy of the nonhomogeneous process

in accordance with our two previous articles (Khribi et al., 2015) and (Khribi et al., 2016).

- The use of spline function instead of Laguerre polynomials for the non-negative function $f(t;\beta)$ which represents the shape of the rate function. It is important to use a function that is very flexible and that would be decreasing quickly as we approach the end of the year and spline function have nice properties with good ability to fit sharply curving shapes.
- The use of the Weibull survival function to reflect the fact that some customers are likely to leave the program over time.

3.3.1 Empirical Tests of the Prediction Model Proposed

In this section, we apply the general Poisson-MaxEnt model using higher moment maximum entropy prior and compared its adequacy to the NB model using the gamma prior proposed by Fredette and Lawless (2007). For this, we explore the performance of our approach by

predicting the number of flights at the end of each month in year 2 to be flown by 5,000 members of such a cohort of top-tier frequent flyers within this loyalty program.

Likelihood Ratio Tests

As suggested in Khribi et al., (2016), the likelihood ratio test (LRT) to determine the value of k in (3.4). Let say we want to compare the 2-moment maximum entropy and the 4-moment maximum entropy priors on the 4-moment and 6-moment maximum entropy priors, then the test statistic is the ratio between the log-likelihood of the null model to the alternative model:

$$\Gamma \quad = \quad -2\log\Big(\frac{l(\boldsymbol{\alpha}_1|N(t))}{l(\boldsymbol{\alpha}_2|N(t))}\Big) \tag{3.12}$$

where $l(\boldsymbol{\alpha}_1|N(t))$ and $l(\boldsymbol{\alpha}_2|N(t))$ are the log-likelihood of the null and alternative models respectively. This is a statistical test for nested models which reject the null hypothesis with a given significance level based on the chi-squared distribution. Through successive testing using the likelihood ratio test (Wilks, 1938), we can determine the number of moments necessary for the k-moment prior in the general Poisson-MaxEnt model.

Table 3.2 present the likelihood ratio test (LRT) results where the last two columns indicate respectively the p-values using 2 and the 4-moment maximum entropy prior model as the null models versus the alternative models with 4 and 6 moments. Note that the last column shows us the number of moments required for our predictive model.

Based on the results in Table 3.2 with a significance level equal to 5%, we can say that the LRT always rejects the model (4) with 6 moments compared with the one with 4 moments. However, it always supports the model (3.4) with 4 moments against the model with 2 moments. This means that the LRT always recommends the use of 4 moments at the end of each month in year 2.

Table 3.3 presents the discrepancy between the real value of $\mathbf{N}_i(t_i, T)$ and its predictor $\hat{N}_i(t_i, T)$ defined by (3.7), where $T = 731$ (the end of 2), using the different values of $t_i = (366+31, 366+31+28, ..., 731)$ where t_i is the number of days at the end of each month

t_i (in days)	p-value of LRT (MaxEnt_2Moments_vs_4Moments)	p-value of LRT (MaxEnt_4Moments _vs_6Moments)	Number of Moments Suggested
397 (13 months)	**< 0.01%**	75.82%	**4**
425 (14 months)	**< 0.01%**	82.59%	**4**
456 (15 months)	**< 0.01%**	89.55%	**4**
486 (16 months)	**< 0.01%**	92.41%	**4**
517 (17 months)	**< 0.01%**	98.74%	**4**
547 (18 months)	**< 0.01%**	99.23%	**4**
578 (19 months)	**< 0.01%**	100%	**4**
609 (20 months)	**0.14%**	100%	**4**
639 (21 months)	**0.75%**	100%	**4**
670 (22 months)	**0.95%**	100%	**4**
700 (23 months)	**2.24%**	100%	**4**
731 (24 months)	**3.87%**	100%	**4**

Table 3.2 The likelihood ratio test (p-value=.05) using data from the loyalty program.

i in year 2. For example, a value of **11.03** in the first line of Table 3.3 means that a prediction for the end of the first month in year 2 based on this model (the general Poisson-MaxEnt model with the 4-moment prior) would be on average **11.03** flights from the real value of $\mathbf{N}_i(t_i, T)$. The likelihood ratio test stopping rule, that is, to stop at 4 moments result is confirmed in Table 3.3, where the average discrepancy values for the general Poisson-MaxEnt model with the 4-moment maximum entropy prior (values in bold font) are always very close to the smallest absolute error discrepancy given by the model using the 6-moment maximum entropy prior.

t_i (in days)	Gamma (Fredette (2007)) $(\widehat{a}_{mle}, \widehat{b}_{mle})$	*MLE 4Moments*	*MLE 6Moments*
397 (13 months)	15.72	**11.03**	10.99
425 (14 months)	13.93	**9.94**	9.92
456 (15 months)	12.37	**8.78**	8.76
486 (16 months)	11.91	**8.03**	8.02
517 (17 months)	9.28	**7.89**	7.87
547 (18 months)	8.57	**6.77**	6.75
578 (19 months)	7.62	**5.49**	5.49
609 (20 months)	6.05	**4.94**	4.92
639 (21 months)	4.43	**4.05**	4.03
670 (22 months)	3.31	**3.00**	3.00
700 (23 months)	1.82	**1.70**	1.70
731 (24 months)	0.00	**0.00**	0.00

Table 3.3 Discrepancy of point predictors with different values of t_i using data from the loyalty program.

As a another example of the usefulness of this approach, let us consider a scenario in which, as she prepares her marketing activities for the fall season, a customer relationship manager of this loyalty program is concerned about deploying extra effort to retain those Gold customers that are in danger of not qualifying for Gold status the next year. In order to target the right customers with a costly special offer, this manager wishes to target those with a moderate chance of actually qualifying for top-tier membership as assessed using data available on August 1st of 2004. For each of these "gold" customers, our model returns the probability that these customers will remain top-tier members in 2006—that is, the probability that they will fly 20 flights or more during 2005. Of course, those frequent flyers having already flown these 20 flights have a probability of 100%. Table 3.4 shows these probabilities for 11 segments according to how likely they are to remain "gold" customers for both predictive models: the model (3.4) using 4-moment maximum entropy prior and the NB model using the gamma prior proposed by Fredette and Lawless (2007). From this table, we notice that the values of the probability of being Gold for customers define by the general Poisson-MaxEnt predictive model with the 4-moment maximum entropy prior (values in bold font) are always closest to the actual proportion of customers who retained top-tier membership by Dec. 31st, 2005. Hence, our predictive model with the 4-moment maximum entropy prior performs better when we compare it to the NB model using the gamma prior where the parameters were estimated using the MLE method.

Probability intervals for customers already qualified	Probability of being Gold for customers with gamma prior	Probability of being Gold for customers with 4-moment prior	Actual proportion of customers who retained top-tier membership by Dec. 31st, 2005
[0-10%[	1.63%	**2.89**%	3.31%
[10-20%[	13.17%	**15.87**%	16.40%
[20-30%]	19.89%	**28.03**%	29.61%
[30-40%[	30.48%	**33.79**%	33.33%
[40-50%[	41.03%	**46.93**%	48.51%
[50-60%[	56.82%	**52.02**%	50.00%
[60-70%[	67.36%	**60.51**%	56.00%
[70-80%[	77.34%	**71.29**%	69.39%
[80-90%[	86.29%	**76.47**%	71.71%
[90-100%[	99.05%	**93.33**%	91.76%
[100%]	100%	**100**%	100%

Table 3.4 Models Fit According to Likelihood of Retaining Top-Tier Frequent Flyer Status.

To further assess the predictive performance of our model compared to the NB model using the gamma prior proposed by Fredette and Lawless (2007), we use the data available up to August 1st, 2005 to extrapolate the rate function of our nonhomogeneous Poisson process between August 1st, 2005 and December 31th, 2005. As Figure 3.3 shows, our model allows rather precise prediction past August 1st. This analysis demonstrates the high degree of validity of using our nonhomogeneous mixed Poisson model for the purposes of forecasting a customer's future purchasing, conditional on his past buying behaviour and his activity to date.

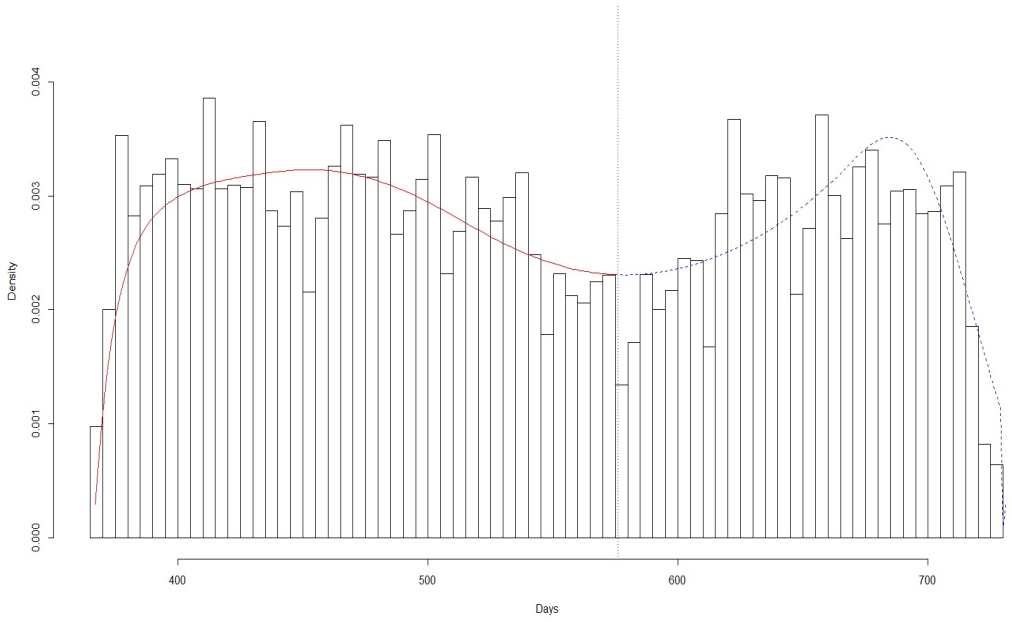

Figure 3.3 Accuracy of the forecasting based on the data available on August 1st ($t = 578$)

Finally, we present a last example of the usefulness of this approach for various scenarios. We can imagine our customer retention manager is interested in predicting the likelihood of remaining top-tier customers at the beginning of each month. Let us consider the example of two Gold customers who both flew 26 flights over the first year. They both have an 86.2% likelihood of remaining Gold customers at the beginning of the second year. Ultimately, Customer A will fly 23 qualifying flights this year thus conserving his top-tier status, whereas Customer B will fly only 19, meaning he will loose his top-tier status at the end of the year.

Figure 3.4 and Figure 3.5 provide the 12 monthly 95% prediction intervals for Customers A and B. The dotted lines on both graphs indicate the total numbers of flights actually flown by the end of the year while the increasing solid curve represents the total number of flights taken at that point in time. As can be seen, and as an additional demonstration of the predictive ability of our model, the forecasted intervals always contain the actual, final number of flights taken for each of those two customers. Of course, the prediction interval also becomes smaller with time, as data accrue regarding both customers' actual behaviour.

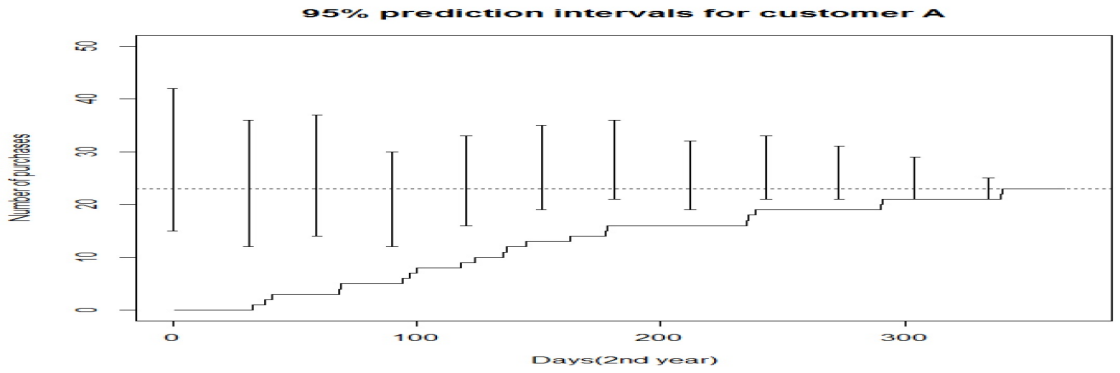

Figure 3.4 95% prediction intervals for customers A

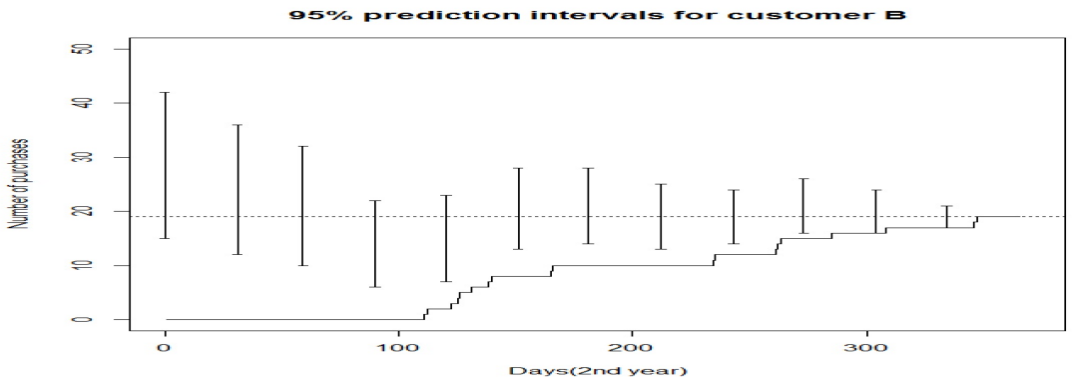

Figure 3.5 95% prediction intervals for customer B

On the basis of their respective flying activities, our model allows us to estimate at any point in time the probability that each of these two customers will take at least 20 flights. For instance, a monthly review would provide the probabilities of taking at least 20 flights before the end of the year for each member (see Table 3.5).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Customer A (23 flights actually taken at the end of the year)	.862	.698	0.798	0.520	0.819	0.963	0.995	0.975	0.999	0.996	1.000	1.000
Customer B (19 flights actually taken at the end of the year)	.862	.698	.485	.074	.097	.502	.562	.308	.332	.630	.428	.122

Table 3.5 Probabilities of taking 20 flights or more during this year, assessed at the beginning of each month based on historical data to date.

As can be seen in Table 3.5, while the probability of Customer A retaining his top-tier status by the end of the year remains high—above 69.8%—throughout the year except for one month, Customer B can be identified as potentially losing his top-tier status as early as April. Considering that only 1 flight actually made the difference in the end, the airline company could have used such approaches as reminding Customer B of the value of his Gold membership as an incentive to fly more in order to retain top-tier benefits into the next year. Adopting such "corrective" actions early on during the year would have likely left enough time for Customer B to better plan his flying activities for the remainder of the year.

According to the results of the different empirical tests applied to our prediction model for the data setting from a major airline company, we can say that the general Poisson-MaxEnt model using 4-moment maximum entropy prior and a spline function for the non-negative function $f(t;\beta)$ with a Weibull survival function $S(t;\beta_1,\beta_2)$ clearly outperformed the NB model proposed by Fredette and Lawless (2007) as a predictive model.

3.4 Summary and Discussion

In this study, we have proposed a predictive model allowing prediction of recurrent events using flexible nonhomogeneous Poisson processes with higher moment maximum entropy priors to model possible heterogeneity amongst the individual units modeled. The motivation for our work lies in the prediction of individual activity level using the data already observed or

other events that occur for individual units or subjects in a population. The database at hand was obtained from the loyalty program of a major commercial airline where the behaviour is observed and stored for each customer. The efficiency of our predictive model is compared to the NB model using the conjugate gamma prior. Also, we have seen throughout this paper that an accurate prediction depends on choosing a satisfactory model for $f(t;\beta))$ in (3.4) representing the shape of event rate functions for individual units or processes. Using the spline functions for $f(t;\beta))$ also makes our approach especially well suited for situations with irregular purchase behaviour, such as seasonal or cyclical products or services. It has been shown that for the database at hand the use of 4-moment maximum entropy prior provides us a realistic prediction model than the one given by the NB model proposed by Fredette and Lawless (2007).

Finally, though some detailed development remains to be done, our predictive model considered here can be extended to others situation where there are costs or other values associated with events and we may wish to predict future costs.

3.5 References

• Amin, A.; Grunske, L. and Colman, A. (2013). An Approach to Software Reliability Prediction Based on Time Series Modeling, *The Journal of Systems and Software*, **86**, 1923–1932.

• Chen, F.; Curran, P.; Bollen, K.A.; Kirby, J.and Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models, *Sociological Methods & Research*, **36**, No.**4**, 462–494.

• England, P. and Verrall, R. (2002). Stochastic Claims Reserving in General Insurance, *British Actuarial Journal,* **8**, 443–518.

• Fredette, M. and Lawless, J.F. (2007). Finite horizon prediction of recurrent events with application to forecast of warranty claims, *Technometrics,* **49**, 66–80.

• Grandell, J. (1997), *Mixed Poisson Processes*, London: Chapman & Hall.

• Harrel, F.E. (2001), *Regression Modeling Strategies*, New York: Springer.

• Khribi, L., Fredette, M. and MacGibbon, B. (2015). The Poisson maximum entropy model for homogeneous Poisson processes, *Communications in Statistics - Simulation and Computation,* **45**, No.**09**, 3435–3456.

• Khribi, L., Fredette, M. and MacGibbon, B. (2016). Choosing between higher moment maximum entropy models and its application to homogeneous point processes, *Communications in Statistics - Simulation and Computation,* to be submitted.

• Lawless, J.F. (1987). Regression Methods for Poisson Process Data, *Journal of the American Statistical Association*, **82**, 808–815.

• Lawless, J.F. and Fredette, Marc (2005). Frequentist prediction intervals and predictive distributions, *Biometrika*, **92**, 529–542.

• Mohammad-Djafari, A.(1992). Maximum likelihood estimation of the Lagrange parameters of the maximum entropy distributions, Appeared in *Maximum Entropy and Bayesian Methods,* Series *Fundamental Theories of Physics*, **50**, 131–139.

• Shannon, C.E.(1948). The mathematical theory of communication. *Bell System Technical Journal,* **27**, 379–423.

• Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual. Version 1.4* (http://www.mrc-bsu.cam.ac.uk/bugs.). Technical Report, Medical Research Council Biostatistics Unit. Cambridge.

• Talvila, E. (2001). Necessary and sufficient conditions for differentiating under the integral sign, *American Mathematical Monthly,* **108**, 544–548.

• Weinstock, R. (1952). *Calculus of Variations - With Applications to Physics and Engineering*, New York, McGraw-Hill.

• Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics*, **9**, No.**1**, 60–62.

CONCLUSION

Dans cette thèse, nous avons proposé de nouveaux modèles prédictif permettant la prédiction des événements récurrents en utilisant des processus homogène et non homogène de Poisson et une possible hétérogénéité entre les unités individuelles modélisées par des effets aléatoires. Nous avons proposé des modèles prédictifs utilisant des techniques bayésienne empiriques et le principe du maximum d'entropie afin de modéliser ces effets aléatoires avec des distributions a priori. Nous avons utilisé un premier modèle utilisant des processus homogène de Poisson, avec comme loi a priori la loi d'entropie maximum à deux moments qui correspond à la loi normale tronquée et nous avons montré qu'elle se compare favorablement au modèle classique de la négative binomiale qui est généralement utilisé pour ce genre de problème utilisant comme loi a priori la loi gamma.

En raison de la condition sur l'utilisation de la loi a priori d'entropie maximum à deux moments, dans notre premier modèle nous avons été contraint à considérer seulement les cas où le coefficient de variation était inférieur ou égal à 1. Ce qui engendre une certaine restriction à l'utilisation de ce premier modèle. Nous avons enlevé cette restriction par l'utilisation des lois d'entropie maximum avec un nombre de moments d'ordre plus élevé et nous l'appliquons dans la prédiction des événements récurrents tout en utilisant des processus de Poisson homogènes et non-homogènes.

Enfin, nous avons évalué la performance de nos modèles par des études de simulation approfondies et par quelques ensembles de données. En particulier, nous avons appliqué notre dernier modèle prédictif sur un ensemble de données réel provenant d'un programme de fidélisation. Nous avons trouvé que notre modèle se compare favorablement en terme d'adéquation au modèle classique de la binomiale négative qui utilise comme loi a priori la loi gamma.

Même si un certain développement beaucoup plus détaillé reste à faire dans nos modèles pré-

dictifs, ces modèles peuvent être étendus à d'autres situations où il y a des coûts ou d'autres valeurs associées à ces événements dans le but de prédire par exemple les coûts futurs.