Exchangeable claim sizes in a compound Poisson type process

Ramsés H. Mena[1] and Luis E. Nieto-Barajas[2]

ramses@sigma.iimas.unam.mx and lnieto@itam.mx

[1]*Department of Probability and Statistics, IIMAS–UNAM, 01000 Mexico, D.F.*

[2]*Department of Statistics, ITAM, Rio Hondo 1, Progreso Tizapan, 01080 Mexico, D.F.*

ABSTRACT. When dealing with risk models the typical assumption of independence among claim size distributions is not always satisfied. Here we consider the case when the claim sizes are exchangeable and study the implications when constructing aggregated claims through compound Poisson type processes. In particular, exchangeability is achieved through conditional independence, using parametric and nonparametric measures for the conditioning distribution. Bayes' Theorem is employed to ensure an arbitrary but fixed marginal distribution for the claim sizes. A full Bayesian analysis of the proposed model is illustrated with a panel type dataset coming from a Medical Expenditure Panel Survey (MEPS).

Key words: Bayesian nonparametric analysis, compound Poisson process, exchangeable claims process, exchangeable sequence, panel data, risk model.

1. Introduction

In insurance and risk modelling, the study of the dynamics through time of the wealth of a business has been object of research for several decades. In the basic setup of collective risk theory, the interest lies in the probability distribution of the reserve process and henceforth, using historical information and a robust model, in predicting potential bankruptcies. The cornerstone tool is based on the modeling of the aggregated claims amount through the compound Poisson process (CPP), here denoted by $\{X_t;\, t \geq 0\}$ and defined through

$$X_t = \sum_{i=1}^{N_t} Y_i, \tag{1}$$

where $\{N_t; t \geq 0\}$ is a homogeneous Poisson process with intensity $\lambda > 0$ and $Y_1, Y_2, \ldots$ a sequence of independent and identically distributed (nonnegative) random variables with a common distribution F, independent of $\{N_t\}$. In insurance applications, $\{N_t\}$ is interpreted as the number of claims performed to the company during the time interval $(0, t]$, and Y_i as the magnitude of the i-th claim. Hence X_t can be seen as the aggregated claims, on the time interval $(0, t]$.

Provided second order moments for the claim distribution exist, some basic characteristics of process (1) are at order

$$\begin{aligned} \mathrm{E}(X_t) &= \lambda t\, \mathrm{E}(Y_i) \\ \mathrm{Var}(X_t) &= \lambda t\, \mathrm{E}(Y_i^2) \\ \mathrm{Cov}(X_t, X_s) &= \mathrm{Var}(X_{t\wedge s}), \end{aligned} \tag{2}$$

where $t \wedge s := \min(s,t)$. Moreover, $\mathrm{Corr}(X_t, X_s) = (t \wedge s)/\sqrt{ts}$, so if we let $t < s$ then $\mathrm{Corr}(X_t, X_s) = \sqrt{t/s}$.

In the collective risk theory it is common to model the *risk process*, $\{Z_t;\, t \geq 0\}$, of a portfolio via the following compensation of a CPP

$$Z_t = r\,t - X_t, \tag{3}$$

where $r > 0$ denotes the constant *gross risk premium rate*. This model constitutes the classical risk model introduced by Lundberg [1] and further analyzed by Cramer [2] and many others. In the case of an insurance company, Z_t takes the interpretation of the profit of the portfolio over the time interval $(0,t]$.

One of the main interests regarding the wealth of a portfolio, with initial capital u and facing risk process (3), is concerning the probability that the *reserve process*, $R_t^u := u+Z_t$, falls below zero. This quantity is usually measured through what is known as the *ruin probability*, $\Psi(u)$, defined through

$$\Psi(u) := \mathrm{P}\left(\inf_{t\geq 0}\{t; R_t^u < 0\} < \infty\right) \tag{4}$$

or through its *finite-horizon* analog

$$\Psi(u,T) := \mathrm{P}\left(\inf_{0\leq t<T}\{t; R_t^u < 0\} < T\right). \tag{5}$$

These quantities are relevant when, in average, the portfolio is facing a profitable business. This situation is measured through the *relative safety loading* ρ, which represents the expected profit/loss relative to the total claim amount, that is

$$\rho := \frac{\mathrm{E}(Z_t)}{\mathrm{E}(X_t)}.$$

Hence an insurance company is willing to have a relatively small ruin probability $\Psi(u)$ when starting a business with $\rho > 0$, bearing in mind that the larger the safety loading the smaller the ruin probability becomes.

Thanks to a well-known renewal argument Feller [3], the non-ruin probability $\varphi(u) := 1 - \Psi(u)$ for the CPP process satisfies $\varphi(u) = E[\varphi(u - rT_1 - X_1)]$, where $T_1 \sim \mathrm{Exp}(\lambda)$ denotes the time of the first claim and $X_1 \sim F$. Hence, it can be proved that

$$\varphi(u) = \frac{\lambda}{r} e^{\lambda u/r} \int_u^\infty e^{-\lambda x/r} \int_0^x \varphi(x-z)F(dz)dx.$$

Apart of the case of exponential claim sizes, or mixture of exponentials, little is known about closed solutions to the above equation. The situation complicates even more for the finite-horizon case, however several approximations and estimators are available, we refer to Rolski et al. [4] for some of these methods.

The literature provides with various generalizations to the above simple compound Poisson process (1), most of them inherited by generalizations of the underlying Poisson process $\{N_t\}$. For example, $\{N_t\}$ could be replaced by a *mixed Poisson process* or a *Cox process* (see, for example, Grandell [5]). These approaches might result in processes with increments no longer being independent. For instance, it can be seen that a stationary mixed Poisson process $\{N_t\}$ (and then the resulting compound process X_t) have exchangeable increments Daboni [6]. Alternative generalizations are in the line of independent increments processes, as can be found in the work by Morales and Schoutens [7], who considered a Lévy process for modeling the aggregate claims X_t. Other generalizations, rather devoted to insurance applications, extend the above reserve process to have, for instance, a non-linear premium income. For these and many other extensions we refer to Rolski et al. [4].

A common scenario is when the insurer only considers the aggregated information of all policies composing a portfolio, e.g. the aggregated claim amounts of all the portfolio's policies together with pre-fixed cut-off times, regardless of whether such claims where performed by the same or different policy holders. In this situation the independence among claims assumption inherent in the CPP approach could, in principle, make some sense. However, as is typically the case in modern products offered by insurance companies, the "collectiveness" is more specialised to more homogeneous groups, and a claim generated by one policy holder may cause a future claim related to the same policy or even claims related to other policies within the same portfolio. This, for instance, is the situation encountered in some health care portfolios where a persistent disease could result in recurrent claims. Clearly this latter case could break with the independence among claims assumption.

Let us envisage a situation where we are able to identify claims made by the same individual j, say $Y_{1j}, Y_{2j}, \ldots$. In this case we would like to model the claiming patterns of individual j by acknowledging a possible dependence among claims related to the same policy. For this reason it is desirable to undertake a framework that includes a possible dependence among claims made by the same individual. Furthermore, if policy-wise information is available, it would also be desirable to consider such disaggregation for modelling purposes. To achieve this goal we will consider a modeling framework that is, in spirit, similar to a mixed model in the sense that a random effect accounts for the temporal correlation in the claims. Within this interpretation, such a random effect will play the role of a latent variable that produces an exchangeable sequence of claims within each individual. At the same time, we will introduce a methodology that allows us to maintain the same marginal distribution for all claims. The main difference of having policy-wise information available, is that we can consider several trajectories of the reserve process, namely individual trajectories, for estimation and modelling purposes. This has a similar statistical interpretation to that of panel studies and results in a robust methodology not only at the aggregate level but also at the average individual one. Hence under this framework we can consider individual (or policy)-wise realizations $X_{t1}, X_{t2}, \ldots, X_{tm}$ in order to build up the corresponding average individual reserve process as well as the aggregated one, say $X_t^a := \sum_{j=1}^m X_{tj}$.

The main objective of this work is to generalize the compound Poisson processes (1) by relaxing the independence assumption in the claims, but keeping the counting process N_t to be a homogeneous Poisson process. The idea of imposing dependence among claims has been considered before by other authors, for instance: Gerber [8] and Promislow [9] imposed a dependence structure through a linear ARMA model; Cossette and Marceau [10] modeled dependence through a Poisson shock model and studied its implications in the ruin probability, and Mikosch and Samorodnitsky [11] model the claims through a stationary stable process and also investigated the impact on ruin probabilities.

The approach we undertake is based on a dependence structure provided via exchangeability assumptions in the claims. It is worth noting that exchangeability has also been used to generalize risk models but in a different direction. Buhlmann [12], Buhlmann [13] and Daboni [6] generalized the counting process and used a mixed Poisson process specially designed to have exchangeable interarrival times. We rather focused on a particular construction that allows us to have dependence among claims while keeping the marginal distribution invariant. This is appealing since in applications, such as risk modelling, well-identified claim distributions are employed. In principle our approach could be also extended to CPP with non-homogeneous or state dependent Poisson process, however we have decided to keep it to the basic reserve process to emphazise the proposed contribution.

The outline of the paper is as follows. In Section 2 we define the exchangeable claims process and provide two ways of defining exchangeable sequences via a parametric and a nonparametric approach. Section 3 describes a methodology to implement Bayesian inference for the parameters in both the traditional compound Poisson process and the proposed approach. Finally in Section 4 we illustrate our model with a full Bayesian analysis of data from the Medical Expenditure Panel Survey (MEPS).

2. Exchangeable claims modeling

Definition 1. Let $\{N_t; t \geq 0\}$ be a simple Poisson process with intensity $\lambda > 0$ and $Y_1, Y_2, \ldots$ a sequence of exchangeable nonnegative random variables with common one-dimensional marginal distribution F, independent of $\{N_t\}$. Then the *exchangeable claims process* (ECP), $\{X_t; t \geq 0\}$, is defined as

$$X_t := \sum_{i=1}^{N_t} Y_i \tag{6}$$

Clearly this new process has no longer independent increments, however it is clearly a Markov process with exchangeable increments. Non random summations of the type defined by (6) have been previously used to generalize the central limit theorem for exchangeable sequences, see for example Klass and Teicher [14].

When analysing real data it is of interest to have an arbitrary but given claim distribution, therefore the issue is how to construct an exchangeable sequence with given marginal distributions. Clearly the easiest case is that of iid claims, however the point here is to allow the model to have a possible dependence structure among the claims.

Based on the representation theorem of de Finetti [15], we can construct a sequence of exchangeable random variables through a conditional independence sequence. More explicitly, the random variables $Y_1, Y_2, \ldots$ are an exchangeable sequence if there exists a latent variable/measure G such that $\{Y_i\}$ for $i = 1, 2, \ldots$ are conditionally independent given G, and G is a random variable/measure (known as the de Finetti's measure) with law described by a distribution/process $\mathcal{P}$. Note that exchangeable variables defined in this fashion are always positive correlated, which suffices for most relevant dependencies found in risk modeling.

Proposition 1. *Let $\{X_t;\, t \geq 0\}$ be an ECP with $Y_i \mid G \sim G$, $i = 1, 2, \ldots$ conditionally independent given G, and $G \sim \mathcal{P}$, then assuming the existence of second order marginal moments we have:*

(i) $E(X_t) = \lambda t E(Y_i)$

(ii) $Var(X_t) = \lambda t E(Y_i^2) + \lambda^2 t^2 Cov(Y_i, Y_j)$

(iii) $Cov(X_t, X_s) = \lambda\,(t \wedge s)\, E[Y_i^2] + \lambda^2\, ts\, Cov(Y_i, Y_j)$

Proof. The whole idea is based on the conditional independence properties.

(i)

$$\mathrm{E}(X_t) = \mathrm{E}\left\{\mathrm{E}\left(\sum_{i=1}^{N_t} Y_i\,|N_t\right)\right\} = \mathrm{E}\,(N_t)\ \mathrm{E}\,(Y_i) = \lambda\, t\, \mathrm{E}(Y_i),$$

where the second equality follows due to the independence between N_t and Y_i.

(ii) First, let us notice that

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}\,\{\mathrm{E}(Y_i \mid G), \mathrm{E}(Y_j \mid G)\} = \mathrm{Var}\{\mathrm{E}(Y_i \mid G)\},$$

hence

$$\begin{aligned}\mathrm{Var}(X_t) &= \mathrm{E}\{\mathrm{Var}(X_t \mid G)\} + \mathrm{Var}\{\mathrm{E}(X_t \mid G)\} \\ &= \lambda\, t\, \mathrm{E}(Y_i^2) + \lambda^2\, t^2\, \mathrm{Var}\{\mathrm{E}(Y_i \mid G)\} \\ &= \lambda\, t\, \mathrm{E}(Y_i^2) + \lambda^2\, t^2\, \mathrm{Cov}(Y_i, Y_j)\end{aligned}$$

(iii)

$$\begin{aligned}\mathrm{Cov}(X_t, X_s) &= \mathrm{Cov}\left\{\mathrm{E}\left(X_t \mid G\right), \mathrm{E}\left(X_s \mid G\right)\right\} + \mathrm{E}\left[\mathrm{Cov}(X_t, X_s \mid G)\right] \\ &= \mathrm{Cov}\left\{\lambda t\, \mathrm{E}(Y_i \mid G), \lambda s\, \mathrm{E}(Y_j \mid G)\right\} + \mathrm{E}\{\mathrm{Var}(X_{t\wedge s} \mid G)\} \\ &= \lambda^2\, ts\, \mathrm{Cov}(Y_i, Y_j) + \lambda\,(t \wedge s)\, \mathrm{E}[Y_i^2].\end{aligned}$$

If we define processes CPP and ECP with the same one-dimensional marginal claims distribution, then comparing their characteristics we note that both processes have the same expected value. However, the variance and covariance are not the same; the variance (covariance) of the ECP is larger than the variance (covariance) of the CPP.

In terms of the correlation, if we let $h(t) := \mathrm{Cov}(Y_i, Y_j)\lambda t + \mathrm{E}[Y_i^2]$ and if $t < s$ then $\mathrm{Corr}(X_t, X_s) = \sqrt{\frac{t}{s}}\frac{h(s)}{\sqrt{h(s)h(t)}}$, which means that the correlation in the ECP is also larger. This behavior is later illustrated in Examples 1 and 2.

It is worth mentioning that in some applications, where heavy tailed distributions are used for claim modeling, the covariance or correlation might not be a good indicator of dependence since typically when using such distributions second moments do not exist, hence one has to resort to other measures of dependence such as Kendall's Tau or Spearman's Rho.

Now we will consider two ways of defining exchangeable sequences of random variables with given marginal distributions, through a parametric and a nonparametric conditioning distributions.

2.1. Exchangeable sequences: Parametric method

The idea is to define an exchangeable sequence $Y_1, Y_2, \ldots$ in such a way that each Y_i will have the same marginal distribution $F(y)$. For the sake of exposition we assume the existence of a density $f(y)$. First, we introduce a latent variable Z with arbitrary conditional density $f(z \mid y)$. Then, we define $f(y \mid z)$ using Bayes' Theorem in the following way:

$$f(y \mid z) = \frac{f(z \mid y) f(y)}{f(z)} \tag{7}$$

with

$$f(z) = \int f(z \mid y) f(y) \mathrm{d}\mu_1(y), \tag{8}$$

where $\mu_1(y)$ represents a reference measure such as counting measure if Y is discrete or the Lebesgue measure if Y is continuous. It is straightforward to show that if we marginalize over Z then,

$$\int f(y \mid z) f(z) \mathrm{d}\mu_2(z) = f(y)$$

as required, where $\mu_2(\cdot)$ is another reference measure acting on Z. Then, if we take $Y_i \mid Z \sim f(y \mid z)$, as in (7), for $i = 1, 2, \ldots$ a sequence of conditional independent random variables given $Z = z$, with marginal distribution for Z, as in (8), then $Y_1, Y_2, \ldots$ is a sequence of exchangeable random variables with marginal densities $f(y)$. A similar idea was used by Pitt et al. [16] to construct stationary autoregressive models with given marginal distributions, although in their construction different Z_i's are used for different Y_i's.

Notice that the possibilities for Z are wide open, for example it could be discrete, continuous, univariate or multivariate. Hence different features of Z will lead to different forms of dependence. Closed form expressions can be obtained when considering conjugate pairs (y, z) in a Bayesian setting.

Example 1. Let us denote by $\mathrm{Ga}(a, b)$ a gamma distribution with mean a/b and by $\mathrm{Gga}(a, b, c)$ a gamma–gamma distribution with mean $cb/(a-1)$. We will define an exchangeable sequence with $\mathrm{Ga}(a, b)$ as marginal distribution by assuming $f(z \mid y) = \mathrm{Ga}(z \mid c, y)$, $c > 0$. In this case we obtain that, $f(y \mid z) = \mathrm{Ga}(y \mid a + c, b + z)$ and $f(z) = \mathrm{Gga}(z \mid a, b, c)$. Therefore, if we take $Y_i \mid Z \sim \mathrm{Ga}(a + c, b + z)$ for $i = 1, 2, \ldots$ conditionally independent given Z and $Z \sim \mathrm{Gga}(a, b, c)$ hence marginally $Y_i \sim \mathrm{Ga}(a, b)$ with $\mathrm{Corr}(Y_i, Y_j) = c/\{(a+c+1)\}$, for all $i \neq j$. Now, since $N_t \sim \mathrm{Po}(\lambda t)$ is the number of claims in $(0, t]$ and $Y_1, Y_2, \ldots$ the claim sizes such that $Y_i \sim \mathrm{Ga}(a, b)$, then $X_t = \sum_{i=1}^{N_t} Y_i$ is the total claims amount in $(0, t]$. Consider two processes, the one with independent claims (X_t^I) and the one with exchangeable claims (X_t^E). Using the moments in (2) and Proposition 1, it follows that the expected value of the total claims amount is,

$$\mathrm{E}\left\{X_t^I\right\} = \mathrm{E}\left\{X_t^E\right\} = \frac{a}{b}\lambda t.$$

The variance of the processes become,

$$\mathrm{Var}\left\{X_t^I\right\} = \frac{a(a+1)}{b^2}\lambda t$$

and

$$\mathrm{Var}\left\{X_t^E\right\} = \mathrm{Var}\left\{X_t^I\right\} + \frac{ac}{b^2(a+c+1)}\lambda^2 t^2,$$

which clearly shows that X_t^E is overdispersed with respect to X_t^I. Finally, the covariance function for $t < s$ takes the form,

$$\mathrm{Cov}\left\{X_t^I, X_s^I\right\} = \mathrm{Var}\left\{X_t^I\right\}$$

and

$$\mathrm{Cov}\left\{X_t^E, X_s^E\right\} = \mathrm{Cov}\left\{X_t^I, X_s^I\right\} + \frac{ac}{b^2(a+c+1)}\lambda^2 ts.$$

In order to have an idea of the implications of using a ECP instead of a CPP in the ruin probability, we performed a simulation study. We considered 5,000 iid realizations of the reserve processes with the above specifications for both the CPP and the ECP, and computed the Monte Carlo estimates of the corresponding finite-horizon ruin probabilities. A finite-horizon realization of reserve process, e.g. on $[0, T]$, was simulated via

$$R_t^u = u + rt - \sum_{i=1}^{N} I(U_i < t)\, Y_i,$$

where $N \sim \mathrm{Po}(\lambda T)$ gives the total number of claims on the interval $[0, T]$, $\{U_i\}_{i=1}^N$ are iid from a uniform distribution on $[0, T]$ (denoting the jump times) and $\{Y_i\}_{i=1}^N$ are iid or exchangeable from the claim distribution for the CPP or ECP respectively. We then approximate the ruin probability as the relative frequencies of those realizations falling below zero. For this experiment the parameters specifications are $r = \lambda = a = c = 1$ and $b = 1.1$, e.g. with a relatively small (but positive) safety loading of $\rho = (r\, b)/(\lambda\, a) - 1 = 0.1$.

Figure 1 shows the behavior of the ruin probability as a function of the initial capital and also as a function of the finite-horizon, the latter for two choices of initial capital, $u = 4$ and $u = 10$. From the top-right graph is evident the existence of a crossing point around $u = 4$, i.e. a point from where ECP causes a larger (and with slower decay) ruin probabilities than those induced by the CPP. In order to further inspect on this issue, the bottom graphs show the behavior of the ruin probabilities as the finite-horizon changes starting from two different choices of initial capital: $u = 4$ (left), i.e. when there is a similar behavior of the ruin probability for both the CPP and the ECP, and $u = 10$ (right), i.e. for a choice of initial capital where the ECP induces a higher ruin probability than the CPP.

Clearly, there is a remarkable variation on the corresponding ruin probabilities. In particular, for this choice of $\mathrm{Ga}(a, b)$ marginal, and dependence induced through $f(z) = \mathrm{Gga}(a, b, c)$, we could conclude that for big initial capitals, the discrepancies between the ruin probabilities tend to be quite substantial, making a process with dependence more riskier. This observation makes sense since the larger the initial capital is, the longer we have to wait for a given trajectory to be ruin, and therefore, the more time we have for the underlying dependence to induce an important difference when compared with the CPP case.

Remark: Notice that a different choice of the conditional distribution $f(z \mid y)$, in the above construction, would lead to a different dependence structure. For instance, choosing $f(z \mid y) = \mathrm{Po}(z \mid c\,y)$ implies that $f(y \mid z) = \mathrm{Ga}(y \mid a+z, b+c)$ and $f(z) = \mathrm{Pga}(z \mid a, b, c)$, where $\mathrm{Pga}(z \mid a, b, c)$ denotes a Poisson-gamma distribution with mean $c\,a/b$. Therefore, for constructing an exchangeable sequence we take $Y_i \mid Z \sim \mathrm{Ga}(a + z, b + c)$ for $i = 1, 2, \ldots$ conditionally independent given Z and $Z \sim \mathrm{Pga}(a, b, c)$ hence $Y_i \sim \mathrm{Ga}(a, b)$ with $\mathrm{Corr}(Y_i, Y_j) = c/\{(b + c)\}$, for $i \neq j$.

2.2. Exchangeable sequences: Nonparametric method

The parametric method described in the previous section has the feature of having a wide variety of choices for the conditional distribution $f(z \mid y)$, leading then to different parametric dependence structures. Alternatively, instead of having a latent random variable Z, we could consider a latent random distribution G, i.e., the conditional density $f(z \mid y)$ will be replaced by a nonparametric density measure $\mathcal{P}$. By doing this, we will be able to define a nonparametric dependence structure within the exchangeable sequence. The nonparametric nature underlied to this construction coveys to a dependence between the variables not depending on the marginal distributions.

Having in mind the previous parametric construction, instead of the latent variable Z we consider a latent random distribution G with conditional random distribution $G \mid Y \sim \mathcal{P}(\cdot \mid y)$. In this case,

$$F(\mathrm{d}y \mid G) = G(\mathrm{d}y)$$

with

$$G \sim \mathcal{P}.$$

Proceeding analogously to the previous construction we would need to choose $\mathcal{P}$ such that

$$\mathrm{E}_{\mathcal{P}}\{G(\mathrm{d}y)\} = \int G(\mathrm{d}y)\mathcal{P}(\mathrm{d}G) = F(\mathrm{d}y).$$

The above is a characteristic property that many probability measures on the space of distributions, used in the Bayesian nonparametric literature, satisfy. This is the case of the seminal Dirichlet process introduced by Ferguson [17] and most of its generalizations such as species sampling models presented by Pitman [18] and normalized random measures with independent increments analyzed in Regazzini et al. [19]. The above feature, is attractive in the Bayesian nonparametric literature, since it allows to set F as the prior guess (mean) at the shape for the realizations of G.

Now, if we take $Y_i \mid G \sim G$, for $i = 1, 2, \ldots$ conditional independent random variables given G, and $G \sim \mathcal{P}$, then $Y_1, Y_2, \ldots$ is a sequence of exchangeable random variables with marginal distributions $F(\mathrm{d}y) = \mathrm{E}_{\mathcal{P}}\{G(\mathrm{d}y)\}$. Mena and Walker [20] also used similar ideas based on nonparametric predictive distributions to define the dynamics of a first order autoregressive processes.

Example 2. In order to define an exchangeable sequence with marginal distributions $\mathrm{Ga}(a, b)$ through the nonparametric method just described, we consider a Dirichlet process $\mathcal{DP}(F/c)$ as the law $\mathcal{P}$ of G, where $1/c > 0$ is the precision parameter and F is a parametric c.d.f. that coincides with the mean of the process G, for details see Ferguson [17]. Now we want that $\mathrm{E}_{\mathcal{DP}}\{G(\mathrm{d}y)\} = F(\mathrm{d}y)$, with F the c.d.f. of a gamma distribution. Thus, if we take $Y_i \mid G \sim G$ for $i = 1, 2, \ldots$ conditionally independent given G, and $G \sim \mathcal{DP}(F/c)$, with $F(\mathrm{d}y) = \mathrm{Ga}(y \mid a, b)\mathrm{d}y$ and $c > 0$ then $Y_i \sim \mathrm{Ga}(a, b)$ with $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Var}_{\mathcal{DP}}(\mu_G)$ and

$$\mu_G = \int yG(\mathrm{d}y),$$

for $i \neq j$.

The distribution of the mean μ_G has been studied by Regazzini et al. [19] and others. These authors provide theoretical expressions for the distribution of the functional μ_G for any centering function F, however, explicit expressions are not available in closed from, except for particular cases of F. On the other hand, there is an alternative way of obtaining dependence properties of an exchangeable sequence modeled by the Dirichlet process.

According to Blackwell and MacQueen [21], the joint distribution of the exchangeable sequence $\{Y_i\}$ where $Y_i \mid G \sim G$ for $i = 1, 2, \ldots$ conditionally independent given G, and $G \sim \mathcal{DP}(F/c)$, can be characterized, after a marginalization of G, by the Pólya urn type updating

$$Y_1 \sim F,$$
$$Y_2 \mid Y_1 \sim \frac{1}{c+1}F + \frac{c}{c+1}F_1$$

and in general,

$$Y_i \mid Y_1, \ldots, Y_{i-1} \sim \frac{1}{c(i-1)+1}F + \frac{c(i-1)}{c(i-1)+1}F_{i-1}(y_i), \tag{9}$$

where $F_i(\cdot)$ is the empirical distribution function (e.d.f.) of the first $i-1$ observations. With this characterization of the exchangeable sequence, induced by the Dirichlet process, we are able to compute the covariance of any pair (Y_i, Y_j).

Proposition 2. *Let $\{Y_i\}$ be an exchangeable sequence such that $Y_i \mid G \sim G$ for $i = 1, 2, \ldots$ conditionally independent given G, and $G \sim \mathcal{DP}(F/c)$, where F is a centering function and $1/c > 0$ the precision parameter. Then,*

$$\mathit{Corr}(Y_i, Y_j) = \frac{c}{c+1}$$

for all $i \neq j$.

Proof. Let $\mu = \mathrm{E}(Y_i)$ and $\sigma^2 = \mathrm{Var}(Y_i)$, which correspond to the expected value and variance of F. Then, using conditional expectation we express $\mathrm{E}(Y_1 Y_2) = \mathrm{E}\{Y_1 \mathrm{E}(Y_2 \mid$

$Y_1)\}$. Now, based on the Pólya urn representation (9) we obtain that $\mathrm{E}(Y_2 \mid Y_1) = \mu/(c+1)+cY_1/(c+1)$ which implies that $\mathrm{E}(Y_1\,Y_2) = \mu^2+\sigma^2 c/(c+1)$. Thus, $\mathrm{Cov}(Y_1, Y_2) = \sigma^2 c/(c+1)$ and $\mathrm{Corr}(Y_1, Y_2) = c/(c+1)$. Therefore, as $Y_1, Y_2, \ldots$ is an exchangeable sequence then $\mathrm{Corr}(Y_i, Y_j) = \mathrm{Corr}(Y_1, Y_2)$, which completes the proof. □

Proposition 2 shows that the correlation induced by the exchangeable sequence $\{Y_i\}$ sampled from the Dirichlet process is, in fact, nonparametric in the sense that it is independent of the marginal distribution of the Y_i's and only depends on the parameter c. This is in contrast with the parametric constructions of exchangeable sequences where the correlation depends on the marginal distribution of the Y_i's.

Example 2 (Continued...) We consider the two processes X_t^I and X_t^E based on independent and exchangeable claims respectively, but now exchangeability is defined through the nonparametric (Dirichlet) construction with $\mathrm{Ga}(a, b)$ marginals. Then from Proposition 1, the moments of the total claims amount in both cases are,

$$\mathrm{E}\left\{X_t^I\right\} = \mathrm{E}\left\{X_t^E\right\} = \frac{a}{b}\lambda t,$$

$$\mathrm{Var}\left\{X_t^E\right\} = \mathrm{Var}\left\{X_t^I\right\} + \frac{ac}{b^2(c+1)}\lambda^2 t^2,$$

and the covariance for $t < s$ is

$$\mathrm{Cov}\left\{X_t^E, X_s^E\right\} = \mathrm{Cov}\left\{X_t^I, X_s^I\right\} + \frac{ac}{b^2(c+1)}\lambda^2 ts,$$

where $\mathrm{Var}\left\{X_t^I\right\}$ and $\mathrm{Cov}\left\{X_t^I, X_s^I\right\}$ are given in Example 1. Again, overdispersion of the process X_t^E with respect to process X_t^I is also clear.

We would like to clarify that the Bayesian nonparametric ideas mentioned here have been employed as a constructive tool rather than for inference procedure, for nonparametric and semiparametric inference on non-homogeneous Poisson processes we refer the reader to Merrick et al. [22], for example.

3. Bayesian inference of CPP and ECP

Once we have proposed a way of constructing exchangeable sequences for defining an ECP, then for a given data set we would like to make inference on the parameters of both, the CPP and the ECP models to compare. The ECP model acknowledges that claims made by the same individual are (or can be) positive correlated. This assumption determines the way our data has to be recorded. That is, for each individual j we require information of all his/her claims $\{Y_{ij}\}$, for $i = 1, \ldots, N_{tj}$, with $N_{tj} \sim \mathrm{Po}(\lambda t)$ independent across j and independent of $\{Y_{ij}\}$. The aggregated claims for individual j, is X_{tj} constructed as in (1) for the independence case or as in (6) for the dependence case, for $j = 1, \ldots, m$. The type of data we require is usually called panel data, where the same set of individuals is followed and measured along time.

The Bayesian approach for making inference has become very popular in several areas, including actuarial sciences (see, for example, Klugman [23]), due to its advantage of combining all available information in a probability distribution. Here we will follow this approach.

3.1. CPP model

In order to set ideas, let us start with the traditional CPP model with independent claim sizes. Let $X_{t1}, \ldots, X_{tm}$ be a collection of independent CPP's as defined in (1), with $Y_{ij} \sim f(y \mid \theta)$ independent for $i = 1, 2, \ldots$ and independent of $N_{tj} \sim \mathrm{Po}(\lambda t)$, for $j = 1, \ldots, m$, with θ and λ the parameters of the claims density and the Poisson processes, respectively.

Due to the independence of the Poisson processes N_{tj} and the claim sizes Y_{ij}, inference for λ and θ can be done separately. Let $J_{1j}, J_{2j}, \ldots, J_{n_j j}$ be the jump times in the $\mathrm{Po}(\lambda t)$ process for individual j, it is well known that the inter-arrival times $W_{1j} = J_{1j}, W_{2j} = J_{2j} - J_{1j}, \ldots$ are independent random variables with $\mathrm{Ga}(1, \lambda)$ distributions. Therefore, inference on λ reduces to the standard estimation problem of the scale parameter in a sample of independent gamma random variables. Thus, a conjugate analysis for λ is obtained when assuming a prior distribution $\pi(\lambda) = \mathrm{Ga}(\lambda \mid \alpha_\lambda, \beta_\lambda)$. Moreover, due to the independence of the processes N_{tj} for all j, the posterior distribution for λ becomes

$$\pi(\lambda \mid \mathbf{w}) = \mathrm{Ga}\left(\lambda \left| \alpha_\lambda + \sum_{j=1}^{m} n_j,\ \beta_\lambda + \sum_{j=1}^{m} \sum_{i=1}^{n_j} w_{ij} \right.\right), \tag{10}$$

which in terms of the jump times $\mathbf{J} = \{J_{ij}\}$ can be expressed as $\pi(\lambda \mid \mathbf{J}) = \mathrm{Ga}\Big(\lambda \mid \alpha_\lambda + \sum_{j=1}^m n_j, \beta_\lambda + \sum_{j=1}^m J_{n_j j}\Big)$.

Now, given $N_{tj} = n_j$, we let $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})$ denote the set of all claims of individual j, $j = 1, \ldots, m$. Then, the likelihood for θ is given by

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_m \mid \theta) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} f(y_{ij} \mid \theta).$$

If we assume that our prior knowledge on θ is summarized in a prior distribution $\pi(\theta)$, then the posterior distribution of θ is obtained through the Bayes' Theorem, i.e.,

$$\pi(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta),$$

with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$. In Section 4, an explicit form of this posterior distribution, will be obtained.

3.2. ECP model

Lets now assume that $X_{t1}, \ldots, X_{tm}$ are independent ECP's as the one given in Definition 1, with $N_{tj} \mid \lambda \sim \mathrm{Po}(\lambda t)$ and $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, \ldots)$ an exchangeable sequence with $f(y \mid \theta)$ marginal distributions. The sequences $\mathbf{Y}_j$'s are independent of each other and independent of the processes N_{tj}, for $j = 1, \ldots, m$. Let us consider two cases, the parametric and the nonparametric construction of the exchangeable sequences:

Parametric case

Recall the parametric construction of the exchangeable sequence with given marginal distributions presented in Section 2.1, that is, for each $j = 1, \ldots, m$ we require a latent variable $Z_j \sim f(z \mid c)$ such that $Y_{ij} \mid Z_j \overset{\text{iid}}{\sim} f(y \mid \theta, z_j)$ for $i = 1, \ldots, n_j$, with (θ, c) parameters of the model.

Given $N_{tj} = n_j$, the likelihood for (θ, c) is given by

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_m \mid \theta, c) = \prod_{j=1}^{m} f(\mathbf{y}_j \mid \theta, c) \tag{11}$$

where

$$f(\mathbf{y}_j \mid \theta, c) = f(y_{1j}, \ldots, y_{n_j,j} \mid \theta, c) = \int \left\{ \prod_{i=1}^{n_j} f(y_{ij} \mid z_j, \theta) \right\} f(z_j \mid c) \mathrm{d}\mu_2(z_j). \tag{12}$$

The marginalization over z_j does not usually have an analytic expression, except for particular cases. When expression (12) is available analytically, inference on (θ, c) is conducted as in the CPP case via the Bayes' Theorem using the likelihood (11). However, if the integral in expression (12) is not analytically available in closed form, we can get around by considering, for the moment, that we have observed $Z_j = z_j$ along with the y_{ij}'s. If we denote $\mathbf{z} = (z_1, \ldots, z_m)$, an extended likelihood has the form

$$f(\mathbf{y}, \mathbf{z} \mid \theta, c) = \prod_{j=1}^{m} f(\mathbf{y}_j, z_j \mid \theta, c) \tag{13}$$

with

$$f(\mathbf{y}_j, z_j \mid \theta, c) = \left\{ \prod_{i=1}^{n_j} f(y_{ij} \mid z_j, \theta) \right\} f(z_j \mid c).$$

If $\pi(\theta, c)$ represents our prior knowledge on (θ, c) then the posterior distribution is given by

$$\pi(\theta, c \mid \mathbf{y}, \mathbf{z}) \propto f(\mathbf{y}, \mathbf{z} \mid \theta, c)\pi(\theta, c).$$

Note that an estimability condition for the parameter c is that $m > 1$.

Remembering that we have assumed that $Z_j = z_j$ was observed. Now, to make posterior inference we implement a Gibbs sampling scheme (see, for example, Smith and Roberts, [24]) with the previous conditional posterior distribution of θ and c as well as the conditional distribution of each Z_j which is given by

$$f(z_j \mid \mathbf{y}, \theta, c) \propto f(\mathbf{y}_j, z_j \mid \theta, c)$$

given in (13), for $j = 1, \ldots, m$. By doing this, we will have a simulated value of Z_j in each iteration. Finally, due to the independence between N_{t_j} and the Y_{ij}'s, the posterior distribution for λ is that given in (10).

Nonparametric case

Let us now consider the nonparametric construction of the exchangeable sequence with given marginal distributions via the Dirichlet process as in Example 2. That is, via a latent measure $G \sim \mathcal{DP}(F/c)$ and $Y_i \mid G \overset{\text{iid}}{\sim} G$.

Given that we have observed $N_{tj} = n_j$, then the likelihood function for (θ, c) is given by

$$f(\mathbf{y} \mid \theta, c) = \prod_{j=1}^{m} f(\mathbf{y}_j \mid \theta, c)$$

with

$$f(\mathbf{y}_j \mid \theta, c) = f(y_{1j}, \ldots, y_{n_j j} \mid \theta, c) = \mathrm{E}_{\mathcal{DP}} \left\{ \prod_{i=1}^{n_j} G(\mathrm{d}y_{ij}) \right\}.$$

Blackwell and MacQueen [21] showed that this joint distribution for the exchangeable sequence can be obtained by the product of expressions coming from the Pólya urn representation (9). Therefore, dropping for the moment the subscript j, we get

$$f(y_1, \ldots, y_n \mid \theta, c) = \prod_{i=1}^{n} \left\{ \left(\frac{1}{1 + c(i-1)} \right) f(y_i \mid \theta) + \left(\frac{c(i-1)}{1 + c(i-1)} \right) \sum_{l=1}^{i-1} \delta_{Y_l}(y_i) \right\},$$

where $\delta_Y(\cdot)$ is the degenerated measure that assigns probability one to the point Y. If we let $y_1^*, \ldots, y_k^*$ the distinct y_i's for $i = 1, \ldots, n$, with $k \leq n$ and after some algebra, the joint distribution for the exchangeable sequence becomes

$$f(y_1, \ldots, y_n \mid \theta, c) = \frac{(1/c)^k \Gamma(1/c)}{\Gamma(1/c + n)} \prod_{i=1}^{k} f(y_i^* \mid \theta), \tag{14}$$

where $\Gamma(\cdot)$ denotes the gamma function.

Having a closer look, we can see that as a function of θ, this likelihood is the same as the likelihood obtained with the traditional CPP but when considering the distinct observations only. This is a feature of the Dirichlet process that allows ties in the observations and thus information on θ only comes from the distinct observations. Moreover, the number of distinct observations k_j, for $j = 1, \ldots, m$, provides information about the parameter c, that is, a smaller value of k_j produces a sharper likelihood for c; however, if no ties are present in any of the exchangeable sequences, the likelihood for c becomes flat.

Now, considering that (θ, c) have prior distribution $\pi(\theta, c)$, then the posterior distribution is simply

$$\pi(\theta, c \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta, c)\pi(\theta, c).$$

Note that given the form of the j-th contribution to the likelihood, (14), if θ and c are independent a-priori then they will also be independent a-posteriori. Again, the posterior distribution for λ is the same as before, given in (10).

Due to the choice of the Dirichlet process to construct the exchangeable sequence, inference on c does not necessarily require several realizations of the ECP as in the parametric case, only one realization of the process would be enough as long as the observed claims sizes had repeated values. This is a consequence of the discreteness of the Dirichlet process (see, Blackwell and MacQueen [21]).

4. Illustration

The Medical Expenditure Panel survey (MEPS) provides information on health care use, expenditures, sources of payment, and health insurance coverage of the US civilian non-institutionalized population. A national representative sample of households is selected every year and information is reported for all members by a single household respondent.

The panel design of the survey provides data for examining person level changes in selected variables including expenditures among others. In particular, MEPS HC-094D (available at: http://www.meps.ahrq.gov) contains information for the 2005 hospital inpatient stays. There are two steps for collecting the information. In the first step, the Household Component, information about each household member is collected. Upon completion of the household interview and obtaining permission from the household respondents, the Medical Provider Component (second step) is carried out. A sample of medical providers are contacted by phone to obtain information that household respondents could not accurately provide.

There are 3341 events (hospital inpatient stays) reported in the database. From those we selected 2110 events which contained complete Medical Provider Component data. We removed the events with missing data and after we added up the claims made in the same day to have a single claim per day, we ended up with 1729 events for which we have the starting date for the event and the total expenditures, resulting from both facility and physician amounts. The 1729 events were made by 66 individuals, which gives an average of 26 events per individual in the year, ranging from 1 to 122 events per person.

We will carry out a comparison when fitting the traditional CPP and the two ECP models introduced in this paper. The expenditures values were transformed to satisfy a parametric assumption. The best power transformation of the data was achieved with a power of 1/4, which produced data that are well modeled by a gamma distribution. After the data was transformed we kept only the first two significant decimal places. The date of the events was also transformed to elapsed days within the same calendar year (2005).

The general setting is the following: We have a sample $X_{t1},\ldots,X_{tm}$ of independent aggregated total expenditures up to time t made by individuals $j=1,\ldots,m$, such that the number of events N_{tj} follows a homogeneous Poisson process $\mathrm{Po}(\lambda t)$ independent of the expenditures sizes Y_{ij} which are gamma distributed with parameters a and b.

We then start by assuming that X_{t_j} is a CPP with independent claims, that is, $Y_{ij}\mid a,b\sim\mathrm{Ga}(a,b)$ are all independent for $i=1,\ldots,n_j$ and $j=1,\ldots,m$. If our prior knowledge on (a,b) can be represented by $\pi(a)=\mathrm{Ga}(a\mid\alpha_a,\beta_a)$ and $\pi(b)=\mathrm{Ga}(b\mid\alpha_b,\beta_b)$ independently, then the posterior distribution, given the sample, is characterized by the conditional distributions

$$\pi(a\mid\mathbf{y},b)\propto\{\Gamma(a)\}^{-n}a^{\alpha_a-1}\left(b^n e^{-\beta_a}\prod_{j=1}^{m}\prod_{i=1}^{n_j}y_{ij}\right)^{a}I(a>0),$$

with $n=\sum_{j=1}^{m}n_j$, and

$$\pi(b\mid\mathbf{y},a)=\mathrm{Ga}\left(b\,\middle|\,\alpha_b+n\,a,\;\beta_b+\sum_{j=1}^{m}\sum_{i=1}^{n_j}y_{ij}\right).$$

Now we will assume that X_{tj} is an ECP with claims $\mathbf{Y}_j=(Y_{1j},\ldots,Y_{n_j j})$ following an exchangeable sequence of gamma random variables. We consider both, the parametric and the nonparametric approaches. For the parametric method we use the construction given in Example 1. That is, take independent latent variables $Z_j \sim \text{Gga}(a, b, c)$, for $j = 1, \ldots, m$, and for each of them define conditionally independent random variables $Y_{ij} \mid Z_j \sim \text{Ga}(a + c, b + z_j)$ for $i = 1, \ldots, n_j$.

For this particular construction of the exchangeable sequence, the joint distribution for $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$, marginalizing over $\mathbf{z}$, is not available explicitly. Therefore we rely on the extended likelihood, as in (13). Now, assuming that (a, b, c) are independent a-priori such that $\pi(a) = \text{Ga}(a \mid \alpha_a, \beta_a)$, $\pi(b) = \text{Ga}(b \mid \alpha_b, \beta_b)$ and $\pi(c) = \text{Ga}(c \mid \alpha_c, \beta_c)$, then the posterior distribution given $(\mathbf{y}, \mathbf{z})$ is characterized by the full conditionals

$$\pi(a \mid \mathbf{y}, \mathbf{z}, b, c) \propto \frac{a^{\alpha_a - 1}}{\Gamma^m(a) \prod_{j=1}^m \Gamma^{(n_j - 1)}(a + z_j)} \left[e^{-\beta_a} b^m \left\{ \prod_{j=1}^m (b + z_j)^{(n_j - 1)} \prod_{i=1}^{n_j} y_{ij} \right\} \right]^a I(a > 0),$$

$$\pi(b \mid \mathbf{y}, \mathbf{z}, a, c) \propto \left\{ \prod_{j=1}^m (b + z_j)^{(n_j - 1)} \right\}^{a+c} b^{\alpha_b + ma - 1} \exp\left\{ -b \left(\beta_b + \sum_{j=1}^m \sum_{i=1}^{n_j} y_{ij} \right) \right\} I(b > 0)$$

and

$$\pi(c \mid \mathbf{y}, \mathbf{z}, a, b) \propto \frac{c^{\alpha_c - 1}}{\{\Gamma(c)\}^m \{\Gamma(a + c)\}^{n-m}} \exp\left[-\beta_c \left\{ \prod_{j=1}^m (b + z_j)^{(n_j - 1)} z_j \prod_{i=1}^{n_j} y_{ij} \right\} \right]^c I(c > 0).$$

Finally we also require the full conditional distribution for the latents z_j's which are

$$f(z_j \mid \mathbf{y}, a, b, c) \propto (b + z_j)^{(n_j - 1)(a+c)} z_j^{c-1} e^{-z_j \sum_{i=1}^{n_j} y_{ij}} I(z_j > 0),$$

for $j = 1, \ldots, m$.

For the construction of an exchangeable sequence with $\text{Ga}(a, b)$ marginals under the nonparametric approach we will use the Dirichlet process as in Example 2. That is, we take $Y_i \mid G \sim G$ for $i = 1, 2, \ldots$ conditionally independent given G and $G \sim \mathcal{DP}(F/c)$ with $F(\mathrm{d}y) = \text{Ga}(y \mid a, b)\mathrm{d}y$. If we use the same prior distribution for (a, b, c) as in the parametric construction, given the exchangeable sequences $\mathbf{y}$, the conditional posterior distributions required to make inference become

$$\pi(a \mid \mathbf{y}, b) \propto \{\Gamma(a)\}^{-k} a^{\alpha_a - 1} \left(b^k e^{-\beta_a} \prod_{j=1}^m \prod_{i=1}^{k_j} y_{ij}^* \right)^a I(a > 0),$$

where $k = \sum_{j=1}^m k_j$ and k_j denote the number of distinct observation in each exchangeable sequence $\mathbf{y}_j$ and $(y_{1j}^*, \ldots, y_{k_j,j}^*)$ the distinct observations for $j = 1, \ldots, m$, and

$$\pi(b \mid \mathbf{y}, a) = \text{Ga}\left(b \,\middle|\, \alpha_b + k\, a,\ \beta_b + \sum_{j=1}^m \sum_{i=1}^{k_j} y_{ij}^* \right).$$

Note that since (a, b) and c are independent a-priori, they are also independent a-posteriori. Moreover, the posterior distribution for c is given by

$$\pi(c \mid \mathbf{y}) \propto \frac{\Gamma^m(1/c)}{\prod_{j=1}^m \Gamma(1/c + n_j)} c^{\alpha_c - \sum_{j=1}^m k_j - 1} e^{-\beta_c c} I(c > 0).$$

Finally, in all cases, the posterior distribution for the intensity parameter λ of the Poisson processes $N_j(t)$ for $j = 1, \ldots, m$, when considering a conjugate analysis, is given by equation (10).

We carried out posterior inference for the expenditures data by implementing Gibbs samplers for the three models, CPP, $\mathrm{ECP_p}$ and $\mathrm{ECP_{np}}$. For sampling from each of the conditional distributions we used random walk Metropolis-Hastings steps, that is, at iteration $t+1$ we took $\theta^* \sim \mathrm{Ga}(1, 1/\theta^{(t)})$ as proposal variate. This proposal distribution is centered at the previous value of the chain and has a variation coefficient of one. A Fortran code of the algorithm is available upon request from the second author. We considered vague prior distributions for all the parameters of the model, i.e. we took $(\alpha_a, \beta_a) = (0.01, 0.01)$, $(\alpha_b, \beta_b) = (0.01, 0.01)$, $(\alpha_c, \beta_c) = (0.01, 0.01)$ and $(\alpha_\lambda, \beta_\lambda) = (0.01, 0.01)$. The Gibbs sampler was run for 600,000 iterations with a burn-in of 100,000 keeping every $50th$ observation after burn-in to reduce the autocorrelation in the chain.

Our main parameter of interest is the one that controls the dependence among observations, that is c. It is worth mentioning that in both (parametric and nonparametric) models a value of $c = 0$ implies that the observations are all independent. However their magnitudes are given in different scales. Posterior distributions of c for the two cases are shown in Figure 2. The left graph corresponds to the parametric construction ($\mathrm{ECP_p}$) where positive values of c are clearly supported by the data. In fact the 5% lower quantile of the distribution is 0.32 which means that the posterior probability that c takes a larger value is 95%. On the other hand, the right graph of Figure 2 contains the posterior distribution of c using the nonparametric construction ($\mathrm{ECP_{np}}$). From here we can observe that only fairly small values of c are supported by the data which means that there is very little or none dependence captured by the nonparametric model.

To better discriminate among the models, we use the logarithm of the pseudo-marginal likelihood (LPML) statistic. This statistic is a measure of the marginal fitting to the data and has been used as a model selection criterion in many different contexts (see, for example, Sinha and Dey [25]). Table 1 summarizes the LPML statistic as well as the posterior estimates of the parameters of the models. The fitting of the $\mathrm{ECP_p}$ provides the best fitting among the three models. This suggests that there is certainly a positive dependence in the claims which has been captured by the $\mathrm{ECP_p}$ model. On the other hand, the worst fitting was obtained by model $\mathrm{ECP_{np}}$, which can be explained by the fact that the dependence in the nonparametric construction, when using the Dirichlet process, relies on the existence of ties in the claim amounts and in the expenditures dataset there are quite few (77 out of 1729).

Posterior estimates of a and b, reported in Table 1 are very similar in all models, being slightly smaller the point estimates in the $\mathrm{ECP_p}$ model. 95% credible intervals are also included in Table 1 for reference. Different values for the hyper parameters of the prior distributions, not reported here, were also studied and gave similar results in terms of fitting of the models and estimators. Therefore, by allowing the claims to be dependent within each individual, via the parametric construction, we are able to better modelling this dataset.

From now on we will concentrate in comparing the independence model (CPP) and the dependence parametric model ($\mathrm{ECP_p}$). There are several ways of appreciating the impact of the parameter c in the $\mathrm{ECP_p}$ model. The correlation among claims made by the same individual is estimated to lie in the interval $(0.021, 0.112)$ with 95% of probability. In order to understand the implications in the ruin probability, we have repeated the same analysis done in Example 1 with a safety loading of 0.1. From Figure 3, it is evident that widening the model to allow for exchangeable claims leads to higher ruin probabilities than those corresponding to the CCP, in both the finite and the infinite horizon case. In particular, a slower decay in the ruin probability, as the initial capital varies, is noted in the $\mathrm{ECP_p}$ when compared to the CPP. Therefore, leading to a more realistic scenario when determining an adequate risk premium rate.

Finally, we carried out a predictive analysis for the aggregated expenditures of a patient in a year. For that we needed to consider the information about the frequencies of occurrence of the claims, modeled by the Poisson processes N_{tj}, $j = 1, \ldots, m$ and in particular by the intensity λ. Considering that we observed 1729 claims in the year, then the posterior distribution for the intensity of the claims per year, λ, is $\mathrm{Ga}(1729.01, 52.7479)$. Therefore, the posterior mean rate is 32.78 claims per person per year.

With the posterior distribution for λ and the posterior predictive distribution for the whole sequence of claims, we obtain the posterior predictive distribution for the aggregated claims in a year for one individual. Figure4 contains normal smoothed density estimators of this aggregated claims distribution in thousands of dollars (back-transforming each claim with a power of 4 adding them up and dividing by one thousand). The solid line corresponds to the aggregated claims distribution obtained with the CPP model whereas the dotted line corresponds to the one obtained with the $\mathrm{ECP_p}$ model. As we can see, the predictive distribution with the $\mathrm{ECP_p}$ is slightly shifted to the left, has heavier tails and is over-dispersed with respect to the one obtained with the CPP model.

For the insurance company, having a model that better represents the data implies a more accurate scenario. For instance, for determining the reserve for a new insuree one could be conservative and take the 95% quantile of the aggregated claims distribution in a year. These values are $606,438$ dollars and $618,370$ dollars for the CPP and the $\mathrm{ECP_p}$ models respectively. Alternatively, the traditional way of determining the reserve is in terms of the posterior mean, which in our case are $394,796$ and $364,120$ dollars respectively. This would result in $30,676$ dollars saving in reserves per person per year.

5. Discussion.

In this paper we introduced a generalization of the compound Poisson process by relaxing the independence assumption in the claims to a more realistic exchangeable assumption. The resulting process has important implications in the ruin probability tending, in average, to ruin faster than when using the traditional CPP process. Moreover, predictive distributions of the aggregated claims are more accurate, thus helping the insurance companies to better determine reserves.

We proposed two general ways of defining exchangeable claims with arbitrary but fixed marginal distributions. This is done by means of conditioning on a latent random variable (parametric approach) or a latent random distribution (nonparametric approach) as in de Finetti's representation Theorem [15]. It is worth emphasizing that this approach is the only way of constructing infinite exchangeable sequences.

The proposed methodology was illustrated for gamma distributed claims, however it can be equally applied to any distribution. In addition, we provided Bayesian procedures to tackle the problem of parameter estimation. An illustration regarding medical expenditures was presented together with possible scenarios of the ruin probability under different initial capitals and policy durations. From this, together with a simulation example, we can conclude that considering an ECP process instead of a CPP we are potentially able to capture more riskier situations and therefore worth to consider when modeling reserve processes.

As we mentioned in Section 2, we employ Bayesian nonparametric ideas as a constructive technique to define exchangeable sequences rather than as an estimation procedure. However, as a byproduct, the nonparametric approach at issue can also be used to elucidate the claim distribution. Instead of marginalizing the latent distribution G as we did for the Dirichlet case, posterior inference could be carried out for the nonparametric G and the posterior predictive distribution of G given the data would be an estimate of the marginal density that generated the claims.

Acknowledgments.

The authors are grateful for the valuable comments of two anonymous referees and an associate editor. This research was partially supported by grants J48072-F and J50160-F from the National Council for Science and Technology of Mexico (CONACYT). Ramsés H. Mena gratefully acknowledge the support of ICER, Turin.

References.

1. Lundberg F. *Försäkringsteknisk Riskutjämning.* F. Englunds boktryckeri A.B.: Stockholm, 1926.

2. Cramér H. *On the mathematical theory of risk.* Skandia Jubilee Volume. Stockholm, 1930.

3. Feller W. *An Introduction to Probability Theory and Its Applications, Volume 2.* Wiley, 1971.

4. Rolski T, Schmidli H, Schmidt V, Teugels J. *Stochastic processes for insurance and finance.* New York: Wiley, 2001.

5. Grandell J. *Mixed Poisson Processes.* Chapman and Hall, 1997.

6. Daboni D. Some models of inference in the risk theory from a Bayesian viewpoint. *ASTIN bull.* 1974; **8**, 38-56.

7. Morales M, Schoutens W. A risk model driven by Lévy processes. *Applied Stochastic Models in Business and Industry* 2003; **19**, 147–167.

8. Gerber HU. Ruin theory in the linear model. *Insurance: Mathematics and Economics* 1982; **1**, 177-184.

9. Promislow SD. The probability of ruin in a process with dependent increments. *Insurance: Mathematics and Economics* 1991; **10**, 99-107.

10. Cossette H, Marceau E. The discrete-time risk model with correlated classes of business. *Insurance: Mathematics and Economics* 2000; **26**, 133-149.

11. Mikosch T, Samorodnitsky G. Ruin probability with claims modeled by a stationary ergodic stable process. *The Annals of Probability* 2000; **28**, 1814-1854.

12. Bühlmann H. *Austauschbare stochatische Variabeln unf ihre Grenzwertsaetze.* Univ. of California Press, 1960.

13. Bühlmann H. *Mathematical Methods in Risk Theory.* Springer, 1970.

14. Klass M, Teicher H. The central limit theorem for exchangeable random variables without moments. *The Annals of Probability* 1987; **15**, 138–153.

15. de Finetti B. Le prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré (Paris)* 1937; **7**, 1–68.

16. Pitt MK, Chatfield C, Walker SG. Constructing first order autoregressive models via latent processes. *Scandinavian Journal of Statistics* 2002; **29**, 657–663.

17. Ferguson T. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**, 209-230.

18. Pitman J. Some developments of the Blackwell-MacQueen urn scheme. *Probability and Game Theory. Papers in honor of David Blackwell*, (ed T.S. Ferguson *et al.*), IMS Lecture Notes Monograph Series, 1996; **30**, 245–267.

19. Regazzini E, Lijoi A, Prünster I. Distributional results for means of random measures with independent increments. *Annals of Statistics* 2003; **31**, 560–585.

20. Mena RH, Walker SG. Stationary autoregressive models via a Bayesian nonparametric approach. *Journal of Time Series Analysis* 2005; **26**, 789-805.

21. Blackwell D, MacQueen JB. Ferguson distributions via pólya urn schemes. *Annals of Statistics* 1973; **1**, 353-355.

22. Merrick JRW, Soyer R, Mazzuchi TA. Are maintenance practices for railroad tracks effective? *Journal of the American Statistical Association* 2005; **100**, 17-25.

23. Klugman SA. *Bayesian Statistics in Actuarial Science: With emphasis on credibility.* New York: Springer, 2003.

24. Smith AFM, Roberts GO. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* B 1993; **55**, 3–23.

25. Sinha D, Dey DK. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association* 1997; **92**, 1195–1212.

Quantity	Model					
	CPP		ECP_p		ECP_{np}	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
LPML	-4124.6	–	-4117.9	–	-4555.8	–
a	11.99	(11.25, 12.72)	11.45	(10.61, 12.24)	11.58	(10.81, 12.48)
b	1.28	(1.20, 1.36)	1.25	(1.17, 1.34)	1.24	(1.15, 1.33)
c	–	–	0.78	(0.27, 1.50)	0.0013	(0.001, 0.002)

Table 1: Model comparison and posterior summaries of (a, b, c) for the CCP and the ECP models. Posterior mean and 95% credible intervals.

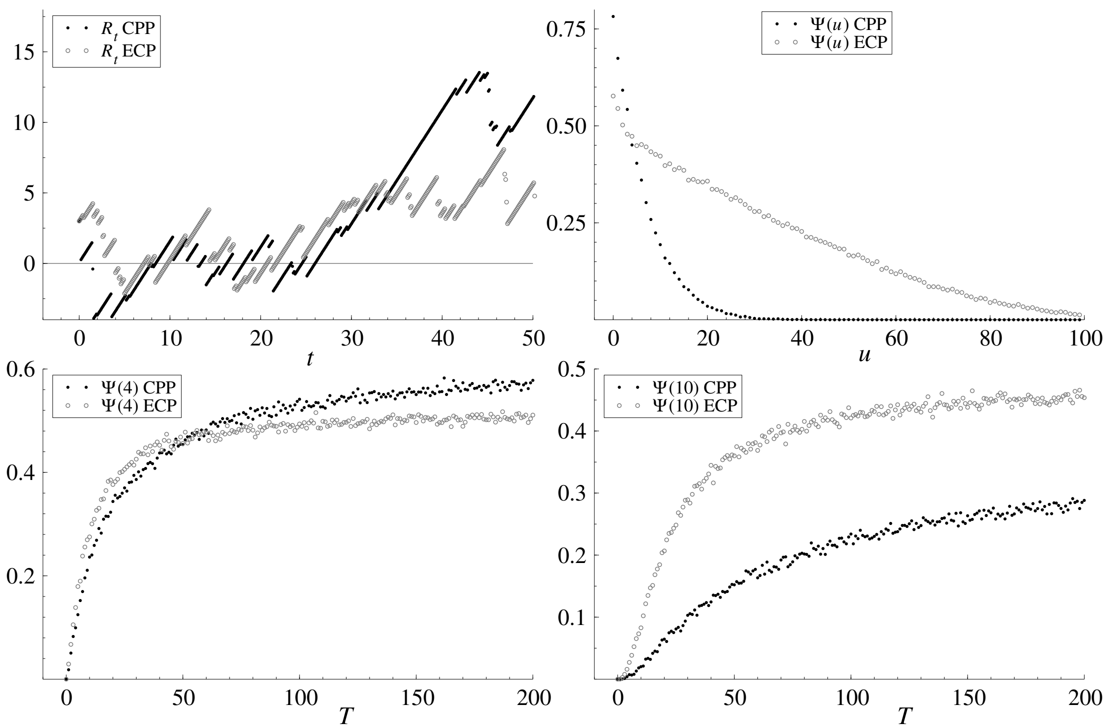

Figure 1: Reserve processes and Monte Carlo (MC) estimates of ruin probabilities for both the CPP and ECP_p models. *Top left*: Two random realizations of the reserve processes. *Top right*: MC estimates of the probability of ruin as the initial capital u changes. *Bottom*: MC estimates of the finite-horizon probability of ruin as the horizon T increases and for a choice of $u = 4$ (*left*) and $u = 10$ (*right*). The MC estimates are based on 5000 iid realizations of the reserve process.

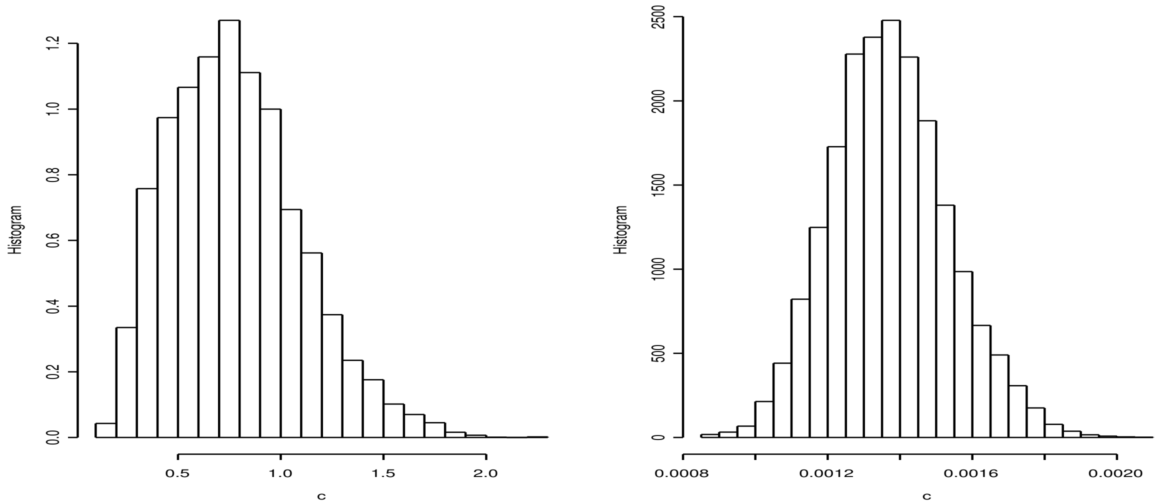

Figure 2: Posterior distribution of dependence parameter c: (*left*) parametric construction and (*right*) nonparametric construction.

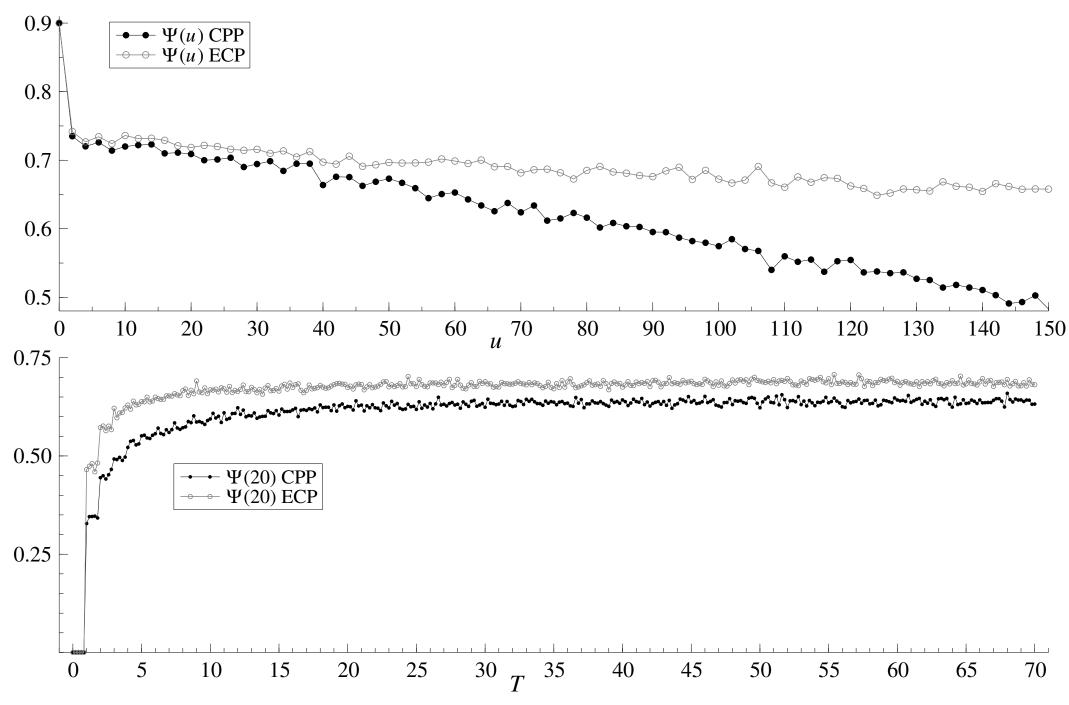

Figure 3: Monte Carlo estimates of ruin probabilities for both the CPP and $\mathrm{ECP_p}$ fitted to the MEPS data. *Top*: Ruin probability as a function of the initial capital u. *Bottom*: Finite horizon ruin probability as the horizon T increases, for $u = 20$.

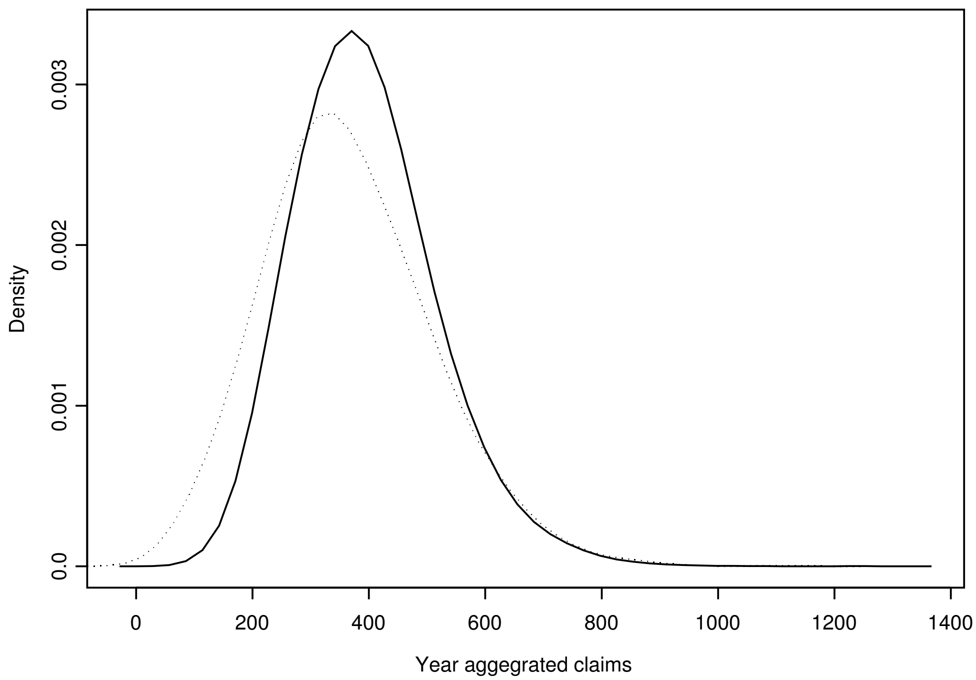

Figure 4: Posterior predictive distribution for the aggregated claims in a year for one individual (in thousands of dollars). (—) CPP, ($\cdots$) $\mathrm{ECP_p}$.