Université de Montréal

Actuarial applications of multivariate phase-type distributions : model calibration and credibility

par

Amin HASSAN ZADEH

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en statistique

septembre 2009

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

Actuarial applications of multivariate phase-type distributions : model calibration and credibility

présentée par

Amin HASSAN ZADEH

a été évaluée par un jury composé des personnes suivantes :

Pierre Duchesne
(président-rapporteur)

Martin Bilodeau
(directeur de recherche)

José Garrido
(co-directeur)

Patrice Gaillardetz
(membre du jury)

Emiliano Valdez
(examinateur externe)

William McCausland
(représentant du doyen de la FES)

Thèse acceptée le:
28 août 2009

SOMMAIRE

Les distributions phase-type (PH) sont utilisées comme modèles de probabilité d'une variable aléatoire positive. Elles remontent aux travaux de Neuts au milieu des années quatre-vingt. Les premières applications se retrouvent en recherche opérationnelle comme modèle pour les temps d'attente en théorie des files d'attente. L'actuariat est un autre domaine où l'on utilise souvent des modèles de probabilité de variables positives d'où l'intérêt des actuaires pour les distributions PH démontré récemment. L'estimation statistique des distributions PH par l'algorithme EM a été proposée au milieu des années quatre-vingt-dix par Asmussen et ses collaborateurs. Les actuaires ont aussi appliqué cette classe de modèles à la théorie du risque et aux probabilités de ruine. Des généralisations aux distributions PH multivariées ont aussi été introduites dans les années quatre-vingt suivant les travaux de Neuts qui ont fait école. Elles peuvent servir à la modélisation de la probabilité de deux ou plusieurs variables positives distribuées conjointement.

Le premier article traite de l'estimateur du maximum de vraisemblance par l'algorithme EM et de tests d'ajustement par le bootstrap paramétrique pour des distributions PH bivariées. Åhlström et ses collaborateurs ont publié en 1999 un algorithme EM pour l'estimation paramétrique de la distribution du temps de rechute en analyse de survie. Ils ont utilisé une distribution PH bivariée dont une composante est plus grande que l'autre avec probabilité un. Même si l'algorithme EM proposé dans cette thèse est semblable, il n'en demeure pas moins que notre modèle est plus général. Nous montrons aussi comment calculer avec autant de précision voulue les coefficients de corrélation de Spearman et de Kendall d'un modèle PH bivarié ajusté à des données. Ces coefficients de corrélation peuvent

alors être comparés aux coefficients non paramétriques de Spearman et de Kendall basés sur les rangs. Des valeurs rapprochées des coefficients paramétriques et non paramétriques sont une indication de la validité du modèle. Un test d'ajustement convergent est construit en comparant la fonction de survie paramétrique (bivariée) ajustée avec la fonction de survie expérimentale au travers d'une statistique de type Cramér-von Mises. Ces résultats sont utilisés pour ajuster une distribution *PH* bivariée à un véritable jeu de données issu du domaine de l'assurance avec pour variables la perte subie (LOSS) et la dépense pour perte allouée après rajustement (ALAE). C'est la première fois à notre connaissance que les distributions *PH* bivariées sont utilisées sur de vraies données. La distribution *PH* bivariée ajustée est ensuite utilisée pour calculer la moyenne et les quantiles de la distribution conditionnelle de la variable ALAE étant donné une valeur de l'autre variable LOSS.

Le deuxième article étend le théorème de Jewell en théorie de la crédibilité à une grande classe de distributions qui sort des distributions exponentielles linéaires et même de la famille exponentielle. Le théorème de Jewell montre que la crédibilité exacte se produit dans la famille exponentielle linéaire univariée et multivariée de distributions conditionnelles, une fois appareillées à la distribution a priori conjuguée appropriée. La crédibilité exacte est étudiée ici dans le cas de distributions *PH* univariées et multivariées. Les chaînes de Markov sous-jacentes sont utilisées, en incluant les paramètres de risque non-observables pour les distributions *PH*.

MOTS CLÉS :

Distributions phase-type ; processus de Markov continu ; chaîne de Markov cachée ; algorithme EM ; bootstrap paramétrique ; crédibilité exacte ; distributions coxiennes ; théorème de Jewell.

SUMMARY

Phase-type (PH) distributions are used as probability models for positive random variables. Their origin stems from the works of Neuts published in the early eighties. The first applications are found in operational research as models for waiting times in the field of queuing theory. Probability models in actuarial science are also fraught with positive variables such as losses and survival times which may explain the recent interest of actuaries in PH distributions. Statistical estimation of PH distributions with the EM algorithm was developed in the mid nineties by Asmussen and his coworkers. Actuaries have also applied this class of models to risk theory and ruin probabilities. Extensions to multivariate PH distributions were also developed in the eighties following the seminal work of Neuts. They can serve as models for the joint probability distribution of two or more positive random variables.

The first paper treats of maximum likelihood estimation by the EM algorithm and goodness-of-fit tests by parametric bootstrap when the model is a bivariate PH distribution. Åhlström and his coworkers published in 1999 an EM algorithm for the parametric estimation of relapse time distributions in survival analysis. They used a bivariate PH distribution with one component greater than the other component with probability one. Although the EM method proposed in this thesis is similar, our model is more general. Moreover, we show how to evaluate with any desired degree of accuracy the Spearman or Kendall corrrelation coefficients of the fitted bivariate PH model. These correlation coefficients can then be compared with the non parametric Spearman or Kendall coefficients based on ranks. A close agreement is an indication of the validity of the model. A consistent goodness-of-fit testing procedure is proposed which compares the fitted (bivariate)

parametric survival function with the empirical survival function using a statistic of the Cramér-von Mises type. A parametric bootstrap algorithm is also provided to obtain the critical region of the proposed test. The results are used to fit real data in the insurance industry relating losses (LOSS) and allocated loss adjustment expenses (ALAE). To our knowledge this is the first time that bivariate *PH* distributions are used to fit real data. The fitted bivariate *PH* distribution is used to obtain the quantiles and the mean of the conditional distribution of the variable ALAE for a given value of the other variable LOSS.

The second paper extends Jewell's theorem in credibility theory to a larger class of distributions, outside of exponential distributions or even the linear exponential family. Jewell's Theorem proves that exact credibility occurs in the univariate and multivariate linear exponential family of conditional distributions, when paired with the appropriate conjugate prior distribution. Here, exact credibility is discussed in a univariate and multivariate *PH* setting. Hidden Markov chains are used, embedding the unobservable risk parameters in the *PH* distributions.

KEY WORDS :

Phase-type distributions; continuous Markov processes; hidden Markov chain; EM algorithm; parametric bootstrap; exact credibility; coxian distributions; Jewell's theorem.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

To my family

Arash

Jino

and

Bayan

who offered me unconditional love and support throughout the course of this thesis.

REMERCIEMENTS

From the formative stages of this thesis, to the final draft, I owe an immense debt of gratitude to my supervisors, Martin Bilodeau and José Garrido. Their sound advice and careful guidance were invaluable throughout this entire thesis process. I also wish to thank Professor Bilodeau for financial support from his NSERC grant.

I would also like to thank to Professors Louis Doray, François Perron, Sabin Lessard, Charles Dugas, Manuel Morales, Richard Duncan, Christian Léger, Alejandro Murua, Roch Roy, Martin Goldstein and Anatole Joffe. All are faculty members of the Département de mathématiques et de statistique of Université de Montréal. Without their time and cooperation this thesis would not have been possible.

For their efforts and assistance, special thanks go as well to my friends Hirbood Assa, Eugene Ursu, Zeinab Mashreghi, Romuald Hervé Momeya Ouabo and Mohammad Bardestani.

Finally, I cannot conclude without mentioning my brother Abdollah and his family, whose extreme generosity will be remembered always.

Chapitre 1

PRELIMINARY NOTIONS ON PHASE-TYPE DISTRIBUTIONS AND CREDIBILITY THEORY

1.1. PHASE-TYPE DISTRIBUTIONS

1.1.1. Sub-intensity matrices and matrix exponentials

Suppose that the real matrix $\mathbf{A} = (a_{ij})_{i,j\in E}$, for $E = \{1, \ldots, l\}$, has eigenvalues $\theta_i = \theta_i(\mathbf{A})$, $i = 1, \ldots, l$. Assume that $|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_l|$. The following result is easily proved.

Lemma 1.1. *If $\boldsymbol{B} = a\boldsymbol{A} + b$ for some real constants a and b, then*

$$\theta_i(\boldsymbol{B}) = a\theta_i(\boldsymbol{A}) + b.$$

The upper bound for the largest eigenvalue θ_1 for a nonnegative matrix is given in the following lemma.

Lemma 1.2. *Let $\boldsymbol{A} = (a_{ij})_{i,j\in E}$ be a nonnegative matrix, i.e. $a_{ij} \geq 0$ for all i, j, then,*

$$|\theta_1| \leq \min\left\{\max_{i\in E}\sum_{j=1}^{l} a_{ij}, \max_{j\in E}\sum_{i=1}^{l} a_{ij}\right\}.$$

Proof. If $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_l)^\top$ is the eigenvector associated with θ_1, then

$$\theta_1\phi_i = \sum_{j=1}^{l} a_{ij}\phi_j, \ i = 1, \ldots, l,$$

and as a result $|\theta_1||\phi_i| \leq \sum_{j=1}^{l} a_{ij} \max_{k\in E} |\phi_k|$, for $i = 1, \ldots, l$. Consequently,

$$|\theta_1| \max_{i\in E} |\phi_i| \leq \max_{i\in E} \sum_{j=1}^{l} a_{ij} \max_{k\in E} |\phi_k|.$$

Thus $|\theta_1| \leq \max_{i\in E} \sum_{j\in E} a_{ij}$. The proof simply uses the fact that $\mathbf{A}^\top$ has the same eigenvalues as $\mathbf{A}$. □

The density and Laplace transform of a phase-type distribution are expressed with a sub-intensity matrix. We present the required preliminaries of sub-intensity matrices:

Definition 1.1. *The matrix $\boldsymbol{T} = (t_{ij})_{i,j\in E}$, is called a sub-intensity or a sub-generator matrix if $t_{ij} \geq 0$, for $i \neq j$, $\sum_{j=1}^{l} t_{ij} \leq 0$, and for at least one $i \in E$, $\sum_{j=1}^{l} t_{ij} < 0$.*

The following lemma gives an upper bound for the real part of the maximal eigenvalue of a sub-intensity matrix. Let $\Re(z)$ denote the real part of the complex number z.

Proposition 1.1. *If $\boldsymbol{T} = (t_{ij})_{i,j\in E}$ is a sub-intensity matrix then, the eigenvalues $\theta_i(\boldsymbol{T}) = 0$ or $\Re(\theta_i(\boldsymbol{T})) < 0$, for $i = 1, \ldots, l$.*

Proof. Let $c > \max_{i\in E}(t_{ii})$. Then, $\mathbf{U} = \mathbf{T} + c\mathbf{I}$ is a nonnegative matrix. Applying Lemma 1.1 and Lemma 1.2 to the matrix $\mathbf{U}$ completes the proof. □

Because $\mathbf{T} = (t_{i,j})_{i,j\in E}$ is nonsingular if and only if 0 is not an eigenvalue then, we have the following corollary.

Corollary 1.1. *A sub-intensity matrix $\boldsymbol{T} = (t_{ij})_{i,j\in E}$ with eigenvalues θ_i is non-singular if and only if $\Re(\theta_i(\boldsymbol{T})) < 0$, for $i = 1, \ldots, l$.*

Definition 1.2. *The exponential of a square matrix $\boldsymbol{A}$ is defined as*

$$\exp(t\boldsymbol{A}) = \sum_{n=0}^{\infty} \frac{(t\boldsymbol{A})^n}{n!}, \ t \in \mathbb{R}. \tag{1.1}$$

In general, it is not an easy task to obtain the exponential of a matrix from the definition. There are many ways to calculate the exponential of a matrix. Moler and Van Loan (1978), examines many ways to calculate the exponential of a matrix to conclude that different forms of matrices require different approaches. The next theorem finds a representation formula for a matrix exponential.

Proposition 1.2. *Let $\theta_1, \ldots, \theta_l$ be the eigenvalues of $\boldsymbol{A} = (a_{ij})_{i,j\in E}$ then,*

$$\exp(t\boldsymbol{A}) = a_1(t)\boldsymbol{A}_1 + \cdots + a_l(t)\boldsymbol{A}_l \tag{1.2}$$

where, $a_k(t)$ and $\boldsymbol{A}_k$ are given recursively by

$$\begin{aligned} a_1(t) &= e^{\theta_1 t}, \\ \boldsymbol{A}_1 &= \boldsymbol{I}, \\ a_k(t) &= \int_0^t e^{\theta_k(t-x)} a_{k-1}(x)dx, \\ \boldsymbol{A}_k &= (\boldsymbol{A} - \theta_1\boldsymbol{I})\ldots(\boldsymbol{A} - \theta_{k-1}\boldsymbol{I}), \end{aligned}$$

for $k = 2, \ldots, l$.

For a proof see Rolski *et al.* (1999), p. 325. We have the following corollary from Proposition 1.2.

Corollary 1.2. *Let $\theta_1, \ldots, \theta_l$ be the eigenvalues of $\boldsymbol{A} = (a_{ij})_{i,j\in E}$. Then, for each $s > \max_{i\in E} \Re(\theta_i)$,*

$$\lim_{t\to\infty} e^{-st}\exp(t\boldsymbol{A}) = \boldsymbol{0}. \tag{1.3}$$

Proof. In view of (1.2), we have to show that

$$\lim_{t\to\infty} \exp(-st)|a_k(t)| = 0, \tag{1.4}$$

for $k = 1, \ldots, l$. Equation (1.4) is true for $k = 1$. Suppose that (1.4) holds for some $k = n - 1 < l$. As a result, for each $\varepsilon > 0$ there exists $\nu > 0$ such that $\exp(-sx)|a_{n-1}(x)| < \varepsilon$ for all $x > \nu$. On the other hand, we have that

$$\begin{aligned} e^{-st}|a_n(t)| \;\leq\; & e^{[\Re(\theta_n)-s]t}\int_0^\nu |e^{-\theta_n x}a_{n-1}(x)|dx \\ & + \int_\nu^t e^{[\Re(\theta_n)-s](t-x)}e^{-sx}|a_{n-1}(x)|dx. \end{aligned}$$

For a fixed point $\nu > 0$, $\lim_{t\to\infty} e^{[\Re(\theta_n)-s]t}\int_0^\nu |e^{-\theta_n x}a_{n-1}(x)|dx = 0$. This implies that the first integrand tends to zero. The second integrand is always less than $\frac{\varepsilon}{s-\Re(\theta_n)}$, and this completes the proof. □

Let $\mathbf{A}(h) = (a_{ij}(h))_{i,j\in E}$ be a matrix function such that all its entries are differentiable functions of h. The matrix derivative of $\mathbf{A}(h)$ is defined by

$$\frac{d\mathbf{A}(h)}{dh} = \lim_{\delta \to 0} \frac{\mathbf{A}(h+\delta) - \mathbf{A}(h)}{\delta}. \tag{1.5}$$

The following lemma states the derivative of a matrix exponential.

Lemma 1.3. *The matrix exponential $e^{\boldsymbol{A}h}$ is differentiable on the whole real line and*

$$\frac{de^{\boldsymbol{A}h}}{dh} = \boldsymbol{A}e^{\boldsymbol{A}h} = e^{\boldsymbol{A}h}\boldsymbol{A}. \tag{1.6}$$

One can prove this lemma by using (1.1) and (1.5). For a proof, see Rolski *et al.* (1999), p. 315. If $\mathbf{A}$ and $\mathbf{B}$ are both differentiable matrix functions then, a differential rule for the product of $\mathbf{A}$ and $\mathbf{B}$ is given in the following lemma.

Lemma 1.4. *If $\boldsymbol{A}(h)$ and $\boldsymbol{B}(h)$ are both differentiable matrix functions then,*

$$\frac{d}{dh}\left[\boldsymbol{A}(h)\boldsymbol{B}(h)\right] = \left[\frac{d}{dh}\boldsymbol{A}(h)\right]\boldsymbol{B}(h) + \boldsymbol{A}(h)\left[\frac{d}{dh}\boldsymbol{B}(h)\right]. \tag{1.7}$$

As in the definition of the matrix derivative, the integral $\int_\nu^t \mathbf{A}(x)dx$ is a matrix with elements $\int_\nu^t a_{ij}(x)dx$, for $\nu < t$. In particular, for the matrix exponential function we have the following lemma.

Lemma 1.5. *If $\boldsymbol{T}$ is a nonsingular matrix then,*

$$\int_\nu^t \exp(x\boldsymbol{T})dx = \boldsymbol{T}^{-1}[\exp(t\boldsymbol{T}) - \exp(\nu\boldsymbol{T})]. \tag{1.8}$$

Moreover, if all the eigenvalues of $\boldsymbol{T}$ have negative real parts then,

$$\int_0^\infty \exp(x\boldsymbol{T})dx = -\boldsymbol{T}^{-1}.$$

Proof. Equation (1.8) is a consequence of (1.6) and (1.7) which implies

$$\frac{d}{dx}\left(\mathrm{T}^{-1}e^{\mathbf{T}x}\right) = e^{\mathbf{T}x}.$$

Let $\nu = 0$ in (1.8) and let $s = 0$ in (1.3) to get $\lim_{t\to\infty}\exp(t\mathrm{T}) = \mathbf{0}$. □

1.1.2. Continuous-time Markov chains

Definition 1.3. *A stochastic process $\{J_t, t \geq 0\}$, defined on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, with values in a countable set E, called the space state of the process, is called a continuous Markov chain if for any finite set $0 \leq t_1 < t_2 < \cdots < t_n <$*

t_{n+1} of times and corresponding set $i_1, i_2, \ldots, i_{n-1}, i, j$ of states in E such that $\mathbb{P}(J_{t_n} = i, J_{t_{n-1}} = i_{n-1}, \ldots, J_{t_1} = i_1) > 0$, we have

$$\mathbb{P}(J_{t_{n+1}} = j | J_{t_1} = i_1, J_{t_2} = i_2, \ldots, J_{t_n} = i) = \mathbb{P}(J_{t_{n+1}} = j | J_{t_n} = i). \tag{1.9}$$

Equation (1.9) is called the Markov property. The Markov property states that in order to calculate the probability of a coming event given some past events, only the most recent past event is relevant. If for all s, t such that $0 \leq s \leq t$ and all $i, j \in E$, the conditional probability $\mathbb{P}(J_t = j | J_s = i)$ depends only on $t - s$, we say that the process $\{J_t, t \geq 0\}$ is homogeneous, or has stationary transition probabilities. In this case, $\mathbb{P}(J_t = j | J_s = i) = \mathbb{P}(J_{t-s} = j | J_0 = i)$ and the function

$$p_{ij}(t) = \mathbb{P}(J_t = j | J_0 = i), \ i, j \in E, \ t \geq 0$$

is called the transition function of the process. All continuous Markov chains discussed in this thesis have stationary transition probabilities. The finite-dimensional probabilities of the process $\{J_t, t \geq 0\}$ can be obtained in terms of the transition function $p_{ij}(t)$ and the initial probability distribution $\alpha_i = \mathbb{P}(J_0 = i)$, $i \in E$. In fact, we have

$$\begin{aligned}
& \mathbb{P}(J_{t_n} = i_n, J_{t_{n-1}} = i_{n-1}, \ldots, J_{t_1} = i_1) \\
= \ & \mathbb{P}(J_{t_n} = i_n | J_{t_{n-1}} = i_{n-1}, \ldots, J_{t_1} = i_1) \mathbb{P}(J_{t_{n-1}} = i_{n-1}, \ldots, J_{t_1} = i_1) \\
= \ & p_{i_{n-1}, i_n}(t_n - t_{n-1}) \mathbb{P}(J_{t_{n-1}} = i_{n-1}, \ldots, J_{t_1} = i_1) \\
& \vdots \\
= \ & \sum_{i_0 \in E} \alpha_{i_0} \prod_{m=1}^{n} p_{i_{m-1}, i_m}(t_m - t_{m-1}),
\end{aligned}$$

where $t_0 = 0$. As the transition function is also a conditional probability, it satisfies the following property :

$$p_{ij}(t) \geq 0, \text{ for all } i, j \in E \text{ and } \sum_{j \in E} p_{ij}(t) = 1. \tag{1.10}$$

It also satisfies

$$p_{ij}(0) = \mathbb{P}(J_0 = j | J_0 = i) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

where δ_{ij} is the Kronecker delta. Finally, for all $s, t \geq 0$, $i, j \in E$,

$$p_{ij}(t+s) = \sum_{k \in E} p_{ik}(s) p_{kj}(t). \tag{1.11}$$

Equation (1.11) is called the Chapman-Kolmogorov equation. For a proof, see Anderson (1991). In order for the chain to go from state i to state j in time $t+s$, it must be in some state at time s. The Chapman-Kolmogorov equation is obtained by using the Markov property and conditioning on the state at time s. The Chapman-Kolmogorov equation shows that if the transition function is known on some interval, $0 < t < t_0$, it is known for all $t > 0$. This fact suggests that the transition probabilities can be determined from their derivatives at 0. If, for $i \neq j$, the limit

$$q_{ij} = \lim_{h \to 0} \frac{p_{ij}(h)}{h}$$

exists then, q_{ij} is called the jump rate from state i to state j. Note that the jump rate is not a function of t. This limit exists for all the cases considered in this thesis. In general, it is possible to construct Markov chains based on jump rates, see Durrett (1999).

In the following, it is shown how to compute the transition probabilities from the jump rates. Using the Chapman-Kolmogorov equation (1.11),

$$\frac{p_{ij}(t+s) - p_{ij}(t)}{s} = \sum_{k \in E, k \neq i} \frac{p_{ik}(s)}{s} p_{kj}(t) + \frac{p_{ii}(s) - 1}{s} p_{ij}(t). \tag{1.12}$$

Using (1.10), note that $1 - p_{ii}(s) = \sum_{k \neq i} p_{ik}(s)$, so that

$$\lim_{s \to 0} \frac{p_{ii}(s) - 1}{s} = -\lim_{s \to 0} \sum_{k \neq i} \frac{p_{ik}(s)}{s} = -\sum_{k \neq i} q_{ik} \equiv -\lambda_i.$$

The limit as s goes to 0 of (1.12) is

$$p_{ij}'(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) - \lambda_i p_{ij}(t). \tag{1.13}$$

Introducing the matrix $\mathbf{Q} = (q_{ij})$, where $q_{ii} = -\lambda_i$, equation (1.13) may be rewritten in matrix form for $p(t) = (p_{ij}(t))_{ij}$ as

$$p'(t) = \mathbf{Q} p(t).$$

This is Kolmogorov's backward equation. This equation solves for

$$p(t) = e^{\mathbf{Q}t}.$$

The matrix $\mathbf{Q}$ is called the infinitesimal generator of the Markov chain. Anderson (1991) gives a rigorous account of continuous-time Markov chains.

1.1.3. Definition of phase-type distribution

A random variable that is defined as the absorption time of an evanescent finite-state continuous-time Markov chain is said to have a phase-type (PH) distribution. The distribution and density functions of a PH distribution can be expressed in terms of the $m \times 1$ initial state distribution vector $\boldsymbol{\pi}$ and the $m \times m$ infinitesimal generator matrix $\mathbf{T}$ of the underlying Markov chain. The pair $(\boldsymbol{\pi}, \mathbf{T})$ is known as a representation of order m of the PH distribution. Since their introduction by Neuts (1981), PH distributions have been used in a wide range of stochastic modeling applications in areas as diverse as telecommunications, teletraffic modeling, biostatistics, queueing theory, risk theory, reliability theory, and survival analysis. Erlang (1917) wrote the first paper to extend the familiar exponential distribution with his method of stages. He defined a nonnegative random variable as the time taken to move through a fixed number of stages, spending an exponential amount of time with a fixed positive rate in each one. Nowadays, we refer to distributions defined in this manner as Erlang distributions. Cox (1955) generalized Erlang's notion by allowing complex parameters. This construction, defines the class of distributions with rational Laplace-Stieltjes transforms (LST), of which the class of PH distributions is a proper subset. These distributions are nowadays also known as matrix-exponential distributions. Neuts (1981) generalized Erlang's method of stages in a different direction. He defined a phase-type random variable as the time taken to progress through the states of a finite-state evanescent continuous-time Markov chain, spending an exponential amount of time with a positive rate in each one, until absorption.

PH distributions have many appealing features. In the following some of them are listed.

(1) They are dense in the class of all distributions defined on the nonnegative real numbers.

(2) The use of *PH* distributions in stochastic models often enables algorithmically tractable solutions to be found. Quantities of interest, such as the distribution and density functions, the Laplace-Stieltjes transform, and the moments of *PH* distributions are expressed simply in terms of the initial phase distribution π and the exponential or powers of the infinitesimal generator $\mathbf{T}$.

(3) Stochastic models, particularly where the exponential distribution is used to model quantities (such as inter-arrival times, service times, or lifetimes), can now be extended with *PH* distributions.

(4) Since the class of *PH* distributions is closed under a variety of operations, such as finite mixtures and convolutions, systems with *PH* inputs often have *PH* outputs.

Neuts and his coworkers, in the late seventies, established much of this modern theory. Neuts (1995) developped queuing theory by *PH* distributions and Asmussen (2000) applied *PH* distributions to risk theory.

Let $\{J_t, t \geq 0\}$ be a Markov process in the finite state space

$$E = \{1, \ldots, m, m+1\},$$

where $1, 2, \ldots, m$ are transient and, thus, $m+1$ is an absorbent state. Then, $\{J_t, t \geq 0\}$ has an infinitesimal generator of the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0}^\top & 0 \end{pmatrix}, \tag{1.14}$$

where $\mathbf{T} = (t_{ij})_{i,j=1,\ldots,m}$ is an $m \times m$ matrix, $\mathbf{t} = (t_i)_{i=1,\ldots,m}$ is an m dimensional column vector and $\mathbf{0}$ is an m dimensional vector of zeros. Superscript $\top$ stands for the transpose of a matrix or vector. Since the rows of an infinitesimal generator must sum to zero, note that $\mathbf{t} = -\mathbf{T}\mathbf{e}$, where $\mathbf{e} = (1, 1, \ldots, 1)^\top$ is the vector of ones. Let $\alpha_i = \mathbb{P}(J_0 = i)$ be the initial state probabilities. Often, it is assumed that the chain does not start in the absorbent state, *i.e.* $\alpha_{m+1} = 0$. In that case,

$\boldsymbol{\alpha}$ can be written as

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\pi} \\ 0 \end{pmatrix}.$$

This condition is assumed in the next definition.

Definition 1.4. *The time until absorption,*

$$X = \inf\{t \geq 0 | J_t = m + 1\} \tag{1.15}$$

is said to have a PH distribution with representation or parameters $(\boldsymbol{\pi}, \boldsymbol{T})$.

The dimension m of $\boldsymbol{\pi}$ is said to be the dimension of the phase-type distribution.

1.1.4. Properties of phase-type distributions

The first property relates to the exponential of the infinitesimal generator matrix.

Proposition 1.3. *Assume the representation (1.14) of the infinitesimal generator. Then,*

$$e^{\boldsymbol{A}s} = \begin{pmatrix} e^{\boldsymbol{T}s} & \boldsymbol{e} - e^{\boldsymbol{T}s}\boldsymbol{e} \\ \boldsymbol{0}^\top & 1 \end{pmatrix}.$$

Proposition 1.4. *If X has a PH distribution with parameters $(\boldsymbol{\pi}, \boldsymbol{T})$ then, the density of X is given by*

$$f(x) = \boldsymbol{\pi}^\top e^{\boldsymbol{T}x}\boldsymbol{t}.$$

Proof : Let $p_{ij}(x) = \mathbb{P}(J_x = j | J_0 = i)$. Then,

$$1 - F(x) = \mathbb{P}(X \geq x) = \sum_{i=1}^{m}\sum_{j=1}^{m} \pi_i p_{ij}(x) = \sum_{i=1}^{m}\sum_{j=1}^{m} \pi_i (e^{\mathbf{T}x})_{ij} = \boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{e},$$

where $(e^{\mathbf{T}x})_{ij}$ is the $ij-$element of the $e^{\mathbf{T}x}$. The density function is the derivative of the cumulative distribution function. Hence, $f(x) = \boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{t}$. □

Proposition 1.5. *The Laplace transform of X is given by*

$$L(s) = E(e^{-sX}) = \boldsymbol{\pi}^\top(-s\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{t},$$

where $\boldsymbol{I}$ is identity matrix of dimension m.

From the definition of the inverse of a matrix, one can write the LST as a ratio of polynomials. The maximum degree of the denominator is m and the degree of the numerator is less than m (because the limit of the Laplace transform as s goes to $-\infty$ is zero). Then, it can be written as follows

$$L(s) = \frac{1 + c_1 s + \cdots + c_{m-1} s^{m-1}}{1 + d_1 s + \cdots + d_m s^m}.$$

Phase-type distributions not only have rational Laplace transforms, but also with some conditions all distributions with rational Laplace transform are of phase-type. See the following theorem of O'Cinneide (1989).

Proposition 1.6. *A distribution defined on $[0, \infty)$ is a PH distribution if and only if*

(1) it is the point mass at zero, or

(2) **a):** *it has a strictly positive density on $(0, \infty)$, and*

b): *it has a rational Laplace transform such that there exists a pole of maximal real part, $-\gamma$, that is real, negative, and such that $-\gamma > \Re(-\xi)$, where $\Re(-\xi)$ is the real part of any other pole.*

Proposition 1.7. *The moment of order n, $n \geq 1$, of X is given by*

$$\mathbb{E}[X^n] = (-1)^n n! \boldsymbol{\pi} \boldsymbol{T}^{-n} \boldsymbol{e}.$$

As a consequence, the full class of *PH* distributions of order m has a parametrization in $2m - 1$ dimensions. This follows from the Cayley-Hamilton theorem from which there is at least one sequence $\lambda_0, \lambda_1, \ldots, \lambda_{m-1}$ such that

$$\mathbf{T}^{-m}\mathbf{e} = \sum_{i=0}^{m-1} \lambda_i \mathbf{T}^{-i}\mathbf{e}.$$

If we fix such a sequence then, those coefficients together with the first $m - 1$ moments determine all the moments recursively. Indeed, it can be seen by pre-multiplying the relation above by $\boldsymbol{\pi}\mathbf{T}^{-n}$ that

$$\frac{(-1)^{n+m}\mathbb{E}[X^{n+m}]}{(n+m)!} = \sum_{i=0}^{m-1} \lambda_i \frac{(-1)^{n+i}\mathbb{E}[X^{n+i}]}{(n+i)!}.$$

Since the Laplace transform near zero is determined by all the moments, it follows that $\lambda_0, \lambda_1, \ldots, \lambda_{m-1}$ and $\mathbb{E}[X], \ldots, \mathbb{E}[X^{m-1}]$ determine the distribution.

When the representation of the *PH* distribution is estimated, overparametrization will occur. From the latter discussion, a *PH* distribution of dimension m has a parametrization with $2m-1$ parameters. However, direct estimation of the representation $(\boldsymbol{\pi}, \mathbf{T})$ requires m^2+m-1 parameters to be estimated.

Neuts (1981) showed that the convolution of two independent *PH* variables with possibly different dimensions is a *PH* variable.

Proposition 1.8. *Suppose that F and G are PH distributions with representations $(\boldsymbol{\alpha}, \boldsymbol{T})$ of order m and $(\boldsymbol{\beta}, \boldsymbol{S})$ of order n, respectively. Then, their convolution $F * G$ is a PH distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{R})$ of order $m+n$, where*

$$\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\alpha} \\ F(0)\boldsymbol{\beta} \end{pmatrix},$$

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{T} & -\boldsymbol{t}\boldsymbol{\beta}^{\top} \\ 0 & -\boldsymbol{S} \end{pmatrix},$$

and $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$.

Note that $F(0)$ is the probability that the chain associated with F starts in the absorbent state, *i.e.* $F(0) = \alpha_{m+1}$. It is also easily proved that a finite mixture of *PH* distributions follows a *PH* distribution.

Proposition 1.9. *If $(p_1; \ldots, p_k)$ is the vector of mixing probabilities and F_j is a PH distribution with representation $(\boldsymbol{\pi}_j, \boldsymbol{T}_j)$, $1 \leq j \leq k$, then, the mixture has the representation with initial state probabilities*

$$\begin{pmatrix} p_1\boldsymbol{\pi}_1 \\ \vdots \\ p_k\boldsymbol{\pi}_k \end{pmatrix}$$

and infinitesimal generator

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{T}_k \end{pmatrix}.$$

For the next theorem we need the following definition.

Definition 1.5. *A PH distribution is called triangular phase-type, or TPH, if it has a representation in which the matrix* $\boldsymbol{T}$ *is of infinitesimal generator upper triangular.*

The most important property of *TPH* distributions found in Assaf and Levikson (1982) is that absorption happens in a bounded number of transitions.

Proposition 1.10. *The TPH class of distributions is the smallest class containing all exponential distributions and which is closed under finite mixtures, finite convolutions and formation of coherent systems.*

1.1.5. Some examples of *PH* distributions

In the following, some well-known probability density functions are represented by phase-type distributions. More examples can be found in Fackrell (2003).

Example 1.1. The exponential distribution with density function $f(x) = \lambda e^{-\lambda x}$ has the representation

$$\begin{aligned} \boldsymbol{\pi} &= 1, \\ \mathbf{T} &= -\lambda. \end{aligned}$$

Example 1.2. The hyper-exponential distribution with probability density function

$$f(x) = \sum_{i=1}^{n} \alpha_i \lambda_i e^{-\lambda_i x}$$

has the representation

$$\begin{aligned} \boldsymbol{\pi} &= (\alpha_1, \ldots, \alpha_n)^{\top}, \quad x > 0, \\ \mathbf{T} &= \begin{pmatrix} -\lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\lambda_n \end{pmatrix}. \end{aligned}$$

Example 1.3. The m-phase Erlang distribution with density function

$$f(x) = \frac{\lambda^m x^{m-1} e^{-\lambda x}}{(m-1)!}, \qquad x > 0,$$

has the representation

$$\begin{aligned}\boldsymbol{\pi} &= (1,0,\ldots,0)^{\top},\\ \mathbf{T} &= \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0\\ 0 & -\lambda & \lambda & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -\lambda & \lambda\\ 0 & 0 & 0 & \cdots & 0 & -\lambda \end{pmatrix},\end{aligned}$$

where m is the dimension of Matrix $\mathbf{T}$.

Example 1.4. A PH distribution is unicycle if it has a representation of the form

$$\begin{aligned}\boldsymbol{\pi} &= (\alpha_1,\alpha_2,\ldots,\alpha_m)^{\top},\\ \mathbf{T} &= \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0\\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -\lambda_{m-1} & \lambda_{m-1}\\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{m-1} & -\lambda_m \end{pmatrix},\end{aligned}$$

where, for $i = 1,\ldots,m-1$, $\mu_i \geq 0$, $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_m$ and $\lambda_m > \sum_{i=1}^{m-1} \mu_i$. In the next example it is exemplified that the representation of a phase-type distribution may not be unique.

Example 1.5. In the following example from Botta *et al.* (1987) the non-uniqueness of the representation happens even for the minimal order or dimension of a PH distribution. All the next three representations lead to the same PH probability density function

$$f(x) = \frac{1}{3}2e^{-2x} + \frac{2}{3}5e^{-5x},\ x > 0.$$

The representations are

$$\boldsymbol{\pi} = (1/3, 2/3)^{\top} \text{ and } \mathbf{T} = \begin{pmatrix} -2 & 0\\ 0 & -5 \end{pmatrix}, \tag{1.16}$$

$$\boldsymbol{\pi} = (1/5, 4/5)^\top \text{ and } \mathbf{T} = \begin{pmatrix} -2 & 2 \\ 0 & -5 \end{pmatrix}, \tag{1.17}$$

and

$$\boldsymbol{\pi} = (0, 1/2, 1/2)^\top \text{ and } \mathbf{T} = \begin{pmatrix} -3 & 1 & 1 \\ 1 & -4 & 2 \\ 1 & 0 & -6 \end{pmatrix}. \tag{1.18}$$

In the representations (1.16) and (1.17), the order of the *PH* distribution is 2, whereas it is 3 in (1.18). In general, a representation which has the minimum order is called minimal. From this example, even the minimal one is not unique. The order of a *PH* distribution is defined as the order of the minimal one

Example 1.6. A PH distribution is said to be acyclic if its matrix $\mathbf{T}$ is upper triangular.

Example 1.7. A PH distribution is said to be Coxian of order p if

$$\boldsymbol{\pi} = (1, 0, \ldots, 0)^\top,$$

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & q_1\lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & q_2\lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_{p-1} & q_{p-1}\lambda_{p-1} \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_p \end{pmatrix},$$

where $0 < q_i < 1$ and $\lambda_i > 0$, $i = 1, \ldots, p$.

As in Proposition 1.6, we have the following theorem about the characterization of Coxian distributions.

Proposition 1.11. *A distribution defined on* $[0, \infty)$ *is a Coxian distribution if and only if*

(1) it is the point mass at zero, or

(2) **a):** *it has a strictly positive density on* $(0, \infty)$*, and*

b): *it has a rational LST with only real and negative poles.*

See O'Cinneide (1991) for a proof. A remarkable result from Cumani (1982) and Dehon and Latouche (1982) establishes that a *PH* variable having an acyclic

Markov chain representation can be uniquely represented by a Coxian distribution with stochastically increasing states, *i.e.* $-\lambda_1 \leq -\lambda_2 \leq \cdots \leq -\lambda_p$. Such processes start at state 1 and can only jump from i to $i+1$ or $p+1$. Therefore, the true parameter dimension is $2p-1$, where p is the dimension of the acyclic PH distribution.

1.1.6. EM algorithm for PH distributions

The EM (Expectation-Maximization) algorithm of Dempster *et al.* (1977) is a general iterative method for finding the maximum-likelihood estimate of the parameters, when the data is incomplete or has missing values. It finds its usefulness when the likelihood function of the incomplete (observed) data is intractable but that of the complete (unobserved or missing) data is of a simpler form which can be analytically optimized. Assume that the data $\mathbf{x}$ is observed and generated by some distributions, say $f(\mathbf{x}|\boldsymbol{\phi})$ with log-likelihood function $l(\boldsymbol{\phi}) = \log f(\mathbf{x}|\boldsymbol{\phi})$. We call $\mathbf{x}$ the incomplete data and refer to $l(\boldsymbol{\phi})$ as the incomplete log-likelihood function. Suppose that an unobserved (complete) data $\mathbf{y}$, where $\mathbf{x} = \mathbf{x}(\mathbf{y})$, has pdf $g(\mathbf{y}|\boldsymbol{\phi})$. Assume

$$Q(\boldsymbol{\phi}'|\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{\phi}}\left[\log g(\mathbf{y}|\boldsymbol{\phi}')|\mathbf{x}\right]$$

exists for all pairs $(\boldsymbol{\phi}', \boldsymbol{\phi})$. The EM iteration $\boldsymbol{\phi}^{(p)} \rightarrow \boldsymbol{\phi}^{(p+1)}$ is defined as follows :
E-step : Compute $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$.
M-step : Find $\boldsymbol{\phi}^{(p+1)}$ that maximizes $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$ over $\boldsymbol{\phi}$.

Simplifications occur when the complete data density function is a member of the exponential family

$$g(\mathbf{y}|\boldsymbol{\phi}) = b(\mathbf{y}) \exp\left[\boldsymbol{\phi}^{\mathsf{T}}\mathbf{t}(\mathbf{y})\right] / a(\boldsymbol{\phi}), \tag{1.19}$$

where $\boldsymbol{\phi}$ is the vector parameter, $\mathbf{t}(\mathbf{y})$ is the vector of the complete data sufficient statistic. If (1.19) holds, simplified expressions found in Dempster *et al.* (1977) for the E and M steps are : E-step : Estimate the complete data sufficient statistics $\mathbf{t}(\mathbf{x})$ by finding

$$\mathbf{t}^{(p)} = \mathbb{E}_{\boldsymbol{\phi}^{(p)}}\left[\mathbf{t}(\mathbf{y})|\mathbf{x}\right].$$

M-step : Determine $\boldsymbol{\phi}^{(p+1)}$ as the solution for $\boldsymbol{\phi}$ of the equation

$$\mathbb{E}_{\boldsymbol{\phi}}\left[\mathbf{t}(\mathbf{y})\right] = \mathbf{t}^{(p)}.$$

By definition a *PH* random variable is the time until absorption in an absorbent state. This can be considered as an incomplete data in the sense that they only provide information about the absorption time of the absorbent state, not about the whole path of the underlying Markov chain, J_t. The initial state, the states that have been visited, and the time spent in each visited state are not observed. Hence, the hidden information can help to maximize the incomplete likelihood function which is untractable. For an observation x from a *PH* variables as defined in (1.15) with representation $(\boldsymbol{\pi}, \mathbf{T})$, the complete information is formulated by the embedded Markov chain of visited states

$$i_0, i_1, \ldots, i_{k-1}, i_k(= m+1),$$

and the sojourn times

$$s_0, s_1, \ldots, s_{k-1}, s_k(= \infty),$$

where k is the number of jumps until hitting $m+1$. A complete observation of the process J_t on the interval $(0, x]$ is represented by

$$\mathbf{y} = (i_0, \ldots, i_{k-1}, s_0, \ldots, s_{k-1}),$$

where $x = s_0 + \cdots + s_{k-1}$. To get the probability density function of $\mathbf{y}$, one needs the probability p_{jl} of jumping from j to l which is given by

$$p_{jl} = \mathbb{P}(i_{n+1} = l | i_n = j) = \begin{cases} 0, & j, l = 1, \ldots, m, j = l, \\ \frac{t_{jl}}{-t_{jj}}, & j, l = 1, \ldots, m, j \neq l, \\ \frac{t_j}{-t_{jj}}, & j = 1, \ldots, m, l = m+1. \end{cases}$$

The density of $\mathbf{y}$ can be derived by Markov chain properties and considering that the time spent in each state i has an exponential distribution with mean $1/\lambda_i$, where $\lambda_i = -t_{ii}$, as in Asmussen *et al.* (1996). Thus,

$$g(\mathbf{y}|\boldsymbol{\theta}) = \pi_{i_0} \exp\{-\lambda_{i_0} s_0\} t_{i_0, i_1} \ldots \exp\{-\lambda_{i_{k-1}} s_{k-1}\} t_{i_{k-1}}, \tag{1.20}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T})$ is the parameter.

Let $J_t^{[1]}, \ldots, J_t^{[n]}$ be n independent realizations of the process. This gives n embedded Markov chains

$$i_0^{[\nu]}, \ldots, i_{k^{[\nu]}-1}^{[\nu]},$$

with corresponding holding times

$$s_0^{[\nu]}, \ldots, s_{k^{[\nu]}-1}^{[\nu]}.$$

The complete data becomes $\mathbf{y} = (\mathbf{y}^{[1]}, \ldots, \mathbf{y}^{[n]})$, where

$$\mathbf{y}^{[\nu]} = (i_0^{[\nu]}, \ldots, i_{k^{[\nu]}-1}^{[\nu]}, s_0^{[\nu]}, \ldots, s_{k^{[\nu]}-1}^{[\nu]}), \ \nu = 1, \ldots, n.$$

The observed incomplete data is the following function of the complete data

$$\mathbf{x}^{[\nu]} = (s_0^{[\nu]} + \cdots + s_{k^{[\nu]}-1}^{[\nu]}).$$

Define

$$\begin{aligned} B_i^{[\nu]} &= \mathbf{1}\left\{i_0^{[\nu]} = i\right\}, \\ Z_i^{[\nu]} &= \sum_{\kappa=0}^{k^{[\nu]}-1} \mathbf{1}\left\{i_\kappa^{[\nu]} = i\right\} s_\kappa^{[\nu]}, \\ N_{ij}^{[\nu]} &= \sum_{\kappa=0}^{k^{[\nu]}-1} \mathbf{1}\left\{i_\kappa^{[\nu]} = i, i_{\kappa+1}^{[\nu]} = j\right\}. \end{aligned}$$

Let

$$\begin{aligned} B_i &= \sum_{\nu=1}^{n} B_i^{[\nu]}, \\ Z_i &= \sum_{\nu=1}^{n} Z_i^{[\nu]}, \\ N_{ij} &= \sum_{\nu=1}^{n} N_{ij}^{[\nu]}, \end{aligned}$$

be the number of Markov processes starting from state i, the total time spent in each state i, and the number of jumps from state i to state j, respectively. Then, the density of the complete data $\mathbf{y}$ is the product of n densities as in (1.20)

$$g(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{m} \left[\pi_i^{B_i} \exp\{t_{ii} Z_i\} \prod_{j=1, j\neq i}^{m+1} t_{ij}^{N_{ij}} \right], \tag{1.21}$$

where $t_{i,m+1} = t_i$. The density (1.21) is a member of a curved multi-parameter exponential family with sufficient statistics

$$B_i,\ Z_i,\ N_{ij},$$

where $i = 1, \ldots, m$, $j = 1, \ldots, m+1$, $i \neq j$. The M-step is given by

$$\hat{\pi}_i = \frac{B_i}{n},\ \hat{t}_{ij} = \frac{N_{ij}}{Z_i},\ i \neq j,\ \hat{t}_i = \frac{N_{i,m+1}}{Z_i},\ \hat{t}_{ii} = -\left(\hat{t}_i + \sum_{j=1, j\neq i}^{m} \hat{t}_{ij}\right).$$

The E-step for an exponential family consists of computing the conditional expectation of the sufficient statistics, given the complete data and the current parameter estimates. If the current parameter estimates at step h of the algorithm is $\boldsymbol{\theta}^{(h)}$, the complete sufficient statistics at the $h+1$ E-step consists in the evaluation of the following conditional expectations

$$\begin{aligned}
B_i^{(h+1)} &= \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[B_i^{[\nu]} | \mathbf{x}^{[\nu]}\right], \\
Z_i^{(h+1)} &= \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[Z_i^{[\nu]} | \mathbf{x}^{[\nu]}\right], \\
N_{ij}^{(h+1)} &= \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[N_{ij}^{[\nu]} | \mathbf{x}^{[\nu]}\right],
\end{aligned}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, m+1$, $i \neq j$. The E-step is the most complicated step. It is given in details in Asmussen *et al.* (1996). Define

$$C(a, b, i, j, \mathbf{T}) = \int_a^b e^{\mathbf{T}(u-a)} \mathbf{E}_{ij} e^{\mathbf{T}(b-u)} du,$$

where $\mathbf{E}_{ij}$ is an $m \times m$ matrix with a one in position (i, j) and zeros elsewhere. The conditional expectations are given as follows :

$$\mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[B_i^{[\nu]} | x^{[\nu]}\right] = \frac{\pi_i^{(h)} \mathbf{e}_i^{\top} e^{\mathbf{T}^{(h)} x^{[\nu]}} \mathbf{T}\mathbf{e}}{f(x^{[\nu]} | \boldsymbol{\theta}^{(h)})},$$

where $\pi_i^{(h)}$ is the ith component of $\boldsymbol{\pi}^{(h)}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[Z_i^{[\nu]}|x^{[\nu]}\right] &= \frac{\boldsymbol{\pi}^{(h)\top}C(0,x^{[\nu]},i,i,\mathbf{T}^{(h)})\mathbf{e}}{f(x^{[\nu]}|\boldsymbol{\theta}^{(h)})}, \\
\mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[N_{ij}^{[\nu]}|x^{[\nu]}\right] &= t_{ij}\frac{\boldsymbol{\pi}^{(h)\top}C(0,x^{[\nu]},i,j,\mathbf{T}^{(h)})\mathbf{e}}{f(x^{[\nu]}|\boldsymbol{\theta}^{(h)})}, \\
\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[N_{i,m+1}^{[\nu]}|x^{[\nu]}\right] &= -t_i\frac{\boldsymbol{\pi}^{(h)\top}e^{\mathbf{T}^{(h)}x^{[\nu]}}\mathbf{T}^{(h)}\mathbf{e}_i}{f(x^{[\nu]}|\boldsymbol{\theta}^{(h)})},
\end{aligned}$$

where $i,j = 1,\ldots,m$, $i \neq j$. The function $C(a,b,i,j,\mathbf{T})$ can be evaluated by numerical methods in ordinary or partial differential equations such as the Runge-Kutta method of order four. More detailed useful numerical methods can be found in Asmussen *et al.* (1996).

1.2. Credibility Theory

Generally speaking, credibility theory is a quantitative tool that allows an insurer to combine the past experience of a policyholder to the pure premium in a risk class or group of risk classes. If the past observed experience of the policyholder indicates a difference in risk to that assumed for the class, then the insurer has to explore this difference to see if it is due to a really different policyholder or it is only due to natural stochastic variation in the risk class. If the policyholder is indeed different, then some credible information can be obtained from the individual experience which is not being considered when the pure premium or manual premium is calculated. In other words, the assumption of homogeneity in the risk class fails.

For example, in car insurance the insurer may assume that the number of accidents in one year follows a Poisson distribution with mean μ, but then the experience of a particular policyholder may have an average $\bar{X}$ that is far from μ. In statistical term, $\bar{X}$ would show a significant difference with μ. In this case, the insurer must consider two facts :

(1) The risk class is not homogeneous. Its heterogeneity should be taken into account.

(2) What share of this difference is due to heterogeneity and to natural random variation?

To combine these two facts, the credibility premium, P_c should be a combination of the manual rate M and the past individual observation summary $\bar{X}$. A very good candidate for P_c is

$$P_c = Z\bar{X} + (1 - Z)M\,, \tag{1.22}$$

where the credibility factor $Z \in [0, 1]$ should be determined. Full credibility occurs when $Z = 1$. Section 1.2.1 deals with limited fluctuations credibility theory which developed at the beginning of twentieth century and represented a first attempt to model credibility in practical situations. The problem with this approach is its lack of a sound underlying mathematical theory justifying the method. As a result, greatest accuracy credibility was developed. It is introduced in Section 1.2.2. This method provides a statistical framework for credibility theory, where the risk parameter has a prior distribution modelling the possible heterogeneity within the portfolio. Both, the classical models of Bühlamnn and Bühlamnn-Straub will be discussed.

1.2.1. Limited fluctuations credibility theory

Limited fluctuations credibility theory was developed in the early part of twentieth century as the first attempt to give quantitative credibility rules. Suppose that $X_1, \ldots, X_n$ represent the past claim experience of a policyholder and are *i.i.d.*, with theoretical mean μ and variance σ^2. The variance of $\bar{X} = \frac{\sum_{j=1}^{n} X_j}{n}$ is $\frac{\sigma^2}{n}$. In this limited fluctuations approach if the variation of $\bar{X}$ about μ is not significant, then full credibility is assigned to $\bar{X}$. In statistical terms, it means that the difference between $\bar{X}$ and μ is small relative to μ with a high probability, *i.e.*, for given small r and $0 < p < 1$ (with r close to 0 and p close to 1)

$$\mathbb{P}\left(|\bar{X} - \mu| < r\mu\right) \geq p,$$

then the full credibility is achieved. Let y_p be defined by

$$y_p = \inf_y \left\{ \mathbb{P}\left(\left| \frac{\bar{x} - \mu}{\sigma\sqrt{n}} \right| < y \right) \geq p \right\},$$

then full credibility occurs if

$$\frac{\sigma}{\mu} \leq \sqrt{\frac{n}{\lambda_0}},$$

where $\lambda_0 = (\frac{y_p}{r})^2$. For more details see Klugman *et al.* (2008), p. 558.

When full credibility does not hold, the linear credibility premium (1.22) is deemed appropriate. One good choice for Z in (1.22) is

$$\frac{n}{n+k},$$

where n is the number of observations in $\bar{X}$ and k is a constant to be chosen. This form of Z tends to 1 as $n \to \infty$. A very elementary way to determine Z in (1.22) is to force the variance of the premium P_c in (1.22) to be controlled at a level, say $\frac{\mu^2}{\lambda_0}$. In this case Z can be expressed by the formula

$$Z = \min\left\{\frac{\mu}{\sigma}\sqrt{\frac{n}{\lambda_0}}, 1\right\}.$$

For more details on limited fluctuations credibility theory see Norberg (1979), Mowbray (1914), Herzog (1999) or Longley–Cook (1962).

1.2.2. Greatest accuracy credibility theory

Greatest accuracy credibility theory is a model–based approach introduced by Bühlmann (1967). In this approach, the risk parameter Θ is modeled by a probability distribution, say Π. The values of Θ varies for different policyholders and this random nature of Θ reflects the heterogeneity within an insurance portfolio. For a given $\Theta = \theta$ the distribution of the number or size of claims in year $i = 1, \ldots, n+1$ is given by $f_{X_i|\theta}(x \mid \theta)$. Usually it is assumed that for a given θ, the $X_1, \ldots, X_{n+1}$ are *i.i.d.* random variables.

The ideal premium rate for the next year $n+1$ should be $\mu_{n+1}(\Theta) = \mathbb{E}[X_{n+1} \mid \Theta]$, but the value of Θ is not known. In a Bayesian context, $\mathbb{E}[X_{n+1} \mid x_1, \ldots, x_n]$ is a valuable substitute with desirable properties. Mathematically, there is no closed form formula for this Bayesian premium, except for some special combinations of the prior distribution, Π and $f_{X|\Theta=\theta}(x \mid \theta)$. Bühlmann (1967) approximates

$\mathbb{E}[X_{n+1} \mid \Theta]$ by a linear function of the past observations $x_1, \ldots, x_n$ with a premium formula of the form :

$$\alpha_0 + \sum_{j=1}^{n} \alpha_i x_i,$$

where α_j, for $j = 1, \ldots, n$ need to be determined. To this end, the α's are chosen such as to minimize the squared error loss, that is

$$Q = \mathbb{E}\left\{ \left[\mu_{n+1}(\Theta) - \alpha_0 - \sum_{j=1}^{n} \alpha_i x_i \right]^2 \right\},$$

where the expectation is taken over the joint distribution of $X_1, \ldots, X_n$ and Θ. Equating $\frac{\partial Q}{\partial \alpha_0}$ to 0 yields the estimators $\tilde{\alpha_j}$ which satisfy

$$\mathbb{E}[\mu_{n+1}(\Theta)] = \mathbb{E}[X_{n+1}] = \tilde{\alpha_0} + \sum_{j=1}^{n} \tilde{\alpha_j} \mathbb{E}[X_j], \tag{1.23}$$

while by taking the partial derivative of Q with respect to α_i and setting to 0 gives

$$\mathbb{E}[\mu_{n+1}(\Theta) X_i] = \tilde{\alpha_0} \mathbb{E}[X_i] + \sum_{j=1}^{n} \tilde{\alpha_j} \mathbb{E}[X_i X_j].$$

The left-hand side of this equation can be written as

$$\begin{aligned}
\mathbb{E}[\mu_{n+1}(\Theta) X_i] &= \mathbb{E}\{\mathbb{E}[\mu_{n+1}(\Theta) X_i | \Theta]\} \\
&= \mathbb{E}\{\mu_{n+1}(\Theta) \mathbb{E}[X_i | \Theta]\} \\
&= \mathbb{E}\{\mathbb{E}[X_{n+1} | \Theta] \mathbb{E}[X_i | \Theta]\} \\
&= \mathbb{E}\{\mathbb{E}[X_{n+1} X_i | \Theta]\} \\
&= \mathbb{E}[X_{n+1} X_i],
\end{aligned}$$

Thus $\partial Q / \partial \alpha_i = 0$ implies

$$\mathbb{E}[X_i X_{n+1}] = \tilde{\alpha}_0 \mathbb{E}[X_i] + \sum_{j=1}^{n} \tilde{\alpha}_j \mathbb{E}[X_i X_j]. \tag{1.24}$$

Multiplying (1.23) by $\mathbb{E}[X_i]$ and subtracting from (1.24) we have

$$\mathrm{Cov}(X_i, X_{n+1}) = \sum_{j=1}^{n} \tilde{\alpha}_j \mathrm{Cov}(X_i, X_j), \; i = 1, \ldots, n. \tag{1.25}$$

Equations (1.23) and (1.25) together are called **normal equations**. In the simplest case, it is assumed that given $\Theta = \theta$, the $X_1, \ldots, X_{n+1}$ are *i.i.d.* random variables. Define

$$\mu(\theta) = \mathbb{E}[X_j|\Theta = \theta]$$

and

$$\nu(\theta) = \mathbb{V}[X_j|\Theta = \theta],$$

where $\mu(\theta)$ and $\nu(\theta)$ are referred to as the hypothetical mean and process variance, respectively. Define also

$$\mu = \mathbb{E}[\mu(\Theta)]\,,$$

$$\nu = \mathbb{E}[\nu(\Theta)]\,,$$

and

$$a = \mathbb{V}[\mu(\Theta)].$$

Bühlmann (1967) shows that

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z\bar{X} + (1 - Z)\mu,$$

where $Z = \dfrac{n}{n+k}$ and k is given by

$$k = \frac{\nu}{a} = \frac{\mathbb{E}[\mathbb{V}(X_j|\Theta)]}{\mathbb{V}[\mathbb{E}(X_j|\Theta)]}. \tag{1.26}$$

The value of Z derived from (1.26) is known as Bühlmann's credibility factor.

The values $\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ also minimize

$$Q_1 = \mathbb{E}\left\{\left[\mathbb{E}[X_{n+1}|X_1, \ldots, X_n] - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j\right]^2\right\}, \tag{1.27}$$

and

$$Q_2 = \mathbb{E}\left\{\left[X_{n+1} - \alpha_0 - \sum_{j=1}^{n} \alpha_j X_j\right]^2\right\}. \tag{1.28}$$

To see this, take derivatives of (1.27) or (1.28) with respect to $\alpha_0, \alpha_1, \ldots, \alpha_n$ and note that the solutions still satisfy the normal equations (1.23) and (1.25). Hence the credibility premium $\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j$ is the best linear estimator of each of the hypothetical mean $\mathbb{E}[X_{n+1}|\Theta]$, the Bayesian premium $\mathbb{E}[X_{n+1}|X_1, \ldots, X_n]$ and the prediction of X_{n+1}.

In Bühmann-Straub's credibility theory model, the classic Bühmann's assumptions are generalized. In this model the conditional variance, $\mathbb{V}[X_j|\Theta = \theta]$ is allowed to be a proportional to $\nu(\theta)$, *i.e.*

$$\mathbb{V}[X_j|\Theta = \theta] = \frac{\nu(\theta)}{m_j},$$

but still $\mathbb{E}[X_j|\Theta = \theta]$ does not depend on j. As a result we have again

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z\bar{X} + (1 - Z)\mu,$$

where $\bar{X} = \sum_{j=1}^{n} \frac{m_j}{m} X_j$, $Z = \dfrac{m}{m+k}$, $m = m_1 + \cdots + m_n$ and k is given by (1.26). For more details on Bühlmann-Straub's model see Klugman *et al.* (2008), p. 588, or Bühlmann and Gisler (2005), p. 77.

1.2.3. Exact credibility

Mayerson (1964) finds that the linear credibility premium is the exact Bayesian premium for some combinations of prior (also called *structural*) and claims distributions. Jewell (1974) extends these exact credibility results to the univariate exponential family of distributions with a proper choice of prior. His main result is a special case of the following theorem.

Proposition 1.12. *(Linear Exponential Family) Suppose that the X_n in $\boldsymbol{X} = (X_1, \ldots, X_{N+1})^\top$ are conditionally independent, given Θ, with common probability density function*

$$f_{X|\Theta}(x_j|\theta) = \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)}, \qquad x_j \in \chi, \theta \in \Omega, \tag{1.29}$$

and the prior density is a natural conjugate

$$\pi(\theta) = \frac{\left[q(\theta)\right]^{-k} e^{\mu k r(\theta)} r'(\theta)}{c(\mu, k)}, \qquad \theta_0 < \theta < \theta_1, \tag{1.30}$$

where $-\infty \leq \theta_0 < \theta_1 \leq \infty$, with $\pi(\theta_0) = \pi(\theta_1) = 0$, $\mu = \mathbb{E}(X)$ and $k = \frac{\mathbb{E}[\mathbb{V}(X|\Theta)]}{\mathbb{V}[\mathbb{E}(X|\Theta)]}$, then exact credibility occurs, with

$$\mathbb{E}(X_{N+1}|x_1, \ldots, x_N) = Z\,\bar{x} + (1 - Z)\,\mu, \tag{1.31}$$

where $Z = \frac{N}{N+k}$.

For a proof see Jewell (1974) and Klugman *et al.* (2008) p. 593.

Landsman and Makov (1998) extends Jewell's Proposition 1.12 to the larger exponential dispersion family. In a parametrization similar to that of Klugman *et al.* (2008), p. 593, its probability density functions are written as :

$$f_{X|\Theta}(x|\theta) = \frac{p(\lambda, x)\, e^{\lambda\, r(\theta) x}}{\big[q(\theta)\big]^{\lambda}}, \qquad x \in \chi, \theta \in \Omega\,. \tag{1.32}$$

The introduction of the dispersion parameter λ makes this a more flexible family of distributions than the linear exponential family ($\lambda = 1$). The natural conjugate prior on Θ remains the same as in (1.30) and exact credibility still occurs, now with $Z = \frac{\lambda N}{\lambda N + k}$. Other properties of (1.32) are that $\mu(\theta) = \mathbb{E}(X \,|\, \Theta = \theta) = q'(\theta)/[r'(\theta)q(\theta)]$ and $\sigma^2(\theta) = \mathbb{V}(X \,|\, \Theta = \theta) = \mu'(\theta)/[\lambda\, r'(\theta)]$. Using the natural conjugate in (1.30) gives $\mu = \mathbb{E}\big[\mu(\Theta)\big]$ and $k = \mathbb{E}\big[\mathbb{V}(X|\Theta)\big]/\mathbb{V}\big[\mathbb{E}(X|\Theta)\big]$. For a comprehensive treatment on the exponential dispersion family, see Tweedie (1984), Nelder and Wedderburn (1972) or Jørgensen (1987).

BIBLIOGRAPHY

Anderson, W. J. (1991). *Continuous-time Markov chains.* Springer-Verlag : New York.

Asmussen, S., Nerman, O., Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.* **23**, 419–441.

Asmussen, S. (2000). *Ruin probabilities.* Advanced Series on Statistical Science & Applied Probability, 2. World Scientific Publishing Co., Inc., River Edge, NJ.

Assaf, D., Levikson, B. (1982). Closure of phase type distributions under operations arising in reliability theory. *Ann. Probab.* **10**, 265–269.

Botta, R. B., Harris, C. M., Marchal, W. G. (1987). Characterizations of generalized hyperexponential distribution functions. *Stochastic Models* **3**, 115–148.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bull.* **4**, 199–207.

Bühlmann, H., Gisler, A. (2005). *A course in credibility theory and its applications.* Springer-Verlag : New York.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Proc. Cambridge Philos. Soc.* **51**, 313–319.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure time distributions. *Microelectronics and Reliability* **22**, 583–602.

Dehòn, M., Latouche, G. (1982). A geometric interpretation of the relations between the exponential and generalized Erlang distributions. *Adv. in Appl. Probab.* **14**, 885–897.

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

Durrett, R. (1999). *Essentials of stochastic processes.* Springer-Verlag : New York.

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineer's Journal* **10**, 189–197.

Fackrell, M. W. (2003). *Characterization of matrix-exponential distributions.* PH. D. thesis : The University of Adelaide.

Herzog, T. (1999). *Introduction to credibility theory, 3rd ed.* Winsted, CT :ACTEX.

Jewell, W. S. (1974). Credible means are exact Bayesian for exponential families. *ASTIN Bull.* **8**, 77–90.

Jørgensen, B. (1987). Exponential dispersion models, with discussion and a reply by the author. *J. Roy. Statist. Soc. Ser. B.* **49**, 127–162.

Klugman, S. A., Panjer, H. H., Willmot, G. E. (2004). *Loss models : From data to decision. Third edition.* John Wiley & Sons : Hoboken, NJ.

Landsman, Z. M., Makov, U. E. (1998). Exponential dispersion models and credibility. *Scand. Actuar. J.* **1**, 89–96.

Longley-Cook, L. (1962). An introduction to credibility theory. *Proceedings of the Casualty Actuarial Society* **XLIX**, 194–221.

Mayerson, L. (1964). A bayesian view of credibility. *Proceedings of the Casualty Actuarial Society* **51**, 85–104.

Moler, C., Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review* **20**, 801–836.

Mowbray, A. H. (1914) How extensive a payroll exposure is necessary to give a dependable pure premium ? *Proceedings of the Casualty Actuarial Society* **I**, 24–30.

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* (General) **135**, 370-384.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models. An algorithmic approach.* Johns Hopkins University Press, Baltimore, Md.

Neuts, M. F. (1995). Matrix-analytic methods in queueing theory. *Probab. Stochastics Ser.*, 265–292. Boca Raton, FL.

Norberg, R. (1979). The credibility approach to experience rating. *Scand. Actuar. J.* **4**, 181–221.

O'Cinneide, C. A. (1989). On nonuniqueness of representations of phase-type distributions. *Comm. Statist. Stochastic Models* **5**, 247–259.

O'Cinneide, C. A. (1991). Phase-type distributions and majorization. *Annals of Applied Probability* **1**, 219–227.

O'Cinneide, C. A. (1999). Phase-type distributions : open problems and a few properties. *Comm. Statist. Stochastic Models* **15**, 731–757.

Rolski, T., Schimdli, H., Schmidt, V., Teugels, J. (1999). *Stochastic processes for insurance and finance.* John Wiley & Sons.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics : Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* (Eds. J.K. Ghosh and J. Roy), Calcutta : Indian Statistical Institute, 579-604.

Chapitre 2

FITTING BIVARIATE LOSSES WITH PHASE-TYPE DISTRIBUTIONS

This paper was submitted for publication in February 2009 to the North American Actuarial Journal which is a refereed journal. The first author is Amin Hassan Zadeh and the coauthor is his Ph. D. supervisor Martin Bilodeau.

ABSTRACT

Maximum likelihood estimation and a (parametric bootstrap) goodness-of-fit test are considered for bivariate phase-type distributions. The initial probability vector and infinitesimal generator matrix are estimated by the EM algorithm. In a special case, the dependence structure of bivariate phase-type distributions is revealed. The results are used to fit a real bi-dimensional data set related to insurance losses (LOSS) and allocated loss adjustment expenses (ALAE). The fitted bivariate phase-type is used to obtain conditional quantiles and mean of ALAE for a given value of LOSS. The bivariate phase-type distribution meets all the requirements listed in Klugman and Parsa (1999).

Key words : Bivariate insurance losses, bivariate phase-type distribution, continuous Markov process, EM algorithm

2.1. INTRODUCTION

Phase-type (PH) random variables are defined as the time until absorption in a set of absorbent states in a continuous time Markov chain environment. Coxian,

Erlang-n, hyper-exponential and mixture of Erlang-n distributions are special cases of *PH* random variables. Neuts (1981) defines the *PH* random variable and establishes its theoretical properties. *PH* distributions are dense among all distributions with positive support. In addition, they have density, Laplace transform and all their moments in closed form and thus, various probability quantities can be obtained easily. Despite the interesting properties of *PH* variables, some difficulties arise in statistical estimation. Non-uniqueness of representations in some *PH* models, as discussed in O'Cinneide (1989), and over-parametrization is briefly mentioned in Asmussen *et al.* (1996). Asmussen *et al.* (1996) study parameter estimation by the EM algorithm, as well as fitting other densities on the positive line with *PH* distributions. In Assaf and Levikson (1982) some properties of *PH* variables in reliability are investigated. Asmussen (2000) applies *PH* distributions to risk theory. In Drekic *et al.* (2004), the distribution of deficit at ruin, in the Sparre Andersen renewal model, with *PH* distributed claim size is considered. Li and Garrido (2004) consider the ruin probability in risk theory for Erlang-n distributions, a special case of *PH* distributions. In Assaf *et al.* (1984), a multivariate *PH* distributions is defined. In Kulkarni (1989) a new class of multivariate *PH* distribution is introduced. In the multivariate case, the structure of dependence under some conditions is studied by Li (2003). The conditional tail expectation for multivariate *PH* distributions is obtained in Cai and Li (2005a).

This paper is organized as follows. Univariate and multivariate *PH* variables, with their properties, are briefly defined in Section 2.2. Section 2.3 covers parameter estimation of bivariate *PH* (*BPH*) distributions via the EM algorithm. In Section 2.4, a method to simulate a *BPH* distribution is used in a small simulation study on the bias and standard deviation of the EM estimator. A (parametric bootstrap) goodness-of-fit test for *BPH* distributions is proposed in Section 2.5. Section 2.6 includes a data analysis of the ALAE data by fitting a *BPH* distribution. It also gives expressions for the conditional quantiles and conditional mean. This article extends the works of Asmussen *et al.* (1996), Assaf *et al.* (1984) and Åhlström *et al.* (1999) to problems of statistical nature in *BPH* distributions,

namely, the statistical estimation by the EM algorithm and a parametric bootstrap goodness-of-fit test. To our knowledge, this is the first paper in which *BPH* distributions are applied in the context of a real data analysis.

2.2. PRELIMINARIES

Consider $\{J_t, t \geq 0\}$ a right continuous Markov process on the finite state space $\Gamma = \{1, 2, \ldots, m, m+1\}$ with initial probability vector $\boldsymbol{\alpha}$ and infinitesimal generator matrix $\mathbf{A}$. Suppose that Γ_1 and Γ_2 are two nonempty stochastically closed subsets ($E \subset \Gamma$ is said to be stochastically closed if, once J_t has entered E, it never leaves) of Γ such that $\Gamma_1 \cap \Gamma_2 = \{m+1\}$ and only the state $m+1$ is absorbent and hence, the absorption into it is certain. As a convention, all vectors are column vectors and superscript $\top$ denotes the transpose of a matrix. The matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0}^\top & 0 \end{pmatrix}, \tag{2.1}$$

where the matrix $\mathbf{T} = (t_{ij})$ is $m \times m$ and $\mathbf{t} = (t_j)$ is an m-dimensional vector. These elements satisfy $t_{ii} < 0$, $i = 1, \ldots, m$, $t_{ij} \geq 0$, $i \neq j$, and $\mathbf{Te} + \mathbf{t} = \mathbf{0}$, where $\mathbf{e}$ is a vector of ones. States $1, \ldots, m$ are transient if and only if $\mathbf{T}$ is nonsingular, see Neuts (1994). In this article, we always suppose that $\alpha_{m+1} = 0$, and hence, $\boldsymbol{\alpha}$ can be written as

$$\boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\pi} \\ 0 \end{pmatrix}.$$

Let X_1 and X_2 be the times until absorption in Γ_1 and Γ_2, respectively. We call the joint distribution of (X_1, X_2) a bivariate *PH* (*BPH*) distribution with representation $(\boldsymbol{\pi}, \mathbf{T}, \Gamma_1, \Gamma_2)$. The marginal distributions of X_1 and X_2 have univariate *PH* distributions.

If (X_1, X_2) has a *BPH* distribution, by using Markov chain theory, it is shown in Assaf *et al.* (1984) that the joint survival function is

$$\mathbb{P}(X_1 > x_1, X_2 > x_2) = \begin{cases} \boldsymbol{\pi}^\top e^{\mathbf{T}x_1} \mathbf{g}_1 e^{\mathbf{T}(x_2 - x_1)} \mathbf{g}_2 \mathbf{e}, & x_2 \geq x_1 \geq 0, \\ \boldsymbol{\pi}^\top e^{\mathbf{T}x_2} \mathbf{g}_2 e^{\mathbf{T}(x_1 - x_2)} \mathbf{g}_1 \mathbf{e}, & x_1 \geq x_2 \geq 0, \end{cases} \tag{2.2}$$

where $\mathbf{g}_k$, $k = 1, 2$, is an $m \times m$ diagonal matrix whose ith diagonal element is 1 if $i \in \Gamma_k^c$, and 0 otherwise. Assaf *et al.* (1984) also provide the Laplace transform of the joint distribution of (X_1, X_2)

$$\begin{aligned}\phi(u_1, u_2) &= \mathbb{E}\left[e^{-u_1 X_1 - u_2 X_2}\right] \\ &= \boldsymbol{\pi}^\top [(u_1 + u_2)\mathbf{I} - \mathbf{T}]^{-1} \left\{\mathbf{G}_2 [u_1 \mathbf{I} - \mathbf{T}]^{-1} \mathbf{T}\mathbf{g}_1 \right. \\ &\quad \left. + \mathbf{G}_1 [u_2 \mathbf{I} - \mathbf{T}]^{-1} \mathbf{T}\mathbf{g}_2 - [\mathbf{T}\mathbf{g}_1\mathbf{g}_2 - \mathbf{G}_1 - \mathbf{G}_2]\right\} \mathbf{e},\end{aligned}$$

where $\mathbf{G}_k = \mathbf{T}\mathbf{g}_k - \mathbf{g}_k\mathbf{T}$, $k = 1, 2$, is the commutator.

In general, the joint distribution F has a singular component on the set $x_1 = x_2$, which can be avoided by supposing that $t_i = 0$, for $i \in \Gamma_1^c \cap \Gamma_2^c$, see Assaf *et al.* (1984). Hereafter, we also suppose that $\pi_i = 0$, for $i \in \Gamma_1 \cup \Gamma_2$, and as a result $\mathbb{P}(X_1 > 0, X_2 > 0) = 1$. By imposing this structure on the initial probability vector $\boldsymbol{\pi}$ we have that $\mathbf{g}_k\boldsymbol{\pi} = \boldsymbol{\pi}$, $k = 1, 2$, and hence, the marginal survival functions can be obtained easily from (2.2).

Using the fact that $de^{\mathbf{T}x}/dx = \mathbf{T}e^{\mathbf{T}x} = e^{\mathbf{T}x}\mathbf{T}$, the density of the absolutely continuous component can be derived from (2.2)

$$f(x_1, x_2) = \begin{cases} \boldsymbol{\pi}^\top e^{\mathbf{T}x_1} \mathbf{G}_1 e^{\mathbf{T}(x_2 - x_1)} \mathbf{T}\mathbf{g}_2 \mathbf{e}, & x_2 \geq x_1 \geq 0, \\ \boldsymbol{\pi}^\top e^{\mathbf{T}x_2} \mathbf{G}_2 e^{\mathbf{T}(x_1 - x_2)} \mathbf{T}\mathbf{g}_1 \mathbf{e}, & x_1 \geq x_2 \geq 0. \end{cases} \tag{2.3}$$

The singular component on $x_1 = x_2$ may be useful in some applications related to life insurance. It is given in Assaf *et al.* (1984) with a further simplification as

$$\begin{aligned}\mathbb{P}(X_1 = X_2 > x) &= \boldsymbol{\pi}^\top e^{\mathbf{T}x} \mathbf{T}^{-1} \left[\mathbf{T}\mathbf{g}_1\mathbf{g}_2 - \mathbf{G}_1 - \mathbf{G}_2\right] \mathbf{e} \\ &= \boldsymbol{\pi}^\top e^{\mathbf{T}x} \mathbf{T}^{-1} \mathbf{g}_1\mathbf{g}_2 \mathbf{T}\mathbf{e}. \end{aligned} \tag{2.4}$$

Hence,

$$\mathbb{P}(X_1 = X_2) = \boldsymbol{\pi}^\top \mathbf{T}^{-1} \mathbf{g}_1\mathbf{g}_2 \mathbf{T}\mathbf{e}$$

is obtained with the evaluation of (2.4) at $x = 0$. Thus, with a correction to the statement made in Assaf *et al.* (1984), the singular part is zero if and only if $\left[\mathbf{T}\mathbf{g}_1\mathbf{g}_2 - \mathbf{G}_1 - \mathbf{G}_2\right] \mathbf{e} = 0$, which is equivalent to $t_i = 0$, $i \in \Gamma_1^c \cap \Gamma_2^c$. When $\pi_i = 0$,

for $i \in \Gamma_1 \cup \Gamma_2$, one can assume without loss of generality that

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}^{(1,2)} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{A}^{(1,2)} & \mathbf{B}^{(1)} & \mathbf{B}^{(2)} \\ \mathbf{0} & \mathbf{A}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{(2)} \end{pmatrix},$$

where the partition corresponds to the three subsets $\Gamma_1^c \cap \Gamma_2^c$, $\Gamma_1 \backslash \{m+1\}$, and $\Gamma_2 \backslash \{m+1\}$. Then, the joint density (2.3) can be rewritten, with a correction to Assaf *et al.* (1984), as

$$f(x_1, x_2) = \begin{cases} -\boldsymbol{\pi}^{(1,2)^\top} e^{\mathbf{A}^{(1,2)} x_1} \mathbf{B}^{(1)} e^{\mathbf{A}^{(1)}(x_2 - x_1)} \mathbf{A}^{(1)} \mathbf{e}, & x_2 \geq x_1 \geq 0, \\ -\boldsymbol{\pi}^{(1,2)^\top} e^{\mathbf{A}^{(1,2)} x_2} \mathbf{B}^{(2)} e^{\mathbf{A}^{(2)}(x_1 - x_2)} \mathbf{A}^{(2)} \mathbf{e}, & x_1 \geq x_2 \geq 0. \end{cases} \tag{2.5}$$

The multivariate version of phase-type random variables (*MPH*) is also defined in Assaf *et al.* (1984). Suppose that $\{J_t, t \geq 0\}$ is a right continuous Markov chain on a finite state space Γ. Let $\Gamma_1, \ldots, \Gamma_n$ be nonempty stochastically closed subsets of Γ, such that $\cap_{i=1}^n \Gamma_i$ has just one member, namely $m+1$, and absorption into $m+1$ is certain. The matrix $\mathbf{A}$ is still the infinitesimal generator as in (2.1). Define $X_k = \inf\{t \geq 0 | J_t \in \Gamma_k\}$, $k = 1, 2, \ldots, n$. We assume that $\pi_i = 0$ for $i \in \cup_{i=1}^n \Gamma_i$. The joint *MPH* distribution of $(X_1, \ldots, X_n)$ has the representation $(\boldsymbol{\pi}^\top, \mathbf{T}, \Gamma_1, \ldots, \Gamma_n)$. For $0 < x_1 \leq x_2 \leq \cdots \leq x_n$,

$$\begin{aligned} S(x_1, \ldots, x_n) &= P(X_1 > x_1, \ldots, X_n > x_n) \\ &= \boldsymbol{\pi}^\top e^{\mathbf{T} x_1} \mathbf{g}_1 e^{\mathbf{T}(x_2 - x_1)} \mathbf{g}_2 \ldots e^{\mathbf{T}(x_n - x_{n-1})} \mathbf{g}_n \mathbf{e}. \end{aligned}$$

The *MPH* distribution is absolutely continuous if and only if $t_{ij} = 0$, whenever $i \in \Gamma_k^c \cap \Gamma_l^c$ and $j \in \Gamma_k \cap \Gamma_l$, where $k \neq l$. The Laplace transform is given in Assaf *et al.* (1984) and it can be used to calculate all the moments. As in the univariate case, *MPH* has the closure property. Let $\mathbf{T} = (T_1, \ldots, T_n)$ and $\mathbf{W} = (W_1, \ldots, W_m)$ be independent *MPH* random vectors. Then, the conjunction $(\mathbf{T}, \mathbf{W}) = (T_1, \ldots, T_n, W_1, \ldots, W_m)$ is an *MPH* random vector. See Marshall and Shaked (1986) for a proof. Moreover, *MPH* distributions are closed under finite

mixture and convolution, see Assaf *et al.* (1984) and Kulkarni (1989). Cai and Li (2005b) give an explicit representation for the convolution of *MPH* distributions. As for univariate *PH*, the class of n-dimensional *MPH* distributions is dense in the set of all distributions on $[0,\infty]^n$, see Assaf *et al.* (1984) for a proof.

2.3. EM ALGORITHM

2.3.1. General EM algorithm

The EM (Expectation-Maximization) algorithm of Dempster *et al.* (1977) is a general iterative method for finding the maximum likelihood estimate of the parameters, when the data is incomplete or has missing values. One of the application of EM algorithms is when maximizing the likelihood function is analytically intractable but the likelihood function can be simplified by assuming the existence of additional but hidden or missing information. The EM algorithm is not guaranteed to find the global maximum, it may converge to a local maximum or even a saddle point of the likelihood surface, see Wu (1983).

Assume that the data $\mathbf{x}$ is observed and generated by some distributions, say $f(\mathbf{x}|\boldsymbol{\phi})$ with log-likelihood function $L(\boldsymbol{\phi}) = \log f(\mathbf{x}|\boldsymbol{\phi})$. We call $\mathbf{x}$ the incomplete data and refer to $L(\boldsymbol{\phi})$ as the incomplete log-likelihood function. Suppose that an unobserved (complete) data $\mathbf{y}$, where $\mathbf{x} = \mathbf{x}(\mathbf{y})$, has pdf $g(\mathbf{y}|\boldsymbol{\phi})$. Assume

$$Q(\boldsymbol{\phi}'|\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{\phi}}[\log g(\mathbf{y}|\boldsymbol{\phi}')|\mathbf{x}]$$

exists for all pairs $(\boldsymbol{\phi}', \boldsymbol{\phi})$. The EM iteration $\boldsymbol{\phi}^{(p)} \rightarrow \boldsymbol{\phi}^{(p+1)}$ is defined as follows :
E-step : Compute $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$.
M-step : Find $\boldsymbol{\phi}^{(p+1)}$ that maximizes $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$ over $\boldsymbol{\phi}$.

Simplifications occur when the complete data density function is a member of the exponential family

$$g(\mathbf{y}|\boldsymbol{\phi}) = b(\mathbf{y}) \exp\left[\boldsymbol{\phi}\mathbf{t}(\mathbf{y})^\top\right] / a(\boldsymbol{\phi}), \tag{2.6}$$

where $\boldsymbol{\phi}$ is the vector parameter, $\mathbf{t}(\mathbf{y})$ is the vector of the complete data sufficient statistic. If (2.6) holds, Dempster *et al.* (1977) present simplified expressions for the E and M steps :

E-step : Estimate the complete data sufficient statistics $\mathbf{t(x)}$ by finding

$$\mathbf{t}^{(p)} = \mathbb{E}_{\boldsymbol{\phi}^{(p)}}[\mathbf{t(y)}|\mathbf{x}].$$

M-step : Determine $\boldsymbol{\phi}^{(p+1)}$ as the solution for $\boldsymbol{\phi}$ of the equation

$$\mathbb{E}_{\boldsymbol{\phi}}[\mathbf{t(y)}] = \mathbf{t}^{(p)}.$$

2.3.2. EM algorithm for the *BPH* distribution

EM algorithms in a Markovian chain environment are not new. There are some works done in this context, including those of Breuer (2002), Ryden (1996), Asmussen *et al.* (1996) and Åhlström *et al.* (1999). The latter one is the EM algorithm for a special case of *BPH* distributions satisfying $x_1 < x_2$ and with censoring conditions on the data. Our work compared to Åhlström *et al.* (1999) might be considered as an incremental work to general *BPH* distributions.

By definition a *BPH* random variable is the time until absorption in stochastically closed subsets Γ_1 and Γ_2. This can be considered as an incomplete data in the sense that they only provide information about the time of hitting Γ_1 and Γ_2, not about the whole path of J_t. The initial state, the states that have been visited, and the time spent in each visited state are not observed. Hence, the hidden information can help to maximize the incomplete likelihood function which is untractable. For the case $x_1 < x_2$, the complete path can be formulated by the embedded Markov chain of visited states

$$i_0, i_1, \ldots, i_{m_1-1}, \ldots, i_{m_2-1}, i_{m_2}(= m+1),$$

and the sojourn times

$$s_0, s_1, \ldots, s_{m_1-1}, \ldots, s_{m_2-1}, s_{m_2}(= \infty),$$

where m_1 is the number of jumps until hitting Γ_1 and m_2 is the number of jumps until hitting the absorbent state $m+1$.

Given an observation (x_1, x_2) of the *BPH* distribution, a complete observation of the process J_t on the interval $(0, x_2]$ is represented by

$$\mathbf{y} = (i_0, \ldots, i_{m_1-1}, \ldots, i_{m_2-1}, s_0, \ldots, s_{m_1-1}, \ldots, s_{m_2-1}),$$

where, $x_1 = s_0 + \cdots + s_{m_1-1}$ and $x_2 = s_0 + \cdots + s_{m_2-1}$. To get the probability density function of $\mathbf{y}$, one needs the probability p_{jk} of jumping from j to k which is given by

$$p_{jk} = \mathbb{P}(i_{n+1} = k | i_n = j) = \begin{cases} 0, & j,k = 1,\ldots,m, j = k, \\ \frac{t_{jk}}{-t_{jj}}, & j,k = 1,\ldots,m, j \neq k, \\ \frac{t_j}{-t_{jj}}, & j = 1,\ldots,m, k = m+1. \end{cases}$$

The density of $\mathbf{y}$ can be derived by Markov chain properties and considering that the time spent in each state i has an exponential distribution with mean $1/\lambda_i$, where $\lambda_i = -t_{ii}$, as in Asmussen *et al.* (1996). Thus,

$$g(\mathbf{y}|\boldsymbol{\theta}) = \pi_{i_0} \exp\{-\lambda_{i_0} s_0\} t_{i_0,i_1} \ldots \exp\{-\lambda_{i_{m_2-1}} s_{m_2-1}\} t_{i_{m_2-1}}, \tag{2.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{T})$ is the parameter.

Let $J_t^{[1]}, \ldots, J_t^{[n]}$ be n independent realizations of the process. This gives n embedded Markov chains

$$i_0^{[\nu]}, \ldots, i_{m_1^{[\nu]}-1}^{[\nu]}, \ldots, i_{m_2^{[\nu]}-1}^{[\nu]}$$

with corresponding sojourn times

$$s_0^{[\nu]}, \ldots, s_{m_1^{[\nu]}-1}^{[\nu]}, \ldots, s_{m_2^{[\nu]}-1}^{[\nu]}, \ \nu = 1,\ldots,n.$$

The complete data becomes $\mathbf{y} = (\mathbf{y}^{[1]}, \ldots, \mathbf{y}^{[n]})$, where

$$\mathbf{y}^{[\nu]} = (i_0^{[\nu]}, \ldots, i_{m_1^{[\nu]}-1}^{[\nu]}, \ldots, i_{m_2^{[\nu]}-1}^{[\nu]}, s_0^{[\nu]}, \ldots, s_{m_1^{[\nu]}-1}^{[\nu]}, \ldots, s_{m_2^{[\nu]}-1}^{[\nu]}), \ \nu = 1,\ldots,n.$$

The observed incomplete data is the following function of the complete data

$$\mathbf{x}^{[\nu]} = (x_1^{[\nu]}, x_2^{[\nu]}) = (s_0^{[\nu]} + \cdots + s_{m_1^{[\nu]}-1}^{[\nu]}, s_0^{[\nu]} + \cdots + s_{m_2^{[\nu]}-1}^{[\nu]}).$$

Define

$$\begin{aligned} B_i^{[\nu]} &= \mathbf{1}\left\{i_0^{[\nu]} = i\right\}, \\ Z_i^{[\nu]} &= \sum_{k=0}^{m_2^{[\nu]}-1} \mathbf{1}\left\{i_k^{[\nu]} = i\right\} s_k^{[\nu]}, \\ N_{ij}^{[\nu]} &= \sum_{k=0}^{m_2^{[\nu]}-1} \mathbf{1}\left\{i_k^{[\nu]} = i, i_{k+1}^{[\nu]} = j\right\}. \end{aligned}$$

Let

$$\begin{aligned} B_i &= \sum_{\nu=1}^{n} B_i^{[\nu]}, \\ Z_i &= \sum_{\nu=1}^{n} Z_i^{[\nu]}, \\ N_{ij} &= \sum_{\nu=1}^{n} N_{ij}^{[\nu]}, \end{aligned}$$

be the number of Markov processes starting from state i, the total time spent in each state i, and the number of jumps from state i to state j, respectively.

Then, the density of the complete data $\mathbf{y}$ is the product of n densities as in (2.7)

$$g(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{m} \left[\pi_i^{B_i} \exp\{t_{ii} Z_i\} \prod_{j=1, j\neq i}^{m+1} t_{ij}^{N_{ij}} \right], \tag{2.8}$$

where $t_{i,m+1} = t_i$. The density (2.8) is a member of a curved multi-parameter exponential family with sufficient statistics

$$B_i,\ Z_i,\ N_{ij},$$

where $i = 1, \ldots, m$, $j = 1, \ldots, m+1$, $i \neq j$. For the complete data, maximum likelihood estimates of the unknown parameters were obtained by Asmussen *et al.* (1996). They can be used in our context and, therefore, the M-step is given by

$$\hat{\pi}_i = \frac{B_i}{n},\ \hat{t}_{ij} = \frac{N_{ij}}{Z_i},\ i \neq j,\ \hat{t}_i = \frac{N_{i,m+1}}{Z_i},\ \hat{t}_{ii} = -\left(\hat{t}_i + \sum_{j=1, j\neq i}^{m} \hat{t}_{ij} \right).$$

However, the E-step differs considerably since the observed and incomplete data is now bivariate. As it was noted in Section 2.3, the E-step for an exponential family consists of computing the conditional expectation of the sufficient statistics, given the complete data and the current parameter estimates.

If the current parameter estimates at step k of the algorithm is $\boldsymbol{\theta}^{(k)}$, the complete sufficient statistics at the $k+1$ E-step consists in the evaluation of the

following conditional expectations

$$B_i^{(k+1)} = \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[B_i^{[\nu]}|\mathbf{x}^{[\nu]}], \tag{2.9}$$

$$Z_i^{(k+1)} = \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Z_i^{[\nu]}|\mathbf{x}^{[\nu]}], \tag{2.10}$$

$$N_{ij}^{(k+1)} = \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[N_{ij}^{[\nu]}|\mathbf{x}^{[\nu]}], \tag{2.11}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, m+1$, $i \neq j$. The most complicated part is the E-step which is derived in the Appendix. All calculations are for the case $x_1 < x_2$; the other case is similar. The final results are given here. For convenience, the same notations are used as in the univariate case in Asmussen *et al.* (1996). To simplify formulas we define the sets $\Gamma_k' = \Gamma_k \backslash \{m+1\}$, $k = 1, 2$, the matrix $\mathbf{E}_{ij}$ with a one in position (i, j) and zeros elsewhere, and the vector $\mathbf{e}_i$ with a one in position i and zeros elsewhere. Also, let

$$C_1(a, b, i, j, \mathbf{T}) = \int_a^b e^{\mathbf{T}(u-a)} \mathbf{E}_{ij} e^{\mathbf{T}(b-u)} du.$$

The conditional expectations in (2.9)-(2.11) are given as follows :

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[B_i^{[\nu]}|\mathbf{x}^{[\nu]}] = \frac{\pi_i^{(k)} \mathbf{e}_i^\top e^{\mathbf{T}^{(k)} x_1^{[\nu]}} \mathbf{G}_1 e^{\mathbf{T}^{(k)}(x_2^{[\nu]} - x_1^{[\nu]})} \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})},$$

where $\pi_i^{(k)}$ is the ith component of $\boldsymbol{\pi}^{(k)}$,

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Z_i^{[\nu]}|\mathbf{x}^{[\nu]}] = \begin{cases} \frac{\boldsymbol{\pi}^{(k)\top} C_1(0, x_1^{[\nu]}, i, i, \mathbf{T}^{(k)}) \mathbf{G}_1 e^{\mathbf{T}^{(k)}(x_2^{[\nu]} - x_1^{[\nu]})} \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}, & i \in \Gamma_1^c \cap \Gamma_2^c, \\ \frac{\boldsymbol{\pi}^{(k)\top} e^{\mathbf{T}^{(k)} x_1^{[\nu]}} \mathbf{G}_1 C_1(x_1^{[\nu]}, x_2^{[\nu]}, i, i, \mathbf{T}^{(k)}) \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}, & i \in \Gamma_1', \\ 0, \text{ otherwise}, \end{cases}$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[N_{ij}^{[\nu]}|\mathbf{x}^{[\nu]}] = \begin{cases} t_{ij} \frac{\boldsymbol{\pi}^{(k)\top} C_1(0, x_1^{[\nu]}, i, j, \mathbf{T}^{(k)}) \mathbf{G}_1 e^{\mathbf{T}^{(k)}(x_2^{[\nu]} - x_1^{[\nu]})} \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}, & i, j \in \Gamma_1^c \cap \Gamma_2^c, \\ -t_{ij} \frac{\boldsymbol{\pi}^{(k)\top} e^{\mathbf{T}^{(k)} x_1^{[\nu]}} \mathbf{E}_{ij} e^{\mathbf{T}^{(k)}(x_2^{[\nu]} - x_1^{[\nu]})} \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}, & i \in \Gamma_1^c \cap \Gamma_2^c, \ j \in \Gamma_1', \\ t_{ij} \frac{\boldsymbol{\pi}^{(k)\top} e^{\mathbf{T}^{(k)} x_1^{[\nu]}} \mathbf{G}_1 C_1(x_1^{[\nu]}, x_2^{[\nu]}, i, j, \mathbf{T}^{(k)}) \mathbf{T}^{(k)} \mathbf{g}_2 \mathbf{e}}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}, & i, j \in \Gamma_1', \\ 0, \text{ otherwise}, \end{cases}$$

$$\mathbb{E}_{\boldsymbol{\theta}^{(k)}}[N_{i,m+1}^{[\nu]}|\mathbf{x}^{[\nu]}] = -t_i \frac{\boldsymbol{\pi}^{(k)\top} e^{\mathbf{T}^{(k)} x_1^{[\nu]}} \mathbf{G}_1 e^{\mathbf{T}^{(k)}(x_2^{[\nu]} - x_1^{[\nu]})} \mathbf{e}_i}{f(\mathbf{x}^{[\nu]}|\boldsymbol{\theta}^{(k)})}.$$

The function $C_1(a, b, i, j, \mathbf{T})$ can be evaluated by numerical methods in ordinary or partial differential equations such as the Runge-Kutta method of order four. Note that for $a < b$,

$$C_1(a, b, i, j, \mathbf{T}) = C_1'(0, b - a, i, j, \mathbf{T}),$$

where

$$C_1'(0, r, i, j, \mathbf{T}) = \int_0^r e^{\mathbf{T}(r-u)} \mathbf{E}_{ij} e^{\mathbf{T}u} du,$$

and

$$\frac{dC_1'(0, r, i, j, \mathbf{T})}{dr} = \mathbf{E}_{ij} e^{\mathbf{T}r} + \mathbf{T} C_1'(0, r, i, j, \mathbf{T}),$$

with initial conditions $C_1'(0, 0, i, j, \mathbf{T}) = \mathbf{0}$.

We end this section with an interesting property that holds when one fits a *BPH* distribution by the EM algorithm.: at each iteration of the EM algorithm, the mean of the fitted *BPH* distribution equals the sample mean. This property was given by Asmussen *et al.* (1996) for *PH* distributions and it also holds in the bivariate case as is now shown. The observations are linear functions of the sufficient statistics,

$$\left(X_1^{[\nu]}, X_2^{[\nu]}\right) = \left(\sum_{i \in \Gamma \setminus \Gamma_1} Z_i^{[\nu]}, \sum_{i \in \Gamma \setminus \Gamma_2} Z_i^{[\nu]}\right).$$

For the component X_1, this implies

$$n\bar{X}_1 = \sum_{\nu=1}^{n} \sum_{i \in \Gamma \setminus \Gamma_1} Z_i^{[\nu]}.$$

Taking conditional expectations, given $\mathbf{x}$, on both sides yields

$$\begin{aligned} n\bar{X}_1 &= \sum_{\nu=1}^{n} \sum_{i \in \Gamma \setminus \Gamma_1} \mathbb{E}_{\boldsymbol{\theta}^{(k)}}[Z_i^{[\nu]} | \mathbf{x}^{[\nu]}] \\ &= \sum_{\nu=1}^{n} \sum_{i \in \Gamma \setminus \Gamma_1} \mathbb{E}_{\boldsymbol{\theta}^{(k+1)}}[Z_i^{[\nu]}] \\ &= \sum_{\nu=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(k+1)}}[X_1^{[\nu]}] \\ &= n\mathbb{E}_{\boldsymbol{\theta}^{(k+1)}}[X_1]. \end{aligned}$$

The same argument applies to the component X_2.

2.4. SIMULATED DATA FROM A *BPH* DISTRIBUTION

Consider a *BPH* distribution with state space $\Gamma = \{1, 2, 3, 4\}$, closed subsets $\Gamma_1 = \{2, 4\}$ and $\Gamma_2 = \{3, 4\}$, $\boldsymbol{\pi} = (1, 0, 0)^\top$, and the matrix

$$\mathbf{T} = \begin{pmatrix} -a_1 & pa_1 & qa_1 \\ 0 & -a_2 & 0 \\ 0 & 0 & -a_3 \end{pmatrix}, \tag{2.12}$$

where $0 < p < 1$, $q = 1 - p$, $a_i > 0$, $i = 1, 2, 3$. This is a special case of the Marshall-Olkin distribution, see Marshall and Olkin (1967). The joint survival function can be written without any matrix exponential using the simplified joint density (2.5)

$$S_{\boldsymbol{\theta}}(x_1, x_2) = \begin{cases} \frac{p}{a_1 - a_2}\left[a_1 e^{-x_1(a_1 - a_2) - x_2 a_2} - a_2 e^{-a_1 x_2}\right] + q e^{-a_1 x_2}, & x_2 \geq x_1 \geq 0, \\ \frac{q}{a_1 - a_3}\left[a_1 e^{-x_2(a_1 - a_3) - x_1 a_3} - a_3 e^{-a_1 x_1}\right] + p e^{-a_1 x_1}, & x_1 \geq x_2 \geq 0. \end{cases}$$

The Pearson correlation coefficient,

$$r_{\boldsymbol{\theta}} = \frac{a_2 a_3 - pq a_1^2}{[a_2^2 + (1 - q^2)a_1^2]^{1/2}[a_3^2 + (1 - p^2)a_1^2]^{1/2}},$$

is obtained from the Laplace transform. It is estimated from the data using the usual sample correlation coefficient $\hat{r}$. The maximum correlation of 1 is obtained by letting a_1 approach 0, whereas the minimum of -1/3 is reached by letting a_2 and a_3 approach 0 and choosing $p = 1/2$. Also, if $a_2 = qa_1$ and $a_3 = pa_1$ then, $r = 0$. In fact, the last case corresponds to two independent exponentials. A more

appropriate measure of dependence for distributions with non-linear regression is Spearman correlation

$$\rho_{\boldsymbol{\theta}} = -3 + 12 \int_0^\infty \int_0^\infty S_{\boldsymbol{\theta}}(x_1, x_2) dF_{\boldsymbol{\theta},1}(x_1) dF_{\boldsymbol{\theta},2}(x_2),$$

where $F_{\boldsymbol{\theta},i}$, $i = 1, 2$, are the two marginal distributions of the joint *BPH* distribution $F_{\boldsymbol{\theta}}$. In the special case $a_2 = a_3$ it is given by

$$\rho_{\boldsymbol{\theta}} = \frac{7a_1a_2^2 + 2a_2^3 - 6a_1^3pq + a_1^2a_2(4p-3)(4p-1)}{2a_1^3 + 7a_1a_2(a_1+a_2) + 2a_2^3}.$$

The minimum Spearman correlation of -3/4 and the maximum of 1 are obtained under the same circumstances as for Pearson correlation. Another appropriate measure is Kendall correlation

$$\tau_{\boldsymbol{\theta}} = -1 + 4 \int_0^\infty \int_0^\infty S_{\boldsymbol{\theta}}(x_1, x_2) dF_{\boldsymbol{\theta}}(x_1, x_2),$$

which for this model is

$$\tau_{\boldsymbol{\theta}} = \frac{a_2a_3 - 2a_1^2pq + a_1(a_3q^2 + a_2p^2)}{(a_1+a_2)(a_1+a_3)}.$$

Kendall $\tau_{\boldsymbol{\theta}}$ for this model varies between -1/2 and 1. For any bivariate distribution, Spearman $\rho_{\boldsymbol{\theta}}$ and Kendall $\tau_{\boldsymbol{\theta}}$ can be consistently estimated from the data using their empirical versions

$$\begin{aligned} \hat{\rho} &= \frac{12}{n(n+1)(n-1)} \sum_{\nu=1}^{n} R_\nu S_\nu - 3\frac{n+1}{n-1}, \\ \hat{\tau} &= \frac{4}{n(n-1)} P_n - 1, \end{aligned}$$

where (R_ν, S_ν) are the pairs of ranks and P_n is the number of concordant pairs. Here, two pairs $(x_1^{[\nu]}, x_2^{[\nu]})$ and $(x_1^{[\gamma]}, x_2^{[\gamma]})$ are said to be concordant when $(x_1^{[\nu]} - x_1^{[\gamma]})(x_2^{[\nu]} - x_2^{[\gamma]}) > 0$.

The EM algorithm was run on 10,000 data sets of size $n = 50$, 100, 200, and 400 generated from a *BPH* distribution with sub-intensity matrix $\mathbf{T}$ as in (2.12) with $p = 0.5$, $a_1 = 0.05$, $a_2 = 0.1$ and $a_3 = 0.1$, which gives correlation coefficients of $r_{\boldsymbol{\theta}} = 0.7836$, $\rho_{\boldsymbol{\theta}} = 0.675$, and $\tau_{\boldsymbol{\theta}} = 0.5$. For the model (2.12), it can be observed that since each of the conditional expectations of the E-step does not depend on

$\boldsymbol{\theta}$ then, the EM algorithm converges in one iteration. This remark gives also an explicit expression for the EM estimator

$$\hat{\mathbf{T}} = \begin{pmatrix} -\frac{n}{\sum_{\nu=1}^{n} \min(x_1^{[\nu]}, x_2^{[\nu]})} & \frac{\sum_{\nu=1}^{n} \mathbf{1}\left\{x_1^{[\nu]} < x_2^{[\nu]}\right\}}{\sum_{\nu=1}^{n} \min(x_1^{[\nu]}, x_2^{[\nu]})} & \frac{\sum_{\nu=1}^{n} \mathbf{1}\left\{x_1^{[\nu]} > x_2^{[\nu]}\right\}}{\sum_{\nu=1}^{n} \min(x_1^{[\nu]}, x_2^{[\nu]})} \\ 0 & -\frac{\sum_{\nu=1}^{n} \mathbf{1}\left\{x_1^{[\nu]} < x_2^{[\nu]}\right\}}{\sum_{\nu=1}^{n} \max(0, x_2^{[\nu]} - x_1^{[\nu]})} & 0 \\ 0 & 0 & -\frac{\sum_{\nu=1}^{n} \mathbf{1}\left\{x_1^{[\nu]} > x_2^{[\nu]}\right\}}{\sum_{\nu=1}^{n} \max(0, x_1^{[\nu]} - x_2^{[\nu]})} \end{pmatrix}.$$

Table 2.1 reports absolute values of biases and standard deviations of the EM estimate computed over the 10,000 replicates. As for any regular maximum likelihood estimator, bias decreases as n^{-1} and becomes negligible compared to the standard deviation which decreases as $n^{-1/2}$.

n		bias			standard deviation		
		0.0010	0.0005	0.0005	0.0074	0.0053	0.0051
50		0	0.0040	0	0	0.0218	0
		0	0	0.0042	0	0	0.0218
		0.0005	0.0002	0.0002	0.0051	0.0036	0.0036
100		0	0.0021	0	0	0.0149	0
		0	0	0.0019	0	0	0.0147
		0.0002	0.0001	0.0001	0.0036	0.0025	0.0025
200		0	0.0011	0	0	0.0104	0
		0	0	0.0011	0	0	0.0103
		0.0991	0.0461	0.0530	0.0025	0.0018	0.0018
400	.001	0	0.6161	0	0	0.0072	0
		0	0	0.5165	0	0	0.0071

TAB. 2.1. Bias and standard deviation of the EM estimator.

Generating of a data set of size n from the *BPH* distribution with representation $(\boldsymbol{\pi}, \mathbf{T}, \Gamma_1, \Gamma_2)$ can be done by repeating n times the following algorithm for one observation (x_1, x_2).

(1) Initialize $x_{min} = 0$, $x_{max} = 0$, and $\boldsymbol{\alpha} = (\boldsymbol{\pi}, 0)$.

(2) Generate a state, $i \in \{1, \ldots, m+1\}$, from a one trial multinomial(1,$\boldsymbol{\alpha}$) distribution.

(3) Generate a sojourn time, s, from the exponential distribution with mean $-1/t_{ii}$.

(4) $x_{max} = x_{max} + s$.

(5) If $i \in \Gamma_1^c \cap \Gamma_2^c$ then $x_{min} = x_{min} + s$.

(6) Calculate the transition probabilities

$$p_{i,k} = \begin{cases} 0, & k = i, \\ \frac{t_{ik}}{-t_{ii}}, & k = 1, \ldots, m, k \neq i, \\ \frac{t_i}{-t_{ii}}, & k = m+1. \end{cases}$$

and let $\boldsymbol{\alpha} = (p_{i,1}, \ldots, p_{i,m+1})$.

(7) Set $j = i$ to store the state previously visited.

(8) Repeat steps (2) to (7) as long as $i \neq m+1$.

(9) If $j \in \Gamma_1$ then $x_1 = x_{min}$ and $x_2 = x_{max}$, otherwise $x_1 = x_{max}$ and $x_2 = x_{min}$.

2.5. GOODNESS-OF-FIT TEST

The statistic used for testing goodness-of-fit is

$$V_n^2 = \sum_{\nu=1}^{n} \left[S_{\hat{\boldsymbol{\theta}}_n}(x_1^{[\nu]}, x_2^{[\nu]}) - \hat{S}_n(x_1^{[\nu]}, x_2^{[\nu]}) \right]^2, \tag{2.13}$$

where $\hat{S}_n$ is the empirical survival function, *i.e.*

$$\hat{S}_n(x_1, x_2) = \frac{1}{n} \sum_{\nu=1}^{n} \mathbf{1} \left\{ X_1^{[\nu]} > x_1, X_2^{[\nu]} > x_2 \right\},$$

and $S_{\hat{\boldsymbol{\theta}}_n}$ is the parametric survival function (2.2) with $\boldsymbol{\theta}$ estimated by the EM algorithm. Large values of V_n^2 are evidence against the parametric model. For regular parametric models, Stute *et al.* (1993) established that the parametric bootstrap of

$$W_n^2 = n \int \left[F_{\hat{\boldsymbol{\theta}}_n}(\mathbf{x}) - \hat{F}_n(\mathbf{x}) \right]^2 dF_{\hat{\boldsymbol{\theta}}_n}(\mathbf{x})$$

is consistent for testing goodness-of-fit. An argument similar to the Lemma in Section 2 of Kiefer (1959) can be used to show that the following version

$$\begin{aligned} W_n^2 &= n \int \left[F_{\hat{\boldsymbol{\theta}}_n}(\mathbf{x}) - \hat{F}_n(\mathbf{x}) \right]^2 d\hat{F}_n(\mathbf{x}) \\ &= \sum_{\nu=1}^{n} \left[F_{\hat{\boldsymbol{\theta}}_n}(\mathbf{x}^{[\nu]}) - \hat{F}_n(\mathbf{x}^{[\nu]}) \right]^2 \end{aligned}$$

can also be bootstrapped consistently. The proposed V_n^2 is obtained by replacing the distribution function by the survival function which has a simpler expression for *BPH* distributions.

The goodness-of-fit bootstrap test of significance level α is performed as follows. Given a sample of size n, $(x_1^{[\nu]}, x_2^{[\nu]})$, $\nu = 1, \ldots, n$, estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_n$ using the EM algorithm for *BPH* distributions, and calculate the goodness-of-fit statistic V_n^2 in (2.13). Then, repeat a large number, say B, of times the following 3 steps.

(1) Generate a bootstrap sample of size n from the *BPH* distribution with parameter $\hat{\boldsymbol{\theta}}_n$ denoted $(\tilde{x}_1^{[\nu]}, \tilde{x}_2^{[\nu]})$, $\nu = 1, \ldots, n$.

(2) Find the EM estimate $\tilde{\boldsymbol{\theta}}_n$ from the bootstrap sample.

(3) Compute the goodness-of-fit statistic

$$\tilde{V}_n^2 = \sum_{\nu=1}^{n} \left[S_{\tilde{\boldsymbol{\theta}}_n}(\tilde{x}_1^{[\nu]}, \tilde{x}_2^{[\nu]}) - \hat{S}_n(\tilde{x}_1^{[\nu]}, \tilde{x}_2^{[\nu]}) \right]^2 .$$

After repeating the previous loop B times, this Monte Carlo simulation produces B (ordered) values :

$$\tilde{V}_{n,(1)}^2 \leq \tilde{V}_{n,(2)}^2 \leq \cdots \leq \tilde{V}_{n,(B)}^2 .$$

The bootstrap test rejects the model when V_n^2 exceeds the $\lceil (1-\alpha)B \rceil$ order statistic, *i.e.* when $V_n^2 > \tilde{V}_{n,(\lceil (1-\alpha)B \rceil)}^2$.

A simulation was conducted to verify the significance level of the bootstrap test. It consisted in generating 2,000 samples of size $n = 200$ from the *BPH* distribution (2.12), each with a different sub-intensity matrix $\mathbf{T}$ generated at random. The parameters were independently generated from the following distributions : a_1, a_2, a_3 uniform on the interval (0,10) and p uniform on (0,1). Each bootstrap

	LOSS	ALAE
mean	41,208	12,588
median	12,000	5,471
standard deviation	102,748	28,146
minimum	10	15
maximum	2,173,595	501,863
0.25 quantile	4,000	2,333
0.75 quantile	35,000	12,577
Pearson $\hat{r}$	0.4022	
Spearman $\hat{\rho}$	0.4519	
Kendall $\hat{\tau}$	0.3154	

TAB. 2.2. Descriptive statistics of ALAE data.

test was done at level $\alpha = 0.05$ with $B = 1,000$ bootstrap samples. The proportion of significant bootstrap tests obtained, out of 2,000, was 0.054, which is close to the intended 0.05 level.

2.6. FITTING ALAE DATA WITH A *BPH* MODEL

This section consists of fitting a *BPH* distribution to insurance company indemnity claims. The data set contains $1,500$ bi-dimensional observations. The variables are LOSS or indemnity payment and ALAE, allocated loss adjustment expenses, which covers expenses attributed to the settlement of individual claims such as claim investigation expenses. See Klugman *et al.* (2008) for more information on the data. Frees and Valdez (1998) and Klugman and Parsa (1999) fitted this data set using copulas. The variable LOSS treated in these two papers is the loss incured to the insured so that this variable is censored when the claim exceeds the policy limit. Here, LOSS is always the indemnity payment which means censoring is ignored. It can be observed from Figure 2.1 that there is a moderate positive sample Pearson correlation of $\hat{r} = 0.4022$. The descriptive statistics are given in Table 2.2.

The variables were rescaled as LOSS/100,000 and ALAE/10,000 so that they are about of the same order of magnitude. This rescaling also makes the elements of the sub-intensity matrix $\mathbf{T}$ not too close to zero, thus improving the numerical stability. To begin the algorithm, a few choices were tried for Γ, Γ_1 and Γ_2. We start with the minimal elements for $\Gamma = \{1,2,3,4\}$, $\Gamma_1 = \{2,4\}$ and $\Gamma_2 = \{3,4\}$ which results in a very poor fit. Then, more elements were added to these three sets. Finally, it was found that the choice $\Gamma = \{1,\ldots,12\}$, $\Gamma_1 = \{5,6,7,12\}$, $\Gamma_2 = \{8,9,10,11,12\}$, yields a good fit. The EM algorithm was iterated 300 times starting with random values of $\boldsymbol{\pi}$ and $\mathbf{T}$. The fitted *BPH* distribution could capture the essential characteristics of the joint distribution, such as the dependence structure.

As shown in Table 2.3, the fitted *BPH* survival function is close to the empirical survival function over the whole domain. The fitted *BPH* distribution has standard deviations for LOSS and ALAE of 102,750 and 28,146, respectively. Pearson, Spearman, and Kendall correlation coefficients are $r_{\hat{\theta}_n} = 0.3932$, $\rho_{\hat{\theta}_n} = 0.4512$ (0.0116), and $\tau_{\hat{\theta}_n} = 0.2968$ (0.0042), respectively. These are very close to the sample statistics in Table 2.2. Since there is no explicit expression for the latter two coefficients, they can be computed numerically, with any desired degree of accuracy, through a simulation. The evaluation of Kendall $\tau_{\hat{\theta}_n}$, for example, was done by jointly simulating 50,000 data points from the fitted (joint) *BPH* distribution, $F_{\hat{\theta}_n}$, and averaging, over these 50,000 values, the survival function, $S_{\hat{\theta}_n}$. As a measure of accuracy, the standard error of the mean accompanies in parentheses this average. Spearman correlation can be evaluated similarly, with the exception that the two components are independently generated from the marginal *PH* distributions of the fitted *BPH* distribution. Also, the means of the fitted *BPH* distribution always equal the sample means as shown in Section 2.3.2. The marginals of the fitted *BPH* distribution are plotted in Figure 2.2. The columns G-H and Frank of Table 2.3 give the two fits obtained in Frees and Valdez (1998) using Pareto distributed marginals in both cases with either the Gumbel-Hougaard copula or the Frank copula. The estimates of the parameters of these two fits were computed by Frees and Valdez (1998) and

(LOSS,ALAE)	$\hat{S}_n$	$S_{\hat{\theta}_n}$	G-H	Frank
(1,000 ;0)	0.9300	0.9262	0.9257	0.9286
(8,000 ;0)	0.5927	0.5981	0.6029	0.6137
(10,000 ;0)	0.5200	0.5425	0.5469	0.5582
(1,000,000 ;0)	0.0007	0.0007	0.0082	0.0088
(10,000,000 ;0)	0.0000	0.0000	0.0000	0.0000
(0 ;1,000)	0.8793	0.8819	0.8659	0.8742
(0 ;10,000)	0.3120	0.3150	0.3237	0.3380
(0 ;100,000)	0.0133	0.0154	0.0121	0.0112
(100 ;10)	0.9932	0.9910	0.9907	0.9910
(100 ;100)	0.9827	0.9807	0.9780	0.9790
(1,000 ;100)	0.9203	0.9162	0.9148	0.9177
(10,000 ;100,000)	0.0120	0.0142	0.0115	0.0097
(1,000 ;1,000)	0.8253	0.8233	0.8157	0.8255
(10,000 ;1,000)	0.4847	0.5047	0.5078	0.5267
(15,000 ;5,000)	0.3013	0.3088	0.3103	0.3356
(5,000 ;15,000)	0.1940	0.1933	0.1953	0.2070
(40,000 ;12,000)	0.1280	0.1263	0.1255	0.1221
(1,000,000 ;100,000)	0.0007	0.0002	0.0039	0.0003
V_n^2		0.1280		

TAB. 2.3. Comparison of survival functions.

were simply used here to compute the survival functions from the distribution functions. The last row is the measure of global fit used for the goodness-of-fit test in Section 2.5 and computed here over all the $n = 1,500$ data points. The Gumbel-Hougaard and Frank models with the censored variable LOSS give fits quite similar to the one obtained with the *BPH* model without censoring. The fit were comparable because there were only about 34 censored observations. Strictly speaking however, such comparisons are difficult to interpret since the variable LOSS has different meanings.

2.6.1. Conditional quantiles and mean

The conditional survival function of X_2, given $X_1 = x_1$, can be obtained from (2.2),

$$\mathbb{P}(X_2 > x_2|x_1) = \begin{cases} \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{G}_1 e^{\mathbf{T}(x_2-x_1)}\mathbf{g}_2\mathbf{e}}{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{T}\mathbf{g}_1\mathbf{e}}, & x_1 < x_2, \\ \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_2}\mathbf{g}_2 e^{\mathbf{T}(x_1-x_2)}\mathbf{T}\mathbf{g}_1\mathbf{e}}{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{T}\mathbf{g}_1\mathbf{e}}, & x_1 > x_2. \end{cases} \tag{2.14}$$

For a given value of LOSS, quantiles of ALAE were calculated based on the conditional survival function (2.14) and presented in Table 2.4 and Figure 2.1. The conditional mean of X_2 given $X_1 = x_1$ is somewhat tedious. For this purpose, by using either (2.14) directly or the conditional probability density function derived from (2.14), one can write a system of differential equations of the first order which can be numerically solved by, *e.g.*, Runge-Kutta methods. The probability density function of X_2 given $X_1 = x_1$ was used here. After some straightforward calculations, the conditional mean is given by

$$\mathbb{E}(X_2|x_1) = \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{G}_1(-\mathbf{T}^{-1} + x_1\mathbf{I})\mathbf{g}_2\mathbf{e} + \boldsymbol{\pi}^\top C_2(x_1, \mathbf{G}_2, \mathbf{T})\mathbf{T}\mathbf{g}_1\mathbf{e}}{-\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{T}\mathbf{g}_1\mathbf{e}},$$

where $C_2(x, \mathbf{G}, \mathbf{T})$ is a function satisfying the differential equation

$$\frac{d}{dx}C_2(x, \mathbf{G}, \mathbf{T}) = C_2(x, \mathbf{G}, \mathbf{T})\mathbf{T} + xe^{\mathbf{T}x}\mathbf{G},$$

with initial conditions $C_2(0, \mathbf{G}, \mathbf{T}) = \mathbf{0}$. The conditional mean is given in the last column of Table 2.4. It is observed in Figure 2.1 that the mean is always greater than the median which reflects the right-skewness of ALAE.

BIBLIOGRAHY

Åhlström, L., Olsson, M., Nerman, O. (1999). A parametric estimation procedure for relapse time distributions. *Lifetime Data Anal.* **2**, 113–132.

Assaf, D., Levikson, B. (1982). Closure of phase type distributions under operations arising in reliability theory. *Ann. Probab.* **10**, 265–269.

Assaf, D., Langberg, N. A., Savits, T. H., Shaked, M. (1984). Multivariate phase-type distributions. *Oper. Res.* **32**, 688–702.

	quantiles					
LOSS	0.05	0.25	0.50	0.75	0.95	mean
10	220	1,240	3,010	6,120	14,590	5,146
100	222	1,240	3,012	6,140	14,590	5,164
1,000	241	1,278	3,088	6,316	15,400	5,356
5,000	318	1,500	3,529	7,324	19,610	6,342
10,000	390	1,895	6,316	9,128	25,320	7,852
25,000	643	3,321	7,364	15,070	36,280	12,243
50,000	954	3,869	9,592	18,914	45,990	15,981
100,000	1,246	5,814	12,567	25,115	89,590	24,034
1,000,000	6,250	29,840	61,930	101,660	181,500	71,721

TAB. 2.4. Conditional quantiles and mean of ALAE given LOSS.

Asmussen, S., Nerman, O., Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.* **23**, 419–441.

Asmussen, S. (2000). *Ruin probabilities.* Advanced Series on Statistical Science & Applied Probability, 2. World Scientific Publishing Co., Inc., River Edge, NJ.

Breuer, L. (2002). An EM algorithm for batch Markovian arrival processes and its comparison to a simpler estimation procedure. Ann. Oper. Res. **112**, 123–138.

Cai, J., Li, H. (2005a). Conditional tail expectations for multivariate phase-type distributions. *J. Appl. Probab.* **42**, 810–825.

Cai, J., Li, H. (2005b). Multivariate risk model of phase type. *Insurance Math. Econom.* **36**, 137–152.

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

Drekic, S., Dickson, D. C. M., Stanford, D. A., Willmot, G. E. (2004). On the distribution of the deficit at ruin when claims are phase-type. *Scand. Actuar. J.* **2**, 105–120.

Frees, E. W., Valdez, E. A. (1998). Understanding relationships using copulas. *N. Am. Actuar. J.* **2**, 1–25.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-V. Mises tests. *Ann. Math. Statist.* **30**, 420–447.

Klugman, S. A., Panjer, H. H., Willmot, G. E. (2008). *Loss models : From data to decision.* Third edition. John Wiley & Sons : Hoboken, NJ.

Klugman, S. A., Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance Math. Econom.* **24**, 139–148.

Kulkarni, V. G. (1989). A new class of multivariate phase type distributions. *Oper. Res.* **37**, 151–158.

Li, H. (2003). Association of multivariate phase-type distributions, with applications to shock models. *Statist. Probab. Lett.* **64**, 381–392.

Li, S., Garrido, J. (2004). On ruin for the Erlang(n) risk process. *Insurance Math. Econom.* **34**, 391–408.

Marshall, A. W., Olkin, I. (1967). A multivariate exponential distribution. *J. Amer. Statist. Assoc.* **62**, 30–44.

Marshall, A. W. Shaked, M. (1986). Multivariate new better than used distributions. *Math. Oper. Res.* **11**, 110–116.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models. An algorithmic approach.* Johns Hopkins University Press, Baltimore, Md.

Neuts, M. F. (1994). *Matrix-geometric solutions in stochastic models. An algorithmic approach.* Dover Publications Inc. : New York.

O'Cinneide, C. A. (1989). On nonuniqueness of representations of phase-type distributions. *Comm. Statist. Stochastic Models.* **5**, 247–259.

Ryden, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. Comput. Statist. Data Anal. **21**, 431–447.

Stute, W., Gonzáles, M. W., Wenceslao, P. Q. M. (1993). Bootstrap based goodness-of-fit tests. *Metrika* **40**, 243–256.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

APPENDIX

We assume a *BPH* distribution for which $\mathbb{P}(X_1 > 0, X_2 > 0) = 1$. As a result, the initial state can be only in $\Gamma_1^c \cap \Gamma_2^c$. All calculations are for the case $x_1 < x_2$. The expressions for the case $x_1 > x_2$ are given without further explanations.

1. Calculations of $\mathbb{E}[B_i|\mathbf{x}]$:

For $x_1 < x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[B_i|\mathbf{x}] &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{1}\{i_0 = i\}|\mathbf{x}] \\
&= \mathbb{P}_{\boldsymbol{\theta}}(i_0 = i|\mathbf{x}) \\
&= \frac{\mathbb{P}_{\boldsymbol{\theta}}(i_0 = i)\,\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} \in d\mathbf{x}|i_0 = i)}{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} \in d\mathbf{x})} \\
&= \frac{\pi_i \mathbf{e}_i^\top e^{\mathbf{T}x_1}\mathbf{G}_1 e^{\mathbf{T}(x_2-x_1)}\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.
\end{aligned}$$

For $x_1 > x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$,

$$\mathbb{E}_{\boldsymbol{\theta}}[B_i|\mathbf{x}] = \frac{\pi_i \mathbf{e}_i^\top e^{\mathbf{T}x_2}\mathbf{G}_2 e^{\mathbf{T}(x_1-x_2)}\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

2. Calculations of $\mathbb{E}[Z_i|\mathbf{x}]$:

For $x_1 < x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}[Z_i|\mathbf{x}] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\int_0^\infty \mathbf{1}\{J_u = i\}\,du|\mathbf{x}\right] \\
&= \int_0^\infty \mathbb{P}_{\boldsymbol{\theta}}(J_u = i|\mathbf{x})\,du \\
&= \int_0^\infty \frac{\mathbb{P}_{\boldsymbol{\theta}}(J_u = i)\,\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} \in d\mathbf{x}|J_u = i)}{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{X} \in d\mathbf{x})}du \\
&= \frac{\boldsymbol{\pi}^\top C_1(0, x_1, i, i, \mathbf{T})\mathbf{G}_1 e^{\mathbf{T}(x_2-x_1)}\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.
\end{aligned}$$

For $x_1 > x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$,

$$\mathbb{E}_{\boldsymbol{\theta}}[Z_i|\mathbf{x}] = \frac{\boldsymbol{\pi}^\top C_1(0, x_2, i, i, \mathbf{T})\mathbf{G}_2 e^{\mathbf{T}(x_1-x_2)}\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

For $x_1 < x_2$, $i \in \Gamma_1'$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[Z_i|\mathbf{x}] &= \int_0^\infty \mathbb{P}_{\boldsymbol{\theta}}\left(J_u = i|\mathbf{x}\right) du \\ &= \int_0^\infty \frac{\mathbb{P}_{\boldsymbol{\theta}}\left(X_1 \in dx_1\right)\mathbb{P}_{\boldsymbol{\theta}}\left(J_u = i|X_1 \in dx_1\right)\mathbb{P}_{\boldsymbol{\theta}}\left(X_2 \in dx_2|J_u = i\right)}{\mathbb{P}_{\boldsymbol{\theta}}\left(\mathbf{X} \in d\mathbf{x}\right)} \\ &= \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{G}_1 C_1(x_1, x_2, i, i, \mathbf{T})\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.\end{aligned}$$

For $x_1 > x_2$, $i \in \Gamma_2'$,

$$\mathbb{E}_{\boldsymbol{\theta}}[Z_i|\mathbf{x}] = \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_2}\mathbf{G}_2 C_1(x_2, x_1, i, i, \mathbf{T})\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

3. Calculations of $\mathbb{E}[N_{ij}|\mathbf{x}]$:

For every small $\epsilon > 0$, N_{ij} can be approximated by

$$N_{ij}^\epsilon = \sum_{k=0}^\infty \mathbf{1}\left\{J_{k\epsilon} = i, J_{(k+1)\epsilon=j}\right\}.$$

For each case we have calculated $\mathbb{E}_{\boldsymbol{\theta}}[N_{ij}^\epsilon|\mathbf{x}]$. The exact value can be obtained by letting $\epsilon \downarrow 0$.

For $x_1 < x_2$, $i, j \in \Gamma_1^c \cap \Gamma_2^c$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[N_{ij}^\epsilon|\mathbf{x}] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{k=0}^\infty \mathbf{1}\left\{J_{k\epsilon} = i, J_{(k+1)\epsilon=j}\right\}|\mathbf{x}\right] \\ &= \sum_{k=0}^{[x_1/\epsilon]-1} \frac{\mathbb{P}_{\boldsymbol{\theta}}\left(J_{k\epsilon} = i, J_{(k+1)\epsilon} = j, \mathbf{X} \in d\mathbf{x}\right)}{\mathbb{P}\left(\mathbf{X} \in d\mathbf{x}\right)} \\ &= \sum_{k=0}^{[x_1/\epsilon]-1} \frac{\mathbb{P}_{\boldsymbol{\theta}}\left(J_{k\epsilon} = i\right)\mathbb{P}_{\boldsymbol{\theta}}\left(J_{(k+1)\epsilon} = j|J_{k\epsilon} = i\right)\mathbb{P}_{\boldsymbol{\theta}}\left(\mathbf{X} \in d\mathbf{x}|J_{(k+1)\epsilon} = j\right)}{\mathbb{P}_{\boldsymbol{\theta}}\left(\mathbf{X} \in d\mathbf{x}\right)}.\end{aligned}$$

Since $e^{\mathbf{T}u}$ is a continuous function and

$$\frac{e^{\mathbf{T}\epsilon} - I}{\epsilon} \to \mathbf{T}, \text{ as } \epsilon \to 0,$$

then,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(J_{(k+1)\epsilon} = j|J_{k\epsilon} = i\right) \to t_{ij},$$

and

$$\mathbb{E}_{\boldsymbol{\theta}}[N_{ij}^\epsilon|\mathbf{x}] \to t_{ij}\frac{\boldsymbol{\pi}^\top C_1(0, x_1, i, j, \mathbf{T})\mathbf{G}_1 e^{\mathbf{T}(x_2 - x_1)}\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

For $x_1 > x_2$, $i, j \in \Gamma_1^c \cap \Gamma_2^c$,

$$\mathbb{E}_\theta[N_{ij}|\mathbf{x}] = t_{ij}\frac{\boldsymbol{\pi}^\top C_1(0, x_2, i, j, \mathbf{T})\mathbf{G}_2 e^{\mathbf{T}(x_1-x_2)}\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

For $x_1 < x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$, $j \in \Gamma_1'$,

$$\begin{aligned}
\mathbb{E}_\theta[N_{ij}^\epsilon|\mathbf{x}] &= \mathbb{P}_\theta\left(J_{x_1-\epsilon} = i, J_{x_1} = j|\mathbf{x}\right) \\
&= \frac{\mathbb{P}_\theta\left(J_{x_1-\epsilon} = i\right)\mathbb{P}_\theta\left(J_{x_1} = j|J_{x_1-\epsilon} = i\right)\mathbb{P}_\theta\left(X_2 \in dx_2|J_{x_1} = j\right)}{\mathbb{P}_\theta\left(\mathbf{X} \in d\mathbf{x}\right)} \\
&= \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}(x_1-\epsilon)}\mathbf{E}_{ii}e^{\mathbf{T}\epsilon}\mathbf{e}_j\mathbb{P}_\theta\left(X_2 \in dx_2|J_{x_1} = j\right)}{\mathbb{P}_\theta\left(\mathbf{X} \in d\mathbf{x}\right)} \\
&\to -t_{ij}\frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{E}_{ij}e^{\mathbf{T}(x_2-x_1)}\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.
\end{aligned}$$

For $x_1 > x_2$, $i \in \Gamma_1^c \cap \Gamma_2^c$, $j \in \Gamma_2'$,

$$\mathbb{E}_\theta[N_{ij}|\mathbf{x}] = -t_{ij}\frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_2}\mathbf{E}_{ij}e^{\mathbf{T}(x_1-x_2)}\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

For $x_1 < x_2$, $i, j \in \Gamma_1'$,

$$\begin{aligned}
\mathbb{E}_\theta[N_{ij}^\epsilon|\mathbf{x}] &= \sum_{k=[x_1/\epsilon]}^{[x_2/\epsilon]-1}\mathbb{P}_\theta\left(J_{k\epsilon} = i, J_{(k+1)\epsilon} = j|\mathbf{x}\right) \\
&= \sum_{k=x_1/\epsilon}^{x_2/\epsilon-1}\frac{\mathbb{P}_\theta\left(X_1 \in dx_1, J_{k\epsilon} = i, J_{(k+1)\epsilon} = j, X_2 \in dx_2\right)}{\mathbb{P}_\theta\left(\mathbf{X} \in d\mathbf{x}\right)} \\
&= \sum_{k=[x_1/\epsilon]}^{[x_2/\epsilon]-1}\mathbb{P}_\theta\left(X_1 \in dx_1\right)\mathbb{P}_\theta\left(J_{k\epsilon} = i, J_{(k+1)\epsilon} = j|X_1 \in dx_1\right) \\
&\quad \cdot\frac{\mathbb{P}_\theta\left(X_2 \in dx_2|J_{(k+1)\epsilon} = j\right)}{\mathbb{P}_\theta\left(\mathbf{X} \in \mathbf{x}\right)}, \\
&\to t_{ij}\frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{G}_1 C_1(x_1, x_2, i, j, \mathbf{T})\mathbf{T}\mathbf{g}_2\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.
\end{aligned}$$

For $x_1 > x_2$, $i, j \in \Gamma_2'$,

$$\mathbb{E}_\theta[N_{ij}|\mathbf{x}] = t_{ij}\frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_2}\mathbf{G}_2 C_1(x_2, x_1, i, j, \mathbf{T})\mathbf{T}\mathbf{g}_1\mathbf{e}}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

For $x_1 < x_2$, $i \in \Gamma_1'$, $j = m+1$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[N_{ij}^{\epsilon}|\mathbf{x}] &= \frac{\mathbb{P}(X_1 \in dx_1)\,\mathbb{P}(J_{x_2-\epsilon} = i|X_1 \in dx_1)\,\mathbb{P}(J_{x_2} = j|J_{x_2-\epsilon} = i)}{\mathbb{P}(\mathbf{X} \in d\mathbf{x})} \\ &\to -t_i \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_1}\mathbf{G}_1 e^{\mathbf{T}(x_2-x_1)}\mathbf{e}_i}{f(\mathbf{x}|\boldsymbol{\theta})}.\end{aligned}$$

For $x_1 > x_2$, $i \in \Gamma_2'$, $j = m+1$,

$$\mathbb{E}_{\boldsymbol{\theta}}[N_{ij}|\mathbf{x}] = -t_i \frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x_2}\mathbf{G}_2 e^{\mathbf{T}(x_1-x_2)}\mathbf{e}_i}{f(\mathbf{x}|\boldsymbol{\theta})}.$$

The *BPH* distribution with a mass on $X_1 = X_2$ occurs rarely in application but is still worthy of mention. Based on the interpretation of the sub-intensity matrix $\mathbf{T}$, this case happens when there is a positive probability of moving from some state in $\Gamma_1^c \cap \Gamma_2^c$ directly to the absorbing state $\Gamma_1 \cap \Gamma_2 = m+1$. Indeed, all calculations will be on the set $\Gamma_1^c \cap \Gamma_2^c$ because the only state visited on $\Gamma_1 \cup \Gamma_2$ is the absorbent state $m+1$. This case is very similar to the univariate case which is treated in details in Asmussen *et al.* (1996). The expectations of the E-step are

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[B_i|\mathbf{x}] &= \frac{\pi_i \mathbf{e}_i e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}}{\boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}},\ i \in \Gamma_1^c \cap \Gamma_2^c, \\ \mathbb{E}_{\boldsymbol{\theta}}[Z_i|\mathbf{x}] &= \frac{\boldsymbol{\pi}^\top C_1(0,x,i,i,\mathbf{T})\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}}{\boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}},\ i \in \Gamma_1^c \cap \Gamma_2^c, \\ \mathbb{E}_{\boldsymbol{\theta}}[N_{ij}|\mathbf{x}] &= t_{ij}\frac{\boldsymbol{\pi}^\top C_1(0,x,i,j,\mathbf{T})\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}}{\boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}},\ i,j \in \Gamma_1^c \cap \Gamma_2^c, \\ \mathbb{E}_{\boldsymbol{\theta}}[N_{ij}|\mathbf{x}] &= -t_i\frac{\boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}_i}{\boldsymbol{\pi}^\top e^{\mathbf{T}x}\mathbf{g}_1\mathbf{g}_2\mathbf{T}\mathbf{e}},\ i \in \Gamma_1^c \cap \Gamma_2^c,\ j = m+1.\end{aligned}$$

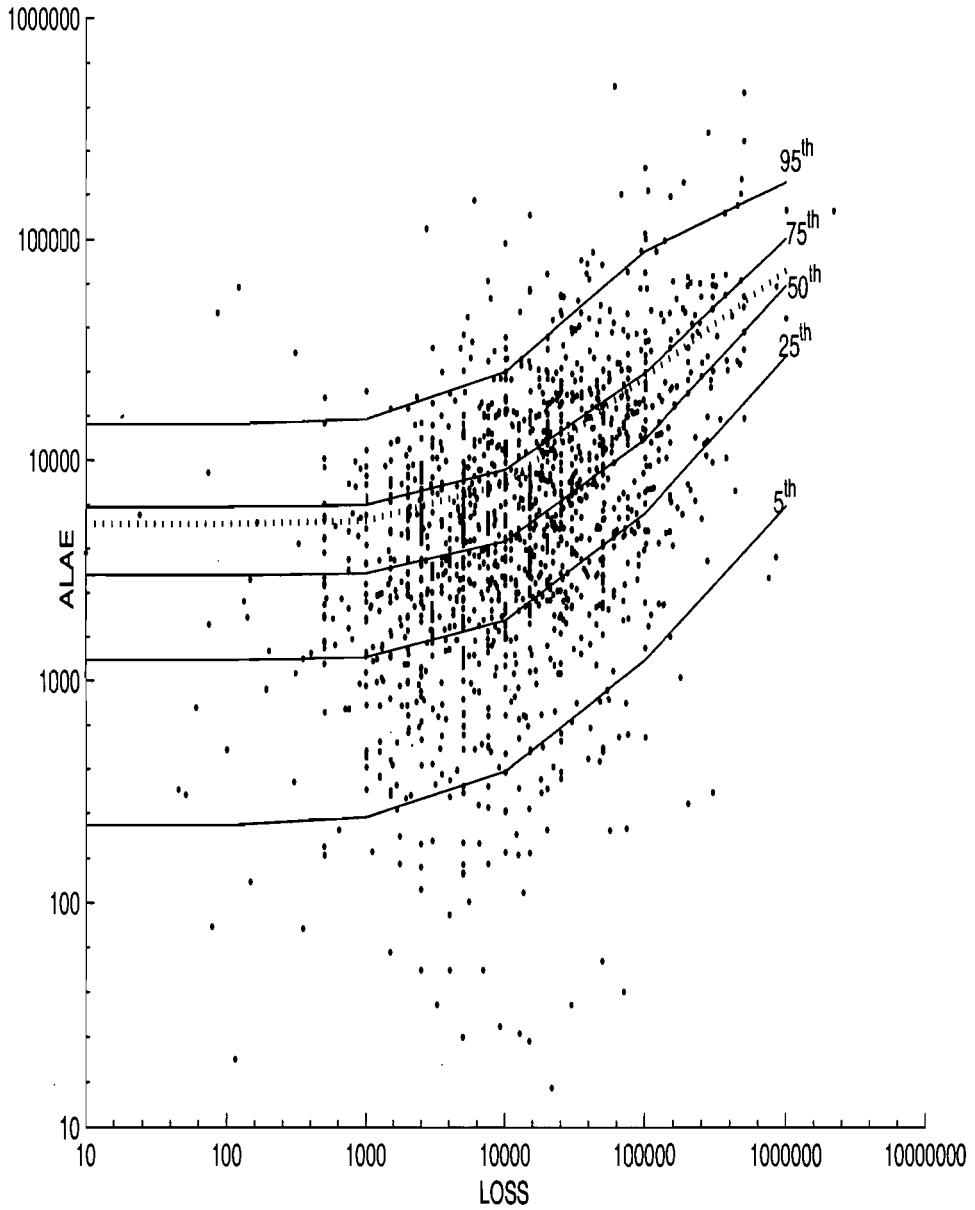

FIG. 2.1. ALAE versus LOSS with curves for conditional quantiles and mean. The dotted curve is the conditional mean of ALAE given LOSS.

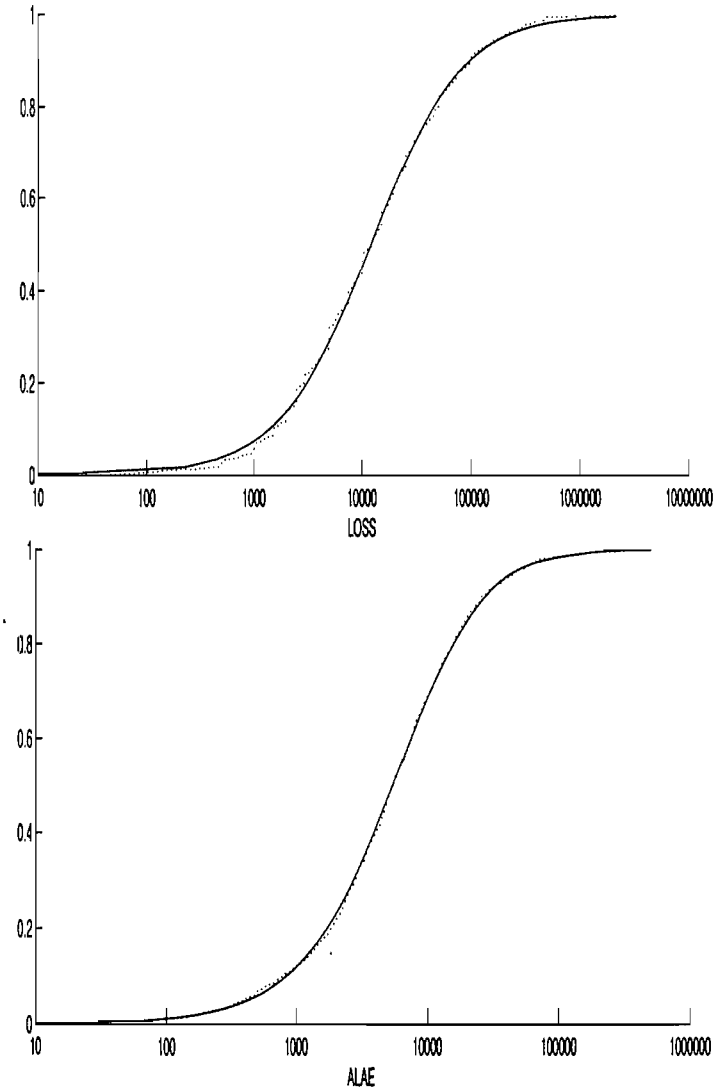

FIG. 2.2. Marginal fitted distribution functions (smooth) and empirical distribution functions (dotted).

Chapitre 3

EXACT CREDIBILITY WITH PHASE-TYPE DISTRIBUTIONS

ABSTRACT

Phase-type random variables are defined as the time until absorption in a set of absorbent states of a Markov chain environment. Coxian, Erlang-n, hyper-exponential and mixtures of Erlang-n random variables are special cases of PH random variables. PH distributions have the property of being dense among all distributions with positive support. This is useful in approximating claim severity distributions by PH distributions.

In credibility theory Jewell (1974a) and Jewell (1974b) prove how exact credibility occurs in the univariate and multivariate exponential family of conditional distributions, when paired with the appropriate conjugate prior. Here exact credibility is discussed in a univariate and multivariate PH setting. Hidden Markov chains are used, embedding the unobservable risk parameters in the PH distribution parameters. This extends the results Jewell outside the exponential family of distributions.

Key words : Phase-Type Distributions, Coxian Distributions, Exact Credibility, Jewell's Theorem, Continuous Markov Processes

3.1. PRELIMINARIES ON CREDIBILITY THEORY

3.1.1. Preview

Credibility theory deals with the calculation of premiums in an insurance portfolio containing heterogeneous risks (contracts) that share some common risk characteristics. Let X_n represents the claim of a particular contract in year n, for $n = 1, \ldots, N$, based on N years of observations. Furthermore, assume that past years premiums were based on a "manual" premium rate, M, possibly different from $\bar{X} = \sum_{n=1}^{N} X_n/N$. If past claims indicate that there is a significant difference between M and $\bar{X}$, then M may not be suitable for the coming period premium. As a result there are two natural questions : (1) how credible is past data ? and (2) what should the premium be next year for this contract, or more precisely, should we consider only the individual contract effects (that is $\bar{X}$), only the portfolio characteristics (i.e. M), or a combination of them ?

A natural choice with ample historical and practical evidence in its favour, is the linear credibility premium, *i.e.*,

$$Z\,\bar{X} + (1 - Z)\,M\,,$$

where the credibility factor Z, $0 \leq Z \leq 1$, is to be determined. This is an important matter that has attracted considerable research interest. The larger Z, the more credible the data in each contract.

The first attempt by Mowbray (1914) was mostly on the first question above. Using introductory statistical methods he identified what sample size is sufficient to consider past observations credible, and assign full credibility ($Z = 1$) to $\bar{X}$. He proposed a formula also for Z when there is only partial credibility (*i.e.* $0 < Z < 1$). This approach is called Limited Fluctuations Credibility Theory. For a deeper analysis on this limited fluctuations approach see Klugman *et al.* (2008) and Norberg (1979).

Bayesian credibility theory has been considered by many authors. To proceed, let us assume that each contract may be characterized by an unobservable risk parameter Θ which varies in value by contract. Θ may be vector-valued. For example a car insurer may categorize policyholders from very good, good and

bad to very bad drivers. A risk classification rule in each category, or rate class, can be for instance the average number of observed accidents in the past N years by the policyholder. Of course Θ varies by contract, therefore it might be considered random with a probability density function $\pi(\theta)$.

It is acceptable to assume that claims in each rating class are conditionally independent given Θ, and have an identical distribution. That means $X_1, \ldots, X_{N+1}$ are conditionally *iid.* with conditional density $f_{X|\Theta}(x|\theta)$, where the claim amounts X_i have been adjusted for inflation or other trends.

The fair premium $\mu(\theta) = \mathbb{E}(X_{N+1}|\Theta = \theta)$ would be ideal to predict the future year net premium, but unfortunately the parameter θ is unobservable. What is known here are the observations for the past N years, hence the next natural candidate for the premium is the predictive mean $\mathbb{E}(X_{N+1}|X_1, \ldots, X_N)$, also called the Bayesian premium. In some cases, under a proper choice of distributions, the Bayesian premium equals the following linear credibility premium

$$M^a = Z\,\bar{x} + (1 - Z)\,\mu\,,$$

where $\mu = \mathbb{E}(X) = \mathbb{E}\big[\mu(\Theta)\big]$ plays the role of the manual premium here. In such cases, it is said that exact credibility occurs.

Bühlmann (1967) shows how to approximate the Bayesian premium with a linear combination of past observed claims, that minimizes the expectation of a square loss function with respect to Θ and $X_1, \ldots, X_N$. As a result he shows that the estimator of the predictive mean is the linear credibility premium with

$$Z = \frac{N}{N + \dfrac{\mathbb{E}_\pi\big[\mathbb{V}(X|\Theta)\big]}{\mathbb{V}_\pi\big[\mu(\Theta)\big]}}\,,$$

where $\mathbb{E}_\pi$ and $\mathbb{V}_\pi$ are the mean and variance with respect to the prior distribution π of Θ. Bailey (1950) and Mayerson (1964) find that the linear credibility premium is the exact Bayesian premium for some combinations of prior (also called *structural*) and claims distributions. Jewell (1974a) extends these exact credibility results to the univariate exponential family of distributions with a proper choice of prior. His main result is a special case of the following theorem.

Theorem 3.1. *Suppose that the X_n in $\boldsymbol{X} = (X_1, \ldots, X_{N+1})^\top$ are conditionally independent, given Θ, with common probability density function*

$$f_{X|\Theta}(x_j|\theta) = \frac{p(x_j)e^{r(\theta)x_j}}{q(\theta)}, \qquad x_j \in \chi, \theta \in \Omega, \tag{3.1}$$

and the prior density is a natural conjugate

$$\pi(\theta) = \frac{\big[q(\theta)\big]^{-k} e^{\mu k r(\theta)} r'(\theta)}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1, \tag{3.2}$$

where $-\infty \leq \theta_0 < \theta_1 \leq \infty$, with $\pi(\theta_0) = \pi(\theta_1) = 0$, $\mu = \mathbb{E}(X)$ and $k = \frac{\mathbb{E}[\mathbb{V}(X|\Theta)]}{\mathbb{V}[\mathbb{E}(X|\Theta)]}$, then exact credibility occurs, with

$$\mathbb{E}(X_{N+1}|x_1, \ldots, x_N) = Z\,\bar{x} + (1 - Z)\,\mu,$$

with $Z = \frac{N}{N+k}$.

For a proof see Jewell (1974a) and Klugman *et al.* (2008), p. 594.

Landsman (1998) extends Jewell's Theorem 3.1 to the larger exponential dispersion family. In a parameterization similar to that of Klugman *et al.* (2008), p. 594, its probability density functions are written as :

$$f_{X|\Theta}(x|\theta) = \frac{p(\lambda, x)\, e^{\lambda r(\theta)x}}{\big[q(\theta)\big]^{\lambda}}, \qquad x \in \chi, \theta \in \Omega. \tag{3.3}$$

The introduction of the dispersion parameter λ makes this a more flexible family of distributions than the linear exponential family ($\lambda = 1$). The natural conjugate prior on θ remains the same as in (3.2) and exact credibility still occurs, now with $Z = \frac{\lambda N}{\lambda N + k}$. Other properties of (3.3) are that

$$\mu(\theta) = \mathbb{E}(X|\Theta = \theta) = \frac{q'(\theta)}{r'(\theta)q(\theta)} \quad \text{and} \quad \sigma^2(\theta) = \mathbb{V}(X|\Theta = \theta) = \frac{\mu'(\theta)}{\lambda\, r'(\theta)}. \tag{3.4}$$

Using the natural conjugate in (3.2) gives $\mu = \mathbb{E}\big[\mu(\Theta)\big]$ and

$$k = \mathbb{E}\big[\mathbb{V}(X|\Theta)\big] / \mathbb{V}\big[\mathbb{E}(X|\Theta)\big].$$

For a comprehensive treatment on the exponential dispersion family, see Tweedie (1984), Nelder and Wedderburn (1972) or Jørgensen (1986) and Jørgensen (1987).

3.1.2. Extensions

If X_n represents the claim in one single contract for period $n = 1, \ldots, N+1$, it is reasonable to assume that for a fixed integer m, X_n can be written as a sum of m independent random variables. For example, $m = 4$ can be interpreted as the number of seasons, where the total amount in one year equals to the sum of claims occurred in each season. Another example could be in car insurance, when for one single policy there are two or more drivers insured on the same car and the total claim for the car equals to sum of claims for each driver. Therefore

$$X_n = \sum_{i=1}^{m} Y_{in}\,, \tag{3.5}$$

where Y_{in} are conditionally independent given a risk parameter Θ, but not necessarily identically distributed. In practice, it may be common that the insurer does not observe or keep a record of the Y_{in}'s, but only of the X_n's. The following theorem shows that if exact credibility occurs for Y_{in}, with a proper prior on Θ, so is true for X_n in (3.5).

Theorem 3.2. *Assume that exact credibility occurs for independent random variables $\{Y_{in}\}_{i=1}^{m}$, with a proper prior distribution of Θ, such that the credibility factors Z are the same for all i, that is,*

$$\mathbb{E}\{Y_{i,N+1}|y_{i1}, \ldots, y_{iN}\} = Z\,\bar{y}_i + (1-Z)\,\mathbb{E}(Y_{in})\,, \qquad \textit{for } i = 1, \ldots, m\,.$$

If $X_n = \sum_{i=1}^{m} \alpha_i\, Y_{in}$, is a weighted average for given constants m and $\alpha_1, \ldots, \alpha_m$, then exact credibility also occurs for X_n with the same credibility factor Z.

Proof. Due to the exact credibility assumption on the $\{Y_{in}\}$, we have that

$$\begin{aligned}\mathbb{E}\{(Y_{1,N+1}, \ldots, Y_{m,N+1})^{\top}|\mathbf{y}_1, \ldots, \mathbf{y}_N\} &= Z\,(\bar{y}_1, \ldots, \bar{y}_m)^{\top} \\ &+ (1-Z)\,\big(\mathbb{E}(Y_{1n}), \ldots, \mathbb{E}(Y_{mn})\big)^{\top},\end{aligned}$$

where $\mathbf{y}_n = (y_{1n}, \ldots, y_{mn})$ and $\bar{y}_i = \sum_{n=1}^{N} y_{in}/N$.

As a result we have

$$\mathbb{E}\{X_{N+1}|\mathbf{y}_1, \ldots, \mathbf{y}_N\} = \frac{Z}{N}\Big(\sum_{n=1}^{N}\sum_{i=1}^{m} \alpha_i y_{in}\Big) + (1-Z)\,\mathbb{E}(X_n)\,.$$

Taking conditional expectations on both sides and keeping in mind that the σ-algebras generated by the X_n's and Y_{in}'s satisfy

$$\sigma(X_1, \ldots, X_N) \subseteq \sigma(Y_{11}, \ldots, Y_{1N}, \ldots, Y_{m1}, \ldots, Y_{mN}),$$

by the tower property of conditional expectation, we get the following result :

$$\mathbb{E}\{X_{N+1} | x_1, \ldots, x_N\} = \frac{Z}{N} \mathbb{E}\{\sum_{n=1}^{N} \sum_{i=1}^{m} \alpha_i y_{in} | x_1, \ldots, x_N\} + (1 - Z)\, \mathbb{E}(X_n),$$

which equals

$$Z\, \bar{x} + (1 - Z)\, \mathbb{E}(X_n).$$

□

Example 3.1. *As an illustrative example where the assumptions of Proposition 3.2 and the conclusions are met, assume that* $Y_{in} = \alpha_i Y$ *where* Y *is a member of the exponential dispersion family in* (3.3) *and* $\alpha_i > 0$, *for* $i = 1, \ldots, m$ *are known coefficients. Then, for any fixed year* $n = 1, \ldots, N$, *the probability density function of a given* Y_{in} *is given by*

$$f_{Y_{in} | \Theta}(y \,|\, \theta) = f_{Y | \Theta}(y/\alpha_i \,|\, \theta)/\alpha_i = \frac{p(\lambda, y/\alpha_i)\, e^{\lambda r(\theta) y/\alpha_i}}{\alpha_i \left[q(\theta)\right]^{\lambda}}, \qquad y > 0.$$

This is not a member of the exponential family in (3.1), *nor of the exponential dispersion family* (3.3). *But we can still condition all* Y_{in}, *for* $i = 1, \ldots, m$, *on the same* θ *and assume that the prior distribution is a member of* (3.2). *Then the posterior distribution, given* $\tilde{y}_N = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$, *becomes*

$$\pi(\theta \,|\, \tilde{y}_N) \propto q(\theta)^{-(\lambda N + k)}\, e^{r(\theta) \left[\frac{\lambda N}{\alpha_i} \bar{y}_i + \mu k\right]}\, r'(\theta),$$

which is of the same form as (3.2), *but for parameters*

$$k^* = \lambda N + k \quad \textit{and} \quad \mu_i^* = \Big[\frac{\lambda N}{\alpha_i} \bar{y}_i + \mu k\Big]/(\lambda N + k) = \frac{\lambda N}{\lambda N + k} \frac{\bar{y}_i}{\alpha_i} + \frac{k}{\lambda N + k} \mu.$$

Hence the posterior mean below is of the same form as the prior mean $\mathbb{E}(Y_{i,N+1}) = \alpha_i\,\mathbb{E}(Y) = \alpha_i\,\mathbb{E}\big[\mathbb{E}(Y \mid \Theta)\big] = \alpha_i\,\mathbb{E}\big[\mu(\Theta)\big] = \alpha_i\,\mu$, *but for parameter* μ_i^* :

$$\begin{aligned}
\mathbb{E}(Y_{i,N+1} \mid \tilde{y}_N) &= \int_{\theta_0}^{\theta_1} \mathbb{E}(Y_{i,N+1} \mid \theta)\,\pi(\theta \mid \tilde{y}_N)\,d\theta \\
&= \alpha_i \int_{\theta_0}^{\theta_1} \frac{q'(\theta)}{r'(\theta)q(\theta)}\,\pi(\theta \mid \tilde{y}_N)\,d\theta\,, \qquad \textit{by (3.4)}\,, \\
&= \alpha_i\,\mu_i^* = \frac{\lambda N}{\lambda N + k}\,\bar{y}_i + \frac{k}{\lambda N + k}\,\alpha_i\,\mu = (1 - Z)\,\bar{y}_i + Z\,\mathbb{E}(Y_{in})\,,
\end{aligned}$$

which means that exact credibility still occurs for Y_{in}, *with the same credibility factor* $Z = \lambda N/(\lambda N + k)$, *for all* $i = 1, \ldots, m$.

One important question arises from the previous theorem : what if m is a random variable ? The credibility problem differs when total claims in a period, X_n, are random sums. We try here to answer this question, at least to some extent.

First suppose that M, the random variable version of m, is independent from Y_{jn} in (3.5) and such that $\mathbb{E}(M) < \infty$. The portfolio interpretation changes as a random M no longer represents the number of seasons or of drivers. Consider instead a group insurance portfolio where the total claims of a given risk class is composed of a random number of iid individual claims Y_{in}. Hence, our portfolio assumptions are as follows :

- $X_n = \sum_{i=1}^{M_n} Y_{in}$, for $n = 1, \ldots, N+1$.
- for fixed i the Y_{in} are conditionally iid random variables, given Θ, for $n = 1, \ldots, N+1$, with conditional mean $\mu_i(\Theta) = \mathbb{E}(Y_{in} \mid \Theta)$ and marginal expectation $\mu = \mathbb{E}[\mu_i(\Theta)]$.
- Prior distributions on Θ in (3.2) are chosen independently such that the exact credibility occurs for Y_{in}, given Θ, with the same credibility factor Z.
- M_n, are *iid* random variables with known probability mass function and independent of Θ and the Y_{in}'s, such that $\mathbb{E}(M_n) < \infty$, for $n = 1, \ldots, N+1$.

In what follows it is shown that under the above conditions, exact credibility occurs.

First denote by $\tilde{m}_N = (m_1, \ldots, m_N)$, then by conditional independence of the Y_{in}'s :

$$\begin{aligned}
\mathbb{E}(Y_{i,N+1} \,|\, \tilde{y}_N, \tilde{m}_N) &= \int_0^\infty y \, f_{Y_{i,N+1}|\tilde{Y}_N,\tilde{M}_N}(y \,|\, \tilde{y}_N, \tilde{m}_N)\, dy \\
&= \int_0^\infty y \, \frac{\int_0^\infty f_{Y_{i,N+1},\tilde{Y}_N,\tilde{M}_N \,|\, \Theta}(y, \tilde{y}_N, \tilde{m}_N \,|\, \theta)\, \pi(\theta)\, d\theta}{f_{\tilde{Y}_N,\tilde{M}_N}(\tilde{y}_N, \tilde{m}_N)}\, dy \\
&= \int_0^\infty \int_0^\infty y \, f_{Y_{i,N+1}\,|\,\Theta}(y \,|\, \theta)\, dy \, \frac{f_{\tilde{Y}_N,\tilde{M}_N \,|\, \Theta}(\tilde{y}_N, \tilde{m}_N \,|\, \theta)\, \pi(\theta)}{f_{\tilde{Y}_N,\tilde{M}_N}(\tilde{y}_N, \tilde{m}_N)}\, d\theta \\
&= \int_0^\infty \mathbb{E}(Y_{i,N+1} \,|\, \Theta = \theta)\, \pi(\theta \,|\, \tilde{y}_N, \tilde{m}_N)\, d\theta = \int_0^\infty \mu_i(\theta)\, \pi(\theta \,|\, \tilde{y}_N, \tilde{m}_N)\, d\theta \,.
\end{aligned}$$

On the other hand, it is easily seen that the posterior distribution is proportional to

$$\begin{aligned}
\pi(\theta \,|\, \tilde{y}_N, \tilde{m}_N) &\propto f_{\tilde{Y}_N \,|\, \Theta,\tilde{M}_N}(\tilde{y}_N \,|\, \theta, \tilde{m}_N)\, \pi(\theta)\,, &(3.6)\\
&\propto e^{r(\theta)(\lambda \sum_{i=1}^N \sum_{j=1}^{m_i} y_{ij} + \mu k)}\, q(\theta)^{-(\lambda \sum_{i=1}^N m_i + k)}\, r'(\theta)\,, &(3.7)
\end{aligned}$$

which is of the same form as the prior distribution, but for $k^* = \lambda \sum_{i=1}^N m_i + k$ and $\mu^* = \frac{\mu k + \lambda \sum_{i=1}^N \sum_{j=1}^{m_i} y_{ij}}{\lambda \sum_{i=1}^N m_i + k}$. Therefore, from Example 3.1,

$$\mathbb{E}(Y_{1,N+1} \,|\, \tilde{y}_N, \tilde{m}_N) = \frac{k}{\lambda \sum_{n=1}^N m_n + k} \mu + \frac{1}{\lambda \sum_{n=1}^N m_n + k} \sum_{n=1}^N \sum_{i=1}^{m_n} y_{in}\,. \qquad (3.8)$$

Taking conditional expectations on both sides of (3.8) with respect to the sigma-field generated by $\tilde{X}_N = (X_1, \ldots, X_N)$ and $\tilde{m}$, yields to

$$\mathbb{E}(Y_{1,N+1} | \tilde{X}_N, \tilde{m}_N) = \frac{N}{\lambda \sum_{n=1}^N m_n + k}\, \bar{x} + \frac{k}{\lambda \sum_{n=1}^N m_n + k}\, \mu\,. \qquad (3.9)$$

By the tower property of conditional expectation we have

$$\begin{aligned}
\mathbb{E}(X_{N+1} \,|\, \tilde{X}_N, \tilde{m}_N) &= \mathbb{E}\Big(\sum_{i=1}^{M_{N+1}} Y_{i,N+1} \,|\, \tilde{X}_N, \tilde{m}_N\Big) \\
&= \mathbb{E}\Big[\mathbb{E}\Big(\sum_{i=1}^{M_{N+1}} Y_{i,N+1} \,|\, \tilde{X}_N, \tilde{m}_N, M_{N+1}\Big) \,|\, \tilde{X}_N, \tilde{m}_N\Big] \\
&= \mathbb{E}\Big[M_{N+1}\, \mathbb{E}\big(Y_{1,N+1} \,|\, \tilde{X}_N, \tilde{m}_N, M_{N+1}\big) \,|\, \tilde{X}_N, \tilde{m}_N\Big] \\
&= \mathbb{E}(M_{N+1} \,|\, \tilde{X}_N, \tilde{m}_N)\, \mathbb{E}(Y_{1,N+1} \,|\, \tilde{X}_N, \tilde{m}_N)\,,
\end{aligned}$$

where $\mathbb{E}(M_{N+1} \mid \tilde{X}_N, \tilde{m}_N) = \mathbb{E}(M_{N+1})$, by independence. Then from (3.9)

$$\mathbb{E}(X_{N+1} \mid \tilde{X}_N, \tilde{m}_N) = \frac{N\mathbb{E}(M_{N+1})}{\lambda \sum_{n=1}^{N} m_n + k}\, \bar{x} + \frac{k\,\mathbb{E}(M_{N+1})}{\lambda \sum_{n=1}^{N} m_n + k}\, \mu\,.$$

Theorem 3.3. *Under the above portfolio assumptions, exact credibility occurs for* $X_n = \sum_{i=1}^{M_n} Y_{in}$ *with credibility factor* $\frac{N\mathbb{E}(M_{N+1})}{\lambda \sum_{n=1}^{N} m_n + k}$.

The credibility factor $\frac{N\mathbb{E}(M_{N+1})}{\lambda \sum_{n=1}^{N} m_n + k}$ is a function of N, and as $N \to \infty$ the credibility factor tends to 1.

Two assumptions are imposed in Theorem 3.3 which may not be hold for some insurance portfolios. The random variables $(M_1, \ldots, M_N)$ may not be observable. For example, in group insurance, the total claims may be recorded but not the number of claims. Independence between the random variables M and Θ is another assumption that may not hold in practice.

The following section presents applications of credibility to random sums with *phase-type* distributions.

3.2. CREDIBILITY THEORY FOR UNIVARIATE *PH*

3.2.1. Phase-type distributions

A random variable that is defined as the absorption time of an evanescent finite-state continuous-time Markov chain is said to have a phase-type (PH) distribution. Since their introduction by Neuts (1995) in 1981, PH distributions have been used in a wide range of stochastic modeling applications in areas as diverse as telecommunications, biostatistics, queueing theory, risk theory, reliability theory, and survival analysis.

Erlang (1917), extended the exponential distribution to Erlang or gamma distributions by the method of stages. He defined an Erlang distribution as a non-negative random variable which equals to summation of a fixed number of exponential random variables with a common parameter θ. Stages refer to the number of exponential random variables summed.

In 1955, Cox (1955) generalized Erlang's notion by allowing for a complex parameter θ. This construction, although it rarely finds a simple probabilistic interpretation, defines the class of distributions with rational Laplace-Stieltjes

transform, of which the class of *PH* distributions is a proper subset. These distributions are nowadays also known as matrix-exponential distributions. Neuts in 1981 generalized in Erlang's method of stages in a different way. He defines a phase-type random variable as the time until absorption into an absorbent state in a continuous-time Markov chain. *PH* random variables have very nice properties which make them tractable and attractive for applications.

PH distributions are dense in the class of all distributions defined on the non-negative real numbers. Quantities of interest, such as the distribution and density functions, the Laplace-Stieltjes transform, and the moments of PH random variables are given in closed form. However, they also have some drawbacks, such as the non-uniqueness of the parameter representation. This creates difficulties for statistical inference with *PH* distributions (see O'Cinneide (1989) and Asmussen *et al.* (1996)). For a comprehensive review of *PH* distributions see Neuts (1995).

Consider $J = \{J_t\,;\, t \geq 0\}$, a right-continuous Markov process on the finite state space $E = \{1, 2, \ldots, \Delta, \Delta+1\}$, with initial probability vector $\boldsymbol{\alpha}$ and infinitesimal generator matrix $\mathbf{A}$, such that only the state $\Delta + 1$ is absorbent. Hence, absorption in it is certain.

The infinitesimal generator matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{B}_0 \\ \mathbf{0} & 0 \end{pmatrix},$$

where the matrix $\mathbf{B} = (b_{ij})$ is $\Delta \times \Delta$ and $\mathbf{B}_0 = (b_{0j})$ is an Δ-dimensional column vector. These elements satisfy the conditions $b_{ii} < 0$, for $i = 1, \ldots, \Delta$, $b_{ij} \geq 0$, for $i \neq j$, and $\mathbf{B}\mathbf{e} + \mathbf{B}_0 = \mathbf{0}$, where $\mathbf{e}$ is a column vector of ones. States $1, \ldots, \Delta$ are transient if and only if $\mathbf{B}$ is nonsingular, see Neuts (1995). In this article, we assume that the initial probability $\alpha_{\Delta+1} = 0$, and hence $\boldsymbol{\alpha}$ can be written as $\boldsymbol{\alpha} = (\boldsymbol{\beta}, 0)$, with $\boldsymbol{\beta}\,\mathbf{e} = 1$.

A random variable X is called a *phase-type* (*PH*), with representation $(\boldsymbol{\beta}, \mathbf{B})$, if it represents the time until absorption into $\Delta + 1$, *i.e.*

$$X = \inf\left\{u\,;\, J_u \in \{\Delta + 1\}\right\}.$$

If X has a PH distribution, using the theory of Markov chains, it can be shown (see Neuts (1995)) that

$$\mathbb{P}\{X > x\} = \boldsymbol{\beta}\, e^{\mathbf{B}x}\, \mathbf{e}\,, \qquad \text{for } x \geq 0\,.$$

Therefore, the probability density function of X is given by

$$\begin{aligned} f_X(x \,|\, \boldsymbol{\beta}, \mathbf{B}) &= -\boldsymbol{\beta}\, e^{\mathbf{B}x}\, \mathbf{B}\, \mathbf{e} \\ &= \boldsymbol{\beta}\, e^{\mathbf{B}x}\, \mathbf{B}_0\,, \qquad \text{for } x \geq 0\,. \end{aligned} \tag{3.10}$$

The pair $(\boldsymbol{\beta}, \mathbf{B})$ is known as a representation of order Δ of the PH distribution.

As a convention, we shall not mention the dimensions of vectors and matrices. It is usually possible to determine them from the context.

Due to the non-uniqueness of the representation of PH distributions, other pairs $(\boldsymbol{\beta}^*, \mathbf{B}^*)$ of same or different dimension may define the same probability density function (see O'Cinneide (1989)).

3.2.2. Credibility theory

In this section we establish that exact credibility occurs for some PH distributions in the univariate and multivariate cases. Instead of considering the probability density function in (3.10) directly, we use a simpler technique. The following lemma is needed to relate a PH random variable with sums of exponential random variables.

Lemma 3.1. *Every* PH *random variable X with representation $(\boldsymbol{\beta}, \boldsymbol{B})$ can be written as a sum of Δ random sums of independent exponential random variables with parameters $\theta_1, \ldots, \theta_\Delta$, where $\theta_i = -b_{ii}$, for $i = 1, \ldots, \Delta$ and Δ is the dimension of matrix $\boldsymbol{B}$.*

Proof. *The observation X of a PH random variable with representation $(\boldsymbol{\beta}, \boldsymbol{B})$ is simply the hitting time of state $\Delta + 1$, which can be written as the sum of the sojourn times S_i in each of the non-observable states $i = 1, \ldots, \Delta$,* i.e.

$$X = \sum_{i=1}^{\Delta} S_i\,, \tag{3.11}$$

where S_i is the total time that the Markov process J spends in state i. Each time J visits state i, it stays a random time (which has exponential distribution with

parameter $\theta_i = -b_{ii}$) in state i and then moves to another state. Hence,

$$S_i = \sum_{j=1}^{M_i} Y_{ij}\,, \tag{3.12}$$

where M_i is the total number of jumps out of state i. Indeed, it counts total jumps to all other states. In other words, $M_i = \sum_{j,j\neq i}^{\Delta+1} M_{ij}$, and M_{ij} is the number of jumps from state i to j, while Y_{ij} are the sojourn times into state j, coming from state i, for $i = 1,\ldots,\Delta$ and $j = 1,\ldots,\Delta+1$. The proof is obtained considering the Y_{ij}'s as independent exponential random variables with parameter θ_i, for $j = 1,\ldots,M_i$. For the joint distribution of M_{ij} and $(S_i)_{i=1}^{\Delta}$ see Asmussen et al. (1996).

□

A *PH* random variable with dimension Δ has $\Delta+\Delta^2-1$ parameters. Choosing the right parameters and giving a prior to these parameters is a delicate problem. The more parameters chosen for the prior distribution, the less tractable the calculations. On the other hand, too few parameters may not provide an accurate description of the risk features.

As seen from Lemma 3.1 every *PH* random variable is written as a sum of random sums of exponentials. Consider the diagonal elements of the sub-intensity matrix $\mathbf{B}$ as our choice for the risk parameter. Hence, we can specify a prior distribution on the parameters of the distribution of Y_{ij} in (3.12). This is a natural choice as each non-diagonal element of $\mathbf{B}$ is a fraction of the diagonal element in the same row, *i.e.*

$$\begin{pmatrix} -\theta_1 & p_{11}\,\theta_1 & \ldots & p_{1\Delta}\,\theta_1 \\ p_{21}\,\theta_2 & -\theta_2 & \ldots & p_{2\Delta}\,\theta_2 \\ \vdots & \ddots & \ddots & \vdots \\ p_{\Delta 1}\,\theta_\Delta & \ldots & \ldots & -\theta_\Delta \end{pmatrix}, \tag{3.13}$$

where $0 \leq p_{lk} \leq 1$ and $\sum_{k=1,k\neq l}^{\Delta} p_{lk} \leq 1$, for $l = 1,\ldots,\Delta$.

Here the mean and variance of Y_{ij} in (3.12), exponentially distributed with parameter θ_i, are given by $\frac{1}{\theta_i}$ and $\frac{1}{\theta_i^2}$. Hence, the bigger θ_i, the smaller the mean of S_i and of X, the *PH* random variable in (3.11). We assume that the coefficients p_{lk} are constants for $k,l = 1,\ldots,\Delta$, $k \neq l$ and that the initial probability vector

$\boldsymbol{\beta}$ is

$$(1, 0, \ldots, 0)\,, \tag{3.14}$$

that is, $J_0 = 1$ with probability 1.

Now assume X_n to be the claim size for period $n = 1, \ldots, N+1$, and that it follows a PH distribution with representation $(\boldsymbol{\beta}, \mathbf{B})$ as in (3.13) and (3.14). The corresponding embedded Markov chains are denoted by $J_1, \ldots, J_{N+1}$. According to Lemma 3.1, we can represent X_n as

$$X_n = \sum_{i=1}^{\Delta} S_{in} = \sum_{i=1}^{\Delta} \sum_{j=1}^{M_{in}} Y_{ijn}\,, \tag{3.15}$$

where M_{in} is the number of jumps from state i in period n and Y_{ijn} is the claim in period i, year n. The risk parameter vector is given by

$$\theta = (\theta_1, \ldots, \theta_\Delta)\,. \tag{3.16}$$

As is natural in credibility theory, we assume that, given $\Theta = \theta$, the $J_1, \ldots, J_{N+1}$ in (3.16) are conditionally independent. As a result, given θ_i, the sojourn time spent in state i after a jump into it (which has an exponential distribution) in period n is independent and identically distributed with the time spent in the state i after a jump into it in period n', for $n' \neq n$ and $n' = 1, \ldots, N+1$.

Mathematically, given θ the set of $Y_{i11}, \ldots, Y_{iM_{i1}1}, \ldots, Y_{i1,N+1}, \ldots, Y_{iM_{i,N+1},N+1}$ are conditionally *iid* exponential random variables with parameter θ_i, for $i = 1, \ldots, \Delta$.

The prior distribution on the risk parameter may be chosen in two ways; univariate or multivariate. In the multivariate case the prior is put on each element of θ in (3.16), independently chosen and from the natural conjugate of an exponential distribution (3.2), *i.e.* with $r(\theta_i) = -\theta_i$ and $q(\theta_i) = 1/\theta_i$ in

$$\pi(\theta_i) = \frac{\left[q(\theta_i)\right]^{-k} e^{\mu_i k r(\theta_i)} r'(\theta_i)}{c(\mu_i, k)}\,, \qquad \theta_0 < \theta_i < \theta_1\,.$$

That is the risk parameter is vector valued with independent but not identically distributed components, as μ_i may vary for different $i = 1, \ldots, \Delta$.

Another possible choice is to consider θ of the form

$$\theta = (\gamma_1 \theta_1, \gamma_2 \theta_1, \ldots, \gamma_\Delta \theta_1)\,, \tag{3.17}$$

where the γ_i's are positive constants and the θ_i's are iid from the univariate natural conjugate prior to the exponential distribution, *i.e.*

$$\pi(\theta_i) = \frac{\theta_i^k \, e^{-k \frac{\mu}{\gamma_i}\theta_i}}{c(\mu, k)}, \qquad \theta_0 < \theta_i < \theta_1 \,. \tag{3.18}$$

Then, for both risk parameter cases, defined in (3.16) or (3.17), Landsman and Makov's extension of Jewell's theorem lets us conclude that exact credibility occurs for claims from this exponential distribution with a natural conjugate prior on the parameter θ. If the distribution of the number of jumps from state i is independent from θ, then by Theorem 3.3 we have that

$$\begin{aligned} \mathbb{E}(Y_{i1,N+1} \,|\, \tilde{y}_{i1}, \ldots, \tilde{y}_{iN}, \tilde{m}_i) &= \frac{m_i}{m_i + k} \frac{\sum_{n=1}^{N} \sum_{l=1}^{m_{in}} y_{iln}}{m_i} + \frac{k}{m_i + k} \mu_i \\ &= a_i \frac{s_{i1} + \ldots + s_{iN}}{m_i} + (1 - a_i) \, \mu_i \,, \end{aligned}$$

where $\tilde{y}_{in} = (y_{i1n}, \ldots, y_{im_{i1}n})$, $\tilde{m}_i = (m_{i1}, \ldots, m_{iN})$, $m_i = \sum_{n=1}^{N} m_{in}$, $a_i = \frac{m_i}{m_i + k}$ and $\mu_i = \mu/\gamma_1$. By Theorem 3.3 and (3.15) we have

$$\begin{aligned} \mathbb{E}(S_{i,N+1} | S_{i1}, \ldots, S_{iN}, \tilde{m}_i) &= \mathbb{E}(M_{i,N+1}) \, a_i \frac{s_{i1} + \ldots + s_{iN}}{m_i} + \mathbb{E}(M_{i,N+1}) \, (1 - a_i) \, \mu_i \,, \\ &= \mathbb{E}(M_{i,N+1}) \, a_i \frac{s_{i1} + \ldots + s_{iN}}{m_i} + (1 - a_i) \, \mathbb{E}(S_{i1}) \,, \end{aligned}$$

Now, again by using the tower property of conditional expectations, Proposition 3.2 and (3.15) we have

$$\begin{aligned} \mathbb{E}(X_{N+1} \,|\, x_1, \ldots, x_N) &= \sum_{i=1}^{\Delta} \mathbb{E}\big(M_{i,N+1} \, a_i \frac{s_{i1} + \ldots + s_{iN}}{m_i} \,|\, x_1, \ldots, x_N\big) \\ &\qquad + \sum_{i=1}^{\Delta} \mathbb{E}\big[S_i(1 - a_i) \,|\, x_1, \ldots, x_N\big] \,. \end{aligned}$$

Then by definition a_i

$$\begin{aligned} &\sum_{i=1}^{\Delta} \mathbb{E}\big(M_{i,N+1} \, a_i \frac{s_{i1} + \ldots + s_{iN}}{m_i} \,|\, x_1, \ldots, x_N\big) \\ &\qquad = \sum_{i=1}^{\Delta} \mathbb{E}\big[\frac{M_{i,N+1}}{m_i + k}(s_{i1} + \ldots + s_{jN}) \,|x_1, \ldots, x_N\big] \,. \end{aligned}$$

If, for any $i = 1, \ldots, N$, the ratio $\frac{M_{i,N+1}}{M_i + k}$ is independent of $X_1, \ldots, X_N$ and $S_{i1}, \ldots, S_{iN}$, then exact credibility occurs under the above assumptions.

Theorem 3.4. *With a proper gamma prior distribution on the diagonal elements of matrix* $\boldsymbol{B}$*, then the Bayesian premium equals the linear credibility premium if* $\frac{M_{i,N+1}}{M_i+k}$ *is independent of* $x_1, \ldots, x_N$.

Now we consider an example of *PH* distributions for which exact credibility of the above type occurs.

Example 3.2. *Assume that the vector of initial probabilities is* $\boldsymbol{\beta} = (1, 0, \ldots, 0)$ *and that the sub-intensity matrix* $\boldsymbol{B}$ *is given by*

$$\boldsymbol{B} = \begin{pmatrix} -\theta_1 & \theta_1 & 0 & \cdots \\ 0 & -\theta_2 & \theta_2 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\theta_\Delta \end{pmatrix}.$$

For this choice of $\boldsymbol{B}$*, the embedded Markov chain always starts from state* 1 *with probability* 1*. Then, it moves from state* i *to state* $i+1$*, to finally end in the absorbing state* $\Delta+1$*. Therefore, the number of jumps from state* i *or the number of jumps into the state* i *is simply one. If the prior distribution of* θ_i*, has the form of* $\pi(\theta_i) = \frac{\theta_i^k e^{-k\,\mu_i\,\theta_i}}{c(\mu_i,k)}$*, that is gamma*$(k+1, k\,\mu_i)$ *distribution, for* $i = 1, \ldots, \Delta$*, then exact credibility occurs for the related* PH *random variable. In this case the credibility factor equals to* $Z = \frac{N}{N+k}$.

3.2.3. Exact credibility for Coxian distributions

Coxian distributions (see Cox (1955)) are a special case of *PH* distributions. The *PH* representation of a Coxian distribution with dimension Δ is given by

$$\mathbf{B} = \begin{pmatrix} -\theta_1 & p_1\,\theta_1 & 0 & \cdots \\ 0 & -\theta_2 & p_2\,\theta_2 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\theta_\Delta \end{pmatrix}, \tag{3.19}$$

and

$$\boldsymbol{\beta} = (1, 0, \ldots, 0)\,. \tag{3.20}$$

Let $J = \{J_u\,;\, u \geq 0\}$ be the embedded Markov chain for the Coxian distribution with space states $\{1, \ldots, \Delta, \Delta+1\}$, where $\Delta+1$ is the absorbing state. Then

$J_0 = 1$ with probability 1, means that it always starts from state 1. It spends there a random exponentially distributed time with parameter θ_1 and then moves either to the absorbing state or to state 2, with probabilities $1 - p_1$ and p_1. If J is in state 2, again it sojourns in this state a random time, distributed exponentially with parameter θ_2, then jumps to state $\Delta + 1$ with probability $1 - p_2$ or to state 3 with probability p_2, and so on.

The number of jumps from state $i = 1, \ldots, \Delta$ is not constant. If state i is visited, then the number of jumps from this state is 1, otherwise it is 0. Hence, the conditions of Theorem 3.4 are not met.

In fact if X is a Coxian random variable with the PH representation $(\boldsymbol{\beta}, \mathbf{B})$ in (3.19) and (3.20), then it can be written as

$$X = \begin{cases} Y_1 & \text{with probability (w.p.) } 1 - p_1\,, \\ Y_1 + Y_2 & \text{w.p. } p_1\,(1 - p_2)\,, \\ \vdots & \vdots \\ Y_1 + Y_2 + \ldots + Y_{\Delta-1} & \text{w.p. } p_1\,p_2 \cdots (1 - p_{\Delta-1})\,, \\ Y_1 + Y_2 + \ldots + Y_{\Delta} & \text{w.p. } p_1\,p_2 \cdots p_{\Delta-1}\,. \end{cases} \tag{3.21}$$

If we define

$$S_i = \sum_{j=1}^{i} Y_j\,,$$

then

$$X = \begin{cases} S_i & \text{w.p. } p_i^* = p_1 \cdots p_{i-1}\,(1 - p_i)\,, \qquad i = 1, \ldots, \Delta - 1\,, \\ S_\Delta & \text{w.p. } p_\Delta^* = p_1 \cdots p_{\Delta-1}\,, \qquad i = \Delta\,. \end{cases}$$

In statistical terms it means that the Coxian random variable X has a distribution that can be written as the mixture of the distributions of the S_i's, *i.e.*

$$f_X(x \,|\, \beta,\, \mathbf{B}) = \sum_{i=1}^{\Delta} p_i^*\, f_{S_i}(x \,|\, \beta, \mathbf{B})\,. \tag{3.22}$$

The exact credibility problem for a Coxian distribution is seen as a special case of the exact credibility problem for a mixture of distributions. To the best of our knowledge, this is an open problem with no general results available for mixtures of distributions. The difficulty is the lack of natural prior conjugates for mixtures of distributions (see Frühwirth-Schnatter (2006), p. 53).

Despite this difficulty with exact credibility for Coxian distributions, their representation in (3.22) shows clear advantages for statistical inference. For example, an EM algorithm for a Coxian distribution can be obtained by adapting the very well-studied EM algorithm for a mixture of distributions (see McLachlan and Peel (2000)).

3.3. EXACT CREDIBILITY FOR MULTIVARIATE PHASE-TYPE DISTRIBUTIONS

Multidimensional random variables often arise in insurance contracts. For example losses and allocated loss adjustment expenses (ALAE) may be seen as bi-dimensional loss variables. Another example is when the number of claims in a portfolio is divided in two categories; one for the number of large or catastrophic losses, say bigger than a given threshold, and another for number of small or ordinary losses.

The concept of multidimensional credibility was first introduced by Jewell (1974b) and Hachemeister (1975). Exact multidimensional credibility theory is discussed in Jewell (1974b). This section investigates the conditions for which the multivariate exponential family, with a proper conjugate prior on the parameters, can lead to exact credibility.

3.3.1. Multivariate phase-type (*MPH*) distributions

Multivariate phase-type distributions are introduced in Assaf *et al.* (1984) as a natural extension of univariate *PH* distributions. Another version of *MPH* distributions is defined in Kulkarni (1989). As in the univariate case, the family enjoys many useful properties, such as being dense in all multivariate distributions with positive support. *MPH* distributions have Laplace transforms in closed form and hence all probabilistic quantities, such as all the moments, are easily derived, see Assaf *et al.* (1984). However, some practical problems arise. The non-uniqueness of the parameterization and over-parameterization are major problems with *MPH* which make the interpretation of parameters difficult, see O'Cinneide (1990) and Assaf *et al.* (1984).

Let $J = \{J_u, u \geq 0\}$ be a right continuous Markov process on the finite state space $E = \{1, 2, \ldots, \Delta, \Delta + 1\}$ with initial probability vector $\boldsymbol{\alpha}$ and infinitesimal generator matrix $\mathbf{A}$. Suppose that $\Gamma_1, \ldots, \Gamma_p$ are p nonempty stochastically closed subsets of E such that $\cap_{i=1}^{p} \Gamma_i = \{\Delta + 1\}$ and only the state $\Delta + 1$ is absorbing and hence, the absorption into it is certain.

The matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{B}_0 \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\mathbf{B} = (b_{ij})$ is an $\Delta \times \Delta$ matrix and $\mathbf{B}_0 = (b_{0j})$ is an Δ-dimensional column vector whose elements satisfy $b_{ii} < 0$, $b_{ij} \geq 0$ for $i \neq j = 1, \ldots, \Delta$ and $\mathbf{Be} + \mathbf{B}_0 = \mathbf{0}$, where $\mathbf{e}$ is a column vector of ones. States $1, \ldots, \Delta$ are transient if and only if $\mathbf{B}$ is nonsingular, see Neuts (1995). In this section we assume that $\alpha_{\Delta+1} = 0$, and hence, $\boldsymbol{\alpha}$ can be written as $\boldsymbol{\alpha} = (\boldsymbol{\beta}, 0)$.

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the time until absorption in $\Gamma_1, \ldots \Gamma_p$, respectively. We call the joint distribution of $\mathbf{X}$ a multivariate *PH*, abbreviated as *MPH*, with representation $(\boldsymbol{\beta}, \mathbf{B}, \Gamma_1, \ldots, \Gamma_p)$. The marginal distributions of $X_1, \cdots, X_p$ are univariate *PH*. If $\mathbf{X}$ has a *MPH* distribution, the probability survival function of $\mathbf{X}$ at a point $x_1 < \cdots < x_p$, can be obtained using the theory of Markov chains as follows :

$$\begin{aligned} \bar{F}_{\mathbf{X}}(x_1, \ldots, x_p \,|\, \boldsymbol{\beta}, \mathbf{B}) &= \mathbb{P}\{X_p > x_p, \ldots, X_1 > x_1 \,|\, \boldsymbol{\beta}, \mathbf{B}\} \\ &= \boldsymbol{\beta} e^{\mathbf{B}x_1} \mathbf{g}_1 e^{\mathbf{B}(x_2 - x_1)} \mathbf{g}_2 \cdots e^{\mathbf{B}(x_p - x_{p-1})} \mathbf{g}_p \mathbf{e}\,. \end{aligned}$$

where the $\mathbf{g}_1, \ldots, \mathbf{g}_p$, are $\Delta \times \Delta$ diagonal matrices whose i-th diagonal elements are 1 if $i \in \Gamma_k^c$, and 0 otherwise (see Assaf *et al.* (1984)) for the details). As a convention, we shall omit the dimensions of vectors, it should be possible to determine them from the context.

The joint probability density function of $X_1, \ldots, X_p$ at a point $x_1 < \cdots < x_p$, can be obtained from (3.23) as follows :

$$f_{\mathbf{X}}(x_1, \ldots, x_p \,|\, \boldsymbol{\beta}, \mathbf{B}) = \boldsymbol{\beta} e^{\mathbf{B}x_1} \mathbf{G}_1 e^{\mathbf{B}(x_2 - x_1)} \mathbf{G}_2 \cdots e^{\mathbf{B}(x_p - x_{p-1})} \mathbf{B} \mathbf{g}_p \mathbf{e}\,,$$

where $\mathbf{G}_k = \mathbf{B}\mathbf{g}_k - \mathbf{g}_k\mathbf{B}$, for $k = 1, \ldots, p$. Similar formulas exist for the probability survival and density functions of any other orders of $(X_1, \ldots, X_p)$ (see Assaf *et al.* (1984)).

Henceforth we assume that the state space E is represented in the following form

$$E = \{1, \ldots, \Delta_1, \ldots, \Delta_k, \Delta_k + 1, \ldots, \Delta, \Delta + 1\} ,$$

where $\Gamma_0 = \{1, \ldots, \Delta_1\}$ and $\Gamma_k = \{\Delta_k + 1, \ldots, \Delta_{k+1}, \Delta + 1\}$, for $k = 1, \ldots, p$, with the convention that $\Delta_{p+1} = \Delta$.

Lemma 3.1 can now be restated for *MPH* distributions.

Lemma 3.2. *Assume that* $\boldsymbol{X} = (X_1, \ldots, X_p)$ *is a* MPH *random variable with representation* $(\boldsymbol{\beta}, \boldsymbol{B})$. *Then each component of* $\boldsymbol{X}$ *can be written as at most* Δ *sums of random sums of independent exponentials with parameters* $\theta_1 = -b_{11}, \ldots, \theta_\Delta = -b_{\Delta\Delta}$, *where* Δ *is the dimension of matrix* $\boldsymbol{B} = (b_{ij})$.

Proof. If $X_{i_1} = \min\{X_i\}_{i=1}^p$, then Γ_{i_1} is the first set to be hit after Γ_0. Therefore $X_{i_1} = \sum_{j \in \Gamma_0} S_j$, where S_j is the sojourn time in state j. Similarly if $X_{i_1} < \cdots < X_{i_{k-1}} < X_{i_k} < \cdots < X_{i_p}$, then

$$X_{i_k} = \sum_{i \in \cup_{j=0}^{k} \Gamma_k} S_i .$$

□

Lemma 3.2 complements the result of Lemma 3.1, in the sense that the marginal distribution of each component is a univariate *PH*.

Henceforth we focus on the special case of *MPH* when $X_1 < \cdots < X_p$ with probability 1.

Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are N observed past claims. If $X_{l,N+1}$ is the *l-th* element of $\mathbf{x}_{N+1}$, then (3.19) can be restated as follows :

$$\begin{aligned} \mathbb{E}(X_{l,N+1} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) \;=\; & \sum_{j \in \cup_{\nu=0}^{l} \Gamma_\nu} \mathbb{E}(M_j^{[N+1]}\, a_j \frac{s_j^{[1]} + \cdots + s_j^{[N]}}{m_j} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) \\ & + \sum_{j \in \cup_{\nu=0}^{l} \Gamma_\nu} \mathbb{E}\{S_j\,(1 - a_j) \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N\} . \end{aligned}$$

As an immediate consequence we get the following the multivariate extension of Theorem 3.4.

Theorem 3.5. *With a proper gamma prior distribution on the diagonal elements of matrix* $\boldsymbol{B}$*, as in* (3.16) *or* (3.17)*, then the Bayesian premium equals the linear credibility premium if* $\frac{M_{i,N+1}}{M_i+k}$ *is independent of* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

The infinitesimal matrix in Example 3.2 provides a good example of *MPH* distributions. In this case we have that

$$\mathbb{E}(\mathbf{X}_{N+1} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{N}{N+k}\,\bar{\mathbf{x}} + \frac{k}{N+k}\,\mathbb{E}(\mathbf{X})\,.$$

3.4. Conclusion

This chapter proposes an extension of Jewell's theorem in credibility theory outside the exponential family. This extension is obtained by modeling claims in a sub-period (called seasons), rather than the total claim in one period. Exact credibility is achieved, both when the number of seasons is fixed or is random.

As an application, we prove that under some assumptions, the Bayesian premium for phase-type claims with a properly chosen parameter prior, equals the linear credibility premium. For phase-type distributions, the time spent in each state $i = 1, \ldots, \Delta$ is interpreted as the claim in season i. Claims per season are unobservable, only the total claim per period is part of the data.

Bibliography

Asmussen, S., Nerman, O., Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.* **23**, 419–441.

Assaf, D., Langberg, N. A., Savits, T. H., Shaked, M. (1984). Multivariate phase-type distributions. *Oper. Res.* **32**, 688–702.

Bailey, A. (1950). Credibility procedures, *Proceedings of the Casualty Actuarial Society* **XXXVII**, 7–23 and 94–115.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bull.* **4**, 199–207.

Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes. *Proc. Cambridge Philos. Soc.* **51**, 313–319.

Erlang, A. K. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineer's Journal.* **10**, 189–197.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* Springer Series in Statistics. Springer : New York.

Hachemeister, C. A. (1974). Credibility for regression models with application to trend ; in credibility, theory and application, *Proceedings of the Berkeley Actuarial Research Conference on credibility.* Academic Press, New York, **24**, 129–163.

Jewell, W. S. (1974a). Credible means are exact Bayesian for exponential families. *ASTIN Bull.* **8**, 77–90.

Jewell, W. S. (1974b). Exact multidimensional credibility. *Mitteilungen der vereinigung schweizerischer Versicherungsmathematiker* **74**, 193–214.

Jørgensen, B. (1986). Some properties of exponential dispersion models. *Scand. Act. J.* **13**, 187–197.

Jørgensen, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B.* **49**, 127–162, with discussion and author's reply.

Klugman, S. A., Panjer, H. H., Willmot, G. E. (2008). *Loss models : From data to decision. Third edition.* John Wiley & Sons : Hoboken, NJ.

Kulkarni, V. G. (1989). A new class of multivariate phase type distributions. *Oper. Res.* **37**, 151–158.

Landsman, Z. M., Makov, U. E. (1998). Exponential dispersion models and credibility. *Scand. Actuar. J.* **1**, 89–96.

McLachlan, G., Peel, D. (2000). *Finite mixture models.* Wiley-Interscience : New York.

Mayerson, L. (1964). A Bayesian view of credibility. *Proceedings of the Casualty Actuarial Society* **51**, 85–104.

Mowbray, A. H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? *Proceedings of the Casualty Actuarial Society* **I**, 24–30.

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. *J. Roy. Statist. Soc. Ser. A.* **135**, 370-384.

Neuts, M. F. (1995). *Matrix-analytic methods in queueing theory.* CRC : Boca Raton, FL, 265–292.

Norberg, R. (1979). The credibility approach to experience rating. *Scand. Actuar. J.* **4**, 181–221.

O'Cinneide, C. A. (1989). On nonuniqueness of representations of phase-type distributions. *Comm. Statist. Stochastic Models* **5**, 247–259.

O'Cinneide, C. A. (1990). Characterization of phase-type distributions. *Comm. Statist. Stochastic Models* **6**, 1–57.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. *Statistics : Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* Calcutta : Indian Statistical Institute, 579-604.

CONCLUSION

Cette thèse a abordé dans un premier temps l'estimation statistique et les tests d'adéquation, aussi appelés test d'ajustement, pour des distributions phase-type (PH) bivariées. Dans un deuxième temps, la thèse a proposé d'utiliser les distributions phase-type univariées ou multivariées dans les modèles de crédibilité en actuariat. Tous les résultats de la thèse ont utilisé la définition d'une distribution PH obtenue de la structure d'une chaîne de Markov en temps continu. La définition d'une distribution PH à partir de la densité (matrix exponential distribution) ne suffit malheureusement pas à la tâche qui nous incombait.

Dans le premier article, on traite de problèmes de nature statistique lors de la modélisation conjointe de deux variables positives et corrélées par une distribution PH bivariée. L'algorithme EM est proposé pour obtenir les estimateurs du maximum de vraisemblance des paramètres inconnus que sont les probabilités initiales des états de la chaîne de Markov et les éléments de la matrice du générateur infinitésimal. Cet algorithme a été programmé en MATLAB et utilisé pour ajuster une distribution PH à un jeu de données de 1500 observations bivariées du domaine de l'assurance. Le calcul des estimateurs est lourd, car il est de nature itérative et nécessite à chaque itération la résolution numérique de systèmes d'équations différentielles par la méthode de Runge-Kutta. Un test d'ajustement dans la lignée des tests de Cramér-von Mises est aussi proposé pour vérifier l'adéquation d'une distribution PH bivariée comme modèle. La statistique du test compare la fonction de survie paramétrique ajustée par l'algorithme EM avec la fonction de survie expérimentale, laquelle est non paramétrique. Une simulation pour un modèle PH bivarié simple d'ordre trois avec 200 observations et 2000

échantillons bootstrap a donné de bons résultats quant au seuil, aussi appelé niveau de signification, du test. Malheureusement, ce test n'a pu être effectué sur les données provenant de l'assurance. Le modèle proposé étant d'ordre onze avec 1500 observations, la lourdeur des calculs dépassait notre puissance de calcul. Une programmation des algorithmes en langage C++, par exemple, pourrait possiblement rendre réalisable le test d'ajustement proposé. On a observé néanmoins que les fonctions de répartition expérimentale et paramétrique des deux distributions marginales sont très rapprochées. La dépendance semble aussi avoir été décrite de manière satisfaisante, car les coefficients de corrélation paramétrique et non paramétrique, c'est-à-dire celui basé sur les rangs, sont aussi très rapprochés pour les mesures de Spearman et de Kendall. Quant au coefficient de corrélation de Pearson, il donne une mesure de la dépendance moins appropriée puisque la régression de la variable ALAE sur la variable LOSS était manifestement non linéaire. En plus de la moyenne conditionnelle, c'est-à-dire de la régression, les distributions *PH* bivariées ont aussi permis de calculer les quantiles de la distribution conditionnelle. Ces quantiles conditionnels sont d'un grand intérêt chez les actuaires comme en fait foi l'utilisation répandue de la mesure de risque de perte connue sous le nom de VaR, l'acronyme de l'expression anglophone Value at Risk.

Les perspectives de recherche envisagées comme suite au premier article comportent deux volets. Le premier volet consisterait à reprogrammer les algorithmes dans un langage compilé et plus efficace du point de vue de la rapidité comme le langage C++. Le langage MATLAB est un langage interprété et plus lent d'exécution. Nous l'avons d'abord utilisé parce qu'il offre déjà des fonctions pour le calcul de la fonction exponentielle d'une matrice et pour la résolution d'équations différentielles. On peut aussi étendre quelque peu le domaine d'application de l'algorithme EM à des modèles où l'une des variables est nulle avec une probabilité non nulle. Dans l'exemple sur l'assurance, la variable ALAE peut être strictement positive même si aucun paiement à l'assuré n'est fait. Le deuxième volet, le plus ambitieux, chercherait à développer des méthodes de sélection de modèles. Un modèle *PH* bivarié comprend en effet plusieurs éléments. On doit d'abord choisir l'ordre du modèle, c'est-à-dire le nombre d'états de la chaîne de Markov, et la

structure de la matrice du générateur infinitésimal. Un autre choix à effectuer est celui des sous-ensembles fermés Γ_1 et Γ_2 dont l'intersection est l'état absorbant. Pour l'instant, nous ne disposons pas d'une véritable méthode de sélection de modèles comme il en existe pour la régression linéaire multiple.

Le deuxième article étend le théorème de Jewell en théorie de la crédibilité à une plus grande classe de distributions que celles des distributions exponentielles linéaires et même de la famille exponentielle. Cette extension est rendue possible grâce à l'emploi des chaînes de Markov sous-jacentes dans la définition même des distributions *PH*. Cette approche permet aux actuaires de décomposer les réclamations totales d'une période donnée en sous-périodes (saisons). Ici, des saisons différentes peuvent avoir des réclamations de distributions distinctes, en autant que le facteur de crédibilité reste le même. Par exemple, si $X_i = \alpha_i Y$, pour $i = 1, \ldots, M$, est la réclamation associée à la saison i, où Y est de la famille exponentielle linéaire ou de dispersion et que $\alpha_i > 0$ est une constante, alors les variables X_i aurons toutes le même facteur de crédibilité, bien que leur distributions diffèrent. La thèse montre aussi que la prime de crédibilité pour le total des réclamations dans la période, *i.e.* $\sum_i X_i$, associée à sa distribution a priori conjugée, est égale à la prime bayésienne.

La crédibilité exacte est obtenue ici dans le cas spécial de distributions *PH* univariées et multivariées avec matrice de sous-intensité bidiagonale et vecteur de probabilités initiales donnant toute la masse à un seul état initial. Si le vecteur de probabilités initiales prend une forme plus générale, le résultat de la thèse peut être généralisé en conditionnant sur l'état initial, *i.e.* les distributions coxiennes. Les chaînes de Markov sous-jacentes sont utilisées, incluant les paramètres de risque non-observables pour les distributions *PH*.

Cet article s'attarde à la prime bayésienne, donc le premier moment des réclamations. Le second moment peut aussi être d'intérêt pour déterminer la précision de l'estimateur de la moyenne. Jewell et Schnieper (1985) prouvent que la prédiction exacte de cette variance est une fonction linéaire de la variance de l'échantillon et de la déviation carrée de la moyenne de la distribution a priori.

Nous croyons, au moins dans certains cas, qu'une modélisation similaire avec des distributions *PH* et *MPH* est possible.

BIBLIOGRAPHY

Jewell, W. S., Schnieper, R. (1985). Credibility approximations for bayesian prediction of second moments. *ASTIN Bull.* **15**, 103–121.