Multivariate phase type distributions - Applications and parameter estimation

Meisch, David; Nielsen, Bo Friis; Leleur, Steen

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis

Multivariate phase type distributions

Applications and parameter estimation

David Meisch

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kongens Lyngby
PHD-2014-331

DTU

Preface

This thesis has been submitted as a partial fulfilment of the requirements for the Danish Ph.D. degree at the Technical University of Denmark (DTU). The work has been carried out during the period from June 1st, 2010, to January 31th 2014, in the Section of Statistic in the Department of Applied Mathematics and Computer Science at DTU. The Ph.D. project has been part of the DTU Compute graduate school ITMAN as well as the UNITE project.

The main supervisor has been Associate Professor Bo Friss Nielsen, DTU Compute. Professor Steen Leleur, DTU Transport, served as a co-supervisor. During my external stay at Universidad Nacional Autónoma de México (UNAM) Professor Mogens Bladt, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), has also been involved in the study. I would like to thank him and IIMAS at UNAM for the hospitality of hosting me during my external stay.

The Ph.D. project was supported by the Danish Council for Strategic Research (grant no. 2140-08-0011). For my research stay at UNAM I received external funding from *Otto Mønsteds Fond*. I would like to take this opportunity to thank for the financial support which made my studies possible.

Kongens Lyngby, January 31, 2014

David Meisch

Summary

Multivariate phase type distributions: Applications and parameter estimation

The best known univariate probability distribution is the normal distribution. It is used throughout the literature in a broad field of applications. In cases where it is not sensible to use the normal distribution alternative distributions are at hand and well understood, many of these belonging to the class of phase type distributions. Phase type distributions have several advantages. They are versatile in the sense that they can be used to approximate any given probability distribution on the positive reals. There exist general probabilistic results for the entire class of phase type distributions, allowing for different estimation methods for the whole class or subclasses of phase type distributions. These attributes make this class of distributions an interesting alternative to the normal distribution.

When facing multivariate problems, the only general distribution that allows for estimation and statistical inference, is the multivariate normal distribution. Unfortunately only little is known about the general class of multivariate phase type distribution. Considering the results concerning parameter estimation and inference theory of univariate phase type distributions, the class of multivariate phase type distributions shows potential for similar great results.

My PhD studies were part of the the work package 3 of the UNITE project.

The overall goal of the UNITE project is to improve the decision support prior to deciding on a project by reducing systematic model bias and by quantifying and reducing model uncertainties.

Research has shown that the errors on cost estimates for infrastructure projects clearly do not follow a normal distribution but is skewed towards cost overruns. This skewness can be described using phase type distributions. Cost benefit analysis assesses potential future projects and depend on reliable cost estimates. The Successive Principle is a group analysis method primarily used for analyzing medium to large projects in relation to cost or duration. We believe that the mathematical modeling used in the Successive Principle can be improved. We suggested a novel approach for modeling the total duration of a project using a univariate phase type distribution. The model is then extended to catch the correlation between duration and cost estimates using a bivariate phase type distribution. The use of our model can improve estimates for duration and costs and therefore help project management to make the optimal decisions.

The work conducted during my PhD studies aimed at shedding light on the class of multivariate phase type distributions. This thesis contains analytical and numerical results for parameter estimations and inference theory for a family of multivariate phase type distributions. The results can be used as a stepping stone towards understanding multivariate phase type distributions better. However, we are far from uncovering the full potential of general multivariate phase type distributions. Deeper understanding of multivariate phase type distributions will open up a broad field of research areas they can be applied to.

This thesis consists of a summary report and two research papers. The work was carried out in the period 2010 - 2014.

Resumé

Multivariate fasetypefordelinger: Anvendelser og parameterestimering

Den bedst kendte univariate sandsynlighedsfordeling er normalfordelingen. Den er grundigt beskrevet i litteraturen inden for et bredt felt af anvendelsesområder. I de tilfælde, hvor det ikke er meningsfuldt at anvende normalfordelingen, findes alternative sandsynlighedsfordelinger som alle er godt beskrevet; mange af disse tilhører klassen af fasetypefordelinger. Fasetypefordelinger har adskillige fordele. De er alsidige forstået på den måde, at de kan benyttes til at tilnærme en vilkårlig sandsynlighedsfordeling defineret på den positive reelle akse. Der eksisterer generelle probabilistiske resultater for hele klassen af fasetypefordelinger, hvilket bidrager til anvendelsen af forskellige estimeringsmetoder på enten klassen af fasetypefordelinger eller dens delklasser. Disse egenskaber gør klassen af fasetypefordelinger til et interessant alternativ til normalfordelingen.

Når det kommer til multivariate problemer, så er den multivariate normalfordeling den eneste generelle fordeling, der tillader parameterestimering og statistisk inferens. Desværre er kendskabet til egenskaberne af den multivariate fasetypefordeling stærk begrænset. Resultaterne for parameterestimering og inferensteori for den univariate fasetypefordeling indikerer et potentiale for lignende gode resultater for klassen af multivariate fasetypefordelinger.

Mit ph.d.-studium var en del af Work Package 3 i UNITE-projektet. UNITE-projektet arbejder mod det overordnede mål at forbedre kvaliteten af beslutningsgrundlaget for projekter. Dette gøres ved at reducere systematisk model bias og ved at beskrive og reducere model usikkerheder generelt. Forskning har vist, at afvigelsen fra omkostningsestimater for infrastrukturprojekter tydeligvis ikke er normaltfordelt men i stedet hælder mod budgetoverskridelser. Denne skævhed kan beskrives med fasetypefordelinger.

Cost-benefit-analyser bruges til at evaluere potentielle fremtidige projekter og til at udvikle pålidelige omkostningsvurderinger. Successiv Princippet er en gruppebaseret analysemetode, der primært bruges til at prædiktere omkostninger og varighed af mellem til store projekter. Vi mener, at den matematiske modellering, der ligger til grund for Successiv Princippet, kan forbedres. Vi foreslår derfor en ny tilgang til modellering af den samlede varighed af et projekt ved hjælp af univariate fasetypefordelinger. Den matematiske model er dernæst udvidet til også at beskrive korrelationen mellem projektvarighed og omkostninger nu baseret på bivariate fasetypefordelinger. Vores model kan anvendes til at forbedre estimater for varighed og omkostninger, og derved hjælpe projekters beslutningstagere til at træffe en optimal beslutning.

Det arbejde, jeg har udført som en del af mit ph.d.-studium, sigtede efter at belyse klassen af multivariate fasetypefordelinger. Denne afhandling indeholder analytiske og numeriske resultater for parameterestimering og inferensteori for en gruppe af multivariate fasetypefordelinger. Resultaterne kan betragtes som et første skridt i retning af en mere tilbundsgående forståelse af multivariate fasetypefordelinger. Vi er imidlertid langt fra at have afdækket det fulde potentiale af generelle fasetypefordelinger. En dybere forståelse af multivariate fasetypefordelinger vil åbne op for et bredt felt af anvendelsesområder.

Afhandlingen består af en opsummerende rapport og to videnskabelige artikler. Det bagvedliggende arbejde var udført i perioden 2010 til 2014.

Acknowledgements

It is always a risky matter to thank people, too easy is somebody forgotten. Therefore, without mentioning any names I would like first of all to thank everybody who helped me during my PhD studies. Everybody who gave me inspiration for my work, and everybody who listened to my problems. Everybody who listened to me when I talked about my work and everybody who listened to me when I talked about my private life. I wish to thank my colleagues, my family and my friends for making my PhD studies such a rewarding experience.

My co-authors Professor Mogens Bladt and Associate Professor Bo Friis Nielsen need to be mentioned especially. Not only did they teach me how to do research they also taught me quite a lot about life in general.

Another person who helped immensely with my academic writing is my host mother Sue Simmons. Without her help, it would have been insufferable to read most of my sentences.

My research regarding the Successive Principle could not have been done without the help of Steen Lichtenberg. He spent countless hours of his private time discussing with me the concept and the challenges of the Successive Principle.

Special thanks belongs to my mother Gisela Schädel, who has been of great moral support, not only during my studies at DTU but also during my studies in Hamburg and all of my life.

Last but not least, I wish to thank my wife Katrine. Not only for her support when I needed it, her patience when I spend my evenings and weekends working, her tolerance when I was tired and exhausted from the studies and unreasonable, but also because she despite all of this chose to marry me. In my opinion this can be considered the greatest I achieved during my PhD studies.

Contents

List of Figures

Symbols and Abbreviations

The following is a list of symbols and abbreviations used in the thesis. Be aware that some symbols will have multiple meanings. However, for each appearance it will be clear from the context which meaning the symbol refers to.

Generally, upper case bold font letters denote matrices, lower case bold font letters denote vectors, lower case italic font letters denote scalars or scalar functions, upper case italic font letters denote random variables, and calligraphic letters denote sets.

The list is not a complete list. Symbols, or meanings of symbols, which only appear a few times, may have been omitted.

Symbols

$\bar{F}_X(\cdot)$	Survival function of the random variable X
$\langle \cdot, \cdot \rangle$	Inner product
$\mathbf{A}^{-1}$	Inverse of the matrix $\mathbf{A}$
$\mathbf{E}(\cdot)$	Expectation
$\mathbf{M}^T$	transpose of matrix $\mathbf{M}$
$\mathbf{0}$	Matrix with all entries being zero
$\mathbf{e}$	Row vector of all ones
$\mathbf{I}$	Identity matrix

$\otimes$	Kroncker product
$\sim$	Distributed as
$\simeq$	Equality in distribution
$F_X(\cdot)$	Cumulative distribution of the random variable X
$f_X(\cdot)$	Probability density function of the random variable X
$I_q(\cdot)$	Modified Bessel function of the first kind
$L(\cdot)$	Likelihood function
$L_0(\cdot)$	log likelihood function
$L_X(\cdot)$	Laplace-Stieltjes transform of the random variable X

Abbreviations

min	Minimal group estimate
mode	Most likely group estimate
BPH	Bilateral phase type
BPH*	Bilateral phase type*
CB	Cost-benefit
CPH	Continuous time phase type
DTU	Technical University of Denmark
EM	Expectation Maximization
LST	Laplace-Stieltjes transform
max.	Maximal group estimate
MBPH*	Multivariate bilateral phase type*
MC	Markov chain
MCMC	Markov chain Monte Carlo
MEB	Merge Event Bias
MJP	Markov jump process

MPH	Multivariate phase type by Assaf et al (1984)
MPH*	Multivariate phase type as defined by Kulkarni (1989)
MVPH	Multivariate phase type as defined by Bladt and Nielsen (2010)
PDE	Partial differential equation
PH	Phase type
SP	Successive Principle
TPH	Triangular phase type
UNITE	Uncertainties in transport infrastructure evaluation

CHAPTER 1

Introduction

The research presented in this thesis has been conducted as part of the UNITE (Uncertainties in Transport Infrastructure Evaluation) project. The aim of the UNITE project is to improve the decision support prior to deciding on a project by reducing systematic model bias and by quantifying and reducing model uncertainties. The motivation for the UNITE project and the work presented in this thesis comes from the recorded inaccuracy of forecasts for planned projects (Lovallo and Kahneman, 2003). It seems amazing that so much inaccuracy exist as "hundreds if not thousands of billions of dollars - public and private- are currently tied up in the provision of new infrastructure around the world" (Flyvbjerg et al, 2003a).

Flyvbjerg et al (2003b) investigated the actual project cost and compared it to the forecast cost for 258 projects. Their result shows that on average the actual cost was 28% higher than the forecasted cost. The reasons for underestimating the project cost cannot be clearly identified. Suggested reasons include optimism bias and anchoring (Lovallo and Kahneman, 2003) and political misrepresentation (Wachs, 1990). A simpler reason could be inadequate models or the unpredictability of the future. Cost underesti-

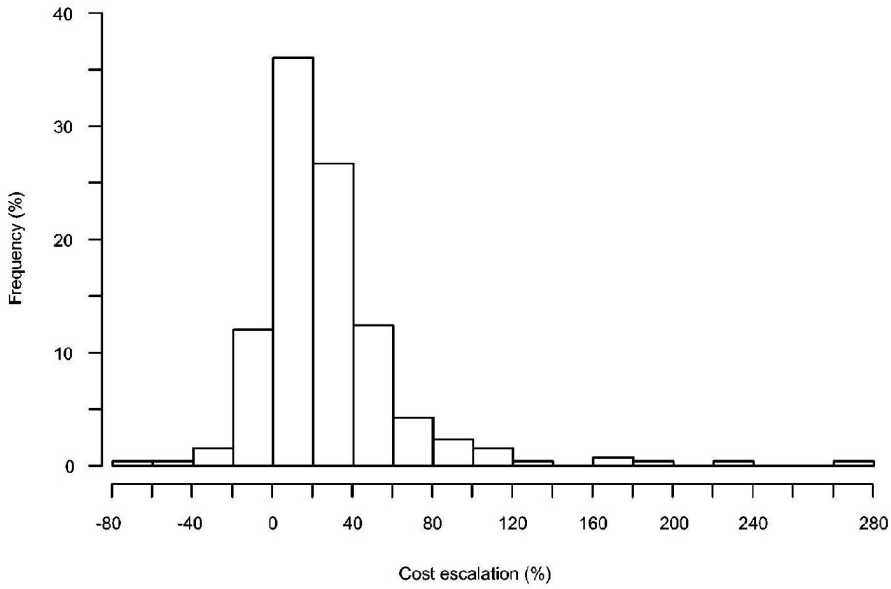


Figure 1.1: Cost escalation in 258 transport and infrastructure projects (constant prices) (Flyvbjerg et al, 2003b)

mation is not the only risk in infrastructure projects, Grimsey and Lewis (2002) summarize at least nine different risks, one of these being operating risk due to delays in construction.

The success of projects are classically defined by completion on time, to budget and with appropriate quality (Williams, 2003). One of many examples for a project that failed at least two of these criteria is the Elbe Philharmonic Hall in Hamburg. The original estimates from 2007, were 114 million Euro construction cost and a completion in 2010. The current estimates are of 789 million Euro and a completion in 2016/2017. It is questionable if the duration or the cost of projects are normally distributed. The underestimation of project costs and project durations is not unique to infrastructure projects, they can be observed through all different

Figure 1.2: Elbe Philharmonic Hall; Example for delay and cost overrun

kinds of projects. Even software projects, which are mainly independent of environmental influences, cannot be properly predicted (Jørgensen and Moløkken-Østvold, 2006). The decision if a project is conducted is often determined using a cost-benefit (CB) analysis. If the actual cost is higher than the predicted cost, an alternative project could have been a wiser choice. Similar arguments can be used in reference to the duration of projects. A delay might, for example cause the loss of market share (electrical consumer products) or revenue (toll roads). Recently, as part of the UNITE project at DTU Transport Morten Skou Nicolaisen (Nicolaisen, 2012) and Jeppe Andersen (Andersen, 2013) have collected data of transport infrastructure projects. The data consist, e.g. of forecast cost as well as actually cost. Further data is now being collected to relate the projects to other factors.

This will be an interesting data base for applying and testing multivariate models.

Current statistical models often assume normally distributed data. However, it is questionable if the duration or the cost of projects are normally distributed. If the data is non-negative and skewed, a convenient choice used in modeling are PH distributions as they can approximate any positive distribution arbitrarily close (Bladt, 2005). Hitherto, when facing multivariate data the only general class of distributions allowing for statistical inference are multivariate normal distributions. Despite the importance of the multivariate normal distribution, there are many cases where data is positive, skewed and clearly non-normally distributed. Examples for this are found in hydrology where simultaneous measurements of precipitation can be analyzed using bivariate Gamma distributions (Yue et al, 2001), in medical trials where they can be used to model the progress of a disease (Ahlström et al, 1999). Multivariate PH distributions are characterized by having PH distributed marginals. They can be used to approximate any multivariate distribution, are straight forward to simulate and have an intuitive probabilistic interpretation.

Considering the value of proper estimation methods for project cost as well as project durations or even joint estimates for cost and duration of a project, it is essential to provide an alternative to the normal and multivariate normal distribution. This need is even more driven by the lack of evidence that costs or durations of projects are normally distributed. Considering univariate data, sufficient alternatives exist, allowing for modeling, estimation and statistical inference for many different applications. For multivariate data, the options are rather limited. The versatility of multivariate PH distributions make them a natural alternative to the multivariate normal distribution. There are very few general results for multivariate PH distribution, one of them being the possibility to calculate all moments analytically (Nielsen et al, 2010).

The research conducted during my PhD studies provides a first step into the direction for estimation and statistical inference in the class of multi-

variate PH distribution as defined by Kulkarni (1989). We have succeeded in estimating parameters and conducting statistical inferences for a subclass of bivariate distributions with Erlang distributed marginals which have a broad field of applications. The inspiration for this research comes from my external stay with Professor Mogens Bladt at the University of Mexico. The results concerning estimations for MPH distributions have been submitted as Appendix A to the "Journal of Stochastic Modeling". The need for estimation arises from mathematical modeling and fitting parameters and models to data. We have altered and extended a project management tool in order to model the entire duration of a project as a PH distributed random variable, allowing us to state all distributional properties directly. Furthermore, we have extended the method in order to deal with multivariate models, primarily in models describing the correlation between cost and duration of projects. The results have been submitted as Appendix B to the "European Journal of Transport and Infrastructure Research", and show how MPH models can improve forecasts for infrastructure models.

In this thesis, Chapter 2 introduces phase type (PH) distributions and different methods for parameter estimation. In Chapter 3, I will introduce multivariate phase type distributions. Chapter 4, will be used to present my contribution to the research regarding parameter estimation for MPH distributions. In Chapter 5, I will present an example for the use of PH and MPH distributions in infrastructure projects. Finally, Chapter 6 will be used to conclude the research conducted during this PhD studies and discuss future areas of research.

I have tried to minimize the number of proofs in this thesis. With the exception of the proof for Lemma 2.6, all proofs that do not refer to other publications are results of my PhD studies.

CHAPTER 2

Univariate phase type distributions

The best known distribution in statistical evaluations is the normal distribution. Especially when modeling error terms it is used frequently. However, the normal distribution being a symmetric distribution with support on the entire real axis makes it less suited to model many natural phenomena, which for example, are not symmetric or perhaps only defined on the positive reals. In these cases PH distributions offer several advantages. One advantage of PH distributions is the possibility to use them to approximate any given distribution with non negative data. Furthermore, they provide an easy stochastic interpretation and closed form solutions for the majority of their statistical properties exist. One typical example is their use in risk theory, for instance (Bladt, 2005), where claims can not be of negative value. Neuts (1975) defined discrete time PH distribution as a probability distribution on the nonnegative integers, "if and only if there exist a finite Markov chain (MC) with a single absorbing state into which absorption is certain, such that for some choice of the initial probabilities this distribution is that of the time till absorption." Continuous PH distributions (CPH) are

defined similarly on the positive reals by use of a continuous time Markov chain. Continuous time Markov chains are also known as Markov jump processes (MJP).

Neuts was the first to publish general results for the distribution of the time until absorption of discrete time as well as continuous time MC. He is known for naming them PH distributions, however, researchers before him investigated special distributions which clearly belong to the class of PH distributions. Some examples are Agner Krarup Erlang, who "observed that the Gamma distribution with an integer valued shape parameter, may be interpreted as a probability distribution constructed by sums of independent exponential random variables" (Neuts, 1975). An application for this distribution can be found in modeling telephone networks (Erlang, 1920). Another pioneer has been David Cox (Cox, 1955) who generalized Erlangs results to cover all distributions with rational Laplace transforms.

A way of stating CPH distribution is by using the representation $(\boldsymbol{\alpha}, \mathbf{T})$ with $\mathbf{T}$ being the sub-intensity matrix of $(X(t); t \geq 0)$ corresponding to the m transient states and $\boldsymbol{\alpha}$ correspondingly being the initial distribution among the transient states, for a random variable τ we write $\tau \sim PH(\boldsymbol{\alpha}, \mathbf{T})$. It should be noted that the representation $(\boldsymbol{\alpha}, \mathbf{T})$ for a given distribution is not necessarily unique.

Definition 2.1. *Continuous Phase type distribution* Let $(X(t); t \geq 0)$ be a Markov jump process on a discrete state space $E = \{1, 2, \ldots, m, m+1\}$ with state $m+1$ being an absorbing state and the states $1, \ldots, m$ being transient. We then define the stochastic variable $\tau = \min\{t \geq 0 : \ X(t) = m+1\}$ as the time of absorption.

Let $\boldsymbol{\alpha}$ be the initial distribution vector on the transient states so that $P(X(0) = i) = \alpha_i$, $\alpha_{m+1} = P(\tau = 0)$ the probability of starting in the absorbing state, and $\mathbf{Q}$ be the generator matrix of $X(t)$ with

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}. \tag{2.1}$$

As $\mathbf{Q}$ is the generator matrix, $\mathbf{T}$ has negative diagonal entries, and for all eigenvalues λ_i holds $\mathrm{Re}(\lambda_i) < 0$. Furthermore $\mathbf{T}$ is invertible and $\mathbf{t} = -\mathbf{T}\mathbf{e}^T$ is called the exit vector. Here $\mathbf{e} = (1, \ldots, 1)$ is a row vector of proper dimension. I will use $\mathbf{0}$ for a matrix of appropriate dimension with all zeros. When stating the representation of a PH distribution, often the minimal representation is used.

Definition 2.2. *Minimal representation of a PH distribution* A representation $(\boldsymbol{\alpha}, \mathbf{T})$ of a PH distribution is minimal if no other representation $(\boldsymbol{\beta}, \mathbf{S})$ exists with $\dim(\mathbf{S}) < \dim(\mathbf{T})$.

Closed form solutions using matrix exponentials for the probability density function $f_X(\cdot)$, the cumulative distribution function $F_X(x) = P(X < x)$ as well as the Laplace-Stieltjes transform (LST) $L_X(\cdot)$ exist.

The exponential of a matrix $\mathbf{M}$ is defined similarly to the definition of the exponential for scalars as

$$exp(\mathbf{M}) = \sum_{i=0}^{\infty} \frac{\mathbf{M}^i}{i!}. \tag{2.2}$$

Efficient numerical evaluation of the matrix exponential can be challenging. There exist several papers on how to calculate the matrix exponential in different ways, see for example Moler and Van Loan (1978) and Moler and Van Loan (2003). Generally it is included in standard numerical software. When calculating the matrix exponential of a sub generator matrix the method of uniformization (e.g. Latouche and Ramaswami (1999)) can be used, resulting in a more stable numerical procedure.

With Equation 2.2 and $\mathbf{e} = (1, \ldots, 1)$, a row vector of proper dimension, these properties can we written as (Neuts, 1975)

$$\begin{aligned} f_X(x) &= \boldsymbol{\alpha} e^{\mathbf{T}x} \mathbf{t} && (2.3) \\ F_x(x) &= 1 - \boldsymbol{\alpha} e^{\mathbf{T}x} \mathbf{e}^T && (2.4) \\ L_X(s) &= E[e^{-Xs}] = \alpha_{m+1} + \boldsymbol{\alpha} \left(s\mathbf{I} - \mathbf{T}\right)^{-1} \mathbf{t}. && (2.5) \end{aligned}$$

Here $\mathbf{M}^T$ denotes the transpose of the matrix $\mathbf{M}$. All moments can be calculated by differentiating the LST and evaluating it at zero. The nth moment of the random variable X is given by

$$E[X^n] = n!\boldsymbol{\alpha}(-\mathbf{T})^{-n}\mathbf{e}^T \tag{2.6}$$

2.1 Closure properties of continuous phase type distributions

The class of CPH is closed under standard operations such as addition, finite mixtures, and order statistic. The closure properties can be exploited when constructing mathematical models using different PH distributions.

Theorem 2.1. Addition of two PH random variables *Let* $X \sim PH(\boldsymbol{\alpha}, \mathbf{T})$ *and* $Y \sim PH(\boldsymbol{\beta}, \mathbf{S})$ *be two independent and PH distributed random variables. Define the random variable* $Z = X + Y$*, then* $Z \sim PH(\boldsymbol{\gamma}, \mathbf{L})$ *with* $\boldsymbol{\gamma} = (\boldsymbol{\alpha}, \alpha_{m+1} \cdot \boldsymbol{\beta})$ *and*

$$\mathbf{L} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \cdot \boldsymbol{\beta} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}. \tag{2.7}$$

Proof. The proof can be found in Neuts (1975). □

The representation $(\boldsymbol{\gamma}, \mathbf{L})$ for Z is not unique. Since $Z = X + Y = Y + X$ also $(\boldsymbol{\gamma}^\star, \mathbf{L}^\star)$ with $\boldsymbol{\gamma}^\star = (\boldsymbol{\beta}, \boldsymbol{\beta}_{m+1} \cdot \boldsymbol{\alpha})$ and

$$\mathbf{L} = \begin{pmatrix} \mathbf{S} & \mathbf{s} \cdot \boldsymbol{\beta} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}. \tag{2.8}$$

is a PH representation for Z. Generally it is assumed that the scalar α_{m+1} is zero, if $\alpha_{m+1} > 0$ it is the probability of the underlying MJP starting in the absorbing state, i.e. $P(X = 0) = \alpha_{m+1}$. It is then said that X has an atom at zero.

It is a well known fact that the finite sum of independent exponentially distributed random variables results in an Erlang distributed random variable. The Erlang distribution is a special case of the Gamma distribution where the shape parameter is integer valued. The probability density function for a random variable Z that is Gamma distributed with shape parameter $k \in \mathbb{R}_+$ and intensity parameter $\lambda \in \mathbb{R}_+$ is defined as:

$$f_Z(z) = \frac{1}{\Gamma(k)} \lambda^k z^{k-1} e^{-\lambda z}. \tag{2.9}$$

For $k \notin \mathbb{N}$, the Gamma distribution is not contained in the class of PH distributions. If $k \in \mathbb{N}$ it is said that the random variable is Erlang k distributed.

Example 2.1. *Erlang distribution* Define two exponentially distributed random variables $X \sim exp(\lambda)$ and $Y \sim exp(\lambda)$. Their PH representation can be given by $(\boldsymbol{\alpha}, \mathbf{T}) = ((1), (-\lambda))$. Furthermore define the random variable $Z = X + Y$. Theorem 2.1 states that Z is as well PH distributed. The PH representation is $(\boldsymbol{\gamma}, \mathbf{L})$ with $\boldsymbol{\gamma} = (1, 0)$ and

$$\mathbf{L} = \begin{pmatrix} -\lambda & \lambda \\ 0 & -\lambda \end{pmatrix}. \tag{2.10}$$

The random variable Z is Erlang two distributed with intensity parameter λ and the representation is unique except for permutation. The density can be given by Equation 2.9 using the general result for PH distributions as

$$f_Z(z) = \boldsymbol{\gamma} e^{\mathbf{L} \cdot z} \mathbf{e}. \tag{2.11}$$

A similar example can be constructed, choosing the random variables X and Y to be again exponentially distributed but with different intensity parameters. In that case Z is said to be generalized Erlang distributed, and unlike Equation 2.9, Equation 2.11 would still be valid. A further generalization is the Coxian distribution, allowing the MJP to reach the absorbing state from any transient state.

Example 2.2. *Coxian distribution* The PH representation of a Coxian distributed random variable X is $(\boldsymbol{\alpha}, \mathbf{T})$ with $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$ and

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & p_1\lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & -\lambda_{k-2} & p_{k-2}\lambda_{k-2} & 0 \\ 0 & 0 & \ldots & 0 & -\lambda_{k-1} & p_{k-1}\lambda_{k-1} \\ 0 & 0 & \ldots & 0 & 0 & -\lambda_k \end{pmatrix}. \tag{2.12}$$

If $p_1 = p_2 = \ldots = p_{k-1} = 1$ the Coaxian distribution is a generalized Erlang k distribution.

Theorem 2.1 implies that the distribution of a weighted sum of independent PH distributed random variables is contained in the class PH. Also finite mixtures of independent and PH distributed random variables is again PH distributed.

Theorem 2.2. Finite mixtures of PH random variables *Let $X_i \sim PH(\boldsymbol{\alpha}_i, \mathbf{T}_i)$, for $i \in \{1, \ldots, k\}$ be independent random variables. Define Z so that $P(Z = X_i) = p_i$ and with that*

$$f_Z(z) = \sum_{i=1}^{k} p_i \boldsymbol{\alpha}_i e^{\mathbf{T}_i z} \mathbf{t}_i. \tag{2.13}$$

Here $\mathbf{t}_i$ is the exit vector of the ith random variable. The distribution of the random variable Z is a convex mixture of PH distributions, and $Z \sim PH(\boldsymbol{\gamma}, \mathbf{L})$ with $\boldsymbol{\gamma} = (p_1\boldsymbol{\alpha}_1, \ldots, p_k\boldsymbol{\alpha}_k)$ and

$$\mathbf{L} = \begin{pmatrix} \mathbf{T}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{T}_k \end{pmatrix}. \tag{2.14}$$

Proof. The proof can be found in Neuts (1975) □

The hyper exponential distribution is one of the most prominent examples of finite mixtures of PH distributed random variables and its distribution can be constructed using Theorem 2.2.

Example 2.3. *Hyper exponential distribution* Let $X_i \sim exp(\lambda_i)$ for $i \in \{1, \ldots, k\}$ be k independent exponentially distributed random variables. Furthermore choose p_i so that $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$, construct the random variable Z so that $P(Z = X_i) = p_i$, then Z is a convex mixture of PH distributed random variables and is itself PH distributed with representation $(\boldsymbol{\gamma}, \mathbf{L})$ where

$$\boldsymbol{\gamma} = (p_1, \ldots, p_n), \tag{2.15}$$

$$\mathbf{L} = \begin{pmatrix} -\lambda_1 & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & -\lambda_n \end{pmatrix}. \tag{2.16}$$

The density can again be stated using the PH representation or directly as $f_Z(z) = \sum_{i=1}^{n} p_i \lambda_i e^{-\lambda_i z}$.

PH distributions are also closed under minimum and maximum operations. It is useful to express the representation of the minimum and the maximum of two PH random variables using the Kronecker product.

Definition 2.3. *Kronecker Product* Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{q \times r}$ and define A_{ij} as the element of the ith row and the jth column of A. Then the Kronecker product $\otimes$ is defined as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} A_{11}\mathbf{B} & \ldots & A_{1m}\mathbf{B} \\ \vdots & \vdots & \vdots \\ A_{n1}\mathbf{B} & \vdots & A_{nm}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{n \cdot q \times m \cdot r}. \tag{2.17}$$

For the minimum of two independent PH distributed random variables the result comes from the following theorem.

Theorem 2.3. Minimum of two independent PH random variables *Here are* $\boldsymbol{\gamma} = \boldsymbol{\alpha} \otimes \boldsymbol{\beta}$ *and*

$$\mathbf{L} = \left(\mathbf{T} \otimes \mathbf{I}_Y + \mathbf{I}_X \otimes \mathbf{S}\right). \tag{2.18}$$

The identity matrix $\mathbf{I}_Y$ *has the dimensions of* $\mathbf{S}$ *and* $\mathbf{I}_X$ *has the dimensions of* $\mathbf{T}$.

Proof. The proof can be found in Neuts (1975). □

Theorem 2.3 can be used to show that the minimum of two exponential distributions is again an exponential distribution.

Example 2.4. *Minimum of two exponentially distributed random variables* Choose $X \sim exp(\lambda_1)$ and $Y \sim exp(\lambda_2)$, define $Z = \min(X, Y)$ then Z is exponentially distributed with density

$$f_Z(z) = (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)z}. \tag{2.19}$$

The random variable Z is of PH type and one representation is $(\boldsymbol{\gamma}, \mathbf{L})$ with $\boldsymbol{\gamma} = (1)$ and

$$\mathbf{L} = \left((-\lambda_1) \otimes \mathbf{I}_Y + \mathbf{I}_X \otimes (-\lambda_2)\right) = \left(-(\lambda_1 + \lambda_2)\right). \tag{2.20}$$

Using Equation 2.3, the density of Z is identical to the result in Equation 2.19

$$f_Z(z) = \boldsymbol{\gamma} e^{\mathbf{L}z}\mathbf{e} = 1e^{-(\lambda_1+\lambda_2)z}(\lambda_1 + \lambda_2). \tag{2.21}$$

The maximum of two random variables can be derived using theorem 2.3 and

$$\begin{aligned} \max(X, Y) &= X + Y - \min(X, Y) \\ &= \min(X, Y) + (X - \min(X, Y)) + (Y - \min(X, Y)). \end{aligned}$$

We know from Theorem 2.3 that $\min(X, Y)$ is of PH type. The difference $X - \min(X, Y)$ is not independent, but gives an idea for how to calculate the maximum of two independent and PH distributed random variables. The difference of two independent and PH distributed random variables can be used to construct bilateral PH (BPH) distributions (Ahn and Ramaswami, 2005). It is an extension of PH distributions to the entire real axis.

Theorem 2.4. Maximum of two independent PH distributed random variables *Let* $X \sim PH(\boldsymbol{\alpha}, \mathbf{T})$ *and* $Y \sim PH(\boldsymbol{\beta}, \mathbf{S})$ *be two independent PH random variables. Then* $Z = \max(X, Y)$ *is again PH distributed with representation* $(\boldsymbol{\gamma}, \mathbf{L})$. *Where* $\boldsymbol{\gamma} = (\boldsymbol{\alpha} \otimes \boldsymbol{\beta})$ *and*

$$\mathbf{L} = \begin{pmatrix} \mathbf{T} \otimes \mathbf{I}_Y + \mathbf{I}_X \otimes \mathbf{S} & \mathbf{I}_X \otimes \mathbf{s} & \mathbf{t} \otimes \mathbf{I}_Y \\ \mathbf{0} & \mathbf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S} \end{pmatrix} \tag{2.22}$$

Proof. The proof can be found in Neuts (1975). □

The first entry of the sub generator matrix $\mathbf{L}$, is the sub generator of the distribution of $\min(X, Y)$. Afterwards the Markov jump process continues with the duration of the remaining variable. The entries $\mathbf{I}_X \otimes \mathbf{s}$ and $\mathbf{t} \otimes \mathbf{I}_Y$ ensure that the remaining lifetime of X or Y continuous in the proper phase.

The exponential distribution is memoryless. The distribution of the maximum of two independent exponentially distributed random variables is therefore a mixture of generalized Erlang distributions.

Example 2.5. *Maximum of two exponentially distributed random variables* Choose again $X \sim exp(\lambda_1)$ and $Y \sim exp(\lambda_2)$, define $Z = \max(X, Y)$. Theorem 2.4 states that random variable Z is PH distributed with representation $(\boldsymbol{\gamma}, \mathbf{L})$ with $\boldsymbol{\gamma} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$ and

$$\mathbf{L} = \begin{pmatrix} -(\lambda_1 + \lambda_2) & \lambda_2 & \lambda_1 \\ 0 & -\lambda_1 & 0 \\ 0 & 0 & -\lambda_2 \end{pmatrix}. \tag{2.23}$$

From Theorem 2.3 we know that the minimum of X and Y is $exp(\lambda_1 + \lambda_2)$ distributed. Depending on which random variable is equal to the minimum is the remaining lifetime either $exp(\lambda_1)$ or $exp(\lambda_2)$ distributed.

It is sufficient to focus on two independent random variables. The extension to more than two random variables is straight forward since

$$\max(x, y, z) = \max(x, \max(y, z)) \tag{2.24}$$

as well as

$$\min(x, y, z) = \min(x, \min(y, z)). \tag{2.25}$$

Aside from being closed under maximum and minimum operations, PH distributions are also closed under general order statistics.

Lemma 2.5. *For $i \in \{1, \ldots, n\}$ let X_i be n independent and PH distributed random variables. Define their order statistics as $\{X_{(1,n)}, \ldots, X_{(n,n)}\}$, where $X_{(1,n)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n,n)} = \max\{X_1, \ldots, X_n\}$. The ith order statistic $X_{(i,n)}$ is the ith smallest random variable and again PH distributed.*

Proof. The proof can be found in Assaf and Levikson (1982). □

The class of all PH distributions is infinitely large and consists of certain subclasses. Further focus will be on the subclass of PH distributions where the sub generator matrix can be expressed as an upper triangular matrix. A quadratic matrix $\mathbf{M}$ can be called upper triangular when all sub diagonal elements are zero

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & \ldots & m_{1n} \\ 0 & m_{22} & \ldots & m_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & m_{nn} \end{pmatrix}. \tag{2.26}$$

2.1.1 Closure properties of PH distributions with upper triangular sub generator Matrix

The sub generator matrix of a PH distribution can be of almost any shape. Only few restrictions hold, e.g. it has to be a quadratic matrix where the diagonal elements are negative, the off diagonal elements are negative, and the row sums have to be smaller or equal to zero. A special subclass contained in the class of PH distribution consist of all PH distributions with an upper triangular sub generator matrix.

Definition 2.4. *Triangular phase pype distribution* A PH distribution is called Triangular PH (TPH) distribution if there exist a representation where its sub generator matrix is an upper triangular matrix.

A PH distribution with non negative initial distribution can be expressed as a TPH if all poles of its LST are real (O'Cinneide, 1991). Upper triangular matrices have several advantages. One example is the calculation of the inverse:

Lemma 2.6. Inverse of an upper triangular block matrix *For a quadratic matrix* $\mathbf{M}$ *of the following structure*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} \tag{2.27}$$

the inverse $\mathbf{M}^{-1}$ *is*

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix}. \tag{2.28}$$

Proof. By direct verification

$$\begin{aligned}
\mathbf{M}\mathbf{M}^{-1} &= \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix} && (2.29) \\
&= \begin{pmatrix} \mathbf{A}\mathbf{A}^{-1} & \mathbf{A}(-\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1}) + \mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{C}\mathbf{C}^{-1} \end{pmatrix} && (2.30) \\
&= \begin{pmatrix} \mathbf{I} & -\mathbf{B}\mathbf{C}^{-1} + \mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} && (2.31) \\
&= \mathbf{I}. && (2.32)
\end{aligned}$$

The opposite direction $\mathbf{M}^{-1}\mathbf{M}$ can be calculated exactly the same way. □

If $\mathbf{M}$ is the sub generator of a TPH then $\mathbf{A}$ and $\mathbf{C}$ are again upper triangular matrices. Successive use of Lemma 2.6 makes the inversion of the sub generator computationally cheap. When calculating the moments of PH distributions (see Equation 2.5), it is necessary to calculate the inverse of the sub generator matrix. For TPH distributions the computational cost of inverting the sub generator is smaller compared to distributions which are not generated by an upper triangular sub generator matrix. Other numerical advantages are at hand. The class of TPH distributions has some of the same closure properties as the complete class of PH distributions.

Theorem 2.7. Closure of the class of TPH *TPH is the smallest class containing all exponential distributions which is closed under finite mixtures, finite convolutions and the formation of coherent systems.*

Proof. The proof can be found in Assaf and Levikson (1982). □

The expression formation of coherent systems come from reliability theory and describe systems that fail once a certain number of components are not functional. A simple example for a PH distribution belonging to the class of TPH is the Erlang distribution from Example 2.1.

Example 2.6. *Sum of two Erlang distributed random variables* Let $X \sim Er(2, \lambda)$ and $Y \sim Er(3, \mu)$ be two independent and Erlang distributed random variables. Define $Z = X + Y$, Z is then PH and more specific THP distributed with representation $(\boldsymbol{\gamma}, \mathbf{L})$ where

$$\begin{aligned} \boldsymbol{\gamma} &= (1, 0, 0, 0, 0), \\ \mathbf{L} &= \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 \\ 0 & -\lambda & \lambda & 0 & 0 \\ 0 & 0 & -\mu & \mu & 0 \\ 0 & 0 & 0 & -\mu & \mu \\ 0 & 0 & 0 & 0 & -\mu \end{pmatrix}. \end{aligned}$$

Obviously, the matrix $\mathbf{L}$ is of upper triangular shape and therefore Z belongs in the class of THP.

Theorem 2.7 makes the class of TPH distributions an interesting subclass. Especially when modeling complex systems, the state space grows fast and direct numerical inversion of the sub generator matrix can be both instable as well as computationally expensive. Another advantage of TPH distributions is the number of transitions before absorption has a finite upper bound. This is an advantage when simulating PH distributions where it is impossible to consider an unbounded number of transitions.

2.2 Fitting PH distributions

Probability distributions are used to describe and predict non-deterministic behavior. A famous example from finance is the Black-Scholes model (Black and Scholes, 1973) which is used for predicting option prices. Fitting probability distributions is about finding the distribution that best suits the given data. The best fit of data can be achieved through different approaches, e.g. moment matching, maximum likelihood estimation and Markov chain Monte Carlo (MCMC) methods.

In this Section we will describe a maximum likelihood related approach for fitting PH distributions, more specifically the expectation maximization

(EM) algorithm. The method seems to be of great potential for the general class of multivariate phase type distributions. The EM algorithm also allows for statistical inference. Asmussen et al (1996) applied the EM algorithm to PH distributions and made it a common tool for fitting PH distributions.

Section 2.2.1 will be used to describe the basic concept of the EM algorithm and to state specific results for its application to PH distributions. Section 2.2.2 is then used to introduce the Fisher information matrix and I will briefly explain how to calculate it via the EM algorithm. There are some difficulties that can arise when fitting PH distributions, e.g. the problem of non unique representations (Neuts, 1975; O'Cinneide, 1989) also known as over parameterization (Fackrell, 2005), making statistical inference infeasible. Therefore it is essential to ensure that the representation is unique with the exception of permutations.

2.2.1 The Expectation Maximization algorithm and its application to parameter estimation of PH distributions

The EM algorithm is generally associated with Dempster et al (1977), however its roots go back to earlier work, e.g. Baum et al (1970) derived maximization techniques for the statistical analysis of Markov chains. Though earlier work has dealt with iterative maximum likelihood estimation techniques, Dempster et al (1977) generalized the specific results and named the method. It is an iterative method for calculating maximum likelihood estimates for probability density functions in cases with missing data or where a direct evaluation of the observed data likelihood function is difficult. As the name indicates, the algorithm contains two different steps. The expectation (E) step replaces the missing (unobserved) data with their conditional expectation given the current parameter estimates as well as the observed data. In the maximization (M) step, maximum likelihood estimates of the parameters are calculated based on the observed data as well as the expectations obtained through the E step. In a more general setting, maximum likelihood estimators $\widehat{\boldsymbol{\Theta}}$ for a parameter vector $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_k)$ maxi-

mize the likelihood function $L(\mathbf{\Theta}; y)$ of a set of independent observations $y = \left\{y^{(1)}, \ldots, y^{(n)}\right\}$. More specific, if $f(y^{(i)}; \mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_k)$ is the probability density function for observation i given the parameter vector $\mathbf{\Theta}$ and y consist of i.i.d. distributed data, the likelihood function can be written as

$$L(\mathbf{\Theta}; y) = \prod_{i=1}^{n} f(y^{(i)}; \mathbf{\Theta}). \tag{2.33}$$

When the likelihood function is twice differentiable for $\mathbf{\Theta}$, then the maximum likelihood estimate $\widehat{\mathbf{\Theta}}$ can only take the values $(\Theta_1, \ldots, \Theta_k)$ for which

$$\frac{\partial L(\mathbf{\Theta}; y)}{\partial \Theta_i} = 0 \quad \forall i \in \{1, \ldots, k\}. \tag{2.34}$$

In case the parameter space is bounded, also the boundaries become candidates for the the maximum likelihood estimates. The likelihood function is locally maximized when the likelihood function around $\widehat{\mathbf{\Theta}}$ is convex, i.e. the Hessian

$$\mathbf{H}\left(L(\mathbf{\Theta}; y)\right) = \begin{pmatrix} \frac{\partial^2 L(\mathbf{\Theta}; y)}{\partial^2 \Theta_1} & \cdots & \frac{\partial^2 L(\mathbf{\Theta}; y)}{\partial \Theta_1 \partial \Theta_k} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 L(\mathbf{\Theta}; y)}{\partial \Theta_k \partial \Theta_1} & \cdots & \frac{\partial^2 L(\mathbf{\Theta}; y)}{\partial \Theta_k \partial \Theta_k} \end{pmatrix} \tag{2.35}$$

is negative definite, in other words all eigenvalues of $H\left(L(\mathbf{\Theta}; y)\right)$ have to be negative.

Often analyzing a transformation of the likelihood function simplifies the analysis at hand. The logarithm is a monotone and continuous function well suited. It is the most important transform for likelihood functions, transforming the product of the probability density functions into a sum of the logarithm of the probability density functions. Analyzing the log transform $L_0(\mathbf{\Theta}; y)$ of the likelihood function $L(\mathbf{\Theta}; y)$ will result in the same maximum likelihood estimates. The log likelihood function is defined as:

$$L_0(\mathbf{\Theta}; y) = log(L(\mathbf{\Theta}; y)) = \sum_{i=1}^{n} \log(f(y^{(i)}; \mathbf{\Theta})). \tag{2.36}$$

If the second derivative of $L_0(\mathbf{\Theta}; y)$ exists, it also exists for the likelihood function and it can be used for obtaining the Fisher information. The Fisher information matrix can be used to estimate the inverse of the variance covariance matrix of the maximum likelihood estimates.

Definition 2.5. *Observed data (Fisher) Information Matrix* The entries of the observed data information matrix $I(\mathbf{\Theta}; y)$ for a parameter vector $\mathbf{\Theta}$ and the observed data y is given by the second derivative of the observed data log likelihood function with respect to $\mathbf{\Theta}$, in other words

$$\mathbf{I}_{ij}(\mathbf{\Theta}, y) = -\frac{\partial^2 L_0(\mathbf{\Theta}, y)}{\partial \Theta_i \Theta_j}. \tag{2.37}$$

Section 2.2.2 will summarize a procedure for calculating the Fisher information matrix for parameter estimates obtained through the EM algorithm reusing calculations used in the EM algorithm. Evaluating the likelihood function through the EM algorithm will result in the same estimates as a direct evaluation of the likelihood function. If more than one local maxima of the likelihood function exist, it is not possible to determine beforehand to which maxima either method will converge.

When using the EM algorithm, it is assumed that two sets of data exist. The observed data y which is incomplete and the complete data x which contains y and an unobserved part z. If the underlying probability distribution function of the data belongs to the family of exponential distributions, e.g. a MJP, it is possible to substitute z with its sufficient statistic (Asmussen et al, 1996). The general procedure for the EM algorithm is described in Algorithm 1.

To fit a PH distribution to given data, the representation $(\boldsymbol{\alpha}, \mathbf{T})$ is to be estimated. Often the only available data at hand is the time of absorption; in this case the likelihood function is

$$L(\alpha, T; y) = \prod_{i=1}^{n} \boldsymbol{\alpha} e^{\mathbf{T} y^{(i)}} \mathbf{t}. \tag{2.38}$$

Algorithm 1: The EM algorithm

1. Choose an initial parameter vector $\boldsymbol{\Theta}_0$ and set i=0.
2. (E-step) Calculate the expectation of z given the current estimate $\boldsymbol{\Theta}_i$ as well as the observed data
$$\hat{z} = E[z|y, \boldsymbol{\Theta}_i]$$
3. (M-step) Find the parameter vector that maximizes the likelihood or log likelihood function
$$\boldsymbol{\Theta}_{i+1} = \text{argmax}_{\boldsymbol{\Theta}} L(\boldsymbol{\Theta}, \hat{z}, y) = \text{argmax}_{\boldsymbol{\Theta}} L_0(\boldsymbol{\Theta}, \hat{z}, y).$$
4. If $|L(\boldsymbol{\Theta}_{i+1}, \hat{z}, y) - L(\boldsymbol{\Theta}_i, \hat{z}, y)| < \epsilon$ stop and choose $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}_{i+1}$ else set $i = i + 1$ and go to 2, the E-step.

A PH distribution is generated by the time until absorption of the underlying MJP, i.e. by the first time the MJP enters the absorbing state. If only one observation $y = (y^{(1)})$ is available the complete data $x = (x^{(1)})$ about the MJP consist of the information about the initial state of the MJP, the sojourn time during each visit for each state, as well as how often it jumped from state i to state j, and the last visited state prior to absorption in state $m + 1$. Assuming that the complete MJP has been observed the number k of jumps prior to absorption is known and with that the complete data can be written as $x = (s_0, S_0, \ldots, s_{k-1}, S_{k-1})$ where S_i is the ith state visited and s_i is the sojourn time of that visit for the MJP. For a PH distribution with representation $(\boldsymbol{\alpha}, \mathbf{T})$ the likelihood function of this observation can be written as

$$L(\boldsymbol{\alpha}, \mathbf{T}; x) = \alpha_{S_0} e^{\mathbf{T}_{S_0 S_0} \cdot s_0} \mathbf{T}_{S_0 S_1} e^{\mathbf{T}_{S_1 S_1} \cdot s_1} \mathbf{T}_{S_1 S_2} \ldots e^{\mathbf{T}_{S k-1 S_{k-1}} \cdot s_{k-1}} \mathbf{t}_{S_{k-1}}. \tag{2.39}$$

The term $\mathbf{T}_{ij}$ refers to the entry that is in the ith row and jth column of

the sub generator matrix $\mathbf{T}$ and $\mathbf{t}_i$ refers to the ith entry in $\mathbf{t}$. Due to the product form of Equation 2.39, the sufficient statistic consists of (B_i, Z_i, N_{ij}) for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, m, m+1\}$ with B_i equal one if the MJP initiated in state i and 0 otherwise, Z_i equal the accumulated sojourn time in state i and N_{ij} be the total number of jumps from state i to state j. If we define $\mathbf{T}_{i\ m+1} = \mathbf{t}_i$, the complete data likelihood and log likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\alpha}, \mathbf{T}; x) &= f(\boldsymbol{\alpha}, \mathbf{T}; x) = \prod_{i=1}^{m} \alpha_i^{B_i} \prod_{i=1}^{m} e^{\mathbf{T}_{ii} Z_i} \prod_{i=1}^{m} \prod_{j=1, j\neq i}^{m+1} \mathbf{T}_{ij}^{N_{ij}} \\ L_0(\boldsymbol{\alpha}, \mathbf{T}; x) &= \sum_{i=1}^{m} B_i \log(\alpha_i) + \sum_{i=1}^{m} \mathbf{T}_{ii} Z_i + \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} N_{ij} \log(\mathbf{T}_{ij}). \end{aligned}$$

The structure of Equation 2.39, makes it straight forward to cover the case where more than one observation is available. Assume that $y = (y^{(1)}, \ldots, y^{(w)})$ and let $B_i^{(k)}$, $Z_i^{(k)}$, and $N_{ij}^{(k)}$ be the sufficient statistics for the kth observation. Since the log likelihood function is linear with respect to the sufficient statistics we can redefine

$$B_i = \sum_{k=1}^{w} B_i^{(k)}, \ Z_i = \sum_{k=1}^{w} Z_i^{(k)}, N_{ij} = \sum_{k=1}^{w} N_{ij}^{(k)}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, m, m+1$ and $j \neq i$. Regardless if one or several observations are at hand, the challenge of applying the EM algorithm lies in calculating $E_{\boldsymbol{\alpha},\mathbf{T}}[B_i^{(k)}|y^{(k)}]$, $E_{\boldsymbol{\alpha},\mathbf{T}}[Z_i^{(k)}|y^{(k)}]$, and $E_{\boldsymbol{\alpha},\mathbf{T}}[N_{ij}^{(k)}|y^{(k)}]$. Asmussen et al (1996) managed to derive closed from solutions for these expectations

using mainly probabilistic arguments and obtained:

$$\begin{aligned}
E_{\boldsymbol{\alpha},\mathbf{T}}[B_i^{(k)}|y^{(k)}] &= \frac{\alpha_i \mathbf{e}_i^T e^{\mathbf{T}y^{(k)}}\mathbf{t}}{\boldsymbol{\alpha} e^{\mathbf{T}y^{(k)}}\mathbf{t}}, \\
E_{\boldsymbol{\alpha},\mathbf{T}}[Z_i^{(k)}|y^{(k)}] &= \frac{\int_0^{y_k} \boldsymbol{\alpha} e^{\mathbf{T}u}\mathbf{e}_i\mathbf{e}_i' e^{\mathbf{T}(y_k-u)}\mathbf{t}du}{\boldsymbol{\alpha} e^{\mathbf{T}u}\mathbf{t}}, \\
E_{\boldsymbol{\alpha},\mathbf{T}}[N_{ij}^{(k)}|y^{(k)}] &= \frac{\mathbf{T}_{ij}\int_0^{y_k} \boldsymbol{\alpha} e^{\mathbf{T}u}\mathbf{e}_i\mathbf{e}_j' e^{\mathbf{T}(y_k-u)}\mathbf{t}du}{\boldsymbol{\alpha} e^{\mathbf{T}y_k}\mathbf{t}}, \\
E_{\boldsymbol{\alpha},\mathbf{T}}[N_{im+1}^{(k)}|y^{(k)}] &= \frac{\mathbf{t}_i\boldsymbol{\alpha} e^{\mathbf{T}y_k}\mathbf{e}_i}{\boldsymbol{\alpha} e^{\mathbf{T}y_k}\mathbf{t}} \text{ for } i,j \in \{1,\ldots,m\}. \qquad (2.40)
\end{aligned}$$

Given the sufficient statistic, differentiating and evaluating $L_0(\boldsymbol{\alpha}, \mathbf{T}; x)$ results in the following maximum likelihood estimates:

$$\widehat{\mathbf{T}}_{ij} = \frac{N_{ij}}{Z_i}, \ \hat{\alpha}_i = \frac{B_i}{w} \quad \forall i,j. \qquad (2.41)$$

The results from Equation 2.41 together with Equation 2.40 can be used in Algorithm 1. However, evaluating the expressions for the conditional expectations directly can be rather costly mainly due to the numerical integration. Alternatively, the equations can be used to establish a linear system of homogeneous differential equations that can be solved using standard methods, such as the Runge-Kutta method (Asmussen, 1992). Yet another approach to reduce the calculation cost is the prior mentioned method of uniformization (Bladt et al, 2011).

2.2.2 Fisher Information Matrix for estimates obtained through the EM algorithm

In order to fit a probability distribution to data and make the distribution usable for practitioners it is essential to estimate the unknown parameters. However, using estimates without knowledge of their quality is a risky game. The EM algorithm is often used in order to obtain maximum likelihood estimates while avoiding to evaluate the observed data likelihood and its derivatives. The EM algorithm has been criticized for the lack of possibility to

calculate the Fisher Information matrix. This issue has been addressed and solved by Oakes (1999). His paper is a standard reference concerning Fisher information in connection to the EM algorithm. McLachlan and Krishnan dedicated an entire book to the EM algorithm (McLachlan and Krishnan, 2007), collecting several methods for obtaining the Fisher information. A more specialized approach comes from Bladt et al (2011), they use the idea of Oakes (1999) and combine it with the method of uniformization to apply the EM algorithm to PH distributions with a minimal representation.

Oakes (1999) calculated the Fisher information by using the complete data likelihood function. Define the score statistics of the complete data x as the gradient of the complete data log likelihood function

$$\boldsymbol{S}_c(x, \boldsymbol{\Theta}) = \frac{\partial L_0(\boldsymbol{\Theta}, x)}{\partial \boldsymbol{\Theta}} \tag{2.42}$$

then the Fisher information matrix can be written as:

$$\mathbf{I}(\hat{\boldsymbol{\Theta}}, y) = E_{\boldsymbol{\Theta}}\left[-\frac{\partial^2 L_0(\widehat{\boldsymbol{\Theta}}, x)}{\partial \boldsymbol{\Theta}, \partial \boldsymbol{\Theta}^T} \middle| y \right] - E_{\boldsymbol{\Theta}}\left[S_c(x, \boldsymbol{\Theta}) S_c(x, \boldsymbol{\Theta})^T | y \right]. \tag{2.43}$$

The second term in Equation 2.43 is often referred to as the missing information due to the incomplete data. The error terms of the estimates are asymptotically multivariate normal distributed, and the inverse of the Fisher information matrix can be used to approximate the variance of the estimates. When estimating parameters that are bound to the positive axis, e.g. when estimating intensity parameters of exponential distributions, the error of the estimates can not be normally distributed. A solution is to use the Fisher information matrix that approximates the error on the estimates for the log parameters to establish confidence intervals. Reversing the log transform then yields sounder confidence intervals for the estimated non-transformed parameters.

When estimating parameters for PH distributions, explicit expressions for calculating the Fisher information matrix can be given (Bladt et al, 2011). Their approach is based on Oakes (1999) splitting the log likelihood function

similarly to Equation 2.43. For observed data $y = (y^{(1)}, \ldots, y^{(w)})$ of a MJP with representation $(\boldsymbol{\alpha}, \mathbf{T})$ on a m dimensional state space they express the Fisher information matrix in terms of first derivatives with respect to $\boldsymbol{\alpha}$ and $\mathbf{T}$ of

$$U_i = \sum_{l=1}^{w} \frac{\mathbf{e}_i^T e^{\mathbf{T}y_l}\mathbf{t}}{\boldsymbol{\alpha} e^{\mathbf{T}y_l}\mathbf{t}} \tag{2.44}$$

$$W_i = \sum_{l=1}^{w} \frac{\boldsymbol{\alpha} e^{\mathbf{T}y_l}\mathbf{t}}{\boldsymbol{\alpha} e^{\mathbf{T}y_l}\mathbf{e}_i} \tag{2.45}$$

$$V_{ij} = \sum_{l=1}^{w} \frac{1}{\boldsymbol{\alpha} e^{\mathbf{T}y_l}\mathbf{t}} \int_0^{y_l} \mathbf{e}_j^T e^{\mathbf{T}(y_l-u)}\mathbf{t}\boldsymbol{\alpha} e^{\mathbf{T}u}\mathbf{e}_i du. \tag{2.46}$$

There is an obvious connection between the results in Equation 2.46 and the results for calculating the conditional expectations from Equation 2.40. These general results concerning the Fisher information matrix when obtaining estimates via the EM algorithm for PH distribution, as well as the analytical results for calculating the conditional expectation of the sufficient statistics, are based on two attributes of PH distributions: The closed form solutions (see for example Equation 2.3) for most stochastic properties and the probabilistic structure of the underlying MJP.

Often, natural pheonoma and therefore available data is not univariate. This causes a demand for similar results for multivariate data. Due to the lack of proper alternatives, the standard approach is to use the multivariate normal distribution.

CHAPTER 3

Multivariate phase type distributions

An alternative to the class of multivariate normal distributions, when analyzing multivariate data, is the class of multivariate PH distributions. The differences between multivariate PH distributions and multivariate normal distributions are similar to the differences between univariate PH distributions and univariate normal distributions. Where multivariate normal distributions are symmetric distributions on $\mathbb{R}^n$, multivariate PH distributions are only defined on $\mathbb{R}^n_+$ and the marginals are skewed. Multivariate phase type distributions are a generalization of the univariate class of PH distributions to higher dimensions. They can be characterized similarly to the multivariate normal distribution where a vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is multivariate normal distributed if and only if every linear combination of the entries of $\mathbf{Y}$ is univariate normal distributed. Hitherto there are three different ways of defining classes of MPH distributions. These definitions are in chronological order from Assaf et al (1984), Kulkarni (1989), and Bladt and Nielsen (2010).

In the following sections, I will present the definitions of the different classes of multivariate PH distributions. Though they are all extensions of the univariate case, they are partly subclasses of each other, they are of different structure and use different abbreviations to clarify which class and construction is used. I will not proceed in chronological order. In Section 3.1 I will introduce the class of multivariate PH distributions by Assaf et al (1984). This class uses several different absorbing states in the construction of the multivariate random vector and is denoted with MPH. Section 3.2 will be used to briefly introduce the class of multivariate phase type distributions as defined by Bladt and Nielsen (2010). This class is defined using projections of the marginal distributions of a multivariate random vector. In order to distinguish this class from the class of MPH distributions we will denote it with MVPH. The focus will be on Section 3.3, introducing the class of multivariate PH distributions defined by Kulkarni (1989). He denoted his class with MPH* to distinguish it from the class of MPH distributions. This class is based on using a MJP with only one absorbing state, resulting in a promising structure for parameter estimation as well as for stochastic modeling.

3.1 Multivariate PH distribution by Assaf et al

Assaf et al (1984) extended the univariate class of PH distributions by one of two straightforward extensions. Where a univariate PH distribution is constructed by an absorbing MJP and the distribution of time until it reaches the absorbing state, they used an absorbing MJP and considered the joint distribution of times until certain subspaces are reached for the first time. The times until these subspaces are reached define the random vector $\mathbf{Y}$.

Definition 3.1. *The class of MPH distributions* Let $\{X(t)\}_{t\geq 0}$ be a MJP on a finite state space E with one absorbing state $m+1$. Let Γ_i, $i \in \{1, 2\ldots, n\}$, be nonempty absorbing subspaces of E, with $\bigcap_i \Gamma_i = m+1$. Define $Y_i = \inf\{t \geq 0 : X(t) \in \Gamma_i\}$ as the first hitting time of Γ_i by the

MJP. The vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is said to be multivariate phase type distributed in the class of MPH distributions.

The following example should clarify the term "'nonempty absorbing subspaces"'.

Example 3.1. *Absorbing subspaces in the class of MPH distributions* Let us consider a bivariate MPH distributed random variable $\boldsymbol{Y} = (Y_1, Y_2)$ where the underlying MJP $\{X(t)\}_{t\geq 0}$ is on the state space $E = \{1, 2, 3, 4, 5\}$, with 5 being the absorbing state, and with sub generator matrix

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ p \cdot \lambda_2 & -\lambda_2 & (1-p) \cdot \lambda_2 & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 \\ 0 & 0 & p \cdot \lambda_4 & -\lambda_4 \end{pmatrix}.$$

In this example $Y_1 = \min(t \geq 0 : X(t) = \{3, 4, 5\})$ and $Y_2 = \min(t \geq 0 : X(t) = 5)$, once the MJP reaches the state 3 it can never jump back to the states 1 and 2. With other words $\Gamma_1 = \{3, 4, 5\}$ and $\Gamma_2 = \{5\}$.

Similar to the class of PH distributions, there exist general procedures for deriving the survival function, the LST, as well as all moments for an MPH distribution. However, these procedures do not result in closed form expressions. Furthermore, this class is closed under similar operations as the univariate class, e.g. finite mixtures (Assaf et al, 1984). Due to the use of overlapping absorbing subsets the sub generator matrix has to consist of block matrices where the lower diagonal blocks consist of zeros.

3.2 Multivariate PH distributions by Bladt and Nielsen

Bladt and Nielsen (2010) defined multivariate PH distributions in relation to univariate PH distributions, similar to how the multivariate normal distribution can be defined in terms of the univariate normal distributions.

Definition 3.2. *The class of MVPH distributions* A vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ follows a multivariate phase type distribution (MVPH) if the inner product $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ has a (univariate) phase type distribution for all non-negative vectors $\boldsymbol{a}$ where $\boldsymbol{a} \neq (0, \ldots, 0)$.

They argued that the class of MPH* distributions by Kulkarni is a subclass of the class of MVPH distributions. This can be shown directly using Definition 3.2 as well as Theorem 6 from Kulkarni (1989). Hitherto it is not clear if it is a strict subset or if the class of MPH* distributions equals the class of MVPH distributions but it can be shown that all distribution contained in the class of MPH* are contained in the class of MVPH (Bladt and Nielsen, 2010) From a mathematical perspective Definition 3.2 is elegant and creates an analogue to the definition of a multivariate Normal distribution. However, Definition 3.2 is not intuitive when modeling stochastic phenomena. The class of MVPH is not well understood and requires further investigation.

3.3 Multivariate PH distributions by Kulkarni

Kulkarni (1989) used a different approach than Assaf et al (1984) as well as than Bladt and Nielsen (2010). He constructed a class of multivariate PH distribution called MPH* by using linear combinations of the occupation times prior to absorption of a univariate PH distribution. Furthermore, he showed that the class MPH by Assaf is a strict subset of MPH*.

Definition 3.3. *Multivariate phase type* distributions* Let $\{X(t), t \geq 0\}$ be a MJP with state space $E = \{1, \ldots, m, m+1\}$ and PH representation $(\boldsymbol{\alpha}, \mathbf{T})$. Define the random variable $\tau = \min(t \geq 0 : \ X(t) = m+1)$ as the time of absorption and define the non-negative reward matrix $\mathbf{R}$

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ r_{m1} & r_{m2} & \ldots & r_{mn} \end{pmatrix} \tag{3.1}$$

and $r_j(i) = r_{ij}$. With that, construct the random variables

$$Y_j = \int_0^\tau r_j\left(X(t)\right) dt,\ 1 \leq j \leq n. \tag{3.2}$$

The random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is then said to follow an MPH* distribution.

I will use Example 3.1 to show how to construct an MPH* representation for any given MPH distribution. This can be understood as an intuitive proof of MPH$\subseteq$MPH*.

Example 3.2. *Deriving an MPH* representation for a given MPH distribution* Let us consider the bivariate MPH distributed random variable $\boldsymbol{Y} = (Y_1, Y_2)$ from Example 3.1, the underlying MJP $\{X(t)\}_{t\geq 0}$ is on the state space $E = \{1, 2, 3, 4, 5\}$, with 5 being the absorbing state, and with sub generator matrix

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ p \cdot \lambda_2 & -\lambda_2 & (1-p) \cdot \lambda_2 & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 \\ 0 & 0 & p \cdot \lambda_4 & -\lambda_4 \end{pmatrix}.$$

In the example the random variables are defined as $Y_1 = \min(t \geq 0 : X(t) = \{3, 4, 5\})$ and $Y_2 = \min(t \geq 0 : X(t) = 5)$. An MPH$^\star(\boldsymbol{\alpha}, \mathbf{T}^\star, \mathbf{R})$ representation is $\boldsymbol{\alpha} = (1, 0, 0, 0)$, $\mathbf{T}^\star = \mathbf{T}$ and

$$\mathbf{R} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \tag{3.3}$$

In a similarl manner it is possible to find an MPH* representation for all distributions belonging to the class of MPH distributions.

The restrictions for the sub generator matrix of an MPH* distribution are the same as in the univariate case, making it possible to extend any univariate PH distribution to an MPH* distribution. It can be shown that any given MPH distribution can be expressed in terms of at least one MPH* representation. When simulating multivariate PH distributions, in case of MPH distributions, the hitting times of different subspaces are essential. When simulating MPH* distributions, the occupation times in the different states prior to absorption of the underlying MJP are of interest.

Recent work by Esparza (2011) shows how the results from the EM algorithm for parameter estimation of univariate PH distributions can be modified to obtain parameter estimates for a subclass of MPH*. The results can only be applied to MPH* distributions when the state space of the underlying MJP can be split into absorbing and overlapping subspaces. For a bivariate PH* distribution the sub generator matrix $\mathbf{T}$ has to be of the following form:

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_2 \end{pmatrix}. \tag{3.4}$$

The structure used by Esparza (2011) is only applicable to a small subclass of the the general class of MPH* distributions.

The class MPH* can be used to construct distributions on the whole reals by using a reward vector that is not necessarily non-negative (Bladt et al, 2013). This class is called the class of bilateral phase type* (BPH*) distributions. The original idea for constructing BPH distribution goes back to Ahn and Ramaswami (2005), and the class can be reconstructed when the reward vector of a BPH* distribution is a non-zero vector with real valued entries. The class of BPH* can easily be extended in a multivariate setting denoted MBPH* (Bladt et al, 2011). Once the class of MPH* is better understood and general results are available, it is likely that these results can be used for analyzing the class of MBPH*.

Numerous examples of multivariate mixtures of exponential distributions can be found in Kotz et al (2000). A number of them are contained in the

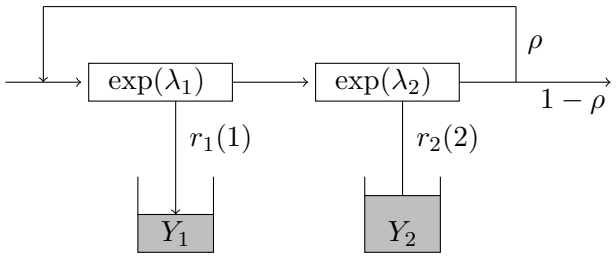


Figure 3.1: Graphical depiction of Kibble's bivariate mixtures of exponential distributions

class of MPH* (Bladt and Nielsen, 2010) and are used as typical examples. One specific bivariate distribution is the Kibble distribution.

Example 3.3. *Kibble's bivariate mixtures of exponential distributions* The Kibble distribution (Kibble, 1941) is a bivariate mixture of Gamma distributions with shape parameter k_1 and k_2, intensity parameters λ_1 and λ_2 and correlation parameter ρ. When $k_1 = k_2 = 1$, an MPH* representation can be given with $\boldsymbol{\alpha} = (1, 0)$ and

$$\mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho \cdot \lambda_2 & -\lambda_2 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.5}$$

This representation can be depicted graphically as shown in Figure 3.1. The probability density function for Kibble's bivariate mixture of exponential distributions can for example be found in Kotz et al (2000). Alternately, it can be derived by conditioning on the number of times the different states are visited. Conditioned on the number of visits, the accumulated occupation times in each state are independent and Erlang distributed. The density of Kibble's bivariate mixture of exponential distributions can be written as

$$f_{Y_1,Y_2}(y_1, y_2) = \lambda_1\lambda_2 e^{-\lambda_1 y1 - \lambda_2 y_2}(1-\rho) I_0(2\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2}) \tag{3.6}$$

where $I_q(z) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+q+1)} \left(\frac{z}{2}\right)^{2i+q}$ is the modified Bessel function of the first kind.

Opposite to the class MPH by Assaf, for distributions contained in the class MPH* there exist no general procedures for constructing the probability density function and the survival function, i.e. the expressions need to be derived individually for each distribution. However, Kulkarni (1989) derived methods for obtaining the LST, for simulation as well as systems of linear partial differential equations (PDE) for the survival function. The survival function $\bar{F}(x)$ of a random variable X is defined as $P(X > x)$:

$$\bar{F}(x) = 1 - F(x). \tag{3.7}$$

The survival function for an MPH* distributed random vector $\boldsymbol{Y}$ can be derived using the survival function conditioned on the initial state of the underlying MJP $(X(t), t \geq 0)$ on the state space $E = \{1, \ldots, m, m+1\}$. It is defined as

$$\bar{F}_i(y_1, \ldots, y_n) = P\left(Y_1 > y_1, \ldots, Y_n > y_n | X(0) = i\right). \tag{3.8}$$

Theorem 3.1. Survival function for MPH* distributions (Kulkarni, 1989) *The functions $\bar{F}_i(y_1, \ldots, y_n)$, $1 \leq i \leq m$, satisfy the following system of simultaneous linear partial differential equations.*

$$\sum_{j=1}^{n} r_j(i) \frac{\partial \bar{F}_i}{\partial y_i} = \sum_{j=1}^{m} \mathbf{T}_{ij} \bar{F}_j, 1 \leq i \leq m. \tag{3.9}$$

Proof. The proof can be found in Kulkarni (1989). □

A master thesis (Qin, 2011) as well as personal research time has been dedicated to find closed form solutions to the PDEs obtained from Theorem 3.1. From a general perspective the results have been rather disappointing. Qin (2011) used a power series extension to derive recursive equations for

calculating the survival function of the Kibble distribution presented in Example 3.3. A generalization of the results to subclasses of MPH* or the whole class has hitherto not been possible.

As mentioned in Chapter 2, part of the versatility of univariate PH distributions is the existence of closed form solutions for most major statistical properties. This allows for derivation of general valid expressions for the EM algorithm. Recent progress in the area of MPH* distributions has been achieved by Bladt and Nielsen (2010) deriving closed form solutions for all cross moments of an MPH* distribution, leading the way towards using general MPH* distributions for stochastic modeling.

Theorem 3.2. Cross-moments for MPH* distributions *The cross-moments* $\mathbf{E}\left(\prod_{i=1}^n Y_i^{k_i}\right)$*, where* $\boldsymbol{Y}$ *follows an MPH* distribution with representation* $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$ *and where* $k_i \in \mathbb{N}$*, are given by*

$$\mathbf{E}\left(\prod_{i=1}^n Y_i^{k_i}\right) = \boldsymbol{\alpha} \sum_{l=1}^{k!} \prod_{i=1}^{k} (-\mathbf{T})^{-1} \Delta(\boldsymbol{r}_{\sigma_l(i)}) \mathbf{e}^T. \tag{3.10}$$

Here $k = \sum_{i=1}^n k_i$*.* $\boldsymbol{r}_j$ *is the* r*th column of* $\mathbf{R}$ *and* σ_l *is one of the* $r!$ *possible ordered permutations of the derivatives, with* $\sigma_l(i)$ *being the value among* $1, \ldots, n$ *at the* i*th position of that permutation. Furthermore,* $\Delta(\boldsymbol{r})$ *is a diagonal matrix with the entries of the vector* $\boldsymbol{r}$ *on the diagonal and zeros elsewhere.*

Proof. Though the theorem is due to Bladt and Nielsen (2010) an explicit proof can be found in Nielsen et al (2010). □

Often only lower order moments and cross-moments are of interest and the results can be simplified to

$$\begin{aligned}
\mathbf{E}(Y_i) &= \boldsymbol{\alpha}(-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_i)\mathbf{e}^T = \boldsymbol{\alpha}(-\mathbf{T})^{-1}\boldsymbol{r}_i \\
\mathbf{E}(Y_i^2) &= \boldsymbol{\alpha}(-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_i)(-\mathbf{T})^{-1}\boldsymbol{r}_i + \boldsymbol{\alpha}(-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_i)(-\mathbf{T})^{-1}\boldsymbol{r}_i = 2\boldsymbol{\alpha}((-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_i)(-\mathbf{T})^{-} \\
\mathbf{E}(Y_iY_j) &= \boldsymbol{\alpha}(-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_i)(-\mathbf{T})^{-1}\boldsymbol{r}_j + \boldsymbol{\alpha}(-\mathbf{T})^{-1}\Delta(\boldsymbol{r}_j)(-\mathbf{T})^{-1}\boldsymbol{r}_i.
\end{aligned}$$

Despite these analytical results, no general class of multivariate PH distributions seems easily fit for parameter estimations. The distributions are often over parameterized and therefore estimates are not unique. In order to obtain parameter estimation methods similar to the ones available for univariate PH distributions it is essential to have general ways of deriving the probability density function as well as a structure allowing for probabilistic arguments.

From the three classes of multivariate PH distributions presented, the class of MPH* distributions is by construction the closest to the univariate class of PH distributions making it favorable for modeling. Furthermore, it has potential for similar parameter estimation methods as used for univariate PH distributions. I provide a summary of our efforts in this regard in Chapter 4.

CHAPTER 4

Parameter estimation via the EM algorithm for the Kibble distribution

Regardless of the fact that the EM algorithm presented in Section 2.2.1 is a widely used tool for fitting PH distributions and additional methods are available, hardly any work has been done for fitting multivariate PH distributions. Very few examples are at hand, e.g. estimating parameters for a clinical trial (Ahlström et al, 1999) and Esparza (2011) derived in her PhD thesis an EM algorithm for special bivariate PH distributions of the class MPH*. Her approach is restricted in the sense that the state space has to be partitionable into absorbing subspaces similar to the requirements for the class MPH.

From a modeling perspective the class of MPH* distributions seems intuitive and straightforward to use. Either states or subsets of the state space can be considered as tasks in a project. These task then contribute differently to several factors, e.g. cost, duration, and environmental impact. Opposite

to the univariate case, there exist no general applicable methods to obtain maximum likelihood estimators for parameters of multivariate phase type distributions. It is a challenging task to derive such general methods, partly caused by the lack of general (analytic) results for obtaining the density of MPH* distributions. Prior attempts during this PhD study to derive closed form solutions for the probability density function and the survival function, similar to the Equations 2.3-2.5 have unfortunately been unsuccessful. Until general results are available, subclasses of MPH* and sometimes even individual distributions, have each to be analyzed individually in order to derive their probabilistic properties.

A special MPH* distribution is the Kibble distribution presented in Example 3.3. It is a bivariate distribution where the marginals are Gamma distributed. With integer valued shape parameters the Kibble distribution belongs to the class of MPH* distributions. For these cases the marginals are Erlang distributed. The distribution was originally derived by Kibble (1941), who calculated the joint distribution of the empirical variances in samples from a bivariate normal distribution. It is a special distribution in the sense that it is broadly applicable and known under different names, e.g. as Moran and Downton's, Jensen's or Gaver's distribution. Several other examples can be found in the book "Continuous Multivariate Distributions" by Kotz et al (2000). Examples for applications can be found several places. Wang and Gosh (2000) use it for competing risk models and Yue et al (2001) show several applications in hydrology, for example flows in different parts of a river. Chatelain et al (2007) use bivariate Gamma distributions to detect changes in radar images. Examples of estimation methods can be found abundantly. Kotz et al (2000) list several moment based estimators and summarizes their advantages and disadvantages. Chatelain et al (2007), apply a maximum likelihood approach using general numerical methods for finding the root of the first derivative of the probability density function in order to estimate the correlation parameter. Iliopoulos et al (2005) as well as Lin et al (2013) suggest a Markov chain Monte Carlo approach to obtain Bayesian estimates.

The vast amount of different estimation methods and areas of application

show the importance of the Kibble distribution and with that the need for accurate parameter estimation methods.

Hitherto, nobody has exploited the fact that under certain conditions the Kibble distribution can be interpreted as an MPH* distribution. The Kibble distribution is therefore an interesting starting point for developing parameter estimation techniques for MPH* distributions as well as statistical inference, with several alternative estimation methods to be used for validation.

For simplicity, this thesis considers first the case where the shape parameters are identical and one. These results can be found in Section 4.1. In Section 4.2 our general results are presented, removing the restriction on the shape parameter. The shape parameters can be of any value as long as the distribution stays in the class of MPH* distributions. These results have been combined in Appendix A and are under review with the "Journal of Stochastic Modeling".

4.1 Bivariate mixtures of exponential distributions

An example for an MPH* representation as well as a graphical interpretation for the Kibble distribution when the shape parameters k_1 and k_2 are one can be found in Example 3.3. The probability density function in Equation 3.6 can be derived using the MPH* interpretation and conditioning on the number N_1 of visits to the state with $exp(\lambda_1)$ distributed sojourn time as well as the number N_2 of visits to the states where the sojourn time is $exp(\lambda_2)$ distributed. For the case that the shape parameters are equal we get $N_1 = N_2$. Conditioned on the number of visits the accumulated sojourn

times Y_1 and Y_2 are independent and Erlang distributed.

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \sum_{n=1}^{\infty} P(N_1=n) f_{Y_1,Y_2}(y_1,y_2|N=n) \\ &= \sum_{n=1}^{\infty} \rho^{n-1}(1-\rho)\lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n-1}}{(n-1)!} e^{-\lambda_2 y_2} \\ &= \lambda_1\lambda_2 e^{-\lambda_1 y 1 - \lambda_2 y_2}(1-\rho) I_0(2\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2}) \end{aligned} \tag{4.1}$$

Section 2.2.1 describes how the EM algorithm can be used to estimate parameters for univariate PH distributions. If we consider one observation for the Kibble distribution (y_1, y_2) the complete data is then $x = (y, N_{12}, N_{21})$. The complete data likelihood function can be written as

$$\begin{aligned} f(x;\boldsymbol{\alpha},\mathbf{T}) &= \alpha_1^{B_1}\alpha_2^{B_2} \cdot e^{-\lambda_1 \cdot Z_1} \cdot e^{-\lambda_2 \cdot Z_2} \lambda_1^{N_{12}} \cdot (p\lambda_2)^{N_{21}} \cdot ((1-\rho)\lambda_2) \\ &= 1 \cdot e^{-\lambda_1 \cdot y_1} \cdot e^{-\lambda_2 \cdot y_2} \lambda_1^{N_1} \cdot (\rho\lambda_2)^{N_1 - 1} \cdot ((1-\rho)\lambda_2). \end{aligned} \tag{4.2}$$

In Equation 4.2 we use the fact that $N_1 = N_2$. An immediate result is that the number of jumps form state 1 to state 2, $N_{12} = N_1$, and the number of jumps form state 2 to state 1, $N_{21} = N_1 - 1$, are linearly connected. Furthermore, the absorbing state 3 can only be reached from state 2 and with that we get directly $N_{23} = 1$. The structure of the model only allows for $\alpha_1 = 1$ as well as $\alpha_2 = 0$. Since $\boldsymbol{\alpha}$ is to be $\boldsymbol{\alpha} = (1, 0)$, it can be omitted in the likelihood function from now on. The likelihood function belongs to the family of exponential distributions and therefore allows us to replace unobserved data with their sufficient statistics, similar to the approach for univariate PH distributions described in Section 2.2.1.

In order to apply the EM algorithm to Kibble's bivariate mixture of exponential distributions it is essential to derive the calculations for the E and the M step. Deriving the M step is straightforward by differentiating the

logarithm of Equation 4.2

$$\frac{\partial \log(f(x,\mathbf{T}))}{\partial \lambda_1} = -y_1 + \frac{N_1}{\lambda_1} \tag{4.3}$$

$$\frac{\partial \log(f(x,\mathbf{T}))}{\partial \lambda_2} = -y_2 + \frac{N_2}{\lambda_2} \tag{4.4}$$

$$\frac{\partial \log(f(x,\mathbf{T}))}{\partial \rho} = \frac{N_1}{\rho} - \frac{1}{1-\rho} \tag{4.5}$$

and finding the root in order to obtain the maximum likelihood estimates which yields

$$\hat{\lambda}_1 = \frac{N_1}{y_1},\ \hat{\lambda}_2 = \frac{N_2}{y_2} \text{ and } \hat{\rho} = \frac{N_1 - 1}{N_1}. \tag{4.6}$$

For a set of w independent observations $\left\{y_1^{(1)}, y_2^{(1)}, \ldots, y_1^{(w)}, y_2^{(w)}\right\}$, with $N^{(v)}$ being the number of jumps associated to the vth observation, the complete data likelihood function can be written as

$$f(x,\mathbf{T}) = e^{-\lambda_1 \cdot (\sum_{v=1}^{w} y_1^{(v)})} \cdot e^{-\lambda_2 \cdot (\sum_{v=1}^{w} y_2^{(v)})} \tag{4.7}$$

$$\cdot\ \lambda_1^{\sum_{v=1}^{w} N_1^{(v)}} \cdot (\rho\lambda_2)^{\sum_{v=1}^{w}(N_2^{(v)}-1)} \cdot ((1-\rho)\lambda_2)^w. \tag{4.8}$$

The maximum likelihood estimates for w observations then become:

$$\hat{\lambda}_1 = \frac{\sum_{v=1}^{w} N^{(v)}}{\sum_{v=1}^{w} y_1^{(v)}}, \tag{4.9}$$

$$\hat{\lambda}_2 = \frac{\sum_{v=1}^{w} N^{(v)}}{\sum_{v=1}^{w} y_2^{(v)}}, \tag{4.10}$$

$$\hat{\rho} = \frac{\sum_{v=1}^{w} \left(N^{(v)} - 1\right)}{\sum_{v=1}^{w} N^{(v)}}. \tag{4.11}$$

These solutions for the M step of the EM algorithm for Kibble's bivariate mixtures of exponential distributions are very similar to the results of the

M step in the EM algorithm for univariate PH distributions from Equation 2.41. The derivation of the E step presents more challenges. One advantage of the Kibble distribution is, the variables Y_1 and Y_2 are the accumulated occupation times in the different states. The only unknown variables of the complete data likelihood are N_1 and N_2, which are only one unknown.

Lemma 4.1. Conditional expectation for Kibbles bivariate mixtures of exponentials *For the Kibble distribution with MPH* representation* $(\boldsymbol{\alpha}, \mathbf{T}, \mathbf{R})$

$$\boldsymbol{\alpha} = (1, 0),\ \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho\lambda_2 & -\lambda_2 \end{pmatrix},\ \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

the conditional expectation of the number of visits N_1 to the state with sojourn time distribution being exp(λ_1), given the observed values y_1 and y_2 as well as the current estimates for λ_1, λ_2, and ρ is

$$E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1|y_1, y_2] = \sqrt{\lambda_1 y_1 \lambda_2 y_2 \rho} \cdot \frac{I_1(2\sqrt{\lambda_1 y_1 \lambda_2 y_2 \rho})}{I_0(2\sqrt{\lambda_1 y_1 \lambda_2 y_2 \rho})} + 1 \tag{4.12}$$

where $I_q(z) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+q+1)} \left(\frac{z}{2}\right)^{2i+q}$ is the modified Bessel function of the first kind.

Proof. For the proof we use standard calculations for conditional probabil-

ities and the definition of the modified Bessel function of the first kind.

$$
\begin{aligned}
E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1|y_1,y_2] &= \sum_{k=1}^{\infty} kP(N_1=k|Y_1=y_1,Y_2=y_2) \\
&= \frac{\sum_{k=1}^{\infty} k f_{Y_1,Y_2|N_1}(y_1y_1|N_1=k)\rho(N_{12}=k)}{f_{Y_1,Y_2}(y_1,y_2)} \\
&= \frac{\sum_{k=1}^{\infty} k(1-\rho)\rho^{k-1}\lambda_1 e^{-\lambda_1 y_1}\lambda_2 e^{-\lambda_2 y2}\frac{(\lambda_1 y_1\lambda_2 y_2)^{k-1}}{((k-1)!)^2}}{\sum_{k=1}^{\infty}(1-\rho)\rho^{k-1}\lambda_1 e^{-\lambda_1 y_1}\lambda_2 e^{-\lambda_2 y2}\frac{(\lambda_1 y_1\lambda_2 y_2)^{k-1}}{((k-1)!)^2}} \\
&= \frac{\sum_{k=0}^{\infty}(k+1)\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} \\
&= \frac{\sum_{k=0}^{\infty}(k)\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} + \frac{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} \\
&= \frac{\sum_{k=1}^{\infty}(k)\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} + 1 \\
&= \frac{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^{k+1}}{((k+1)!(k)!)}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} + 1 \\
&= \frac{\sqrt{\lambda_1 y_1\lambda_2 y_2\rho}\sum_{k=0}^{\infty}\frac{(\sqrt{\lambda_1 y_1\lambda_2 y_2\rho})^{2k+1}}{((k+1)!(k)!)}}{\sum_{k=0}^{\infty}\frac{(\lambda_1 y_1\lambda_2 y_2\rho)^k}{((k)!)^2}} + 1 \\
&= \sqrt{\lambda_1 y_1\lambda_2 y_2\rho}\cdot\frac{I_1(2\sqrt{\lambda_1 y_1\lambda_2 y_2\rho})}{I_0(2\sqrt{\lambda_1 y_1\lambda_2 y_2\rho})} + 1
\end{aligned}
$$

□

As described in Section 2.2.2 the EM algorithm for univariate PH distributions allows for obtaining the Fisher information matrix directly. It is not

possible to proceed in a similar fashion when dealing with MPH* distributions. However, the Fisher information matrix can be calculated following the calculations on page 108-109 in McLachlan and Krishnan (2007)

$$\mathbf{I}(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\rho}, y) = E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}\left[\begin{pmatrix} \frac{N_1}{\lambda_1^2} & 0 & 0 \\ 0 & \frac{N_1}{\lambda_2^2} & 0 \\ 0 & 0 & \frac{N_1-1}{\rho^2} + \frac{1}{(1-\rho)^2} \end{pmatrix} \middle| y \right]$$
$$+ \quad E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}\left[\begin{pmatrix} -y_1 + \frac{N}{\lambda_1} \\ -y_2 + \frac{N_1}{\lambda_2} \\ \frac{N_1-1}{\rho} - \frac{1}{(1-\rho)} \end{pmatrix} \cdot \begin{pmatrix} -y_1 + \frac{N_1}{\lambda_1} & -y_2 + \frac{N_1}{\lambda_2} & \frac{N_1-1}{\rho} - \frac{1}{(1-\rho)} \end{pmatrix} \middle| y \right].$$

The conditional expectation of N_1, $E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1|y_1, y_2]$ is used in the EM algorithm and is therefore directly available for the calculation of the Fisher information matrix. For this special case of the Kibble distribution calculating $E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1^2|Y_1 = y_1, Y_2 = y_2]$ results in a similar expression.

Lemma 4.2. Conditional second moment for Kibbles bivariate mixture of exponentials *For the Kibble distribution with MPH* representation* $(\alpha, \mathbf{T}, \mathbf{R})$*,*

$$\boldsymbol{\alpha} = (1, 0), \ \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho\lambda_2 & -\lambda_2 \end{pmatrix}, \ \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

the conditional second moment of the number of visits N_1 to the state where the sojourn time follows an exp(λ_1) distribution, given the observed values y_1 and y_2 as well as the current estimates for λ_1, λ_2, and ρ is

$$E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1^2|y_1, y_2] \quad = \quad 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + 1 + \rho\lambda_1\lambda_2 y_1 y_2 \quad (4.13)$$

Proof. The proof uses the same arguments are the proof to Lemma 4.1.

$$E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1^2|y_1,y_2] = \sum_{n=1}^{\infty} nP(N=n|y_1,y_2) \tag{4.14}$$

$$\begin{aligned}
&= \frac{\sum_{n=1}^{\infty} n^2\rho^{n-1}(1-\rho)\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}\lambda_2\frac{(\lambda_2 y_2)^{n-1}}{(n-1)!}e^{-\lambda_2 y_2-\lambda_1 y_1}}{\sum_{n=1}^{\infty}\rho^{n-1}(1-\rho)\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}\lambda_2\frac{(\lambda_2 y_2)^{n-1}}{(n-1)!}e^{-\lambda_2 y_2-\lambda_1 y_1}} \\
&= \frac{\sum_{k=0}^{\infty}\frac{(k+1)^2}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}} \\
&= \frac{\sum_{k=0}^{\infty}\frac{k^2}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}} + 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N_1|y_1,y_2]\right) + 1 \\
&= \frac{\sum_{k=0}^{\infty}\frac{k}{(k-1)!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}} + 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2]\right) + 1 \\
&= \frac{\sum_{k=1}^{\infty}\frac{k}{(k-1)!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}} + 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2]\right) + 1 \\
&= \frac{\sum_{s=0}^{\infty}\frac{s+1}{(s)!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+1}}{(s+1)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^k}{(k)!}} + 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2]\right) + 1 \\
&= 2\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2]\right) + 1 + \rho\lambda_1\lambda_2 y_1 y_2
\end{aligned} \tag{4.15}$$

□

The result in Equation 4.13 simplifies the calculation of the Fisher information matrix for Kibble's bivariate mixture of exponential distributions immensely. It consists of a combination of numerical results already obtained in the last step of the EM algorithm. Lemma 4.1 and Lemma 4.2 can be used to implement the EM algorithm and evaluate the statistical soundness of the estimates when applied to simulated data. For this special case of the Kibble distribution Kotz et al (2000) list several moment

based estimators of which we used a selection for comparison in the paper in Appendix A. Furthermore, Chatelain et al (2007) have suggested a direct estimation method evaluating the first derivative of the density function using standard numerical methods. Their methods can also be applied for identical shape parameters not equal to one. The comparison delivers a repeating pattern where the estimates obtained via the EM algorithm equal the ones obtained through the method obtained by Chatelain et al (2007). This pattern is expected since both methods are used to calculate maximum likelihood estimates. The maximum likelihood estimates are clearly better suited than the moment-based estimates. A representative example can be found in table 4.1 including the estimated standard deviation σ_ρ for ρ which was calculated using the results from the EM algorithm. The correlation parameter used to generate the data is in the column headed ρ, the estimates calculated via the EM algorithm are in the column headed ρ_{EM}, and ρ_{Ch} is used to denote the estimates calculated via the approach by Chatelain et al (2007).

Table 4.1: Comparing different estimation methods for the case of 1000 data points, $\lambda_1 = 1$, $\lambda_2 = 10$

ρ	ρ_{EM}	ρ_{Ch}	σ_ρ
0.2	0.1827	0.1827	0.0061
0.4	0.3828	0.3828	0.0090
0.6	0.6113	0.6113	0.0084
0.8	0.8088	0.8088	0.0051

The clear advantage of using the EM algorithm, though it is computational more expensive, is the possibility to calculate the Fisher information matrix. If the MPH* representation is unique, the Fisher information matrix allows us to evaluate the soundness of our estimates.

Lemma 4.3. Uniqueness of representation for the Kibble distribution *The estimates obtained for Kibbles bivariate geometrical mixtures of exponentials via the EM algorithm are unique with the exception of permutations.*

Proof. Let $\simeq$ describe the identity in distribution of two MPH* distributions and assume

$$\begin{aligned}\boldsymbol{\alpha} &= (1,0),\ \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho\cdot\lambda_2 & -\lambda_2 \end{pmatrix},\\ \boldsymbol{\beta} &= (\beta_1,\beta_2),\ \mathbf{U} = \begin{pmatrix} -\mu_1 & r_1\cdot\mu_1 \\ r_2\cdot\mu_2 & -\mu_2 \end{pmatrix},\ \mathbf{R} = \mathbf{I}.\end{aligned}$$

To prove uniqueness, we compare the Laplace transform of the two representations:

$$\begin{aligned}
& MPH^\star(\boldsymbol{\beta},\mathbf{U},\mathbf{R}) \simeq MPH^\star(\alpha,\mathbf{T},\mathbf{R})\\
\iff\ & \forall\ s\ L_{\boldsymbol{\beta},\mathbf{U}}(s) = L_{\boldsymbol{\alpha},\mathbf{T}}(s)\\
\iff\ & \frac{\lambda_1\lambda_2(1-\rho)}{s_1s_2+s_2\lambda_1+s_1\lambda_2+\lambda_1\lambda_2(1-\rho)}\\
=\ & \frac{\beta_1s_2\mu_1(1-r_1)\mu_2+\beta_2s_1\mu_1\mu_2(1-r_2)+\mu_1\mu_2(1-r_1r_2)}{s_1s_2+s_2\mu_1+s_1\mu_2+\mu_1\mu_2(1-r_1r_2)}\\
\iff\ & \beta_1\mu_1\mu_2(1-r_1)s_1s_2^2+\beta_2\mu_1\mu_2(1-r_2)s_1^2s_2\\
+\ & \lambda_1\beta_1\mu_1\mu_2(1-r_1)s_2^2+\lambda_2\beta_2\mu_1\mu_2(1-r_2)s_1^2\\
+\ & (\lambda_1\beta_2\mu_1\mu_2(1-r_2)+\mu_1\mu_2(1-r_1r_2)+\beta_1\lambda_2\mu_1\mu_2(1-r_1)-\lambda_1\lambda_2(1-\rho))\,s_1s_2\\
+\ & (\lambda_1\mu_1\mu_2(1-r_1r_2)+\beta_1\lambda_1\lambda_2(1-\rho)\mu_1\mu_2(1-r_1)-\mu_1\lambda_1\lambda_2(1-\rho))\,s_2\\
+\ & (\lambda_2\mu_1\mu_2(1-r_1r_2)+\beta_2\lambda_1\lambda_2(1-\rho)\mu_1\mu_2(1-r_2)-\mu_2\lambda_1\lambda_2(1-\rho))\,s_1=0\\
\implies\ & \{\beta_1=0\ \vee\ r_1=1\}\ \wedge\ \{\beta_2=0\ \vee\ r_2=1\}\\
\implies\ & \mu_1=\lambda_1,\ \mu_2=\lambda_2,\ \{\beta_1=1,\ \beta_2=0,\ r_1=1,\ r_2=\rho\}\\
\vee\ & \{\beta_1=0,\ \beta_2=1,\ r_1=\rho,\ r_2=1\}.
\end{aligned}\tag{4.16}$$

Therefore the MPH* representation for Kibble's bivariate mixture of exponential distributions is unique, with the exception of permutations. □

The uniqueness of the representation is essential for further statistical analysis of the estimators and enables us to use the Fisher information matrix to

estimate the variance covariance matrix of the estimated parameters. The variance covariance matrix can for example be used to calculate the correlation between the intensity parameters and the correlation parameter. In our numerical example it is negative as expected, since an increase of ρ correlates to more visits to the different transient states before absorption of the underlying MJP and influences directly the accumulated sojourn times.

We have shown how to use an EM algorithm to calculate maximum likelihood estimates for a bivariate mixture of exponential distributions. Using the MPH* interpretation, it is possible to obtain more general results. The results presented in this section are very promising.

4.2 Bivariate mixtures of Erlang distributions

The results from Section 4.1 can be extended to the cases where the shape parameters of the Kibble distribution are not restricted to being equal to one. The method suggested by Chatelain et al (2007) also removes the restriction of shape parameters equal to one, but unlike our approach they continue to assume that the shape parameters are identical. With the following matrices

$$\begin{aligned} \mathbf{T} &= \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho\lambda_2 & -\lambda_2 \end{pmatrix}, \ \mathbf{T}^\star = \begin{pmatrix} 0 & 0 \\ (1-\rho)\lambda_2 & 0 \end{pmatrix}, \ \mathbf{t} = \begin{pmatrix} 0 \\ (1-\rho)\lambda_2 \end{pmatrix} \\ \mathbf{R} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \text{and } \mathbf{R}_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}, \end{aligned} \tag{4.17}$$

an MPH* representation can be given with

$$\mathbf{T}_E = \begin{pmatrix} \mathbf{T} & \mathbf{T}^\star & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{T} & \mathbf{T}^\star & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{T} & \mathbf{T}^\star & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{T} & \mathbf{t} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\lambda_2 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & 0 & -\lambda_2 & \lambda_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & \cdots & 0 & -\lambda_2 \end{pmatrix}, \; R_E = \begin{pmatrix} \mathbf{R} \\ \vdots \\ \vdots \\ \mathbf{R} \\ \mathbf{R}_2 \\ \vdots \\ \vdots \\ \mathbf{R}_2 \end{pmatrix}.$$

assuming that $\mathbf{T}_E$ contains q $\mathbf{T}$ matrices and l additional states where the sojourn times are $exp(\lambda_2)$ distributed then the underlying MJP visits the states $\{1, \ldots, 2q, \ldots, 2q + l\}$ prior to absorption. The initial distribution is $\boldsymbol{\alpha} = (1, 0, \ldots, 0)$ of dimension $2 \cdot q + l$. The probability density function can be written as:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \frac{\lambda_1\lambda_2 e^{-\lambda_1 y1 - \lambda_2 y_2}}{(q-1)!}(1-\rho)^q \frac{(\lambda_1 y_1 \lambda_2 y_2)^{q-1}(\lambda_2 y_2)^l}{(\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2})^{q+l-1}} \\ &\cdot \; I_{q+l-1}(2\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2}). \end{aligned} \tag{4.18}$$

With $I_q(z) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+q+1)} \left(\frac{z}{2}\right)^{2i+q}$ being the modified Bessel function of the first kind. For this general case, Lemma 4.1 can be generalized.

Theorem 4.4. Conditional expectation for Kibbles bivariate mixtures of Erlang distributions *For the Kibble distribution contained in the class of MPH* distributions with intensity parameters λ_1 and λ_2, and $q, l \in \mathbb{N}$ the conditional expectation of the number of visits N to the states with sojourn time distribution being $exp(\lambda_1)$, given the observed values y_1 and y_2 as well as the current estimates for λ_1, λ_2, and ρ, is*

$$E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2] = \sqrt{\rho\lambda_1\lambda_2 y_1 y_2} \cdot \frac{I_{l+q}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{l+q-1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})} + q \tag{4.19}$$

Proof. The proof is similar to the proof of Lemma 4.1.

$$
\begin{aligned}
E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2] &= \sum_{n=q}^{\infty} nP(N=n|y_1,y_2) \\
&= \frac{\sum_{n=q}^{\infty} n\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}e^{-\lambda_1 y_1}\lambda_2\frac{(\lambda_2 y_2)^{n+l-1}}{(n+l-1)!}e^{-\lambda_2 y_2}}{\sum_{n=q}^{\infty}\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}e^{-\lambda_1 y_1}\lambda_2\frac{(\lambda_2 y_2)^{n+l-1}}{(n+l-1)!}e^{-\lambda_2 y_2}} \\
&= \frac{\sum_{k=0}^{\infty}\frac{k+q}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+l+q-1)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+l+q-1)!}} \\
&= \frac{\sum_{k=1}^{\infty}\frac{1}{(k-1)!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}} + q \\
&= \frac{\sqrt{\rho\lambda_1\lambda_2 y_1 y_2}\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})^{2k+l+q}}{(k+l+q)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})^{2k+l+q-1}}{(k+l+q-1)!}} + q \\
&= \sqrt{\rho\lambda_1\lambda_2 y_1 y_2}\cdot\frac{I_{l+q}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{l+q-1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})} + q \qquad (4.20)
\end{aligned}
$$

□

Furthermore Lemma 4.2 can also be generalized to the general case.

Theorem 4.5. Conditional second moment for Kibbles bivariate mixtures of Erlang distributions *For the Kibble distribution described in Theorem 4.4, the conditional second moment of the accumulated number of visits N to the states where the sojourn time follows an $exp(\lambda_1)$ distribution, given the observed values y_1 and y_2 as well as the current estimates for λ_1, λ_2, and ρ, is*

$$
\begin{aligned}
E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N^2|y_1,y_2] &= \rho\lambda_1\lambda_2 y_1 y_2\cdot\frac{I_{q+l+1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{q+l-1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})} \\
&+ (2q+1)\left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1,y_2]\right) + q(q-1). \qquad (4.21)
\end{aligned}
$$

Proof. The proof is similar to the proof of Lemma 4.2

$$
\begin{aligned}
& E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N^2|y_1.y_2] = \sum_{n=q} P(n^2 = q|y_1, y_2) \\
= & \frac{\sum_{n=q}^{\infty} n^2 \binom{n-1}{q-1} \rho^{n-q}(1-\rho)^q \lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n+l-1}}{(n+l-1)!} e^{-\lambda_2 y_2}}{\sum_{n=q}^{\infty} \binom{n-1}{q-1} \rho^{n-q}(1-\rho)^q \lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n+l-1}}{(n+l-1)!} e^{-\lambda_2 y_2}} \\
= & \frac{\sum_{k=0}^{\infty} \frac{k^2}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+l+q-1}}{(k+l+q-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+l+q-1}}{(k+l+q-1)!}} + 2q \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q^2 \\
= & \frac{\sum_{k=1}^{\infty} \frac{k}{(k-1)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}} + 2q \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q^2 \\
= & \frac{\sum_{s=0}^{\infty} \frac{s+1}{(s)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+1}}{(s+l+q)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}} + 2q \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q^2 \\
= & \frac{\sum_{s=0}^{\infty} \frac{s}{(s)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+1}}{(s+l+q)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}} + \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2] - q\right) \\
+ & 2q \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q^2 \\
= & \frac{\sum_{j=0}^{\infty} \frac{1}{(j)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{j+2}}{(j+l+q+1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+l+q-1)!}} + \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2] - q\right) \\
+ & 2q \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q^2 \\
= & \rho\lambda_1\lambda_2 y_1 y_2 \cdot \frac{I_{q+l+1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{q+l-1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})} \\
+ & (2q+1) \left(E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1, y_2]\right) + q(q-1) \qquad (4.22)
\end{aligned}
$$

□

With $E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N^2|y_1.y_2]$ and $E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}[N|y_1.y_2]$ the Fisher information matrix

for the estimates obtained through the EM algorithm can be calculated as

$$\mathbf{I}(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\rho}, y) = E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}\left[\begin{pmatrix} \frac{N}{\lambda_1^2} & 0 & 0 \\ 0 & \frac{N+l}{\lambda_2^2} & 0 \\ 0 & 0 & \frac{N-q}{\rho^2} + \frac{q}{(1-\rho)^2} \end{pmatrix} \middle| y \right]$$
$$+ \quad E_{\boldsymbol{\alpha},\mathbf{T},\mathbf{R}}\left[\begin{pmatrix} -y_1 + \frac{N}{\lambda_1} \\ -y_2 + \frac{N+l}{\lambda_2} \\ \frac{N-q}{\rho} - \frac{q}{(1-\rho)} \end{pmatrix} \cdot \begin{pmatrix} -y_1 + \frac{N}{\lambda_1} & -y_2 + \frac{N+l}{\lambda_2} & \frac{N-q}{\rho} - \frac{q}{(1-\rho)} \end{pmatrix} \middle| y \right].$$

These results have been used to implement an EM algorithm with input parameters λ_1, λ_2, ρ, l, q, and w that simulates w bivariate observations and then runs the EM algorithm in order to obtain maximum likelihood estimates of the parameters λ_1, λ_2 and ρ. The results are in terms of accuracy and computational time similar to the case with $q = 1$ and $l = 0$.

Kibbles bivariate mixtures of Erlang distributions are closed under addition, if the intensity parameters are identical, and they are in the broader sense a subclass of the class of MPH* distributions. Theorem 4.4 and Theorem 4.5 enable the use of the EM algorithm to estimate parameters for this family of MPH* distributions as well as to calculate statistical inference on the estimated parameters. With this we have presented a first step towards general estimation methods and inference theory in the class of MPH* distributions. The next chapter will be used to give an example of the general applicability of MPH* distributions.

CHAPTER 5

Using phase type distributions in project management based on the Successive Principle

The Successive Principle (SP) is a group estimation procedure that is primarily used in Scandinavia and Great Britain. It was developed at the University of Denmark by Glahn and Lichtenberg (1984) and is comprehensively summarized and explained in the book "Successive Principle: A quality assurance technique for Schedules and Budgets" (Lichtenberg, 2000). It has similarities with the Program Evaluation and Review Technique developed by the United States Navy (Fazar, 1959), but its analyzing method has several specific characteristics, e.g. the use of subjective probabilities as well as the assumption that the duration or the cost of a subtask follows an Erlang distribution. The argument for choosing Erlang distributions is that they are not symmetric and only have support on the positive real axis. The SP consists of two parts; first a group analysis discusses, and evaluates a project establishing estimates using subjective probabilities; second

the subjective estimates are used to create a mathematical model for the project under consideration.

Section 5.1 outlines the procedure suggested by Lichtenberg (2000) to obtain unbiased group estimates. Section 5.2 discusses some of the mathematical assumptions behind the SP. Finally, in Section 5.3 we will use some of these assumptions and suggest a new way for dealing with lead times and show how to use the closure properties of PH distributions to model the duration of an entire project, resulting in a general PH distribution with representation $(\boldsymbol{\alpha}_D, \mathbf{T}_D)$. Given the PH representation we then construct an MPH* model to capture the correlation between the duration and the cost of a project. Analogously, it is possible to extend the model to more than two variables.

5.1 The general concept of the SP

The different aspects of the group estimation part of the SP have been shortly summarized by Lichtenberg (2006). The focus of our research is on the mathematical perspective. However, it is of importance to understand how estimates can be obtained through the SP. Therefore we only state a simplified version of the group estimation phase used in the SP, e.g we will not elaborate on the group psychology behind the analyzing estimation. The center of a group analysis is the persons participating. Therefore, a group analysis always begins with establishing a proper analysis group. The SP suggests can be used with groups of different size. Recommendable are groups consisting of four to twenty-five people. Lichtenberg (2000) states that the analysis process looses its drive when the group consists of more than ten people. It is also important that not too few people are chosen. The SP depends on estimates that cover a wide range, when having too few people in the analysis group this becomes unrealistic. As the subjective probabilities obtained from each member of the group are the basis for everything to follow, the group analysis is an essential step.

The idea of group analysis is not new. Hill (1982) discussed the advantages

and disadvantages of group analysis and collected several sources confirming the positive effect of knowledge sharing via group discussions. Ideally, for the SP, the group consists of young and old, female and male, optimistic and pessimistic as well as experts and amateurs. The right mixture of participants is one key element. Then, the project under consideration is discussed and potential general uncertainties are evaluated. This is followed by establishing a break down structure and by identifying its most uncertain elements. This often results in the identification of physical subtasks of the project at hand, but might also construct artificial tasks. Not all elements of a project affect the cost or duration equally. Often this effect mirrors Pareto's Principle (Humphreys, 2005) where the overall uncertainty of a project is due to a minority of its subtasks. The idea of using a break down structure to identify and rank the most uncertain events has also been used by Hillson (2003) but in a different setting than the SP. Further successive detailing and analyzing of the elements of a project is assumed to result in independent items or subtasks as well as subjective triple estimates, estimates for the most likely value , minimum value and the maximum value, by each member of the analysis group. The individual estimates are then used to produce group estimates. For the most likely group value (*mode*), the group average of the most likely values is calculated. For the minimum and maximum values, the groups minimum (*min*) and maximum values (*max*) are taken. These group estimates are then used to calculate the mean value and the standard deviation for each of the critical items. To simplify the future text we will use the term subtasks regardless of the item under consideration being an actual subtask or not, e.g. motivation. The motivation of the employees during a project can influence the duration severely. This can therefore either be attributed to the individual subtask at hand or summarized as a generic task in itself.

5.2 Mathematical assumptions of the SP

The SP assumes that the duration or the cost of a subtask follows an Erlang seven $Er(7, \lambda)$ distribution. Example 2.1 shows how to construct an Erlang

two distribution with intensity parameter λ as the sum of two identical, independent, and exponentially distributed random variables. A PH representation for an Erlang seven distribution can be given with $\boldsymbol{\alpha} = (1, 0, 0, 0, 0, 0, 0)$ and

$$\mathbf{T} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & -\lambda & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & -\lambda & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & \lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & \lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & -\lambda & \lambda \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda \end{pmatrix}. \tag{5.1}$$

The estimates obtained through the procedure explained in Section 5.1 are handled as if they were the mode, the 1% quantile, and the 99% quantile for each subtask. Assuming the distribution which generates these estimates, the local mean (m) and the local standard deviations (s) of these Erlang seven distributions can be calculated as (Lichtenberg, 2000)

$$\begin{aligned} m &= \frac{min + 2.9 \times mode + max}{4.9} \\ s &= \frac{max - min}{4.65}. \end{aligned} \tag{5.2}$$

Let m_i be the expected value for subtask i and s_i be the standard deviation of subtask i. First we consider the case of cost estimates. We assume the group analysis has identified n independent subtasks which have the largest influence on the overall uncertainty of the projects cost. It is assumed these subtasks are independent of each other. If two subtask are depending on each other, they are further investigated and split in smaller subtasks until independence can be achieved. The total expected cost C can be calculated as:

$$C = \sum_{i=1}^{n} m_i. \tag{5.3}$$

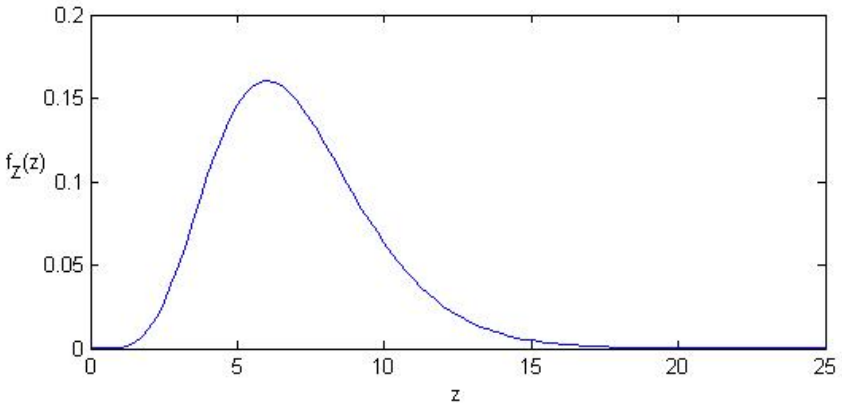


Figure 5.1: PDF of an Er(7,1) distribution

Furthermore, the project standard deviation S can be expressed as:

$$S = \sqrt{\sum_{i=1}^{n} s_i^2}. \tag{5.4}$$

These values are expressed using general results for sums of independent random variables. The SP then assumes that the total cost of a project is normally distributed with mean C and standard deviation S. The justification of this procedure is based on the central limit theorem. Even though Erlang distributions are skewed (see Figure 5.1), a sum of Erlang distributed random variables will converge towards the normal distribution. With increasing shape parameter an Erlang distribution will loose its skewness. In Chapter 2 we have shown that the sum of Erlang distributed random variables follows a generalized Erlang distribution. This means that the distribution of the total cost follows a generalized Erlang distribution where the state space of the underlying MJP is of dimension $n \times 7$. Therefore, the distribution of the total cost of a project will converge reasonable fast to a normal distribution.

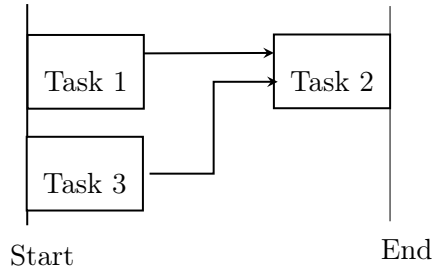


Figure 5.2: A subtask with two predecessors

Special care has to be taken when estimating the total duration of a project. Generally, the SP only considers the subtasks that lie on the time critical, and one or two near critical, paths. The subtasks on these paths either run parallel or successively and interact with each other, i.e. depend on the completion or partial completion of one or more predecessors. These interactions do not influence the total cost of a project, but do make it more complex to create a proper mathematical model for the duration of the project. The procedure for obtaining the triple estimates, the estimates for min, $mode$, and max, in the first step of the SP is the same when modeling the durations of a project as when modeling the cost of a project. The structure for the case that a subtask needs two predecessors to be completed before it can be started can be seen Figure 5.2. A simple approach is to assume deterministic durations for each subtask. The (expected) total duration of this project can then be calculated by taking the maximum of the (expected) durations of subtask 1 and subtask 3 and add the (expected) duration of task 2. It is a rare occasion in real life that a subtask takes a fixed amount of time. If the duration of each subtask is random and not deterministic, this simple approach is naive. This fact is only strengthened by the complexity of external and internal factors affecting the duration of a subtask. This is one reason why the SP assumes the duration of a subtask is random and making it possible that the subtask which has the shorter expected duration takes longer to complete than the subtask with the longer

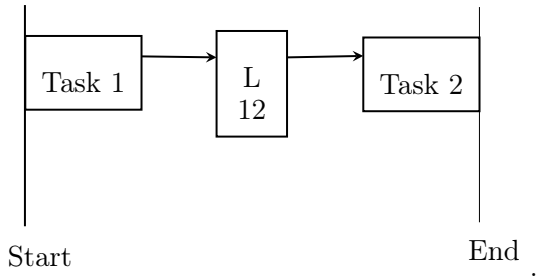


Figure 5.3: Lead time between subtask 1 and subtask 2

expected duration. In order to highlight this phenomena, let us abstract from the idea of subtasks and consider two independent random variables D_1 and D_3. Without loss of generality, assume $E[D_1] < E[D_3]$ then

$$\begin{aligned} P(D_3 < E[D_3]) &\geq P(D_1 < E[D_3], D_3 < E[D_3]) \\ &= P(D_1 < E[D_3])P(D_3 < E[D_3]). \end{aligned} \tag{5.5}$$

This is generally true, if the support of the distributions of the random variables D_1 and D_3 overlaps. It represents the fact that it is less likely for two subtask to finish on time than it is for one subtask. When estimating durations of projects, the inequality in Equation 5.5 needs to be accounted for. This can be done by introducing a merge event bias (MEB) increasing the total duration of a project. The MEB can either be calculated directly (Gong and Rowings, 1995) or approximated. The latter is the standard approach in the SP.

The second challenge when estimating the duration of a project, is the existence of lead times. Lead times are generally understood as the time a subtask can start prior to completion of a predecessor. Figure 5.3 expresses the lead time graphically as an independent subtask with a random duration. Solutions for dealing with the MEB exist. However, it can be solved more elegantly when the duration of the subtasks follow PH distributions. When facing randomly distributed lead times with support on the entire

positive real axis it is necessary to expand the mathematical modeling used in the SP. We suggest the assumption that the durations of subtasks as well as of lead times follow Erlang distributions. We go even further and use only the fact that Erlang distributions belong to the class of PH distributions. We then model the duration of the entire project directly using a general PH distribution. A PH model has the advantage that it enables the same modeling procedure for projects with very few subtasks as for medium to large projects. When relying on the central limiting theorem the project at hand may not consist of too few subtasks.

5.3 Modeling structures of subtasks arising in projects

Our research focused primarily on improving the mathematical modeling of the duration of a project. In 2000, Lichtenberg and Partners conducted an analysis using the SP for an large international IT development project. Their results are based on the traditional calculations used in the SP. The focus of the analysis was on final delivery and they identified several problems as well as solutions. The group consisted of 12 members that spend two days analyzing the project. Steen Lichtenberg was so kind to supply us with the original estimates as well as the structure of the project. The critical and near critical paths determined through the group analysis step of the SP can be seen in Figure 5.4. Our ultimate goal was to use the closure properties of PH distributions to create a PH model for the total duration of the project and compare the results to the original estimates.

As argued in Section 5.1 two complications arise when modeling the total duration of a project, to model the time until completion of several possible predecessors as well as possible lead times.

The problem of having two or more predecessors for a single subtask can easily be solved using the closure properties of the class of PH distributions. Let us assume a project, as described in Figure 5.2, where the final subtask can not start before its two predecessors are completed. The duration of

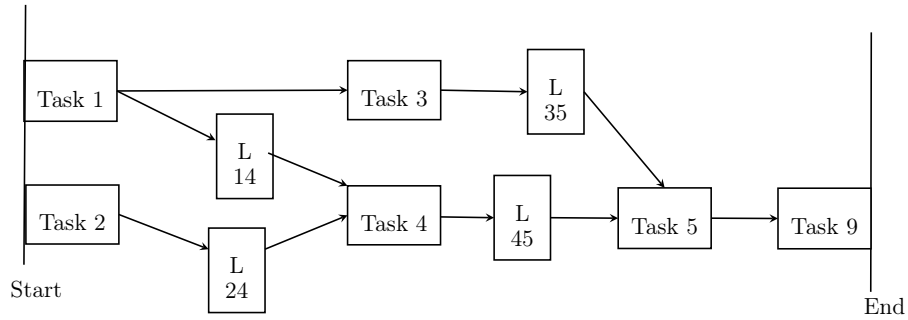


Figure 5.4: Subtasks on the critical and near-critical paths for a software project

the two predecessors can be modeled using the maximum of two PH distributed random variables. We know from Theorem 2.1 that it is again PH distributed. The maximum of the two predecessors is then added to the duration of the final subtask in order to calculate the total duration of the project. This is an addition of two PH distributed random variables, resulting again in a PH distributed random variable (Theorem 2.4). Hence, the duration of this project is PH distributed.

We define D as the time from start until the end of a project, i.e. the total duration of the project. The duration of subtask i is defined as D_i and its distribution can be described using the PH representation $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$. If we consider a project consisting of three subtask as displayed in Figure 5.2 D can easily be calculated as

$$D = \max\{D_1, D_3\} + D_2. \tag{5.6}$$

The distribution of D can be given by its PH representation $(\boldsymbol{\alpha}_D, \mathbf{T}_D)$ with

$$\mathbf{T}_D = \begin{pmatrix} \mathbf{T}_1 \otimes \mathbf{I}_3 + \mathbf{I}_1 \otimes \mathbf{T}_3 & \mathbf{I}_1 \otimes \mathbf{t}_3 & \mathbf{t}_1 \otimes \mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_1 & \mathbf{0} & \mathbf{t}_1 \cdot \alpha_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_3 & \boldsymbol{t}_3 \cdot \alpha_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{T}_2 \end{pmatrix}$$

and

$$\boldsymbol{\alpha}_D = (\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_3, 0, \ldots, 0).$$

The second and more challenging complication when making a stochastic model for the duration of a project is the occurrence of lead times Lij between subtask i and j. We will denote the duration of lead time Lij as L_{ij}. We assume that $L_{ij} \geq 0$, generally speaking lead times can reduce the total duration of a project. This can be achieved by subtracting the lead times in the appropriate places during the calculation of the total duration. If a project has a certain delay with respect to the completion of one or more predecessors it can be modeled using the addition of two PH random variables as shown in Theorem 2.1. Since lead times enable the initiation of a subtask prior to completion of its predecessor, but not prior to the initiation of the predecessor it is assumed that

$$E[L_{ij}] < E[D_i]. \tag{5.7}$$

This is a reasonable approach if the lead times and duration of the subtasks are deterministic or their support is only on a subset of the real numbers. Unlike the triangular distribution or other distributions commonly used to model the duration of subtasks and lead times, PH distributions have support on the whole real axis and therefore it follows that

$$\begin{aligned} P(D_i < L_{ij}) &> 0, \\ P(D_j < L_{ij}) &> 0, \text{ and} \\ P(D_i + D_j - L_{ij} < 0) &> 0. \end{aligned} \tag{5.8}$$

The expression $D_i + D_j - L_{ij}$ can be expressed as a bilateral PH distribution (Ahn and Ramaswami, 2005). The possibility of negative duration of two

subtask caused by the lead time as in Equation 5.8 is from a mathematical perspective unproblematic, however from a project management perspective troubling and unrealistic. Our solution for solving the challenge of lead times is to redefine lead times.

Definition 5.1. *Alternative definition of lead times* The alternative lead time is defined as the time a subtask j is running parallel to its predecessor i and its duration is denoted with $\tilde{L}_{ij}$.

The motivation for this definition is two-fold. Firstly, in our opinion it seems unnatural to model a lead time as the time a successor can start prior to the completion time of its predecessor. It is not the duration until the completion of the predecessor that is crucial for a possible start, but the level of completion of the predecessor. With Definition 5.1 the start of a successor of a subtask depends on how long it has been running and therefore which level of completion has been archived. Secondly, it allows us to model the duration of a project using only maximum operation as well as addition of independent and PH distributed random variables. For the project described in Figure 5.4 the new structure using alternative lead times can be found in Figure 5.5. In this case $\widetilde{\text{Task}}$ 1 is not the same subtask as Task 1 but only the part that had to be completed prior to the start of subtask 4. The duration of a the subtask $\widetilde{\text{Task}}$ i is denoted with $\tilde{D}_i$. The new local means can be calculated using the result from the classical analysis, as done in Appendix B. If done so, one should ensure that the expected duration until total completion of subtask i is still the same as in the original analysis. Alternatively, or preferably, they can be derived directly during the group estimation procedure of the SP. To demonstrate the general concept of creating a PH model using alternative lead times we consider a project that follows the structure presented in Figure 5.3.

Example 5.1. The project represented in Figure 5.3 can be restructured using alternative lead times as in Figure 5.6, with $E[\tilde{D}_1] = E[D_1] - E[L_{12}]$, $E[\tilde{L}_{12}] = E[L_{12}]$, and $E[\tilde{D}_2] = E[D_2]$. Let the PH representation for the duration of the subtasks $\tilde{D}_i$ and the alternative lead time $\tilde{L}_{12}$ be $(\tilde{\boldsymbol{\alpha}}_i, \tilde{\mathbf{T}}_i)$

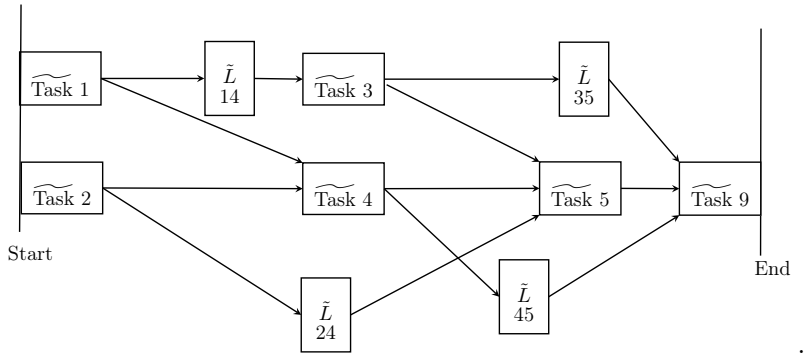


Figure 5.5: Software project from Figure 5.4, restructured by the use of alternative lead times

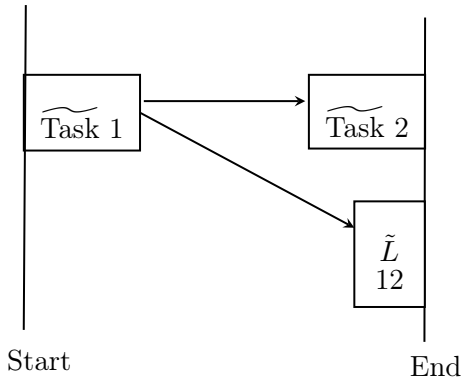


Figure 5.6: Alternative lead time between subtask 1 and 2

and $(\tilde{\boldsymbol{\alpha}}_{12}, \tilde{\mathbf{T}}_{12})$ respectively. The PH distribution of the total duration D can be represented with $(\boldsymbol{\alpha}_D, \mathbf{T}_D)$ where

$$\mathbf{T}_D = \begin{pmatrix} \tilde{\mathbf{T}}_1 & \tilde{\boldsymbol{t}}_1 \cdot (\tilde{\boldsymbol{\alpha}}_2 \otimes \tilde{\boldsymbol{\alpha}}_{12}) & 0 & 0 \\ 0 & \tilde{\mathbf{T}}_2 \otimes \mathbf{I}_{12} + \mathbf{I}_2 \otimes \tilde{\mathbf{T}}_{12} & \mathbf{I}_2 \otimes \tilde{\mathbf{t}}_{12} & \tilde{\mathbf{t}}_2 \otimes \mathbf{I}_{12} \\ 0 & 0 & \tilde{\mathbf{T}}_2 & 0 \\ 0 & 0 & 0 & \tilde{\mathbf{T}}_{12} \end{pmatrix} \tag{5.9}$$

and $\boldsymbol{\alpha}_D = (\tilde{\boldsymbol{\alpha}}_1, 0, \ldots, 0)$ of appropriate dimension. The expected duration in this example is the same as the expected duration using the traditional lead time. Our model ensures that subtask 1 as well as subtask 2 have to be completed in order to determine the total duration of the project. This can be expressed as:

$$D = \tilde{D}_1 + \max\left\{\tilde{D}_2, \tilde{L}_{12}\right\}, \tag{5.10}$$

and can capture the case that subtask 1 proceeds normally until the start of subtask 2 and then experiences a severe delay.

If the duration of the subtasks and the lead times are Erlang 7 distributed, already the state space of this small example grows to a dimension of 70 and with that the transition matrix is 70×70. It is problematic to state the transition matrix for a complete project as in Figure 5.5, where the state space is of dimension 3353. Even when reusing the sub generator matrices of the subtasks and lead times, the transition matrix becomes too big to be printed on paper in a readable manner. However, it is not a problem to deal with these matrices numerically and we can calculate all distributional properties of the PH distributed random variable D, e.g. the mean, quantiles and the probability density function. The modeling steps only use the assumption of PH distributed durations for the subtasks as well as for the lead times. In the SP these durations are $Er(7, \lambda)$ distributed and their intensity parameter λ can be calculated using the local mean m obtained through the SP,

$$\lambda = \frac{7}{m}. \tag{5.11}$$

Having the PH representation for D, the distribution of the time until completion as well as all moments and quantiles can easily be calculated. One interesting result for the software project under consideration (as presented in Figure 5.5) is the probability of the total duration of the project being longer than expected

$$P(D > E[D]) = 0.4727. \tag{5.12}$$

If D was normally distributed the result in Equation 5.12 should be 0.5. This result therefore speaks against the assumption that the total duration of a project is normal distributed. Assuming different PH distributions, distributions which are more skewed than Erlang 7 distribution, the result will be even more extreme. Aside from the distributional properties, the modeling also results in the sub generator matrix T_D and initial distribution α_D for $D \sim PH(\alpha_D, T_D)$ which can be used for further analysis.

5.4 Modeling correlation between duration and cost

Most projects are characterized by a strong dependents between their duration and their cost, may it be actual cost or loss in revenue. A review of the correlation between cost and duration for construction projects can be found in Ogunsemi and Jagboro (2006). Different factors, e.g. wages and cost of acquisitions, will always affect the cost. Some of them are time independent, some are directly connected to the duration of a subtask, and some are based on a general delay of the project. Not all factors contributing to the total cost of a project underlie the same variability. We have focused on the costs that are directly depending on the duration of the subtasks at hand.

We assume that the cost a subtask generates consists of a constant part and a part that grows linearly with the duration of the subtask, e.g. salary or the cost for renting equipment. We define the time dependent cost of subtask i as $K_i = c_i \cdot D_i$. With these assumptions and the result from

Section 5.2 it is possible to create a bivariate PH model contained in the class of MPH* distributions that we introduced in Chapter 3. Our model describes the correlation between time dependent cost of the subtasks on the critical and near-critical paths and the duration of a project. With proper parameters at hand, the extension to a bivariate model is only the beginning and the method described can be used to include other impact factors into the model, e.g the environmental impact of the project. These parameters could be estimated directly using the group estimation procedure during the group analysis of the SP, however, for the model it is irrelevant how the parameters are derived. Alternatively it is possible to estimate these parameters, for example via the EM algorithm, by use of data from former projects. However, collecting data for infrastructure projects is a challenging task, see Nicolaisen (2012) and Andersen (2013).

The next step in our modeling approach is to construct the reward matrix. The reward column vector for the total duration of the project consists of ones in every entry. The construction of the reward vector for the cost is not as straightforward. The entries depend on the subtasks running parallel and are therefore a sum of the prior estimated c_i. For Example 5.1 the MPH* representation can be given with

$$\mathbf{T}_D = \begin{pmatrix} \tilde{\mathbf{T}}_1 & \tilde{\mathbf{t}}_1 \cdot (\tilde{\boldsymbol{\alpha}}_2 \otimes \tilde{\boldsymbol{\alpha}}_{12}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{T}}_2 \otimes \mathbf{I}_{12} + \mathbf{I}_2 \otimes \tilde{\mathbf{T}}_{12} & \mathbf{I}_2 \otimes \tilde{\mathbf{t}}_{12} & \tilde{\mathbf{t}}_2 \otimes \mathbf{I}_{12} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{T}}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{T}}_{12} \end{pmatrix},$$

$$\mathbf{R}_{DC} = \begin{pmatrix} \mathbf{e}^T & c_1 \cdot \mathbf{e}^T \\ \mathbf{e}^T & (c_1 + c_2) \cdot \mathbf{e}^T \\ \mathbf{e}^T & c_2 \cdot \mathbf{e}^T \\ \mathbf{e}^T & c_1 \cdot \mathbf{e}^T \end{pmatrix}.$$

Here $\mathbf{e}^T$ is a column vector of appropriate dimension and can vary in dimension in order to fit the dimensions of the sub matrices in $\mathbf{T}_D$. Having the MPH* representation allows us to evaluate all moments and cross moments directly (see Theorem 3.2) as well as to obtain other probabilistic quantities

of the project through stochastic simulation. To our knowledge there exist no other method that allows project management to evaluate the risk of delays and cost overruns simultaneously,

CHAPTER 6

Conclusions

This chapter will be used to summarize the work done during my PhD study and to highlight the main contributions of my research. Finally, I will give some suggestions for further research.

This PhD project has been part of the UNITE project which has the overall goal to identify issues with forecasts for infrastructure projects and to improve the available forecasting methods. The general topic of my research was multivariate PH distributions. Only very little is known about the probabilistic properties of multivariate PH distributions. My research focused especially on the class of MPH* distributions, introduced in Chapter 3. The original idea of my PhD study was to derive general results and to apply these results to improve the mathematical modeling of infrastructure projects. During the research we realized that it was not feasible to produce general results for the entire class of MPH* distributions. Consequently the efforts shifted towards a more specific approach by focusing on the Kibble distribution.

The main contribution in regard to the Kibble distribution is the generalization of the EM algorithm, well known for parameter estimation for

univariate PH distributions, and applying it to a family of MPH* distributions. Our results are a first step towards general parameter estimation methods for the whole class of MPH* distributions. Aside from applying the EM algorithm to a family of MPH* distributions, we have developed a new method for parameter estimation for the Kibble distribution. The method exploits the fact that under certain conditions the Kibble distribution is contained in the class of MPH* distributions. Our approach enables us to derive expressions for the Fisher information matrix and with that allows for statistical inference. Our results are presented in Chapter 4 and in the paper in Appendix A. Prior estimation methods always assumed equal shape parameters, whereas our method overcomes this restriction.

Although our results bring better understanding of a family of MPH* distributions they are far from being generalizable towards the whole class of MPH* distributions. In order to obtain more general estimation methods, it is essential to derive general formulas for the probability density function for distributions contained in the class of MPH* distributions, or at least some of its subclasses. Furthermore, the EM algorithm and its implementations are often criticized for its computation times. Our approach is no exception of it, especially when compared to the other estimation methods at hand. Here, further work can be done, e.g. by optimizing the numerical implementation of our EM algorithm or using different parameters when initiating the EM algorithm.

Many continuous multivariate distributions are known. Several of them are multivariate PH distributions. However, this fact is often not exploited when using them in mathematical modeling. The Successive Principle, presented in Chapter 5, is quite unique with its assumption that the duration or cost of a subtask is Erlang distributed. When modeling the cost or duration of an entire project, this fact is neglected. We suggested a new modeling approach, based on the fact that the Erlang distribution belongs to the class of PH distributions, introduced in Chapter 2. The result is a model that describes the total duration of a project as one PH distribution. In order for the model to remain in the class of PH distributions, we defined an alternative lead time. Alternative lead times model the relation between

a subtask and its predecessor based on the completion of the predecessor. In our opinion alternative lead times are a more intuitive way of modeling the start of a subtask prior to the completion of its predecessor. Having a univariate PH model for the duration we extended the model to capture the correlation between duration and cost of a project. Our model is an example of an application and the versatility of MPH* distributions. The results are presented in Chapter 5 and Appendix B. Especially when modeling durations and costs of projects, both values restricted to be positive, it is an improvement over models based on the normal distribution.

We changed the modeling steps underlying the Successive Principle and applied our method to a rather simple project and already here the state space of the underlying Markov jump process grows fast. In order to make our method applicable it is essential to develop a tool that takes the structure of a project and translates it directly into a sub generator matrix for the entire project using only the maximum and the addition of PH distributed random variables. With the current state of our research we have derived a mathematical model that more accurately describes the interactions between the subtasks of a project. From a mathematical point of view this is a valuable result. From a practical point of view the improvement gained through our model is rather small compared to the overall uncertainties when dealing with subjective estimates. However, our modeling approach is not restricted to using subjective probabilities but only to using PH distributions. How the parameters of these PH distributions are estimated is irrelevant. Making our method, especially once a proper tool for deriving the sub generator matrix is developed, useful for decision makers.

My research in general has shed light on the class of MPH* distributions and will hopefully inspire further research towards a more general understanding of this versatile class of distributions.

Bibliography

Ahlström L, Olsson M, Nerman O (1999) A parametric estimation procedure for relapse time distributions. Lifetime Data Analysis 5(2):113–132

Ahn S, Ramaswami V (2005) Bilateral phase type distributions. Stochastic models 21(2-3):239–259

Andersen J (2013) The Shaping of Transport Model-Based Knowledge Production. Institut for Samfundsudvikling og Planlægning, Aalborg Universitet

Asmussen S (1992) Phase-type representations in random walk and queueing problems. The Annals of Probability pp 772–789

Asmussen S, Nerman O, Olsson M (1996) Fitting phase-type distributions via the em algorithm. Scandinavian Journal of Statistics pp 419–441

Assaf D, Levikson B (1982) Closure of phase type distributions under operations arising in reliability theory. The Annals of Probability pp 265–269

Assaf D, Langberg NA, Savits TH, Shaked M (1984) Multivariate phase-type distributions. Operations Research 32(3):688–702

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics 41(1):164–171

Black F, Scholes M (1973) The pricing of options and corporate liabilities. The journal of political economy pp 637–654

Bladt M (2005) A review on phase-type distributions and their use in risk theory. Astin Bulletin 35(1):145–167

Bladt M, Nielsen BF (2010) Multivariate matrix-exponential distributions. Stochastic models 26(1):1–26

Bladt M, Esparza LJR, Nielsen BF (2011) Fisher information and statistical inference for phase-type distributions. Journal of Applied Probability 48:277–293

Bladt M, Esparza LJR, Nielsen BF (2013) Bilateral matrix-exponential distributions. In: Matrix-Analytic Methods in Stochastic Models, Springer, pp 41–56

Chatelain F, Tourneret JY, Ferrari A, Inglada J (2007) Bivariate gamma distributions for image registration and change detection. IEEE Trans Image Processing 16(7):1796–1806

Cox DR (1955) A use of complex probabilities in the theory of stochastic processes. In: Proc. Camb. Phil. Soc, Cambridge Univ Press, vol 51, pp 313–319

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Series B (Methodological) pp 1–38

Erlang A (1920) Telefon-ventetider. et stykke sandsynlighedsregning. Matematisk Tidsskrift B 31:25–42

Esparza LJR (2011) Maximum likelihood estimation of phase-type distributions. IMM-PHD-2010-245, Technical University of Denmark (DTU)

Fackrell M (2005) Fitting with matrix-exponential distributions. Stochastic Models 21(2-3):377–400

Fazar W (1959) Program evaluation and review technique. The American Statistician 13(2):10

Flyvbjerg B, Bruzelius N, Rothengatter W (2003a) Megaprojects and risk: An anatomy of ambition. Cambridge University Press

Flyvbjerg B, Skamris Holm MK, Buhl SL (2003b) How common and how large are cost overruns in transport infrastructure projects? Transport Reviews 23(1):71–88

Glahn JE, Lichtenberg S (1984) Successive principle: A quality assurance technique for schedules and budgets. Organizing and Managing Construction, CIB W-65 1984, Proceedings - 4th International Symposium on Organization and Management of Construction Volume 2: Computers, Productivity pp 557–567

Gong D, Rowings J (1995) Calculation of safe float use in risk-analysis-oriented network scheduling. International Journal of Project Management 13(3):187–194, DOI 10.1016/0263-7863(94)00004-V

Grimsey D, Lewis MK (2002) Evaluating the risks of public private partnerships for infrastructure projects. International Journal of Project Management 20(2):107–118

Hill GW (1982) Group versus individual performance: Are n+1 heads better than one? Psychological Bulletin 91(3):517

Hillson D (2003) Using a risk breakdown structure in project management. Journal of Facilities Management 2(1):85–97

Humphreys KK (2005) Conducting project risk analysis, how to do it and how not to do it. ICEC International Roundup 19(2):3–13

Iliopoulos G, Karlis D, Ntzoufras I (2005) Bayesian estimation in kibble's bivariate gamma distribution. Canadian Journal of Statistics 33(4):571–589

Jørgensen M, Moløkken-Østvold K (2006) How large are software cost overruns? a review of the 1994 chaos report. Information and Software Technology 48(4):297–301

Kibble WF (1941) A two-variate gamma type distribution. The Indian Journal of Statistics 5(2):137–150

Kotz S, Balakrishnan N, Johnson NL (2000) Continuous Multivariate Distributions, 2nd edn. Wiley-Interscience, New York

Kulkarni VG (1989) A new class of multivariate phase type distributions. Operations Research 37(1):151

Latouche G, Ramaswami V (1999) Introduction to matrix analytic methods in stochastic modeling, vol 5. Siam

Lichtenberg S (2000) Proactive Management of Uncertainty using the Successive Principle, 1st edn. PF Forlag, Lyngby, Denmark

Lichtenberg S (2006) The successive principle–a scientific crystal ball for management. The IPSI BgD Transactions on Internet Research 2(1):33–39

Lin YJ, Lio Y, Ng HKT (2013) Bayes estimation of moran–downton bivariate exponential distribution based on censored samples. Journal of Statistical Computation and Simulation 83(5):837–852

Lovallo D, Kahneman D (2003) Delusions of success. Harvard business review 81(7):56–63

McLachlan G, Krishnan T (2007) The EM algorithm and extensions, vol 382. John Wiley & Sons

Moler C, Van Loan C (1978) Nineteen dubious ways to compute the exponential of a matrix. SIAM review 20(4):801–836

Moler C, Van Loan C (2003) Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM review 45(1):3–49

Neuts MF (1975) Probability distrubutions of phase type. In: Liber Amicorum Professor Emeritus H. Florin, Department of Mathematics, University of Louvian, Belgium, pp 173–206

Nicolaisen M (2012) Forecasts: Fact or Fiction? Department of Develpment and Planning, Aalborg University

Nielsen BF, Nielson F, Nielson HR (2010) Model checking multivariate state rewards. In: Quantitative Evaluation of Systems (QEST), 2010 Seventh International Conference on the, IEEE, pp 7–16

Oakes D (1999) Direct calculation of the information matrix via the em algorithm. Journal of the Royal Statistical Society Series B (Statistical Methodology) 61:479–482

O'Cinneide CA (1989) On non-uniqueness of representations of phase-type distributions. Stochastic Models 5(2):247–259

O'Cinneide CA (1991) Phase-type distributions and invariant polytopes. Advances in Applied Probability 23(3):515–535

Ogunsemi D, Jagboro G (2006) Time-cost model for building projects in nigeria. Construction Management and Economics 24(3):253–258

Qin S (2011) Partial differential equations related to multivariate phase type distributions. Master's thesis, Technical University of Denmark, DTU Informatics, E-mail: reception@imm.dtu.dk, Asmussens Alle, Building 305, DK-2800 Kgs. Lyngby, Denmark, URL http://www.imm.dtu.dk/English.aspx, supervised by Associate Professor Bo Friis Nielsen, bfn@imm.dtu.dk, DTU Informatics, and David Meisch

Wachs M (1990) Ethics and advocacy in forecasting for public policy. Business & professional ethics journal pp 141–157

Wang CP, Gosh M (2000) Bayesian analysis of bivariate competing risk models. Indian Journal of Statistic 62:388–401

Williams T (2003) Assessing extension of time delays on major projects. International Journal of Project Management 21(1):19–26

Yue S, Ouarda T, Bobee B (2001) A review of bivariate gamma distributions for hydrological application. Journal of Hydrology 246(1):1–18

Appendices

APPENDIX A

Paper 1

Estimation of the Kibble Distribution using the EM algorithm

Authors:
Mogens Bladt, David Meisch, Bo Friis Nielsen

Submitted to:
"Journal of Stochastic Modeling"

ESTIMATION OF THE KIBBLE DISTRIBUTION USING THE EXPECTATION-MAXIMIZATION ALGORITHM

MOGENS BLADT, DAVID MEISCH, AND BO FRIIS NIELSEN

Abstract. In this paper we use the expectation-maximisation (EM) algorithm to estimate the parameters of Kibble's bivariate Gamma distribution for the case where the shape parameters are integer valued. In this case, the Kibble distribution can be interpreted as a special case of a multivariate phase–type distribution allowing us to identify the estimation problem as a case of incomplete data. The EM algorithm ensures numerical stability and provides indirectly the information matrix which may be used for the calculation of confidence regions.

1. Introduction

Multivariate gamma distributions have widespread applications. Wang and Gosh (2000) [20] use them for competing risk models and Yue et al (2001) [21] show several applications in hydrology, for example flows in different parts of a river. Chatelain et al (2007) [7] use bivariate Gamma distributions to detect changes in radar images.

1991 *Mathematics Subject Classification.* Primary .

Key words and phrases.

The main focus in this paper will be on a specific bivariate Gamma distribution, namely the Kibble distribution. Kibble (1941) [11] calculated the joint distribution of the empirical variances in samples from a bivariate normal distribution. We will focus on the case where the shape parameters are integer valued. Gamma distributions with integer valued shape parameters are frequently referred to as Erlang distributions. The Kibble distribution becomes then a bivariate distribution where the marginal distributions are Erlang distributions and where the joint density is

$$\begin{aligned} f(y_1, y_2) &= \frac{\lambda_1\lambda_2 e^{-\lambda_1 y_1 - \lambda_2 y_2}}{(q-1)!}(1-\rho)^q \frac{(\lambda_1 y_1 \lambda_2 y_2)^{q-1}(\lambda_2 y_2)^l}{(\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2})^{q+l-1}} \\ &\quad \cdot I_{q+l-1}(2\sqrt{\rho\lambda_1 y_1 \lambda_2 y_2}), \end{aligned}$$

where $I_q(z) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+q+1)} \left(\frac{z}{2}\right)^{2i+q}$ is the modified Bessel function of the first kind. Several mathematical models have been developed and later proven to be special cases of the Kibble distribution. For example Downton (1970) [9] and Moran (1969) [14] analysed a special case of the bivariate exponential distribution. Here a successive damage model for two identical components is used, meaning that both components fail after having received the same number of geometrically distributed shocks. The shocks arrive according to two independent Poisson processes, one for each component. These special cases, known under

different names, e.g. as Moran and Downton's, Jensen's or Gaver's distribution can be found in Kotz, Balakrishnan and Johnson (2000) [19]. Bladt and Nielsen (2010) [6] show that the Kibble distribution is contained in the class of multivariate phase-type distribution (MPH) of the type MPH* introduced by Kulkarni (1989) [12].

The main contribution of our paper is to give a maximum likelihood-based alternative to previous primarily moment-based estimators for the Kibble distribution when the shape parameters are integer valued. By exploiting the probabilistic structure of MPH* we can allow for different shape parameters for the marginal distribution. The EM algorithm can be used to estimate all parameters as well as the Fisher information matrix.

Various, primarily moment-based, estimators for the correlation parameter in the Kibble distribution when the shape parameters are one, i.e. Kibble's bivariate exponential distribution, have been introduced. For details see Kotz, Balakrishnan and Johnson (2000) [19]. Chatelain et al [7] showed that the first derivative of the log likelihood function can be reduced to linear combinations of functions related to the confluent hyper geometric function. In order to calculate the root in $[0, 1[$ a Newton Rhapson procedure initiated by a standard correlation coefficient estimator is used. A Bayesian approach somewhat related to our

EM approach was presented in Iliopoulos et al (2005) [10] and modified to deal with censored data in Lin et al (2013) [13].

The structure of the paper is as follows. In Section 2, we give a short introduction to phase type (PH) and multivariate phase type (MPH) distributions. Then we give a review of alternative estimators when the shape parameters are one. In Section 4, we describe the EM algorithm for PH distributions and modify it, to be used for the Kibble distribution, and we present a way to calculate the Fisher information matrix. In Section 5, we focus on the case where the shape parameters are one, apply the EM algorithm to simulated data, and compare it to the estimation methods, introduced in Section 2, specially developed for Kibble's bivariate mixture of exponential distributions. The paper is concluded in Section 6.

2. Background

We will use the interpretation of the Kibble distribution as a multivariate phase type distribution of the type $MPH^\star$ to develop the estimation algorithm. We give a brief introduction to phase type distributions referring the reader to Neuts (1975) [16], Neuts (1981) [17], or Bladt (2005) [4] for further information. Let $\{X(t)\}_{t\geq 0}$ be a Markov jump process on a discrete state space $E = \{1, 2, \ldots, m, m+1\}$ with

state $m+1$ being an absorbing state and the states $1, \ldots, m$ being transient. We then define $\tau = \min\{t \geq 0 : \ X(t) = m+1\}$ as the time of absorption. The distribution of τ is phase-type as termed by Neuts [16]. With T being the sub-intensity matrix of $\{X(t)\}_{t\geq 0}$ corresponding to the m transient states and α correspondingly being the initial distribution among the transient states, we write $\tau \sim PH(\alpha, T)$. The pair (α, T) is called a representation of the phase type distribution. In general a representation for a phase type distribution is not unique.

Kulkarni (1989) [12] introduced MPH distributions using the occupation times $(H_1, \ldots, H_m)$ of the Markov jump process $X(t)$ in the states $\{1, \ldots, m\}$. The multivariate vector $Y = (Y_1, \ldots, Y_n)$ is defined using linear combinations of these occupation times

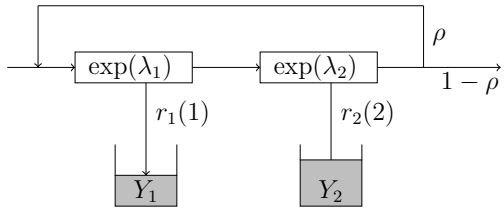
$$Y_j = \sum_{i=1}^{m} r_{ij} \cdot H_i = \int_0^{\tau} r_j\left(X(t)\right) dt.$$

The parameters $r_j(i) = r_{ij} \geq 0$ for $j \in \{1, \ldots, n\}$ and $i \in \{1, \ldots, m\}$ are the weights of the linear combinations alternatively interpreted as rates for rewards collected in each state visited by the underlying Markov jump process. The vector Y is said to be multivariate phase-type, $MPH^{\star}(\alpha, T, R)$ distributed, where $R = \{r_{ij}\}$ is called the reward matrix.

We will now demonstrate that the Kibble distribution with integer shape parameter can be expressed by a negative binomial mixture of gamma distributions with integer shape parameters. In its simplest form with $q = 1$ and $l = 0$ the Kibble distribution reduces to a bivariate exponential distribution. The MPH* representation for Kibble's bivariate exponential distribution is (α, T, R) with $\alpha = \begin{pmatrix} 1, 0 \end{pmatrix}$ and

$$T = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho\lambda_2 & -\lambda_2 \end{pmatrix}, \; R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This representation can be depicted graphically by



.

The interpretation of the figure is as follows. The system is entered from the left into the first box marked with $\exp(\lambda_1)$. This box represents a delay or sojourn time governed by an exponential distribution with intensity λ_1. Reward is accumulated continuously with a constant rate of $r_1(1)$ to the variable Y_1 during visits to this box. Upon leaving the box a similar box with exponential intensity λ_2 is entered. During visits to this box reward is accumulated to the variable Y_2 with rate $r_2(2)$. After the second box is left the first box is reentered with probability

ρ or the system is left with probability $1-\rho$. In the latter case reward accumulation is ceased. We can write

$$Y_i = r_{ii} \sum_{j=1}^{N} H_{ij},$$

where N is the geometrically distributed number of visits to each of the two states and H_{ij} is the duration of the jth visit to state i, $i = 1, 2$. The marginal distributions of the Y_i's are thus geometric mixtures of Erlang distributions well-known to be exponential. The dependence between the two variables is due to the feedback mechanism parameterised by ρ that ensures that Y_1 and Y_2 have the same number of exponential terms in the random sums. By serially connecting q identical copies of the diagram one obtains variables Y_1 and Y_2 that are Erlang distributed due to their nature as a sum of a negative binomially distributed number of exponential random variables. Finally, we could have additional exponential phases with no feedback mechanism ascribing reward to either Y_1 or Y_2. We will add ℓ phases with exponential rate λ_2 ascribing reward to Y_2 and none phases of this kind ascribing reward to Y_1. In this case we have an MPH* representation

with

$$T_E = \begin{pmatrix} T & T^\star & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & T & T^\star & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & T & T^\star & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & T^\star & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \dots & T & t^\star & \cdots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & \cdots & -\lambda_2 \end{pmatrix}, \; R_E = \begin{pmatrix} R \\ R \\ R \\ \vdots \\ R \\ R \\ R_2 \\ \vdots \\ R_2 \end{pmatrix}.$$

with

$$T^\star = \begin{pmatrix} 0 & 0 \\ (1-\rho)\lambda_2 & 0 \end{pmatrix}, t^\star = \begin{pmatrix} 0 \\ (1-\rho)\lambda_2 \end{pmatrix} \text{ and } R_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}.$$

Using the probabilistic structure given through the MPH* representation we can derive the density for Kibble's bivariate Gamma distribution. Let us assume there are q blocks with sub generator matrix T and an additional ℓ states where the sojourn times are $\exp(\lambda_2)$ distributed. The underlying jump process visits the states $\{1, \dots, 2q, \dots, 2q+\ell\}$ before it is being absorbed. Let N_i be the number of times state i is visited. The occupation time in state i conditioned on the number of visits is then Erlang distributed. If we define the random variable

$N = \sum_{i=1}^{q} N_{2i-1}$, as the number of times the states where the sojourn time is given by the parameter λ_1 are visited, then N follows a negative binomial distribution with failure probability $1-\rho$ and q failures. Conditioned on N, the accumulated occupation times in the states parameterised with λ_1 and λ_2 are independent and Erlang distributed with intensity λ_1 and λ_2 respectively. The density function can be written as

$$
\begin{aligned}
& f(y_1, y_2) = \sum_{n=q}^{\infty} P(N=n) f(y_1, y_2 | N=n) \\
= & \sum_{n=q}^{\infty} \binom{n-1}{q-1} \rho^{n-q} (1-\rho)^q \lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!} e^{-\lambda_2 y_2} \\
= & \lambda_1 \lambda_2 e^{-\lambda_1 y_1 - \lambda_2 y_2} (1-\rho)^q \sum_{n=q}^{\infty} \frac{(n-1)!}{(q-1)!(n-q)!} \rho^{n-q} \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} \frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!} \\
= & \frac{\lambda_1 \lambda_2 e^{-\lambda_1 y_1 - \lambda_2 y_2}}{(q-1)!} (1-\rho)^q \sum_{k=0}^{\infty} \frac{1}{k!} \rho^k \frac{(\lambda_1 y_1)^{k+q-1} (\lambda_2 y_2)^{k+\ell+q-1}}{(k+\ell+q-1)!} \\
= & \frac{\lambda_1 \lambda_2 e^{-\lambda_1 y_1 - \lambda_2 y_2}}{(q-1)!} (1-\rho)^q \frac{(\lambda_1 y_1 \lambda_2 y_2)^{q-1} (\lambda_2 y_2)^{\ell}}{(\sqrt{\rho \lambda_1 y_1 \lambda_2 y_2})^{q+\ell-1}} \sum_{k=0}^{\infty} \frac{(\sqrt{\rho \lambda_1 y_1 \lambda_2 y_2})^{2k+\ell+q-1}}{k!(k+\ell+q-1)!} \\
= & \frac{\lambda_1 \lambda_2 e^{-\lambda_1 y_1 - \lambda_2 y_2}}{(q-1)!} (1-\rho)^q \frac{(\lambda_1 y_1 \lambda_2 y_2)^{q-1} (\lambda_2 y_2)^{\ell}}{(\sqrt{\rho \lambda_1 y_1 \lambda_2 y_2})^{q+\ell-1}} I_{q+\ell-1}(2\sqrt{\rho \lambda_1 y_1 \lambda_2 y_2})
\end{aligned}
$$

where $I_q(z) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(i+q+1)} \left(\frac{z}{2}\right)^{2i+q}$ is the modified Bessel function of the first kind.

3. Moment based estimators for $q = 1$ and $l = 0$

When $q = 1$ and $l = 0$ Y_1 and Y_2 are exponentially distributed. Several estimators for ρ have been derived for this case. Let us consider w different sets of observations $Y^{(j)} = (Y_1^{(j)}, Y_2^{(j)})$ for $j \in \{1, \dots, w\}$, where $Y_1^{(j)}$ is the accumulated time spent in states with sojourn time distribution $\exp(\lambda_1)$ and $Y_2^{(j)}$ is the accumulated time spent in the states with sojourn time distribution being $\exp(\lambda_2)$ respectively. We define the average of our observed times:

$$\bar{Y_i} \;=\; \frac{1}{w}\sum_{j=1}^{w} Y_i^{(j)},\ i \in \{1, 2\}$$

as well as the sample correlation coefficient

$$\tilde{\rho}_0 = \frac{\sum_{j=1}^{w}(Y_1^{(j)} - \bar{Y_1})(Y_2^{(j)} - \bar{Y_2})}{\sqrt{\sum_{j=1}^{w}(Y_1^{(j)} - \bar{Y_1})^2 \cdot \sum_{j=1}^{w}(Y_2^{(j)} - \bar{Y_2})^2}}.$$

Furthermore we need

$$\tilde{\rho} = \frac{\sum_{j=1}^{w}(Y_1^{(j)} - \bar{Y_1})(Y_2^{(j)} - \bar{Y_2})}{w\bar{Y_1}\bar{Y_2}}.$$

With this the estimators obtained from Kotz, Balakrishnan and Johnson [19] are

1. $\tilde{\rho}_1 = \begin{cases} 0 & \text{if } \tilde{\rho} \leq 0 \\ \tilde{\rho} & \text{if } 0 < \tilde{\rho} \leq 1, \\ 1 & \text{if } \tilde{\rho} > 1. \end{cases}$

2. $\tilde{\rho}_2 = \tilde{\rho}\left(1 + \frac{3}{w}\right) - \frac{1}{w}\tilde{\rho}^2$

3. $\tilde{\rho}_3 = \begin{cases} 0 & \text{if } -1 \leq \tilde{\rho}_0 \leq 0, \\ \tilde{\rho}_0 & \text{if } 0 < \tilde{\rho}_0 \leq 1. \end{cases}$

4. $\tilde{\rho}_4 = \begin{cases} 0 & \text{if } \tilde{\rho}_0(1 + \frac{2}{w}) - \frac{1}{w}(\tilde{\rho}_0^2 + \tilde{\rho}_0^3) \leq 0 \\ \tilde{\rho}_0(1 + \frac{2}{w}) - \frac{1}{w}(\tilde{\rho}_0^2 + \tilde{\rho}_0^3) & \text{if } 0 < \tilde{\rho}_0(1 + \frac{2}{w}) - \frac{1}{w}(\tilde{\rho}_0^2 + \tilde{\rho}_0^3) \leq 1, \\ 1 & \text{if } \tilde{\rho}_0(1 + \frac{2}{w}) - \frac{1}{w}(\tilde{\rho}_0^2 + \tilde{\rho}_0^3) > 1. \end{cases}$

5. $\tilde{\rho}_5 = \begin{cases} 0 & \text{if } \frac{-(w-3)+\sqrt{(w-3)^2+4w\tilde{\rho}}}{2} \leq 0 \\ \frac{-(w-3)+\sqrt{(w-3)^2+4w\tilde{\rho}}}{2} & \text{if } 0 < \frac{-(w-3)+\sqrt{(w-3)^2+4w\tilde{\rho}}}{2} \leq 1, \\ 1 & \text{if } \frac{-(w-3)+\sqrt{(w-3)^2+4w\tilde{\rho}}}{2} > 1. \end{cases}$

6. $\tilde{\rho}_{5J} = \begin{cases} 0 & \text{if } w\tilde{\rho}_5 - (w-1)\tilde{\rho}_{5(\cdot)} \leq 0 \\ w\tilde{\rho}_5 - (w-1)\tilde{\rho}_{5(\cdot)} & \text{if } 0 < w\tilde{\rho}_5 - (w-1)\tilde{\rho}_{5(\cdot)} \leq 1, \\ 1 & \text{if } w\tilde{\rho}_5 - (w-1)\tilde{\rho}_{5(\cdot)} > 1. \end{cases}$

   where $\tilde{\rho}_{5(\cdot)} = \frac{1}{w}\sum_{i=1}^{w} \tilde{\rho}_{5(i)}$.

7. $\tilde{\rho}_6 = \begin{cases} 0 & \text{if } \hat{\rho}^\star \leq 0 \\ \hat{\rho}^\star & \text{if } 0 < \hat{\rho}^\star \leq 1, \\ 1 & \text{if } \hat{\rho}^\star > 1. \end{cases}$

   where $\tilde{\rho}^\star = C + D - \frac{1}{3}$,

   $C = \left(-\frac{B}{2} + \sqrt{\frac{A^3}{27} + \frac{B^2}{4}}\right)^{\frac{1}{3}}$,

   $D = \left(-\frac{B}{2} - \sqrt{\frac{A^3}{27} + \frac{B^2}{4}}\right)^{\frac{1}{3}}$,

   $A = w - \frac{7}{3}$,

The estimators $\tilde{\rho}_{5(i)}$ and $\tilde{\rho}_{6(i)}$ are determined using $\tilde{\rho}_5$ and $\tilde{\rho}_6$ leaving out the i-th observation. Al-Saadi and Young (1980) [1] modified the moment-based estimator $\tilde{\rho}$, originally proposed by Nagao and Kadayo (1971) [15] to $\tilde{\rho}_1$. Their estimator can be used to test for $\rho = 0$. For large w, Al-Saadi and Young proposed $\tilde{\rho}_2$ in order to reduce bias. Furthermore they proposed $\tilde{\rho}_3$ as a sample coefficient-based estimator and improved it to $\tilde{\rho}_4$. They compared the estimators and concluded that $\tilde{\rho}_2$ and $\tilde{\rho}_4$ have for small samples a much smaller bias then $\tilde{\rho}_1$ as well as $\tilde{\rho}_3$, with an exception for cases where ρ is of small value.

The estimator $\tilde{\rho}_5$ was suggested by Balakrishnan and Ng (2001) [3], and $\tilde{\rho}_{5J}$ is simply the jack knife version of $\tilde{\rho}_5$. They showed that $\tilde{\rho}_{5J}$ reduces the bias substantially. They also introduced $\tilde{\rho}_6$ and state " ... in terms of MSE, we find that $\tilde{\rho}_6$ to be the best among all the estimators considered ... "'. Based on these remarks one would expect $\tilde{\rho}_{5J}$ and $\tilde{\rho}_6$ to be best performing among the moment estimators.

4. The EM algorithm

The EM algorithm was formulated by Dempster et al (1977) [8] and since then it has become a standard method for dealing with incomplete data. The algorithm is an iterative method for calculating maximum

likelihood estimates in cases with missing data or where a direct evaluation of the observed data likelihood function is difficult. The algorithm consists of two main steps. In the E (expectation) step the expectation of the log likelihood of the unobserved data given the current estimate of the parameters as well as the observed data is calculated. In the M (maximisation) step the expected log likelihood function is maximised to derive new parameters.

For distributions belonging to the exponential family one can replace unobserved data with their sufficient statistic. A fully observed Markov jump process belongs to the exponential family of distributions. Given incomplete data, for PH distributions the absorption time, we calculate the expectation of the sufficient statistics of the unobserved data and use the expected sufficient statistics as if they were obtained from observed data. Asmussen et al (1996) [2] applied the EM algorithm to phase-type (PH) distributions. We now describe their approach. The sufficient statistics consist of the number of realizations starting in the different states, the number of jumps between the states as well as the total occupation time in each state.

Observing the complete Markov jump process of a PH distribution is equivalent to observing the embedded Markov chain, i.e. the states the Markov jump process is in after each jump $I_0, \cdots, I_{t-1}, I_t = m+1$

and the sojourn times $S_0, \cdots, S_{t-1}, S_t = \infty$ in these states, where t is the number of jumps until absorption. This enables us to describe the complete information as $x = (i_0, \cdots, i_{t-1}, s_0, \cdots, s_{t-1})$ with $s_0 + s_1 + \cdots + s_{t-1} = \tau$.

If we consider a PH distribution with sub-generator matrix T and initial distribution α the complete data likelihood for one observation can be written as

$$f(x;\ \alpha,\ T) = \prod_{i=1}^{m} \alpha_i^{B_i} \prod_{i=1}^{m} e^{t_{ii} Z_i} \prod_{i=1}^{m} \prod_{j=1, j\neq i}^{m+1} t_{ij}^{N_{ij}},$$

with B_i being one if $\{X(t)\}_{t\geq 0}$ started in i and zero otherwise, Z_i being the total time spent in state i, and N_{ij} the total number of jumps from i to j. If more than one observation is available the complete data likelihood can be used unchanged, except B_i now being the number of observations starting in i instead of being an indicator function. The advantage of the EM algorithm for the univariate case is that all calculations in the E step can be done using matrix analytic methods. When dealing with MPH* distributions this property vanishes. There exists no general explicit formulation for the density of MPH* distributions, this makes it impossible to derive similar general results as in the univariate case.

Let us consider an MPH* distribution where the reward matrix R is a diagonal matrix, i.e. the only non zero entries are on the diagonal of the matrix. For this special case the two representations MPH$^\star(\alpha, T, R)$ and MPH$^\star(\alpha, R^{-1}T, I)$ relate to the same distribution. The likelihood function becomes

$$f(x;\ \alpha,\ T,\ R) = \prod_{i=1}^{m} \alpha_i^{B_i} \prod_{i=1}^{m} e^{t_{ii}\frac{y_i}{r_{ii}}} \prod_{i=1}^{m} \prod_{j=0, j\neq i}^{m} \left(\frac{t_{ij}}{r_{ii}}\right)^{N_{ij}}.$$

The complete data likelihood function is similar to the univariate case and the maximum likelihood estimates can easily be calculated. Furthermore we can assume $R = I$ for any model with diagonal reward matrix. If we consider a reward matrix with non diagonal structure, the likelihood function can become more complex but the maximisation will always be possible whereas it might become impossible to get an analytic solution for the conditional expectations.

We now describe how to derive the formulas needed in the expectation step in order to use the EM algorithm to estimate the parameters for the Kibble distribution. The variables Y_1 and Y_2 directly give the accumulated occupation times of the underlying jump process in the states with sojourn time distribution being $\exp(\lambda_1)$ and $\exp(\lambda_2)$.

A complete observation would also contain B_1, B_2, N with $N = \sum_i^q N_{2i-1}$ where N_i is the number of visits to the states with sojourn

time distribution $\exp(\lambda_1)$. Due to the structure of the underlying Markov jump process the only information missing is N.

The complete likelihood for one observation can be written as

$$f(x;T) = 1 \cdot e^{-\lambda_1 \cdot y_1} \cdot e^{-\lambda_2 \cdot y_2} \lambda_1^N \cdot (p\lambda_2)^{N-q} \cdot ((1-p)\lambda_2)^q \cdot \lambda_2^\ell.$$

We have immediately that the maximum likelihood estimates based on the complete sample x are

$$\hat{\lambda}_1 = \frac{N}{y_1}, \quad \hat{\lambda}_2 = \frac{N+\ell}{y_2}, \quad \hat{\rho} = \frac{N-q}{N}.$$

We consider a set of observations $\left\{y^{(1)}, \cdots, y^{(w)}\right\}$ with $y^{(v)} = \left(y_1^{(v)}, y_2^{(v)}\right)$, and let $N^{(v)}$ be the number of times the underlying Markov jump process visits the states with sojourn time distribution parameterised with λ_1. The likelihood function can be written as

$$\begin{aligned} f(x,T) &= e^{-\lambda_1 \cdot (\sum_{v=1}^w y_1^{(v)})} \cdot e^{-\lambda_2 \cdot (\sum_{v=1}^w y_2^{(v)})} \\ &\cdot \lambda_1^{\sum_{v=1}^w N^{(v)}} \cdot (p\lambda_2)^{\sum_{v=1}^w (N^{(v)}-r)} \cdot ((1-p)\lambda_2)^{w \cdot r} \cdot \lambda_2^{w \cdot \ell} \end{aligned}$$

and the estimates for the set of observations $\left\{y^{(1)}, \cdots, y^{(w)}\right\}$ are

$$\bar{\lambda}_1 = \frac{\sum_{v=1}^w N^{(v)}}{\sum_{v=1}^w y_1^{(v)}},$$

$$\begin{aligned}\bar{\lambda}_2 &= \frac{\sum_{v=1}^{w}\left(N^{(v)}+\ell\right)}{\sum_{v=1}^{w} y_2^{(v)}},\\ \bar{\rho} &= \frac{\sum_{v=1}^{w}\left(N^{(v)}-r\right)}{\sum_{v=1}^{w} E[N^{(v)}|y^{(v)}]}.\end{aligned}$$

In order to apply the EM algorithm we are left with calculating $E[N|y]$.

$$\begin{aligned}E_{\alpha,T,R}[N|y] &= \frac{\sum_{n=q}^{\infty} n\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}e^{-\lambda_1 y_1}\lambda_2\frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!}e^{-\lambda_2 y_2}}{\sum_{n=q}^{\infty}\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\lambda_1\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}e^{-\lambda_1 y_1}\lambda_2\frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!}e^{-\lambda_2 y_2}}\\ &= \frac{\sum_{n=q}^{\infty} n\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}\frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!}}{\sum_{n=q}^{\infty}\binom{n-1}{q-1}\rho^{n-q}(1-\rho)^q\frac{(\lambda_1 y_1)^{n-1}}{(n-1)!}\frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!}}\\ &= \frac{\frac{1}{(q-1)!}\sum_{k=0}^{\infty}\frac{k+q}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+\ell+q-1)!}}{\frac{1}{(q-1)!}\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+\ell+q-1)!}}\\ &= \frac{\sum_{k=0}^{\infty}\frac{k}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+\ell+q-1)!}+q\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+\ell+q-1)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+q-1}}{(k+\ell+q-1)!}}\\ &= \frac{\sum_{k=1}^{\infty}\frac{1}{(k-1)!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}}+q\\ &= \frac{\sqrt{\rho\lambda_1\lambda_2 y_1 y_2}\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})^{2k+\ell+q}}{(k+\ell+q)!}}{\sum_{k=0}^{\infty}\frac{1}{k!}\cdot\frac{(\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})^{2k+\ell+q-1}}{(k+\ell+q-1)!}}+q\\ &= \sqrt{\rho\lambda_1\lambda_2 y_1 y_2}\cdot\frac{I_{\ell+q}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{\ell+q-1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}+q,\end{aligned}$$

with $I_q(z)=\sum_{i=0}^{\infty}\frac{1}{i!\Gamma(i+q+1)}\left(\frac{z}{2}\right)^{2i+q}$ being the modified Bessel function of the first kind.

4.1. **Fisher Information Matrix.** The Fisher information matrix is well known for estimating the inverse of the variance covariance matrix. Having the variance covariance matrix we can give confidence intervals for our estimates. There are several ways to calculate the Fisher information matrix. For example Oakes (1999) [18] shows how to obtain the Fisher information matrix directly from the EM algorithm while Bladt, Esparza and Nielsen (2011) [5] derive explicit formulas for the EM algorithm applied to univariate PH distributions. For the parameter vector Ψ the observed Fisher information matrix of the estimates $I(\hat{\Psi}, y)$ can be calculated as the second derivative of the observed log likelihood function evaluated using the estimated parameters. Often second derivatives are difficult to evaluate, and alternatively the Fisher information matrix can be written as

$$I(\hat{\Psi}, y) = \mathbb{I}_c(\hat{\Psi}, y) - [E_\Psi \left\{ S_c(Y;\Psi) S_c^T(Y;\Psi) \right\} | y]_{\Psi=\hat{\Psi}}$$

with $\mathbb{I}_c(\hat{\Psi}, y)$ being the negative expectation of the second derivative of the complete data log likelihood function and $S_c(Y;\Psi)$ being the complete data score statistic, the first derivative of the complete data log likelihood. In our example the complete data log likelihood is

$$\begin{aligned} \log f(x,T) &= -\lambda_1 y_1 - \lambda_2 y_2 + N\log(\lambda_1) \\ &+ (N+\ell)\log(\lambda_2) + (N-q)\log(\rho) + q\log(1-\rho). \end{aligned}$$

And with that the Fisher information matrix can be written as

$$I(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\rho}, y) = E_{\alpha,T,R}\left[\left[\begin{pmatrix} \frac{N}{\lambda_1^2} & 0 & 0 \\ 0 & \frac{N+\ell}{\lambda_2^2} & 0 \\ 0 & 0 & \frac{N-q}{\rho^2} + \frac{q}{(1-\rho)^2} \end{pmatrix} \middle| y \right]\right]$$

$$+ \quad E_{\alpha,T,R}\left[\left[\begin{pmatrix} -y_1 + \frac{N}{\lambda_1} \\ -y_2 + \frac{N+\ell}{\lambda_2} \\ \frac{N-q}{\rho} - \frac{q}{(1-\rho)} \end{pmatrix} \cdot \begin{pmatrix} -y_1 + \frac{N}{\lambda_1} & -y_2 + \frac{N+\ell}{\lambda_2} & \frac{N-q}{\rho} - \frac{q}{(1-\rho)} \end{pmatrix} \middle| y \right]\right].$$

The only random variables in this equation are N and N^2. We get $E_{\alpha,T,R}[N|y]$ directly from the expectation step in our algorithm. This leaves us with deriving

$$\begin{aligned}
E_{\alpha,T,R}[N^2|y] &= \frac{\sum_{n=q}^{\infty} n^2 \binom{n-1}{q-1} \rho^{n-q}(1-\rho)^q \lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!} e^{-\lambda_2 y_2}}{\sum_{n=q}^{\infty} \binom{n-1}{q-1} \rho^{n-q}(1-\rho)^q \lambda_1 \frac{(\lambda_1 y_1)^{n-1}}{(n-1)!} e^{-\lambda_1 y_1} \lambda_2 \frac{(\lambda_2 y_2)^{n+\ell-1}}{(n+\ell-1)!} e^{-\lambda_2 y_2}} \\
&= \frac{\frac{1}{(q-1)!} \sum_{k=0}^{\infty} \frac{(k+q)^2}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+\ell+q-1}}{(k+\ell+q-1)!}}{\frac{1}{(q-1)!} \sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+\ell+q-1}}{(k+\ell+q-1)!}} \\
&= \frac{\sum_{k=0}^{\infty} \frac{k^2}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+\ell+q-1}}{(k+\ell+q-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k+\ell+q-1}}{(k+\ell+q-1)!}} + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \frac{\sum_{k=0}^{\infty} \frac{k^2}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \frac{\sum_{k=0}^{\infty} \frac{k}{(k-1)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+r-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \frac{\sum_{k=1}^{\infty} \frac{k}{(k-1)!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{s=0}^{\infty} \frac{s+1}{s!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+1}}{(s+\ell+q)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \frac{\sum_{\ell=0}^{\infty} \frac{\ell}{s!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+1}}{(s+\ell+q)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + \left(E_{\alpha,T,R}[N|y] - q\right) \\
&+ 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \frac{\sum_{\ell=0}^{\infty} \frac{1}{s!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{s+2}}{(s+\ell+q+1)!}}{\sum_{k=0}^{\infty} \frac{1}{k!} \cdot \frac{(\rho\lambda_1\lambda_2 y_1 y_2)^{k}}{(k+\ell+q-1)!}} + \left(E_{\alpha,T,R}[N|y] - q\right) \\
&+ 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2 \\
&= \rho\lambda_1\lambda_2 y_1 y_2 \cdot \frac{I_{q+\ell+1}(2\sqrt{\rho\lambda_1\lambda_2 y_1 y_2})}{I_{q+\ell-1}(2z)} \\
&+ \left(E_{\alpha,T,R}[N|y] - q\right) + 2q\left(E_{\alpha,T,R}[N|y]\right) + q^2
\end{aligned}
$$

5. Examples - Mixtures of exponential distributions

For the special case of a bivariate exponential distribution we will compare our result to the selection of estimates listed in Johnson, Balakrishnan and Kotz (2001) [19] introduced in Section 3. We verify the accuracy of our estimates using the method suggested by Chatelain et al (2007) [7]. To compare the EM algorithm to the parameter estimators for ρ introduced in Section 2 we simulate bivariate data by first drawing w geometrically distributed random numbers $G_1, \ldots, G_w$ and afterwards, drawing Erlang-G_i random numbers with parameter λ_1 and λ_2, respectively for $i = 1, \ldots, w$. We investigate the cases where

the intensities are approximately equal and where they are quite different, along with different values for $0 < \rho < 1$. We produced data sets by simulating with $\rho_1 = 0.2$, $\rho_2 = 0.4$, $\rho_3 = 0.6$, and $\rho_4 = 0.8$ for $(\lambda_1, \lambda_2) = (0.7, 1.3)$ and for $(\lambda_1, \lambda_2) = (1, 10)$ to represent the scenario where the intensities are approximately the same and the scenario where the intensities are clearly different. We created data sets with 50, 250, and 1000 data points.

The marginal distributions are exponential by construction. To give an impression of the bivariate distribution and our data we have compared bivariate histograms of our data sets with 1000 data points to contour lines from the pdf for the Kibble distribution. For each data set, we have scaled the data by ρ to make the graphical comparison easier. We consider the case $\lambda_1 \approx \lambda_2$. Figure 1 shows the data together with contour plots of the Kibble distribution. We get a similar picture when we look at the case $\lambda_1 << \lambda_2$.

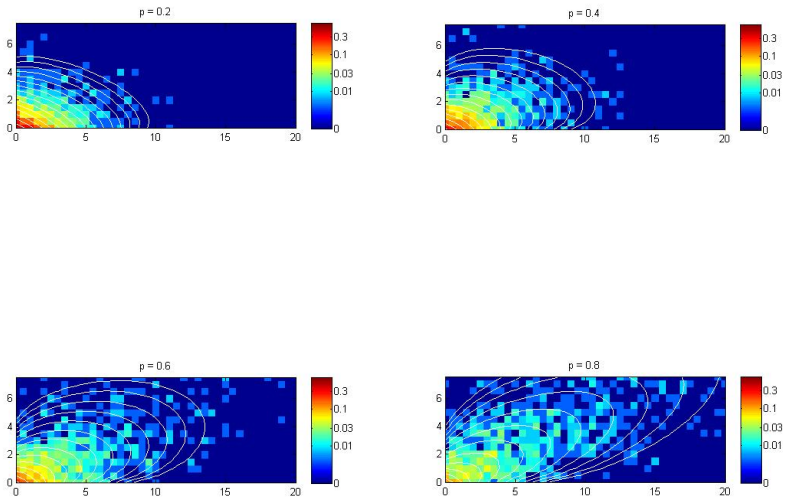


FIGURE 1. Contour lines of the pdf and bivariate histogram of the simulated data for $\lambda_1 = 0.7$ and $\lambda_2 = 1.3$

Tables 1-6 present our results.

TABLE 1. 50 data points, $\lambda_1 = 0.7$, $\lambda_2 = 1.3$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.1529	0.1529	0.2879	0.3035	0.1959	0.2028	0.3043	0.2030	0.2957	0.1845
0.4	0.4692	0.4692	0.1516	0.1603	0.2399	0.2480	0.1608	0.2483	0.1529	0.2426
0.6	0.6898	0.6898	0.4532	0.4763	0.5978	0.6103	0.4773	0.6102	0.4547	0.5919
0.8	0.8416	0.8416	1.0000	1.0000	0.8675	0.8741	1.0000	0.8739	1.0000	0.8901

TABLE 2. 250 data points, $\lambda_1 = 0.7$, $\lambda_2 = 1.3$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.1832	0.1832	0.1583	0.1601	0.1652	0.1663	0.1601	0.1664	0.1592	0.1638
0.4	0.5039	0.5036	0.3899	0.3940	0.4255	0.4279	0.3940	0.4279	0.3922	0.4242
0.6	0.5859	0.5859	0.7689	0.7758	0.6476	0.6500	0.7758	0.6500	0.7829	0.6610
0.8	0.8085	0.8085	0.8358	0.8430	0.8245	0.8262	0.8430	0.8261	0.8453	0.8284

TABLE 3. 1000 data points, $\lambda_1 = 0.7$, $\lambda_2 = 1.3$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.2209	0.2209	0.2151	0.2157	0.2188	0.2192	0.2157	0.2192	0.2156	0.2188
0.4	0.4446	0.4446	0.3593	0.3603	0.3880	0.3885	0.3603	0.3886	0.3599	0.3878
0.6	0.5769	0.5769	0.6497	0.6512	0.6201	0.6207	0.6512	0.6207	0.6516	0.6213
0.8	0.7744	0.7744	0.7775	0.7792	0.7905	0.7910	0.7792	0.7910	0.7795	0.7913

TABLE 4. 50 data points, $\lambda_1 = 1$, $\lambda_2 = 10$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.0735	0.0735	0.1263	0.1336	0.1404	0.1455	0.1340	0.1457	0.1214	0.1470
0.4	0.4974	0.4974	0.3125	0.3293	0.3556	0.3664	0.3302	0.3666	0.3255	0.3433
0.6	0.6417	0.6417	1.0000	1.2558	0.7967	0.8058	1.0000	0.8055	1.0000	0.8197
0.8	0.8095	0.8095	1.0000	1.0000	0.8989	0.9042	1.0000	0.9039	1.0000	0.9156

As expected, in general the estimates come closer to the true parameter value the more observations are available. In general, we can say that the estimates for ρ by the EM algorithm are at least as good as

TABLE 5. 250 data points, $\lambda_1 = 1$, $\lambda_2 = 10$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.1809	0.1809	0.1439	0.1456	0.1692	0.1704	0.1456	0.1704	0.1453	0.1703
0.4	0.3953	0.3953	0.3204	0.3238	0.3340	0.3361	0.3238	0.3361	0.3227	0.3348
0.6	0.6430	0.6430	0.7283	0.7350	0.6966	0.6988	0.7350	0.6988	0.7352	0.6990
0.8	0.8017	0.8017	0.7636	0.7704	0.8052	0.8070	0.7705	0.8070	0.7687	0.8072

TABLE 6. 1000 data points, $\lambda_1 = 1$, $\lambda_2 = 10$

ρ_{real}	ρ_{EM}	ρ_{Ch}	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	ρ_{5J}	ρ_{6J}
0.2	0.1827	0.1827	0.1779	0.1784	0.1811	0.1814	0.1784	0.1814	0.1785	0.1821
0.4	0.3828	0.3828	0.3439	0.3448	0.3532	0.3537	0.3448	0.3537	0.3445	0.3532
0.6	0.6113	0.6113	0.5821	0.5835	0.6049	0.6055	0.5835	0.6055	0.5833	0.6050
0.8	0.8088	0.8088	0.7729	0.7747	0.8083	0.8087	0.7747	0.8087	0.7745	0.8090

the other estimates, and it seems that the EM algorithm handles the case where $\lambda_1 = 1 << \lambda_2 = 10$ better than the other methods. The calculation time is, at least for small data samples, reasonable.

5.1. **Precision of estimates.** A weakness of MPH* distributions is that most distributions do not have a unique MPH* representation. This makes it difficult to compare parameters directly and makes it insensible to calculate precision estimates like confidence intervals. We now show that any estimator describing the correlation in Kibble's bivariate exponential distribution for a given reward matrix R is unique

except for permutations. Let $\sim$ describe the identity in distribution of two MPH* distributions and assume

$$\begin{aligned}
\alpha &= (1,0),\ T = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \rho_2 \cdot \lambda_2 & -\lambda_2 \end{pmatrix}, \\
\beta &= (\beta_1, \beta_2),\ U = \begin{pmatrix} -\mu_1 & r_1 \cdot \mu_1 \\ r_2 \cdot \hat{\lambda}_2 & -\mu_2 \end{pmatrix}, R = I.
\end{aligned}$$

In order to prove the uniqueness, we will compare the Laplace transform of the two representations:

$$\begin{aligned}
& MPH^\star(\beta, U, R) \sim MPH^\star(\alpha, T, R) \\
\Leftrightarrow\quad & \forall\ s\ L_{\beta,U}(s) = L_{\alpha,T}(s) \\
\Longleftrightarrow\quad & \frac{\lambda_1\lambda_2(1-\rho_2)}{s_1s_2 + s_2\lambda_1 + s_1\lambda_2 + \lambda_1\lambda_2(1-\rho_2)} \\
=\quad & \frac{\beta\alpha_1 s_2\mu_1(1-r_1)\mu_2 + \beta_2 s_1\mu_1\mu_2(1-r_2) + \mu_1\mu_2(1-r_1r_2)}{s_1s_2 + s_2\mu_1 + s_1\mu_2 + \mu_1\mu_2(1-r_1r_2)} \\
\Longleftrightarrow\quad & \beta_1\mu_1\mu_2(1-r_1)s_1s_2^2 + \beta_2\mu_1\mu_2(1-r_2)s_1^2s_2 \\
+\quad & \lambda_1\beta_1\mu_1\mu_2(1-r_1)s_2^2 + \lambda_2\beta_2\mu_1\mu_2(1-r_2)s_1^2 \\
+\quad & (\lambda_1\beta_2\mu_1\mu_2(1-r_2) + \mu_1\mu_2(1-r_1r_2) + \beta_1\lambda_2\mu_1\mu_2(1-r_1) - \lambda_1\lambda_2(1-\rho_2))\, s_1s_2 \\
+\quad & (\lambda_1\mu_1\mu_2(1-r_1r_2) + \beta_1\lambda_1\lambda_2(1-\rho_2)\mu_1\mu_2(1-r_1) - \mu_1\lambda_1\lambda_2(1-\rho_2))\, s_2 \\
+\quad & (\lambda_2\mu_1\mu_2(1-r_1r_2) + \beta_2\lambda_1\lambda_2(1-\rho_2)\mu_1\mu_2(1-r_2) - \mu_2\lambda_1\lambda_2(1-\rho_2))\, s_1 = 0 \\
\Longrightarrow\quad & \{\beta_1 = 0\ \vee\ r_1 = 1\}\ \wedge\ \{\beta_2 = 0\ \vee\ r_2 = 1\} \\
\Longrightarrow\quad & \mu_1 = \lambda_1,\ \mu_2 = \lambda_2,\ \{\beta_1 = 1,\ \beta_2 = 0,\ r_1 = 1,\ r_2 = \rho_2\}
\end{aligned}$$

$$\vee \quad \{\beta_1 = 0,\ \beta_2 = 1,\ r_1 = \rho_2,\ r_2 = \rho_1\}.$$

With the uniqueness of representation we know it is safe to compare the estimated correlation parameter with the one used to generate the data and to give confidence intervals for the estimates.

For the estimates obtained using the EM algorithm we calculated the Fisher information matrix. Since ρ is the critical parameter, we will state the estimated standard deviation σ_ρ for ρ. In Tables 7 and 8 we show the results for the examples with 1000 data points.

TABLE 7. 1000 data points, $\lambda_1 = 0.7$, $\lambda_2 = 1.3$

ρ	0.2	0.4	0.6	0.8
ρ_{EM}	0.2209	0.4446	0.5769	0.7744
σ_ρ	0.0070	0.0092	0.0087	0.0059

TABLE 8. 1000 data points, $\lambda_1 = 1$, $\lambda_2 = 10$

p	0.2	0.4	0.6	0.8
p_{EM}	0.1827	0.3828	0.6113	0.8088
σ_ρ	0.0061	0.0090	0.0084	0.0051

The correlation between ρ and the intensity parameters is negative as expected. An increase of ρ correlates to more jumps before absorption of the underlying Markov jump process. The number of visits to

the states with different sojourn time distributions directly influences the random variables Y_1 and Y_2 to counter this effect the intensity parameters λ_1 and λ_2 would have to be decreased.

6. Discussion

We have adapted the EM algorithm to be used to estimate parameters for a bivariate mixture of Erlang distributions also known as the Kibble distribution. Our focus has been on the estimation in the Kibble distribution taking advantage of the interpretation as an MPH* distribution. For this reason we have focused on integer shape parameters. The EM algorithm is, however, also applicable in the general case only the interpretation as a bivariate phase type distribution is lost. With the results from Section 4.1 it is possible to give confidence intervals for the estimates.

For the special case where the shape parameters are one we have applied the EM algorithm to simulated data and compared it to a selection of moment based estimators specially developed for Kibble's bivariate mixture of exponential distributions as well as a direct maximum likelihood approach. The EM algorithm calculates a maximum likelihood estimate and is compared to the moment based estimators more robust

towards variations in the parameters as well as data size. We have validated our experimental results by implementing a standard Newton Raphson method like the one used by Chatelain et al [7]. Comparing the EM algorithm to the direct approach used by Chatelain (2007) [7] the direct approach seems to somewhat outperform the EM algorithm with respect to speed, while the EM algorithm guarantees convergence and give a closed form expression for the Fisher information matrix.

In order to apply our method in situations where the computation time is crucial it would be useful to investigate the option to speed up the EM algorithm. Right now it is initiated with random parameters. It is possible that choosing the sample correlation coefficient for ρ as an initial value would increase the convergence speed. In cases with identical shape parameters, this can be taken so far as to use the direct approach by Chatelain (2007) [7] and use our results only to calculate the Fisher information matrix. Another interesting problem for further investigation would be to analytically calculate the convergence rate of our EM algorithm.

There exist no general applicable method to obtain maximum likelihood estimators for multivariate phase-type distributions. This paper is a first step in estimating for this class of distributions. Future work should investigate the possibility to generalise our results, ideally to

produce similar general results as for univariate phase type distributions. This might be more challenging than expected. First steps can definitely be done by extending our results to more general mixtures of Erlang distributions.

Acknowledgement

David Meisch and Bo Friis Nielsen would like to thank the Danish Council for Strategic Research for the support through the UNITE project under grant 2140-08-0011. Mogens Bladt acknowledges the support from the Mexican Research Council, Conacyt, grant 48538. David Meisch would like to thank Sue Simmons for her support.

References

[1] S.D. Al-Saadi and D.H. Young. Estimators for the correlation coefficient in a bivariate exponential distribution. *Journal of Statistical Computation and Simulation*, 11(1):13–20, 1980.

[2] S Asmussen, O Nerman, and M Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistic*, 23(4):419–441, DEC 1996.

[3] N. Balakrishnan and Hon Keung. Tony Ng. Improved estimation of the correlation coefficient in a bivariate exponential distribution. *Journal of Statistical Computation and Simulation*, 68(2):173–184, 2001.

[4] Mogens Bladt. A review on phase-type distributions and their use in risk theory. *Astin Bulletin*, 35(1):145–167, 2005.

[5] Mogens Bladt, Luz Judith R Esparza, and Bo Friis Nielsen. Fisher information and statistical inference for phase-type distributions. *Journal of Applied Probability*, 48A:277–293, 2011.

[6] Mogens Bladt and Bo Friis Nielsen. Multivariate matrix-exponential distributions. *Stochastic Models*, 26(1):1–26, 2010.

[7] F. Chatelain, J.-Y. Tourneret, A. Ferrari, and J. Inglada. Bivariate gamma distributions for image registration and change detection. *IEEE Trans. Image Processing*, 16(7):1796–1806, July 2007.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

[9] F. Downton. Bivariate exponential distributions in reliability theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(3):pp. 408–417, 1970.

[10] George Iliopoulos, Dimitris Karlis, and Ioannis Ntzoufras. Bayesian estimation in Kibble's bivariate gamma distribution. *The Canadian Journal of Statistics*, 33(4):571–589, 2005.

[11] W. F. Kibble. A two-variate gamma type distribution. *The Indian Journal of Statistics*, 5(2):137–150, 1941.

[12] V. G. Kulkarni. A new class of multivariate phase type distributions. *Operations Research*, 37(1):151, 1989.

[13] Yu-Jau Lin, Y. L. Lio, and Hon Keung Tony Ng. Bayes estimation of MoranâĂŞDownton bivariate exponential distribution based on censored samples. *Journal of Statistical Computation and Simulation*, 83(5):837–852, 2013.

[14] P. A. P. Moran. Statistical inference with bivariate gamma distributions. *Biometrika*, 56(3):pp. 627–634, 1969.

[15] M Nagao and M Kadoya. Two bivariate exponential distribution its numerical table for engineering application. *Bulletin of Disaster Prevention Research Institute*, 20(178):183–215, 1971.

[16] M. F. Neuts. Probability distrubutions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*, pages 173–206, Department of Mathematics, University of Louvian, Belgium, 1975.

[17] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models an Algorithmic Approach*, volume 2 of *John Hopkins Series in Mathematical Sciences*. The John Hopkins University Press, 1981.

[18] David Oakes. Direct calculation of the information matrix via the em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61:479–482, 1999.

[19] N. Balakrishnan S. Kotz and N. L. Johnson. *Continuous Multivariate Distributions*. Wiley-Interscience, New York, 2nd edition edition, 2000.

[20] Chen-Ping Wang and Malay Gosh. Bayesian analysis of bivariate competing risk models. *Indian Journal of Statistic*, 62:388–401, 2000.

[21] S. Yue, T.B.M.J. Ouarda, and B. Bobée. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology*, 246(1-4):1–18, 2001.

Institute for Applied Mathematics and Systems, National University of Mexico, A.P. 20-726, 01000 Mexico, D.F., Mexico

E-mail address: bladt@sigma.iimas.unam.mx

Department for Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

E-mail address: dame@dtu.dk

Department for Applied Mathematics and Computer Science, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

E-mail address: bfn@imm.dtu.dk

APPENDIX B

Paper 2

On the use of phase type distributions in the Successive Principle

Authors:
David Meisch, Bo Friis Nielsen

Submitted to:
"European Journal of Transport and Infrastructure Research"

On the use of phase type distributions in the Successive Principle

David Meisch[1,*], Bo Friis Nielsen[1]

[1] Technical University of Denmark, Department of Applied Mathematics and Computer Science, DK 2800 Kgs. Lyngby, Denmark

Email: dame@dtu.dk

Abstract

The "Successive Principle" is an analysis method that has been widely applied in Scandinavia particularly for many transport and other infrastructure projects. We demonstrate how the calculations performed in the "Successive Principle" fit into the broad class of phase type distributions such that any value related to the distribution can be calculated. Particularly, one can calculate the standard deviation and upper quantiles exactly.

Our method can deal with lead times to compensate for the early start of sub-projects. Due to the phase type formulation we are able to calculate distribution properties simultaneously for durations and costs. We demonstrate our methodology by a case study that has previously been analysed in the Successive Principle framework. As the Successive Principle is based on Erlang 7 distributions we use these distributions in the example. However, any phase type distribution would be suitable. We also provide a small constructed example to illustrate the potential of simultaneous analysis of cost and duration.

Keywords*:* cost estimation, duration estimation, group analysis, phase type distribution, multivariate phase-type distribution, Successive Principle

[*] Corresponding author

1. Introduction

Forecasts for the duration and cost of infrastructure projects have been constantly and remarkably inaccurate as stated by Kahneman and Lovallo (Kahneman & Lovallo, 2003) and Wachs (Wachs, 1990). The suggested reasons are, e.g. optimism bias, anchoring (Kahneman & Lovallo, 2003) and political misrepresentation (Wachs, 1990). Throughout time this has been a repeating pattern. Flyvbjerg (Flyvbjerg, 2006) mentions and discards several reasons for this unchanging pattern in recent decades. As it is a well-known fact that forecasts are often inaccurate, it is a pressing research problem to develop tools for project management that can deliver accurate estimations for cost, demand, duration, and other variables.

Flyvbjerg (Flyvbjerg, 2006) suggests "Reference Class Forecasting" to avoid not only optimism bias but also political misrepresentation. The idea is to analyse projects based on similar, already conducted, projects. This is only applicable in cases where a reference already exists. Another solution is to consider group analysis which includes the experience of the group members and is not necessarily a strictly quantitative analysis. Hill (Hill, 1982) quotes several authors saying "group processes could lead to process gain". The idea of group analysis is to increase precision by knowledge sharing. It might be problematic to find appropriate specialists for the analysis or to prevent them from influencing each other. Another approach for obtaining project forecasts is to determine the uncertainties and then focus on the ones with major impact. Humphreys (Humphreys, 2005) associates the number of subtasks in a project that affect the overall outcome by a relevant percentage to Pareto's Principle (also known as the 80-20 rule) indicating that the majority of the overall uncertainty comes from very few items. Hillson (Hillson, 2003) uses a Risk Breakdown Structure to obtain a ranked list that identifies the most uncertain events and defines it as "A source-oriented grouping of project risks that organises and defines the total risk exposure of the

project". The Successive Principle (SP) (Lichtenberg, 2000) combines many of these ideas. The SP is a group analysis method that is widely used throughout Scandinavia. It divides projects in several subtasks. It is assumed that durations and costs of subtasks are Erlang distributed, more specifically Erlang 7 distributed. The method can be used for project cost, cost-benefit analysis and duration-related estimates. The goal is to avoid several pitfalls of a traditional analysis by using an inhomogeneous group of people guided by a professional. The Successive Principle has been of use internationally with most cases in Scandinavia for several decades and has been constantly improved by researchers and its users.

The main contribution of this paper is to apply the phase type methodology to the SP in order to obtain distributional properties beyond the mean. Being able to calculate the probability of delays is very important when assessing the risk of a project. The assumption of the SP that the duration of a subtask follows an Erlang distribution can be generalised to the assumption that durations follow general phase type (PH) distributions. The PH assumption allows for very general distributions enhancing the applicability of the SP particularly one can perform sensitivity analysis to see consequence of different distributional assumptions.

The phase type interpretation allows for the complete calculation of all distributional properties. In addition, we suggest a different approach for dealing with lead times in order to stay in the class of PH distributions. We use our approach to examine a real case that has previously been analysed with the SP. In order to make this comparison we use the same distribution used in the SP, the Erlang 7 distribution. Finally, we show how to use multivariate PH (MPH) distributions to model the dependence between cost and duration of a project.

2. PH and MPH distributions

The Erlang distributions used in the SP belong to the broader class of phase-type (PH) distributions.

PH distributions can be used to approximate any given distribution and are defined as the distribution of time until absorption in an absorbing continuous time Markov chain (Pinsky & Karlin, 2011) also known as a Markov jump process. To be more precise, let $X(t)$ be a Markov jump process on a discrete state space $E = \{1,2, \dots, m, m+1\}$ with $\mathrm{m}+1$ being an absorbing state and the states $\{1, \dots, \mathrm{m}\}$ being transient states. We then define $\tau = \min\{\mathrm{t} \geq 0\colon \mathrm{X(t)} = \mathrm{m}+1\}$ as the time of absorption. The distribution of τ is of phase type as termed by Neuts (Neuts, 1975). With T being the sub-intensity matrix of X(t) corresponding to the m transient states and α correspondingly being the initial distribution vector among the transient states, we write $\tau \sim \mathrm{PH}(\alpha, T)$. We call (α, T) a representation for the PH distribution. The sub intensity matrix T is of dimension $\mathrm{m} \times \mathrm{m}$ and α is of dimension m. Closed form solutions for the density, the cumulative distribution, as well as all moments exist.

In this paper we will present only fundamental results concerning PH distributions that are used in our modelling approach. For more results on PH distributions the reader is referred to (Neuts M. F., 1981) or (He, 2014).

Define $\mathbf{1}$ a column vector with ones in every entry, and $T^0 = (-T\mathbf{1})$ then the probability density function f(x) as well as the cumulative distribution F(x) and the expectation of τ can be written as

$$f(\tau) = \alpha e^{Tx} T^0 \quad (1)$$

$$F(\tau) = 1 - \alpha e^{Tx} \mathbf{1} \quad (2)$$

$$E(\tau) = -\alpha T^{-1} \mathbf{1}. \quad (3)$$

Erlang distributions are defined through two different parameters. The shape parameter k and the intensity parameter λ. Their probability density function as well as their cumulative distribution is well known

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \quad (4)$$

$$F(x) = 1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\lambda x} (\lambda x)^n. \quad (5)$$

Furthermore, let a random variable X be Erlang distributed with shape parameter k and intensity parameter λ, we then write $X \sim \mathrm{Er}(k, \lambda)$ and the $\mathrm{PH}(\alpha, T)$ representation can be given by

$$T = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda & \lambda \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix} \quad (6)$$

and $\alpha = e_1 = (1,0,\dots,0)$. The dimension of T is $k \times k$ and the dimension of α is k. PH distributions are closed under order statistics. Let $X \sim \mathrm{PH}(\alpha_X, T_X)$ and $Y \sim \mathrm{PH}(\alpha_Y, T_Y)$ then $\max(X, Y) \sim \mathrm{PH}(\alpha_{\max(X,Y)}, T_{\max(X,Y)})$ with

$$\alpha_{\max(X,Y)} = (\alpha_x \otimes \alpha_y) \quad (7)$$

$$T_{\max(X,Y)} = \begin{pmatrix} T_X \otimes I_Y + I_X \otimes T_Y & I_X \otimes T_Y^0 & T_X^0 \otimes I_y \\ 0 & T_x & 0 \\ 0 & 0 & T_Y \end{pmatrix}. \quad (8)$$

Here $\otimes$ is the Kronecker product (Neuts, 1975) and I_j is an identity matrix of proper dimension.

Assaf and Levikson (Assaf & Levikson, 1982) showed that the class of PH distributions is closed under the three basic operations finite mixtures, finite convolutions, and formation of coherent systems of independent components. A coherent system is a system that fails once a

certain number of components are not functional. They also showed that the subclass of PH distributions with an upper triangular sub generator matrix is closed under these operations. This means as long as the subtasks are Erlang distributed, we can always model the entire project as a PH distribution with an upper triangular sub-generator matrix. This has certain advantages, for example the certainty of a bounded number of states being visited as well as some numerical advantages when inverting the sub-generator matrix T. When analysing the dependence between duration and cost of a project we are dealing with a bivariate variable. Bivariate variables can be modelled using MPH distributions. Kulkarni (Kulkarni, 1989) defines MPH distributions based on univariate PH distributions. Kulkarni denotes the occupation times in the states $\{1, \dots, m\}$ of the absorbing Markov jump process $X(t)$ as $(H_1, \dots, H_m)$. He then constructs the multivariate vector $Y = (Y_1, \dots, Y_n)$ by using linear combinations of the occupation times. To be more precise $Y_i = \sum_{\{j=1\}}^{m} r_{\{ji\}} \cdot H_j = \int_0^\tau r_i\,(X(t))dt$, with $r_i(j) = r_{ij} \geq 0$ for $i \in \{1, \dots, n\}$ and $j = \{1, \dots, m\}$ being the weights of the linear combinations, in other words the reward collected in each state visited by the underlying Markov jump process. The vector Y is then said to be multivariate phase type*, $MPH^*(\alpha, T, R)$ distributed, where $R = (r_{ji})$ is called the reward matrix.

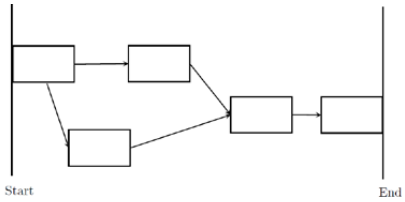
3. Modelling project subtasks using PH distributions

We will focus on the duration assessment as this methodologically is more demanding due to the effects of lead time and merge event bias (MEB). The term lead time refers to a subtask starting prior to the completion of its predecessor. The MEB is a correction term to compensate for the fact that the task with shortest expected duration will actually be the longest when one models the duration of two parallel subtasks by one single task. We will

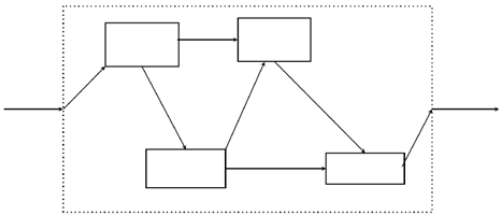
show how to use given parameters for the duration of each subtask to model the entire duration of a project using PH distributions. By using calculations for general PH distributions we completely eliminate the need for MEB as we model the projects evolvement explicitly. For the lead times we suggest an approach that keeps the model for the total duration in the class of PH distributions.

3.1 Modelling different structures arising in projects

A common way of representing the subtasks of a project and their interrelationship is using a state diagram like the one below.



Each box represents a subtask and one is particularly interested in the duration of these. Similarly, one can graphically describe PH distributions using state diagrams now with the requirement that the occupation time in each block is exponentially distributed. However, by grouping a number of exponential blocks together we can model the duration of a subtask using general PH distributions.

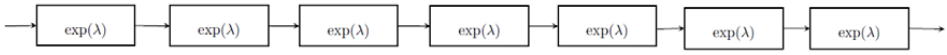


The PH distribution depicted above could for example model the duration of a subtask in the state diagram. Due to the coherent system property, assuming durations of subtasks to be PH distributed will immediately imply that the total duration of the project is also PH distributed.

The assumption in the SP is that the duration D_i of a subtask i is $Er(7,\lambda_i)$ distributed. We will represent a PH distribution with its representation (α_i, T_i) . For an $Er(7,\lambda_i)$ the representation can be given with $\alpha_i = (1,0,\dots,0)$ of dimension 7 and

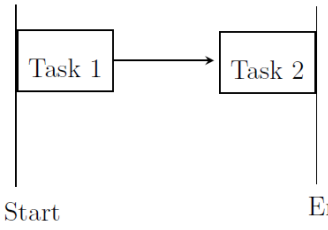
$$T_i = \begin{pmatrix} -\lambda_i & \lambda_i & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda_i & \lambda_i \\ 0 & 0 & \dots & 0 & -\lambda_i \end{pmatrix}. \qquad (9)$$

This can graphically be depicted as:



We will use these building blocks (α_i, T_i) to create a model for D the total duration of the project.

Let us consider a simple project consisting of two subtasks running successively

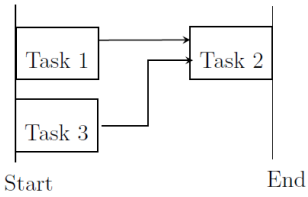


Obviously, $D = D_1 + D_2$ and with that $D \sim PH(e_1, T_D)$ whith $e_1 = (1,0\dots,0)$ of appropriate dimension and

$$T_D = \begin{pmatrix} T_1 & T_1^0 \alpha_2 \\ 0 & T_2 \end{pmatrix}. \qquad (10)$$

We can calculate the expected duration of the whole project as the sum of the expectations of the subtasks $E[D] = \alpha_D(-T)^{-1}\mathbf{1} = E[D_1] + E[D_2] = \alpha_1(-T_1)^{-1}\mathbf{1} + \alpha_2(-T_2)^{-1}\mathbf{1}$.

In general $\alpha_D(-T)^{-1}\mathbf{1}$ will not simplify similarly. If we consider a project where subtask 2 can only start once subtasks 1 and 3 are completed



i.e. after the maximum of two random variables, calculating the expected duration of the project is not as straight forward. This is due to the variability of the durations, making it possible that the subtask with the shorter expected duration takes longer time to complete. Different analysis methods, e.g. SP introduces a merge event bias MEB. The PERT approach does not directly account for the MEB, it can be approached by using simulation (Douglas & Cox, 1978) or numerical algorithms (Pontrandolfo, 2000). Kulkarni and Adlakha (Kulkarni & Adlakha, Markov and Marko-Regenerative Pert Networks, 1986) solve the problem by assuming exponentially distributed duration for subtasks and by constructing a Markov jump process.

This approach can be extended by allowing for general PH distributions and with that $D \sim PH(e_1, T_D)$ where

$$T_D = \begin{pmatrix} T_1 \otimes I_3 + I_1 \otimes T_3 & I_1 \otimes T_3^0 & T_1^0 \otimes I_3 & 0 \\ 0 & T_1 & 0 & T_1^0\alpha_2 \\ 0 & 0 & T_3 & T_3^0\alpha_2 \\ 0 & 0 & 0 & T_2 \end{pmatrix}. \qquad (11)$$

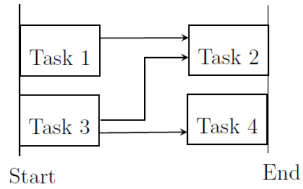
The matrix $T_1 \otimes I_3 + I_1 \otimes T_3$ represents the distribution of time until one of the subtasks 1 or 3 is completed. The entries $I_1 \otimes T_3^0$ and $T_1^0 \otimes I_3$ are the exit rates and ensure the continuing subtask will continue in the right state. Finally $T_1^0\alpha_2$ and $T_3^0\alpha_2$ ensure the proper start of subtask 2.

The expected time of the project E[D] is not simply the sum of the expected times of the subtasks as

$$E[D] = E[\max\{D_1, D_3\}] + E[D_2] \quad (12)$$
$$\neq \max\{E[D_1], E[D_3]\} + E[D_2]. \quad (13)$$

The inequality $E[\max\{D_1, D_3\}] \neq \max\{E[D_1], E[D_3]\}$ is what creates the need for the MEB in the SP. One advantage of modelling the entire project duration as a PH distribution is that no correction term is necessary.

We can extend the complexity by assuming task 2 can only start when tasks 1 and 3 are completed, but task 4 can start immediately when task 3 is completed

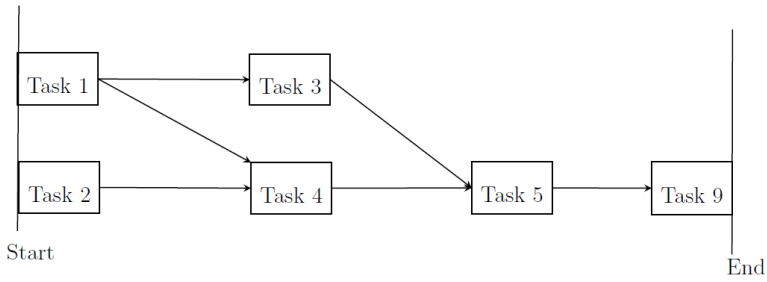


Clearly $\mathrm{E[D]} = \mathrm{E[\max\{D_3 + D_4}\,, \max\{D_1, D_3\} + D_2\}]$ The sub generator matrix can be given as

$$T = \begin{pmatrix} T_1 \otimes I_3 + I_1 \otimes T_3 & I_1 \otimes T_3^0 & T_1^0 \otimes I_3 & 0 & 0 & 0 & 0 \\ 0 & T_1 \otimes I_4 + I_1 \otimes T_4 & 0 & I_1 \otimes T_4^0 & T_1^0 \otimes I_4 & 0 & 0 \\ 0 & 0 & T_3 & 0 & T_3^0 \cdot (\alpha_2 \otimes \alpha_4) & 0 & 0 \\ 0 & 0 & 0 & T_1 & 0 & T_1^0 \alpha_2 & 0 \\ 0 & 0 & 0 & 0 & T_2 \otimes I_4 + I_2 \otimes T_4 & I_2 \otimes T_4^0 & T_2^0 \otimes I_4 \\ 0 & 0 & 0 & 0 & 0 & T_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & T_4 \end{pmatrix}. \quad (14)$$

Cases where a subtask has more than two predecessors can be modelled using the well-known fact $\max\{\mathrm{X_1}, \mathrm{X_2}, \mathrm{X_3}\} = \max\{\mathrm{X_1}, \max\{\mathrm{X_2}, \mathrm{X_3}\}\}$.

An example for the structure of such a project is



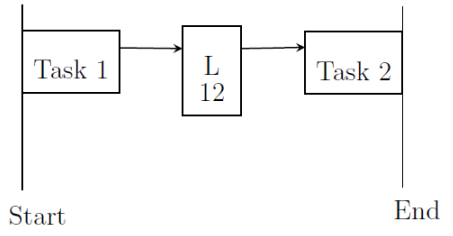
We can model the project using the sub generator matrix

$$\begin{pmatrix} T_1 \otimes I_2 + I_1 \otimes T_2 & I_1 \otimes T_2^0 & T_1^0 \otimes I_2 \otimes \alpha_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & T_1 & 0 & T_1(\alpha_3 \otimes \alpha_4) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & T_2 \otimes I_3 + I_2 \otimes T_4 & T_2^0 \otimes I_3 \otimes \alpha_4 & I_2 \otimes T_3^0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & T_3 \otimes I_4 + I_3 \otimes T_4 & 0 & I_3 \otimes T_4^0 & T_3 \otimes I_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & T_2 & 0 & T_2^0 \alpha_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & T_3 & 0 & T_3^0 \alpha_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & T_4 & T_4^0 \alpha_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_5 & T_5^0 \alpha_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_9 \end{pmatrix} \quad (15)$$

and e_1 as the initial probability distribution vector.

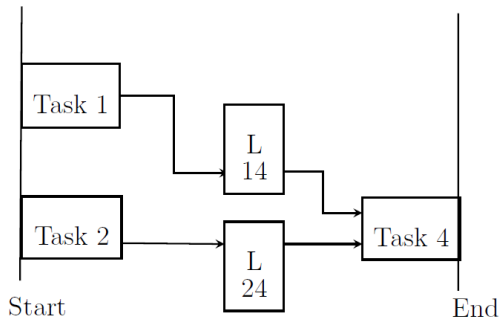
3.2 Modelling lead times

As mentioned earlier, it is not uncommon to assume that a subtask can start before the predecessor is completed. The event that subtask j can start before subtask i is completed is denoted with Lij and the duration of this event is given with L_{ij}. This time can be either deterministic or random. In the SP it is assumed that this time random. A graphical representation can be found below.



One obvious way of modelling the total duration would be by $D = D_1 - L_{12} + D_2$. However, in this case D can potentially be negative, as L_{12} can attain any positive real value. As we are considering the duration of a project it is desirable to have $P(D > 0) = 1$. As L_{12} reduces the contribution of D_2 to the total duration of the project an alternative way of modelling D is $D = D_1 + D_2 - \min\{D_2, L_{12}\}$ and with that $P(D > 0) = 1$. This way L_{12} cannot reduce the total duration by more than the duration of subtask 2, however with L_{12} and D_1 being random variables it is still possible that $D_1 < L_{12} < D_2$ which can be interpreted as subtask 2 beginning before subtask 1 has started. This is an unrealistic event that can be avoided by modelling the total duration as $D = D_1 + D_2 - \min\{D_1, D_3, L_{12}\}$. In this case D is again PH distributed as shown by Ahn and Ramaswami (Soohan & Ramaswami, Bilateral phase type distributions, 2005). However, the sub-generator will not be of upper triangular form.

If we have two predecessors to a subtask and both allow for lead times as in the following diagram



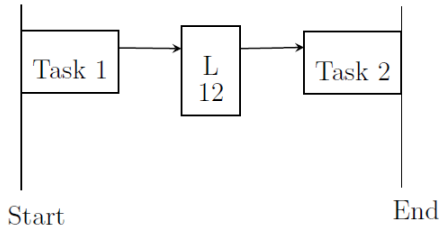
calculating the total duration is rather difficult. Examples can be found in (Lichtenberg, 2000) or in (Gong & Hugsted, 1993). Following the argument as before the total duration of the project is

$$D = \max\{D_1 + D_4 - \min\{D_1, D_4, L_{14}\}, D_2 + D_4 - \min\{D_2, D_4, L_{24}\}\} \quad (16)$$

$$= \max\{D_1 - \min\{D_1, T_4, L_{14}\}, D_2 - \min\{D_2, D_4, L_{24}\}\} + T_4 \quad (17)$$

Clearly the two terms we are taking the maximum of are not independent and therefore we are not able to state a general PH representation for D. It is not even clear if allowing more general PH distributions than Erlang distributions for the duration of subtasks that the maximum will stay in the class of PH distributions. In order to obtain more general results, we decided not to follow this path further but to suggest a different approach which will fit better to the simplicity that makes the SP appealing and will ensure an upper triangular sub generator matrix for the total duration of the project when the subtasks follow PH distributions generated by an upper triangular matrix.

If we go back to the model



we can reformulate the lead time L_{12} from being the time subtask 2 can start before completion of subtask 1, to $\tilde{L}_{12}$ being the time subtask 1 has to run after the start of subtask 2. Furthermore we have to define the time $\widetilde{D}_i$ until the first successor of subtask i can start. If subtask i has no successor it is the time until completion of the subtask. It is assumed that subtasks prior to a certain subtask without a lead time connection have to be completed before it can begin. With that we can model the total duration $\widetilde{D}$ and calculate

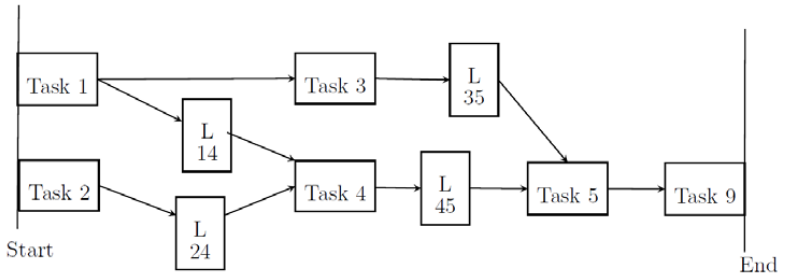
$$\tilde{D} = \tilde{D}_1 + \max\left\{\tilde{L}_{12} + \tilde{D}_2\right\} = \tilde{D}_1 + \max\left\{\tilde{L}_{12} + D_2\right\}. \qquad (18)$$

In this case we have $E[D] = E[\widetilde{D}]$ and $\widetilde{D}_i$ can either be directly estimated by the analysis group responsible for the parameter choices of the SP or by the values traditionally obtained through the SP. One further advantage is that this way of modelling reproduces the actual

behaviour of a project more accurately. In reality it is not possible to wait until the completion of a subtask, and from there derive how many time units ago the successor could have been initiated. Contrary a subtask will run and after a certain level of completion is reached the successor can start, while the predecessor still has to be completed and might face further delays.

4. Case Study: Modelling a project using PH distributions

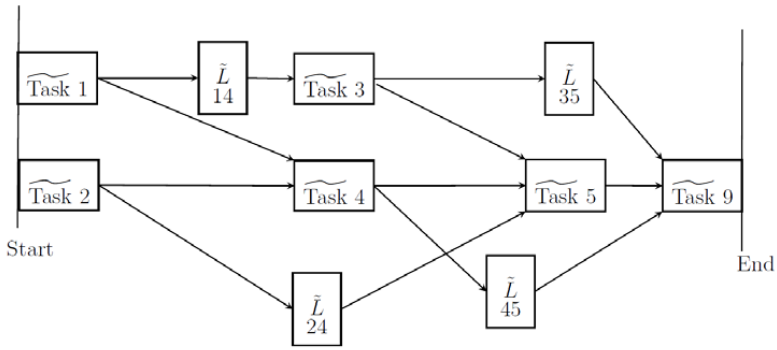
In 2000 Lichtenberg & Partners produced a schedule analysis for a large international IT development project using the SP. The group analysis identified the uncertainties of the project and concluded the following diagram represents the critical paths and near critical paths of the project



Every subtask except for subtask nine represents a physical part of the project. Subtask nine represents motivation or any other general conditions influencing the projects duration. Using this structure we have that the L_{ij}'s contribute with a negative value to the total duration of the project. More specifically the group came up with the following values

$$E[D_1] = 12.4\ E[D_2] = 15.8\ E[D_3] = 7\ E[D_4] = 6.4 \quad (19)$$
$$E[D_5] = 4.2\ E[D_9] = 2\ E[L_{14}] = -3\ E[L_{24}] = -3.2 \quad (20)$$
$$E[L_{45}] = -2.6\ E[L_{35}] = -1.9 \quad (21)$$

In order to only have subtasks that contributed with positive time to the duration of the project we will restructure the model to



with

$$E[\tilde{D}_1] = E[D_1] + E[L_{14}] = 9,4 \quad (22)$$
$$E[\widetilde{L}_{14}] = -E[L_{14}] = 3 \quad (23)$$
$$E[\tilde{D}_2] = E[D_2] + E[L_{24}] = 12.6 \quad (24)$$
$$E[\widetilde{L}_{24}] = -E[L_{24}] = 3,2 \quad (25)$$
$$E[\tilde{D}_3] = E[D_3] + E[L_{35}] = 5,1 \quad (26)$$
$$E[\widetilde{L}_{35}] = -E[L_{35}] = 1,9 \quad (27)$$
$$E[\tilde{T}_D] = E[D_4] + E[L_{45}] = 3,8 \quad (28)$$
$$E[\widetilde{L}_{45}] = -E[L_{45}] = 2,6 \quad (29)$$
$$E[\tilde{D}_5] = E[D_5] = 4,2 \quad (30)$$
$$E[\tilde{D}_9] = E[D_9] = 2. \quad (31)$$

To comply with the analysis performed by Lichtenberg & Partner we choose the specific PH distribution of the subtasks to be Erlang 7 distributed and calculate the intensity parameters based on the expectations of the original subtask. This will cause a slight loss of information and for further analysis we suggest to directly estimate the parameters for $\widetilde{D}_i$ and $\tilde{L}_{ij}$ during the process of the group analysis.

To simplify the sub generator T_D matrix we combine the subtask 3 and the lead time 14 to the task 143 with sub generator matrix

$$\tilde{T}_{143} = \begin{pmatrix} \tilde{T}_{14} & T_{14}^{0} \cdot e_1 \\ 0 & \tilde{T}_3 \end{pmatrix}. \tag{32}$$

Even for this simple project, when modelling the total duration of the project using the methods presented in this paper, the sub generator matrix will consist of 28 different combinations of the sub generator matrices for the independent subtask. The state space for the entire project contains 3353 states. It is not a problem to handle this numerically, however we decided not to state the sub generator matrix T_D.

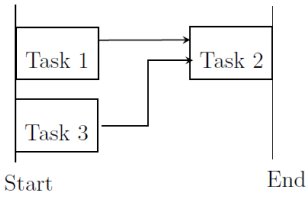
Numerically we can immediately get

$$E[D] = 25.9962 \tag{33}$$
$$Var[D] = 24.2235 \tag{34}$$
$$P(D > E[D]) = 0.4727 \tag{35}$$

as well all other moments, quantiles, and probabilities. The mean value obtained in the original analysis was 25.5 month. Our result is slightly higher; this can be explained by avoiding approximations of the MEB.

5. Correlation between cost and time

Frequently there is some dependence between the cost and the duration of a project. It even goes so far that the cost of construction projects has been used to estimate the necessary time for the project (Lichtenberg S. , 1974). Several methods accounting for the dependence in terms of cost and time have been reviewed by Ogunsemi and Jagboro (Ogunsemi & Jagboro, 2006). Let us again consider a simple project



each subtask generates cost due to the duration of the subtask, e.g. man hours and use of equipment, this allows us to model part of the projects cost related to the duration of each subtask.

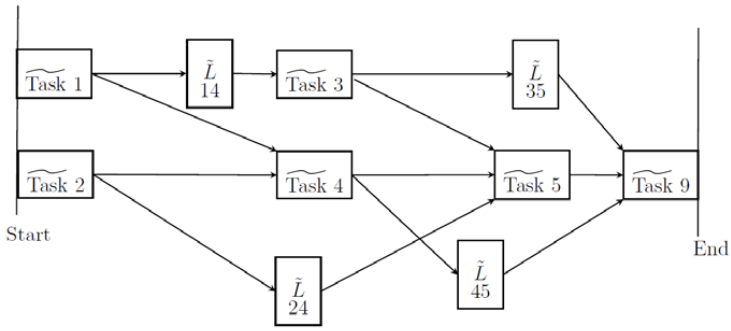
We define the time dependent cost of subtask i as $K_i = c_i \cdot D_i$, a constant multiplied with the duration of the subtask. Then the total time dependent cost C of the project is $C = \sum_i K_i$. The scalars c_i can be derived through several methods. One obvious choice is to estimate them during the group analysis session of the SP.

The total duration depending cost of the project with 3 subtasks is simply $C = K_1 + K_2 + K_3$. The interest now is in the duration as well as the time dependent cost. This leaves us with modelling the project as an MPH* distribution, with $\alpha = e_1$ and

$$T_D = \begin{pmatrix} T_1 \otimes I_3 + I_1 \otimes T_3 & I_1 \otimes T_3^0 & T_1^0 \otimes I_3 & 0 \\ 0 & T_1 & 0 & T_1^0 \alpha_2 \\ 0 & 0 & T_3 & T_3^0 \alpha_2 \\ 0 & 0 & 0 & T_2 \end{pmatrix}, \quad R = \begin{pmatrix} \mathbf{1} & \mathbf{1} \cdot (c_1 + c_3) \\ \mathbf{1} & \mathbf{1} \cdot c_1 \\ \mathbf{1} & \mathbf{1} \cdot c_3 \\ \mathbf{1} & \mathbf{1} \cdot c_2 \end{pmatrix}. (37)$$

With other words $R_{i1} = 1$. R_{i2} is the sum of the time dependent cost parameters c_j of the subtasks running simultaneous at this stage.

This extension can easily be applied to our remodelled project



The main work is already done, as it is to derive the sub generator matrix T_D. Adapting the reward matrix is straight forward. Having an MPH* model does not allow us to evaluate the density directly but only by numerical approximation. Nevertheless, we can calculate all moments as well as cross moments directly as shown by Bladt and Nielsen (Bladt & Nielsen, 2010). Furthermore simulation is straight forward and the marginal distributions are also PH distributions.

Small analytic example for analysing the dependence between cost and time

In order to further illustrate the approach we will consider an analytical example. Let us consider a small infrastructure project where a pedestrian bridge over a river is to be build. We are interested in the expected duration and cost of the project. Only material available without cost is being used. Therefore the only cost under consideration is working hours. The construction of the bridge consists of three parts, accruing the material, building the foundation and finishing of the bridge. Furthermore we assume these task have to be completed successively by different craftsmen with different wages, c_1, c_2, c_3. The durations undergo stochastic variation and for simplicity we assume they are exponentially distributed with intensities λ_1, λ_2 and λ_3. Therefore the MPH* representation can be given with

$$\alpha = (1,0,0), \qquad T = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & -\lambda_3 \end{pmatrix}, R = \begin{pmatrix} 1 & c_1 \\ 1 & c_2 \\ 1 & c_3 \end{pmatrix}$$

Let X be the duration of the project and Y the cost, the cross moment is then

$$E[X,Y] = \alpha(-T)^{-1}\Delta(R_1)(-T)^{-1}R_2 + \alpha(-T)^{-1}\Delta(R_2)(-T)^{-1}R_1$$

$$= \frac{2c_1}{\lambda_1^2} + \frac{2c_2}{\lambda_2^2} + \frac{2c_3}{\lambda_3^2} + \frac{c_1 + c_2}{\lambda_1\lambda_2} + \frac{c_1 + c_3}{\lambda_1\lambda_3} + \frac{c_2 + c_3}{\lambda_2\lambda_3}$$

Where $\Delta(R_i)$ is a diagonal matrix with the i-th column of R on the diagonal.

From the cross moment we can derive the covariance and the correlation.

6. Discussion

The SP consists of two separate parts; a data generating part and an analysis part. It is the analysis part we address in this paper. We have focused on data obtained from the SP, however, data could just as well have been obtained from other sources. In reality quality data on infrastructure projects is scarce and due to the nature no replications are at hand. See the UNITE publication by Nicolaisen (Nicolaisen, 2012) for a thorough discussion of the problems associated with data gathering.

In case one has data at hand the parameters of phase type distributions can be estimated using the Expectation-Maximisation method see Asmussen et al. (Asmussen, Nerman, & Olsson, 1996). As mentioned above data might not be available, however, the approach of (Asmussen, Nerman, & Olsson, 1996) can still be applied if one is able to suggest an initial distribution, which can then be fitted by minimising the so-called Kulback-Leibler distance expressing the degree of similarity between distributions The phase-type base approach can be used to improve sensitivity analyses and hence the accuracy of cost-benefit calculations which have been brought forward through Salling and Banister (Salling & Banister, 2009).

An important aspect of projects that with today's methods is hard to quantify is the environmental impact in correlation to construction time and cost. Furthermore the trend in society is that environmental impact is considered increasingly important. Properly quantified one should be able to include environmental impact in the multivariate phase type model.

However, it is doubtful that the duration of all subtasks can be described by the same distribution like the Erlang 7 distribution. It is a clear advantage to model an entire project as a PH distribution, as it allows for more general PH distributions for the durations of subtasks without any major changes. The only significant disadvantage of our modelling approach is the increased cost of modelling. The complexity of our model asks for a tool which takes the structure of a project and translates it into a transition matrix. This will be a necessity to make our modelling approach an everyday tool. In our opinion the cost of developing a tool is minor compared to the benefits project management can gain from it.

We have extended our model to the bivariate case, correlating duration and cost of a project.

If it is assumed that the duration and the cost of subtasks follow Erlang distributions it is possible to estimate the intensity parameters as well as the dependence between cost and duration using the expectation-maximization algorithm (Meisch, 2014).

So far analysing the depends between the cost and duration of a project has only been done theoretically and still has to prove its worth in real life.

One direction for further research is to implement the bivariate analysis of duration and cost and try it in several group analyses. The data obtained from these group analyses can then be used to revise our modelling approach.

7. Conclusion

We have generalised the assumption of the SP that the duration of a subtask in a project follows an Erlang distribution to model the total duration of a project using a PH distribution. This allows us to evaluate the density, quantiles as well as all moments of the duration of a project. In order to stay in the class of PH distributions, we have suggested a new way of dealing with lead times, which might be seen as more natural. As expected, compared to the classical approach of the SP, our model results in a higher expected total duration of a project. In turn we avoid the need for estimating the merge event bias and can calculate the distribution of the duration of the project directly. Furthermore, we have shown how to use a bivariate PH distribution to model the dependence between the cost of a project and the duration of its subtasks. We believe that our approach is a step towards making project forecasts more reliable.

Acknowledgment

David Meisch and Bo Friis Nielsen would like to thank the Danish Council for Strategic Research for the support through the UNITE project under grant 2140-08-0011. Furthermore we would like to thank Steen Lichtenberg for his support to our research.

References:

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistic*, *23*(4), 419-441.

Assaf, D., & Levikson, B. (1982). Closure of phase type distributions under operations arising in reliability theory. *The Annals of Probabilty* , 265-269.

Bladt, M., & Nielsen, B. F. (2010). Multivariate matrix-exponential distributions. *Stochastic Models*, *26*(1), 1-26.

Douglas, D. E., & Cox, J. L. (1978). PERT and simulation. *Conference on Winter simulation* (pp. 89-98). Miami Beach: IEEE Press.

Flyvbjerg, B. (2006). From nobel prize to project management: Getting risk right. *Project Management Journal*, *37*(3), 5-15.

Gong, D., & Hugsted, R. (1993). Time-uncertainty analysis in project networks with a new merge-event time-estimation technique. *International Journal of Project Management*, 165-173.

He, Q.-M. (2014). *Fundamentals of matrix-analytic metods.* New York: Springer.

Hill, G. W. (1982). Group versus individual performance: are n=1 heads better than one? *Psychological Bulletin, 91*(3), 517-539.

Hillson, D. (2003). Using a risk breakdown structure in project management. *Journal of Facilities Management, 2*(1), 85-97.

Humphreys, K. K. (2005). Conducting project risk analysis, how to do it and how not to do it. *ICEC International Roundup, 19*(2), 3-13.

Kahneman, D., & Lovallo, D. (2003). Delusions od Success. *Harvard buisiness review, 81*(7), 56-63.

Kulkarni, V. G. (1989). A new class of multivariate phase-type distributions. *Operation Research, 37*, 151-158.

Kulkarni, V. G., & Adlakha, V. G. (1986). Markov and Marko-Regenerative Pert Networks. *Operations Research*, 769-781.

Lichtenberg, S. (1974). *Dr. Technisis Thesis: Project Planning - a Third Generation Approach.* Kogens Lyngby: DTU.

Lichtenberg, S. (2000). *Proactive Management of Unertainty using the Successive Principle.* Kongens Lynbgy: PF Forlag.

Meisch, D. (2014). *PhD Thesis: Multivariate phase type distributions - Applications and parameter estimation.* Kongens Lyngby: Technical University of Denmark.

Neuts, M. F. (1975). Probability distributions of phase type. In R. Holvoet, *Liber Amicorum Professor Emeritus H. Florin* (pp. 173-206). Louvian: Katholieke Universiteit Leuven.

Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach.* Baltimore: The Johns Hopkins University Press.

Nicolaisen, M. S. (2012). *PhD Thesis: Forecasts: Fact or Fiction?* Aalborg: Department of Develpment and Planning, Aalborg University.

Ogunsemi, D., & Jagboro, G. (2006). Time-cost model for building projects in Nigeria. *Construction Management and Economics, 24*(3), 253-258.

Pinsky, M., & Karlin, S. (2011). *An introduction to stochastic modeling.* Burlington: Academic press.

Pontrandolfo, P. (2000). Project duration in stochastic networks by the PERT-path technique. *International Journal of Project Management*, 215-222.

Salling, K. B., & Banister, D. (2009). Assessement of large transport infrastructure projects: The CBA-DK model. *Transportation Research part A: Policy and Practice, 43*(9-10), 800-813.

Soohan, A., & Ramaswami, V. (2005). Bilateral phase type distributions. *Stochastic models, 21*(2-3), 239-259.

Soohan, A., & Ramaswami, V. (2005). Bilateral phase type distributions. *Stochastic models*, 239-259.

Wachs, M. (1990). Ethics and advocay in forecasting for puplic policy. *Business and Porfessional Ethics Journal, 9*(1/2), 141-157.