Universidad
Carlos III de Madrid
www.uc3m.es

# *TESIS DOCTORAL*

# *Maxentropic and Quantitative methods in Operational Risk Modeling*

**Autor:**

***Erika Gomes Gonçalves***

**Director/es:**

**Silvia Mayoral**

**Henryk Gzyl**

**DEPARTAMENTO DE ECONOMÍA DE LA EMPRESA**

Getafe, Junio de 2016

Universidad Carlos III de Madrid
www.uc3m.es

**TESIS DOCTORAL**

# Maxentropic and Quantitative methods in Operational Risk Modeling

***Autor:*** *Erika Gomes Gonçalves*

**Director/es:** **Silvia Mayoral**

**Henryk Gzyl**

Firma del Tribunal Calificador:

Firma

Presidente: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Secretario: (Nombre y apellidos)

Getafe, de de

Universidad Carlos III de Madrid (UC3M)
PhD in Business and Quantitative Methods.
Department of Business.

PhD THESIS

# MAXENTROPIC AND QUANTITATIVE METHODS IN OPERATIONAL RISK MODELING

**Author:** Erika P. Gomes Gonçalves
**Advisors:** Silvia Mayoral
Henryk Gzyl

June, 2016

# Contents

# Acknowledgements

This thesis would not have been possible without the support and assistance of a number of people. I would like to express my gratitude to my supervisors, Dr. Silvia Mayoral and Dr. Henryk Gzyl, for their thoughtful ideas and suggestions, without forgetting their invaluable support, during the course of research.

I gratefully acknowledge the funding received towards my PhD from the Department of Business (UC3M) and for their help in the successful ending of this thesis

Finally, I wish to thank my family and my friends for encouraging me and supporting me during the course of this thesis.

# Abstract

In risk management the estimation of the distribution of random sums or collective models from historical data is not a trivial problem. This is due to problems related with scarcity of the data, asymmetries and heavy tails that makes difficult a good fit of the data to the most frequent distributions and existing methods.

In this work we prove that the maximum entropy approach has important applications in risk management and Insurance Mathematics for the calculation of the density of aggregated risk events, and even for the calculation of the individual losses that come from the aggregated data, when the available information consists of an observed sample, which we usually do not have any information about the underlying process.

From the knowledge of a few fractional moments, the Maxentropic methodologies provide an efficient methodology to determine densities when the data is scarce, or when the data presents correlation, large tails or multimodal characteristics. For this procedure, the input would be the sample moments $E[e^{-\alpha S}] = \mu(\alpha)$ or some interval that encloses the difference between the true value of $\mu(\alpha)$ and the sample moments (for eight values of the Laplace transform), this interval would be related to the uncertainty (error) in the data, where the width of the interval may be adjusted by convenience. Through a simulation study we analyze the quality of the results, considering the differences with respect to the true density and in some cases the study of the size of the gradient and the time of convergence. We compare four different extensions of Maxentropic methodologies, the Standard Method of Maximum Entropy (SME), an extension of this methodology allows to incorporate additional information through a reference measure, called Method of Entropy in the Mean (MEM) and two extensions of the SME that allow introduce errors, called SME with errors or SMEE.

Although our motivating example come from the field of Operational Risk analysis, the developed methodology may be applied to any branch of applied sciences.

**Keywords:** maximum entropy methods, decompounding; density reconstructions; risk.

# List of Tables

# List of Figures

# Chapter 1

# Introduction

In banking and financial industries, during the last few decades, there has been a collective effort to develop a conceptual framework to characterize and quantify risk, in particular in Operational Risk, in order to reserve money to cover possible large-scale losses and ensure in this way the solvency and economic stability of the financial system. In the field of insurance modeling, the same effort has been realized to set premiums and optimal reinsurance levels, using historical data of claims or losses. The differences between these two sectors lies in the availability of data. Typically in insurance, the data is sufficient for a statistical description of the process, while, in the area of Operational Risk, the sample sizes are always small. So, depending on the given sample size the results may vary widely.

An important measure used in operational risk is Value-at-Risk, which estimates the worst possible loss event that can occur with a certain probability over a given time period. The regulations state that the allocated capital charge should correspond to a 1-in-1000 (quantile .999) years worst possible loss event, which is a much larger time period than the window over which banks have been collecting operational loss data (usually between five and fifteen years).

To calculate regulatory capital, we have to calculate the distribution of the losses so a quantitative model is required to model the peculiarities of the loss distribution: high kurtosis, severe right skewness and excessive heavy-tailedness. To model this kind of distributions is quite complicated due to the impossibility of modeling 'tail events' because in many cases no data is available to characterize the distribution in the tail, so there is the possibility that

the estimates understate the capital, that from a regulatory perspective should be avoided. On the other hand, overestimation is also an issue to take into account in banking, due to the opportunity costs of not investing the available capital. So, for the correct estimation of quantiles, we have to model the distribution of interest in the whole range of loss values.

Motivated by the aforementioned issues, we propose to model the total losses by maximizing an entropy measure. The Maxentropic methodologies that we will describe in this thesis can handle with the problems discussed before and they can provide a good density reconstruction over the entire range of values.

Entropy, in probability theory, can be interpreted as a measure of the degree of uncertainty or unpredictability of a probabilistic system, where information is the key element that drives the modeling. Anything we cannot explain is assigned as much uncertainty (maximum entropy) as possible. A good example comes from the field of finance where the hypothesis that markets respond to the level of uncertainty or the element of surprise is analogous to entropy. In the last two decades, this concept which has its original role in statistical physics, has found important applications in several fields, and this measure has been an important tool for portfolio selection, as a measure of the degree of diversification, and more widely in asset pricing, to tackle the problem of extracting asset probability distributions from limited data. Zhong et. al. (2013) make a review of the many applications of this concept in finance.

However, to extend this concept to other problems and fields, with different quantification processes other assumptions are required, according to the nature of the problem. Finding the correct density is quite important in insurance and risk applications, for example, for the estimation of Value at Risk. Historically, risk decision-makers have had a preference for methodologies that make some parametric assumptions about the sample distribution based on (incomplete) information that the economic theory provides, which usually do not fit well enough. In many cases, densities such as normal, log-normal, Pareto and gamma distributions are assumed to analyze behavior of the random variables of interest, that in most of the cases incorrectly modeled fat tails and asymmetries that are crucial to explain financial variables, without take into account economic or monetary politics changes, and problems related with the quality and availability of the data.

Entropy as a quantitative measure of ignorance or uncertainty can be maximized mathemat-

ically to find the probability distribution that best avoids unnecessary assumptions, without having to insert any structure or correlation beyond the required by the given constraints and that better represents the available information. This measure reaches its maximum value when we do not have any information or restrictions, and in this case (by the principle of insufficient reason of Laplace) the occurrence of the events are assigned to be equally likely (following an uniform distribution), but reaches its minimum and vanishes for known or certain events. In the case where we have partial information, the probability distribution that better model it, is the one that maximizes the entropy measure under certain assumptions or distribution characteristics formulated by constrains, in an optimization problem. The quality of the results will be linked to the chosen constraints, which will be responsible to update the probabilities from a previous distribution (the uniform).

This concept was used by Jaynes (1957) for the first time as a method of statistical inference in the case of an under-determined problem where it was necessary to choose the probability distribution of a discrete random variable $X$ between all possible distributions, when the available information was the expected value of $X$ and the fact that $p_i \leq 0$ and $\sum_i p_i = 1$, in this case there were infinitely many proper probability distributions satisfying this information. So, through the concept of Shannon entropy he selects the alternative that best represented the given information. The essence of our methodology falls within this framework.

The interest of this thesis lies into modeling the distribution of the accumulated damage or loss, namely random sums or compound model through maximum entropy methodologies. Computing such distributions is not a trivial problem, and it has received much attention for a long time. Textbooks as Panjer (2006), McNeil et al. (2005) and Rolski (1999) have done a big effort to collect and describe much of the methodologies used to calculate this kind of distributions in insurance and risk measurement, but Maxentropic techniques like those proposed here are not yet widely used.

The basic techniques used to obtain the aggregated distribution range from analytic, as for example methods that involve the direct inversion the Laplace transform, to numerical, as for example methods that consist in the calculation of convolution of discretized densities, or the use of Fourier transforms. The choice of the method is dictated by the specific objective to be achieved. In the case of operational risk where sometimes it is necessary to compute

the 0.999 quantile of the aggregate loss distribution, the accuracy and efficiency of the chosen method is an important issue. With maximum entropy methodologies some of the problems of misspecification, complexity and computational effort are considerably reduced when they are compared with the common approximations found in the literature.

Additionally, the techniques found in the literature works with the knowledge of two ingredients, the severity and the frequency of the risk events, the proposed methodology works in several levels with and without the knowledge or those elements, through the use of aggregated data, or even may be used to obtain the severity (or individual loss) distribution through the knowledge of the frequency and the aggregated losses.

In this thesis the motivating example comes from the field of risk analysis, but the statement of the problem might as well come from any branch of applied sciences. The estimation of the underlying loss process provides a starting point to design policies, set premiums and reserves, calculate optimal reinsurance levels and calculate risk pressures for solvency purposes in insurance and risk management. Also, this is useful in structural engineering to describe the accumulated damage of a structure, just to mention one more possible application.

We apply the methodology to a problem consisting of the determination of a probability density, where the available information is represented through a few values of its numerically-determined Laplace transform or, that is the same, the knowledge of a few of its fractional or non-integer moments of all orders (up to some measurement errors stemming from insufficient data). This is the novelty of the methodology, that is based on the results of Lin (1992) that will be explained in Chapter 2.

In, general, the methodology that is proposed to determine the distribution of the density $f_S$, may be summarized in two basic steps: first, set the constraint through the estimation of the Laplace transform $E[e^{-\alpha_i S}]$ for a few non-integer values of the transform parameter $\alpha$, and convert the problem of inversion of the Laplace transform into a fractional moment problem, that would content, as we indicate before, the available information. The second stage consists in solving the fractional moment problem by means of the maximum entropy method.

To complete the first step, the Laplace transform defined as an expected value is rewritten as

$$\psi(\alpha) = E[e^{-\alpha S}] = P(S = 0) + P(S > 0)E[e^{-\alpha S}|S > 0] \tag{1.1}$$

where $P(S=0)=P(N=0)=p_o$, and where $p_o$ has to be estimated from the data, or supplied by the analyst. Setting

$$\mu(\alpha)=\frac{\psi(\alpha)-p_o}{1-p_o}$$

thus, the problem becomes to determine $f_S$ from the integral constraint

$$E[e^{-\alpha S}|S>0]=\int_0^\infty e^{-x\alpha_j}f_S(x)dx=\int_0^1 y^{\alpha_j}f_Y(y)dy=\mu(\alpha_j),\quad j=0,...,K. \tag{1.2}$$

where the change of variables $y=e^{-x\alpha}$ is required and $K$ is the number of moments. Our distribution (assuming it exists) will be given by the constraint defined above. Then, the optimization problem consists in maximizing the entropy measure: $S(f)=-\int_0^1 f_Y(y)lnf_Y(y)dy$, satisfying the restriction (1.2). Thus, the essence of the maximum entropy is to transform the Laplace inverse problem into a problem of maximizing a concave function (the entropy) on some convex set.

The details related with the number of moments, values for alpha and the optimization process will be described in the next chapter.

Shall see later, in Chapter 2, that by introducing some changes into the formulation described before, we can introduce additional information about the distribution $S$, incorporating the subjective opinion of the modeler that may be though of as a first guess of the distribution that we are calculated.

As, we stated before, this thesis provides a thorough description of the Maxentropic methodology in the sensible case of operational risk. We explore this through three levels of aggregation. Here, each of the chapters of this thesis is related with one level of aggregation.

In order to have a better understanding of the problem, let us denote by $B$ an integer to be interpreted as the number of business lines of a financial institution in the context of operational risk. At the very last level (or third level) of aggregation the total losses are defined as

$$S=\sum_{b=1}^{B}S_b, \tag{1.3}$$

where for $b=1,...,B$ the (positive) random variable $S_b$ denotes the total risk in business line $b$. For each business line, let $H_b$ denotes the number of risk types contributing to $S_b$. Thus, the second level of aggregation is

$$S_b=\sum_{h=1}^{H_b}S_{b,h}, \tag{1.4}$$

where $S_{b,h}$ denotes the total loss in business line $b$ due to risk type $h$. Now, the most basic level of aggregation (or first level) follows from the standard actuarial modeling, also used in the advanced methodology in operational risk is

$$S_{b,h} = \sum_{k=1}^{N_{b,h}} X_{b,h,k} \tag{1.5}$$

where the $N_{b,h}$ are integer valued random variables denoting the frequency of events in the selected modeling time interval, and the $\{X_{b,h,k}, |k \geq 1\}$ are random variables denoting the losses (or severity) in the $k-$th individual risk event of type $h$ in business line $b$. The usual modeling assumption is that the $X_{b,h,k}$ are independent among themselves and $N_{b,h}$. But the $N_{b,h}$ may exhibit dependence among themselves.

This thesis is structured in six chapters. Chapter 2 starts describing in depth the mathematical aspects that lie behind the choice of the Maxentropic methods as methodology to solve the distributional problem discussed before, starting from the first level of aggregation and considering the case where we do not know the closed form of the Laplace transform of the compound sum. This case occurs, for example, when the frequency and severity are known and follow a Poisson - LogNormal distributions, respectively. Then, the calculation of the expected values only relies in the numerical form of the Laplace over the collected data, this poses the problem related with the sample errors that we try to reach in Chapter 3. The approaches discussed here are extended to the case when we do not have any information about the severity (individual losses) and we only have data of the total losses and the frequency to obtain such distribution.

The two methods that are used in this chapter are the Standard method of Maximum Entropy (SME) and the Maximum Entropy in the Mean (MEM), although equivalents, they differ on the choice of a prior or reference measure. This chapter explores the quality of the reconstructions obtained through those methods, and here we make the first attempt to decompose or disaggregate the random sums in its frequency/severity distributions, through a probabilistic inverse problem.

The solutions should be robust, so small changes in the input (the positive random variable $S$) should produce small changes in the reconstructed $f_S$. This is stated with a series of simulations. Additionally, the correct specification of the densities obtained are evaluated through techniques like PIT-based test of uniformity, Kolmogorov-Smirnov, Anderson-Darling tests and

Berkowitz (2001) tests. Regarding tests for independence, it is considered the Ljung-Box test, and a version of Berkowitz's (2001) test, in order to validate the absence of serial correlation in the PITs. To overcome the limitations of these tests (establish in the Appendix A.1) we extend our evaluations using visual tools as correlograms, reliability diagrams and marginal calibration plots, that along with the VaR and TVaR estimates and the decompounding procedure works as additional check-up of the quality of the results. Additionally, after fitting the model we analyze how the estimators perform on new data.

An interesting feature of the Maxentropic methodology is its potential to incorporate errors in the data. So, in Chapter 3, this is examined and two possible ways of deal with sample errors are compared, within the framework of the SME method, when the input is the numerical approximation of the Laplace transform. The differences between the two approaches lies in the input and the output of these methodologies. One of these methods provides an additional output that works as an estimator of the additive error.

In Chapter 4 is proposed a methodology that tries to solve a problem that is particularly common in firms in the early stages of operational risk data modeling, and occurs when the data collection procedure does not distinguish between sub-populations or sources of risk and either between the size and number of events of a compound model. The methodology that is called decompounding in Chapter 2 is extended here to separate the risk sources and identify the statistical nature of the frequency and severity through the use of maximum entropy. This is combined with frequency models of the type $(a, b, 0)$ and clusters algorithms like k-means and expectation-maximization algorithm. In general, this is an intractable problem because there is often insufficient information in the aggregate data. In this chapter we tackle the problem by first attempting to determine the frequency of individual risk factors and subsequently estimating the distribution of the severity through a two step procedure. This is an application that can be a topic of great interest to financial regulators, who often, only have access to highly aggregated data.

Upon reaching Chapter 5 we will have concluded that maximum entropy based techniques are powerful tools to determine density distributions. Here, the pending question is: What happen when the data is scarce, as is common in baking?. At this point we know that the Maxentropic density depends upon the data used. In this chapter we analyze the variability of

the reconstructed densities to examine the impact of this variability on the estimation of VaR or the TVaR, favored tool of financial regulators for bank capital adequacy, and extensively used in fields such as insurance and fund management. The results showed that even when the quantity of data is small, the Maxentropic methods correctly fit the tails of the distribution, where the less frequent events occurs and where the fit is important to calculate risk measures, like VaR and TVaR. Although, the uncertainty is an important issue, we could see that when re-sampling is a possibility the fit is quite good. Finally, in Chapter 6 we briefly discuss how the Maxentropic method could be used to capture dependencies between random sums in the last level of aggregation. The results are compared with a standard procedure (the convolution method) in order to proof how the maximum entropy performs easily with less effort. Lastly, in Chapter 7, we conclude and summarize the main findings of this thesis.

# Chapter 2

# Two Maxentropic approaches to determine the probability density of compound risk losses

## 2.1 Introduction

This chapter introduce the theoretical aspects that are behind two different approximations of the Maxentropic procedures used to determine the probability distribution of a compound random variable, that describes the accumulated or total number of losses generated in Operational risk and/or in the insurance industry, using as input only a finite number of empirically determined fractional moments of the exponential. The two methods that are developed in this chapter are the Standard method of Maximum Entropy (SME) and the method of Maximum Entropy in the Mean (MEM), that would be also used to develop, through an inverse probabilistic procedure, the distribution of the individual losses from the knowledge of the compound loss. In this chapter we start describing the first step of the methodology, this is, the definition of the constraint used for the optimization problem, introduced in Chapter 1. After that we describe in detail the methodology.

To be specific, suppose that the frequency of losses in a given period of time is described by a compound random variable of the type $S = \sum_{j\geq 0}^{N} X_j$, where $N$ is a Poisson random variable of intensity $\ell$ with $G(z) = \sum_0^\infty z^n P(N = n)$ the generating function of the frequency of the

events $N$, and $\{X_j,\ \text{for } j = 1, ..., N\}$ denote the individual losses which are independent and identically distributed, where $\phi(\alpha) = E[e^{-\alpha X}]$ is its Laplace transform.

From an abstract point of view, the implementation of the Maxentropic methods fall within the techniques to invert Laplace transform from a few values of the parameter along the real axis. Then, the starting point will be the Laplace transform of $S(N)$ (or $S$ for short)

$$E[e^{-\alpha_i S}] = \psi(\alpha_i) = \int_0^\infty e^{-\alpha_i s} dF_S(s), \quad i = 1, ..., K. \tag{2.1}$$

which will be calculated analytically, through the probability generation function $G(\phi(\alpha))$, in the case where we know the distribution of the underlying process, or numerically, through the calculation of the sample fractional moments.

If we set $Y = e^{-S}$, we transform the Laplace inversion problem into the problem of inferring the density $f_Y(y)$ from its $\alpha_i$ fractional moments, where $\alpha_i$ is a real number. For that, we think of the previous identity as follows

$$\psi(\alpha_i) = E[Y^\alpha] = \int_0^1 y^{\alpha_i} dF_Y(y), \quad i = 1, ..., K, \tag{2.2}$$

Note now that, since the distribution $F_S$ of $S$ has a point mass $p_o$ at $S = 0$, in order to relate the $\psi(\alpha)$ to the density $f_Y(y)$ of $Y$, it is necessary to condition out the mass at $\{Y = 1\}$ (or the mass of $F_S$ at $\{S = 0\}$). For that consider the conditional version

$$E[e^{-\alpha_i S} \mid \quad S > 0] = \int_0^1 y^{\alpha_i} f_Y(y) dy = \frac{\psi(\alpha_k) - p_o}{1 - p_o} := \mu(\alpha_i), \quad i = 1, ..., K. \tag{2.3}$$

which defines the $\mu(\alpha_i)$ that will be the input for the Maxentropic methods. Once $f_Y$ has been determined, in order to recover $f_S$ it is necessary to apply the change of variables $y = e^{-s}$ to obtain $f_S(s) = e^{-s} f_Y(e^{-s})$.

In Table 2.1 we give some examples of analytic expressions for the $\alpha$ order fractional moments. Here we observe the case for a simple Gamma distribution and a Poisson-Gamma compound distribution. Through the calculation of analytically fractional moments we can mimic any of the most used probabilistic models, as Gamma, Exponential, Weibull as well as any compound distribution.

Nevertheless, in the case where the individual losses follow a lognormal distribution, Laplace transform techniques are hard to implement because, to begin with, the Laplace transform of

| Distribution | Probability density function | Parameters | $\psi(\alpha)$ |
|---|---|---|---|
| Gamma | $f_X(x) = b^a x^{a-1} e^{-bx}/\Gamma(a)$ | $a > 0,\ b > 0$ | $b^a(\alpha + b)^{-a}$ |
| Poisson-Gamma | $f_X(x) = b^a x^{a-1} e^{-bx}/\Gamma(a)$ | $a > 0,\ b > 0$ | $exp((-\ell((1 - b^a(\alpha + b)^{-a})))$ |
| | $p_k = \frac{e^{-\ell}\ell^k}{k!}$ | $k \geq 1$ | |

Table 2.1: Laplace Transform of Gamma & Poisson-Gamma distributions

a compound lognormal density is unknown. To wit, consider the effort to compute it approximately carried out by Leipnik (1991). Thus, the example that it is considered in this chapter is such that a direct analytical solution is not known, and the standard numerical procedures to determine them are hard to implement. Here, we propose a direct procedure to solve the problem of determining the density of the total sum of lognormal individual losses when the frequency is of the Poisson type through the use of a simulated sample.

Besides that, the importance of the lognormal distribution lies on that is frequently used to model individual claims in various classes of insurance business and in risk theory to model losses caused by different risk events. The fact that it has a heavy tail is important, because it allow us consider the possibility of describing very large claims, which correspond to losses that threaten the solvency of an insurance company or a bank. This has important implications for the determination of premiums, risk reserves and reinsurance (Crow et al., 1988).

Thus, for this case and any other where there is not a close form of the Laplace transform, the required value has to be estimated from observed values of the aggregated sums. It is at this point where errors may enter to the model due to problems with the data collection process. So, if samples of a random variable are available, the unbiased estimator of the fractional moments would be:

$$E[e^{-\alpha_k S}] = \psi(\alpha_k) = \frac{1}{T}\sum_{t=1}^{T} e^{-\alpha_k S_t} \tag{2.4}$$

where $k = 1, ...K$, $t = 1, 2, ..., T$, and $T$ denotes the size of the data collected over some long period of the data collection. In the numerical examples we considered a sample size big enough for not worry about the statistical errors. The potential errors in input would be considered in the next chapter.

In the next sections, two versions of maximum entropy procedures will be described and

used to determine the distribution of the individual losses from the knowledge of the total losses $S$. Also, we will explore the case where there is a historical record of total losses and a model for the frequency, that allow us, through Maxentropic methodologies, to decompound (or to disaggregate) the distribution of total losses ($S$) to obtain the distribution of individual losses ($X$). This could be useful for a risk manager that may want to know it in order to apply any particular corrective loss prevention policy.

In summary, this chapter is organized as follows: in Section 2.2 we discuss the basis behind the choice of the values of $\alpha_k$ and the number of moments $K$ for the calculation of the fractional moments of equation (2.3), later we describe briefly the basic details of the SME and MEM methods in Section 2.3. In Section 2.4, we show the results of the implementation of the SME and MEM approaches to determine the distribution of total losses. Section 2.5 is devoted to the computation of two of the most commonly used risk measures, namely Value-at-Risk or $VaR$ and the Tail-Value-at-Risk or $TVaR$ using the Maxentropic density as loss probability density. This could be interesting for risk managers who may consider insuring operational risk losses to decrease the capital charges. Section 2.6 is devoted to the complementary problem of decompounding, by means of which the distribution of individual losses is obtained from the total loss. In Section 2.7, we present some concluding remarks. Finally, in the Appendix A.1 of this thesis we provide a quick overview of the statistical tests, graphical tools and error measures used to verify the robustness and the quality of the results.

## 2.2 The fractional moment problem

In this section with deal with the issue of determine which and how many moments to consider as input of the Maxentropic methodology. Here we discuss only the basis and more important results of the moment problem, but a more complete discussion can be found in Gzyl & Tangliani (2012) and Gzyl et al. (2013).

The idea behind the choice of the fractional moments is in the work of Lin (1992). He asserts that the knowledge of an infinite sequence of non-integer or fractional moments determines the underlying distribution. This result is establish through the following two theorems:

*Theorem 2.1.* Let $F_Y$ be the distribution of a positive random variable $Y$. Let $\alpha$ be a sequence of positive and distinct numbers in $(0, A)$ for some A, satisfying $lim_{k\to\infty}\alpha_k = \alpha_0 < A$. If $E[Y^A] < \infty$ , the sequence of moments $E[Y^{\alpha_k}]$ characterizes $F_Y$.

*Theorem 2.2.* Let $F_Y$ be the distribution of a random variable $Y$ taking values in $[0, 1]$. Let $\alpha_k$ be a sequence of positive and distinct numbers satisfying $lim_{k\to\infty}\alpha_k = 0$ and $\sum_{n\geq 1}\alpha_k = \infty$. Then the sequence of moments $E[Y^\alpha]$ characterizes $F_Y$.

Thus, all the information about our problem can be compressed in a set of moments obtained through the Laplace transform. The selection of those moments should be in a way that we use only those that are the more relevant or informative.

According to Lin's theorems we have two alternatives for the choice of the $\alpha_i$ powers of equation (2.3). For any $\alpha_i$ the sequence $\alpha_i = (1/K)\alpha_0$ or $\alpha_i = (1-1/K)\alpha_0$ meet the conditions. $\alpha_0$ may be integer or real.

Now, it is necessary to select the optimal number of finite fractional moments to use. This can be solved through Theorem 2.3, also called Entropy-convergence Theorem.

*Theorem 2.3.* Let $\{\alpha_j\}_{j\geq 0}$ be a sequence satisfying Lin's theorem 2.2 and let $f_K^*$ be the optimal solution. Then the Maxentropic approximation $f_K^*$ converges in entropy to $f_Y$ in the sense that

$$lim_{K\to\infty}S[f_K^*] = S[f_Y]$$

where

$$S[g] := -\int_0^1 f_Y(y)lnf_Y(y)dy$$

denotes the entropy of a probability density $g$, and $0 \geq S[f_1] \geq .., . \geq S[f_K] \geq ... \geq S[f]$, i.e. the uncertainty decreases conforming $K$, the number of moments, increases.

To keep our explanation simple, we combine Theorem 2.3 with Pinker's inequality to assert that

$$\int_0^1 |f_K^*(y) - f_Y(y)|dy \leq \sqrt{2(S[f_K^*] - S[f_Y])} \to 0$$

as $K \to \infty$.

Clearly, an arbitrarily large number of moments $K$ decrease the distance between the optimal solution $f_K^*(y)$ and the real density $f_Y(y)$. So, it is clear that after a specific number of moments, the additional information introduced is not visible, this is strictly related with the computer arithmetic. Thus, it is useless consider moments of orders higher than a certain value. Experimentally the number of moments $K$ that best compress the information is eight. See Gzyl & Tangliani (2010) to understand the mathematical details behind these theorems.

Here, it is important to highlight why we use fractional moments instead of integer moments. The entropy-convergence theorem gives a good criteria to determine the optimal number of moments. The results in Gzyl & Tangliani (2010), showed through the observation of the difference $S[f_K^*] - S[f_Y]$, that the effect of the integer moments on the decrease of the entropy require more moments of higher order in comparison with the fractional moments. This means, that we require almost the double of integer moments to find a decrease of the order $1 \times 10^{-2}$ in the entropy. Thus, a good density reconstruction can be obtained with a reduced number of fractional moments.

## 2.3 The Maxentropic approaches

This section establish the basis of the SME and MEM methods used to solve the problem of finding the density of the total severity from the knowledge of a small number of fractional moments. Both Maxentropic procedures yield similar and good quality results. The procedure of maximum entropy in the mean is more general and contains the standard maximum entropy method as particular case, and it is useful to know about its existence.

### 2.3.1 The Standard method of Maximum Entropy (SME)

This is a variational procedure, based on the results of Jaynes (1957), to solve the (inverse) problem consisting of finding a probability density $f_Y(y)$ (on $[0, 1]$ in this case), satisfying the

following integral constraints:

$$\int_0^1 y^{\alpha_k} f_Y(y)dy = \mu_Y(\alpha_k) \qquad \text{for} \qquad k = 0, 1, ..., K. \tag{2.5}$$

Here we set $\alpha_0 = 0$ and $\mu_0 = 1$ to take care of the natural normalization requirement on $f_Y(y)$. The intuition is rather simple: The class of probability densities satisfying (2.5) is convex. One can pick up a point in that class one by maximizing a concave functional (an "entropy") that achieves a maximum in that class. That extremal point is the "Maxentropic" solution to the problem. In our case, the functional is given by

$$S(f) = -\int_0^1 f(y) ln f(y) dy$$

When the problem has a solution can be expressed in terms of the Lagrange multipliers as

$$f_K^*(y) = \exp\left(-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}\right) \tag{2.6}$$

in which the number of moments $K$ appears explicitly and $\boldsymbol{\lambda^*} = (\lambda_1^*, ..., \lambda_K^*)$ is a $K-$dimensional vector. Clearly, the generic form of the normalization factor is given by

$$Z(\boldsymbol{\lambda}) = \frac{1}{e^{-\lambda_0}} = \int_0^1 e^{-\sum_{k=1}^{K} \lambda_k y^{\alpha_k}} dy. \tag{2.7}$$

With this notation the generic form of the solution looks like

$$f_K^*(y) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\sum_{k=1}^{K} \lambda_k^* y^{\alpha_k}} = e^{-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}}. \tag{2.8}$$

To complete, it remains to specify how the vector $\boldsymbol{\lambda}^*$ can be found. For that one has to minimize the dual entropy:

$$\Sigma(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \ln Z(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{\mu}_Y \rangle \tag{2.9}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the standard Euclidean scalar product and $\boldsymbol{\mu}$ is the $K-$vector with components $\mu_k$, and obviously, the dependence on $\boldsymbol{\alpha}$ is through $\boldsymbol{\mu}_Y$. Thus, the Lagrangian encapsulates all the information of the primal problem.

When the solution of this dual problem exists, we have

$$S(f^*) = -\int_0^1 f^*(y) \ln f^*(y) dy = \Sigma(\boldsymbol{\lambda}^*, \boldsymbol{\mu}) = \ln Z(\boldsymbol{\lambda}^*) + \langle \boldsymbol{\lambda}^*, \boldsymbol{\mu}_Y \rangle \tag{2.10}$$

### 2.3.2 The method of Maximum Entropy in the Mean (MEM)

The MEM provides another interesting approach to solve the problem of determining $f_Y(y)$ such that equation (2.2) holds. It can be summed up by saying that it consists in a technique to obtain $f_Y(y)$ in its discretized version, as the expected value of an auxiliary probability distribution which is determined by an entropy maximization procedure. To implement MEM numerically, the first step consists of discretizing the problem. To discretize (2.5) set $x_j = (1/M)f((j-1)/M)$ and $A_{i,j} = (\frac{2j-1}{2M})^{\alpha_i}$, for $j = 1, ..., M$. This leads to a linear system of equations like:

$$\sum_{j=1}^{M} A_{i,j}x_j = \mu_i, i = 0, 1, ..., K \text{ with } x_j \geq 0 \text{ and } j = 1, ..., M. \tag{2.11}$$

The first factor in front of the definition of $x_j$ comes from the discretization $dy \approx 1/M$, and $A_{i,j}$ is obtained as the midpoint approximation of $y^{\alpha_i}$ for the chosen partition. We have added a normalization constraint by choosing $\alpha_0 = 0$. Note that with this choice we have $A_{0,j} = 1$ for $j = 1, ..., M$ and $\sum_{j=0}^{M} x_j = \mu_0 = 1$ as normalization condition for $\alpha_0 = 0$. Observe that if we consider a partition of size 200, as we have $K + 1$ moments, we have a system of $K + 1$ equations to determine 200 unknowns subject to a positive constraint. Actually, the original problem consists of $K + 1$ equations to determine a continuous function, so the discretization may seem to be an improvement from the dimensionality point of view.

To continue, to take care of the positivity constraint, consider a space $\Omega = [0, \infty)^M$ with its Borel sets $\mathcal{F}$. Denote by $\boldsymbol{\xi} = (\xi_1, ...\xi_M)$ a generic point in $\Omega$ and define the coordinate maps $X_j(\boldsymbol{\xi}) : \Omega \to [0, \infty)$ by $X_j(\boldsymbol{\xi}) = \xi_j$. On $(\Omega, \mathcal{F})$ place a reference measure $dQ(\boldsymbol{\xi})$ and we search for a measure $dP = \rho dQ$ such that

$$\sum_{j=1}^{M} A_{i,j}E_P[X_j] = \mu_i \text{ for } i = 0, 1, ..., K. \tag{2.12}$$

The choice of the measure $Q$ is up to the modeler, and it may be thought of as the first guess of $P$. The only restriction upon it is that the convex hull generated by its support is $\Omega$. The purpose of that condition is to ensure that any strictly positive density $\rho(\boldsymbol{\xi})$ of $P$ respect to $Q$ is such that

$$E_P[X_j] = \int_\Omega \xi_j \rho(\boldsymbol{\xi}) d\boldsymbol{\xi} \in [0, \infty),$$

that is, the positivity constraint is automatically satisfied. The other constraints will be achieved by a special choice of $P$. With the notations introduced before, it is easy note that the class of probabilities

$$\mathcal{P} = \{P << Q \text{ such that (2.12) holds true}\}$$

is a convex, closed set if not empty, which we suppose to be the case. On this set we define the entropy function

$$S_Q(P) = -\int_\Omega \rho(\boldsymbol{\xi}) \ln\left(\rho(\boldsymbol{\xi})\right) dQ(\boldsymbol{\xi})$$

whenever the integral is finite or $\infty$ otherwise.

$$\text{Find } P^* \in \mathcal{P} \text{ which maximizes } S_Q(P). \tag{2.13}$$

From now on the routine is pretty much as in the previous section. It is clear that the MEM uses the SME as a stepping stone. The problem is similar but in another setup. The generic solution is of the type

$$\rho^*(\boldsymbol{\xi}) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\langle \boldsymbol{A^t}\boldsymbol{\lambda}^*, \boldsymbol{\xi}\rangle} \tag{2.14}$$

where, again $\boldsymbol{\lambda}^*$ is obtained minimizing the dual entropy function

$$\Sigma(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \ln Z(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle, \tag{2.15}$$

where, recall, $\boldsymbol{\lambda}$ is a $K$dimensional vector, and $\boldsymbol{\mu}$ is a $K$dimensional vector of constraints defined (2.11). This time, the function $Z(\boldsymbol{\lambda})$ is defined by

$$Z(\boldsymbol{\lambda}) = \int_\Omega e^{-\langle \boldsymbol{A^t}\boldsymbol{\lambda}^*, \boldsymbol{\xi}\rangle} dQ(\boldsymbol{\xi}).$$

That (2.14) is the density that maximizes the entropy when the minimum of the Lagrange-Fenchel dual of the entropy, given by (2.15), is achieved in the interior of its domain, is a standard result in duality theory presented in Chapter 4 of Borwein and Lewis (2000).

#### 2.3.2.1 Exponential reference measure

Set $\Omega = [0,\infty)^M$ and on $\Omega$ we put a reference measure $Q$ defined by

$$dQ(\xi) = \prod_{j=1}^{M} \zeta e^{-\zeta\xi_j}$$

where $\zeta$ is set by the modeler. Where the normalization factor is

$$Z(\lambda) = E_Q[e^{<\lambda,AX>}] = \prod_{j=1}^{M}\left(\int_0^{\zeta} e^{-(A^t\lambda)\xi_j}\zeta_j e^{-\zeta\xi_j}d\xi_j\right) = \prod_{j=1}^{M}\frac{1}{(\zeta+(A^t\lambda)_j)}$$

Notice that the first order condition for the extremality of $\lambda^*$ reads like

$$\frac{\partial}{\partial\lambda_i}\Sigma(\lambda) = 0 = -\sum_{j=1}^{M} A_{i,j}\frac{1}{(\zeta+(A^t\lambda^*)_j)} + \mu_i$$

which clearly indicates that the Maxentropic estimator of the solution $x_j$ to (2.11) is

$$x_j^* = \frac{1}{(\zeta+(A^t\lambda^*)_j)} \tag{2.16}$$

This is the particular case of the gamma reference measure with $\varrho = 1$. When $\varrho > 1$ the reference measure is

$$Q(\xi) = \prod_{j=1}^{N}\frac{\zeta^{\varrho}\xi_j^{\varrho-1}e^{-\zeta\xi_j}}{\Gamma(\varrho)}(d\xi_j)$$

and the solution is $x_j^* = \frac{1}{(\zeta+(A^t\lambda^*)_j)^{\varrho}}$

#### 2.3.2.2 Poisson reference measure

We shall illustrate the MEM when the reference measure is a Poisson distribution. Recall that the $x_j$ are to be estimated as expected values with respect to a probability $P$ to be determined as we explained before. As said, in each interval of the partition we consider the following reference measure:

$$q(d\xi) = e^{-\eta}\sum_{k\geq 0}\frac{\eta^k}{k!}\epsilon_{\{k\}}(d\xi).$$

Here we use $\epsilon_{\{a\}}(d\xi)$ to denote the unit point mass (Dirac delta) at $a$. Certainly the convex hull of the set non-negative integers is $[0,\infty)$. Notice that the Laplace transform of the Poisson distribution is $\exp(-\eta(1-e^{-\alpha}))$. Thus, the normalization factor now is

$$Z(\boldsymbol{\lambda}) = \prod_{j\geq 0}^{M}\exp\Big(-\eta\Big(1-e^{-(\boldsymbol{A^t\lambda})_j}\Big)\Big)$$

from which we obtain

$$\Sigma(\boldsymbol{\lambda}) = -\eta \sum_{j-1}^{M} \left(1 - e^{-(\boldsymbol{A^t \lambda})_j}\right) + < \boldsymbol{\lambda}, \boldsymbol{m} >$$

Notice now that if $\boldsymbol{\lambda}^*$ minimizes that expression, then the estimated solution to (2.11) is

$$x_j^* = e^{-(\boldsymbol{A^t \lambda}^*)_j} \tag{2.17}$$

Do not forget, $\boldsymbol{A^t \lambda}_j = \sum_{i=0}^{8} \lambda_i A_{i,j}$. Recall as well that $A_{0,j} = 1$ for $j = 1, ..., M$. As $\sum_{j=0}^{M} x^*(j) = 1$, we can rewrite (2.17) as

$$x_j^* = \frac{e^{-(\hat{\boldsymbol{A}}^{\boldsymbol{t}} \hat{\boldsymbol{\lambda}}^*)_j}}{z(\hat{\boldsymbol{\lambda}}^*)} \tag{2.18}$$

where we redefined $\hat{\mathbf{A}}$ as the matrix obtained from $\mathbf{A}$ by deleting the zero-th row, and $\hat{\boldsymbol{\lambda}}$ as the $8-$dimensional vector obtained by deleting $\lambda_0$. To complete,

$$z(\hat{\boldsymbol{\lambda}}^*) = e^{-\lambda_0} = \sum_{j=1}^{M} e^{-\left(\hat{\boldsymbol{A}}^{\boldsymbol{t}} \hat{\boldsymbol{\lambda}}^*\right)_j}.$$

Recall that $x_j^* = (1/M) f_Y^*((j-1)/M)$, from which we determine $f_Y^*(y)$ approximately by interpolation.

## 2.4 Numerical examples

The examples that are considered in this section consists of a compound random variable in which the frequency of events during a given period of time is described by a Poisson random variable $N$ of parameter $\ell$, and the i.i.d. individual losses $X_j, j \geq 1$ are distributed according to a common lognormal distribution ($X \sim logN(\mu, \sigma^2)$. In this case the Laplace transform of $S(N)$ cannot be computed analytically, and has to be estimated numerically.

We proceed as follows:

1. We generate a sample of size 8.000 from a compound distribution $S(N) = \sum_{1 \leq j}^{N} X_j$, where $\{N \geq 0 | \quad N = n_1, n_2, \ldots, n_n\}$, are the sizes of the lognormal samples with mean $\mu$ and standard deviation $\sigma$, and $n$ is the total number of Poisson samples $n_k$ of parameter $\ell$. Using the simulated losses, we calculate the moments $\mu(\alpha_i)$, which are the input data needed to obtain the Maxentropic distributions.

2. We apply the SME and MEM approaches described in Sections 2.1 and 2.2, using eight fractional moments $\mu(\alpha_i)$ of the exponential of the variable, where $\alpha_i = \frac{1.5}{i}$ with $i = 1, \ldots, 8$. That is $K = 8$ is the number of non-zero moments. The minimization procedure is carried out using the Barzilai-Borwein's (1988) algorithm.

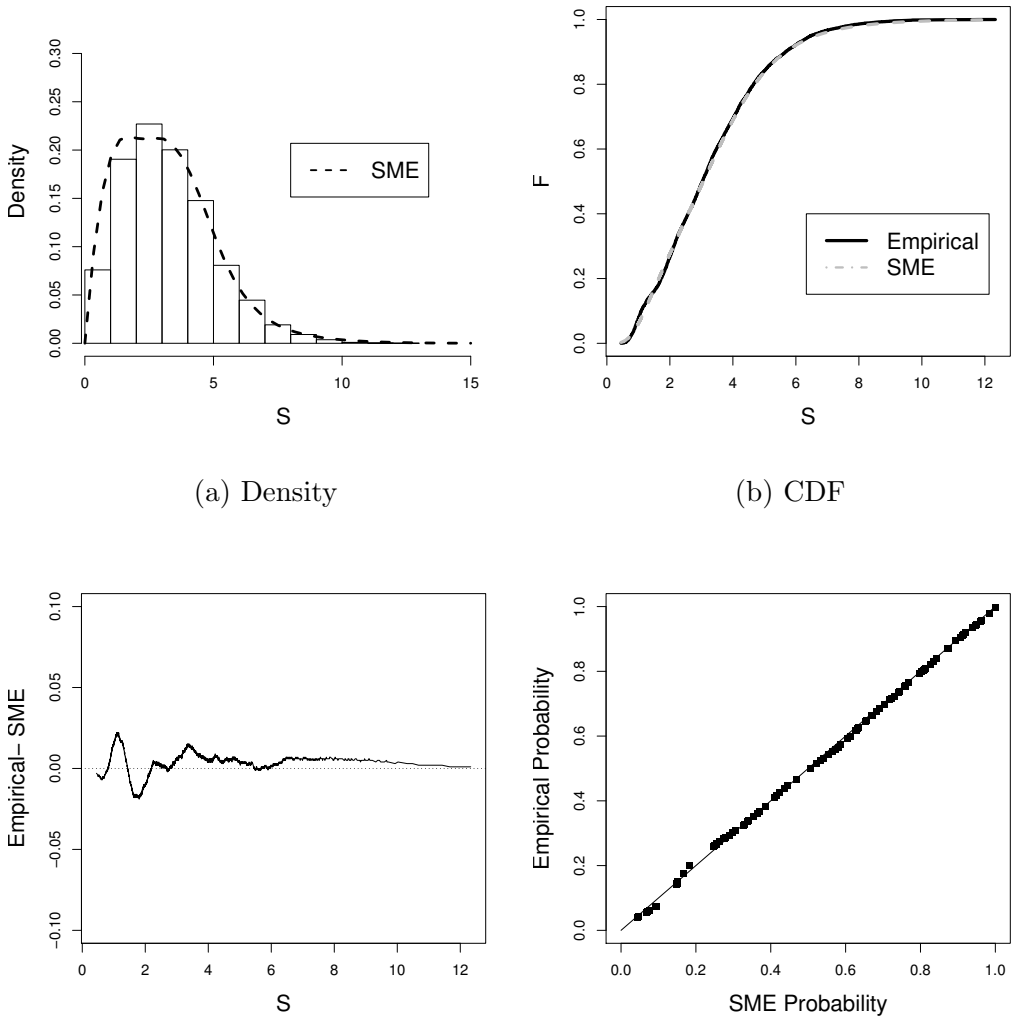
3. Each Maxentropic approach is implemented in five cases. One is described with full detail and called Case (1), characterized by the parameters $\ell = 3$, $\mu = 0$, $\sigma = 0.25$, and the other four (Cases 2-5) are used to verify the robustness of the Maxentropic approaches with respect to changes in the parameters. Thus, we consider two cases with the same lognormal parameters $\mu = 0, \sigma = 0.25$, but different Poisson parameters (Cases 2-3) and other two cases with the same Poisson parameter $\ell = 3$ and different lognormal parameters (Cases 4-5).

4. Once the Maxentropic densities have been obtained, we evaluate the quality of the reconstructions using a variety of criteria (described in the Appendix A.1) for the two independent data sets, namely, the observed data set of size 8000 and a test data set of size between 1000 and 1500. These datasets may contain zero values that are discarded from the analysis through equation (2.3), with $p_o = P(S = 0) = e^{-\ell}$.

### 2.4.1 Numerical Reconstructions with the SME approach

This section starts with Case 1 described by the parameters $\ell = 3$, $\mu = 0$ and $\sigma = 0.25$. In Figure 2.1 we display the density, the distribution function, the marginal calibration and reliability diagram of the SME based reconstruction. These plots allow us to observe the performance of the SME method with respect to the empirical histogram and the empirical (cumulative) distribution function.

Notice in Figure 2.1a that the SME density and the histogram of the observed data appear to be close. The same can be seen in Figure 2.1b for the (cumulative) distribution function of the SME and the observed data, whose differences seem to be imperceptible. Figure 2.1c displays the marginal calibration diagram that allows us to observe the differences between the (cumulative) distribution functions of the SME reconstruction and the observed data. This figure shows that the differences are not larger than 0.022. Additionally, Figure 2.1d shows the




(a) Density

(b) CDF

(c) Marginal calibration

(d) Reliability Diagram

Figure 2.1: SME results. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

reliability diagram of the observed frequency against the SME density.

Even though the results of the Figure 2.1 seems to indicate a good reconstruction, as measure of closeness we also consider the $L_1$ and $L_2$ norms of the distances between the reconstructed density and the empirical density, as well as the MAE and RMSE errors detailed in the Appendix A.1 at the end of this thesis. In Table 2.2 we display the results of the computations. These values confirm the plausibility of the good quality of the reconstruction displayed in Figure 2.1.

It is convenient to test density reconstructions for correct calibration. This consists of testing whether the inverse probability transforms (PIT) is independent and uniformly distributed.

| Approach | $L_1$ | $L_2$ | MAE | RMSE |
|---|---|---|---|---|
| SME | 0.1225 | 0.0598 | 0.0071 | 0.0089 |

Table 2.2: Errors SME approach. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

Deviations from uniformity may indicate a poor reconstruction. In Figure 2.2 we display the PIT transformation and correlograms of different powers for Case (1). As we can see, the PIT histogram seems to be uniform and the correlation plots do not show any sign of dependence.

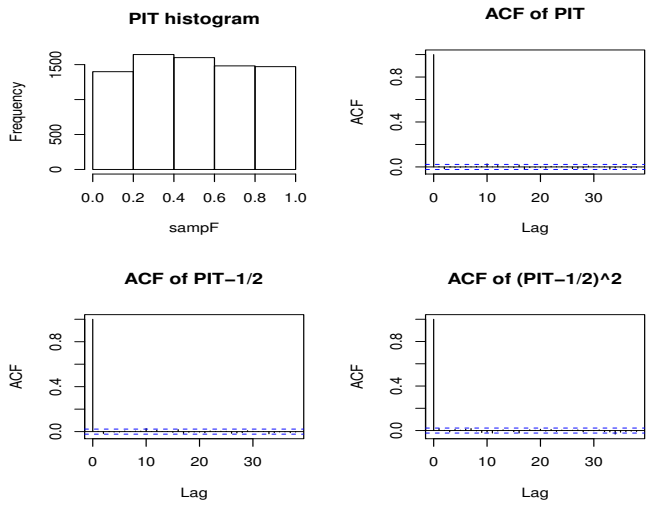
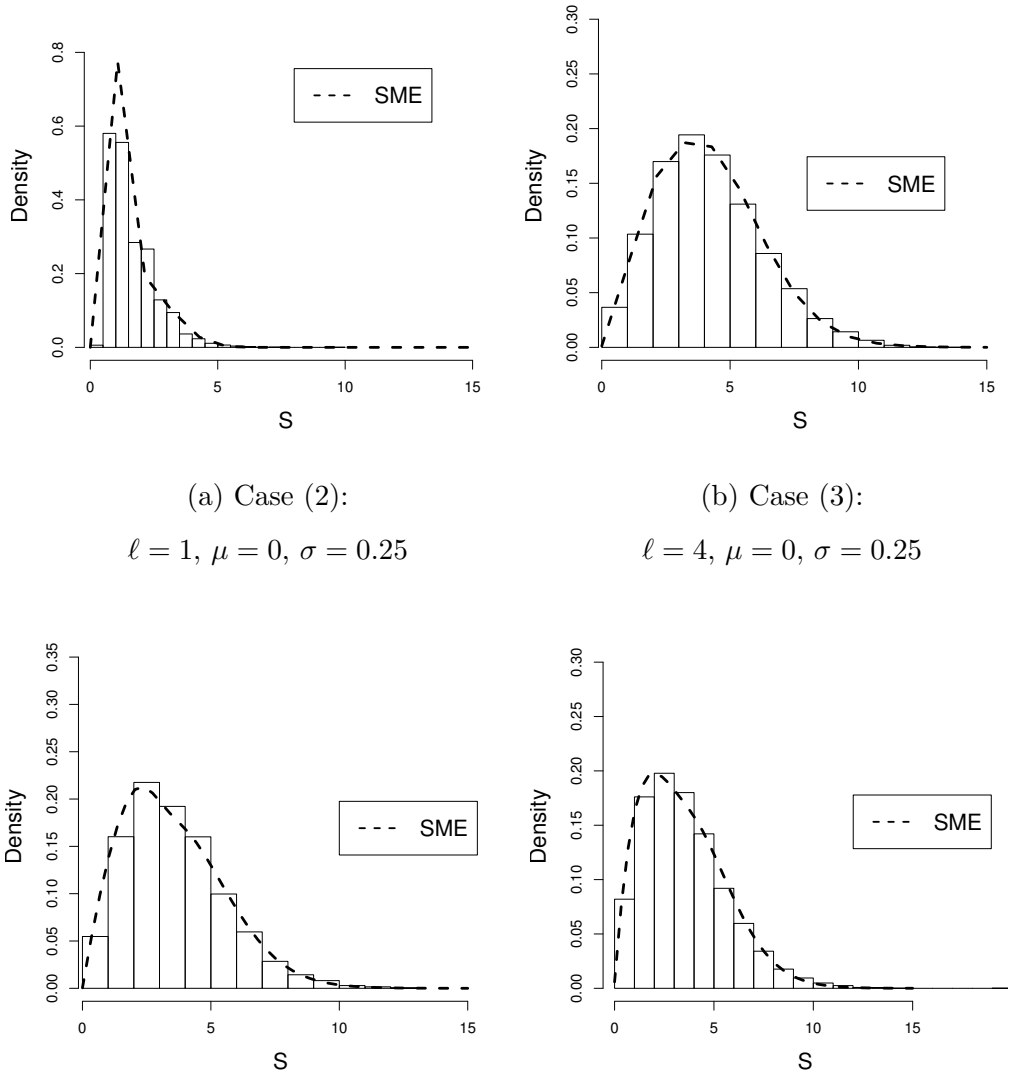



Figure 2.2: Probability integral transform (PIT) histogram and sample autocorrelation functions for the SME approach. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$.

As said, to test robustness of the SME method, we performed reconstructions with simulated data for other values of the parameters. In each of the cases we went through the same routine: we generated the data, computed the moments and carried out the Maxentropic procedure. In Figure 2.3 we display the different SME reconstructions, along with the histogram of the observed data. The same criteria that we used before to measure quality of the Case (1) yield consistent results in these cases as well.

Also, the distances between the estimated and the empirical densities in the $L_1$ and $L_2$ norms, and the MAE and RMSE values, for Cases (2) to (5) considered in Figure 2.3, are listed in Table 2.3. They suggest that the reconstructions are reasonably good. Besides, in Table 2.3 we can see that in Cases (3) and (4) the values of the $L_1$, $L_2$, MAE and RMSE are smaller




(a) Case (2): $\ell = 1$, $\mu = 0$, $\sigma = 0.25$

(b) Case (3): $\ell = 4$, $\mu = 0$, $\sigma = 0.25$

(c) Case (4): $\ell = 3$, $\mu = 0.1$, $\sigma = 0.25$

(d) Case (5): $\ell = 3$, $\mu = 0$, $\sigma = 0.5$

Figure 2.3: SME densities for different Poisson-Lognormal parameters

than those observed in Case (1), and in Cases (2) and (5) are larger.

In order to evaluate the ability of the SME method to perform well on data that does not were used to fit the model, we calculate the error and distances of the SME densities with a test data set, that is, an independent and smaller sample from the same population as the observed data. The corresponding $L_1$, $L_2$, MAE and RMSE results for all the cases consider in this paper are shown in Table 2.4. They seem to show similar results to the obtained with the observed

| Error | Case (2) | Case (3) | Case (4) | Case (5) |
|---|---|---|---|---|
| $L_1$-norm | 0.2649 | 0.0947 | 0.1196 | 0.1105 |
| $L_2$-norm | 0.2099 | 0.0399 | 0.0563 | 0.0516 |
| MAE | 0.0216 | 0.0038 | 0.0074 | 0.0058 |
| RMSE | 0.0257 | 0.0047 | 0.0094 | 0.0064 |

Table 2.3: Errors SME approach. Cases: (2)-(5)

data set displayed in Tables 2.2 and 2.3. These tables indicate that the obtained Maxentropic approximation does not suffer of overffiting and performs well on unobserved data.

| Error | Case (1) | Case (2) | Case (3) | Case (4) | Case (5) |
|---|---|---|---|---|---|
| L1-norm | 0.1223 | 0.2103 | 0.0960 | 0.1206 | 0.1471 |
| L2-norm | 0.0649 | 0.1847 | 0.0408 | 0.0591 | 0.0651 |
| MAE | 0.0109 | 0.0216 | 0.0126 | 0.0095 | 0.0120 |
| RMSE | 0.0147 | 0.0259 | 0.0140 | 0.0121 | 0.0171 |

Table 2.4: Errors SME approach. Validation set

For each choice of parameters, we applied the different non-parametric statistical tests described in the Appendix A.1. The results obtained are displayed in Table 2.5. Three stars mean that at 1% of significance we do not reject the null hypothesis of no differences between the empirical and Maxentropic density, and two stars mean that we do not reject the null hypothesis at 5% of significance. Most of the cases pass the tests at 5% of significance, and the rest of the tests do it at 1% of significance.

### 2.4.2 Numerical Reconstructions with the MEM approach

In this section we describe the results of the implementation of the MEM approach described in Section 2.2. We consider the same compound sum as before, with parameters $\ell = 3$, $\mu = 0$ and $\sigma = 0.25$ respectively. Here we use the MEM with a Poisson reference measure with parameters $\eta = 2$, and a partition in $[0, 1]$ of size $M = 200$, as described in Section 2.2.

In Figure 2.4 we display the density, the distribution function, the marginal calibration and reliability diagram of the MEM reconstruction and the observed data set. In panel 2.4a we show

| **Criterion** | Case (1) | Case (2) | Case (3) | Case (4) | Case (5) |
|---|---|---|---|---|---|
| KS test of uniformity | 1.51*** | 1.28** | 0.87** | 1.17** | 1.24** |
| Anderson-Darling test: | 1.95** | 2.44** | 1.46** | 1.51** | 1.39** |
| Cramér- Von Mises test: | 0.30** | 0.44** | 0.19** | 0.13** | 0.29** |
| Berkowitz Test: : | 5.74** | 2.17** | 4.43** | 4.55** | 1.19** |
| Jarque-Bera Test : | 1.34** | 8.98*** | 2.93** | 4.27** | 3.67** |

Table 2.5: *Statistics* of SME approach. Test set

the SME and MEM reconstructions along with the histogram of the observed data, whereas in panels 2.4b, 2.4c and 2.4d, we display the distribution function, the marginal calibration diagram and reliability diagram respectively, here we only show the MEM reconstruction along with the empirical distribution. In Figure 2.4c we observe minor fluctuations about zero, with values not greater than 0.027 in absolute value. Such small fluctuations indicate a good fit.

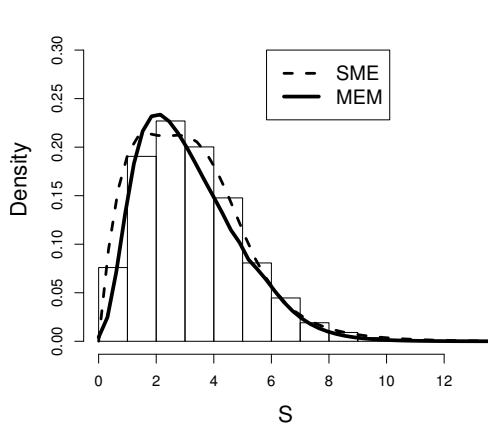
In Table 2.6 we show the numerical distances between the reconstructions in the $L_1$, $L_2$, SME and MEM distances, in order to see how different the Maxentropic densities are from the observed data set. Clearly, in Table 2.6 the SME reconstruction seems to be a little better than the MEM reconstruction, but even so, the MEM approximation provides us with a quite good reconstruction.

| Approach | L1 | L2 | MAE | RMSE |
|---|---|---|---|---|
| SME | 0.1225 | 0.0598 | 0.0071 | 0.0089 |
| MEM | 0.1279 | 0.0609 | 0.0086 | 0.0109 |

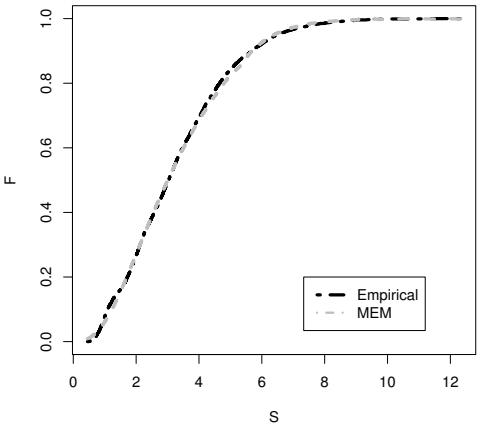
Table 2.6: Errors of SME and MEM approaches. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

In Figure 2.5 we display the PIT transformation and correlograms of different powers for the Case (1). As we can see, the PIT histogram seems to be uniform and the correlation plot does not show any sign of dependence.

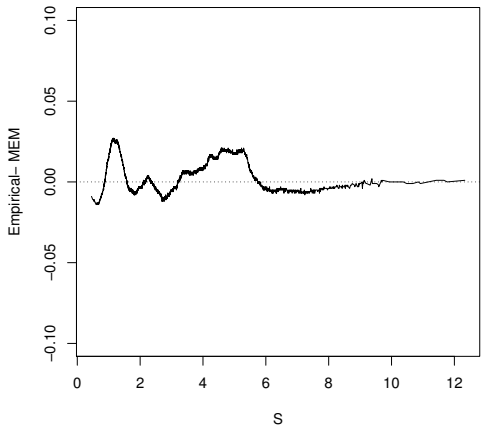
In Figure 2.6 we display the results of MEM and SME approaches, along with the empirical histogram of the compounded sum for different values of the parameters $\ell$, $\mu$ and $\sigma$, in order to see the differences between the reconstructions and the observed data. We point out that



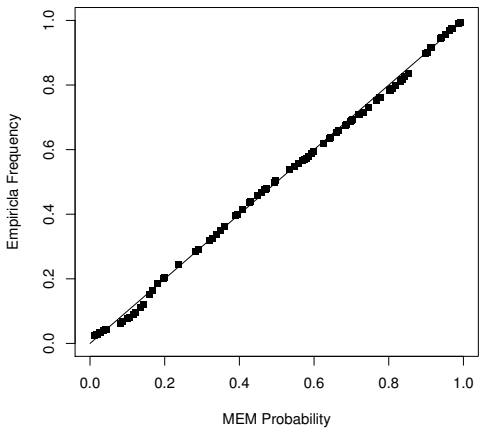

(a) Density




(b) CDF




(c) Marginal calibration




(d) Reliability Diagram

Figure 2.4: MEM results ($\eta = 2$, $M = 200$). Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

the results of the MEM method are very sensitive to the interpolation scheme used to display the results and to the size of the partition employed, which makes it harder to fit the tails of the distribution.

In Table 2.7 we list the $L_1$ and $L_2$ distances between reconstructed and empirical densities, along with the MAE and RMSE distances between the MEM-distribution function and the empirical distribution function. These errors are in most of the cases a little larger than those obtained for the SME reconstructions. An exception occurs for the Case (2) where the MEM reconstruction is better than the density obtained with the SME.

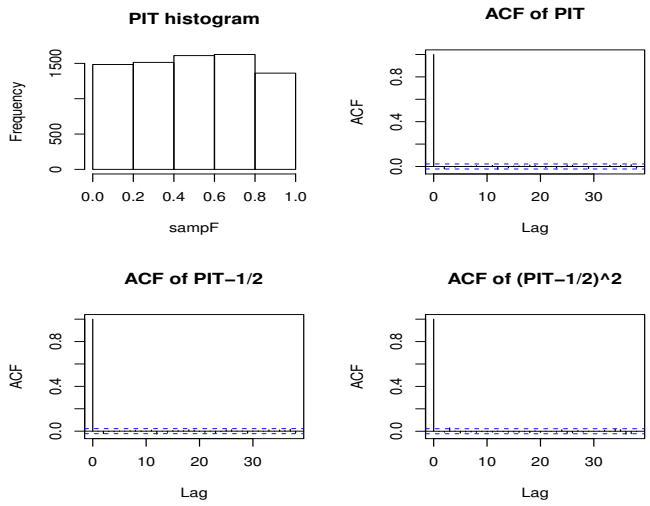


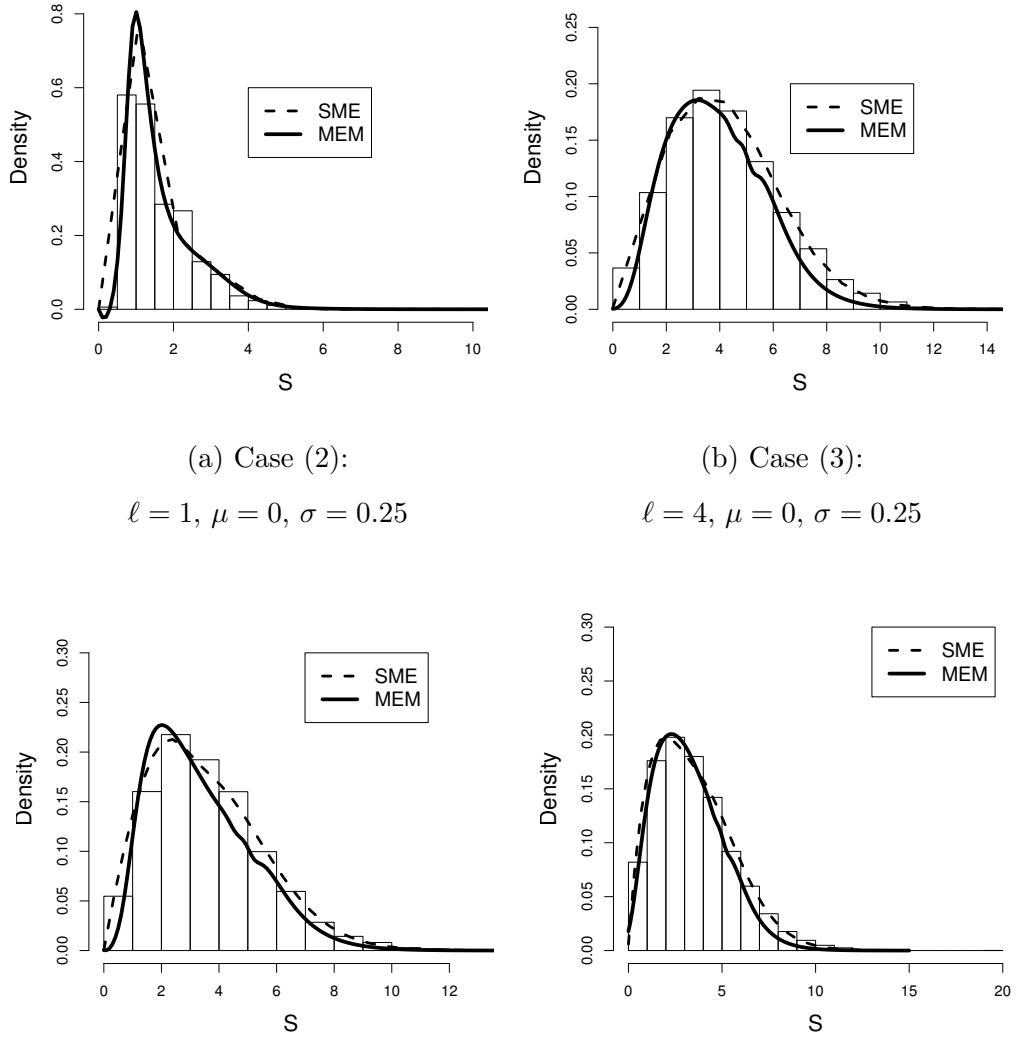

Figure 2.5: Probability integral transform (PIT) histogram and sample autocorrelation functions for the MEM approach. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$.

| **ERROR** | **Case (2)** | | **Case (3)** | | **Case (4)** | | **Case (5)** | |
|---|---|---|---|---|---|---|---|---|
| | SME | MEM | SME | MEM | SME | MEM | SME | MEM |
| L1-norm | 0.2649 | 0.2560 | 0.0947 | 0.1952 | 0.1196 | 0.1652 | 0.1105 | 0.1498 |
| L2-norm | 0.2099 | 0.2091 | 0.0399 | 0.0857 | 0.0563 | 0.0770 | 0.0516 | 0.0605 |
| MAE | 0.0216 | 0.0182 | 0.0038 | 0.0172 | 0.0074 | 0.0123 | 0.0058 | 0.0114 |
| RMSE | 0.0257 | 0.0221 | 0.0047 | 0.0248 | 0.0094 | 0.0145 | 0.0064 | 0.0166 |

Table 2.7: Errors of SME and MEM approaches. Cases: (2)-(5)

Using the test set we can observe how does the MEM reconstruction perform on unobserved data. In Table 2.8 we display the values of the $L_1$, $L_2$, MAE and RMSE measures of error. Clearly, these results are similar to those displayed in Table 2.7 proving the good performance of the obtained reconstructions.

In Table 2.9 we display the results of statistical tests applied to the test set. As in Section 2.4.1 those marked with asterisks do no reject the null hypothesis of equality between distributions. As we can appreciate, most of the tests seems to validate our reconstruction at 5% of significance.




(a) Case (2):
$\ell = 1$, $\mu = 0$, $\sigma = 0.25$

(b) Case (3):
$\ell = 4$, $\mu = 0$, $\sigma = 0.25$

(c) Case (4):
$\ell = 3$, $\mu = 0.1$, $\sigma = 0.25$

(d) Case (5):
$\ell = 3$, $\mu = 0$, $\sigma = 0.5$

Figure 2.6: SME & MEM densities for different Poisson-Lognormal parameters

## 2.5 Risk measures

In this Section we present the computation of the $VaR$ and $TVaR$ of the total loss $S$ corresponding to the parameters $\ell = 3$, $\mu = 0$ and $\sigma = 0.25$ (i.e., Case (1)), at various confidence levels. These results also serve to test the potential of the SME and MEM approaches. In this example we compare the quantiles of the reconstruction against the sample quantiles.

| Error | Case (1) | Case (2) | Case (3) | Case (4) | Case (5) |
|---|---|---|---|---|---|
| L1-norm | 0.1896 | 0.2057 | 0.1580 | 0.1598 | 0.1751 |
| L2-norm | 0.1370 | 0.1866 | 0.0781 | 0.0763 | 0.0704 |
| MAE | 0.0131 | 0.0186 | 0.0201 | 0.0170 | 0.0161 |
| RMSE | 0.0150 | 0.0225 | 0.0223 | 0.0201 | 0.0198 |

Table 2.8: Errors of the MEM approach. Test set

| **Critical Values** | | | | | |
|---|---|---|---|---|---|
| **Test** | Case (1) | Case (2) | Case (3) | Case (4) | Case (5) |
| KS test of uniformity | 1.07** | 1.11** | 1.22** | 1.53*** | 1.23** |
| Anderson-Darling test: | 2.28** | 1.68** | 3.71*** | 3.31*** | 2.99*** |
| Cramér-v. Mises test: | 0.30** | 0.31** | 0.41** | 0.25** | 0.39** |
| Berkowitz Test: | 7.74** | 5.27** | 10.04*** | 6.94** | 1.54** |
| Jarque-Bera Test: | 10.25 | 7.78*** | 3.75** | 2.28** | 9.20*** |

Table 2.9: *Critical Values* of MEM approach. Test set

For the calculation of $VaR$ and $TVaR$ at the confidence level $\gamma$ of the SME and MEM reconstructions, we use the theoretical definition of VaR and TVaR simplified in the following lemma taken from Rockafellar and Uryasev (2000).

**Lemma 1.** *The function*

$$a \to U(a) = a + \frac{1}{1-\gamma}\int_a^\infty (t-a) f_S^*(t) dt$$

*defined on* $(0, \infty)$ *is convex in* $a$*, achieves its minimum at* $VaR_\gamma$ *and its minimum value is* $TVaR_\gamma(S) = E^*[S|S > VaR_\gamma]$*. Then,* $f_S^*$ *denotes the optimal Maxentropic density.*

In Tables 2.10 and 2.11 we display the resulting computations for a collection of $\gamma$'s. The lasts columns are corresponding to $VaR$ and $TVaR$ of the simulated sample (observed data set) along with their confidence levels at 95%. Additionally we include the absolute difference between the VaR and TVaR of the observed data with those from the VaR and TVaR obtained from the reconstructions. The symbol * indicate those values that are close to the empirical value.

In order to calculate the empirical VaR and TVaR we consider $S > 0$ ordered in increasing size $s_1 \leq s_2 \leq s_T^*$, where $T^*$ is the number of sample points with $S > 0$. Then we estimated $VaR$ as $\widehat{VaR}_\gamma(S) \approx S([T^*(\gamma)])$, where $[a]$ denotes the integer part of the real number $a$. The estimate of the TVaR is obtained from the same ordered list of values as

$$\widehat{TVaR}_\gamma = \frac{1}{T^* - [T^*\gamma] + 1} \sum_{j=[T^*(\gamma)]}^{T^*} s_j$$

The confidence intervals for the VaR and TVaR were calculated by resampling without replacement using subsamples of size 90% of the total data size.

| VaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval | |
| $\gamma$ | SME | MEM | Empirical | SME error | MEM error | $VaR_{inf}$ | $VaR_{sup}$ |
| 0.900 | 5.657* | 5.657* | 5.672 | 0.015 | 0.015 | 5.587 | 5.763 |
| 0.910 | 5.798* | 5.818* | 5.809 | 0.011 | 0.009 | 5.725 | 5.920 |
| 0.920 | 5.939* | 5.980* | 5.968 | 0.029 | 0.012 | 5.872 | 6.055 |
| 0.930 | 6.081* | 6.141* | 6.118 | 0.037 | 0.023 | 6.021 | 6.227 |
| 0.940 | 6.222* | 6.303* | 6.299 | 0.077 | 0.004 | 6.202 | 6.379 |
| 0.950 | 6.505* | 6.465* | 6.474 | 0.031 | 0.009 | 6.377 | 6.614 |
| 0.960 | 6.788* | 6.788* | 6.759 | 0.029 | 0.029 | 6.617 | 6.908 |
| 0.970 | 7.071* | 7.273* | 7.122 | 0.051 | 0.151 | 6.955 | 7.334 |
| 0.980 | 7.495* | 7.919 | 7.583 | 0.088 | 0.336 | 7.428 | 7.767 |
| 0.990 | 8.485* | 8.566* | 8.384 | 0.101 | 0.182 | 8.078 | 8.593 |
| 0.995 | 9.051* | 9.061* | 9.016 | 0.035 | 0.045 | 8.747 | 9.210 |
| 0.999 | 9.192 | 10.182* | 10.34 | 1.148 | 0.158 | 9.686 | 11.43 |

Table 2.10: Comparison of VaR for the SME and MEM reconstructions. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

In Tables 2.10 and 2.11 the asterisks indicate that the calculated VaR and TVaR for the reconstructions belongs to the empirical confidence interval at the 95% level, also shown in each table. The rather good agreement of the results displayed in these tables are a further indication of the quality of the Maxentropic methods.

| TVaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval | |
| $\gamma$ | SME | MEM | Empirical | SME error | MEM error | $TVaR_{inf}$ | $TVaR_{sup}$ |
| 0.900 | 6.817* | 6.833* | 6.839 | 0.022 | 0.006 | 6.721 | 6.948 |
| 0.910 | 6.930* | 6.967* | 6.961 | 0.031 | 0.006 | 6.846 | 7.079 |
| 0.920 | 7.055* | 7.041* | 7.095 | 0.040 | 0.054 | 6.974 | 7.232 |
| 0.930 | 7.192* | 7.121* | 7.245 | 0.053 | 0.124 | 7.109 | 7.389 |
| 0.940 | 7.344* | 7.303* | 7.417 | 0.073 | 0.114 | 7.276 | 7.562 |
| 0.950 | 7.513* | 7.528* | 7.622 | 0.109 | 0.094 | 7.459 | 7.781 |
| 0.960 | 7.935* | 7.814* | 7.874 | 0.061 | 0.060 | 7.700 | 8.045 |
| 0.970 | 8.205* | 8.220* | 8.188 | 0.017 | 0.032 | 7.989 | 8.373 |
| 0.980 | 8.533* | 8.516* | 8.601 | 0.068 | 0.085 | 8.405 | 8.817 |
| 0.990 | 8.959* | 8.912* | 9.262 | 0.303 | 0.350 | 8.968 | 9.555 |
| 0.995 | 9.556* | 9.606* | 9.843 | 0.287 | 0.237 | 9.465 | 10.222 |
| 0.999 | 10.543* | 11.215* | 11.167 | 0.624 | 0.048 | 10.341 | 11.848 |

Table 2.11: Comparison of TVaR for the SME and MEM reconstructions. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

**Comment** The calculation of the VaR and the TVaR measures using the SME and MEM densities, may be considered to be an additional test to evaluate the quality of the reconstruction. This is done by comparing these values with the empirical VaR and TVaR obtained from the observed data.

## 2.6 Decompounding

It may not be always possible to observe frequency and severity separately. That is, even though the frequency of events is recorded, only the total or aggregated loss data is available. Nevertheless, the risk analyst may want to know the distribution of individual losses, because it is at that level where loss prevention or mitigation is applied. The Maxentropic approaches that we have developed here allow us to determine the distribution of the individual loss as well.

In our case we know how to compute numerically the Laplace transform $\psi(\alpha)$ of the total losses $S$, and we also have an analytic expression for the generating function $G(z)$ of the frequency of the events $N$. From these ingredients, we can obtain the Laplace transform $\phi(\alpha)$ of the individual losses, which we can use as starting point to determine the probability distribution of individual losses. The relationship between the various Laplace transforms is contained in equation (2.3), from which we obtain

$$\phi(\alpha_k) = \tfrac{1}{l} ln(\psi(\alpha_k)) + 1,$$

where, to use what we have developed above, we may write

$$\psi(\alpha_k) = e^{-\ell} + (1 - e^{\ell}) \int_0^1 y^{\alpha_k} f_Y^*(y) dy,$$

for the Laplace transform of the aggregate losses, $\ell$ being the parameter of the observed Poisson frequency, which is known in our case, and $f_Y^*$ is the Maxentropic probability density of the total losses, which we have already determined.

To exemplify our procedure, we shall use Case (1), characterized by the following parameters: $\ell = 3$, $\mu = 0$ and $\sigma = 0.25$. Having determined numerically $f_S^*$, and computed $\phi(\alpha)$ as mentioned before, we can apply the SME and MEM procedures to obtain the probability density of individual losses. The resulting individual densities are shown in Figure 2.7. The comparison with the known probability density that was used to generate the total losses is another consistency test of the procedure. The true distributions of individual losses is a lognormal density with parameters $\mu = 0$, $\sigma = 0.25$, and is included in Figure 2.7.

In Tables 2.12 and 2.13 we present the results of several measures of quality of reconstruction. Here the SME and MEM reconstructions are compared with the true lognormal density and with the observed (simulated) data. We include the comparison between the reconstructed and true density to the data obtained by simulation, but keep in mind that such data might not be available. The true density is to be used as a benchmark to test the quality of the Maxentropic procedures. Note that the best results are obtained for the MEM-reconstruction, whose errors are smaller, namely, 0.0087 and 0.0105 respectively, as measured by the MAE and RMSE distances to the empirical data.

The $L_1$ and $L_2$ distances between the reconstructions by both procedures and the histogram are similar to those between the true distribution and the histogram. Notice that in the

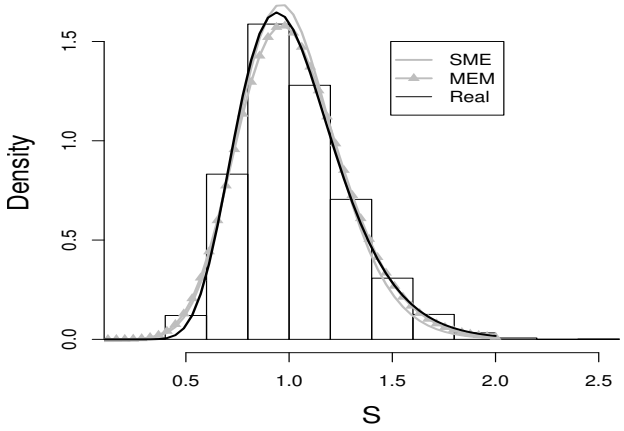



Figure 2.7: Density of the individual losses obtained by SME & MEM approaches. Case (1): $\ell = 3$, $\mu = 0$, $\sigma = 0.25$

left column of the second table are equal because they measure distance between the true density and the histogram. The distances between the Maxentropic density and the histogram, measured by the $L_1$ and the $L_2$ are 0.16 and 0.18 respectively. Notice that they are similar to the same distances between the true density and the histogram. These good results play the role of a check up test for the SME and MEM procedures applied throughout this paper.

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | Real density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SME | 0.0042 | 0.0047 | 0.0127 | 0.0257 | 0.0143 | 0.0232 |
| MEM | 0.0042 | 0.0047 | 0.0087 | 0.0105 | 0.0096 | 0.0113 |

Table 2.12: MAE and RMSE values of the individual losses calculated by SME & MEM approaches

## 2.7 Concluding comments

We examined the performance of two Maxentropic methods for the reconstruction of the probability density of total losses in a standard actuarial model of interest to the insurance and banking industries. The two methods are based on the maximization of an entropy functional

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | Real density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SME | 0.1640 | 0.1865 | 0.1886 | 0.1992 | 0.0679 | 0.0574 |
| MEM | 0.1640 | 0.1865 | 0.1659 | 0.1818 | 0.0621 | 0.0624 |

Table 2.13: $L_1$ and $L_2$ distances of the individual losses calculated by SME & MEM approaches

defined on an appropriate space of probabilities. The input for each procedure essentially consists of the value of the Laplace transform of the total loss distribution, which in our case has been determined numerically. This is particularly useful in cases like the one treated here, in which there is no easy analytic way of computing the Laplace transform of the individual severity

One important reason for using the Maxentropic methods, is that they provide good reconstructions based on a very small amount of information. In our case this means the knowledge of the Laplace transform at $K = 8$ real values of its transform's parameter. Besides providing quite reasonable representations of the density of total loss, the Maxentropic methods admit extensions that can be used to handle measurement error or scarce data situations. This will be discussed in Chapter 3.

To asses the quality of our method, we carried out a variety of statistical tests, including an indirect consistency test consisting of an application of a decompounding procedure to determine the probability density of the individual losses, when only the total loss is recorded and a model is available for the frequency of events. The statistical consistency tests were passed quite satisfactorily.

At this point we mention the reason for the large size of the data sample. We considered a very large sample for not to worry about the sample errors built into the numerical determination of the Laplace transform. When the size of the sample data is small, we have to take the potential errors in the input (that is, the fractional moments) into account. The method of maximum entropy can be extended to take care of that possibility. We shall take up this problem in Chapter 3.

As we explain in the introduction our procedure works in several levels. Suppose for example that we have the following problem: Suppose that one knows that there are several sources of

risk, each one with similar distribution, and suppose that only the total number of events and total combined loss are recorded. Within our framework we can determine the frequency distribution of each type of risk and the distribution of the severities (or individual risks). We will developed in more detail this idea in Chapter 4.

The potential applicability of the Maxentropic methodology to problems in operational risk and insurance is enormous, specially in the cases where the historical data could have problems of multi-modality, heavy tails, or may include extreme events or economic shocks, for which the estimation of a probability density may be extremely difficult to compute.

# Chapter 3

# Density Reconstructions with Errors in the Data

## 3.1 Introduction

We discuss in the previous chapters that an important problem in many applications of probability is the determination of the probability density of a positive random variable when the information available consists of an observed sample. For example, it can be either an exit time or a reaction time, the accumulated losses or or accumulated damage and so on. A standard technique, related to a variety of branches of analysis, consists of use of the Laplace transform. But some times such technique may fail because the transform cannot be determined as in the case of the lognormal variable. In this regard, see the efforts in Leipnik (1991) to determine the Fourier-Laplace transform of the lognormal variable. One is then led to search for techniques to invert Laplace transforms from a few of its values determined numerically. And that is the reason why we chose to use a sample generated from a Poisson-lognormal as data to test the methods that we propose.

How we established before, we are interested in a method to obtain a probability density $f_S(s)$ from the knowledge of the values of the Laplace transform

$$E[e^{-\alpha_i S}] = \mu(\alpha_k), \quad i = 1, ..., K. \tag{3.1}$$

To be specific, the positive random variable $S$ may denote the total loss of some kind of losses accumulated during a given time interval, and the density $f_S(s)$ is the object that we are

after. Due to the importance of this problem for the insurance industry, there has been quite a large amount of effort devoted to finding systematic ways to compute $f_S(s)$ from the knowledge of the ingredients of some model relating the $S$ to more basic quantities, like frequency of losses and individual severities. See Panjer (2006) for example, for a relatively recent update on methods to deal with that problem.

We should mention at the outset that if the Laplace transform $E[e^{-\alpha S}]$ was known as a function on the positive real axis, a variety of methods to determine $f_S$ exist. Among them, the use of maximum entropy techniques that bypass the need to extend the Laplace transform into a complex half-plane. We saw in Chapter 2 that the standard maximum entropy method (SME) to solve this problem is simple to implement.

As we explained in Chapter 2, in many cases, $E[e^{-\alpha S}]$ has to be estimated from observed values $s_1, ...., s_T$ of $S$, so, the only knowledge that may be available to us is the total loss in a given period. It is at this point where errors come in, because in order to determine numerically $\mu(\alpha)$ we have to use a random sample. If we were somehow determining $\mu(\alpha)$ by means of some experimental procedure, then an error might come in through the measurement process.

For this reason, the problem that we want to address is now be restated as:

$$\text{Find } \ F_S \ \text{ such that } \ E[e^{-\alpha_i S}] = \int_0^\infty e^{-\alpha_i s} dF_S(s) \in C_i \ \ i = 1, ..., K. \tag{3.2}$$

where $C_i$ is some interval enclosing the true value of $\mu(\alpha_i)$ for $i = 1, ..., K$. These intervals are related to the uncertainty (error) in the data. For us these will be the statistical confidence intervals, but as we are not using them for statistical estimation of a mean, but as a measure of some experimental error, we adjust the width of the interval to our convenience.

To transform the problem into a fractional moment problem, note that, since $S$ is positive, we may think of $Y = e^{-S}$ as a variable in $[0, 1]$ whose density $f_Y(y)$ we want to infer from the knowledge of an interval in which its fractional moments fall in, that is, from

$$\int_0^1 y^\alpha f_Y(y) dy \in C_i, \ \ i = 1, ..., K. \tag{3.3}$$

where $C_i$ denotes now an interval around the true but unknown moments $\mu(\alpha_i)$ of $f_Y$.

To obtain $f_Y$ using the constraint (3.3), the basis of the SME method described in Chapter 2 may be used. See Cherny et al. (2003) for a rigorous proof of the basic existence and representation of the SME method and for applications in statistical mechanics. See as well

Csiszár (1975) and Csiszár (1986) for a different rigorous proofs of these results. See also Kapur (1996) for an interesting collection of applications in a large variety of fields.

Extensions of the method of maximum entropy to handle errors in the data do not seem to have received much attention despite their potential applicability. Two such possible extensions, explored in the framework of the method of maximum entropy in the mean (MEM) as applied to linear inverse problems with convex constraints, were explored in Gzyl and Velásquez (2011). Here we want to provide alternative ways to incorporate methods to deal with errors in the data to solve (3.3) within the framework of the SME method. The difference between the two methods that we analyze in this chapter lies in that one of them provides us with an estimator of the additive error.

This chapter is organized as follows. In the next section we present the two extensions of the SME, and in the Section 3.3 we apply them to obtain the probability density $f_S(s)$ from the knowledge of the interval where the fractional moments of $Y = e^{-S}$ fall in. We point out two features of our simulations at this stage: we consider one with a relatively small sample and one with a larger sample. The exact probability density from which the data is sampled is to be used as a benchmark against which the output of the procedures is compared. We should mention that the methods we present here are a direct alternative to the methods based on the method of maximum entropy in the mean (MEM). For an application of that technique to determination of risk measure from the knowledge of mispriced risks, see Gzyl and Mayoral (2010).

## 3.2 The Maxentropic approaches

As was mentioned before, in each subsection of Section 3.2 we consider a different way to extend the SME. In Section 3.2.1 we present the extension of the method of maximum entropy to include errors in the data, later in Section 3.2.2 we present another version which additionally allows us estimate the additive error in the data.

### 3.2.1 Extension of the standard Maxentropic approach without error estimation (SMEE-1)

Here we present an extension of the variational method originally proposed by Jaynes (1957), based on an idea proposed in Gamboa (1998), to solve the (inverse) problem consisting of finding a probability density $f_Y(y)$ (on $[0,1]$ in this case), satisfying the following integral constraints:

$$\int_0^1 y^{\alpha_k} f_Y(y)dy \in C_k \qquad \text{for} \qquad k = 0, ..., K. \tag{3.4}$$

where the interval $C_k = [a_k, b_k]$, around the true but unknown $\mu_Y(\alpha_k)$, is determined from the statistical analysis of the data for each of the moments. For $k = 0$ only we set $C_0 = \{1\}$ since for $\alpha_0 = 0$ we have $\mu_0 = 1$ to take care of the natural normalization requirement on $f_Y(y)$. To state the extension denote by $\mathcal{D}$ the class of probability densities $g(y)$ satisfying (3.4). This class is convex. On this class define the entropy by $S(g) = -\int_0^1 g(y)\ln(g(y))dy$, whenever the integral is finite (or $-\infty$ if not). Now, to solve (3.4), we extend Jaynes' method as follows. The problem now is:

$$\text{Find} \;\; g^* = argsup\{S(g)|g \in \mathcal{D}\} \tag{3.5}$$

To dig further into this problem, let us introduce some notation. Let $g$ denote any density and denote by $\mu_g(\boldsymbol{\alpha})$ the vector of $\boldsymbol{\alpha}$ moments of $g$. Set $\mathbf{C} = C_0 \times ... \times C_K$. And, for $\mathbf{c} \in \mathbf{C}$, let us denote by $\mathcal{D}(\mathbf{c})$ the collection of densities having $\mu_g(\boldsymbol{\alpha}) = \mathbf{c}$. With these notations, and following the proposal in Gamboa (1988), then we carry on the maximization process sequentially, and restate the previous problem as

$$\text{Find} \;\; g^* = argsup\{argsup\{S(g)|g \in \mathcal{D}(\mathbf{c})\}|\mathbf{c} \in \mathbf{C}\}. \tag{3.6}$$

The idea behind the proposal is clear: First, solve a maximum entropy problem for each $\mathbf{c} \in \mathbf{C}$ to determine $g^*_{\mathbf{c}}$, and then, maximize over $\mathbf{c} \in \mathbf{C}$ to determine the $\mathbf{c}^*$ such that $g^*_{\mathbf{c}^*}$ yields the maximum entropy $S(g^*_{\mathbf{c}^*})$ over all possible moments in the confidence set.

Invoking the standard argument, we know that when the inner problem has a solution, it is of the type

$$g^*_{\mathbf{c}}(y) = \exp\left(-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}\right) \tag{3.7}$$

in which the number of moments $K$ appears explicitly. It is usually customary to write $e^{-\lambda_0^*} = Z(\boldsymbol{\lambda}^*)^{-1}$, where $\boldsymbol{\lambda^*} = (\lambda_1^*, ..., \lambda_K^*)$ is an $K-$dimensional vector. Recall as well that

the normalization factor is given by

$$Z(\boldsymbol{\lambda}) = \int_0^1 e^{-\sum_{k=1}^K \lambda_k y^{\alpha_k}} dy. \tag{3.8}$$

With this notation the generic form of the solution looks like

$$g^*_{\mathbf{c}}(y) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\sum_{k=1}^K \lambda^*_k y^{\alpha_k}} = e^{-\sum_{k=0}^K \lambda^*_k y^{\alpha_k}}. \tag{3.9}$$

To complete, it remains to specify how the vector $\boldsymbol{\lambda}^*$ can be found. For that one has to minimize the dual entropy:

$$\Sigma(\boldsymbol{\lambda}, \mathbf{c}) = \ln Z(\boldsymbol{\lambda}) + < \boldsymbol{\lambda}, \mathbf{c} > \tag{3.10}$$

where $< \mathbf{a}, \mathbf{b} >$ denotes the standard Euclidean scalar product. It is also a standard result that

$$\sup\{S(g) | g \in S(g) = \mathbf{c}\} = \inf\{\Sigma(\boldsymbol{\lambda}, \mathbf{c}) | \boldsymbol{\lambda} \in \mathbb{R}^K\}$$

With this the double minimization process can be restated as

$$\sup\{\sup\{S(g) | g \in \mathcal{D}(\mathbf{c})\} | \mathbf{c} \in \mathbf{C}\} = \sup\{\inf\{\Sigma(\boldsymbol{\lambda}, \mathbf{c}) | \boldsymbol{\lambda} \in \mathbb{R}^K\} | \mathbf{c} \in \mathbf{C}\}.$$

And invoking the standard minimax argument, restate it as

$$\sup\{\sup\{S(g) | g \in \mathcal{D}(\mathbf{c})\} | \mathbf{c} \in \mathbf{C}\} = \inf\{\sup\{\Sigma(\boldsymbol{\lambda}, \mathbf{c}) | \mathbf{c} \in \mathbf{C}\} | \boldsymbol{\lambda} \in \mathbb{R}^K\}.$$

Now, due to the special form of $\Sigma(\boldsymbol{\lambda}, \mathbf{c})$, it suffices to compute $\sup\{< \boldsymbol{\lambda}, \mathbf{c} > | \mathbf{c} \in \mathbf{C}\}$. Then, it is simple to verify that $\sup\{< \boldsymbol{\lambda}, \rho > | \rho \in [-1, 1]^K\} = \|\boldsymbol{\lambda}\|_1$. Consider the affine mapping $T(\mathbf{c}) = \mathbf{D}\mathbf{c} + \mathbf{h}$, where $\mathbf{D}$ is diagonal with elements $2/(b_k - a_k)$ and $h_k = -(a_k + b_k)/(b_k - a_k)$. This maps $[a_1, b_1] \times ... \times [a_K, b_K]$ bijectively onto $[-1, 1]^K$. With this is easy to see that

$$\delta^*_{\mathbf{C}}(\boldsymbol{\lambda}) \equiv \sup\{< \boldsymbol{\lambda}, \mathbf{c} > | \mathbf{c} \in \mathbf{C}\} = \|\mathbf{D}^{-1}\boldsymbol{\lambda}\|_1 - < \mathbf{D}^{-1}\boldsymbol{\lambda}, \mathbf{h} > .$$

Explicitly,

$$\|\mathbf{D}^{-1}\boldsymbol{\lambda}\|_1 = \sum_{i=1}^K \frac{(b_i - a_i)|\lambda_i|}{2} \quad \text{and} \quad < \mathbf{D}^{-1}\boldsymbol{\lambda}, \mathbf{h} > = -\sum_{i=1}^K \frac{(b_i + a_i)\lambda_i}{2}.$$

As a first step towards a solution, we have the simple

**Lemma 2.** *With the notations introduced before, set*

$$\Sigma(\boldsymbol{\lambda}) = \ln(Z(\boldsymbol{\lambda})) + \delta^*_{\mathbf{C}}(\boldsymbol{\lambda})$$

*Then* $\Sigma(\boldsymbol{\lambda})$ *is strictly convex in* $\boldsymbol{\lambda}$.

Observe that $\partial\delta^*_{\mathbf{C}}(\boldsymbol{\lambda})/\partial\lambda_i$ is defined except at $\lambda_i = 0$, where it is sub-differentiable (see Borwein et al. (2000)). Actually

$$\begin{aligned}
\partial\delta^*_{\mathbf{C}}(\boldsymbol{\lambda})/\partial\lambda_k &= b_k \ \text{ when } \ \lambda_k > 0, \\
\partial\delta^*_{\mathbf{C}}(\boldsymbol{\lambda})/\partial\lambda_k &= a_k \ \text{ when } \ \lambda_k < 0, \\
\partial\delta^*_{\mathbf{C}}(\boldsymbol{\lambda})/\partial\lambda_k &\in (a_k, b_k) \ \text{ when } \ \lambda_k = 0
\end{aligned} \tag{3.11}$$

And to close up, we have

**Theorem 1.** *Suppose that the infimum $\boldsymbol{\lambda}^*$ of $\Sigma(\boldsymbol{\lambda})$ is reached in the interior of the set $\{\boldsymbol{\lambda} \in \mathbb{R}^K | Z(\boldsymbol{\lambda}) < \infty\}$. Then the solution to the maximum entropy problem (3.6) is*

$$g^*(y) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\sum_{k=1}^K \lambda_k^* y^{\alpha_k}} = e^{-\sum_{k=0}^K \lambda_k^* y^{\alpha_k}}. \tag{3.12}$$

*Due to the computation above, it is clear that*

$$\int_0^1 y^{\alpha_k} g^*(y) dy \in [a_k, b_k], \ \ \textit{for} \ \ k = 1, ..., K.$$

**Comment:** This is a rather curious result. Intuitively at the minimum, we have $\nabla_{\boldsymbol{\lambda}} \Sigma(\boldsymbol{\lambda}) = 0$, then according to (3.11), if all $\lambda_i \neq 0$, the Maxentropic density $g^*(y)$ has moments equal to one of the end points of the confidence intervals. If all $\lambda_i = 0$, the density $g^*(y)$ is uniform, and the reconstructed moments $\int_0^1 y^{\alpha_k} g^*(y) dy$ are anywhere in $[a_k, b_k]$.

### 3.2.2 Extension of the standard Maxentropic approach with error estimation (SMEE-2)

In this section we present an extension of the method of maximum entropy that allows us to estimate the errors in the measurement, and to circumvent the concluding comments in the previous section, that is to obtain estimated moments different from the end points of the confidence intervals. Instead of supposing that $f_Y(y)$ satisfies (3.4) we present a totally different approach. To restate the problem, consider the following argument. If we suppose that the measurement error in the determination of the $k-$th moment lies within the confidence interval $C_k = [a_k, b_k]$, then the unknown estimate of the measurement error can be written as

$p_k a_k + (1-p_k) b_k$ for appropriate but unknown $p_k, q_k = 1 - p_k$. We propose to extend the original problem to (if the extension is not clear, see the Appendix A.2)

$$\int_0^1 y^{\alpha_k} f_Y(y) dy + p_k a_k + (1 - p_k) b_k = \mu_k \qquad \text{for} \qquad k = 1, ..., K. \tag{3.13}$$

The idea behind the proposal is clear. In order to obtain the $\mu_k'$s either experimentally or numerically, we average over a collection of observations (simulations) and the errors in each average additively. Thus the observed value of $\mu_k$, consists of the true (unknown) moment which is to be determined, plus an error term that has to be estimated as well. This time we search for a density $f_Y(y)$ with $y \in [0, 1]$, and numbers $0 < p_k < 1$, $(k = 1, ..., K)$ such that (3.13) holds. This computational simplification is possible when the support of the distribution of errors is bounded. To compress the notations we write probability distributions concentrated on $\{a_k, b_k\}$ as $q_k(dz) = p_k \epsilon_{a_k}(dz) + (1 - p_k)\epsilon_{b_k}(dz)$ and the probability that we are after is a mixture of continuous and discrete distributions. To determine it we define, on the appropriate space of product probabilities (see the Appendix A.2), the entropy

$$S(g, \mathbf{p}) = -\int_0^1 g(y) \ln(g(y)) dy - \sum_{k=1}^{K} \left( p_k \ln p_k + (1 - p_k) \ln(1 - p_k) \right). \tag{3.14}$$

Needless to say, $0 < p_k < 1$, for $k = 1, ..., K$. with all these notations, our problem becomes

$$\begin{aligned} &\text{Find probability density } g^*(y) \text{ and numbers } \ 0 < p_k < 1 \ \text{ satisfying:} \\ &\text{Constraints (3.13) and the normalization constraint } \textstyle\int_0^1 f_Y(y) dy = 1. \end{aligned} \tag{3.15}$$

The solution can be expressed in terms of the Lagrange multipliers

$$\begin{aligned} g^*(y) &= \frac{e^{-\sum_{k=1}^{K} \lambda_k^* y^{\alpha_k}}}{Z(\boldsymbol{\lambda}^*)} \\ p_k &= \frac{e^{-a_k \lambda_k^*}}{e^{-a_k \lambda_k^*} + e^{-b_k \lambda_k^*}}. \end{aligned} \tag{3.16}$$

Here, the normalization factor $Z(\boldsymbol{\lambda})$ is as before. This time the vector $\boldsymbol{\lambda}^*$ of Lagrange multipliers is to be found minimizing the dual entropy

$$\Sigma(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \sum_{k=1}^{K} \ln \left( e^{-a_k \lambda_k} + e^{-b_k \lambda_k} \right) + < \boldsymbol{\lambda}, \boldsymbol{\mu} > \tag{3.17}$$

Once $\boldsymbol{\lambda}^*$ is found, the estimator of the measurement error is, as implicit in (3.13), given by

$$\epsilon_k = \frac{a_k e^{-a_k\lambda_k^*} + b_k e^{-b_k\lambda_k^*}}{e^{-a_k\lambda_k^*} + e^{-b_k\lambda_k^*}}. \tag{3.18}$$

Notice that, although the formal expression for $g^*(y)$ is the same as that for the first method, the result is different because the $\boldsymbol{\lambda}^*$ is found minimizing a different functional.

For not to interrupt the flow of ideas, we compile the basic model behind the results that we just presented as well as some simple but necessary computations in the Appendix A.2.

## 3.3 Numerical implementations

Here we suppose that the random variable of interest follows a lognormal distribution with known parameters $\mu = 1$ and $\sigma = 0.1$. Even though it is a simple example, it is rather representative of the type of distributions appearing in applications. As said, our data will consist in simulated data, and we shall consider two data sets of different sizes.

To produce the two examples do the following:

1. Simulate two samples $\{s_1, ..., s_T\}$ of sizes $T = 200$ and $T = 1000$ from the lognormal distribution.

2. For each sample we compute its $\alpha-$moment and its confidence interval using the standard statistical definition. That is, we compute

$$\bar{\mu}_i = \frac{1}{T}\sum_{k=1}^{T} e^{-s_j\alpha_i}, \quad \alpha_i = 1.5/i, \quad i = 1, ..., 8.$$

   and the confidence interval as specified below.

3. After obtaining each Maxentropic density, we use standard statistical test to measure the quality of the density reconstruction procedure.

Table 3.1 shows the error intervals, which we take to be the 10% confidence interval for the mean obtained respectively using the standard definition, that is as $\left(\bar{\mu}_i - z_{0.45}\frac{sd_i}{sqrt(n)}, \bar{\mu}_i + z_{0.55}\frac{sd_i}{sqrt(n)}\right)$, where $sd_i$ is the sample standard deviation and $\bar{\mu}_i$ is the sample mean of $Y^{\alpha_i}$ the simulated samples of size 200 and 1000.

| $\mu_i$ | **Sample size 200** | **Sample size 1000** |
|---|---|---|
| $\mu_1$ | [0.0175, 0.0177] | [0.0181, 0.0182] |
| $\mu_2$ | [0.1302, 0.1307] | [0.1318, 0.1321] |
| $\mu_3$ | [0.2559, 0.2565] | [0.2578, 0.2581] |
| $\mu_4$ | [0.3592, 0.3599] | [0.3611, 0.3614] |
| $\mu_5$ | [0.4405, 0.4412] | [0.4423, 0.4426] |
| $\mu_6$ | [0.5048, 0.5054] | [0.5065, 0.5068] |
| $\mu_7$ | [0.5565, 0.5570] | [0.5580, 0.5583] |
| $\mu_8$ | [0.5987, 0.5992] | [0.6001, 0.6004] |

Table 3.1: Error Intervals of $S$ for sample sizes of 200 and 1000

In Table 3.3, we list the moments of $S$ for the two sample sizes. As mentioned before, the error intervals are the unique input for the Maxentropic method of density without error estimation, whereas the moments are needed for both the SME and the Maxentropic method with error estimation. Recall as well that the second method forces the estimation error in the $i-$th moment to lie in the corresponding error interval which is centered at the observed sample moment.

| Size | Moments of S | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ | $\mu_7$ | $\mu_8$ |
| 200 | 0.0176 | 0.1304 | 0.2562 | 0.3596 | 0.4409 | 0.5051 | 0.5568 | 0.5990 |
| 1000 | 0.0181 | 0.1319 | 0.2579 | 0.3612 | 0.4424 | 0.5066 | 0.5581 | 0.6002 |

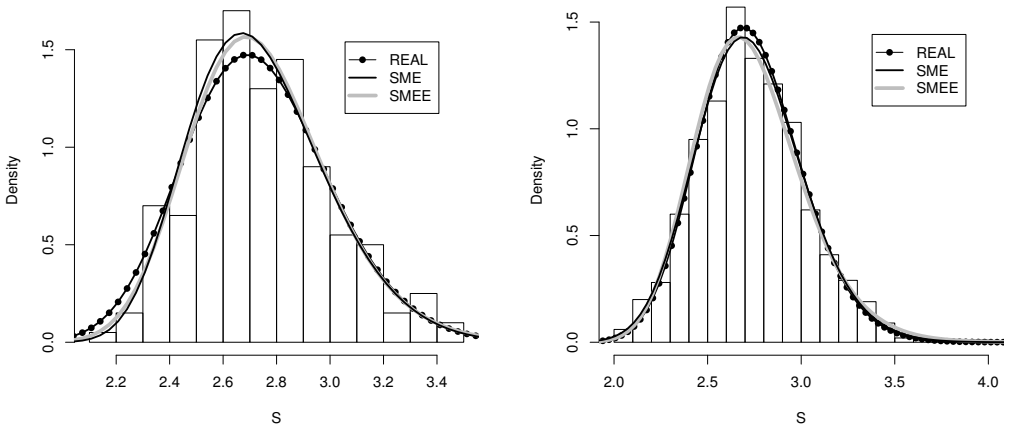
Table 3.2: Moments of $S$ for different sample sizes

### 3.3.1 Reconstruction without error estimation

We now present the results of implementing the first approach, namely the density reconstruction without estimating the additive noise.

The two panels of Figure 3.1 display the true density from which the data was sampled, the histogram that was obtained, the density reconstructed according to the SME method as well

as the density obtained according to the first extension of the SME (labelled SMEE-1). The left panel contains the results obtained for a sample of size 200 whereas the right panel the reconstruction obtained for a sample of size 1000.




(a) Results with a sample of size 200 (b) Results with a sample of size 1000

Figure 3.1: Density of the individual losses obtained by SME and SMEE-1 approaches for different samples sizes.

Even though the moments of the density obtained with the SME method coincide with the empirical moments, the moments of the density obtained by the first reconstruction method do not need to coincide with the sample moments. They only need to fall within (or at the boundary) of the error interval, which is centered at the sample moments. In Table 3.3 we list the moments that those densities determine for each sample size.

| Size | Moments of S | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ | $\mu_7$ | $\mu_8$ |
| 200 | 0.0176 | 0.1302 | 0.2559 | 0.3592 | 0.4405 | 0.5048 | 0.5565 | 0.5987 |
| 1000 | 0.0181 | 0.1319 | 0.2579 | 0.3612 | 0.4424 | 0.5066 | 0.5581 | 0.6002 |

Table 3.3: Moments of the Maxentropic densities reconstructed according to SMEE-1 approach

Let us now run some simple quality tests. As we have the true density from which the data was sampled, we can perform three comparisons. We compute the actual $L_1$ and $L_2$ distances

between the densities and between the densities and the histograms. Actually this is a bin-dependent computation and not a truly good measure of the quality of reconstruction, but we carry it out as a consistency test.

In Table 3.4 we display the results of the computations. The distances between the empirical densities (histograms and true density) and the Maxentropic densities are calculated according to the L1 and L2 distances described in the Appendix A.1.

Note that the distances between the true density and the Maxentropic reconstruction are much smaller that those between the Maxentropic or true and the histogram, and that the distance of the true and the Maxentropic distances and the histogram are similar. Thus, the reconstruction methods are performing well.

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | True density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SMEE-200 | 0.1599 | 0.1855 | 0.1504 | 0.1753 | 0.0527 | 0.0583 |
| SME- 200 | 0.1599 | 0.1855 | 0.1449 | 0.1727 | 0.0668 | 0.0761 |
| SMEE-1000 | 0.1042 | 0.1077 | 0.1052 | 0.1158 | 0.0619 | 0.0577 |
| SME-1000 | 0.1042 | 0.1077 | 0.0973 | 0.1044 | 0.0307 | 0.0289 |

Table 3.4: $L_1$ and $L_2$ distances between the True density, Maxentropic densities (SME & SMEE-1) and the histogram

Another measures of quality of reconstruction are the $L_1$ and $L_2$ distances between cumulative distribution functions, these measures are $MAE$ and $RMSE$. To distinguish them from the standard distances, $MAE$ stands for "mean average error" and $RMSE$ stands for "root mean square error". These measures are described in the Appendix A.1. The results are displayed in Table 3.5.

It is intuitively clear, and confirmed in the results displayed in the table, that the larger the size of the sample the better the estimation results for both methods, that is, the SME and SMEE.

In the Table 3.6 we can see details of the convergence of the method used according to the sample size. Clearly the SMEE method involves less iterations and machine time, but involves a lower value for the gradient.

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | True density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SMEE-200 | 0.0158 | 0.0199 | 0.0104 | 0.0149 | 0.0106 | 0.0115 |
| SME -200 | 0.0158 | 0.0199 | 0.0089 | 0.0103 | 0.0105 | 0.0115 |
| SMEE-1000 | 0.0064 | 0.0076 | 0.0072 | 0.0090 | 0.0104 | 0.0119 |
| SME -1000 | 0.0064 | 0.0076 | 0.0043 | 0.0056 | 0.0053 | 0.0060 |

Table 3.5: MAE and RMSE distances between the True density, Maxentropic densities (SME & SMEE-1) and the histogram

| **Approach** | **SMEE (200)** | **SME (200)** | **SMEE (1000)** | **SME (1000)** |
|---|---|---|---|---|
| **time** | 1.79 min. | 4.33 min. | 1.40 min | 3.25 min. |
| **iterations** | 693 | 1380 | 512 | 1223 |
| **min gradient** | $1.06\times10^{-4}$ | $1.86\times10^{-7}$ | $8.81\times10^{-5}$ | $1.72\times10^{-8}$ |

Table 3.6: Convergence of SME and SMEE-1 methods for different sample sizes.

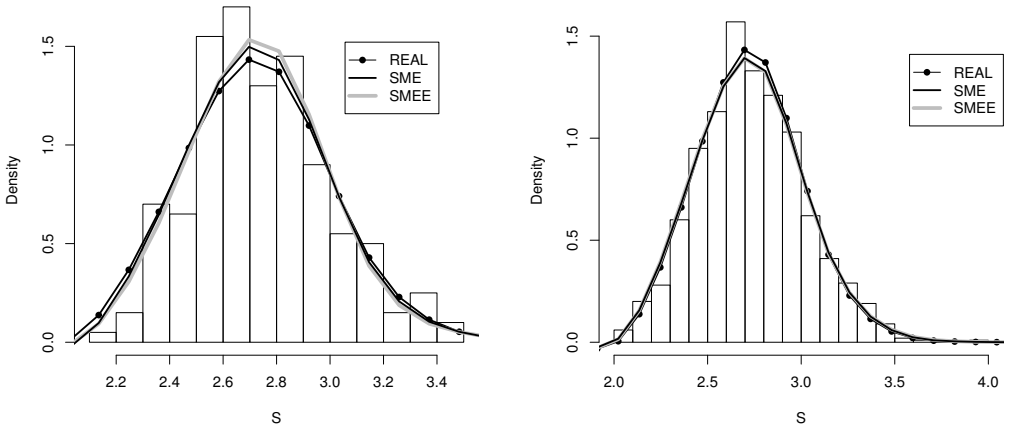
To close, let us consider Table 3.3 once more. We commented after Theorem 1, that when the multipliers were non zero, the reconstructed moments are the end points of the confidence intervals. This is not borne out by the results in that table because at the minimum the norm of the gradient is $\sim 10^{-4}$ and not exactly 0, and this tiny error explains the slight differences. Not only that, since the first method yields the boundary points of the confidence intervals as reconstructed moments, the corresponding Maxentropic density is expected to differ more from the true one than that of the center of the confidence interval.

### 3.3.2 Density reconstruction with error estimation

Recall that this time we are provided with moments measured (estimated) with error, and that we have prior knowledge about the range of the error about the moment. We are know interested in determining a density, the moments that it determines and an estimate of the additive measurement error.

In each panel of Figure 3.2 we again display 4 plots: Along with the true density and the histograms generated by sampling from it, the reconstructed Maxentropic density (labeled

SME) determined by the original sample moments, and the Maxentropic densities (labeled SMEE) provided by the second procedure, for each error interval and each sample size. Visually the SMEE densities are more close to the SME density this time.




(a) Results with a sample of size 200 (b) Results with a sample of size 1000

Figure 3.2: Density of the individual losses obtained by SME and SMEE-2 approaches for different samples sizes.

This time the optimal Lagrange multipliers $\lambda_i^*, \quad i = 1, ..., 8$, determine both the maximum entropy distributions as well as the weights of the endpoint of the confidence intervals for the estimation of the noise. With the multipliers one obtains the maximum entropy density, from which the true moments can be computed as $\hat{\mu}_k = \int_0^1 y^{\alpha_k} f_Y^*(y)dy$. The values obtained for each type of confidence interval and for each sample size are presented in Table 3.7.

| Size | Moments of S | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ | $\mu_7$ | $\mu_8$ |
| 200 | 0.0176 | 0.1304 | 0.2561 | 0.3595 | 0.4408 | 0.5051 | 0.5567 | 0.5989 |
| 1000 | 0.0181 | 0.1319 | 0.2579 | 0.3612 | 0.4424 | 0.5066 | 0.5581 | 0.6002 |

Table 3.7: Moments determined by SMEE-2 for samples of 200 and 1000

Table 3.8 displays the estimated errors as well as the corresponding weights of the endpoints of the error intervals. Keep in mind that the estimated error in the determination of each moment as well as the estimated moment add up to the measured moment.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| S | 200 | $p_k$ | 0.5143 | 0.5003 | 0.4985 | 0.4979 | 0.4975 | 0.4973 | 0.4972 | 0.4970 |
| I | | $\epsilon_k$ | -1.6×$10^{-6}$ | -1.4×$10^{-7}$ | 8.1×$10^{-7}$ | 1.2×$10^{-6}$ | 1.4×$10^{-6}$ | 1.4×$10^{-6}$ | 1.5×$10^{-6}$ | 1.4×$10^{-6}$ |
| Z | 1000 | $p_k$ | 0.5016 | 0.5071 | 0.5036 | 0.5008 | 0.4987 | 0.4972 | 0.4960 | 0.4951 |
| E | | $\epsilon_k$ | -1×$10^{-7}$ | -1.5×$10^{-6}$ | -1×$10^{-6}$ | -2.4×$10^{-7}$ | 3.6×$10^{-7}$ | 7.8×$10^{-7}$ | 1×$10^{-6}$ | 1.2×$10^{-6}$ |

Table 3.8: Weights and estimated and errors determined by the SMEE-2 approach

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | True density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SMEE-200 | 0.1599 | 0.1855 | 0.1474 | 0.1748 | 0.0761 | 0.0804 |
| SME-200 | 0.1599 | 0.1855 | 0.1449 | 0.1727 | 0.0668 | 0.0761 |
| SMEE-1000 | 0.1042 | 0.1077 | 0.0977 | 0.1051 | 0.0325 | 0.0301 |
| SME-1000 | 0.1042 | 0.1077 | 0.0973 | 0.1044 | 0.0307 | 0.0289 |

Table 3.9: L1 and L2 distances between the True density, Maxentropic densities (SME & SMEE-2) and the histogram

To measure the quality of the reconstructions we compute $L_1$ and $L_2$ distances between densities as well as the distance between distribution functions. These are displayed in Tables 3.9 and 3.10 below. Again, the distances between densities and histograms depend on the bin sizes. For a sample size of 1000 the results of the method SME and SMEE seem to coincide better and being more closer to the lognormal curve than the histogram (simulated data).

To finish, consider Table 3.11 in which we can show the details of the convergence of the

| Approach | Hist. vs. True density | | Hist. vs. Maxent. | | True density vs. Maxent. | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SMEE-200 | 0.0158 | 0.0199 | 0.0094 | 0.0127 | 0.0134 | 0.0163 |
| SME-200 | 0.0158 | 0.0199 | 0.0089 | 0.0103 | 0.0105 | 0.0115 |
| SMEE-1000 | 0.0064 | 0.0076 | 0.0043 | 0.0056 | 0.0057 | 0.0062 |
| SME-1000 | 0.0064 | 0.0076 | 0.0043 | 0.0056 | 0.0053 | 0.0060 |

Table 3.10: MAE and RMSE distances between the True density, Maxentropic densities (SME & SMEE-2) and the histogram

SMEE versus the SME in the second case. The two leftmost columns correspond compare the SMEE versus the SME for a sample of size 200 whereas the two rightmost columns compare the performance for a sample of size 1000. All things considered, it seems that the second method, that is simultaneous determination of density and measurement errors, has better performance.

| **Approach** | **SMEE (200)** | **SME (200)** | **SMEE (1000)** | **SME (1000)** |
|---|---|---|---|---|
| **time** | 37.48 seg. | 4.33 min. | 18.95 seg. | 3.25 min. |
| **iterations** | 220 | 1380 | 112 | 1223 |
| **min gradient** | $1.07\times10^{-10}$ | $1.86\times10^{-7}$ | $5.82\times10^{-10}$ | $1.72\times10^{-8}$ |

Table 3.11: Convergence of the SMEE-2 approach used for different samples sizes.

## 3.4 Concluding remarks

We presented and compared two possible ways of incorporating errors in the data into the usual maximum entropy formalism to determine probability densities. They correspond to two different demands about the resulting output of the method.

The first method goes as follows: For each $\mathbf{c}$ in the constraint space $\mathbf{C}$, one solves a maximum entropy to obtain a density $g^*(\mathbf{c})$ having $\mathbf{c}$ as constraint. The one varies $\mathbf{c}$ to find a $\mathbf{c}^*$ such that the entropy $S(g^*(\mathbf{c}^*))$ of $g^*(\mathbf{c}^*)$ is maximal among all the $S(g^*(\mathbf{c}))$. Standard duality theory is invoked along the way to obtain the $g^*(\mathbf{c}^*)$ without having to actually determine the $\mathbf{c}^*$.

The second method uses both the moments and a range for the error as input data. In the first case we simply obtain a density whereas in the other we obtain both a density as well as an estimate of the error.

Recall that the only input needed for the first method is a constraint range, and that we compared with SME with moments estimated from a large sample to have an idea of the performance of both procedures. We shall explore a more realistic case (small data set) elsewhere. In the example that we considered, both methods provide satisfactory results. The minimization procedure converges more slowly and the norm of the gradient (which is a measure of the reconstruction error) is much larger in the first one of them. This is probably due to the fact that the Hessian matrix has very small eigenvalues near the minimum, making the function too

flat there. Actually, both Hessians have very small determinant, but that corresponding to the first method is about $10^{-20}$ times smaller.

This is reflected in the reconstruction error measured by the size of the gradient of the dual entropy $\Sigma(\boldsymbol{\lambda})$ at the minimum.This is due to the flexibility of the method. Even though the $L_1$ and $L_2$ differences between the empirical density and the Maxentropic densities are small in both procedures, the second method yields a better fit, and an estimate of the measurement error as well.

# Chapter 4

# Maxentropic approach to decompound aggregate risk losses

## 4.1 Introduction

A risk manager may be faced with the following problem: she/he has obtained loss data collected during a year, but the data only contains the total number events and the total loss for that year. She/he suspects that there are different sources of risk, each occurring with a different frequency, and wants to identify the frequency with which each type of event occurs and if possible, the individual losses at each risk event.

The purpose of this chapter is to examine a combination of disentangling and decompounding procedures, to get as close as possible to that goal. The first step of the methodology is the disentangling procedure. Here, a preliminary analysis is carried out to determine the number of risks groups present. Once that is decided, the underlying model for the frequency of each type of risk is worked out. After that we use the Maxentropic techniques as second step or decompounding stage to determine the distribution of individual losses from the observed total loss.

Now, let $N_h$ denote the number losses of certain type labeled by $h = 1, ..., H$, occurring in a given time interval, $H$ being the total number of risk sources, and $X_{h,k}, k \geq 1$ a sequence of positive real-valued random variables that represent the size of the $k-$th loss of type $h$. With this notation, the quantity of interest is the compound model given by

$$S = \sum_{h=1}^{H} S_h \text{ where } S_h = \sum_{k=1}^{N_h} X_{h,k} \quad \text{with} \quad S_k = 0 \quad \text{if} \quad N_h = 0. \tag{4.1}$$

Here, $h = 1, ...., H$ labels the business line/type of the institution. In (4.1), it is supposed for each $h$ that the $X_{h,k}, k \geq 1$ are independent and identically distributed and independent of the $N_h$, and we shall denote by $X_h$ a random variable having that common distribution.

An interesting problem confronting the risk manager occurs at the level of the $S_h$. It may be the case that the frequency of events of type $h$ is actually the sum of at least two different types of events, i.e., it may happen that $N_h = N_h^1 + N_h^2$ during the observation period, even though the individual losses in each case may be the same. In this case, the risk analyst may want to know how many types of risk events of each type occur, and if possible, what is the distribution of individual losses. This is of interest because it is at that level where loss prevention or mitigation may be applied. This question motivates the problem that we shall examine here by a combination of two procedures: one (called disentangling) which consists of determining how many risk sources are there and their frequencies, and the other, the decompounding, is used to determine what is the statistical nature of the individual severities. Both procedures will be explained in Section 4.2, in which we establish the methodological aspects of the paper. We shall see that unless we know that the distributions of individual losses are the same, all that one is able to obtain is a mixture of distributions, a kind of collective individual loss distribution, which becomes the true individual distribution only when it is known that all are the same.

The disentangling procedure was proposed in Gomes and Gzyl (2014), and we briefly review it here for the sake of completeness. The decompounding problem was treated for the first time by Buchmann and Grüubel (2003). There, they estimate the frequency rate parameter and the individual severity distribution, in an univariate compound Poisson distribution, this was done through a discrete Panjer inversion. Later on, a different approach to the problem was presented in Van Es et al. (2007), who proposed a kernel type, non-parametric density estimator to obtain the individual losses of an univariate and bivariate compound Poisson distribution. They report good results in the numerical examples presented. Bøgsted and Pitts (2010) extended the work of Buchmann and Grübel (2004), and constructed a generalization of a non-parametric estimator for any compound base process given any parametric model for

the distribution of the frequencies. We shall carry out the decompounding procedure using the maximum entropy method, for which the input shall be the Laplace transform of the aggregated loss $S$.

The remainder of the paper is organized as follows: Section 4.2 is devoted to a more detailed description of the methodology. Finally, Section 4.3 exhibit a few numerical examples and a brief discussion of the results, after that we gather some concluding remarks.

## 4.2 Methodological preliminaries

In this section we lay down the basic framework in which we shall be working. We refer to Figure 4.1 for graphic display of the setup. As mentioned, the number of risk events during a certain period of time is recorded as well as the total loss during the period. It is known, or suspected, that there may be more of one source of risk present in the recorded aggregated loss, thus the first issue to take care of, is to determine the number of different risk types and their statistical nature (Disentangling Step). Once this is achieved, the next order of business is to determine the distribution of the individual risks (Decompounding Step).

In actual practice, before applying the disentangling procedure to determine the statistical nature of the various frequencies of events, a first step consisting of determining how many different type $H$ of events are, will be applied. Once we know that value ($H = 2$ in the example displayed in Figure 4.1), we apply the disentangling procedure described in Section 4.2.1. The output of this preprocessing stage is the specification of $H = 2$ random variables.

Now, if we suppose that the distribution of the individual losses of each type are the same, invoking (4.1), it is clear, for the independence assumption, that the Laplace transform $\psi(\alpha)$ of the total loss is given by

$$\psi(\alpha) = E[e^{-\alpha S}] = E[e^{-\alpha S_1}]E[e^{-\alpha S_2}] = G_{N_1}(\phi(\alpha))G_{N_2}(\phi(\alpha)) \tag{4.2}$$

where, of course, $\phi(\alpha) = E[e^{-\alpha X}]$ denotes the Laplace transform of the individual losses. Thus, once we have determined the distributions of the $N_i,\ i = 1, 2$, and computed $\psi(\alpha)$ from the data, we can solve (4.2) for $\phi(\alpha)$ and use it as input in the Maxentropic methods to obtain the probability density of the individual losses from it. This is the second stage of the process

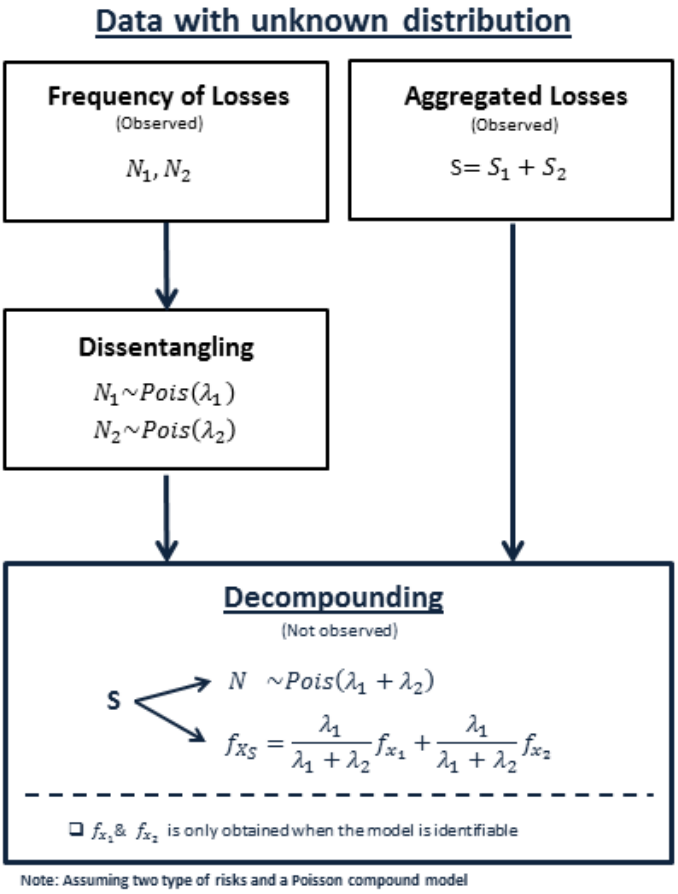



Figure 4.1: Disentangling & Decompounding procedure

displayed in Figure 4.1.

To exemplify, suppose that the frequencies are Poisson $(\ell_i)$, for $i = 1, ..., H$, and that the individual severities are distributed according to $f_{X_i}$. In this case, a simple computation shows that in this case (4.2) becomes

$$\psi(\alpha) = e^{-\sum_{h=1}^{H} \ell_h (E[e^{-\ell X_h}]-1)} = e^{-\ell(E[e^{-\alpha \hat{X}}]-1)} \tag{4.3}$$

where $\ell = \sum_{h=1}^{H} \ell_h$, and $\hat{X}$ is a random variable, where $f_{\hat{X}} = \sum_{h=1}^{H} \frac{\ell_h}{\ell} f_{X_h}$. That is, the aggregated risk is the result of compounding (1) a risk produced with a Poisson parameter equal to $\ell$ and (2) an individual loss with density equal to the weighted average of the individual densities.

Following Wang (1997) we may extend the previous result as follows

**Proposition 1.** *Suppose that the $H$ compound risks, to be aggregated, have frequencies $N_h$ with a common mixing distribution $F(\theta)$ such that, given $\theta$ the $N_{|\theta} \sim P(\theta \ell_h)$ which are independent.*

*Let* $S = \sum_{h=1}^{H} S_h$, *then* $S \sim \sum_{n=1}^{N} \hat{X}_n$ *with* $N_{|\theta} \sim P(\theta\ell)$, *where* $\ell = \sum_{h=1}^{H} \ell_h$, *and as before, the* $\hat{X}_n$ *have common density* $\sum_{h=1}^{H} \frac{\ell_h}{\ell} f_{X_h}$.

*Proof.* The proof is actually easy, and hinges on the fact that $\frac{\ell_i}{\ell} = \frac{\theta\ell_i}{\theta\ell}$ is independent of the mixing parameter $\theta$, thus proceeding as if to prove (4.3) we would be led to

$$\psi(\alpha) = \int e^{-\theta\ell(E[e^{-\alpha\hat{X}}]-1)} dF(\theta) = E[e^{-\alpha \sum_{n=1}^{N} \hat{X}_n}]$$

as claimed □

Notice that for the proposed mixing we have $\ell_i / \sum_j \ell_j = E[N_i] / \sum j E[N_j]$. This type of mixing appears as well in what Wang (1997) calls common shock models, which include correlated Poisson frequencies. The class of frequencies, for which this is valid, includes the Poisson and Negative Binomial random variables. These results were turned into a proposal to define an equivalent individual loss distribution by Wang, which we state as

**Definition 1** (Equivalent mixture)**.** *Let* $f_{X_h}$ *denote the common probability density of the individual losses for the compound losses of the* $h-th$ *type,* $h = 1, ..., H$, *and let* $N_h$ *denote the respective loss frequency. The aggregated individual loss has a density given by*

$$f_{\hat{X}} = \sum_{h=1}^{H} \frac{E[N_h]}{E[N_{agg}]} f_{X_h}, \tag{4.4}$$

*which we shall refer to as the "equivalent mixture", where* $E[N_{agg}] = \sum_{h=1}^{H} E[N_j]$ *is the mean of the aggregate loss frequency and by* $\hat{X}$ *we denote the individual loss which compounded according to* $N_{agg}$, *yields the total aggregate risk* $S$.

Thus, the equivalent mixture of (individual) densities, is the density of a random variable, which compounded with the total frequency of events, yields the total observed losses. Our goal will be to determine the distributions of the $N_i$ and the equivalent mixture of densities, from the knowledge of the aggregate risk data.

In the general case, we are within the scope of the identifiability problem considered for example in Chapter 3 of Titterington et al. (1985). So, unless we know that the individual losses of the risk sources are identically distributed, we shall not be able to obtain the distribution of

the individual losses of each risk type. At the most, we shall be able to obtain the equivalent mixture of densities given by (4.4). At this point it is interesting to advance, that the equivalent mixture coincides reasonably well with the distribution obtained by the decompounding procedure in the numerical experiments that we carried.

### 4.2.1 Disentangling frequencies

Consider the situation in which the frequency data $N$ is such that we cannot distinguish between two or more sub-populations of risk sources. In this case it is necessary to separate those sub-populations in order to calculate the distribution of total losses for each type of risk. For that, essentially a two stage procedure is necessary. The underlying supposition is that the frequency of events of all risks can be modeled by a distribution in the $(a, b, 0)$ Panjer class as described in Table 4.1. The members of this family are characterized by a common recursion formula and two parameters that can be determined by a linear regression.

**Recursion formula**

$$p_k/p_{k-1} = a + b/k; \qquad k \geq 1$$

where $p_k = P(N = k)$ for $k \in \mathbb{N}$, and $p_k = N_k / \sum_k N_k$

| | **Density** | $a$ | $b$ | $p_0$ |
|---|---|---|---|---|
| Poisson | $p_k = \frac{e^{-\ell}\ell^k}{k!}$ | $0$ | $\ell$ | $e^{-\ell}$ |
| Binomial | $p_k = \binom{m}{k} q^k(1-q)^{m-k}$ | $-\frac{p}{1-p}$ | $(n+1)\frac{p}{1-p}$ | $(1-p)^n$ |
| Neg. Binomial | $p_k = \binom{k+r-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k$ | $\frac{\beta}{1+\beta}$ | $(r-1)\frac{\beta}{1+\beta}$ | $(1+\beta)^{-r}$ |
| Geometric | $p_k = \left(\frac{1}{1+\beta}\right) \left(\frac{\beta}{1+\beta}\right)^k$ | $\frac{\beta}{1+\beta}$ | $0$ | $(1+\beta)^{-r}$ |

where $k, \beta, \ell, r, n \in \mathbb{N}$

Table 4.1: Relation between parameters $a$ and $b$ for discrete family distributions

The recursion $p_k/p_{k-1} = a + b/k$ mentioned in Table 4.1 can be rewritten as $kr(k) = ak + b$, where we set $r(k) = p_k/p_{k-1}$. Then, the plot $kr(k)$ vs. $k$ may suggest if there is more than one subpopulation present and this may indicate the parametric family to which it belongs. For example, if the plot consists of a line with zero slope, i.e. $a = 0$, this suggests that the underlying probability distribution is a Poisson, on the contrary if the slope is positive or

negative ($a > 0$ or $a < 0$), then the underlying distribution is Negative Binomial or Binomial, respectively. The identification of these distributions will depend of the size of the data set and the degree of overlap. Additionally, identification of the models entails a visual comparison of plots with different scales, to see how steeper are the slopes of the data points.

The first stage consists of an initial analysis of the data in order to check for inconsistencies, this exploratory analysis may include a way to determine whether there are more than one sub-populations in the data. This can be done by criteria like the Akaike information criterion (AIC) proposed in Akaike (1973)-(1974), or the modified Akaike information criterion (AIC3) in Bozdogan (1987), the Bayesian integrated criterion (BIC) in Schwarz (1978), the integrated classification likelihood BIC type (ICL-BIC), see Biernacki et al. (2000) or McLachlan and Peel, (2004).

In Gomes et al. (2014), we supposed that we knew how many sub-populations were present. In the present work, we apply additional techniques like the projection pursuit analyzed in Caussius and Ruiz Gazen (2009), Stone (2004), Peña and Prieto (2001 a,b) and Werner (2003), combined with the criteria mentioned before to determine the number of clusters (that is, the number of regression lines) present. Also, one can apply the elbow method described in Horthon and Everritt (2000) or the negentropy criterion described in Ramaswamy (1993) or Swait (2011) as confirmation tests of the number of clusters in the mixture.

Once we know the number of clusters, we can proceed and apply a combination of EM and $k$-means procedures to infer the values of the coefficients $a$ and $b$ for each regression. Then, recurring to Table 4.1, we determine the parameters of the corresponding distribution. The quality of the parameters obtained is determined by a chi-squared test. As mentioned before, the value of the slope is a key indicator of which of the $(a, b, 0)$ models should be selected.

Additionally, it is also possible to compare the appropriateness of the distribution by looking at the relationship between the variance and the mean, or visualizations. These may help to improve the preliminary results obtained with the Panjer diagnostic display. We want to insist on the fact that this is not a mechanical procedure. It may involve handling some parameters by trial and error and visual comparisons with different family distribution plots, which may also help in the identification of the underlying model. But, as stressed in Panjer (2006), we must remember that whatever model we select it is only an approximation of reality, so our goal

is to determine a model that is good enough, to use to answer the question being considered, which depends on the particular application. Besides, the decision can be influenced by the success of particular models in similar situations, or the value of a particular model for its intended use.

To sum up, this methodology gives quite reasonable estimates especially for large samples and small values of $H$. Unfortunately, this procedure involves handling some parameters by either trial and error, which could be complicated in cases in which the counting distribution are too close from each other, distorting the overall picture. Even so, this will be useful as starting point for other methodologies or combinations of them. Additionally, previous knowledge about the data might be useful for the analysis. For example, in operational risk, the Poisson as well as the Negative Binomial models are adequate to describe the frequency of the losses. That is, previous experience should be brought in to determine a parsimonious model.

### 4.2.2 Decompounding Model

As we mentioned before, sometimes it may not be possible to record the individual severities in each risk event separately, then the only available data consists of aggregated losses. Despite the fact that this information may be enough to infer the distributions of the aggregated loss, interesting details and valuable information for risk management is hidden. This motivates the question: Is it possible to infer the distribution of the individual severities from that of the aggregated loss?. Below, we shall see that in some cases this is possible, and that maximum entropy methods provide an efficient way to do it.

In the case in which all individual losses have a common distribution, the connection between the distribution of individual severities and the aggregate loss, is contained in the extended version of (4.2), namely

$$\psi(\alpha) = E[e^{-\alpha S}] = \prod_{i=1}^{H} E[e^{-\alpha S_i}] = \prod_{i=1}^{H} G_{N_i}(\phi(\alpha)) \tag{4.5}$$

Thus, we can solve for $\phi(\alpha)$ in terms of $\psi(\alpha)$, and then we can apply the Maxentropic procedures described in Chapters 2 and 3 to determine the distribution of individual severities from it. When it makes sense to suppose that the densities of the individual losses are all equal, that is, that $\phi_1(\alpha) = ... = \phi_H(\alpha) \equiv \phi(\alpha)$, we may solve (4.5) for $\phi(\alpha)$, either analytically or

numerically. Otherwise we have to resort on (4.4) to obtain the equivalent mixture of densities.

In Table 4.2, we show the relationship between $\phi(\alpha)$ and $\psi(\alpha)$ for the different models of frequency distributions in the $(a, b, 0)$ class for the case of only one source of risk. But in the general case we must use (4.5).

| **Starting point for the decompounding procedure** | |
|---|---|
| Distributions | $\phi(\alpha_k)$ |
| Poisson($\ell$) | $\frac{1}{\ell} ln(\psi(\alpha_k)) + 1$ |
| Binomial ($n$) | $\frac{\psi(\alpha_k)^{(1/n)} - (1-p)}{p}$ |
| Neg. Binomial($r$, $\beta$) | $\frac{1-\psi(\alpha_k)^{(-1/r)}}{\beta} + 1$ |
| Geometric($\beta$) | $\frac{1-\psi(\alpha_k)^{-1}}{\beta} + 1$ |

Table 4.2: *Laplace transform of the Severities*

We point out that the Poisson case and the case covered by Proposition (1) are of special interest for us. In this case we are in the framework of the list of examples of Table 4.2. In this case we (1) compute numerically the Laplace transform of the aggregate losses, (2) use the disentangling procedure to identify the mixture of frequencies and then (3) we use the decompounding procedure to obtain the individual severity distribution or a mixture of probability densities. In this case, with some more information we may be able to obtain the individual severities for each risk.

Thus, $\psi(\alpha)$ is calculated through

$$\psi(\alpha_k) = E(e^{-\alpha S}) = \frac{1}{T} \sum_{n=1}^{T} e^{-\alpha S_n} \tag{4.6}$$

in which $T$ is the number of data points, $S_n$ is the $n-$th loss recorded and $\alpha_k$ is the $k-$th value of the Laplace transform parameter. Actually, this computation is the first step for the determination of the probability density of the aggregate loss distribution, as we stated in Chapter 2 and 3. A detail to be kept in mind, is the relationship between the $\psi(\alpha)$ and the fractional moments of the density $f_S$, contained in the following simple lemma.

**Lemma 3.** *Let $S$ be a compound random variable describing losses, and suppose that it has a*

*density* $f_S$. *Then*

$$\psi(\alpha) = E[e^{-\alpha S}] = P(N=0)+(1-P(N=0))\int_0^1 y^\alpha f_Y(y)dy. = P(N=0)+(1-P(N=0))\mu(\alpha).$$

The proof consists of conditioning out the event $\{S=0\} = \{N=0\}$ and making the change of variables $Y = e^{-S}$. The probabilities $P(N=0)$ are obtained in the disentangling step.

## 4.3 Maximum entropy methods

The maxent methodologies that we use, in this chapter, are the standard maximum entropy method (SME) described in Section 2.3.1 of Chapter 2 and the maximum entropy with errors (SMEE) discussed in Section 3.2.1 of Chapter 3.

## 4.4 Numerical examples

We will illustrate the use of our methodology considering several cases, in which we include different data features and procedures. The results for each of the cases were obtained by simulating samples corresponding to different frequencies and severities from the same or different (continuous) distribution, and then aggregating the resulting compound sums. The pseudo-code could be detailed as:

**Sample generation procedure**

- For each $h : h = 1, ..., H$, generate a random loss count $N_h$ from a discrete distribution $F_{N_h}$ with mean $E[N_h]$ and variance $VaR[N_h]$.
- For each $h$, generate a sequence of individual losses $X_{h,n}$, for $n = 1, ..., N_h$ from a continuous distribution $F_{X_h}$.
- Set $S_h = \sum_{n=1}^{N_h} X_{h,n}$, loss for each line $h$.
- Set $S = \sum_{h=1}^{H} S_h$, which is the total compound loss that we need.

To test the proposed methodology, we consider different combinations of frequencies and individual severity distributions. In several of the cases, we use the Poisson or the Negative Binomial as counting distributions because they are rather common in insurance and operational

risk models, and more important, because they are members of the $(a, b, 0)$ class. The title of each subsection describes the combination of frequency/individual severity used. In the cases considered we describe in more detail the disentangling and decompounding procedures, and then we just summarize the results obtained.

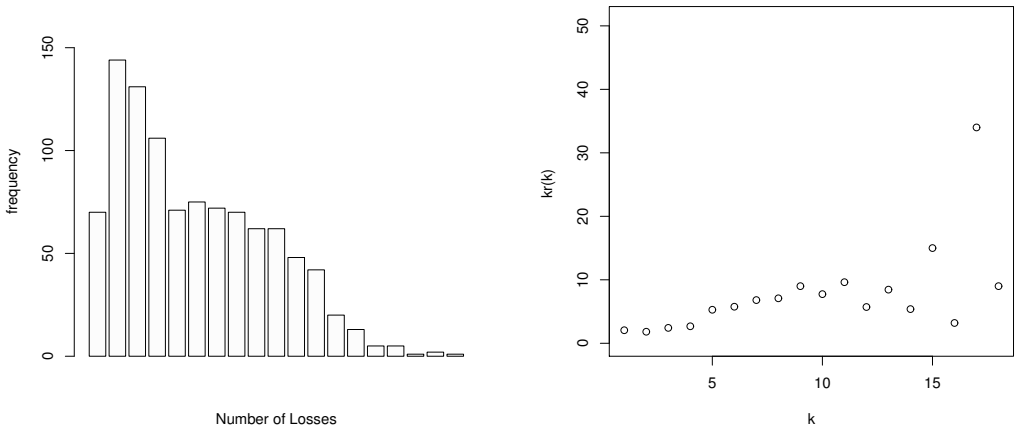
In all numerical examples we shall measure the quality of the reconstructions by several distances between the reconstructed density and the histogram as well as between the reconstructed density and the equivalent density mixture. For that we shall use the $L_1$ and $L_2$ distances with the caveat that such comparison depends on the bins assigned to the histograms, and two distances, the MAE and the RSME that use data points and cumulative distributions for comparison.

### 4.4.1 Case 1: Poisson frequencies and Lognormal individual losses

For this case we suppose that the aggregate risk has two sources, the frequencies of each of them being a Poisson distribution with parameters $\ell_1 = 2$, $\ell_2 = 8$, and that the individual severities $X_1$, $X_2$ follow a common lognormal distributions, $X \sim lognormal(-1, 0.25)$. All the variables are supposed independent. Besides we consider a sample of size 500 to compute $S = S_1 + S_2$, and all that we record, or consider as given data, is the total number of risk events and the total loss in each sample.

#### 4.4.1.1 Disentangling procedure

This is the first step in our methodology and we shall describe it in much detail. In Figure 4.2a we show a histogram of the frequency of losses. We need to determine from the simulated data the number of sub-populations and their distributions. If we suppose that they belong to the $(a, b, 0)$ class and we make use of the plot based on the Panjer recursion formula, we observe that patterns appear as shown in Figure 4.2b. There we observe a group of values around the value two, and another group of points that is more dispersed. Also, towards the right end of the abscissa axis we observe some larger values. Additionally, the groups observed in Figure 4.2b show little slope, and the only significant increase in the values occurs at the end. This seems indicate that the underlying distribution could be a mixture of Poisson distributions. On the other hand, by re-scaling the vertical axis of the Figure 4.2b the slope of the points




(a) Frequencies of the losses (b) Panjer recursion plot

Figure 4.2: Histogram of frequencies and Panjer Recursion plot. Case 1.

would look steeper, and an univariate Negative Binomial distribution can also be a possible candidate. But notice that the shape of the histogram of Figure 4.2a does not seem to suggest that.

To get an idea about the number of possible groups that are present in the data sample, it is customary to utilize a variety of information criteria measures like the AIC, AIC3, BIC and ICL-BIC values (see the notes in Table 4.3). These measures address the goodness-of-fit of the clustering method, and they are defined so that bigger is better. Additionally, we estimate the negentropy measure which indicates how well discriminated or separated the classes seem to be, based on posterior probabilities (as before the higher the measure, the better). The results of these estimations are listed in Table 4.3, where we consider a number of components that varies between one and six. This range was selected according to the result obtained with the sum of squared errors (SSE) for a broad number of clusters. This methodology is known in the clustering literature as the Elbow Method and it consists of calculating the sum of the squared distance between each member of a cluster to its cluster centroid, in order to give us an idea of the possible number of groups.

Table 4.3 shows that the information criteria is not very conclusive, these values indicate that the number of groups are between three and five. On the basis of the negentropy measure we decided to begin supposing that there are four sub-populations present.

| N. of components | AIC | AIC3 | BIC | ICL-BIC | Negentropy |
|---|---|---|---|---|---|
| 1 | 255.2 | 259.2 | 259.0 | 259.0 | 1 |
| 2 | 253.6 | 260.6 | 260.2 | 266.1 | 0.7763 |
| 3 | 231.5 | 241.5* | 240.9* | 244.9 | 0.9043 |
| 4 | 229.4 | 242.4 | 241.7 | 243.7* | 0.9616* |
| 5 | 228.4* | 244.4 | 243.5 | 247.8 | 0.9292 |
| 6 | 229.8 | 248.8 | 247.7 | 251.4 | 0.9458 |

* Value of g given by criterion.

$AIC = -2loglik + 2H$, $AIC3 = -2loglik + 3H$

$BIC = -2loglik + Hlog(n)$, $ICL = BIC + 2 * ENT$

$Negentropty = 1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{H} -p_{il} \cdot ln(p_{il})}{n \cdot ln(H)}$

$p_{il}$: is the posterior probability that the element $i$ being in group $l$

$H$: number of groups, $n$: number of elements in the data

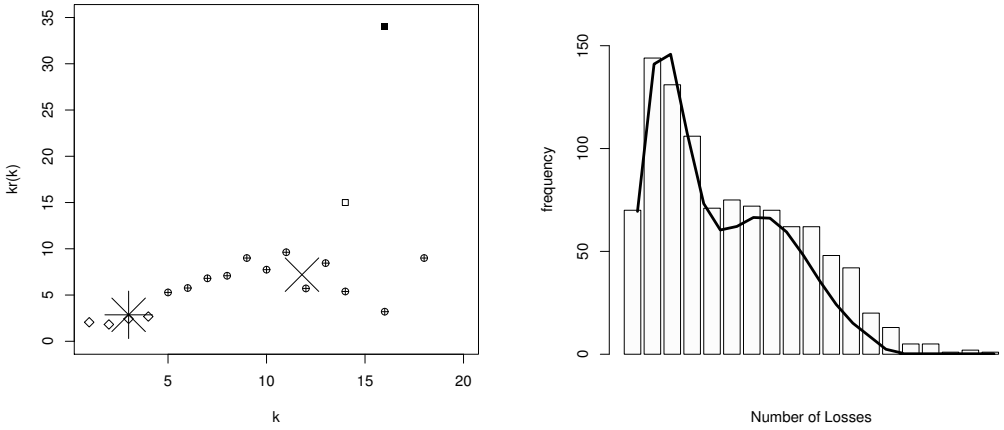
Table 4.3: *AIC, AIC3, BIC, ICL-BIC & Negentropy values. Case 1.*

Additionally, more advanced methodologies may be brought into the analysis. For example, the projection pursuit method of Peña and Prieto (2001a,b) gives us a possible clustering without the need of introduce in advance the number of groups. This algorithm detects four groups in our data, two of them being clusters and the other two are isolated points or outliers (the largest values). This result is equivalent to the one obtained with the EM algorithm when the input is the discrete data and the number of groups is $H = 4$. The results are displayed in Figure 4.3a.

In Figure 4.3a we see that the two largest clusters have almost zero slope, this indicates that the two groups follow a Poisson distribution. Combining the results with the $k$-means the Poisson parameters are the centers with values of 2.08 and 7.64 for each group. Rounding these values, the density obtained is showed in Figure 4.3b as a mixture of Poisson distributions with different parameters.

Looking at Figure 4.3b we conclude that this density mixture provides a satisfactory fit to the data. Considering the $\chi^2$ goodness-of-fit test statistic we find a satisfactory fit ($\chi^2_o$=9.512, $\rho$=0.5747, 11 degrees of freedom ). Additionally, comparing this result with a simple Poisson and a simple Negative Binomial model, the mixture model always gives superior fit.

Now that we have identified the frequency distributions, we proceed to calculate the severity distribution of the individual losses through Maxentropic methodologies. The input is the numerical computation of the Laplace transform (as in eq. 4.6).




(a) Disentangling Result (b) Mixture of Poisson distributions

Figure 4.3: Results: Disentangling step. Case 1.

Thus, by (4.3) we model the combined frequency $N$ by a Poisson distribution with $E[N] = E[N_1] + E[N_2] = \ell_1 + \ell_2 \approx 10$ and $Var[N] = Var[N_1] + Var[N_2] = \ell_1 + \ell_2 \approx 10$, because we are supposing that they are independent. Then, the combined individual severity distribution may be calculated by

$$f_{\hat{X}} = \frac{E(N_1)}{E(N_1) + E(N_2)} \times f_{X_1} + \frac{E(N_2)}{E(N_1) + E(N_2)} \times f_{X_2}$$

where $f_{X_1}$ and $f_{X_2}$ are the probability densities of the individual severities that are unknown. As this is usually the case, by the use of the Maxentropic methodology we can decompound the aggregate losses to obtain the distribution of the individual severities that produces the observed aggregate loss. We take up this task next.

#### 4.4.1.2 Decompounding procedure

Once we have determined a model for the frequency distribution, we implement the second stage of the procedure. We consider two versions on the Maxentropic methodology: reconstruction with and without error in the data.

The inputs for the Maxentropic methods are shown in Table 4.4. Here we observe the numerical calculation of the Laplace transform $\psi(\alpha)$ of the aggregate losses along with the calculation of the Laplace transform $\phi(\alpha)$ of the individual severity using the fact that the

combined frequency is Poisson and the computations displayed in Table 4.2.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\psi(\alpha_k)$ | 0.0064 | 0.0515 | 0.1229 | 0.1978 | 0.2671 | 0.3284 | 0.3818 | 0.4283 |
| $\phi(\alpha_k)$ | 0.4959 | 0.7034 | 0.7903 | 0.8380 | 0.8680 | 0.8886 | 0.9037 | 0.9152 |

Table 4.4: Numerical approximation of the Laplace transform of the aggregate losses and individual losses. Case 1.

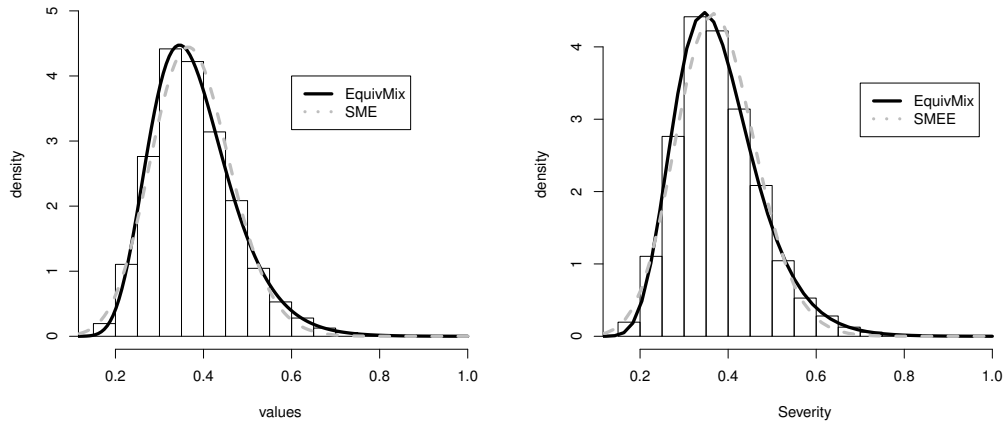
To implement the Maxentropic methodology for data with errors we need an interval for the numerical determination of the Laplace transform of the individual losses, which are shown in Table 4.5, these intervals are small and should contain the values of Table 4.4, they can be determined at the 1% or at the 5% confidence level using a bootstrapping procedure.

| $\phi(\alpha_k)$ | Confidence Interval |
|---|---|
| $\phi(\alpha_1)$ | [0.4948, 0.4966] |
| $\phi(\alpha_2)$ | [0.7032, 0.7039] |
| $\phi(\alpha_3)$ | [0.7903, 0.7907] |
| $\phi(\alpha_4)$ | [0.8380, 0.8383] |
| $\phi(\alpha_5)$ | [0.8680, 0.8682] |
| $\phi(\alpha_6)$ | [0.8886, 0.8888] |
| $\phi(\alpha_7)$ | [0.9037, 0.9038] |
| $\phi(\alpha_8)$ | [0.9151, 0.9153] |

Table 4.5: Intervals for the Laplace of the simulated individual losses. Case1.

The results of the Maxentropic methodology for the inputs of the Tables 4.4 and 4.5 are shown in Figure 4.4. In the left panel of Figure 4.4 we display the results of the SME approach when the entries Table 4.4 are used as inputs. All curves seem to be close to the histogram of simulated data and to the equivalent mixture. In Table 4.9 we show the MAE and RMSE distances between them, and in Table 4.10 we show the $L_1$ and $L_2$ distances between them. Keep in mind that this comparison is possible because the simulated data and the equivalent mixture density are known to us, but in actual practice this may be unavailable.

The right panel of Fig. 4.4 displays the results the application of the SMEE to the inputs of Table 4.5. All the curves seem to be close to the histogram and the equivalent mixture density




(a) Results with SME (b) Results with SMEE

Figure 4.4: Results: Decompounding Step. Case 1.

| Approach | Hist. vs. Real density | | Hist. vs. Maxent | | Real density vs. Maxent | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SME | 0.0024 | 0.0029 | 0.0129 | 0.0143 | 0.0129 | 0.0143 |
| SMEE | 0.0024 | 0.0029 | 0.0143 | 0.0162 | 0.0142 | 0.0159 |

Table 4.6: MAE and RMSE distances between reconstructed densities, original histogram and densities. Case 1.

(the mixture of lognormals). The distances between the Maxentropic reconstructions and the data are shown in Tables 4.6 and 4.7.

#### 4.4.1.3 Identifiability

We know from (4.3) that the severities follow a mixture of distribution of the form

$$f_{\hat{X}} = \frac{E(N_1)}{E(N_1)+E(N_2)} \times f_{X_1} + \frac{E(N_2)}{E(N_1)+E(N_2)} \times f_{X_2}$$

the Maxentropic methods give us the value of $f_{\hat{X}}$ and the disentangling methodology give us the weights of the distribution by assumption, and we can say that $f_{X_1}$ and $f_{X_2}$ are identifiable. If the two severities were different then the mixture is said to be unidentifiable, because there will be only one equation but two unknowns, in this case $f_{X_1}$ and $f_{X_2}$.

| Approach | Hist. vs. Real density | | Hist. vs. Maxent | | Real density vs. Maxent | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SME | 0.1120 | 0.2148 | 0.1443 | 0.2573 | 0.1116 | 0.1740 |
| SMEE | 0.1120 | 0.2148 | 0.1469 | 0.2633 | 0.1152 | 0.1822 |

Table 4.7: L1 and L2 distances between reconstructed densities, original histogram and densities. Case 1.

### 4.4.2 Case 2: Poisson frequencies with Beta and Frechet for the individual losses

This example is similar to Case 1, except that now the densities of the individual losses are quite different among themselves. They are specified as follows: For the frequencies we consider $N_1 \sim Poisson(1)$ and $N_2 \sim Poisson(2)$, and the individual loss distributions are, respectively, $X_1 \sim Frechet(2, 4)$ and $X_2 \sim Beta(1, 5)$. This time the sample size was 1000. Simulating a sample from these populations we obtain the input for the problem, and we apply the disentangling/decompounding procedure once more.

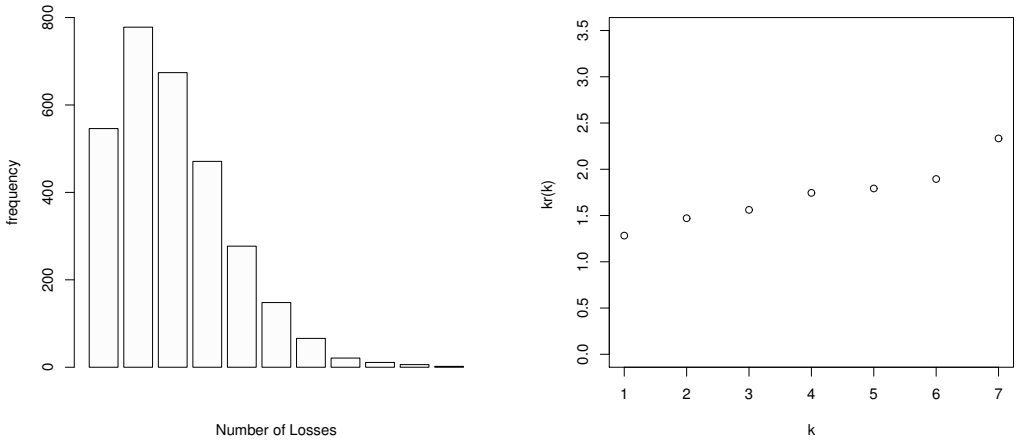
#### 4.4.2.1 Disentangling procedure

The results of applying the Panjer recursion are shown in Figure 4.5b. There we can observe an increasing pattern, but the increase does not seem to be large. If we truncate the values resulting of the Panjer recursion, we can see two horizontal lines (without slope) with y-values equal to 1 and 2, which suggests that we have mixture of Poisson distributions with different values of the parameters. The result of the EM algorithm is presented in Figure 4.6a and the mixture distribution is displayed in Figure 4.9b.

#### 4.4.2.2 Decompounding procedure

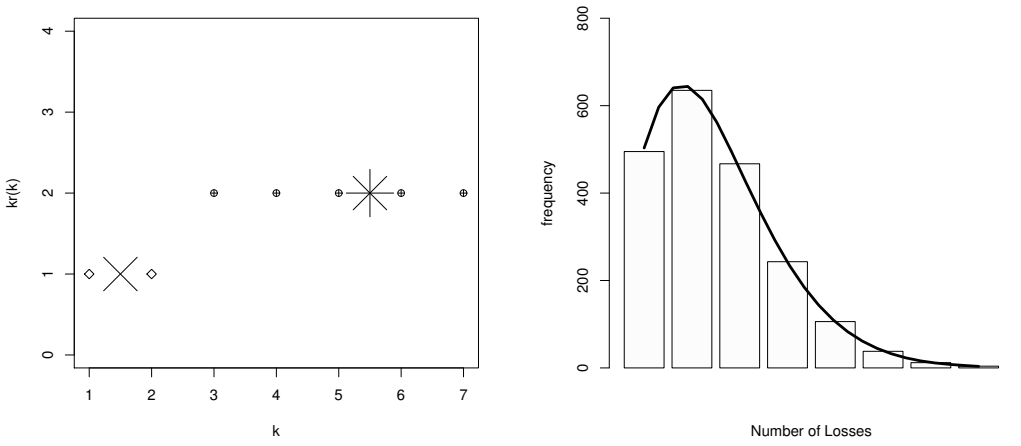
The inputs for the maxent methods are shown in Table 4.8. In Figures 4.7a and 4.7b respectively, we present the result of the decompounding procedure using the SME and the SMEE methods, along with the plot of the equivalent mixture of densities.

Even though the reconstruction in the middle region is not that satisfactory, the tail behavior is quite good. But keep in mind that we are using only $K = 8$ fractional moments to




(a) Frequencies of the losses (b) Panjer recursion plot

Figure 4.5: Histogram of frequencies and Panjer Recursion plot. Case 2.



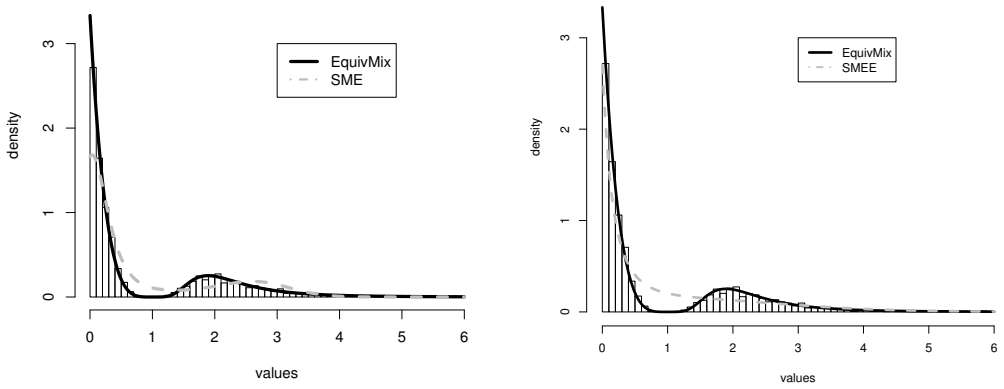

(a) Disentangling Result (b) Mixture of Poisson distributions

Figure 4.6: Results: Disentangling step. Case 2.

| $M$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\psi(\alpha_k)$ | 0.2117 | 0.3164 | 0.3998 | 0.4673 | 0.5221 | 0.5672 | 0.6048 | 0.6365 |
| $\phi(\alpha_k)$ | 0.4824 | 0.6163 | 0.6944 | 0.7463 | 0.7833 | 0.8109 | 0.8323 | 0.8494 |

Table 4.8: Numerical approximation of the Laplace transform of the aggregate losses and individual losses. Case 2.

carry out the reconstruction. Perhaps the incorporation of more information may yield better reconstruction results, but the given eight moments are those used for all the reconstructions that we carried out.




(a) Decompounding result. (b) Results with SMEE

Figure 4.7: Results: Decompounding step. Case 2.

| Approach | Hist. vs. Equi. Mix. | | Hist. vs. Maxent | | Equiv. Mix. vs. Maxent | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SME | 0.0193 | 0.0233 | 0.0595 | 0.0709 | 0.0776 | 0.0908 |
| SMEE | 0.0193 | 0.0233 | 0.0515 | 0.0643 | 0.0658 | 0.08024 |

Table 4.9: MAE and RMSE distances between reconstructed densities, original histogram and densities. Case 2.

| Approach | Hist. vs. Equi. Mix | | Hist. vs. Maxent | | Equiv. Mix. vs. Maxent | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SME | 0.4328 | 0.5999 | 0.4179 | 0.3954 | 0.4551 | 0.3913 |
| SMEE | 0.4328 | 0.5999 | 0.4978 | 0.4763 | 0.4554 | 0.3472 |

Table 4.10: L1 and L2 distances between reconstructed densities, original histogram and densities. Case 2.

### 4.4.3 Case 3: Binomial frequencies and Beta, Gamma, Weibull, Frechet individual losses

We consider now a more elaborate example. It consists of a larger collection of risks, each having a different individual loss density. We shall generate the data by aggregating four sources of risk in a sample of size 1000. We suppose that the individual frequencies are of the binomial type: $N_1 \sim Binomial(15, 0.50)$, $N_2 \sim Binomial(30, 0.65)$, $N_3 \sim Binomial(30, 0.90)$ and $N_4 \sim Binomial(40, 0.8)$. The individual losses to be aggregated have (in the same order) distributions given by $X_1 \sim Beta(1, 25)$, $X_2 \sim Weibull(1, 0.1)$, $X_3 \sim Frechet(0.01, 2)$ and $X_4 \sim Gamma(0.1, 3)$.

Note that now they are of quite different nature. For the disentangling procedure it is convenient to have Table 4.11, in which we collect a summary of the parameters of the binomial distributions used for the simulation.

| - | $\mu$ | $\sigma^2$ | $p$ | $n$ |
|---|---|---|---|---|
| $Group_1$ | 7.5 | 3.75 | 0.5 | 15 |
| $Group_2$ | 19.5 | 6.825 | 0.65 | 30 |
| $Group_3$ | 27 | 2.7 | 0.90 | 30 |
| $Group_4$ | 32 | 6.4 | 0.8 | 40 |

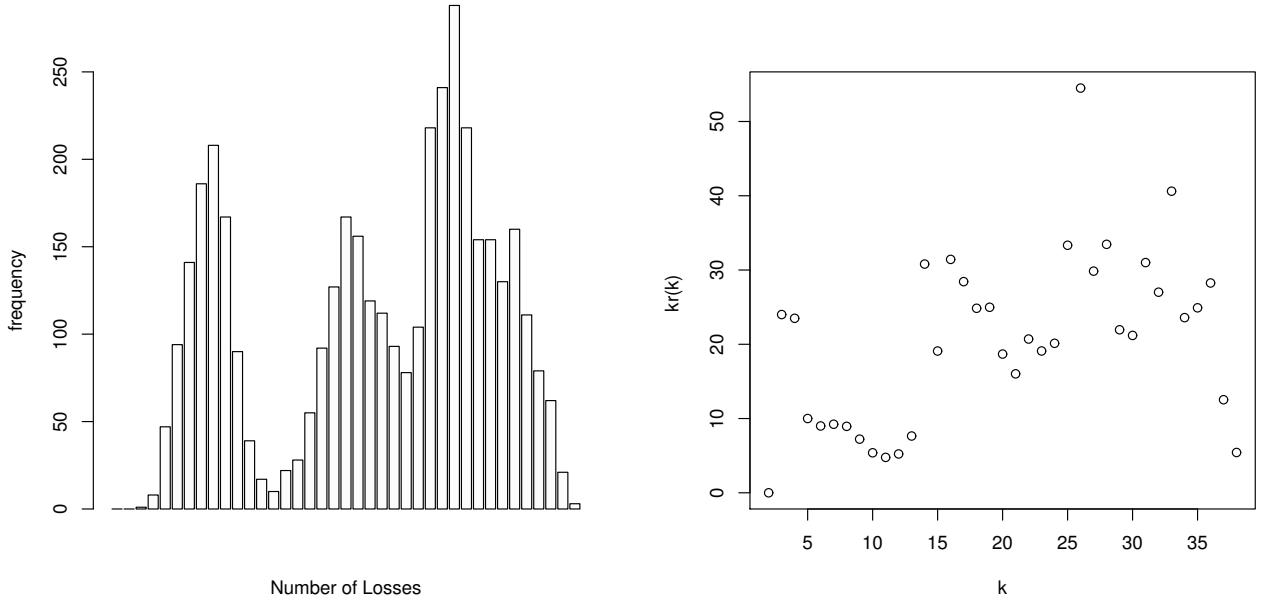
Table 4.11: Binomial parameters of each group. Case 3.

#### 4.4.3.1 Disentangling procedure

We proceed as in the previous examples. In Figure 4.8a we can see the histogram of the frequency of events and next to it, in (4.8b), the result of the plot according to the Panjer recursion. Here we can observe different groups of points that seem to have a negative slope. The result obtained applying the EM algorithm is presented in Figure 4.9a, and the mixture distribution may be seen in Figure 4.9b.

In Table 4.8 we present a summary of the estimated parameters obtained with the regression formulation discussed in before along with a trial and error search in order to improve the results.

We considered five groups according to the result of the Negentropy measure, as the other




(a) Frequencies of the losses (b) Panjer recursion plot

Figure 4.8: Histogram of frequencies and Panjer Recursion plot. Case 3

four criteria does not seem to be consistent as observed in Table 4.12. In the left panel of Figure 4.9a we observe the five groups obtained by the EM procedure. The first set of points is excluded and for the calculations the other four sets of data are used.

| **N. of components** | **AIC** | **AIC3** | **BIC** | **ICL-BIC** | **Negentropy** |
|---|---|---|---|---|---|
| 1 | 573.8 | 577.8 | 580.4 | 580.4 | 1 |
| 2 | 564.2 | 572.2* | 577.5* | 586.0* | 0.8330 |
| 3 | 561.1 | 573.1 | 581.1 | 593.6 | 0.8457 |
| 4 | 560.5 | 576.5 | 587.1 | 597.2 | 0.9011 |
| 5 | 557.8* | 577.8 | 591.1 | 597.8 | 0.9434* |
| 6 | 558.7 | 582.7 | 598.7 | 608.0 | 0.9291 |

* Value of g given by criterion.

$AIC = -2loglik + 2H$, $AIC3 = -2loglik + 3H$

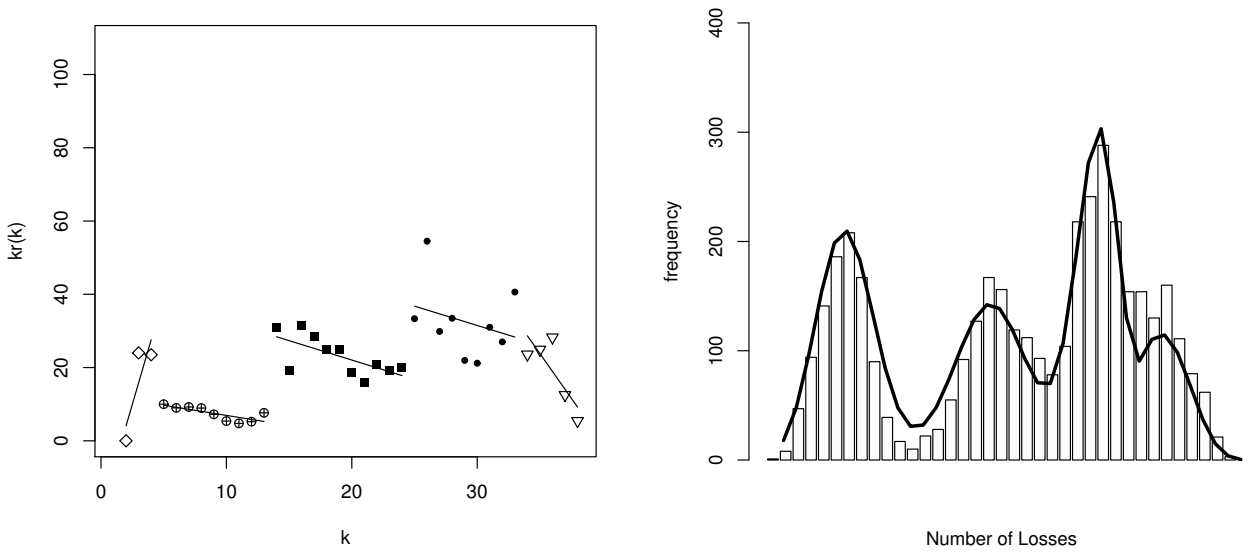
$BIC = -2loglik + Hlog(n)$, $ICL = BIC + 2 * ENT$

$Negentropy = 1 - \frac{\sum_{i=1}^{n}\sum_{l=1}^{H} -p_{il} \cdot ln(p_{il})}{n \cdot ln(H)}$

$p_{il}$: is the posterior probability that the element $i$ being in group $l$

$H$: number of groups, $n$: number of elements in the data

Table 4.12: *AIC, AIC3, BIC, ICL-BIC & Negentropy values. Case 3.*




(a) Disentangling Result

(b) Mixture of Binomial distributions

Figure 4.9: Results: Disentangling Results. Case 3.

| - | $\mu$ | $\sigma^2$ | $p$ | $n$ |
|---|---|---|---|---|
| $Group_1$ | 7.67 | 4.94 | 0.36 | 22 |
| $Group_2$ | 19.2 | 7.68 | 0.60 | 32 |
| $Group_3$ | 27.3 | 2.457 | 0.91 | 30 |
| $Group_4$ | 32.37 | 5.50 | 0.83 | 39 |

Table 4.13: Results: Binomial parameters of each group. Case 3.

#### 4.4.3.2 Decompounding procedure

To obtain the severity of the individual losses we make use of both the SME and the SMEE procedures. In Table 4.14 we list the values of $\psi(\alpha)$ obtained by the simulation procedure and the corresponding values of $\phi(\alpha)$ to be used as input to the decompounding procedure.

For the SMEE procedure we consider as input an interval which include the values of $\phi(\alpha)$ of Table 4.15. This interval is usually small. We obtained by bootstrapping with replacement.

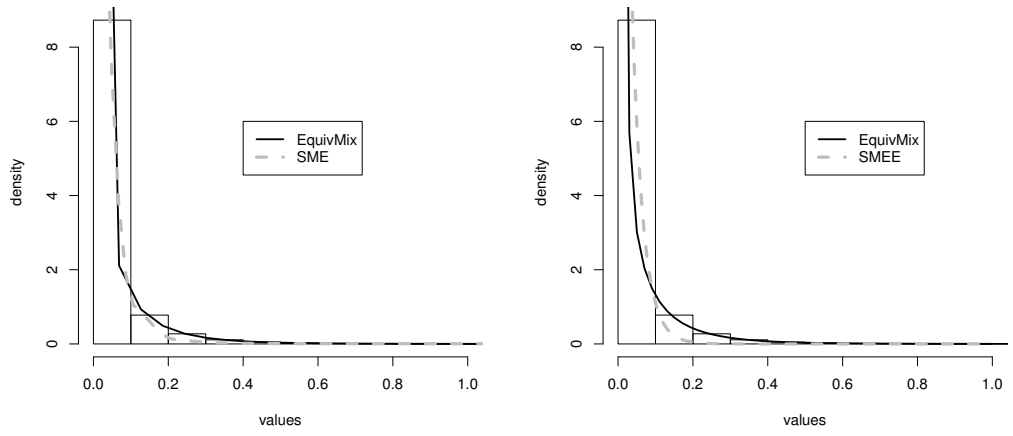
The results of the Maxentropic methodology for the inputs of Tables 4.14 and 4.15 are shown in Figure 4.10.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\psi(\alpha_k)$ | 0.0062 | 0.0688 | 0.1624 | 0.2524 | 0.3302 | 0.3957 | 0.4506 | 0.4971 |
| $\phi(\alpha_k)$ | 0.9425 | 0.9694 | 0.9791 | 0.9841 | 0.9872 | 0.9893 | 0.9908 | 0.9919 |

Table 4.14: Numerical approximation of the Laplace transform of the aggregate losses and individual losses. Case 3.

| $\phi(\alpha_k)$ | **Confid. Interval** |
|---|---|
| $\phi(\alpha_1)$ | [0.9425, 0.9426] |
| $\phi(\alpha_2)$ | [0.9694, 0.9695] |
| $\phi(\alpha_3)$ | [0.9791, 0.9792] |
| $\phi(\alpha_4)$ | [0.9841, 0.9842] |
| $\phi(\alpha_5)$ | [0.9871, 0.9872] |
| $\phi(\alpha_6)$ | [0.9892, 0.9893] |
| $\phi(\alpha_7)$ | [0.9908, 0.9909] |
| $\phi(\alpha_8)$ | [0.9919, 0.9920] |

Table 4.15: Intervals for the Laplace of the simulated individual losses. Case 3.




(a) Results with SME (b) Results with SMEE

Figure 4.10: Results: Decompounding step. Case 3.

In Table 4.16 we show MAE and RMSE distances between the reconstructed densities, original histogram and the mixture of densities computed from the input data. In Table 4.17

| Approach | Hist. vs. Equiv. mix | | Hist. vs. Maxent | | Equiv. mix vs. Maxent | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SME | 0.2120 | 0.2474 | 0.1270 | 0.1355 | 0.1052 | 0.1285 |
| SMEE | 0.2120 | 0.2474 | 0.1772 | 0.1854 | 0.0894 | 0.1009 |

Table 4.16: MAE and RMSE distances between reconstructed densities, original histogram and densities. Case 3.

| Approach | Hist. vs. Equiv. mix | | Hist. vs. Maxent | | Equiv. mix vs. Maxent | |
|---|---|---|---|---|---|---|
| | L1-norm | L2-norm | L1-norm | L2-norm | L1-norm | L2-norm |
| SME | 0.3943 | 1.9240 | 0.3163 | 1.1964 | 0.3147 | 1.1085 |
| SMEE | 0.3943 | 1.9240 | 0.3490 | 1.2403 | 0.3353 | 1.1201 |

Table 4.17: L1 and L2 distances between reconstructed densities, original histogram and densities. Case 3.

we show the distances between the densities in the L1 and L2 norms.

In the Tables 4.16 and 4.17 we can observe a decrease in the value of the errors when we use as input the interval of $\phi(\alpha_k)$ in Table 4.15.

## 4.5 Concluding comments

We presented a two stage procedure to determine the probability density of individual losses corresponding to an observed aggregate loss. The first consisting of determining how many risk sources are present and their distributions, and the second to determine the distribution of individual losses when all risk sources have a common severity, or an equivalent individual severity when not.

In the several numerical experiments that we performed, when the severities (individual losses) of the different sources are different among themselves but the frequencies are of the Poisson or Negative Binomial type, the method works quite well. We considered a case in which the frequencies were Poisson with different parameters and the individual distributions had a common Gamma distribution. The results were as expected and the individual density was perfectly reconstructed. Another example considered consisted of a family of different negative

binomial frequencies of events and the individual losses were two different Pareto distributed losses. The disentangling procedure yielded the parameters of the frequencies quite well and the individual losses turned out to be distributed according to the equivalent mixture.

But when the frequencies are of different types, the method does not necessarily work. There may be instances in which, even though the disentangling of the frequencies might be possible, the nature of the underlying individual losses can be such that no equivalent loss might be determined. This happened when we considered the case in which the data comes from the following combination: The frequencies were distributed according to $N_1 \sim Poisson(2)$, $N_2 \sim NBin(10, 0.5)$ and $N_3 \sim Bin(25, 0.8)$, and the corresponding loss densities according to $X_1 \sim Weibull(2, 4)$, $X_2 \sim Frechet(2, 4)$ and $X_3 \sim Beta(1, 5)$.

Notice that if we knew that the individual frequencies were the same, we could at least attempt to solve (4.5) numerically to obtain $\phi(\alpha)$, but this may not be always possible.

Also, we considered a small number of risk sources because when this number is large, the disentangling procedure is hard to implement, in the cases where the amount of data is small. Actually, this is an important detail, because this is a real issue in operational risk analysis. This, combined with the fact that not all of the individual severities have the same distribution, is a handicap of the method. Then, the performance of the two procedures depends on the availability of data, which is a scarce resource in banking.

We finish by addressing several issues. One relates to an important methodological aspect: How does the estimator perform on new data? In terms of the numerical examples, how do the estimators fitted on one half of the data set, perform on predicting the other half of a data set? These two are related and these have a direct answer: The procedures are the same but have to be applied afresh to each new data set. As with operational risk: data sets can be augmented as time goes on. More data would mean better estimates of the loss distributions.

Another interesting issue is: Real data sets for operational risk will contain information not just on the risks but on how the data was collected. How to deal with this? Note that the problems and the methods to solve them, use very specific data, namely total number of events during a given period and total loss during that period. If the information on how the data is collected affects those numbers, and if that can be quantified, we could perhaps treat the resulting available data as data with errors in the measurement process.

An issue that we do not deal with, but will be the matter of the forthcoming chapter, is the relationship between the size of the sample data and the variability of the Maxentropic density in the regulatory capital estimation.

# Chapter 5

# LDA: Analysis of the sample dependence in density reconstruction by Maxentropic methods

The method of maximum entropy has proven to be a powerful tool to determine probability densities from a few values of its Laplace transform. This is so, even when the amount of data to compute numerically the Laplace transform is small. But in this case the variability of the reconstruction due to the sample variability in the available data can lead to quite different results. It is the purpose of this note to explore and quantify as much as possible the variability of the densities reconstructed by means of two Maxentropic methods: the standard maximum entropy method and its extension to incorporate data with errors.

The power of the method that we apply here, stems from the possibility of inverting the Laplace transform of a positive random variable $S$ from the knowledge of a few values of its Laplace transform $\psi(\alpha) = E[e^{-\alpha S}]$ at a small set of values of the parameter $\alpha$ by recasting the problem into a fractional moment problem on $[0, 1]$ after the transformation $y = e^{-s}$. Furthermore, an interesting feature of the methodology is that the statistical error in the estimation of $\psi(\alpha)$ can be incorporated into the procedure, as was developed in Chapter 3, and that the procedure itself makes no assumptions about the statistical nature of the data.

The pending question of interest in operational risk is: What about in the banking industry, where the amount of data is small?

This is a natural issue, because the Maxentropic density depends upon the data used. To state the aim of this paper, on one hand, the maximum entropy technique does provide us with excellent densities, but they vary considerably with the data. It is our aim here to analyze the variability of the reconstructed densities, and on the other to examine the impact of this variability on the estimation of the regulatory capital.

To state the problems that we are interested in, let us begin saying that we are interested in compound variables $S = \sum_{k=0}^{N} X_k$, where $N$ is an integer valued random variable describing the (yearly, say) frequency of events, and $X_k$ is the individual severity of each loss. What the analyst observes each year is a collection $\{n; x_1, ..., x_n\}$, where $n$ is the number of risk events and $\{x_1, ..., x_n\}$ are the losses occurring at each event. The aggregate loss for that year is $\sum_{k=0}^{n} x_k$. When $n = 0$ there were no losses, the sum is empty and $s = 0$. Suppose that the record consists of $T$ years of data. From these, the analyst has to determine the distribution of losses, which is the intermediate step in the calculation of regulatory capital or some other measure of risk, or perhaps when some insurance premium is to be calculated. When we need to specify the year $j$ we shall write $(n_j, x_1, ..., x_{n_j})$ and $s_j = \sum_{k=0}^{n_j} x_k$. For us, an observed sample (of losses) will be a vector of size $T$, $\omega = (s_1, ..., s_T)$.

The Laplace transform of $S$ is estimated by

$$\psi(\alpha) = \frac{1}{T} \sum_{j=1}^{T} e^{-\alpha s_j} \tag{5.1}$$

where $S_j$ denotes the losses experienced during the $j-$th year. Since the distribution function of $S$ has a probability $P(N = 0) = P(S = 0) > 0$ at $S = 0$, to determine the probability density of the losses we have to condition out this event and replace $\psi(\alpha)$ by

$$\mu(\alpha) = \frac{\psi(\alpha) - P(N = 0)}{1 - P(N = 0)} \tag{5.2}$$

where $P(N = 0)$ is the estimated as the fraction of the number of years of observation in which there were no losses. Notice that if we use the change of variables $y_j = e^{-s_j}$ we can rewrite (5.1) as

$$\psi(\alpha) = \frac{1}{T} \sum_{j=1}^{T} y_j^{\alpha} \tag{5.3}$$

which is the empirical version of

$$\psi(\alpha) = \int_0^1 y^{\alpha} dF_Y(y) = \int_0^{\infty} e^{-\alpha x} dF_S(x).$$

With this notation our problems consists of finding a density $f_Y(y)$ on $[0,1]$ such that

$$\int_0^1 y^\alpha f_Y(y)dy = \mu(\alpha),$$

and once $f_Y(y)$ is at hand, the change of variables $f_S(x) = e^{-x} f_Y(e^{-x})$ provides us with the desired density.

We shall use the notation $f_T(\omega, x)$ to denote the Maxentropic density reconstructed form the collection of moments (5.2). We recall how to obtain $f_T$ in Section 5.2 below. Note that as (5.2) depends on the sample $\omega$, then $f_T$ depends on $\omega$. To restate our goals, there are three things that we want to understand, or develop intuition about. First, how much does $f_T$ change when we change $\omega$. Second, how much do some basic risk measures change when we change $\omega$, and third, what happens when $T$ becomes very large.

The connection between the moments $\mu(\alpha, \omega)$ is quite non-linear, so the study of the variability of $f_T$ is not that simple, nevertheless, a few things can be said. We shall carry out this in Section 5.2, while in Section 5.3 we examine this issue by numerical simulations. The data that we use as input consists of an aggregation of risks of different nature (second level of aggregation), so it is not a simple compound model as that considered in Chapter 2 and 3.

## 5.1 The maximum entropy inversion techniques

In this chapter we consider two of the approaches discussed in Chapter 2 and 3. First, the standard maximum entropy (SME) discussed in Section 2.2.1 of Chapter 2 and its extension discussed in Section 3.3.2 of Chapter 3.

### 5.1.1 The standard Maxentropic approach with error estimation

The idea behind the proposal is clear. In order to obtain the $\mu_k$s either experimentally or numerically, we average over a collection of observations (simulations) and the errors in each average additively. Thus, the observed value of $\mu_k$, consists of the true (unknown) moment which is to be determined, plus an error term that has to be estimated as well. The errors are supposed to fall in an interval $[-c, c]$ independently of the sample ($c$ may decrease as the sample gets larger, but is otherwise independent of the sample). The reason for this choice is

any point in $[-c, c]$ can be written as $-cp + c(1-p)$ for some $0 < p < 1$. similarly as we have described in Section 3.3.2 of Chapter 3.

This computational simplification is possible when the support of the distribution of errors is bounded. To compress the notations we write probability distributions concentrated on a two point set $\{-c, c\}$ as $q_k(dz) = p_k\epsilon_{-c}(dz) + (1-p_k)\epsilon_c(dz)$, that is, the probability that we are after is a mixture of continuous and discrete distributions.

As we saw in Chapter 3, the usual variational argument in Jaynes (1957), or the more rigorous proofs in Maslov and Cherny (2003), yield

$$\begin{aligned} g^*(y) &= \frac{e^{-\sum_{k=1}^{K}\lambda_k^* y^{\alpha_k}}}{Z(\boldsymbol{\lambda}^*)} \\ p_k &= \frac{e^{c\lambda_k^*}}{e^{c\lambda_k^*}+e^{-c\lambda_k^*}}. \end{aligned} \tag{5.4}$$

Thus, the vector $\boldsymbol{\lambda}^*$ of Lagrange multipliers is to be found minimizing the dual entropy.

Once $\boldsymbol{\lambda}^*$ is found, the estimator of the measurement error is

$$\epsilon_k = \frac{-ce^{c\lambda_k^*} + ce^{-c\lambda_k^*}}{e^{c\lambda_k^*} + e^{-c\lambda_k^*}}. \tag{5.5}$$

Notice that, although the formal expression for $g^*(y)$ is the same as that for the SME, the result is different because $\boldsymbol{\lambda}^*$ is found minimizing a different functional. For more detail about this matter we direct the reader to Chapters 2 and 3.

### 5.1.2 Remarks on minimization process

Let us examine some properties of $\Sigma(\boldsymbol{\lambda}, \boldsymbol{\mu})$ related to the minimization process. First, observe that in our set up, it is a strictly convex function and twice continuously differentiable function, defined on all $\mathbb{R}^K$. If we denote by

$$\phi(\boldsymbol{\lambda}) = \begin{cases} -\nabla_{\boldsymbol{\lambda}} \ln Z(\boldsymbol{\lambda}) \\ -\nabla_{\boldsymbol{\lambda}} \left( \ln Z(\boldsymbol{\lambda}) + \sum_{k=1}^{K} \ln\left(e^{c\lambda_k} + e^{-c\lambda_k}\right) \right) \end{cases} \tag{5.6}$$

then the first order condition for $\boldsymbol{\lambda}^*$ to be a minimizer is that $\phi(\boldsymbol{\lambda}^*) = \boldsymbol{\mu}$. In practice this is not of much help and the minimization of $\sum(\lambda, \mu)$ has to be carried out numerically.

Notice that $\phi(\boldsymbol{\lambda})$ is differentiable and its Jacobian (the negative of the Hessian matrix of $\ln(Z(\boldsymbol{\lambda}))$ is the negative of the covariance matrix of the fractional powers $y^{\alpha_i}$, that is of the

positive definite matrix with components $\mu(\alpha_i+\alpha_j)-\mu(\alpha_i)\mu(\alpha_j)$ with respect to the maximum entropy density.

Observe as well that $(\boldsymbol{\mu})$ has all of its components bounded by 1, thus the image of $[0,1]$ by $\phi^{-1}$, which is continuous due to finiteness of the Hessian matrix, is compact. Thus the following assumption, that plays a role in the analysis carried out in the next section is natural:

**Assumption A** *Suppose that there is a ball $B(0,R)$ in $\mathbb{R}^K$ such that all solutions to $\phi(\boldsymbol{\lambda}^*)=\boldsymbol{\mu}$ lies there for every $(\boldsymbol{\mu})\in[0,1]^K$. Suppose as well that the Hessian matrix*

$$\frac{\partial^2\Sigma(\boldsymbol{\lambda}^*,\boldsymbol{\mu})}{\partial\lambda_i\partial\lambda_j}=-\frac{\partial\phi_i}{\partial\lambda_j}(\boldsymbol{\lambda}^*)$$

*has its upper eigenvalue uniformly bounded above, and its minimum eigenvalues uniformly bounded away from 0 in $B(0,R)$.*

## 5.2 Variability of the reconstructions

As mentioned before, there several issues to take care of. The simplest of all is to make sure that when the sample is large, the reconstructed $f_K$ converges to the reconstruction obtained from the moments of the exact Laplace transform. This is rather direct if we suppose that the samples are independent.

**Lemma 4.** *Suppose that the risks observed during $T$ consecutive years are independent of each other. Denote by $\mu_T(\alpha)$ the moments computed as in (5.2) and $\mu_e(\alpha)$ the exact moments. Denote as well $\boldsymbol{\lambda}_T^*$ and $\boldsymbol{\lambda}_e^*$ the corresponding minimizers of $\sum(\lambda,\mu)$ and by $f_T(x)$ and $f_e(x)$ the Maxentropic densities.*

*Then $\mu_T(\alpha)\to\mu_e(\alpha)$ and $\boldsymbol{\lambda}_T^*\to\boldsymbol{\lambda}_e^*$, and therefore $f_T(x)\to f_e(x)$ when $T\to\infty$.*

*Proof.* The proof hinges on an application of the law of large numbers combined with an invocation to the continuity of $\phi^{-1}$ based on Assumption A. □

To relate the sample variability of $f_T$ to the sample variability of the $\mu(\alpha)$, starting from $\boldsymbol{\lambda}_T^*=\phi^{-1}(\boldsymbol{\mu})$, and applying the chain rule, it follows that up to terms of $o(\delta\boldsymbol{\mu})$

$$\delta\boldsymbol{\lambda}_T^*=\mathbf{D}\delta\boldsymbol{\mu} \tag{5.7}$$

where $\mathbf{D}$ is the inverse matrix of the Jacobian of $\phi$ evaluated at $\boldsymbol{\lambda}_T^*$. Recall that the Maxentropic solution to the inverse problem is

$$f_T(y) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\sum_{k=1}^{K} \lambda_k^* y^{\alpha_k}} = e^{-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}}. \tag{5.8}$$

in which a reference to the sample size $T$ is made explicit. It is to make the state of the following results where the independence of the intervals of the sample enters. If we had considered such intervals to be sample dependent, the function $\phi$ defined in (5.6) would be sample dependent and the arguments that follow would not be true. With these notations another application of the chain rule leads to the proof of the following simple lemma.

**Lemma 5.** *With the notations introduces before, up to terms that are $o(\delta\boldsymbol{\mu})$, using the computation in (5.7)*

$$f_T(x) - f_e(x) = \sum_{i,j=1}^{K} \left(\mu(\alpha_i) - e^{-x\alpha_i}\right) f_e(x) D_{i,j} \delta\mu_j. \tag{5.9}$$

*Here $K$ is the number of moments used to determine the density and $\delta\mu_j = \mu_T(\alpha_j) - \mu_e(\alpha_j)$.*

This result makes explicit the deviation of $f_T$ from its exact value $f_e$, up to first order in the $\delta\mu$. Notice if we integrate both sides of (5.9) with respect to $x$ we get 0 on both sides. And if we multiply both sides both $e^{x\alpha_k}$ both sides integrate to $\delta\mu_k$.

If we think of the $f_T$ as random observations, they happen to be vectors in a convex set in $L_1(dx)$. The values $f_T(x)$ at each point can be thought as the components of those vectors, and results of the type of the central limit theorem for such values bear out in the simulations that we carried out.

But these results are of not much use for banking practice because neither the exact moments $\mu_e(\alpha)$ are known, neither is the Maxentropic density $f_e$ that they determine. Another potentially useful result goes along the following lines. Suppose that a bank has a relatively large number of agencies, call it $N_A$ say, and that each of them has been collecting data for the last $T$ years. If all risks befalling these agencies can be considered identically distributed, the computations in Lemma 5, lead to an interesting consequence. To establish it, we need to introduce extra labeling. Denote by $\mu_{T,m}(\alpha_i)$ the $i-$th moment computed from the $T-$year data at branch $m = 1, ..., N_A$ of the bank. Denote as well, by $f_{T,m}(x)$ the density reconstructed from the data from branch $m$ of the bank. We have

**Lemma 6.** *Set*

$$\hat{f}_T = \frac{1}{N_A} \sum_{m=1}^{N_A} f_{T,m}(x) \tag{5.10}$$

*Then* $\hat{f}_T \approx f_e(x)$.

*Proof.* Let us rewrite (5.9) in the proposed notation:

$$f_{T,m}(x) - f_e(x) = \sum_{i,j=1}^{K} \left(\mu_e(\alpha_i) - e^{-x\alpha_i}\right) f_e(x) D_{i,j} \left(\mu_{T,m}(\alpha_j) - \mu_e(\alpha_j)\right).$$

Summing over $m$ we obtain

$$\hat{f}_T(x) - f_e(x) = \sum_{i,j=1}^{K} \left(\mu_e(\alpha_i) - e^{-x\alpha_i}\right) f_e(x) D_{i,j} \left(\mu_e(\alpha_j) - \mu_e(\alpha_j)\right) = 0$$

up to terms of order $o(\delta\boldsymbol{\lambda})$, which concludes the proof of this intuitive result. □

As measure of the variability we shall estimate the $L_1$ difference between the approximate reconstruction $f_T$ for each sample and the true $f_e$ density. For this, the following results come in handy. An interesting way to compare two densities $f$ and $g$ (on $[0, \infty)$ say) is in terms of their Kullback's divergence (or relative entropy) defined by

$$K(f, g) = \int f(x) \ln \left(\frac{f(x)}{g(x)}\right) dx. \tag{5.11}$$

Two properties of (5.11) which are important for us here are collected under

**Proposition 2.** *Consider two densities $f$ and $g$ on $[0, \infty)$ and let the Kullback divergence among them be defined by (5.11). Then*
(a) $K(f, g) \geq 0$ and $K(f, g) = 0 \Leftrightarrow f = g \quad a.e$
(b) (Kullback's inequality ) $\frac{1}{4} \left(\|f - g\|_1\right)^2 \leq K(f, g)$.

Part (a) is an easy consequence of Jensen's inequality and part (b) is a guided exercise in Kullback (1959).

Now, with the notations introduced at the beginning of this section consider to begin with $K(f_e, f_T)$. Using the representation provided by equations (5.8), (5.11) and $\sum(\lambda, \mu)$, it takes a simple computation to verify that,

$$K(f_e, f_T) = \ln Z(\boldsymbol{\lambda}_T^*) + \langle \boldsymbol{\lambda}_T^*, \boldsymbol{\mu}_e \rangle - \ln Z(\boldsymbol{\lambda}_e^*) - \langle \boldsymbol{\lambda}_e^*, \boldsymbol{\mu}_e \rangle \tag{5.12}$$

Combining this Proposition 2 with Lemma 4, we obtain

**Proposition 3.** *With the notations introduced above we have*

$$\|f_T - f_e\|_1 \leq (4K(f_e, f_T))^{1/2} \to 0 \ \ as \ \ T \to \infty. \tag{5.13}$$

**Comment** A similar result would have been obtained if we had considered $K(f_T, f_e)$. Of interest to us is that (5.13) allows us to estimate the sample size variability of the reconstruction for a given sample.

Actually, we can combine Lemma 4, equation (5.12) and Proposition (3) to further relate the deviations $\|f_T - f_e\|_1$ to the deviations of $\boldsymbol{\mu}_T$ from $\boldsymbol{\mu}_e$.

**Proposition 4.** *With the notations introduced above we have*

$$\|f_T - f_e\|_1 \leq (2\langle \delta\boldsymbol{\mu}, \mathbf{D}\delta\boldsymbol{\mu}\rangle)^{1/2} \to 0 \ \ as \ \ T \to \infty. \tag{5.14}$$

*Here $\delta\boldsymbol{\mu}$ and $\mathbf{D}$ are as before.*

The proof is simple, and drops out from the fact that up to terms of order $o((\delta\boldsymbol{\mu})^2)$, $K(f_e, f_T) = \langle \delta\boldsymbol{\mu}, \mathbf{D}\delta\boldsymbol{\mu}\rangle$.

## 5.3 Numerical Results

This section is organized as follows. We begin by describing the sample generation process. To the large sample we apply the standard maximum entropy procedure and reconstruct a density from $K = 8$ values of its Laplace transform. Since the full sample is very large, the effect of the sample variability is small. We call that density, the *true density* and use it as reference to test the behavior of the densities reconstructed from smaller samples. After obtaining the "true" density, to analyze the sample variability we first plot the densities obtained for the smaller samples along the true density, and we present the different measures of sample variability. After that, we carry out an analysis of the sample variability of the regulatory capital, computed as the VaR and TVaR of each reconstructed density.

### 5.3.1 The sample generation process

As data set we considered an aggregation of losses generated with different compound variables, to model an aggregate risk coming from different risk sources. The frequencies of the

compound variables were chosen as: Two Poisson random variables with parameters $\ell_1 = 80$ and $\ell_2 = 60$, four binomial frequencies of parameters $(n_3 = 70, p_3 = 0.5)$, $(n_4 = 62, p_4 = 0.5)$, $(n_5 = 50, p_5 = 0.5)$ and $(n_6 = 76, p_6 = 0.5)$. We considered as well two negative binomial frequencies with parameters $(n_7 = 80, p_7 = 0.3)$ and $(n_8 = 90, p_8 = 0.8)$.

The corresponding individual losses were chosen as independent identically distributed random variables listed as follows. By $X_k$ we denote the type of the individual loss compounded according to the $k-$th frequency in the previous list. $X_1$ and $X_4$ Champernowne densities with parameters $(\alpha_1 = 20, M_1 = 85, c_1 = 15)$ and $(\alpha_4 = 10, M_4 = 125, c_4 = 45)$. $X_2$ was chosen as a lognormal of parameters $(-0.01, 2)$. Then $X_3$ and $X_8$ were chosen to have fat tails and to follow a Pareto distribution with (shape, scale)-parameters given by $(10, 85)$ and $(5.5, 5550)$. Next in the list are $X_5$ and $X_6$, chosen to be Gamma distribution with (shape,scale) parameters given respectively by $(4500, 15)$ and $(9000, 35)$. And finally, $X_7$ was chosen to be a Weibull of type $(200, 50)$. This is summarized in Table 5.1

| $S_h$ | $N_{bh}$ | $X_{bh}$ |
|---|---|---|
| $S_1$: | POISSON ($\ell = 80$) | CHAMPERNOWNE ($\alpha = 20$, $M = 85$, $c = 15$) |
| $S_2$: | POISSON ($\ell = 60$) | LogNormal ($\mu$ = -0.01, $\sigma = 2$) |
| $S_3$: | BINOMIAL($n = 70$, $p = 0.5$) | PARETO( $shape = 10$, $scale = 85$) |
| $S_4$: | BINOMIAL ($n = 62$, $p = 0.5$) | CHAMPERNOWNE ($\alpha = 10$, $M = 125$, $c = 45$) |
| $S_5$: | BINOMIAL ($n = 50$, $p = 0.5$) | GAMMA($shape = 4500$, $rate = 15$) |
| $S_6$: | BINOMIAL ($n = 76$, $p = 0.5$) | GAMMA($shape = 9000$, $rate = 35$) |
| $S_7$: | NEGATIVE BINOMIAL ($r = 80$, p = 0.3) | WEIBULL($shape = 200$, $scale = 50$) |
| Tail: | NEGATIVE BINOMIAL ($r = 90$, p = 0.8) | PARETO($shape = 5.5$, $scale = 5550$) |

Table 5.1: Inputs for the simulation of $S$

The size of the total sample was 5000, from which different subsamples were chosen to study the sample variability.

As the size of the resulting aggregate losses have order of magnitude $10^4$, and since the Maxentropic methods use the Laplace transforms of the aggregate loss as starting point, we scale the data prior to the application of the procedure, and reverse the scaling to plot the resulting densities. The simplest scaling consists of dividing by $10^4$ as we did. Another possibility is to introduce a scaled variable defined by $s - \min(s))/\max(s) - min(s)$, where of course, $\max(s)$

and min($s$) stand for the maximum and minimum values realized in the sample.

### 5.3.2 The true Maxentropic density

In Figure 5.1 we present the result of applying both Maxentropic methods, the SME and the SMEE to the full data set. As intuitively expected, since the data set is very large, the Maxentropic reconstructions agree quite well with the simulated (empirical) data.

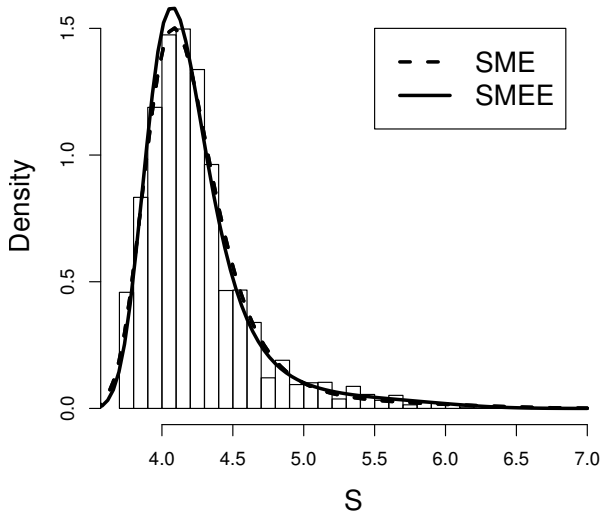



Figure 5.1: Histogram of total losses $\times 10^4$ (5000 data points) and Maxentropic density.

To determine how well the densities obtained approximate the data, we consider two measures, the mean absolute error (MAE) and the square root mean error (RMSE). See the Appendix A.1 and the reference Hyndman and Koehler (2006) for more about such measures.

| Approach | **MAE** | **RMSE** |
|---|---|---|
| SMEE | 0.005928 | 0.006836 |
| SME | 0.006395 | 0.009399 |

Table 5.2: MAE and RMSE for a sample size of 5000

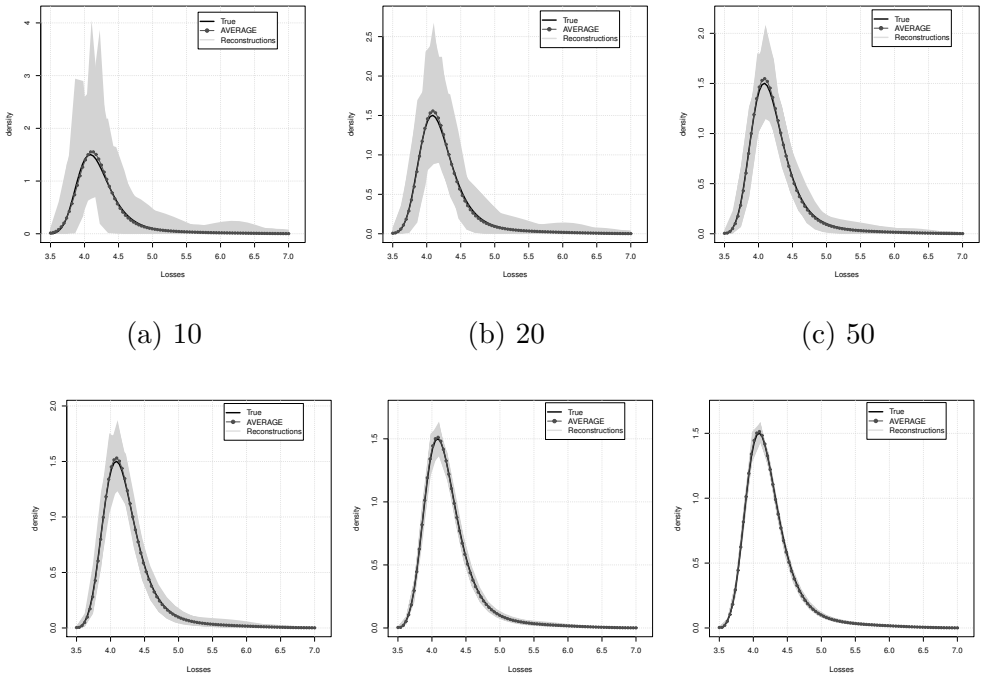
In Table 5.2 we show the results of the computation of MAE and RSME to the full data set and how the well the Maxentropic densities fit the empirical data. Clearly, both Maxentropic

procedures yield quite good reconstructions.

At this point we add that the reconstruction error in the Maxentropic procedure, that is, the norm of the gradient of the dual entropy $\sum(\lambda, mu)$, was less than $10^{-6}$ in all reconstructions. This value measures the quality of the reconstructions, and is used as stopping criterion in the minimization of the dual entropy.

### 5.3.3 The sample variability of the densities

In this section we study how does the Maxentropic density approximate the true density. For each sample, of a fixed size, we applied both the SME and the SMEE, to obtain a density. Let us describe separately the results obtained by each method.




(a) 10 (b) 20 (c) 50

(d) 100 (e) 500 (f) 1000

Figure 5.2: SME reconstructions with different sample sizes

The results displayed in the panels of Figure 5.2 should be understood as follows. To study the sample variability, we applied the Maxentropic procedures to 200 samples of different sizes. For example in the first panel, the gray shadow consists of the plot of the 200 densities produced as output of the Maxentropic procedure estimated form the moments produced by 200 samples of size $T = 10$. In each panel we plot as well the average density $\hat{f}_T$ as well as the true density.

| **Size** | **Error Meas.** | 1st Qu. | Mean | 3rd Qu. | Area | AVERAGE |
|---|---|---|---|---|---|---|
| 10 | MAE | 0.0500 | 0.0880 | 0.1150 | 2.625 | 0.0092 |
| | RMSE | 0.0590 | 0.1010 | 0.1310 | | 0.0120 |
| 20 | MAE | 0.0359 | 0.0619 | 0.0811 | 1.523 | 0.0089 |
| | RMSE | 0.0400 | 0.0702 | 0.0915 | | 0.0116 |
| 50 | MAE | 0.0203 | 0.0377 | 0.0510 | 0.9545 | 0.0082 |
| | RMSE | 0.0237 | 0.0429 | 0.0575 | | 0.0106 |
| 100 | MAE | 0.0144 | 0.0266 | 0.0343 | 0.696 | 0.0053 |
| | RMSE | 0.0169 | 0.0304 | 0.0393 | | 0.0066 |
| 200 | MAE | 0.0126 | 0.0194 | 0.0247 | 0.5375 | 0.0053 |
| | RMSE | 0.0148 | 0.0225 | 0.0285 | | 0.0067 |
| 500 | MAE | 0.0081 | 0.0128 | 0.0163 | 0.3258 | 0.0055 |
| | RMSE | 0.0101 | 0.0153 | 0.0194 | | 0.0076 |
| 1000 | MAE | 0.0067 | 0.0093 | 0.0108 | 0.2033 | 0.0054 |
| | RMSE | 0.0087 | 0.0115 | 0.0132 | | 0.0078 |

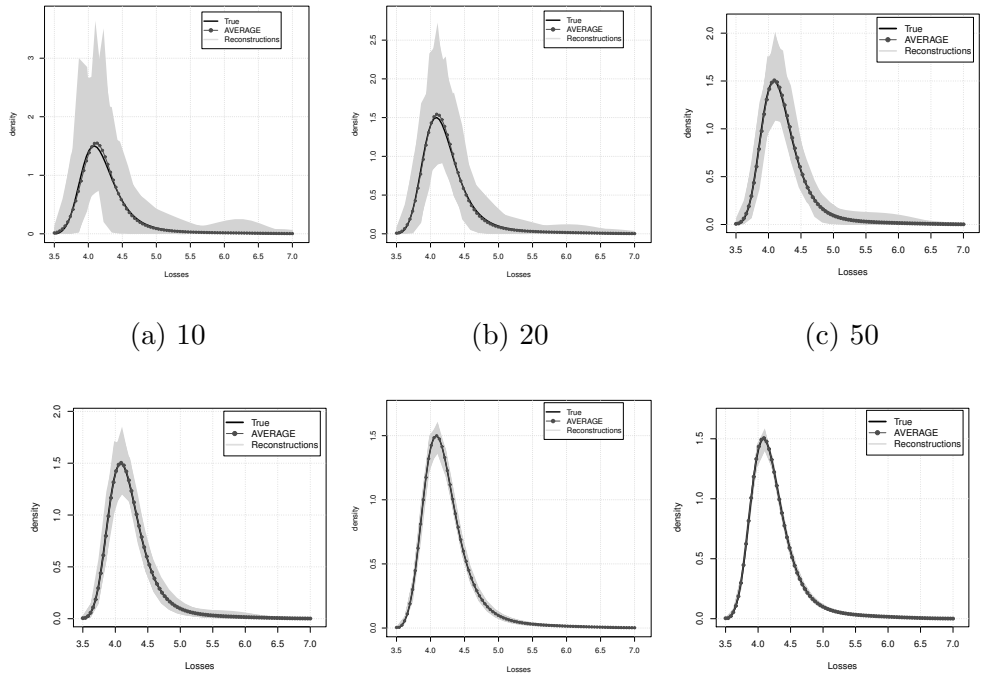
Table 5.3: MAE & RMSE of SME for different sample sizes

Also, for each panel we average the moments and use them as input for the maximum entropy method and we display the density obtained from such average moments as well as the density obtained from the the exact moments. That the last two mentioned coincide is not a surprise according to the results in the previous section. As we move from panel to panel, the amount of data used to compute the moments increases and they are closer to the true values. This is the law of large numbers in action, thus the spread in the indeterminacy of the true density decreases, as is clearly seen by the shrinking size of the gray shadow.

The entries of Table 5.3 are as follows. The first two columns describe the size of the sample and the error measure being listed. For the second three, a histogram of the MAE and RSME errors was made for the samples of each size, and the percentiles listed were computed. Again, each error measure was computed as described in the previous section. In column number six we list the area of the gray shadow, computed in the obvious way. To obtain the last column, we averaged the moments over 200 samples and we obtained the corresponding Maxentropic density, then we compared it to the empirical density of the sample used.

To obtain the next figure, we repeat the same procedure as before, except that this time we apply

the SMEE method. This in order to test whether specifying a measurement error we can improve the reconstructions. In the six panels of Figure 5.3 we followed the same procedure described before.




(a) 10 (b) 20 (c) 50

(d) 100 (e) 500 (f) 1000

Figure 5.3: SMEE reconstructions with different sample sizes

Similarly to the previous case, as we increased the amount of data, the reconstructions improved, but as shown in Table 5.4 the improvement, contrary to expectations, did not become more significant. The entries in Table 5.4 were produced and have the same meaning as those in Table 5.3.

In Table 5.5 we summarize the variability of the reconstructions depending on the size of the sample in a lightly different way. There we list how does the mean MAE and the area of the gray shadow for various sample sizes. The improvement of the reconstructions as the sample size increases is apparent.

### 5.3.4 Computation of the regulatory capital

This section is devoted to the computation of the two most used risk measures, namely, the $VaR$ and the $TVaR$, which are used to determine the regulatory capital. We explained how to compute these risk measures in Chapter 2. We present several tables. The idea is provided an intuition about the possible variability of the quantities that are essential in risk measurement.

| **Size** | **Error Meas.** | 1st Qu. | Mean | 3rd Qu. | Area | AVERAGE |
|---|---|---|---|---|---|---|
| 10 | MAE | 0.0419 | 0.0690 | 0.0898 | 2.619 | 0.0069 |
| | RMSE | 0.0514 | 0.0784 | 0.1030 | | 0.0110 |
| 20 | MAE | 0.0360 | 0.0620 | 0.0820 | 1.759 | 0.0066 |
| | RMSE | 0.0420 | 0.0705 | 0.0918 | | 0.0109 |
| 50 | MAE | 0.0198 | 0.0378 | 0.0500 | 1.044 | 0.0065 |
| | RMSE | 0.0240 | 0.0430 | 0.0582 | | 0.0102 |
| 100 | MAE | 0.0142 | 0.0267 | 0.0353 | 0.690 | 0.0060 |
| | RMSE | 0.0168 | 0.0306 | 0.0398 | | 0.0082 |
| 200 | MAE | 0.0125 | 0.0196 | 0.0247 | 0.552 | 0.0063 |
| | RMSE | 0.0147 | 0.0229 | 0.0270 | | 0.0072 |
| 500 | MAE | 0.0083 | 0.0131 | 0.0165 | 0.294 | 0.0058 |
| | RMSE | 0.0101 | 0.0156 | 0.0199 | | 0.0083 |
| 1000 | MAE | 0.0068 | 0.0093 | 0.0109 | 0.200 | 0.0057 |
| | RMSE | 0.0082 | 0.0114 | 0.0133 | | 0.0082 |

Table 5.4: MAE & RMSE of SMEE results for different sample sizes

In Table 5.6 we present a comparison of the values of the $VaR$ and the $TVaR$ computed from an empirical sample of size 5000, and that computed from the SME and SMEE densities at the 95% and the 99% confidence levels. The sample considered to build Table 5.6 is large for operational risk purposes, although data sets corresponding to large disasters may comprise even smaller data sets.

In Table 5.7 we consider two measures of variability of the $VaR$ and the $TVaR$ computed from the Maxentropic densities obtained for 200 samples of the indicated sizes. In each cell we present the mean and the variance (within parentheses) of each risk measure, for each sample size. And we do it at the 95% and the 99% confidence levels.

To finish, consider Table 5.8 in which we compute the "true" $VaR$ and the 'true" $TVaR$ of small samples. Recall that the "true" density of a small sample of size $T$ was obtained averaging the densities of 200 samples of size $T$, and was shown to become closer to the true density of the total loss as $T$ increased. We see that the same happens to the $VaR$ and the $TVaR$, as described in Table 5.8.

| | 10 | | 100 | | 1000 | |
|---|---|---|---|---|---|---|
| | Mean (MAE) | Area | Mean (MAE) | Area | Mean (MAE) | Area |
| SME | 0.0880 | 2.625 | 0.0266 | 0.696 | 0.0092 | 0.2033 |
| SMEE | 0.0690 | 2.619 | 0.0267 | 0.690 | 0.0093 | 0.2000 |

Table 5.5: Summary of results

| | $\gamma$ | **Empirical** | **SME** | **SMEE** |
|---|---|---|---|---|
| VaR | 0.950 | 5.05 | 4.935 | 5.004 |
| | 0.990 | 5.72 | 5.755 | 5.772 |
| TVaR | 0.950 | 5.45 | 5.443 | 5.461 |
| | 0.990 | 6.05 | 6.0207 | 6.014 |

Table 5.6: Comparison of VaR and TVaR at 95% and 99% for a unique sample of size 5000

| Size | VaR(95%) | | TVaR(95%) | | VaR(99%) | | TVaR(99%) | |
|---|---|---|---|---|---|---|---|---|
| | SME | SMEE | SME | SMEE | SME | SMEE | SME | SMEE |
| 10 | 4.96 | 4.87 | 5.30 | 5.156 | 4.331 | 5.283 | 4.328 | 5.634 |
| | (0.4530) | (0.6740) | (0.5678) | (0.7984) | (1.083) | (0.8597) | (0.995) | (1.3138) |
| 20 | 4.96 | 4.91 | 5.30 | 5.282 | 4.518 | 5.502 | 4.633 | 5.818 |
| | (0.4536) | (0.5200) | (0.5678) | (0.6515) | (1.004) | (0.7537) | (1.004) | (0.8053) |
| 50 | 4.97 | 4.925 | 5.39 | 5.386 | 5.003 | 5.688 | 5.162 | 6.058 |
| | (0.2902) | (0.3254) | (0.382) | (0.4286) | (0.988) | (0.5463) | (1.069) | (0.5617) |
| 100 | 4.96 | 4.931 | 5.44 | 5.43 | 5.457 | 5.779 | 5.694 | 6.016 |
| | (0.1994) | (0.2258) | (0.251) | (0.2794) | (0.705) | (0.3626) | (0.768) | (0.3476) |
| 200 | 4.97 | 5.013 | 5.45 | 5.46 | 5.624 | 5.766 | 5.871 | 6.016 |
| | (0.1537) | (0.1828) | (0.1902) | (0.1995) | (0.395) | (0.2343) | (0.417) | (0.2498) |
| 500 | 4.95 | 4.93 | 5.45 | 5.45 | 5.708 | 5.822 | 5.972 | 6.017 |
| | (0.08963) | (0.1038) | (0.1136) | (0.1249) | (0.153) | (0.1539) | (0.159) | (0.1385) |
| 1000 | 4.95 | 4.95 | 5.45 | 5.45 | 5.729 | 5.828 | 5.977 | 6.064 |
| | (0.05815) | (0.07033) | (0.07088) | (0.07765) | (0.109) | (0.1065) | (0.107) | (0.09838) |

Table 5.7: Mean and standard deviation of the VaR and TVaR for 200 samples of different sizes

## 5.4 Concluding remarks

We have seen that the Maxentropic techniques are quite suitable to determine a density from a few values of its Laplace transform by recasting the problem as a fractional moment problem. Not

| *Size* | VaR(95%) | | TVaR(95%) | | VaR(99%) | | TVaR(99%) | |
|---|---|---|---|---|---|---|---|---|
| | SME | SMEE | SME | SMEE | SME | SMEE | SME | SMEE |
| 10 | 4.81 | 4.82 | 5.32 | 5.35 | 4.54 | 5.64 | 5.97 | 5.97 |
| 100 | 4.95 | 4.92 | 5.42 | 5.38 | 5.67 | 5.72 | 5.99 | 6.05 |
| 500 | 4.95 | 4.95 | 5.49 | 5.45 | 5.78 | 5.79 | 6.09 | 6.09 |
| 1000 | 5.01 | 4.95 | 5.50 | 5.45 | 5.78 | 5.79 | 6.05 | 6.05 |

Table 5.8: VaR and TVaR for the average of the Maxentropic densities for different sample sizes.

only that, the Maxentropic procedures provide for more freedom of action: on the one hand the can help us to cope with a case in which the traditional methods are not that good when there is not enough data.

Maxentropic methods are useful to estimate loss densities when we are aggregating total losses from different sources. This is more useful when the amount of data is large. The importance of this possibility is that it does not make any use of the statistical dependence among the risks.

The Maxentropic method, combined with a simulation or re-sampling procedure provides us with reasonable measures of the variability of quantities of interest in risk measurement. For example, if we had 200 data points, we could re-sample a large number of times, and the average over the densities determined by each sub-sample to obtain rather accurate estimate of relevant risk measures.

# Chapter 6

# Maximum entropy approach to the loss data aggregation problem

## 6.1 Introduction

In a previous chapters, we tested the performance of a maximum-entropy-based method to obtain the density of a compound random variable. We showed that this is a powerful method to obtain probability densities of aggregate losses when the available data consists of the aggregate loss collected during several time intervals.

At the base, there is aggregation to compute the total loss of a given type of risk (first level of aggregation, used in Chapters 2 and 3), then there is aggregation to compute the total loss within a given business line (second level of aggregation, see Chapters 4 y 5, and finally, there is the aggregation of the risk over the different business lines (third level of aggregation).

Much of the work in the standard "classical" literature in the subject, concerns the first level of aggregation. See the volumes by Panjer (2006) or Schevchenko (2011) for example, or consider Aue and Kalkbrener (2006), Brockmann and Kalkbrener (2010) and references therein, to list but very few. We mention as well the work of Shen and Jia (2012) in which the method of maximum entropy is used to estimate loss distributions in a way unrelated to our approach. We mention at this point, that the term aggregation is also used to refer to the important problem consisting of merging data from different sources. That is not addressed here.

At the third level of aggregation, i.e. when the data is made up of aggregated losses caused by different types of risk events in different business lines, as is usual in operational risk, it is possible to

obtain the density through two possibilities, that would be explored in this chapter. When the data is well collected, that is, when the losses are collected as a joint vector, the Maxentropic techniques are quite suitable for finding the probability density of the aggregated loss. When the data is not well collected, the Maxentropic procedure provides us with the marginal densities, which can then be coupled by means of some appropriate copula, and then carry on one of the two procedures that we apply. At any rate, the two possibilities hinge in an essential way on the Maxentropic technique to determine a probability density from its Laplace transform. This is due to the fact that such techniques provide us with analytical expressions for the densities, which makes many numerical procedures easier to implement.

It is the aim of this chapter, to examine and compare these alternative ways of solving the problem of determining the density of aggregate losses.

A specific motivation for the work reported below comes from the need to compute the regulatory capital for operational risk losses according to the directives of the Basel Committee. To comply with the rules, the risk analyst has to gather data and, the output of the modeling process is the estimated operational risk regulatory capital. A systematic way of completing that task, is to compute the total loss density and then use it to determine the VaR or the TVaR at a high confidence level.

Another instance in which the distribution of a random variable describing total losses is interesting, comes from the insurance industry, where the density is used to price risk premiums to insure the loss.

To come back to the problem from the operational risk point of view, there are two problems common to most methods to determine a density of losses from the data collected by a bank: On one hand, if the amount of data collected by the bank is not large, one has to quantify the variability, or the sample dependence, of the density reconstructed as we saw in Chapter 5. On the other hand, even if the amount of data is large, we may not know the nature of the statistical dependency among the data. It turns out that, when the data is properly collected, that modeling aspect may not be necessary for the Maxentropic technique.

The rest of this note is organized as follows. In the remainder of this section (Subsection 6.1.1) we establish the mathematical problem to be solved. In Sections 6.2 and 6.3, we describe the data generation procedure used to explore the two possibilities discussed before and then, we present the numerical examples. We conclude with some remarks in Section 6.4.

### 6.1.1 Problem setup

As noted, our aim is to determine the probability density of a variable $S$ which is obtained after several levels of aggregation. Let us denote by $B$ an integer to be interpreted as the number of business lines of a financial institution. At the last level of aggregation we have

$$S = \sum_{b=1}^{B} S_b, \tag{6.1}$$

where for $b = 1, ..., B$ the (positive) random variable $S_b$ denotes the total risk in business line $b$. For each business line, let $H_b$ denote the number of risk types contributing to $S_b$. At the second level of aggregation we have

$$S_b = \sum_{h=1}^{H_b} S_{b,h}, \tag{6.2}$$

where $S_{b,h}$ denotes the total loss in business line $b$ due to risk type $h$. Now we come to the most basic level, that of describing $S_{b,h}$ as a compound random variable. We follow the standard, actuarial like modeling used in the advanced methodology in operational risk, and set

$$S_{b,h} = \sum_{k=1}^{N_{b,h}} X_{b,h,k} \tag{6.3}$$

where the $N_{b,h}$ are integer valued random variables denoting the frequency of events in the selected modeling time interval, and $\{X_{b,h,k}, |k \geq 1\}$ are random variables denoting the losses in the $k-$th individual risk event of type $h$ in business line $b$. This is the most basic level of aggregation. The usual modeling assumption is that the $X_{b,h,k}$ are independent among themselves and of the $N_{b,h}$. But the $N_{b,h}$ may exhibit dependence among themselves.

As we saw in the previous chapters, the methodology that we propose to determine the distribution of the density $f_S$ of the total aggregate loss consists of two basic steps: first, estimate the Laplace transform $E[e^{-\alpha_i S}]$ for a few (we use $K = 8$) real values of the transform parameter $\alpha$, and then convert the problem of inversion of the Laplace transform into a fractional moment problem. The second stage consists in solving the fractional moment problem by means of the maximum entropy method. To determine the probability density of losses, we note that

$$\psi(\alpha) = E[e^{-\alpha S}] = P(S = 0) + P(S > 0)E[e^{-\alpha S}|S > 0]$$

where $P(S = 0) = P(N = 0) = p_o$, where $p_o$ has to be estimated from the data, or supplied by the analyst. Setting

$$\mu(\alpha) = \frac{\psi(\alpha) - p_o}{1 - p_o}$$

our problem becomes to determine $f_S$ from

$$\int_0^\infty e^{-x\alpha_j} f_S(x)dx = \int_0^1 y^{\alpha_j} g_S(y)dy = \mu(\alpha_j), \quad j = 0, ..., K. \tag{6.4}$$

The middle term is obtained from the first after the change of variables $y = e^{-x\alpha}$. We set $\alpha_0 = 0$ and $\mu_0 = 1$ to take care of the natural requirement on $f_Y(y)$.

For this chapter any of the variations used in Chapters 2 and 3 are valid. For this chapter we will use the SME methodology, for details please review Chapter 2.

## 6.2 Numerical procedures

There are several issues to be aware of in the numerical examples for each level of aggregation. Let us begin with the simplest issue, namely the number of business lines and the number of risk types. We consider $B = 3$ business lines and $H_b = 7$ risk types for each business line. At the most basic level of aggregation we consider a moderately large amount of data to avoid the sample variability issues considered in Chapter 5. At that level, to give the model some sort of realism, we considered several frequency models and several different individual loss densities. To go from that to the second aggregation level, and to compute the total loss within business line, we used a copula to create dependency among the frequency of events within each business line. Thus, for each $b = 1, 2, 3$, we generate frequency of events samples $(N_{b,1}, ..., N_{b,7})$, and for each frequency sample we generate individual losses according to each number of events to obtain the $S_{b,h}$ according to (6.3) and aggregate these into $S_b$ according to (6.2). In this way we introduce dependence at this level.

To obtain the total loss according to (6.1), the institution should have collected/recorded the joint values of the losses $(S_1(t), S_2(t), S_3(t))$, where $t = 1, 2, ..., T$, where $T$ denotes some long period of data collection. We then estimate $\psi(\alpha)$ as

$$\hat{\psi}(\alpha) = \frac{1}{T}\sum_{t=1}^{T} \exp\left(-\alpha \sum_{b=1}^{B} S_b(t)\right),$$

and then apply the Maxentropic routine explained in Section 2.1 to obtain the probability density of the total loss.

Since, this is not what is usually done, let us suppose that the losses per business line have been properly collected, but not as a data vector, but as its separate components. At this point, we can bring in the Maxentropic technique described in Section 2.1 and obtain the densities $f_{S_b}$, after which there are two possibilities:

**1** Bring in a copula that appears reasonable, couple them to obtain a joint density $f(x_1, x_2, x_3)$ with

which to compute the Laplace transform of $S = S_1 + S_2 + S_3$, and use this transform within combined with the Maxentropic technique to obtain the probability density $f_S$.

**2** Use the copula and apply the iterated convolution procedure proposed in Alexander (2003) to obtain the density $f_S$.

We shall apply both of these possibilities for the purpose of comparison, and compare the results obtained to those obtained by the Maxentropic procedure, supposing that the data was properly collected.

### 6.2.1 The data generation procedure

First we describe the data generation process in some more detail, and then describe the specifics of the procedure.

1. For each fixed $b = 1, 2, 3$ we take care of the first two levels of aggregation and the first level of dependency as follows. For the first step simulate $h = 7$ compound distributions $S_h = \sum_{k=1}^{N_{b,h}} X_{b,h,k}$ where $N_{b,h}$ will follow several frequency distributions that are combined with copulas to create the first level of dependencies. The copulas used are Gaussian with different values of correlations $\rho = 0.3, 0.5$ and $0.8$. To describe individual losses, we considered some of the commonly used continuous distributions, to wit, the Champernowne, Pareto, gamma, lognormal and Weibull distributions. These were appropriately compounded up to obtain $S_h$.

2. Step 1 is repeated $B$ times to obtain $S_b$. Then, they are summed up to obtain $S$ in case they can be supposed to be independent. To consider dependencies go to step 3.

3. To create the dependence between $S_b$, the densities of each $S_b$ are determined applying the Maxentropic method. These will be used to obtain the total loss density $f_S$ using the coupling procedures described in Section 6.2.2.

The specific details, gathered in Tables 6.1, 6.2 and 6.3, are as follows: For $S_1$ and $S_2$ we use the same inputs, except that to generate $S_2$ we added large losses (greater than 14000) from a Pareto distribution. This was done to add kurtosis to the sample and to make the distribution heavier at the tail. We might think about this in two different ways. On the one hand, we can regard this as data about large losses in a given business line, collected by the institution making the analysis, or obtained from other institutions without attributing it to any risk type, or as data, obtained similarly, but associated with risk events of some specific type. Besides that, to generate dependence among the business lines we coupled the frequencies as follows: We invoked a normal copula with parameter 0.8 to couple $N_{11}$, $N_{12}$ and $N_{13}$, a normal copula with parameter 0.5 for $N_{14}$, $N_{15}$, and one with parameter

0.3 for $N_{16}$, $N_{17}$. The same was done to generate $S_2$. For $S_1$ and $S_2$ we use only those values greater than 11000. For each $S_i$ we simulate 9000 data points.

| $S_{bh}$ | $N_{bh}$ | $X_{bh}$ |
|---|---|---|
| $S_{11}$: | POISSON ($\ell = 20$) | CHAMPERNOWNE ($\alpha = 25$, $M = 85$, $c = 15$) |
| $S_{12}$: | POISSON ($\ell = 40$) | PARETO ($shape = 10$, $scale = 85$) |
| $S_{13}$: | BINOMIAL ($n = 70$, $p = 0.5$) | PARETO ($shape = 10$, $scale = 85$) |
| $S_{14}$: | BINOMIAL ($n = 62$, $p = 0.5$) | CHAMPERNOWNE ($\alpha = 10$, $M = 100$, $c = 45$) |
| $S_{15}$: | BINOMIAL ($n = 50$, $p = 0.5$) | GAMMA ($shape = 1500$, $rate = 15$) |
| $S_{16}$: | BINOMIAL ($n = 56$, $p = 0.5$) | GAMMA ($shape = 4000$, $rate = 35$) |
| $S_{17}$: | NEGATIVE BINOMIAL ($r = 60$, $p = 0.3$) | WEIBULL ($shape = 100$, $scale = 15$) |

Table 6.1: Inputs for the simulation of $S_1$

In Table 6.2 we only mention the details of the (very) large losses added to a sample obtained by the same procedure used to obtain $S_1$, in order to simulate potential large loses in the second business line.

| | $N_{bh}$ | $X_{bh}$ |
|---|---|---|
| TAIL: | NEGATIVE BINOMIAL ($r = 90$, $p$ =0.8) | PARETO ( $shape = 5.5$, $scale = 2100$) |

Table 6.2: Inputs for the simulation of $S_2$

In Table 6.3 we show the distributions used for the simulation of $S_3$, in this case all the distributions are independent. To the data generated for this case, we attach a tail at 35000 and we take only values greater than 32000 from the generated data set.

**NOTE:** To apply the Maxentropic procedure for the calculation of the densities, we have to consider empirical Laplace transforms of large numbers. To avoid overflow and underflow, we divide the losses by an appropriate scaling factor, and after the density is obtained, we undo the scaling. This was done in Chapter 5.

### 6.2.2 The coupling procedure

Copulas have become an ubiquitous method to introduce dependence among random variables. To introduce some notations, we briefly recall the copula method, and then how we use it to generate data and to compare the output of the Maxentropic procedure with the output obtained applying the

| $S_{bh}$ | $N_{bh}$ | $X_{bh}$ |
|---|---|---|
| $S_{31}$: | POISSON ($\ell = 80$) | CHAMPERNOWNE ($\alpha = 20$, $M = 85$, $c = 15$) |
| $S_{32}$: | POISSON ($\ell = 60$) | LogNormal ($\mu$ = -0.01, $\sigma = 2$) |
| $S_{33}$: | BINOMIAL($n = 70$, $p = 0.5$) | PARETO( $shape = 10$, $scale = 85$) |
| $S_{34}$: | BINOMIAL ($n = 62$, $p = 0.5$) | CHAMPERNOWNE ($\alpha = 10$, $M = 125$, $c = 45$) |
| $S_{35}$: | BINOMIAL ($n = 50$, $p = 0.5$) | GAMMA($shape = 4500$, $rate = 15$) |
| $S_{36}$: | BINOMIAL ($n = 76$, $p = 0.5$) | GAMMA($shape = 9000$, $rate = 35$) |
| $S_{37}$: | NEGATIVE BINOMIAL ($r = 80$, p = 0.3) | WEIBULL($shape = 200$, $scale = 50$) |
| Tail: | NEGATIVE BINOMIAL ($r = 90$, p = 0.8) | PARETO($shape = 5.5$, $scale = 5550$) |

Table 6.3: Inputs for the simulation of $S_3$

copula method. Within our framework, a copula $c(u_1, u_2, u_3)$ is just the joint cumulative distribution function of some probability law on $[0, 1]^3$. If we denote by $F_b$ the cumulative distribution associated with $f_{S_b}$, we define the joint density that the copula induces by

$$F_c(x_1, x_2, x_3) = c(F_1(x_1), F_2(x_2), F_3(x_3)). \tag{6.5}$$

For more information about copulas and some examples, see Chapter 5 in McNeil et al. (2005). If $c(u_1, u_2, u_3)$ is supposed to be continuously differentiable with respect to the three variables, the coupled density obtained from $F_c$ is

$$f_c(x_1, x_2, x_3) = \frac{\partial^3}{\partial u_1 \partial u_2 \partial u_3} c(F_1(x_1), F_2(x_2), F_3(x_3)) f_{S_1}(x_1) f_{S_2}(x_2) f_{S_3}(x_3). \tag{6.6}$$

#### 6.2.2.1 The sequential convolution procedure

When a copula and individual densities are available, one can obtain the density of the sum from the joint density by repeated convolution. This simple observation provides a rather effective tool. For a description of the recursive procedure, we refer the reader to the work of Alexander (2003). The recursive process goes as follows.

1. First we shall derive the distribution of the sum of a pair of values $S_1 + S_2$ from the joint density $f_{S_1,S_2}(s_1, s_2) = f_{S_1}(s_1) f_{S_2}(s_2) c(s_1, s_2)$, where $f_{S_1}$ and $f_{S_2}$ are the marginal densities and $c(s_1, s_2)$ is a given dependency structure defined through a copula model $C$.

2. The probability density of the sum $L_{12} = S_1 + S_2$ is given by the convolution integral $f_{S_1+S_2} = \int_{s_1} f_{S_1,S_2}(s_1, l_{12} - s_1) = \int_{s_2} f_{S_1,S_2}(s_2, l_{12} - s_2)$

The steps 1 and 2 are repeated for the rest of the sum, $L_{123} = S_1 + S_2 + S_3 = L_{12} + S_3$, until the density of the sum $S$ is calculated. Actually, what makes this interesting in the present setup, is the

fact that copulas are usually simple functions to deal with, and the maximum entropy method provides an analytic expression for the marginal densities $f_{S_b}$. The availability of such analytical expressions for the marginal densities makes the computations with copulas much easier.

### 6.2.3 Maxentropic density reconstructions

As we said, to obtain the probability density of the total loss we shall follow two routes. For the first one, we obtain the (marginal) probability densities of the losses in each line of activity of the institution, and then produce a joint density by means of a copula, which is used to compute the Laplace transform of the total loss $\sum S_b$, from which the total loss density is obtained. The computation of the Laplace transform of $\sum S_b$ is carried out by means of a large fine discretization of the $(s_1, s_2, s_3)$ space:

$$\psi(\alpha) = E[e^{-\alpha_k(S_1+S_2+...+S_B)}] = \sum_{i=1}^{N-1} e^{-(s_{1i}+s_{2i}+...+s_{Bi})\alpha_k} f(s_{1i}, s_{2i}, ..., +s_{Bi})\Delta s_1 \Delta s_2 ... \Delta s_B$$

where $N$ is the number of partitions used in the discretization and

$f(s_1, s_2, ..., s_B) = c[F_1(s_1), ..., F_1(s_B)] \prod_{i=1}^{B} f_{S_i}(x_i)$ the joint distribution, $c$ is the density of the copula model $C$, and $f_{S_1}, ..., f_{S_B}$ are marginal densities.

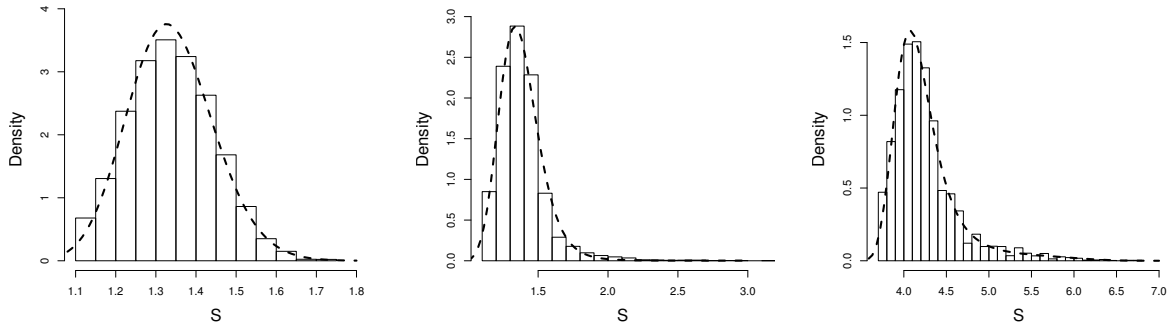
Once the $\psi(\alpha)$ is at hand, we apply the SME procedure described in Chapter 2, to obtain the density of the sum $f_{S_1+S_2+...+S_B}$.

The second approach exploits the knowledge of the joint density to obtain the density of the sum by recursive convolution.

## 6.3 Description of the numerical results

The first step in the data generation process consists of generation the losses per business line. The details of the data generation process were explained in Section 6.2.1. In each of the three panels of Figure 6.1 we display the histogram and probability density of $S_b$ for $b = 1, 2, 3$ obtained applying the SME method. As we indicated before, each of these is the sum of seven different types of events correlated among them. These three densities will be the inputs for the examples which we shall develop below.

Consider Table 6.4 in which we show the MAE (mean average error) and RMSE (root mean square error) distances between the reconstructed densities $f_{S_i}$ and the histogram. The agreement is quite good.




(a) $S_1$ (b) $S_2$ (c) $S_3$

Figure 6.1: Losses for each line of activity, reconstructed by SME

| | *Error* | |
|---|---|---|
| $S$ | $MAE$ | $RMSE$ |
| $S_1$ | 0.0054 | 0.0072 |
| $S_2$ | 0.0241 | 0.0282 |
| $S_3$ | 0.0061 | 0.0071 |

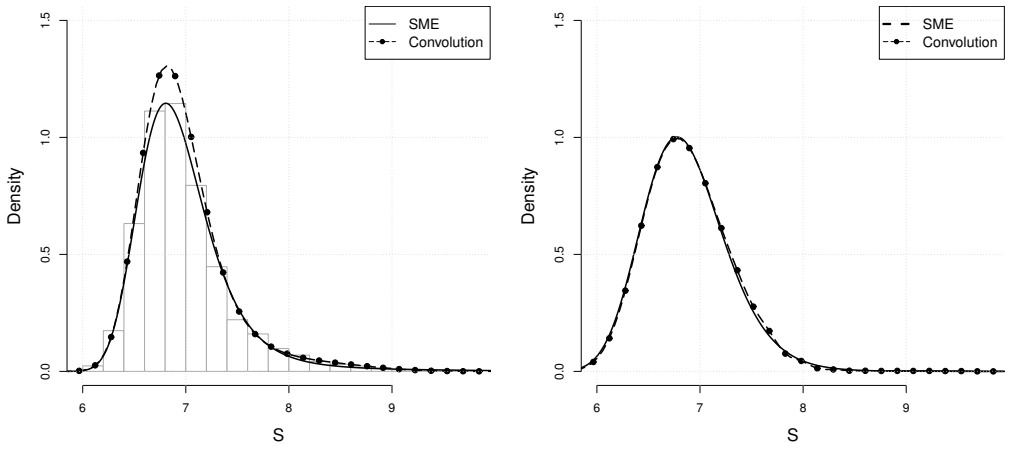
Table 6.4: MAE and RMSE distances between the histogram and the SME density

### 6.3.1 First case: a copula is known

For the results we present below, we suppose that we know that the losses in the different business lines can be coupled by some copula.

Since the output of the maximum entropy procedure is an analytical expression for each density, a joint density of $(S_1, ..., S_b)$ can be produced as explained before. With that joint density we can do two things. We can compute the Laplace transform of $S_1 + S_2 + ... + S_b$, and use it as input for the Maxentropic routine. Or we can use the iterated convolution procedure explained in Alexander (2003) to obtain the probability density of the sum. We do both and compare the results.

In Figure 6.2 we can observe the resulting densities with different combinations of copulas and correlations for the convolution and SME approach. In the first panel we plot the histogram of the sum supposing that the losses are independent. The continuous line in each of the panels, we plot the Maxentropic reconstruction of the density, and we use a dashed-dotted line to show the reconstruction obtained with the convolution technique. The copulas employed to simulate the data used as inputs for the other three are as follows: two Gaussian copulas with $\rho = 0.5$ and $\rho = 0.8$ and a Student copula with parameters $\rho_i = 0.7$, $\nu = 10$. The caption of each panel mentions which copula was used.




(a) Independent Copula

(b) Gaussian, $\rho_i = 0.5$

(c) Gaussian, $\rho_i = 0.8$

(d) t-copula, $\rho_i = 0.7$, $\nu = 10$

Figure 6.2: Density of the sum obtained by convolutions and the SME for different types of copulas and correlations

We add that the t-student copula is frequently used in the risk literature to model tail dependency.

In Table 6.5 we compare the reconstruction errors measured by MAE and RSME between the Maxentropic and the convolution technique. The measure of discrepancy were calculated as explained in Appendix A.1. These results confirm the visual results apparent in Figure 6.2.

| | *Error* | |
|---|---|---|
| Copula | $MAE$ | $RMSE$ |
| $Independent$ | 0.027 | 0.039 |
| $Gaussian$, $\rho = 0.5$ | 0.004 | 0.005 |
| $Gaussian$, $\rho = 0.8$ | 0.004 | 0.005 |
| $t - student$, $\rho = 0.7$ , $\nu = 10$ | 0.004 | 0.005 |

Table 6.5: MAE and RMSE distances between SME & Convolution approaches for several cases

Next, in Table 6.6 we present the confidence intervals for the empirical VaR and TVaR calculated with a large sample (of size 9000), and the VaR and TVaR computed from the density of the sum corresponding to the independent copula.

| | $VaR$ | | | | $TVaR$ | | | |
|---|---|---|---|---|---|---|---|---|
| $S$ | 0.95 | | 0.99 | | 0.95 | | 0.99 | |
| | Emp. | Conf. Int. | Emp. | Conf. Int. | Emp. | Conf. Int. | Emp. | Conf. Int. |
| | 7.877 | 7.849-7.910 | 8.601 | 8.560 -8.652 | 8.301 | 8.239 - 8.363 | 8.888 | 8.799 -8.979 |

Table 6.6: VaR, TVaR and confidence intervals for the simulated S of size 9000

To finish, in Table 6.7 we collect some results of numerical comparisons of the VaR and TVar for different density reconstructions.

### 6.3.2 Second case: The data is well collected

As we said, this corresponds to the case in which the institution collected the joint losses of each risk type, in order to obtain the aggregate loss per line of business, that is, the institution collected the vectors $(S_1, ..., S_b)$ and then computed the total loss $S = \sum S_b$. We shall present three examples organized as follows. For each we consider a copula with which we couple the marginal Maxentropic densities. Then, we generate data from the joint density and determine the density of total loss by two methods. First, using the Maxentropic procedure, and second using the convolution procedure with

| Copula | $VaR$ | | | | $TVaR$ | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.95 | | 0.99 | | 0.95 | | 0.99 | |
| | SME | Conv. | SME | Conv. | SME | Conv. | SME | Conv. |
| *Independent* | 7.793 | 7.828 | 8.698 | 8.603 | 8.527 | 8.332 | 9.190 | 8.922 |
| *Gaussian*, $\rho = 0.5$ | 7.560 | 7.672 | 8.045 | 7.983 | 7.868 | 7.850 | 8.272 | 8.292 |
| *Gaussian*, $\rho = 0.8$ | 7.682 | 7.672 | 8.136 | 8.138 | 7.9568 | 7.978 | 8.379 | 8.435 |
| $t - student$, $\rho = 0.7$, $\nu = 10$ | 7.672 | 7.707 | 8.138 | 7.983 | 7.936 | 7.914 | 8.348 | 8.317 |

Table 6.7: Comparison of VaR & TVaR for the SME & Convolution approaches for several cases

two different copulas: once with the right copula (to test the result provided with the Maxentropic procedure), and once with different copula to examine the effect of the wrong copula choice. The cases will be referred to using the name of the copula employed to generate the data. The point of the exercise is to show that, if the data was well collected, the Maxentropic procedure will be the right tool to use.

#### 6.3.2.1 Case 1: The Gumbel copula

For this example, the data consists of 1000 data vectors $(S_1, S_2, S_3)$ from the distribution obtained when we couple the marginals with a Gumbel copula. In Figure 6.3 we display the histogram of $S_1 + S_2 + S_3$, along with the Maxentropic reconstruction of the total loss densities obtained by the three methods mentioned before. The copulas used for comparison are the Clayton and Gumbel copula with parameters equal to 2 and 3, respectively.

The SME seems to produce better results in general than the convolution approach. In Table 6.8 we list the MAE and RSME between the empirical distribution and the densities obtained by the SME, and the convolution method with the Gumbel and Clayton copulas.

| Error | SME | Convolution (Gumbel) | Convolution (Clayton) |
|---|---|---|---|
| MAE | 0.005107 | 0.005453 | 0.02512 |
| RMSE | 0.006291 | 0.007283 | 0.03088 |

Table 6.8: Errors of the SME. Case 1.

In Tables 6.9 and 6.10 the values SME err. & Conv. err. are defined as the absolute difference from the empirical VaR and/or TVaR. The values of empirical VaR and TVaR are the sample quantiles of

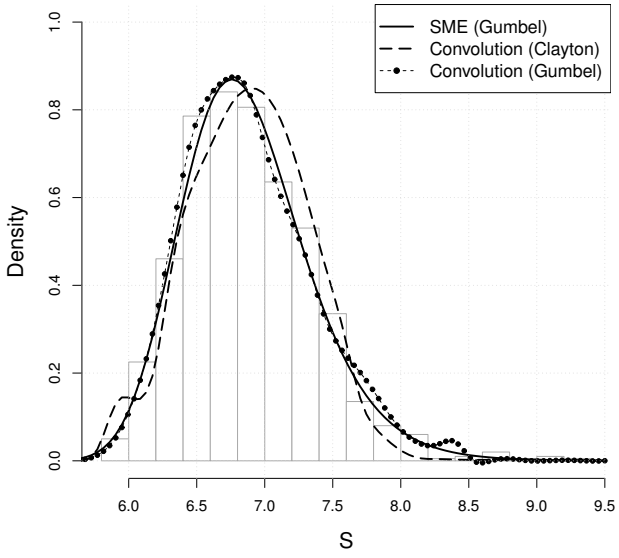



Figure 6.3: Reconstructed densities. Case 1.

the simulated sample and its confidence intervals are obtained using a bootstrap method. The VaR and TVaR of the methodologies used (SME and Convolution) are calculated as in Chapter 2, only for losses strictly greater than zero.

| VaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Empirical | SME err. | Conv. err. | $VaR_{inf}$ | $VaR_{sup}$ |
| 0.900 | 7.431 | 7.569 | 7.491 | 0.060 | 0.078 | 7.442 | 7.522 |
| 0.950 | 7.707 | 7.707 | 7.704 | 0.003 | 0.003 | 7.641 | 7.757 |
| 0.990 | 8.259 | 8.259 | 8.231 | 0.028 | 0.028 | 8.102 | 8.603 |
| 0.995 | 8.397 | 8.397 | 8.672 | 0.275 | 0.275 | 8.231 | 8.804 |
| 0.999 | 8.810 | 8.672 | 8.999 | 0.189 | 0.327 | 8.689 | 9.065 |

Table 6.9: Comparison of VaR for the SME & Convolution approaches. Case 1.

#### 6.3.2.2 Case 2: The Gaussian copula with negative correlation

We present the result of repeating the routine with a Gaussian copula that has negative correlation. In Figure 6.4 we can observe the histogram of the simulated data (500 data points) along with the reconstructions using a Gaussian copula with correlations of $-0.2$. We can also observe the result

| TVaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Empirical | SME err. | Conv. err. | $TVaR_{inf}$ | $TVaR_{sup}$ |
| 0.900 | 7.817 | 7.834 | 7.836 | 0.019 | 0.002 | 7.718 | 7.963 |
| 0.950 | 8.032 | 8.030 | 8.091 | 0.059 | 0.061 | 7.922 | 8.262 |
| 0.990 | 8.498 | 8.408 | 8.769 | 0.271 | 0.361 | 8.344 | 8.974 |
| 0.995 | 8.687 | 8.499 | 8.875 | 0.188 | 0.371 | 8.473 | 9.147 |
| 0.999 | 9.092 | 8.896 | 9.160 | 0.068 | 0.264 | 8.650 | 9.254 |

Table 6.10: Comparison of TVaR for the SME & Convolution approaches. Case 1.

obtained by applying the convolution technique using a t-student with correlations equal to -0.2 and 2 degrees of freedom as a guess. The numerical comparisons are displayed in Tables 6.11, 6.12 and 6.13.

To measure the quality of the procedures we compare the different reconstructions with the histogram of the simulation, using MAE & RMSE errors and VaR and TVaR calculations.

The SME seems to produce better results in general than the convolution approach.

| Error | SME | Convolution (Gaussian) | Convolution (t-Student) |
|---|---|---|---|
| MAE | 0.007989 | 0.01071 | 0.01430 |
| RMSE | 0.009605 | 0.01264 | 0.01652 |

Table 6.11: Errors of the SME. Case 2.

| VaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Empirical | SME err. | Conv.err. | $VaR_{inf}$ | $VaR_{sup}$ |
| 0.900 | 7.237 | 7.293 | 7.236 | 0.001 | 0.057 | 7.212 | 7.263 |
| 0.950 | 7.399 | 7.293 | 7.365 | 0.034 | 0.072 | 7.309 | 7.389 |
| 0.990 | 7.682 | 7.569 | 7.658 | 0.024 | 0.089 | 7.516 | 7.689 |
| 0.995 | 7.803 | 7.707 | 7.719 | 0.084 | 0.012 | 7.595 | 7.856 |
| 0.999 | 8.175 | 8.534 | 8.601 | 0.426 | 0.067 | 7.689 | 8.926 |

Table 6.12: Comparison of VaR for the SME & Convolution approaches. Case 2.

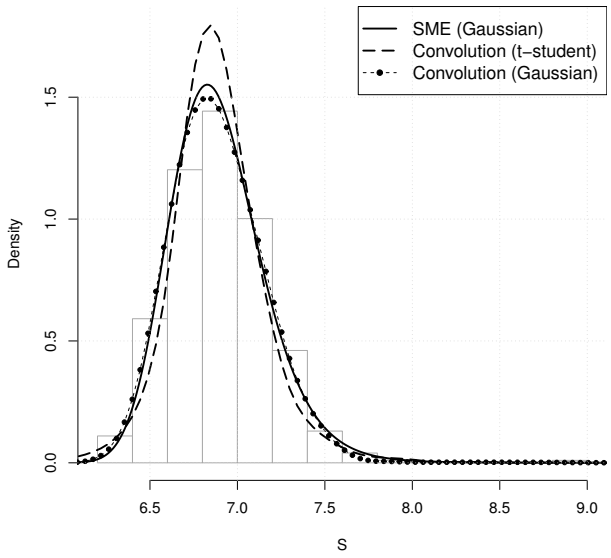



Figure 6.4: Density of the sum obtained by convolutions and SME. Case 2.

| TVaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Emp. | SME err. | Conv.err. | $TVaR_{inf}$ | $TVaR_{sup}$ |
| 0.900 | 7.439 | 7.404 | 7.419 | 0.020 | 0.015 | 7.336 | 7.536 |
| 0.950 | 7.571 | 7.514 | 7.549 | 0.022 | 0.035 | 7.415 | 7.735 |
| 0.990 | 7.892 | 7.837 | 7.920 | 0.028 | 0.083 | 7.551 | 8.443 |
| 0.995 | 8.052 | 8.089 | 8.047 | 0.005 | 0.042 | 7.578 | 8.926 |
| 0.999 | 8.529 | 8.758 | 8.334 | 0.195 | 0.424 | 7.658 | 8.926 |

Table 6.13: Comparison of TVaR for the SME & Convolution approaches. Case 2.

#### 6.3.2.3 Case 3: The Gaussian copula with positive correlation

Too finish we repeat the procedure, this time with a Gaussian copula that has a positive correlation. In Figure 6.6 we display the histogram of $S_1 + S_2 + S_3$, obtained with 500 of simulated data points, along with the reconstructions using a Gaussian copula with positive correlations of 0.8. In that figure we plot as well the result of applying the convolution technique using a t-student with correlations equal to 0.8 and 2 degrees of freedom as a guess. The numerical comparisons are summed up in Tables 6.14, 6.15 and 6.16.

As measure of the quality of the reconstructions we compute the MAE & RMSE errors between

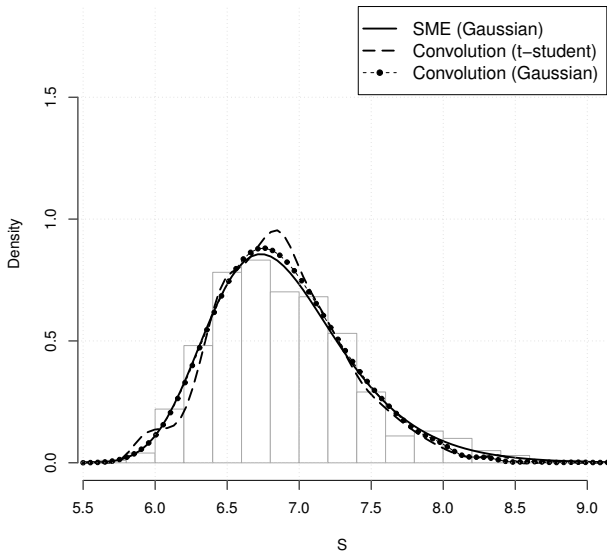



Figure 6.5: Reconstructed densities

Figure 6.6: Density of the sum obtained by convolutions and SME. Case 3.

the several densities obtained and the histogram. The results are shown in Table 6.14.

| Error | SME | Convolution (Gaussian) | Convolution (t-Student) |
|---|---|---|---|
| MAE | 0.007271 | 0.008513 | 0.01355 |
| RMSE | 0.008857 | 0.011520 | 0.01671 |

Table 6.14: Errors of the SME. Case 3

Another comparison is provided in Table 6.15 and 6.16, in which, respectively, the empirical VaR and TVaR are listed next to the values obtained from the densities determined by the Maxentropic and convolution procedures.

The SME seems to show better results in general than the convolution approach.

## 6.4 Concluding remarks

From the examples that we presented, it is clear that the Maxentropic procedure is a conceptually simple, fast and efficient method for determining the probability density of aggregate losses for the purpose of risk measurement, regulatory or economic capital determination and risk premia

| VaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Empirical | SME err. | Conv.err. | $VaR_{inf}$ | $VaR_{sup}$ |
| 0.900 | 7.569 | 7.517 | 7.523 | 0.046 | 0.006 | 7.460 | 7.596 |
| 0.950 | 7.845 | 7.672 | 7.885 | 0.040 | 0.213 | 7.793 | 7.967 |
| 0.990 | 8.397 | 8.138 | 8.246 | 0.151 | 0.108 | 8.143 | 8.511 |
| 0.995 | 8.534 | 8.293 | 8.514 | 0.020 | 0.221 | 8.205 | 8.578 |
| 0.999 | 8.948 | 9.069 | 8.842 | 0.106 | 0.227 | 8.511 | 8.842 |

Table 6.15: Comparison of VaR for the SME & Convolution approaches. Case 3.

| TVaR | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Approaches | | | Errors | | Confidence Interval 95% | |
| $\gamma$ | SME | Conv. | Emp. | SME err. | Conv.err. | $TVaR_{inf}$ | $TVaR_{sup}$ |
| 0.900 | 7.915 | 7.787 | 7.940 | 0.025 | 0.153 | 7.773 | 8.094 |
| 0.950 | 8.149 | 7.978 | 8.163 | 0.014 | 0.185 | 7.985 | 8.312 |
| 0.990 | 8.666 | 8.435 | 8.504 | 0.162 | 0.069 | 8.186 | 8.714 |
| 0.995 | 8.826 | 8.677 | 8.611 | 0.215 | 0.066 | 8.230 | 8.842 |
| 0.999 | 9.182 | 9.389 | 8.710 | 0.472 | 0.679 | 8.246 | 8.842 |

Table 6.16: Comparison of TVaR for the SME & Convolution approaches. Case 3.

computation.

An important moral of our analysis, is that it is worthwhile to devise data collection procedures in such a way that the joint empirical is explicitly incorporated in the losses. We saw that when this is done, the application of the Maxentropic procedure become a routine matter.

Even when the losses are not well recorded, the use of copulas in conjunction with the Maxentropic technique, provides an efficient method to tackle the problem of density determination for the LDA in the AMA approach to determine operational risk regulatory capital.

# Chapter 7

# Final Remarks

Finding the correct density is quite important because this provides a starting point to design policies, set premiums and reserves, calculate optimal reinsurance levels and calculate risk pressures for solvency purposes in insurance and risk management. Usually, risk decision makers and regulators make some parametric assumptions about the sample distribution, that in most of the cases incorrectly modeled fat tails and asymmetries that are crucial to explain the financial variables of interest, without take into account problems related with the quality and availability of the data. The error in the estimation of the parameters (that is worsened by the lack of data, common problem in Banking), even the small ones, contaminate the final density, skewing any measure that can be obtained through that density, as for example risk measures like VaR and TVaR in risk management.

In this work we provide several extensions of the maximum entropy methodologies that handle the problems mentioned before, providing a good density reconstruction over the entire range of values, modeling correctly the tail and the body of the distribution, through the use of a few number of fractional moments obtained from the data.

In this thesis we contribute to the literature in several aspects. In Chapter 2, we studied two Maxentropic methodologies called Standard Maximum Entropy method (SME) and Maximum Entropy in the Mean method (MEM) to find the density distribution of aggregated loss data using the Laplace transformation as an input. The considered example was a Poisson-LogNormal process. This case was chosen because it is not known the closed form of the Laplace transform of the compound sum. Then, to solve this case, we considered the numerical form of the Laplace over the simulated data, and we evaluated the results through the use of visualizations and several tests. The use of a numerical approximation poses the problem related with the sample errors, limitation that we try to address in Chapter 3. Additionally, in this paper we made the first attempt to decompose the random sums in

its frequency/severity distributions through the use of Maxentropic methodologies. The idea behind this paper is to prove the simplicity and value of the Maxentropic methods for the calculation of the regulatory capital through the estimation of risk measures like Value at Risk (VaR) and Tail value at risk (TVaR).

Later in Chapter 3, we examined two possible ways of deal with sample errors, within the framework of the SME method. The differences between the two approaches lay on the output of these methodologies. One of these methods provided an additional output that worked as an estimator of the additive error.

Next in Chapter 4, we extend the methodology that was called decompounding in Chapter 2, to separate the risk sources and to identify the statistical nature of the frequency and severity through the use of maximum entropy and clustering methods.

Additionally, in Chapter 6 we explore how to include correlation between losses, when one have several levels of aggregation, as is usual in Operational Risk Management.

At this point, we have concluded that maximum entropy based techniques are rather powerful to determine density distributions in risk. So, the pending question was: What happen when the data is scarce, as is common in baking?. At this point we know that the Maxentropic density depends upon the data used. So, in Chapter 5, we analyze the variability of the reconstructed densities to examine the impact of this variability on the estimation of VaR and TVaR, which are favored tool of financial regulators for bank capital adequacy, and that is extensively used in fields such as insurance and fund management. The results showed that even with a small quantity of data we can obtain a density distribution defined over the entire range of values.

The methodologies used in this thesis can be extended to calculate premiums or estimate the optimal price of a derivative. There are uncountable applications of this methodology and further research would be done in this line.

# Appendix

## A1. Quality of the reconstructions: Visual tests and goodness of fit tests

Once a density has been determined, it is necessary to test whether it is consistent with the data. The evaluation process is inherently a statistical problem, which involves exploring, describing, and making inferences about data sets containing observed and estimated values. Exploratory tests to asses the quality of a reconstruction, include visual comparisons through the use of graphical tools like reliability and calibration plots which measures the agreement between the estimation and the observed data.

Many of the comparisons that we perform in this thesis are made with respect to a simulated compound random variable whose probability density is unknown. In general, we chose a large enough sample to provide us with a good approximation of the data used as input to the Maxentropic methods. Throughout the Chapter 2 we referred to this sample as "the observed data". On the other hand, we also perform comparisons using a sample that we call "the test set", which is an independent data set that comes from the same population as the observed data, and which is used to audit the results. This is done in order to avoid the problem of over-fitting, which occurs when the model gives better results for the "observed" set than for other data set which comes from the same population. Besides this helps us to evaluate the ability of the Maxentropic density to perform well on unobserved data. Let us now list the different tests that we performed.

#### $L_1$ and $L_2$ distances

We computed the $L_1$ and $L_2$ distances between the histogram and the Maxentropic (reconstructed) density by means of

$$L_1 = \sum_{k=0}^{G-1} \int_{b_k}^{b_{k+1}} |f_S^*(s) - f_n(s)|ds + \int_{b_G}^{\infty} |f_S^*(s)|ds$$

$$L_2 = \sqrt{\sum_{k=0}^{G-1} \int_{b_k}^{b_{k+1}} (f_S^*(s) - f_n(s))^2 ds + \int_{b_G}^{\infty} (f_S^*(s))^2 ds}$$

where $b_k$ and $b_{k+1}$ are the boundaries of the bins in the histogram, $G$ is the number of bins, $f_S^*$ is the reconstructed density and $f_n$ is the density obtained from the histogram (i.e. (frequency in the bin $k$)/(size of the data set)). This measure has the disadvantage of depending on the location and the number of bins of the histogram.

**MAE and RMSE distances**

To overcome the bin dependency we computed the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) which measure the distance between the cumulative distribution functions of the empirical and the reconstructed densities. These are computed as follows

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |F_S^*(s_j) - F_n(s_j)|$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (F_S^*(s_j) - F_n(s_j))^2}$$

These measures do not depend on the choice of bins to create the histogram, but only on the sample data, see Hyndman et al. (2006).

**PIT based analysis**

A different way to use the empirical and reconstructed cumulative distribution functions (CDF), called probability integral transform (PIT) by Rosenblatt (1952), relies on the fact that $F_S(S)$ is uniformly distributed on $[0, 1]$, Thus, if the $s_j$ are the sample points examining the collections $F_S^*(s_j)$, for uniformity, or deviation thereof, we decide whether the reconstruction may have failed to capture some aspect of the underlying data generation process.

To test uniformity and independence in the PIT test, a visual inspection of a PIT-histogram and autocorrelation plots is used along with additional tests like the KS-test, the Anderson-Darling test, and the Cramér-Von Mises test (Tay et al., 2000). Additionally, we also consider the Berkowitz back

test approach, which consists of taking the inverse normal transformation of the PIT and then applying a joint test for normality and independence, that is sometimes combined with the normality test of Jarque-Bera. Let us briefly describe some of these tests now.

**Kolmogorov-Smirnov test**

The Kolmogorov-Smirnov test is a test of uniformity which verifies the differences of fit between the empirical distribution function (EDF) and the estimated (reconstructed) distribution function, using the largest absolute observed distance between them,

$$D_n = \underset{s}{sup}|F_n(s_j) - F_S^*(s_j)|$$

where $n$ is the number of data points; $\{s_j|j = 1, ..., n\}$ are the sample data points of the total losses $S$; $F_n(\cdot)$ is the (cumulative) empirical distribution function; and $F_S^*(\cdot)$ is the Maxentropic (cumulative) distribution function. $\sqrt{n}D_n$ is the statistic used to perform the hypothesis. A problem with this test is that the KS statistic depends on the maximum difference, without considering the whole estimated distribution. This is important when the differences between distributions are suspected to occur only at the upper or lower end of their range (Cruz, 2002). This may be particularly problematic in small samples. Besides, little is known about the impact of the departures from independence on $D_n$, i.e., if we are not sure about the independence of the sample, we would not be sure what is the meaning of the results of the test. There exist other EDF tests, which in most situations are more effective than the simple Kolmogorov-Smirnov test. Further details about the K-S test can be seen in Stephens (1974) or in Ruppert (2004).

**Anderson-Darling test**

This is a more sophisticated version of the KS approach, which emphasizes more the tails of the distribution than the KS test, Cruz (2002), and it is based on the quadratic difference between $F_n(s)$ and $F_S^*(s)$. Here the AD statistic is computed as follows:

$$A_n^2 = n \int_{-\infty}^{\infty} |F_n(s) - F_S^*(s)|^2 \Psi(s) f^*(s) ds,$$

where $\Psi(s) = \frac{1}{F_S^*(s)(1-F_S^*(s))}$ is a weight function; $n$ is the number of data points, and $\{s_j|j = 1, ..., n\}$ the observed (simulated in our case) total loss $S$ sample, $F_n(\cdot)$ is the empirical (cumulative) distribution function and $F_S^*(\cdot)$ is the (cumulative) Maxentropic distribution function.

When $\Psi(s) = 1$, the Anderson-Darling (AD) statistic reduces to the statistic which is known as the Cramér-von Mises statistic. The Anderson-Darling (AD) statistic behaves similarly to the Cramér-von Mises statistic, but is more powerful to test whether $F_S^*(s)$ departs from the true distribution in the tails, especially when there appear to be many outlying values. For goodness-of-fit testing, departure in the tails is often important to detect, and $A_n^2$ is the recommended statistic Marsaglia et al., (2004) or Thas (2010).

**Berkowitz back test**

Berkowitz (2001) proposed the transformation $z_n = \Phi^{-1}(\int_{-\infty}^{s_n} f_S^*(s)ds) = \Phi^{-1}(F(s_n))$, to make the data i.i.d standard normal under the null hypothesis. This allows to make use of powerful battery of available normality tests, instead of relying on uniformity tests. Besides that, Berkowitz back test provides a joint test of normality and independence. See Christoffersen (2012) for more.

It is usually convenient to supplement the Berkowitz test with at least one additional test for normality, for example Jarque-Bera test. This extra test ensures that we test for the predicted normal distribution, and not just for the predicted values of the parameters $\rho$, $\mu$ and $\sigma$.

**Jarque-Bera test**

The standard Jarque-Bera-test is a test of normality. This test uses a sum of the empirical skewness coefficient and the empirical kurtosis as statistic to test for formality, or deviations thereof.

Unfortunately, the standard JB statistic is very sensitive to extreme observations, due to the fact that the empirical moments are known to be very sensitive to outliers; and the sample variances is more affected by outliers than the mean, disturbing the estimations of the sample skewness and kurtosis. To solve the problem a robust modification of the Jarque-Bera test was proposed by Gel et al. (2008), which utilizes the robust standard deviation (namely the average absolute deviation from the median (MAAD)) to estimate a more robust kurtosis and skewness from the sample (Gel et al.,2008). Consider Ruppert (2011) as well.

**Correlograms**

Tests like KS and AD do not prove independence, so to asses whether the probability integral

transformation (PIT) of the data is i.i.d, we use a graphical tool, the correlogram, which helps in the detection of particular dependence patterns and can provide information about the deficiencies of the density reconstructions Diebold et al., (1998).

**Reliability Diagram or QQ-plots**

This plot serves to determine the quality of a fit by the proximity of the fitted curve to the diagonal, the closer the better the approximation, deviations from the diagonal gives the conditional bias. Additionally, this plot could indicates problems as over-fitting, when the fitted curve lies below the diagonal line and under-fitting when the fitted curve lies above the line. See Härdle et al. (2002).

**Marginal Calibration Plot**

The calibration plot is based on the idea that a system is marginally calibrated if its estimations and observations have the same (or nearly the same) marginal distribution. Then, the graphical device is a plot of $F_S^*(s_j) - F_n(s_j)$ versus $s_j$. Under the hypothesis of marginal calibration, we expect minor fluctuations about zero. The same information may be visualized in terms of the quantiles $Q(F_S^*(\cdot), q) - Q(F_n(\cdot), q), \quad q \in (0,1)$ of the functions $F_S^*(\cdot)$ and $F_n(\cdot)$, see Gneiting et al. (2007).

# A2. Further mathematical details: Chapter 3

Here, we complete the basic modeling process missing in Section 3.2.2 and then some elementary computations useful for the minimization procedures.

To put the error estimation procedure into a model, consider the space $\Omega := [0,1] \times \mathbf{C}$ on which we consider the obvious $\sigma-$algebra of Borel subsets, and consider the reference measure $dQ_0(dy, d\boldsymbol{\xi}) = dy \prod (\epsilon_{a_i}(d\xi_i) + \epsilon_{b_i}(d\xi_i))$. On $\Omega$ we define the "generalized" moments

$$\phi : \Omega \to \mathbb{R}, \text{ where } \phi(y, \boldsymbol{\xi}) = y^{\alpha_i} + \xi_i$$

where $\xi_i$ is the $i-$th coordinate of $\boldsymbol{\xi}$. Also, when $P << Q_0$, then there exists a density $f(y)$ and numbers $p_k$ for $k = 1, ..., K$ such that $dP(dy, d\boldsymbol{\xi}) = f(y)dy \prod (p_i \epsilon_{a_i}(d\xi_i) + (1-p_i)\epsilon_{b_i}(d\xi_i))$. The maximum entropy problem can now be stated as

$$\text{Find } P << Q_0, \text{ that maximizes } S_{Q_0}(P) \text{ given by (3.14) and } E_P[\phi_i] = \mu_i \;\; i = 1, ..., K. \quad \text{(A.2.1)}$$

Now the procedure is standard, and the result is stated in the easy to prove, but important

**Theorem 2.** *Suppose that the infimum $\boldsymbol{\lambda}^*$ of $\Sigma(\boldsymbol{\lambda})$ given by (3.17) is reached in the interior of the set $\{\boldsymbol{\lambda} \in \mathbb{R}^K | Z(\boldsymbol{\lambda}) < \infty\}$. Then the solution to the maximum entropy problem (A.2.1) is given by (3.16)*

*Proof.* All it takes is to differentiate (3.17) with respect to $\lambda_i$ equate to 0 and read the desired conclusion.

**Comment** This is a rather diluted of the general result presented in Cherny (2003) or Jaynes (1957), but is enough to keep us going.

In the next two subsections we add the explicit computation of the gradient of $\Sigma(\boldsymbol{\lambda})$ for those to use gradient based methods for its minimization.

## Derivative of $\Sigma(\boldsymbol{\lambda})$ when reconstructing with data in a confidence interval

Having determined the confidence interval $[a_i, b_i]$ for each $i = 1, ..., 8$ the next step is to minimize $\Sigma(\boldsymbol{\lambda})$. For the first case, in case of need, here is the derivative. It invokes (3.11)

$$\frac{\partial \Sigma(\boldsymbol{\lambda})}{\partial \lambda_i} = -\int_0^1 y^{\alpha_i} \frac{e^{-\sum_{k=1}^8 \lambda_k y^{\alpha_k}}}{Z(\boldsymbol{\lambda})} dy + k_i.$$

where $k_i = b_i$ if $\lambda_i > 0$ and $k_i = a_i$ if $\lambda_i < 0$, and in the rare case that $\lambda_i = 0$, choose $k_i \sim U(a_i, b_i)$. And as before, $Z(\boldsymbol{\lambda}) = \int_0^1 e^{-\sum_{i=1}^8 \lambda_i y^{\alpha_i}} dy$.

Once the minimizer $\boldsymbol{\lambda}^*$ has been found, the Maxentropic density (in the changed variables is

$$g^*(y) = \frac{e^{-\sum_{i=i}^8 \lambda_i^* y^{\alpha_i}}}{Z(\boldsymbol{\lambda})}$$

when the change of variables $f^*(t) = e^{-t} g^*(e^{-t})$.

## Derivative of $\Sigma(\boldsymbol{\lambda})$ when reconstructing with error estimation

This time the derivatives of $\Sigma(\boldsymbol{\lambda})$ are a bit different. On one hand

$$\frac{\partial \Sigma(\boldsymbol{\lambda})}{\partial \lambda_i} = -\int_0^1 y^{\alpha_i} \frac{e^{-\sum_{i=1}^8 \lambda_i y^{\alpha_i}}}{Z(\boldsymbol{\lambda})} dy - \frac{a_i e^{-a_i \lambda_i^*} + b_i e^{-b_i \lambda_i^*}}{e^{-a_i \lambda_i^*} + e^{-b_i \lambda_i^*}} + \mu_i.$$

Once the minimizing $\boldsymbol{\lambda}^*$ has been found, the routine should be the same as before. That is, use (3.17) to obtain the density and plot along with the result obtained in the previous section.

# Bibliography


[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In B. N. Petrov and F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.

[3] Alexander, C. (2003). A better approach to operational risk aggregation. *Operational Risk & Regulation*. Available online at `http://www.risk.net/data/basel/pdf/basel_or_0403_tech.pdf`.

[4] Aue, F. & Kalkbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk, *Jour. Operat. Risk*, 1, 49-93.

[5] Barzilai, J., & Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8, 141-148.

[6] Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19, 465-474.

[7] Berliner, B. (1984). Some Thoughts on Uncertainty, the Transfer of Knowledge and Insurability, *The Geneva Papers on Risk and Insurance*, 9, 380-395

[8] Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22, 719-725.

[9] B$\phi$gsted, M., & Pitts, S. M. (2010). Decompounding random sums: a nonparametric approach. *Annals of the Institute of Statistical Mathematics*, 62, 855-872.

[10] Borwein, J.M. & Lewis, A.S (2000). Convex Analysis and Nonlinear Optimization, *CMS Books, Springer Verlag*, New York.

[11] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

[12] Brocket, P.L. (1991). Information theoretic approach to actuarial science: a unification and extension of relevant theory and applications, *Transac. Soc. Actuaries*, 43, 73-135.

[13] Brockmann, M. & Kalkbrener, M. (2010). On the aggregation of risk, *Jour. Operat. Risk*, 12, 45-68.

[14] Buchmann, B., & Grübel, R. (2003). Decompounding: an estimation problem for Poisson random sums. *Annals of statistics*, 1054-1074.

[15] Buchmann, B., & Grübel, R. (2004). Decompounding Poisson random sums: recursively truncated estimates in the discrete case. *Annals of the Institute of Statistical Mathematics*, 56, 743-756.

[16] Caussinus, H. & Ruiz-Gazen, A. (2009) Exploratory Projection Pursuit, in Data Analysis (ed G. Govaert), *ISTE, London, UK*. doi: 10.1002/9780470611777.

[17] Cherny, A. & Maslov, V. (2003). *On minimization and maximization of entropy functionals in various disciplines*, Theory of Probab. and its Applic, 17, 447-464.

[18] Christoffersen, P.F. (2012). Elements of financial risk management, *Springer-Verlag*, Berlin.

[19] Crow, E. L., & Shimizu, K. (1988). Lognormal distributions: Theory and applications, *Marcel Dekker*, New York.

[20] Cruz, M. G. (2002). Modeling, measuring and hedging operational risk, *John Wiley & Sons*, New York.

[21] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 146-158.

[22] Csiszár, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 768-793.

[23] Den Iseger, P.W, Smith, W.A.J. & Dekker, R. (1997). Computing compound distributions faster. *Insurance: Mathematics and Economics*, 20, 23-34.

[24] Diebold, F. X., Gunther, T. A. & Tay, A. S. (1998), Evaluating density forecasts with applications to financial risk management, *International Economic Review*, 39, 863-883.

[25] Embrechts, P. & Frei, M. (2007). Panjer recursion versus FFT for compound distributions, *Math. Meth. of Operational Res.*, 69, 497-508.

[26] Esscher, F. (1932) On the probability function in the collective theory of risk, *Skandinavisk Aktuarietidskrift*, 15, 175-195.

[27] Gamboa, F. Minimisation de l'information de Kullback et minimisation de l'entropie sous une contrainte quadratique. *CRAS Paris, Sèrie I Math.*, 1988, 306, 425-427.

[28] Gel, Y. R., & Gastwirth, J. L. (2008). A robust modification of the Jarque-Bera test of normality. *Economics Letters*, 99(1), 30-32.

[29] Gomes, E., & Gzyl, H. (2014). Disentangling Frequency Models, *Journal of Operational Risk*, **9**(2), 3-21.

[30] Gomes-Gonçalves, E, Gzyl, H., Mayoral, S. (2014). Density Reconstructions with Errors in the Data. *Entropy*, 16, 3257-3272.

[31] Gomes-Gonçalves, E. Gzyl, H., Mayoral, S. (2015a). Two maxentropic approaches to determine the probability density of compound risk losses. *Insurance: Mathematics and Economics*, 62, 42-53.

[32] Gomes-Gonçalves, E., Gzyl, H. & Mayoral, S. (2015b). Maxentropic approach to decompound aggregate risk losses. *Insurance: Mathematics and Economics*, 64, 326-336.

[33] Gomes-Gonçalves, E., Gzyl, H., & Mayoral, S. (2015c). LDA: Analysis of the Sample Dependence in Density Reconstruction by Maxentropic Methods. *Available at SSRN 2664214*.

[34] Gomes-Gonçalves, E., Gzyl, H., & Mayoral, S. (2016). Maximum Entropy Approach to the Loss Data Aggregation Problem. *Journal of Operational Risk*, 11(1).

[35] Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 243-268.

[36] Gzyl , H (1995). The Method of Maximum Entropy, *World Scient. Pubs., Singapore.*

[37] Gzyl, H. & Mayoral, S. A method for determining risk aversion functions from uncertain market prices of risk *Insurance: Mathematics and Economics*, 2010, 47, 84-89.

[38] Gzyl, H., Novi-Inverardi, P.L. & Tagliani, A. (2013). A comparison of numerical approaches to determine the severity of losses. *Journal of Operational Risk*, 8 pp.3-15.

[39] Gzyl, H. & Recht, L. (2006). Geometry on the space of probabilities II: projective spaces and exponential families, *Revista Matemática Iberoamericana*. 3, 833-849

[40] Gzyl, H., & Tagliani, A. (2010). Hausdorff moment problem and fractional moments. *Applied Mathematics and Computation*, 216(11), 3319-3328.

[41] Gzyl, H., & Tagliani, A. (2012). Determination of the distribution of total loss from the fractional moments of its exponential. *Applied Mathematics and Computation*.

[42] Gzyl, H. & Velásquez, Y. (2011). Linear Inverse Problems: The maximum entropy connection, *World Scientific Pubs*, Singapore, Singapore.

[43] Haberman, S., Kallaf-Allah, M. & Verrall, R. (2011) Entropy, longevity and the cost of annuities, *Insurance: Mathem. and Econ.*, 48, 197-204.

[44] Härdle, W., Kleinow, T., & Stahl, G. (2002). Applied quantitative finance, *Springer-Verlag*, Berlin.

[45] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4), 679-688.

[46] Hothorn, T., & Everitt, B. S. (2009). A handbook of statistical analyses using R (Vol. 12). *CRC Press*.

[47] Jaynes, E. T., (1957a), Information Theory and Statistical Mechanics, Phys. Rev., 106, 620.

[48] Jaynes, E. T., (1957b), Information Theory and Statistical Mechanics II, Phys. Rev., 108, 171.

[49] Kapur, J. N. (1989). Maximum-entropy models in science and engineering. *John Wiley & Sons*.

[50] Le Courtois, O., & Quittard-Pinon, F. (2008). Fair valuation of participating life insurance contracts with jump risk. *The Geneva Risk and Insurance Review*, 33, 106-136.

[51] Leipnik, R. B. (1991). On lognormal random variables: I-the characteristic function. *The Journal of the Australian Mathematical Society*. Series B. Applied Mathematics, 32(03), 327-347.

[52] Li, J. S-h. (2010). Pricing longevity risk with the parametric bootstrap: A maximum entropy approach, *Insurance: Mathem. Econ.*, 47 176-186.

[53] Lin, G.D. (1992). Characterizations of distributions via moments, *Sankhya: The Indian Journal of Statistics*, 54, 128-132.

[54] Marsaglia, G., & Marsaglia, J. (2004). Evaluating the anderson-darling distribution. *Journal of Statistical Software*, 9, 1-5.

[55] Martin-Löf, A. (1986) Entropy, a Useful Concept in Risk Theory, *Scandinavian Actuarial Journal*, 69, 223-35.

[56] McLachlan, G., & Peel, D. (2004). Finite mixture models. *John Wiley & Sons.*

[57] McNeil, A., Frey, R. & Embrechts, P. (2005) *Quantitative Risk Management*, Princeton Univ. Press, Princeton, NJ.

[58] Mosteller, D.C., & Tukey, J.W. (Eds.) (2011). Exploring data tables, trends, and shapes (Vol. 101). *John Wiley & Sons.*

[59] Panjer, H. H. (2006). Operational Risk: Modeling Analytics. *Wiley*, New York.

[60] Peña D., Prieto F. J.(2001a). Cluster Identification Using Projections, *Journal of the American Statistical Association*, 96, 1433-1445.

[61] Peña D. & Prieto F. J.(2001b). Cluster identification routines for the paper Cluster Identification Using Projections, *J. Amer. Stat. Assoc.*, 96, 1433-1445.

[62] Peña, D., Prieto, F.J. & Rocke, D.M., (2001). Discussion of Multivariate Outlier Detection and Robust Covariance Matrix Estimation *Technometrics*, 43, 300-303.

[63] Ramaswamy, V., Wayne Desarbo, D. Reibstein & W. Robinson (1993). An Empirical Pooling Approach for Estimating Marketing Mix Elasticities with PIMS Data, *Marketing Sciences*, 12,103-124

[64] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of risk*, 2, 21-42.

[65] Rolski T. , Schmidli H. , Schmidt V. , Teugles J. (1999). Stochastic Processes in Insurance and Finance, *John Wiley & Sons*, New York.

[66] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3), 470-472.

[67] Ruppert, D. (2011). Statistics and data analysis for financial engineering (pp. 183-216). New York, NY: *Springer*.

[68] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464.

[69] Shen, P. & Jia, W. (2012) Maximum entropy based operational risk distribution estimation of commercial banks,in *Management Science and Engineering* (ICMSE), 2012 International Conference IEEE, 1398-1403.

[70] Shevchenko, P. V. (2011). Modelling operational risk using Bayesian inference. *Springer Science & Business Media*.

[71] Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347), 730-737.

[72] Stone, J. V. (2004). Independent component analysis: a tutorial introduction. *Massachusetts Institute of Technology*.

[73] Swait, J. D. (2011). Discrete choice theory and modeling. T*he Oxford Handbook of the Economics of Food Consumption and Policy*, 119.

[74] Tay, A. S., & Wallis, K. F. (2000). Density forecasting: a survey. *Journal of forecasting*, 19(4), 235-254.

[75] Thas, O. (2010). Comparing distributions. *New York: Springer*.

[76] Temnov, G. & Warnung, R. (2008) A comparison of risk aggregation methods for operational risk, *Jour. Operat. Risk*, 3, 3-23.

[77] Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). Statistical analysis of finite mixture models. *Wiley*, Chichester

[78] Van Es, B., Gugushvili, S., & Spreij, P. (2007). A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13, 672-694.

[79] Wang, A.S. (1997). Aggregation of correlated risk portfolios: Models and algorithms.*Casualty Actuarial Society*, pp. 848-939.

[80] Werner, M. (2003). Identification of multivariate outliers in large data sets. *Doctoral dissertation, University of Colorado at Denver.*

[81] Zhou, R., Cai, R., & Tong, G. (2013). Applications of entropy in finance: A review. *Entropy*, 15(11), 4909-4931.