Insurance: Mathematics and Economics 71 (2016) 145–153

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Loss data analysis: Analysis of the sample dependence in density reconstruction by maxentropic methods

Erika Gomes-Gonçalves [a], Henryk Gzyl [b,*], Silvia Mayoral [a]

[a] *Department of Business Administration, Universidad Carlos III de Madrid, Spain*

[b] *Centro de Finanzas, IESA, Caracas, Bolivarian Republic of Venezuela*

ARTICLE INFO

Article history:
Received September 2015
Received in revised form
August 2016
Accepted 24 August 2016
Available online 8 September 2016

Keywords:
Loss distributions
Loss data analysis
Maximum entropy density reconstruction
Sample dependence of density estimation
Sample dependence of risk measures

ABSTRACT

The problem of determining probability densities of positive random variables from empirical data is important in many fields, in particular in insurance and risk analysis. 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 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 issues that we consider are of special interest for the advanced measurement approach in operational risk, which is based on loss data analysis to determine regulatory capital, as well as to determine the loss distribution of risks that occur with low frequency.

1. Introduction and preliminaries

One of the methodologies that banks can use to determine regulatory capital for operational risk is the advanced measurement approach, which is based on the possibility of determining the operational risk capital from the probability density of the yearly losses. Actually the problem of determining the probability density of compound losses has received a lot of attention since a long time ago in the literature devoted to insurance matters. But since the Basel Committee proposals to measure and manage operational risk, it had a revival. See Cruz (2002), Panjer (2006) or Shevchenko (2011) for a variety of aspects about the problem and for procedures to obtain the probability density of aggregate losses from the historical data. The two volumes just mentioned are part of a large body of literature devoted to the theme. To mention just a few papers rapidly cascading into a large pool of literature, consider Aue and Kalkbrenner (2006), Temnov and Warnung (2008) and Brockmann and Kalkbrenner (2010)

In a series of previous papers, see Gzyl et al. (2013), we explored the usefulness of the maximum entropy based (maxentropic) procedure to determine the density of aggregate losses, and compared the procedure to standard procedures like Fourier inversion techniques, direct computation of the total loss density by convolution or reconstruction from integral moments.

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 α by recasting the problem into a fractional moment problem on $[0, 1]$ after the transformation $x = e^{-s}$. 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 developed in Gomes-Gonçalves et al. (2014), and that the procedure itself makes no assumptions about the statistical nature of the data.

That the maxentropic methods work when the amount of data is large, was the subject matter of Gomes-Gonçalves et al. (2015a,b). In the first of these, the aim was to examine the performance of two maxentropic methods to determine the density of aggregate losses. In the second, it was supposed that the losses may be produced by different sources of risk, that is, that may have different types of events producing losses at different rates, but the available data consists of the total loss. The problem in this case is to disentangle the different sources of risk, and to determine the nature of the individual losses.

In our previous work, we have seen that maximum entropy based techniques are quite powerful to determine density

* Corresponding author.
E-mail addresses: epgomes@emp.uc3m.es (E. Gomes-Gonçalves), henryk.gzyl@iesa.edu.ve (H. Gzyl), smayoral@emp.uc3m.es (S. Mayoral).

http://dx.doi.org/10.1016/j.insmatheco.2016.08.007

distributions when the amount of data is large. The maxentropic techniques work equally well when the amount of data is small, a situation which happens when analyzing operational risk data for example. In this case, one expects the resulting densities to depend on the sample used to compute the moments. Fortunately, the dependence of the maxentropic density on the sample is such that its variability can be analyzed explicitly. It is our aim here to analyze the variability of the reconstructed densities, and on the other hand to examine the impact of this variability on the estimation of the regulatory capital.

To state the problems with which we shall be concerned, let us begin saying that we are interested in compound variables of the type $S = \sum_{k=0}^{N} X_k$, 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 $\{n; x_1, \dots, x_n\}$, where n is the number of risk events and $\{x_1, \dots, x_n\}$ are the losses occurring at each event. The aggregate loss for that year is $s = \sum_{k=0}^{n} x_k$. When $n = 0$ there were no losses, the sum is empty and $s = 0$. Suppose that the record consists of M 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 $(n_j, x_1, \dots, x_{n_j})$ and $s_j = \sum_{k=0}^{n_j} x_k$. For us, an observed sample (of losses) will be an (M) vector $\omega = (s_1, \dots, s_M)$.

The Laplace transform of S is estimated by

$$\psi(\alpha) = \frac{1}{M}\sum_{j=1}^{M} e^{-\alpha S_j} \tag{1}$$

where S_j denotes the losses experienced during the jth year. Later on we shall consider the moments corresponding to K values of the parameter α. 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

$$\mu(\alpha) = \frac{\psi(\alpha) - P(N=0)}{1 - P(N=0)} \tag{2}$$

where $P(N = 0)$ is 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 $y = e^{-s}$, we can rewrite (1) as

$$\psi(\alpha) = \frac{1}{M}\sum_{j=1}^{M} y^{\alpha} \tag{3}$$

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 problem 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 $f_Y(y)$ has been obtained, the change of variables $f_S(x) = e^{-x} f_Y(e^x)$ provides us with the desired density.

As we shall have to emphasize the dependence of f_Y on the size M of the sample, we shall drop the Y and simply write f for it, and we shall use the notation $f_M(\omega, x)$ to denote the maxentropic density reconstructed from the collection of K moments as in (2). We describe how to obtain f_M in Section 2 when we explain the maximum entropy methods. Note that as (2) depends on the sample ω, then f_M depends on ω. To further specify our goals, there are three things that we want to understand, or develop intuition about. First, how much does f_M change when we change ω. Second, how much do some basic risk measures change when we change ω, and third, what happens as M becomes very large.

As we shall recall in Section 2, the connection between the moments $\mu(\alpha, \omega)$ is quite non-linear, the study of the variability of f_M is not that simple, nevertheless, a few things can be said. We shall carry this out in Section 3, while in Section 4 we examine this issue by numerical simulations. The data that we use as input consists of an aggregation of risks of different nature, so it is not a simple compound model as that considered in our previous work.

We close this section mentioning that there are other methods to deal with the problem of inferring loss densities from data, some simpler and some more elaborate. Consider the well known Panjer recursion technique, or the fast Fourier transform as described, say in Embrechts and Frei (2007) or Shevchenko (2011), or the cubic interpolation spline proposed in den Iseger et al. (1997). One of the interesting features of the representation provided by the maxentropic approach is that it provides an explicit analytic representation of the density which can be used as a starting point for a systematic analysis of the sample dependence

2. The maximum entropy inversion techniques

We shall describe two complementary approaches to the density reconstruction problem. First, the standard maximum entropy (SME) method and then the standard maximum entropy with error (SMEE) in the data, which is useful to cope with the issue of data uncertainty.

2.1. The standard maximum entropy method

The procedure 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:

$$\int_0^1 y^{\alpha_k} f_Y(y) dy = \mu_Y(\alpha_k) \quad \text{for } k = 0, 1, \dots, K \tag{4}$$

seems to have been originally proposed in Jaynes (1957). 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 (4) is convex. One can pick up a point in that class one by maximizing (or minimizing) a concave (convex) functional (an "entropy") that achieves a maximum (minimum) in that class. That extremal point is the "maxentropic" solution to the problem. It actually takes a standard computation to see that, when the problem has a solution, it is of the type

$$f(y) = \exp\left(-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}\right) \tag{5}$$

which depends on the α's through the λ's. It is usually customary to write $e^{-\lambda_0^*} = Z(\boldsymbol{\lambda}^*)^{-1}$, where $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_K^*)$ is a K-dimensional vector. Clearly, the generic form of the normalization factor is given by

$$Z(\boldsymbol{\lambda}) = \int_0^1 e^{-\sum_{k=1}^{K} \lambda_k y^{\alpha_k}} dy. \tag{6}$$

With this notation, the generic form of the solution looks like

$$f^*(y) = \frac{1}{Z(\boldsymbol{\lambda}^*)} e^{-\sum_{k=1}^{K} \lambda_k^* y^{\alpha_k}} = e^{-\sum_{k=0}^{K} \lambda_k^* y^{\alpha_k}}. \tag{7}$$

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

$$\Sigma(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \ln Z(\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{\mu}_Y \rangle \tag{8}$$