

# Estimating copula functions from limited data

David Allwright, Vera Hazelwood (Smith Institute)<sup>1</sup>

Trevor Maynard (Lloyds of London)<sup>2</sup>

16 November 2006

**Smith** *institute*  
*for industrial mathematics and system engineering*

**LLOYD'S**

---

<sup>1</sup> Emails: [David@smithinst.co.uk](mailto:David@smithinst.co.uk), [Vera@smithinst.co.uk](mailto:Vera@smithinst.co.uk), address: Smith Institute, Surrey Technology Centre, Surrey Research Park, Guildford, Surrey GU2 7YG

<sup>2</sup> Email: [Trevor.Maynard@lloyds.com](mailto:Trevor.Maynard@lloyds.com)

# 1 Introduction

It is common to wish to make a probabilistic model for a real-world scenario in circumstances where there is only limited data on which to build the model. One example might be in traffic modelling in a communication system: the rare events (very large jobs that will overflow a buffer) do not happen often, and yet one has to design a system that handles such a situation well enough to satisfy some required guarantee of quality of service. Another example would be where one has to prove that a safety system has a very low probability of failure, in circumstances where there is very little observed data in the tails of the distribution that would give rise to failure. But the application that we shall discuss in this paper is in insurance, where one may wish to have reserves that will only fail to cover a year's losses with probability 1/200, and yet there may only be 10 years' data on which to base the assessment. We shall address the limits of what can be done in such cases, and in particular what are the limits on estimating *correlated* risks (or equivalently copula functions) with limited data.

## 2 The univariate case

Although our aim is to consider correlated losses, we begin by considering the question of a *single* loss: the univariate case as opposed to multivariate. So this loss is modelled as a real-valued random variable  $L$  with an unknown probability density function  $f(x)$  and cumulative distribution function  $F(x)$ . The circumstances outlined in the introduction are that we observe  $n$  independent losses  $L_1, L_2, \dots, L_n$  and from them we want to estimate some point in the upper tail of the distribution. We denote this point by  $L^* = L^*(p)$ , and it is the loss such that

$$p = \Pr(L \leq L^*) = \int_{-\infty}^{L^*} f(x) dx = F(L^*). \quad (1)$$

We are thinking particularly of the case where  $p$  is close to 1, so  $q = 1 - p = \Pr(L > L^*)$  is small, and  $L^*$  is in the upper tail of the distribution: the probability of a loss greater than  $L^*$  is only  $q$ , which would be say 0.005 if losses greater than  $L^*$  are supposed to occur on average once every 200 years.

We are going to estimate  $L^*$  by some function  $L^*h$  of the  $n$  observed losses, and it is critical that our estimate  $E$  should not *underestimate* the true value  $L^*$ . Giving an estimate that is too small is the thing to avoid: in the insurance case it means that our reserves are inadequate more often than once in 200 years, or in the safety application it would mean that the true risk of failure was over the threshold designed in the safety case. (Of course, there are concerns in the other direction — one does not want the reserves to be maintained unnecessarily large, or to design a safety system that is unduly conservative — but in our context these are *secondary* considerations.) So in mathematical terms, if  $F(E) \geq p$  that is all right: the probability in the tail beyond  $E$  is at most  $q$ . But the situation we wish to avoid is  $F(E) < p$ , the situation where there is actually *more* probability in the upper tail than the required value  $q$ .

Of course, since the  $L_i$  are random we are never going to be *certain* that  $E$  does the job, and instead what we must aim for is to have a certain confidence level attaching to  $E$ , say a 95% confidence level. Notice that this confidence level is different from the percentage point  $p$  that we are trying to estimate. We shall call the confidence level  $(1 - \alpha)$  so for a 95% confidence level we take  $\alpha = 0.05$ . Then to say that our  $E$  is a 95% confidence upper bound for the true  $L^*$  is to say that

$$\Pr(E \geq L^*) \geq 1 - \alpha, \quad \Pr(E < L^*) \leq \alpha. \quad (2)$$

In these probabilities, it is  $E$  that is the random variable, since it is a function of the data  $L_i$ . So we are taking a “frequentist” approach, attempting to establish a confidence interval  $[E, \infty)$  for  $L^*$ , not a Bayesian credibility interval.

Does such an estimate  $E$  exist or not? The critical thing is that we do not know the form of the distribution of  $L$ . If we knew the form of the loss distribution, then it may be possible to find  $E$ , and one example would be if losses are uniformly distributed over  $[0, L_{\max}]$  for an unknown value of  $L_{\max}$ . In fact then

$$E_u = p \max(L_i) \alpha^{-1/n} \quad (3)$$

ensures (2), since  $L^* = pL_{\max}$  and so

$$\Pr(E_u < L^*) = \Pr(\max(L_i) < L_{\max} \alpha^{1/n}) = (\alpha^{1/n})^n = \alpha. \quad (4)$$

For instance, in the case  $n = 10$ ,  $q = 0.005$  and  $\alpha = 0.05$  that we have mentioned, the estimate is  $E_u = 1.343 \max(L_i)$ . Equally, if losses are exponentially distributed with density  $\lambda e^{-\lambda x}$  in  $x > 0$  for an unknown parameter  $\lambda$  then we need to choose  $E_e = K \sum L_i$  where  $\Gamma_n(-\log q/K) = \alpha$ , where  $\Gamma_n$  is the cumulative distribution function of the  $\Gamma(n, 1)$  distribution. For the same example data,  $E_e = 0.977 \sum L_i$ .

But in general, if we do *not* know the form of the distribution of  $L$  then there is a simple criterion for whether or not such an estimator  $E$  can be found:

1. If  $p^n \leq \alpha$  then such an estimator  $E$  can be found. Indeed  $E = \max(L_i)$  satisfies (2).
2. If  $p^n > \alpha$  then such an estimator  $E$  cannot be found. In fact for any function  $E(L_1, \dots, L_n)$  there are loss distributions for which (2) fails. These distributions include, for instance, mixtures of two exponentials.

To show the first point here is trivial since if  $E = \max(L_i)$  then

$$\Pr(E < L^*) = \Pr(\text{each } L_i < L^*) = F(L^*)^n = p^n \leq \alpha. \quad (5)$$

The second point is the more interesting result: it essentially says that if  $p^n > \alpha$  then we do not have enough data to estimate the  $p$ -point of the distribution with confidence  $1 - \alpha$ .

The idea that proves this second point can be illustrated by radioactivity. Suppose there is a radioactive element and we want to find the  $p = 0.995$  point of the distribution of its decay times, the time  $T^*$  after which out of 200,000,000 atoms there would only be about

1,000,000 left. We want to do this with  $1 - \alpha = 95\%$  confidence, but we are only given the observed decay times of  $n = 10$  randomly chosen atoms. How do we proceed—in particular, is this the question we discussed before, given that radioactive decay times are exponentially distributed? Well, if there is just *one* isotope, then yes, the analysis above gave the exact answer  $E_e$ . But if there may be two isotopes then we can't estimate  $T^*$  with the required confidence level. For suppose the element consists of one isotope A which is a proportion  $c$  of the total and has short half-life, and another isotope B which has much longer half-life. If  $c < p$  then the  $p$ -point of the distribution is determined predominantly by the half-life of isotope B. But the  $n$  atoms in our random sample will all be isotope A with probability  $c^n$ , so there is at least a probability  $c^n$  that we underestimate  $T^*$ . But if  $p^n > \alpha$  then it is possible that  $p^n > c^n > \alpha$ , and so it is impossible to achieve the required confidence level.

Clearly there is nothing particular about the exponentials here: the same argument would apply if  $L$  is a mixture of any two given distributions with unknown relative scaling, or unknown separation. The crucial point is that if  $c < p$  then there is a probability  $c^n$  that  $L_1, \dots, L_n$  do not explore the tail beyond  $c$ , which is where  $L^*(p)$  is determined; so if  $c^n > \alpha$  then we cannot estimate  $L^*(p)$  with confidence  $1 - \alpha$ .

In thinking about the applicability of this result, it is important that the distributions are not at all unusual—one cannot escape the conclusion by restricting to (say) unimodal distributions. In particular if a mixture of two distributions is regarded as a reasonable distribution to fit to loss data then the result very definitely applies.

Another way of thinking of the result is that it shows that attempts to extrapolate from  $L_1, L_2, \dots, L_n$  out into the tail of the distribution are flawed unless there is a definite known scientific reason for the form of the distribution, and that form is amenable to the kind of analysis given above for the uniform and exponential cases.

Another way of thinking of the result is that it shows that we cannot in general do better than  $\max(L_i)$  to estimate a point in the upper tail. Any estimate that attempts to push further out into the tail, by taking  $\max(L_i) + k_0 + k_1(\max(L_i) - \min(L_i))$  for some positive constants  $k_0, k_1$ , may be better in a particular case, but cannot cope with *all* distributions. There is no scientific reason for choosing a particular value of  $k$  unless we know something quite specific about the form of the distribution.

In fact even if the distribution is say a 2-parameter Pareto with

$$f(x) = \frac{\beta a}{(1 + \beta x)^{a+1}} \text{ in } x > 0, \quad (6)$$

with unknown  $a$  and  $\beta$ , then no estimator  $E$  can be found when  $p^n > \alpha$ .

### 3 The multivariate case

We now move on to the multivariate case, and in fact we shall describe the bivariate case only, since it illustrates all the essential ideas. So, suppose we have  $n$  observed pairs  $(L_i, M_i)$  from a bivariate distribution of losses, and from this data we want to estimate a pair of

thresholds  $L^*$  and  $M^*$  with the property that  $L > L^*$  or  $M > M^*$  is small, in fact

$$\Pr(L > L^* \text{ or } M > M^*) \leq q. \quad (7)$$

To state this in terms of the cumulative distribution function, we want

$$1 - q = p = F(L^*, M^*) = \Pr(L \leq L^* \text{ and } M \leq M^*) = \int_{-\infty}^{L^*} \int_{-\infty}^{M^*} f(x, y) dx dy. \quad (8)$$

As before we shall not be able to do this certainly, all we can aim for is a certain level of confidence in our estimates, so if we let  $E$  be our estimate of  $L^*$  and  $F$  be our estimate of  $M^*$  then we want

$$\Pr(E < L^* \text{ or } F < M^*) < \alpha. \quad (9)$$

The new thing in the multivariate case is that of course  $E$  can depend not only on the  $L_i$  but on the  $M_i$ , and equally  $F$  can depend on the  $L_i$  as well as the  $M_i$ . Following the univariate example we certainly expect that there will be some restriction on whether or not we can produce such estimates  $E$  and  $F$ , and the question is, What will those restrictions be? In fact the answer is just the same: such estimates will exist if  $p^n < \alpha$  but will not (in general) if  $p^n > \alpha$ . If instead of (7) we wanted to have a bound  $L^*$  such that

$$\Pr(aL + bM > L^*) < q, \quad \text{or} \quad \Pr(\max(aL, bM) > L^*) < q, \quad (10)$$

then again the same result applies. These cases might represent situations where we are exposed to different proportions  $a$  and  $b$  of the losses  $L$ ,  $M$ . The underlying reason for each of these results is simply that we can apply the univariate result to the random variable  $aL + bM$ , or  $\max(aL, bM)$ . If the joint distribution of  $(L, M)$  is unknown (or even if it is known but with some unknown scaling parameter like the half-life ratio above) then the distribution of  $aL + bM$  is also unknown and the reasons in the univariate case still apply.

## 4 Copula function estimation

We now turn to the question of estimating the copula function for a pair of variables. So if the variables are  $(L, M)$  as above, and if  $F_L$  is the marginal distribution function of  $L$ , and  $F_M$  of  $M$ , then the variables  $U = F_L(L)$ ,  $V = F_M(M)$  are each uniform on  $[0, 1]$  but they are correlated. The copula function is

$$C(u, v) = \Pr(U \leq u, V \leq v), \quad (11)$$

and the joint density of  $(U, V)$  on the unit square  $0 \leq u \leq 1$ ,  $0 \leq v \leq 1$  is

$$c_{UV}(u, v) = 2C(u, v)uv. \quad (12)$$

So the joint distribution function of the losses  $L$ ,  $M$  is expressed in terms of the marginals and the copula as

$$F_{LM}(l, m) = C(F_L(l), F_M(m)). \quad (13)$$

This equation captures the general aim, or viewpoint, that

... copulas can be used to provide multivariate dependence structure separately from the marginal distributions. [1]

This separation of the individual marginal distributions from the coupling of the two variables together is an attractive aspect of this way of viewing non-independent random variables. However, when we come to estimating  $C$  from some data, say  $(L_i, M_i)$  for  $i = 1, 2, \dots, n$ , the separation is less clear cut and we can make several points about it. There are, in general terms, two possible approaches to this fitting:

1. Joint fitting, where we fit the marginals and the copula jointly to the whole data;
2. Fitting the marginals first and then the copula.

We comment on each of these in turn.

## 4.1 Joint fitting

For the process of joint fitting, suppose we wish to model the marginal distribution of  $L$  by a density function of some particular form  $f_L(l; \alpha)$  depending on a parameter (or set of parameters)  $\alpha$ . The corresponding distribution function will be denoted by  $F_L(l; \alpha)$ . Similarly,  $M$  is modelled by a density function  $f_M(m; \beta)$  and distribution function  $F_M(m; \beta)$ . The copula will be modelled by a function  $C(u, v; \theta)$  with parameter  $\theta$ . The joint density function of  $(L, M)$  then is

$$f_{LM}(l, m; \alpha, \beta, \theta) = (2\ln) C(F_L(l; \alpha), F_M(m; \beta); \theta) \quad (14)$$

$$= c(F_L(l; \alpha), F_M(m; \beta); \theta) f_L(l; \alpha) f_M(m; \beta). \quad (15)$$

Then the maximum likelihood method of estimating the parameters would be to choose  $\alpha$ ,  $\beta$ ,  $\theta$  to maximize

$$\prod_{i=1}^n f_{LM}(L_i, M_i; \alpha, \beta, \theta) = \prod_{i=1}^n c(F_L(L_i; \alpha), F_M(M_i; \beta); \theta) f_L(L_i; \alpha) f_M(M_i; \beta). \quad (16)$$

The maximization of this over  $\alpha$  now clearly depends not only on the  $L_i$ , but also on the  $M_i$ , through the copula factor  $c$ . This is somewhat different from the point of view quoted earlier. In fact the copula function is modelling the interdependence between  $L$  and  $M$ , and so when maximum likelihood estimation uses this model, the model responds by *using* that hypothesised interdependence to give coupling between the observed values of  $M$  and the marginal distribution of  $L$ . (Although we have formulated this explanation in terms of maximum likelihood estimation, it is clear that the same phenomenon will occur if some alternative goodness-of-fit criterion is used.) If the model form of the copula is good, if there is some scientific reason why we know it has a certain form with parameters that we are uncertain of, then it could be right to take this joint estimation approach. But unless we are sure that we really want this cross-coupling between the data for one variable and the estimated marginal distribution of the other, then it may be better to consider the alternative approach.

## 4.2 Fitting marginals, then copula

This approach is more in keeping with the separation of marginals and copula that was referred to above. If the distributions are labelled with the parameters in the preceding section, then the distribution of  $L$  would be determined by choosing the parameter  $\alpha$  to maximize the marginal likelihood

$$\prod_{i=1}^n f_L(L_i; \alpha), \quad (17)$$

and similarly for the choice of  $\beta$  from the  $M_i$ . Then, with those values of  $\alpha$  and  $\beta$  fixed, the copula parameter  $\theta$  are determined to maximize

$$\prod_{i=1}^n c(U_i, V_i; \theta), \quad (18)$$

where  $U_i = F_L(L_i; \alpha)$  and  $V_i = F_M(M_i; \beta)$ . Thus the  $U_i$  are the result of mapping the observed losses  $L_i$  into the unit interval by their estimated distribution function, and  $V_i$  similarly for the losses  $M_i$ .

There are two points we may make about this:

1. The  $U_i$  are not directly observed data: they already depend on first fitting a distribution to the actual observed losses  $L_i$ . They will occur in the same order as the  $L_i$ , since the distribution function is increasing, but they are best viewed as random variables. If the distribution parameters are obtained exactly, then the  $U_i$  have a joint density that is constant on the region  $0 < u_1 < u_2 < \dots < u_n < 1$  if we choose that particular ordering. The mode of  $U_i$  is  $(i - 1)/(n - 1)$  and the mean of  $U_i$  is  $i/(n + 1)$ . The “empirical distribution” determined by the  $L_i$  effectively takes  $U_i$  to be  $(2i - 1)/(2n)$ , intermediate between these values.
2. If the wrong marginal distribution is fitted, then this will introduce bias in the fitting of the copula function. A simple example illustrating how this could happen is provided by the data in Figure 1. This data was in fact generated by taking the two variables independent, and each uniformly distributed over  $[100, 400]$ . However, suppose a fitting process assumed that each variable had a uniform distribution over  $[0, L_{\max}]$  and estimated  $L_{\max}$  as 400. Then the scaled points will predominantly lie towards the north east corner of the unit square, and so a copula function will fit a positive correlation between the variables. In effect, the copula function will try to compensate for any inaccuracy in fitting the marginal distributions.

Neither of these is easy to deal with, but they do bring out the fact that fitting copula functions to data has several points that one should be aware of that make it more complicated than fitting a univariate distribution to data.

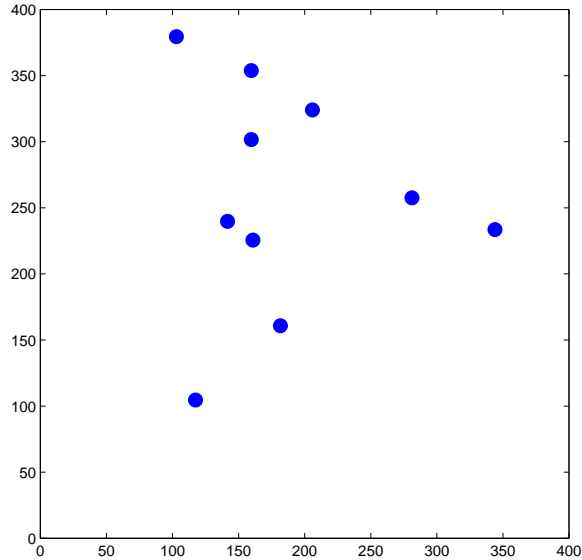


Figure 1: Sample of bivariate data.

### 4.3 Fitting a joint distribution

It is perfectly possible to fit a multivariate distribution to data, and this should at least be considered as an alternative to the marginals-and-copula approach. For instance, the Gaussian (normal) and lognormal distributions have well-known multivariate forms, and distributions like the hyperexponential and Pareto can be generalized easily to give them corresponding multivariate forms. For instance, a multivariate version of a  $k$ -component hyperexponential distribution would have density  $\sum_{i=1}^k p_i w_i \exp(-a_i \cdot x)$  where the  $p_i$  are probabilities,  $a_i$  are vectors with positive entries, and  $w_i$  is the product of the entries of  $a_i$ . Equally a multivariate analogue of the Pareto distribution would have density  $\sum_{i=1}^k p_i w_i' / (1 + a_i \cdot x)^{\alpha_i}$ .

To illustrate this approach, we used a set of pseudodata that is intended to provide representative values of insurance losses for 8 different lines of business over a year. Data were provided for 1000 independent draws from this distribution, so this could be taken as representing data from 100 independent companies over a 10-year period. In practice of course one will have much less data than this, and it is unlikely to be realistic to treat different companies in the same year as having independent losses. The first few lines of the data set are given in Table 1.

If we fit a bivariate lognormal distribution to the Property and Motor data then we obtain a distribution with contours as shown in Figure 2. In fact a goodness of fit test shows that this provides a perfectly acceptable fit to this data.

Casualty	Casualty Treaty	Property	Property Treaty	Marine	Motor	Energy	Aviation
121	95	92	70	34	155	77	43
240	152	163	86	48	195	128	68
123	70	87	70	49	124	59	62
110	55	100	74	55	140	69	40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: First few lines of table of pseudo data for annual losses from 8 lines of business.

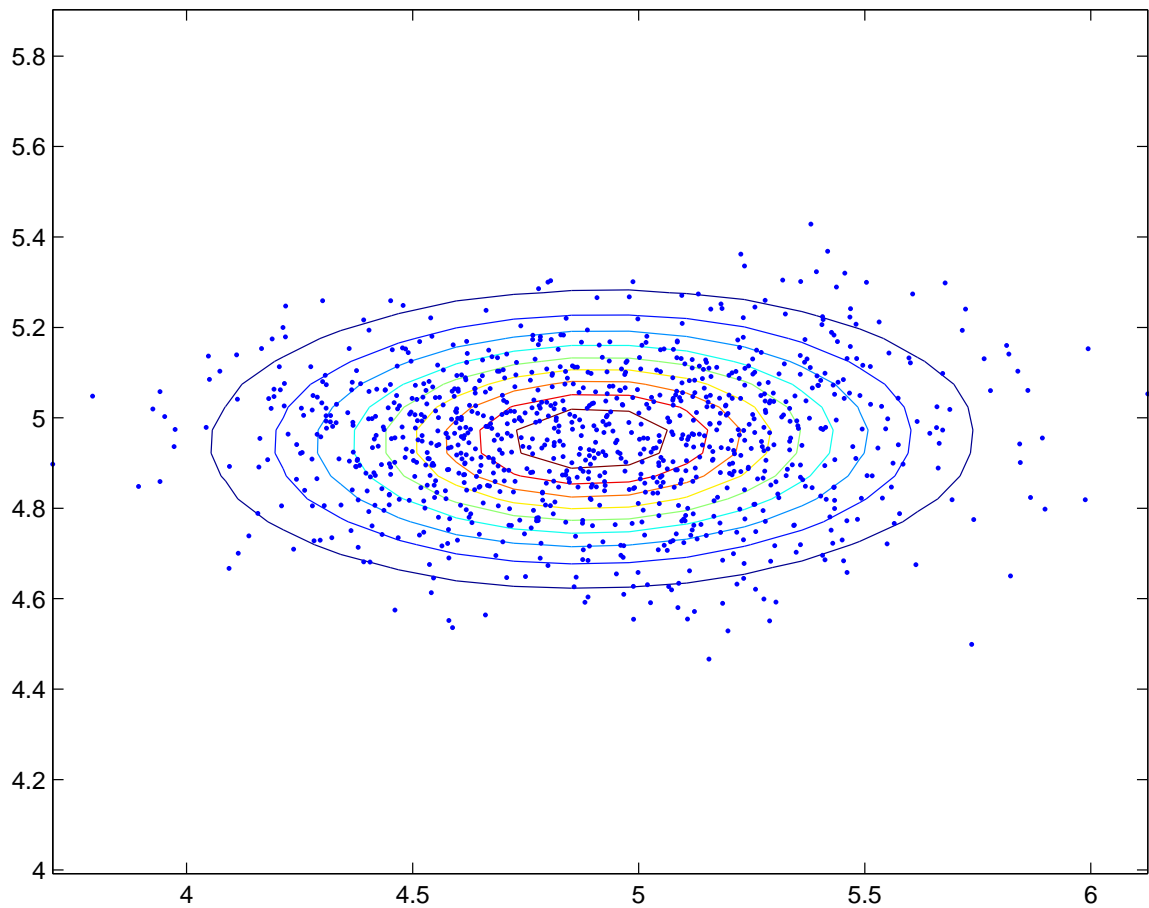


Figure 2: Fitting a bivariate lognormal distribution to the Property and Motor pseudo data. The contours contain 0.1, 0.2,  $\dots$ , 0.9 of the probability.

## 5 Fitting copula functions to the pseudodata

In fitting copula functions to the pseudodata, we shall here use the empirical distribution function as described earlier. So if  $L_1, L_2, \dots, L_n$  are the observed losses, then we arrange them in order

$$L_{i(1)} < L_{i(2)} < \dots < L_{i(n)}, \quad (19)$$

and take the variables in the unit interval  $U_i$  to be

$$U_{i(1)} = \frac{1}{2n}, \quad U_{i(2)} = \frac{3}{2n}, \quad \dots, \quad U_{i(n)} = \frac{2n-1}{2n}. \quad (20)$$

The kind of result arising from maximum likelihood estimation is illustrated with an example, where we have used the FGM copula function

$$C(u, v; \theta) = uv(1 + \theta(1-u)(1-v)), \quad (21)$$

where the parameter  $\theta$  has to lie between  $-1$  and  $1$ . If we have data values  $(u_i, v_i)$  (obtained from the empirical distribution function as above) then the log-likelihood function is

$$\log L(\theta) = \sum_{i=1}^n \log C(u_i, v_i; \theta). \quad (22)$$

The maximum likelihood estimate of  $\theta$  is the value  $\theta_m$  such that

$$\log L(\theta_m) = \max\{\log L(\theta) : -1 \leq \theta \leq 1\}, \quad (23)$$

and a confidence interval  $[\theta_-, \theta_+]$  around that can be estimated by

$$\log L(\theta_{\pm}) = \log L(\theta_m) - A^2/2 \quad (24)$$

where  $A$  is chosen as the normal deviate giving a 2-sided tail probability of  $\alpha$ . (For instance,  $A = 1.96$  if  $\alpha = 0.05$ .) Figure 3 now shows the maximum likelihood and confidence interval results, for  $n$  varying from 10 to 1000 (at a sequence of approximately geometrically spaced steps, so the horizontal axis has been given a logarithmic scale). It is clear that the confidence interval does not narrow very rapidly as  $n$  increases, even though we have effectively given it some artificial narrowing by assuming that the  $u_i$  and  $v_i$  are known instead of random. We can go further and consider this parameter estimation for each pair of variables in the table, and we obtain the data shown in Figure 4. It is clear that in most cases the interval does not narrow rapidly. In fact for  $n = 100$  most of the confidence intervals have narrowed from the initial range  $[-1, 1]$  to an interval of length about 1. The only cases where a decisive trend is visible are for

- Casualty and Casualty Treaty;
- Property and Property Treaty;
- Marine and Energy.

A fundamental point is the dimensionality of course: 10 points distributed over a square tell us less about a function over that square than 10 points in the corresponding problem along a line. And when we go to higher dimensionality, the problem is compounded.

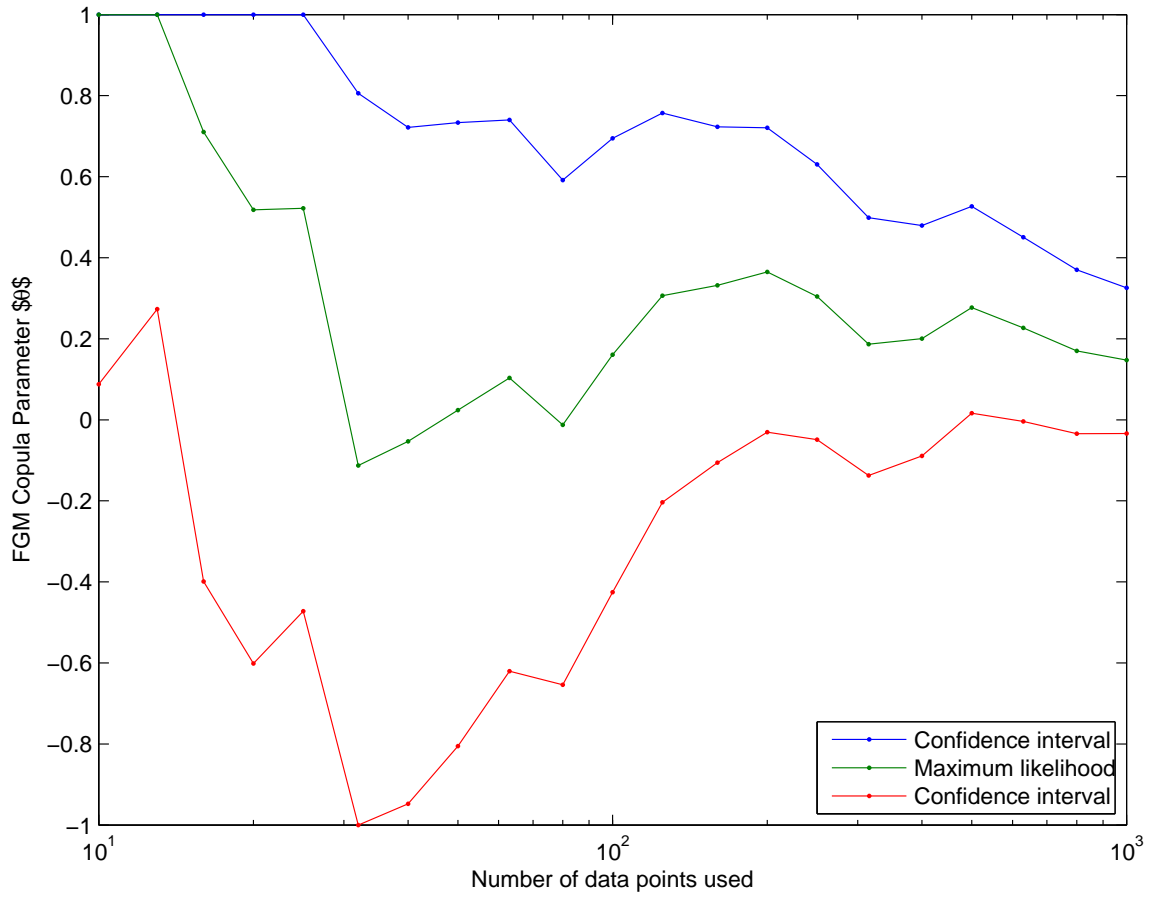


Figure 3: Maximum likelihood value of  $\theta$  (middle plot) and upper and lower ends of 95%-confidence interval, for estimating the parameter  $\theta$  in the FGM copula from  $n$  data points. (Property and Motor pseudo data).

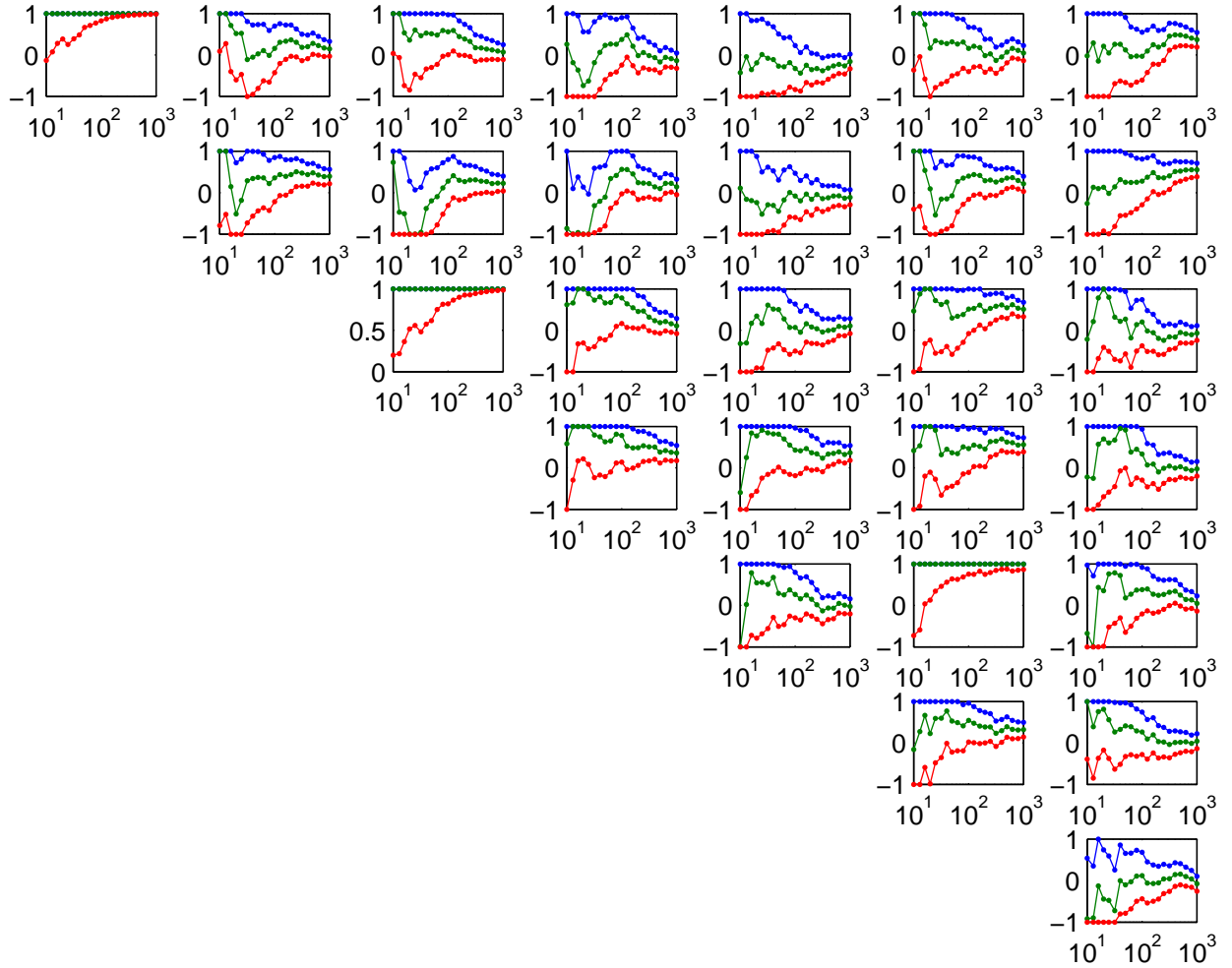


Figure 4: Maximum likelihood values of  $\theta$  and confidence intervals for each pair of variables in the table of pseudodata.

## 6 Further questions

There are clearly many questions that we have left unanswered here, but among them we draw attention to two:

- If the marginal distributions are unknown and we observe bivariate data  $(L_i, M_i)$  for  $i = 1, 2, \dots, n$ , what is the maximum likelihood estimate of the copula function? This will only depend on the ordering of the  $L_i$  relative to the  $M_i$ , but it is quite a basic question that one would like to know the answer to.
- What are the fundamental limits imposed by data size on copula estimation? Here the interest would be in obtaining a result corresponding to that earlier about estimating percentage points in the tail of an unknown distribution. Is there a sharp result analogous to the criterion “ $p^n < \alpha$ ” that determines how much data we need in order to locate a given point of a copula function with a given confidence level?

## 7 Conclusions

Clearly limited data presents a serious problem in estimating percentage points in the tails of a distribution, and in estimating copula functions. Nevertheless, one has to proceed somehow and so, in addition to being fully aware of the difficulties, one needs various approaches (which might well be combined in practice).

One of these approaches is to use as much additional information and expert knowledge as possible in addition to the raw loss data. For instance, under this heading would come studying the effects of climate change, earthquake risk, the influence of changing regulatory structures, and so on, and particularly studying the effects of these on the risk of extreme events, and on the interdependencies between different risks. These kind of models would then have to be used to give further information on the tails of the distributions of interest. It must always be borne in mind though that the data in the bulk of the distribution does not force the form of the tail. The tail may be governed by quite different effects from the bulk—effects about which the data contains little information.

A second approach is to get more detailed data, in particular disaggregated data. For instance, a year’s loss is the sum of 12 monthly losses and if those monthly losses can be treated as independent and the individual data obtained, then in some cases there may be scope for this to lead to narrowed confidence intervals.

## References

- [1] Enjoy the joy of copulas. Jun Yan. University of Iowa.  
<http://www.stat.uiowa.edu/techrep/tr365.pdf>