The Actuarial Profession
making financial sense of the future

GIRO Conference and Exhibition 2010
Ajay Chhabra (Aspen) and Pietro Parodi (Willis)

Dealing with sparse data

Practical challenges and techniques

12-15 October 2010

Working Party Introduction

- Background to Working Party

- What are we hoping to achieve?

- Working Party:
  - Ajay Chhabra (Chair)
  - Pietro Parodi
  - Tom Day

- Acknowledgements:
  - David Menezes
  - Isobel Prowen
  - Joseph Lo

Important Note:
The ideas discussed in this presentation are those of the presenters, and are not necessarily reflective of views or practices of their respective employers.

Overview

I. Sparse data – setting the scene

II. Techniques to deal with sparsity

III. Case Studies
   a. Loss Severity Parameterisation
   b. Elastic Net Regularisation
   c. Copula Calibration
   d. Reverse Stress Testing

IV. Assessment of probabilities

V. Conclusions

I. Sparse data - setting the scene

Why investigate sparsity?

Sparsity affects us all

All non-life actuaries have to deal with sparsity: data are systematically pushed to the limit of sparsity

Sparsity represents a commercial opportunity

Data-rich problems seldom lead to massive profits for insurers…

Sparsity makes us actuaries valuable

Actuaries are valued for their judgement and business knowledge, not for their grasp of statistics – which is probably grasped more firmly by other professionals

How does data sparsity arise?

When the risk is new	Historical data have just started being collected
When the risk changes with time	Historical data become quickly irrelevant
When we are "in the tail"	Especially relevant in reinsurance, commercial insurance, capital modelling
When we have many dimensions	When many dimensions are involved, there arises the "dimensionality curse", by which an exponentially larger data set is required to achieve the same level of accuracy
When data is just not there	There could be many reasons for this…

What does sparsity entail?

Parameter uncertainty	The accuracy of the parameters degrade as the number of data points decreases, e.g. the error on the mean is $\propto 1/\sqrt{n}$
Model uncertainty	Our ability to discriminate between models decreases as the number of data points decreases
Bias	E.g. goodness of fit tests are biased, MLE is biased, etc
Breakdown of most traditional quantitative actuarial methods	Many of the traditional quantitative methods used by actuaries, e.g. reserving methods, claims inflation estimation methods, etc, break down when data are too sparse

Across-the-board examples of sparsity

Pricing

- Severity model selection with few data points
- Rating factor selection with many factors and few data points
- Reinsurance pricing

Reserving

- Lack of historical experience for a class of business
- Sparse claims triangles, e.g. number of claims above a high threshold in reinsurance

Capital Modelling / Risk Management

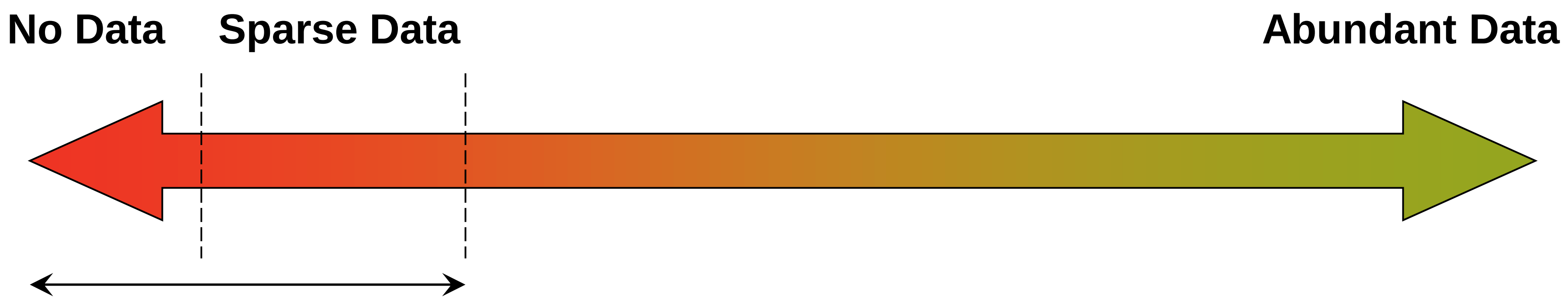
- Capital requirements at high percentiles (e.g. 99.5%)
- Dependency modelling (e.g. copula calibration)
- Stress/scenario testing

II. Techniques to deal with sparsity

How can we deal with sparsity?

- “Expert Judgement”
- Benchmarks / collateral data
- Extreme Value Theory
- Statistical Learning Theory
- …

- GLMs / Multivariate Regression
- Principal Component Analysis
- Maximum likelihood estimation
- Curve-fitting software
- “Expert Judgement”



No Data **Sparse Data** **Abundant Data**

There is no Holy Grail!

Role of Expert Judgement

Modelling = Data + Knowledge

- Where data are sparse, a pure frequentist approach doesn't work
- We are all (consciously or otherwise) Empirical Bayesians!
  - Prior: “Expert Judgement” and market data
  - Posterior: Informed by prior and existing data

There is no substitute for Expert Judgement

- **Key challenges:**
  - How can we trust judgement?
  - How can we explain and communicate judgement?

Large losses – Extreme Value Theory

The tail won’t probably fit

Most distribution fits break down for large losses and there’s a need to model the tail separately. However, the tail is almost by definition sparse.

The tail is always a GPD… no model uncertainty then!

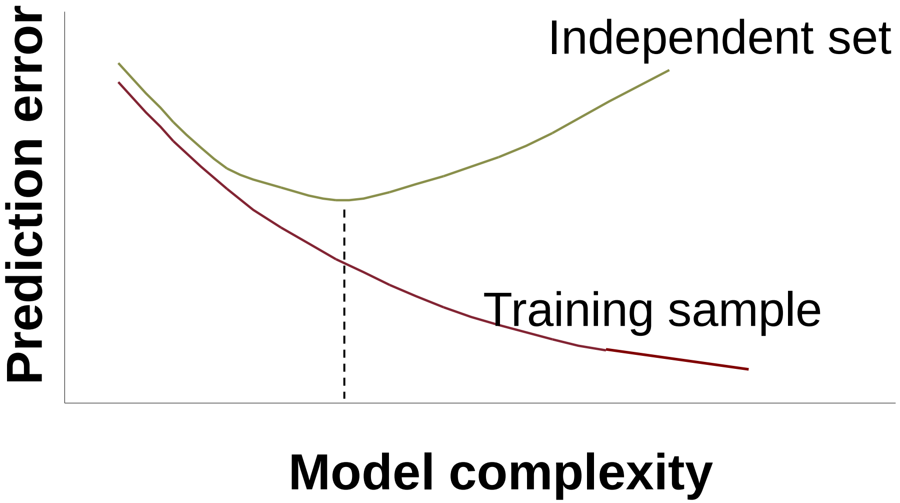
The main results of extreme value theory is that all distributions behave *asymptotically* as a Generalised Pareto Distribution.

But sparsity affects extreme value theory, too

- Parameter uncertainty creates a **prediction horizon**
- There is a significant bias on the “shape” parameter → tail classification becomes a problem, underestimation is likely

Statistical learning theory
A rigorous framework for model selection

Modelling is the ultimate *ill-posed problem:* infinitely many models can be fitted to the same data, and a solution is only possible by imposing constraints (e.g. simplicity)



Structural risk = Dist ($\mathcal{M}$, $\mathcal{D}$) + Penalty(d,n)

($\mathcal{M}$ = model, $\mathcal{D}$ = data**)**

Examples

AIC = - 2 log L + 2 d

BIC = - 2 log L + log(n) d

MDL = **BIC**

Statistical learning theory

Sparsity-based regularisation schemes

An alternative idea for model selection

Instead of minimising just the distance between the model and the data, minimise a regularised distance, which includes a penalty term on the parameters β, e.g.

$$\text{ExpectedPredictionError} = \left\| Y - f_\beta(X) \right\|_{l_2}^2 + \lambda \left\| \beta \right\|_{l_2}^2$$

Enforcing variable sparsity, and dealing with data sparsity

Some regularisation schemes also perform variable selection automatically, keeping the complexity at bay; and some are devised to work with many variables (e.g. rating factors) but very few data points

Maximum Entropy

Provides a means of imposing prior constraints on the form of a model, whilst maximising ‘uninformedness’ (“entropy”) of the rest of the model.

Entropy for a continuous distribution with density *p*(*x*):

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx$$

Examples:

Constraints	**Maximum Entropy Distribution**
Upper / Lower Bound: [a, b]	U(a,b)
Mean (μ) / SD (σ)	N(μ, σ^2)
Mean (μ)	Exp(1/ μ)

Maximum Entropy
Past and future applications

- Used extensively in the field of inference
  - Bayesian techniques often use ‘entropic priors’, which is a tightening of the classic ‘uninformed priors’

- EMB have made use of the maximisation of ‘relative entropy’ in their Economic Scenario Generator
  - Allows for user-imposed constraints on the statistical properties of the modelled economic scenarios

- Worthwhile area for further research – likely to be useful for curve-fitting, and could be used to address the parameter uncertainty in the tail of the GPD

III. Case study

a. Severity model selection

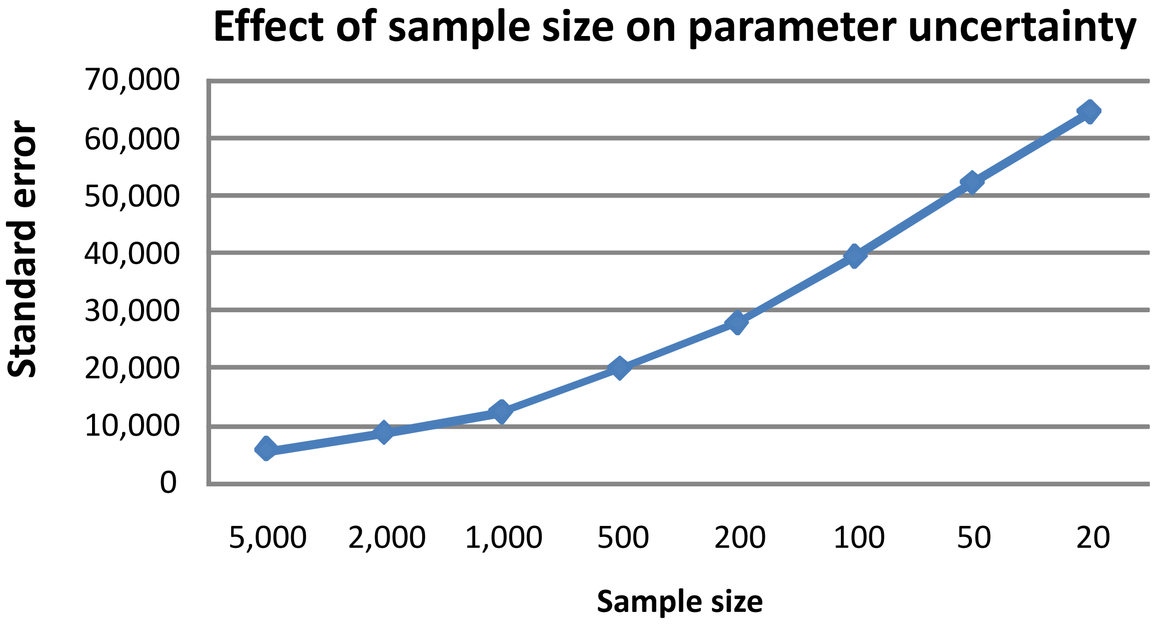
The problem: Determine the severity model and its parameters for a given data set

As the problem becomes sparser, at least three things happen:

1. parameter uncertainty increases…
2. … the mean is underestimated in the majority of samples…
3. … and model uncertainty becomes critical

Sparsity causes parameter uncertainty

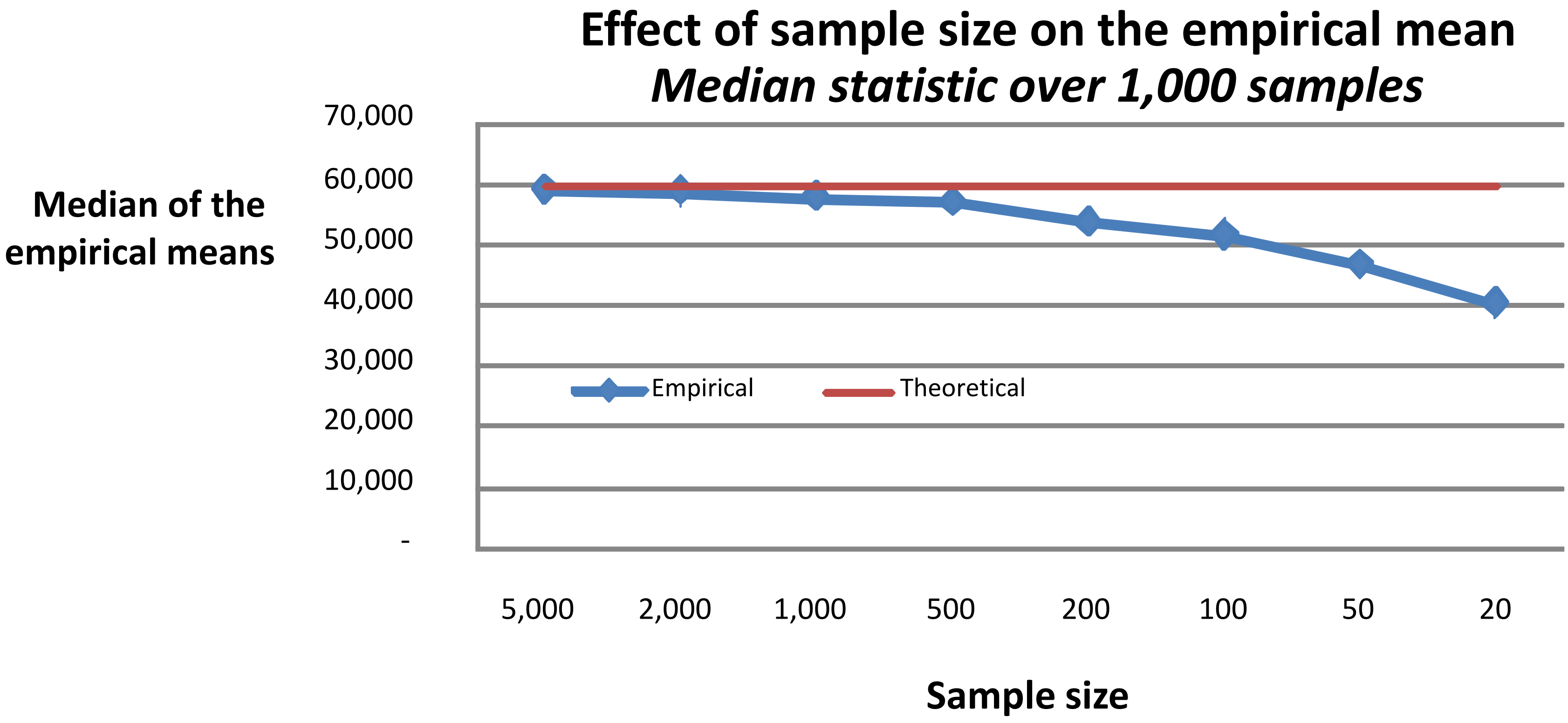
The standard error on the parameters increases as dataset size decreases. It eventually drowns the estimate itself…



- Artificially generated data from a Lognormal distribution with $\mu=9$, $\sigma=2$
- Theoretical mean is £59,876
- 1,000 samples for each selected sample size

Sparsity causes parameter bias

Bias = systematic downward (upward) error in parameter estimation



The theoretical average for LogN(9,2) is £59,876

Sparsity causes model uncertainty

		AIC	Least Squares	KS	Kuiper	AD
N=1,000	Lognormal ranks…	#1	#2	#2	#2	#1
	… and the winner is:	LogN	Beta tr.	Beta tr.	Beta tr.	LogN
	# of parameters:	2	4	4	4	2
N=100	Lognormal ranks…	#5	#9	#9	#9	#10
	… and the winner is:	Pareto	Burr	Burr	Burr	Loglog
	# of parameters:	2	3	3	3	2
N=20	Lognormal ranks…	#4	#9	#9	#10	#5
	… and the winner is:	Plog inv	Beta tr	Beta tr	Beta tr	Gam tr
	# of parameters:	2	4	4	4	3

Results obtained with Risk Explorer, with 22 distributions to pick from, using MLE

Yes, but… does it really matter?

	LogN(9,2)	*LogN(8.86,1.78)*	*Beta transf*	*Burr*
Mean	60,945	34,722	16,448	244,294
95%	218,750	131,815	78,649	138,859
99.50%	1,439,375	701,616	78,649	1,891,458
99.90%	3,941,002	1,715,139	78,649	13,883,349

Curve fits based on N=20

It does.

Lessons learned

- Unconstrained distribution fitting to select the best distribution just doesn’t work
  - Models with more parameters will be chosen more easily
- One must also be aware of the bias introduced…
- …and of course of the parameter uncertainty

Solutions? [1-2]

1. Restrict the number of admissible models

- Use experience or theoretical reasons, e.g. EVT
- Exclude distributions with undesirable properties
- Only distributions with some rationale?
- Even choosing something because it is traditional is better than scatter-gunning!

2. Punish model complexity

- Statistical learning theory: **any model more complex than necessary makes poorer predictions**
- AIC works better because it punishes complexity:

$$-2\log\Pr(\text{data})+2\times\text{no\_of\_parameters}$$

Solutions? [3]

3. Use a Bayesian approach to model selection

i. Assign a prior probability to each model, $\mathbf{Pr}(\mathcal{M}_j)$

ii. Select the model with the largest posterior distribution given the data:

$$\mathcal{M}^* = \operatorname{argmax} \mathbf{Pr}(\mathcal{M}_j \mid D) = \operatorname{argmax} \mathbf{Pr}(D \mid \mathcal{M}_j)\, \mathbf{Pr}(\mathcal{M}_j)$$

How to estimate the prior probabilities?

- Possible empirical approach: consider models effective on larger clients: e.g. 70% of large clients used LogN, 30% used Burr
- Another possible approach: use market analysis, and use $\mathbf{Pr}(\mathcal{M}_j \mid D_{\text{mkt}})$ as the prior

Solutions? [4-5]

4. Add the bias back	The bias can be estimated by simulating from the fitted distribution. By adding the bias back, the solution now won't maximise likelihood anymore but will be unbiased… one can't have it both ways This is especially important for percentiles
5. Validate against an independent set	• In statistical learning theory, there is *always* a training set (used for parameter estimation), and a *test set* used for selection and validation • This will call the complex models' bluff!

Severity models revisited

Method 1: Use a restricted set and punish complexity

1. Exclude inverse distributions, threshold distributions, distributions with negative values such as Normal
2. Rank distributions according to AIC

Distr	-log Pr(D|M)	AIC	No of param
LogN	215.12	219.12	2
Burr	213.36	219.36	3
Weibull	216.71	220.71	2
Gamma	218.77	222.77	2
Loglogistic	219.01	225.01	3
Frechet	225.05	229.05	2
Paralogistic	229.7	233.7	2
Gumbel	244.54	248.54	2

Note: LogN is ranked #2 in this restricted set according to LS,KS,AD, Kuiper

Severity models revisited

Method 2: Use a Bayesian approach to model uncertainty

1. E.g. $\Pr(\mathcal{M}_j)$ = % of *large* clients for which $\mathcal{M}_j$ is the best model

2. Choose the model with the largest posterior $\Pr(\mathcal{M}_j|D)$

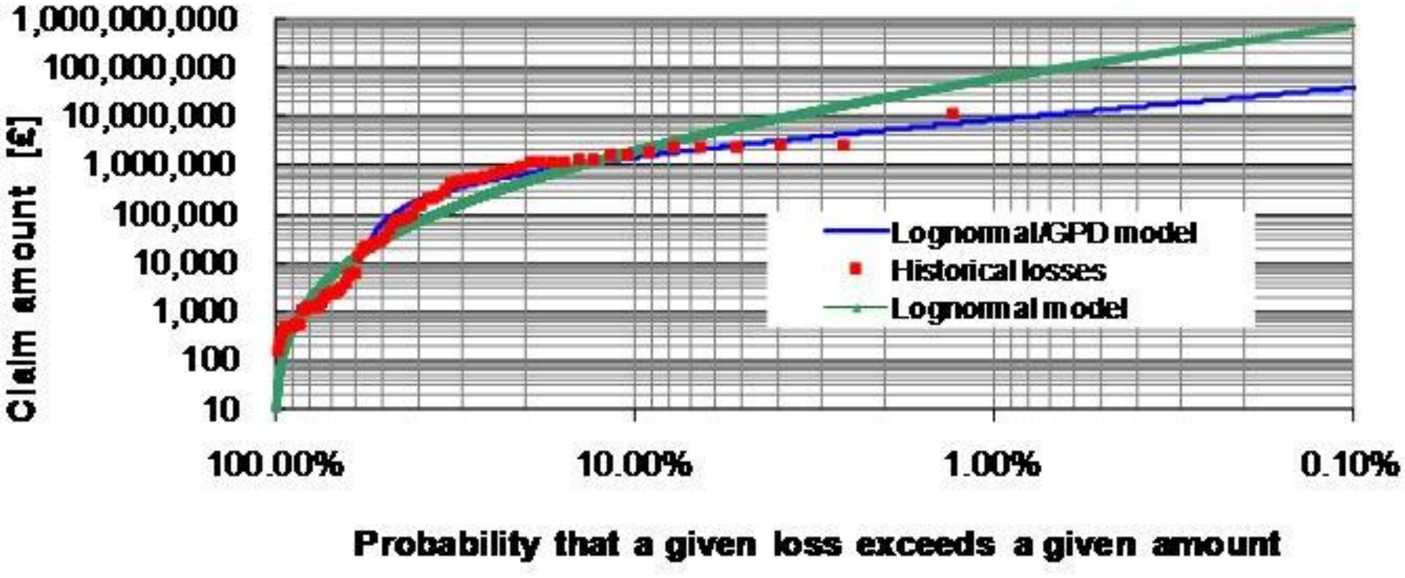
Distr	$-\log\Pr(D\mid\mathcal{M})$	$\Pr(\mathcal{M})$	$\Pr(\mathcal{M}\mid D)$
LogN	215.12	60%	49.7%
Burr	213.36	10%	48.2%
Weibull	216.71	10%	1.7%
Gamma	218.77	20%	0.4%
Loglogistic	219.01	0%	0.0%
Frechet	225.05	0%	0.0%
Paralogistic	229.7	0%	0.0%
Gumbel	244.54	0%	0.0%

3. Correct parameters for the bias (if applicable)

What if we are modelling the tail?
Extreme value theory to the rescue

Pickands-Balhema-deHaan theorem:

The tail of any distribution can be modelled as a Generalised Pareto distribution.



⇒ This solves the model uncertainty problem!

Sparsity-related problems with EVT

- What is the tail, actually? Some residual model uncertainty…
- You still have parameter uncertainty, which leads to a *prediction horizon*
- There is a strong bias on the parameters, which may lead to behaviour-switching (e.g. from power law to exponential to finite support)

III. Case studies

b. Rating factors selection

Rating factors selection and regularisation

No. of data points (N) vs no. of rating factors (p)

Rating factor selection suffers from the dimensionality curse

As the number of possible rating factors increase, the number of data points needed to have a stable variable selection process increases exponentially ($N \sim c^{p}$).

From multi-way analysis to regularisation

GLM represents great progress to tackle sparsity with respect to multi-way analysis. However, it still assumes that $N >> p$. There exist techniques, as elastic net regularisation (Zou & Hastie, 2005), which address the situation where $p >> N$.

Want to know more? Talk C6 on Thursday: "Regularisation: An efficient and simple approach to rating factors selection"

III. Case studies

c. Copula calibration

Copulas in Dependence Modelling

- Copulas are increasingly in use by non-life insurers as a tool to model non-linear correlations and particularly tail dependence

- Purely statistical approach to dependencies

- Copula family needs to be carefully selected by consideration of desirable features:
  - Tail-dependency
  - Symmetry / asymmetry
  - Simplicity

- Given the form of the copula, parameters need to be selected:
  - Maximum likelihood and other statistical approaches
  - “Judgement”

Copula Example: Sample Data Set

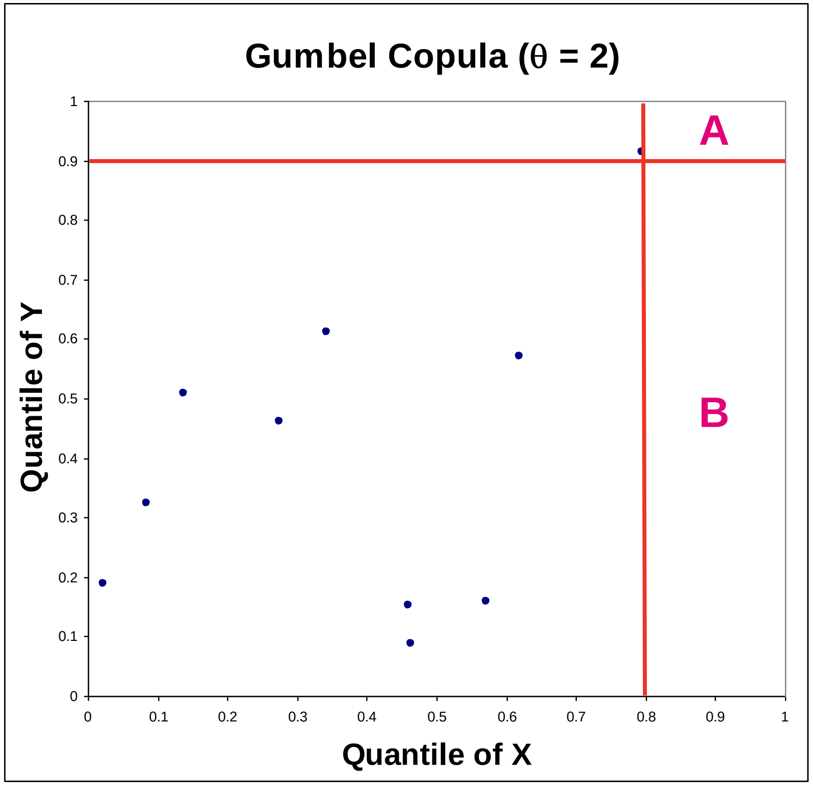
- Simulation of 10 observations from Gumbel Copula, with θ = 2

- Should exhibit heavy dependence in the upper tail.

- Require estimate of:
  P(Y > 90th %ile | X > 80th %ile) – i.e. P(A | A ∪ B)

- **Model Sample (*n* = 10):**
  Kendall τ = 0.2 ⇒ θ = 1.25
  P(Y > 90th %ile | X > 80th %ile) = **25%**

- **Actual Model**
  θ = 2 ⇒ Kendall τ = 0.5
  P(Y > 90th %ile | X > 80th %ile) = **42%**



Why has it gone so wrong?

- Kendall τ correlation coefficient for Gumbel copula is driven by dependence in the tail

- Kendall τ grossly understated, due to insufficient tail observations

- Estimate of the only parameter, θ, is driven by estimate of τ

Without a large sample of observations, relying solely on statistical methods fitted to data does not work in the tails of joint distributions!

Is there a place for copula modelling in non-life insurance?

- Modelling of dependence using ‘structural’ correlations (i.e. causal drivers) should be used to the greatest extent possible, within bounds of pragmatism and explainability

- Copulas remain useful tools for modelling residual non-linear dependency, which may not readily be captured by structural correlations alone

- Key is to calibrate based on judgemental estimates of tail dependence

- Must be able to communicate clearly the implications of the judgements employed

  - ***“Our capital model assumes that the probability of a global fall in stock markets of more than 35% is x%, and the probability that this is combined with our D&O book giving rise to losses exceeding $400m is y%.”***

Expert Judgement applied to Bayesian Problems

- We have established that there is no substitute for expert judgement, and we are all “Empirical Bayesians”

- Bayesian problems themselves can be assisted by prior judgement, particularly those which require probability estimates of extreme events
  - e.g. Curve fitting in the tail: “what does ‘1-in-100’ look like?”
  - e.g. (Reverse) stress / scenario testing: what is the probability of a particular combination of adverse events?

- Will illustrate that judgement can be more than just a “finger in the air”, and the thought process can be made transparent by explicit Bayesian formulation

III. Expert Judgement

d. Case Study: (Reverse) Stress Testing

Stress & Scenario Testing (SST)

- Challenge is to quantify both the probability and severity of extreme adverse events

- Key tool for risk management
  - Inform capital model design or parameterisation
  - Validate outputs of capital model
  - Identify unmodelled risks, which can then be managed

- Not just a qualitative exercise!

- Mathematically, the problem reduces to the estimation of joint and conditional probabilities of adverse quantifiable events

- For this, Bayes' Theorem is the key

Bayes' Theorem

- P(A, B) = P(A | B) P(B)
  = P(B | A) P(A) (by symmetry)
- Invaluable because estimation of joint probabilities is **extremely** difficult
- We generally do not have a good intuition for joint probabilities, except in the trivial case of independence: P(A, B) = P (A) P(B)
- We are much better at estimating **conditional probabilities**, though these too can give rise to paradoxes of intuition, if not framed correctly
- Bayes' Theorem decomposes the problem of estimating a highly unknown quantity into a problem of estimating better understood quantities
- It will always be necessary to make subjective probability assessments, but we can at least make the problem simpler!

Reverse Stress Test Example

- Required to think hard about risks to your company, including those that are reasonably foreseeable, and may cause your business model to fail, and estimate the likelihood of it occurring
  (See GIRO Workshop E5: Reverse Stress Testing)

- Example Reverse Stress Test:
  - ***In a particular year, there is an active hurricane season, which produces 2 severe US windstorm events of Katrina/Rita/Wilma magnitude. To make things worse, an earthquake of Richter 9.0 strikes in California causing widespread losses across across both property and liability classes of business. Concerns relating to the impact of the natural disasters on US economic output lead to a sharp fall in the dollar relative to other major currencies. A number of reinsurers default, causing an increase in the level of net losses.***

- How do you estimate the probability of this?

Reverse Stress Test Example

- Events:
  - Windstorm 1 (A)
  - Windstorm 2 (B)
  - California Earthquake (C)
  - Fall in dollar (D)
  - Reinsurers defaulting (E)

- We are required to estimate:

P(A, B, C, D, E)

$$
\begin{aligned}
&= P(E \mid A, B, C, D)\, P(A, B, C, D) && \text{(Bayes)} \\
&= P(E \mid A, B, C)\, P(A, B, C, D) && \text{(Conditional Independence)} \\
&= P(E \mid A, B, C)\, P(D \mid A, B, C)\, P(A, B, C) && \text{(Bayes)} \\
&= P(E \mid A, B, C)\, P(D \mid A, B, C)\, P(A, B)\, P(C) && \text{(Independence)} \\
&= P(E \mid A, B, C)\, P(D \mid A, B, C)\, P(B \mid A)\, P(A)\, P(C) && \text{(Bayes)}
\end{aligned}
$$

- Much more intuitive now!

Reverse Stress Test Example

- Notice how a judgement of ‘causality’ has been used:
  1. To decide that Californian Earthquakes happen independently of windstorms
  2. To decide that reinsurers defaulting and the fall of the dollar are conditionally independent, given the occurrence of the natural catastrophes

- Bayesian networks are directed acyclical graphs to describe prior judgements relating to ‘causality’ and conditional probability, and are a powerful formalisation of the thought processes described above

- By decomposing the joint event into its constituents, we not only gain transparency, but may be able to place useful upper and lower bounds on the overall probability by considering bounds of the constituent parts

Reverse Stress Test Example

- Events:

Windstorm 1	(A)
Windstorm 2	(B)
California Earthquake	(C)
Fall in dollar	(D)
Reinsurers defaulting	(E)

P(A)	20%–30%
P(B | A)	50%–80%
P(C)	2%–5%
P(D | A, B, C)	60%–80%
P(E | A, B, C)	60%–80%

- Return Period: 1-in-130 to 1-in-1300(!)

- Bounds are not tight, but may not be completely useless
  - Capital model with 50,000 modelled scenarios should give between 60 and 480 scenarios of this sort. How does that stack up?

Reverse Stress Test Example

P(A)	20%–30%
P(B | A)	50%–80%
P(C)	2%–5%
P(D | A, B, C)	60%–80%
P(E | A, B, C)	60%–80%

- Return Period: 1-in-130 to 1-in-1300

- Which assumption is most sensitive?
  - …turns out to be one of the “marginal” rather than “conditional” probabilities
  - Counterintuitive result?
  - Helps focus attention on risk and probability assessment

- This exercise highlights the extent of subjectivity in assessing extreme probabilities, but the Bayesian approach helps in formulating and exposing the thought processes behind the judgemental assumptions

IV. Assessment of Probabilities

Judgemental Assessments of Probability

- Humans are notoriously bad at assessing the probability of events, particularly joint events.

- Interpreting probability is a non-trivial challenge:

  - “The probability of a fair coin, when tossed, coming up heads 3 times in a row is 12.5%”

  - “The probability of England winning the World Cup at some point in the next 40 years is 12.5%”

- There is a difference between the **purely frequentist** view of probability and the **purely subjective** view of probability, which describes the level of confidence in an uncertain occurrence

Subjective Probability

- As subjective probabilities express degrees of belief, they are highly judgemental, and will give range to a wide range of estimates
- Where multiple experts exist, all of whom have at least some degree of credibility, polling may be able to produce a better estimate (“Wisdom of Crowds”)
- It is at least as important for the judgement to be understood and communicated in a transparent manner, so that it can be challenged
    - In the end, a model is just a formalisation of a set of assumptions and beliefs…though it does give us useful information about the logical *consequences* of our beliefs

Eliciting Probability Judgements (1)

Determining Probabilities Directly

- While probability assessments are inherently subjective, there are plenty of ways we can avoid common pitfalls to elicit judgement by asking questions in the right way:

- **Return Period pitfalls**
  - Avoid asking for return periods directly (“What is your ‘1-in-100’ loss?”)
  - Common interpretation as ‘1-in-100 years’ is severely flawed
  - If we have to use a time-oriented concept, consider: “How many events per year?” before asking the “how many years?” question

- **Comparison with ‘everyday’ probabilities**
  - “Is it more or less likely than rolling a 7 from a pair of dice 3 goes running?” (i.e. ~“1-in-200”)

Eliciting Probability Judgements (2)

Assessing conditional relationships

- **Assessing independence**
  - Poorly phrased questions
    "Is A likely to happen independently of B?" "Is A independent of B?"
    "Is B caused by A?"
  - Better question:
    "If you knew about whether A would happen or not, would it change your view of the likelihood/riskiness of B?"

- **Assessing conditional probabilities**
  - Bayes' Theorem lets us elicit conditional probabilities in either direction
  - Which is easier to estimate:

P(D&O Loss Ratio exceeds plan | Stock Market Crash)
or
P(Market crash | D&O Loss Ratio exceeds plan)?

  - Where causal relationship is known, "causal" direction is **much** more intuitive!

Eliciting Probability Judgements

Assessing conditional relationships

- **Erroneous interpretation of ‘temporal dependence’:**
  - Lack of understanding of “given”
  - Poorly phrased question:
    “What is the chance of A happening given that B has happened?”
  - Better question:
    “If, at some point in the future, you were to know that B were the case, what do you think would be the chance that A would at the same time also be the case?”

Eliciting Probability Judgements (4)

Cognitive Bias

- Biases due to human / behavioural influences and factors
- ~70 different types identified on Wikipedia!
- Overconfidence
  - Can assess by requiring ‘confidence interval’ estimates of uncertain quantities for which real values known
  - Possibility for bias correction(?)
- Underconfidence
  - “Have no idea.” / “Your guess is as good as mine!”
  - Test intuition of probabilities as before against everyday events
  - Decompose event into ‘causal’ sub-events where possible

Eliciting Probability Judgements (5)

Validation Step

- **Validation**
  - Should check results of model / risk assessment exercise by presenting consequences of subjective assumptions back to assumption-setters
  - Check for consistency and general ‘feeling’ of reasonableness
  - May find inconsistencies such as:

    $P(A) > P(B)$ and $P(B) > P(C)$, but $P(A) < P(C)$
  - May lead to re-iteration of probability assessment exercise

V. Conclusions

Conclusions (1)

- Sparsity must be addressed in a Bayesian framework, whether formalised or not – a frequentist approach just won't do

- One must be aware of all that sparsity entails: parameter uncertainty, model uncertainty, bias (possibly), and the breakdown of many of the quantitative methods that actuaries use

- There is a surprising number of **quantitative techniques** that can assist us when dealing with sparsity, ranging from extreme value theory to statistical learning theory

- Some level of expert judgement or prior constraints is essential for quantitative techniques perform well

Conclusions (2)

- **Qualitative approaches** tend to be centred on the subjective assessment of risky events, given sparse or no data…but Bayesian techniques can help the process of eliciting expert judgement

- We should be aware of cognitive bias and minimise the error in expert judgement caused by asking the wrong questions!

- Ultimately, for all approaches, transparency and interpretability of results are key

- Expert judgements should not go unchecked, and the importance of validation should not be underestimated

Questions or comments?

Expressions of individual views by members of The Actuarial Profession and its staff are encouraged.