CIRANO

Northfield Conference

Williamburg June 24-26, 2011

Modeling Risk with Limited Internal Data

Bryan Campbell

Vice-President, Finance Group, CIRANO

Department of Economics, Concordia University

Eric Stam

Project Director, Finance Group, CIRANO

Agenda

1. Context : Operational Risk Advanced Measurement Approach
2. Model I: Scenarios Only
3. Model II: Scenarios in Estimation
4. Model III: Scenarios as Filters
5. Conclusions

- **Operational Risk**

  Risk of loss resulting from inadequate failed or failed internal processes, people and systems, or from external events. [Basel II]

  Seven loss categories: Internal fraud; external fraud; employment practices and workplace safety; clients, products and business practices; damage to physical assets; business disruptions and systems failures; execution, delivery and process management.

  Eight business lines.

  Time horizon: one year

- **Advanced Measurement Approach [AMA]**

  Find Loss Distribution for relevant cells among 56 Basel categories.

  Aggregate to overall Loss Distribution [via copulas !].

  Determine appropriate tail quantile.

- **Modeling Strategy**

  Convolute frequency distributions [Poisson] with severity distribution.

- **Basel II directives**

Model must use internal data, relevant external data, scenario analysis and factors reflecting the business environment and internal control systems. Bank must have sound approach for weighting these 4 inputs

Banks "must use scenario analysis of expert opinion in conjunction with external data to evaluate [their] exposure to high-severity events. This approach draws on the knowledge of experienced business managers and risk management experts to derive reasoned assessments of plausible severe losses."

- **Some wisdom**

"A key lesson learned challenges is the need to add a forward perspective to historic data analysis – a key element of AMA"

"Let a thousand flowers bloom"

- **Aside : Current State of Play**

No international consensus on how to integrate internal, external data; nor consensus on how to use scenarios. Basel III more or less silent on Ops Risk.

- **Aside : Current State of Play**

Emerging idea: pool internal and external data to estimate, then use scenarios / judgment to validate. Apparent advantage: first step neutral and then use scenarios to stress test.

- **CIRANO models**

Scenarios are used in essential but different ways in VaR computations obtained by:

Model I: Can be implemented without data. Data used as prompts.

Model II: Integration of internal data, external data and scenarios in estimation.

Model III: Scenarios used to filter possibilities consistent with external data.

These are now discussed in turn. The focus is not so much the quest for VaR. Rather on how we can use manager input [eg, subjective probabilities] in model construction.

- **Implementation features**

  Questions are put to managers who work in small teams [3 - 5] chosen by a project administrator. Each is aware who is participating.

  The questions are straightforward and fall within the experience of the participants. Each knows how the others are responding as a group.

  The questions follow a sequence determined by the respondent's previous answers. There is a 10-question max.

- **Questions Posed**

  The first question concerns the median loss [$].

  Subsequent questions are yes/relative to median.

  A final question concerns the max loss.

  The important point is that the issues addressed fall within the experience of the respondent.

  There follows a mapping of the sequence of question responses onto distributions.

CIRANO

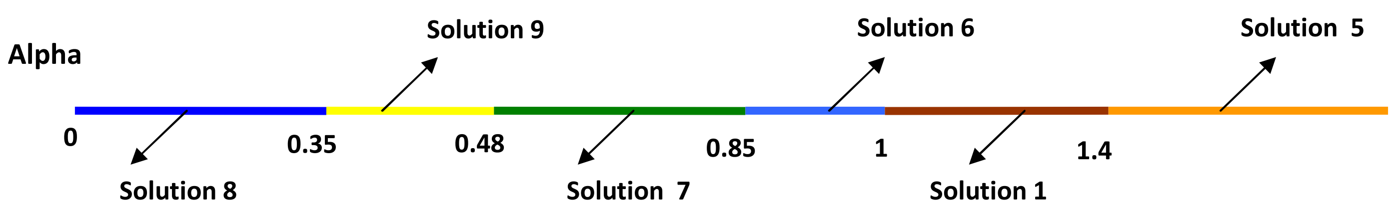
Model I : Scenarios Only

	+ 25% < Med/4	+ 25% < Med/2	+ 10% < Med/2	+ 25% > 2Med	+ 10% > 3Med	+15% > 3Med	+ 25% > 3Med	+ 25% > 4Med	Distributions
Sol #1	No	Yes (2)	Yes	No (1)	No	No	No	No	**Gamma**
Sol #2	No	No (2)	Yes (3)	Yes (1)	Yes	Yes	No	No	**Pareto 1**
Sol #3	No	No (2)	No (3)	Yes (1)	Yes	Yes	No	No	**Pareto 1**
Sol #4	No	No (2)	No	No (1)	Yes (3)	Yes	No	No	**Pareto 1 Lognormal**
Sol #5	No	No (2)	No	No (1)	No (3)	No	No	No	**Gamma Pareto 1 Lognormal**
Sol #6	No	Yes (2)	Yes	Yes (1)	Yes	No (4)	No (3)	No	**Gamma Lognormal**
Sol #7	No	Yes (2)	Yes	Yes (1)	Yes	Yes (4)	No (3)	No	**Gamma Lognormal Pareto 2**
Sol #8	Yes	Yes (2)	Yes	Yes (1)	Yes	Yes	Yes (3)	Yes (4)	**Gamma Pareto 1 Pareto 2 Lognormal**
Sol #9	Yes (5)	Yes (2)	Yes	Yes (1)	Yes	Yes	Yes (3)	No (4)	**Gamma**
Sol #10	No (5)	Yes (2)	Yes	Yes (1)	Yes	Yes	Yes (3)	No (4)	**Pareto 1 Pareto 2 Lognormal**

- **Mapping from Solution Space to Gamma Parameter Space**

Median : **1000 $**

Gamma distribution consistent with solutions : **1, 5, 6, 7, 8, 9**.



Gamma can’t handle solutions 2, 3, 4,10

There is a solution associated with each value of the parameter Alpha.

- **Similar results holds for lognormal, Paretos, Weibull.**

- **Second step - weight with Pareto if Max from respondent too large**

Solution 9			Q(maxAdmin)	Weighting (1.5 million simulations)			VaR (100 000 simulations)		Center					
				β Pareto	ω (weight on Gamma)	QMax Pareto	99%	99.9%	M/4 +25%	M/2 +25%	M 50%	2xM +25%	3xM +25%	4xM -25%
Gamma	Max (99.99%)	Max Admin							27	37	2000	35	25	19
Alpha : 0.47	74k	3M												
Beta: 9902														
							1.2M	1.4M						
Négative Binomiale	Frequency month	Max Frequency 40 losses												
N:2														
P:0.136														
Precision : 99%														
	Q(0.9999)	Value Q(maxAdmin)												
Weight1	3M	3M	0.9999	0.932	91.74%	0.999	5.9M	48M	27	37	2000	35	25	19
Weight2	653k	2.9M	0.99999	1.330	82.82%	0.9999	1.9M	5M	26	37	2000	34	25	19
Weight3	259k	2.7M	0.999999	1.75	74.07%	0.99999	1.4M	2.2M	25	36	2000	34	24	18
Weight4	166k	1.7M	0.9999999	2.18	66.30%	0.999999	1.2M	1.6M	24	36	2000	33	24	17

Judgment must be exercised on the choice of quantile to associate with Max.

- **Incorporating Internal Data**

Data as respondent: The answers to the question concerning the median and the dispersion of losses around the median would be answered by the data itself. The respondent wouldn't know « who » among his colleagues in the group was answering according to the data. This approach is appropriate with a sizeable data set.

Data as prompt: The data is presented as part of the scenario prompts. So when the question concerning the median is asked, the data value would appear on the screen.

This approach is suitable when is sparse: here the respondent would view the data as suggestive, confirming or not his/her intuitions. In practice, the data is blended with qualitative information.

- **Incorporating External Data**

The qualitative features of extreme external losses could be presented to respondents as part of the scenario used to elicit the Max. This has been our approach.

More quantitatively, we could use external data as a statistic for the Max. One idea: interpret the average of extreme losses as providing a measure of Expected Shortfall given a quantile. Let this measure be a constraint in the weighting determination.

Overview Model I

- Scenario questions from a fixed set enable a systematic selection procedure among a class of distributions.
- A small number of questions are quite effective.
- There is an underlying completeness result: any sequence of responses determines a distribution from the class of distributions and any distribution from the class corresponds to a sequence of questions.
- The set of potential solutions is small.
- [The model has been used as part of a Basel Pillar II review of operational risk procedures. The VaR determined from this approach cohered if not illuminated the number obtained from the standard calculation of regulatory capital for operational risk used by the bank.]

- **Preliminaries**

One drawback to Model I is that each respondent determined his/her own VaR. In practice, we have used interpreted the totality of responses as an interval,

It is natural to try to estimate over the set of responses before proceeding to the VaR determination.

The form of the question posed in this model involves both a frequency and severity dimension. Model II extracts a severity distribution from a set of responses.

The questions remain within the experience of the managers involved in the scenario sessions. External data is used in the determination of extreme events. The threshold, however, is estimated from the scenarios.

Internal data may be added to the estimation procedure.

- **Scenario Workshops**

Objective: Determine a range, and a distribution over this range, for mid-tail losses, based on expert opinion in the context of the business line and its risk environment looking forward.

Participants in the business line give their estimates [privately] of the [smallest] 1-in-5 year, 1-in-10 year, 1-in-25 year and 1-in-50 year losses in a workshop environment. x is a 1-in-n loss if losses greater than x occur with an annual frequency of 1/n.

Subjective data are arranged in pairs (λ_i, x_i), where λ_i, is the annual frequency of losses greater than x_i ; i = 1, 2, 3 and 4, corresponding to these four questions.

Aside. The relation between loss frequencies and the severity distribution is described by: $\lambda(x) = \lambda S(x)$, where λ is the frequency of losses, $\lambda(x)$ is the frequency of losses greater than x and S(x) is the probability that a given loss exceeds x.

Easy to show: if a loss is greater than 1-in-5 year, there is a 10% chance that it exceeds the 1-in-50 year subjective estimate.

- **Estimation –strategy**

We begin: $\lambda_i = \lambda S(x_i, \theta) \;\Rightarrow\; x_i = S^{-1}\left(\frac{\lambda_i}{\lambda}; \theta\right).$

Here θ is the vector of parameters that specify a distribution within a certain class of distributions such as the lognormal. Once we have estimates of the parameters, the 1-in-50 loss estimate is given by:

$$\hat{x}_4 = S^{-1}\left(\frac{\lambda_4}{\hat{\lambda}}; \hat{\theta}\right).$$

The expansion of the parameter set to include λ permits us to exploit the form of the scenario questions.

- **Estimation-Moment Condition**

$$\ln\left(x_i^{(k)}\right) = \ln\left(S^{-1}\left(\frac{\lambda_i}{\lambda}; \theta\right)\right) + \grave{o}_i^{(k)}, \qquad \text{for i = 1, … , 4.}$$

Here we assume the respondent *k* gives the true answer (in log form) with noise to the questions concerning losses with the given frequencies. There are K respondents.

- **Estimation- GMM with Updating**

$$\left(\hat{\lambda},\hat{\theta}\right)=\underset{(\lambda,\theta)}{\arg\min}\,\overline{g}_K(\lambda,\theta)'V_K(\lambda,\theta)^{-1}\,\overline{g}_K(\lambda,\theta).$$ See below:

The optimization is solved relative to a class of distributions S parameterized by θ. For each class, we obtain estimates of (λ, θ) that yield estimates of the values x_i. At CIRANO, we have used the lognormal, Pareto and Generalized Pareto classes.

$$g_k(\lambda,\theta)\equiv\left(\ln\left(\frac{1}{x_1^{(k)}}S^{-1}\left(\frac{\lambda_1}{\lambda},\theta\right)\right),\ldots,\ln\left(\frac{1}{x_4^{(k)}}S^{-1}\left(\frac{\lambda_4}{\lambda},\theta\right)\right)\right)' \qquad \overline{g}_K(\lambda,\theta)\equiv\frac{1}{K}\sum_{k=1}^{K}g_k(\lambda,\theta)$$

$$V_K(\lambda,\theta)\equiv\frac{1}{K}\sum_{k=1}^{K}\left(g_k(\lambda,\theta)-\overline{g}_K(\lambda,\theta)\right)\left(g_k(\lambda,\theta)-\overline{g}_K(\lambda,\theta)\right)'$$

The weighting matrix is important. Suppose that an overestimation of the 1-in-5 year loss generally leads to an overestimation of the 1-in-10, 1-in-25, and 1-in-50 year losses. If this is the case, we do not want to over-penalize for correlated errors. Allowing for a correlation structure between the error terms gives more efficient estimates.

- **Integrating Internal Data**

We add one point which in applications is the 1-in1 year loss estimate. The estimate itself is taken from the data. One task is to determine the variance of this sample estimate so that it can be included as part of the general GMM procedure. In our work the variance is determining via bootstapping.

GMM will then determine the parameters of the distribution describing the severity of losses between model estimates of the 1-in-1 year point and the 1-in-50 year point. [NB There is a delicate point concerning the weight applied to the added data point; discretion is the better part of valour here-another day.]

- **Integrating External Data**

Banks must rely on external data to capture the severity tail. Generally, such data does not report all losses but the probability of a loss being included increases with its severity. Stochastic threshold models have been developed to deal with this problem.

Due to a limited amount of external losses for some units of measure, we assume the stochastic threshold distribution to be the same across all units of measure and estimate the stochastic threshold distribution using all external data. Then we estimate using the standard maximum likelihood methodology the severity tail for different units of measures but using only external losses above the 1-in-50 years loss estimate obtained via GMM. A Pareto distribution is assumed.

- **Simulation of the Capital Estimates**

The frequency distribution is assumed to be Poisson with annual frequency λ . Losses fall be between $[0,T)$, in $[T,\tau)$, and in $[\tau,\infty)$ [where T, τ are the thresholds obtained from the GMM estimation] with frequencies $\lambda-\lambda_T$, $\lambda_T-\lambda_\tau$, and λ_τ that rae independent Poisson. The annual loss frequency is estimated internal data by dividing the number of losses by the number of years.

The VaR quantile is computed using Monte-Carlo simulation.

1. Draw $n_I \sim Poisson\left(\hat{\lambda}-\hat{\lambda}_T\right)$, $n_S \sim Poisson\left(\hat{\lambda}_T-\hat{\lambda}_\tau\right)$, and $n_E \sim Poisson\left(\hat{\lambda}_\tau\right)$.
2. Randomly draw n_I from internal losses below $\hat{T}$.
3. Draw n_S observations from the GMM fitted distribution. conditional on being in $\left[\hat{T},\hat{\tau}\right)$.
4. Draw n_E observations from the fitted Pareto distribution.
5. Sum up all losses to get the annual losses.

Repeat many times.

Overview Model II

- Here scenarios are used in a novel way enabling three severity distributions to be stitched together to reflect characterizations of losses based on internal, scenario and external data respectively. The points of contact of the three distributions are themselves estimated.
- The scenarios are the basis of an estimation procedure that determines mid-range severity. Internal data can be incorporated in this procedure.
- Moreover, the frequency of losses that fall within each distribution is determined naturally from the empirical frequency of losses calibrated to the form of the scenario questions.

- **Preliminaries**

In a business risk modeling exercise, we pooled de-meaned revenue data from a number of banks over a number of years. A t-distribution with unknown scaling parameter and degrees of freedom was estimated was estimated using weighted maximum likelihood (the weights reflecting proximity to the activity of our client bank. The VaR was read off the estimation and combined with the bank's mean to determine an appropriate economic capital level for the bank.

The presupposition that the same t-distribution fits all banks seemed lame.

How could subjective judgment be brought to bear on the problem?

- **Modeling Framework- Introduction**

We assume that losses follow a distribution with density function $f(x|\theta)$. For example, $f(x|\theta)$ could be lognormal with the usual mean and variance parameters. The objective is to find estimates of the parameters.

Even though it would be unreasonable to assume that banks share the same loss distribution, it would foolish to conclude that such external data is irrelevant. Accordingly, we let θ take on different values for each bank and we introduce a second distribution to describe the vector of parameters: this involves the density function $\pi\,(\theta\,;\Phi)$ called the hyper-distribution, where the parameters in Φ are called the hyper-parameters.

In this vein, we could assume that losses follow lognormal distributions with parameters given by (μ_i, σ_i), while $(\mu_i, \ln\sigma_i)$ follow bivariate normal distributions with mean μ_θ and covariance Σ_θ. Accordingly, $\theta_i = (\mu_i, \ln\sigma_i)$ represent the parameters and are different for each bank i, while $\Phi = (\mu_\theta, \Sigma_\theta)$ represent the hyper-parameters which are estimated using external data.

- **Modeling Framework- Some Details**

Assuming losses are independent, the likelihood function of losses for each bank is given by:

$$\begin{aligned}\ell_i &= f\left(x_{i1}, x_{i2}, \ldots, x_{iN_i}; \Phi\right) \\ &= \int_\theta f\left(x_{i1}, x_{i2}, \ldots, x_{iN_i}, \theta; \Phi\right) d\theta \\ &= \int_\theta f\left(x_{i1}, x_{i2}, \ldots, x_{iN_i} \mid \theta\right) \pi(\theta; \Phi) d\theta \\ &= \int_\theta \prod_{j=1}^{N_i} \left\{ f\left(x_{ij} \mid \theta\right) \right\} \pi(\theta; \Phi) d\theta\end{aligned}$$

Pooling all external data, the likelihood function becomes:

$$\ell = \prod_{i=1}^{K} \ell_i = \prod_{i=1}^{K} \int_\theta \prod_{j=1}^{N_i} \left\{ f\left(x_{ij} \mid \theta\right) \right\} \pi(\theta; \Phi) d\theta$$

The hyper-parameters in Φ may be estimated by maximizing the likelihood function. In practice, we are faced with numerical issues. The integral inside the likelihood function may not have a tractable solution, in which case it needs to be computed numerically. Furthermore, the likelihood is maximized over the hyper-parameters in Φ, which means that the integral needs to be computed at each iteration. However, numerical approximation techniques may be used when adopting a multivariate normal hyper-distribution.

- **Constraining the Parameter Space via Expert Opinion**

The hyper-distribution density $\pi\left(\theta;\hat{\Phi}\right)$ offers a way to measure the credibility of loss distribution $f(x|\theta)$ relative to losses observed in the industry. The universe of credible loss distributions may be narrowed based on expert opinion. It is important to formulate questions that can be easily understood by practitioners. We design the questionnaire such that the general formulation of questions may be adopted without any knowledge of the loss distribution. During the execution of the questionnaire, questions are generated from the estimated hyper-distribution and answers from previous questions.

We opt for simple yes/no questions: Is the probability of a loss exceeding X more than 25%? This is an alternative way of asking a question about the 75% percentile. There is no need to worry about the specific value in X, which will be determined from the hyper-distribution. It is still important to exercise judgment in formulating proper questions (e.g. avoid asking about the 99% percentile).

- **Execution of the Questionnaire**

Parameters are simulated from the hyperdistribution. Each question is calibrated so that half the simulations are consistent with any of the answers.

"Is the probability of a loss exceeding X more than 25%?" The 75% percentile of the loss distribution is computed for each simulated parameters. The sample median of the 75% percentiles replaces the value X. If the answer to this question is yes, then the 75% percentile needs to be lower than X, and all simulations with a 75 % percentile lower than X are thrown away. Having taken the median of the percentiles, this is exactly half the sample. The same is true if the answer is no, in which case all simulations with a 75 % percentile greater than X are thrown away.

The resulting sample is still simulated from the hyper-distribution but reflects the constraints imposed by the answer. This may be interpreted as an updating process, and the new sample becomes a basis for following question. The questionnaire adapts to the answers given at every step. Half the sample is dropped after each question, and the parameter space shrinks quickly. Assuming expert opinion is reliable, the parameter space shrinks around the true parameter.

Assuming a multivariate normal hyper-distribution, a numerical procedure based on convolution of the covariance matrix may be used instead of simulation.

CIRANO

- **Validation of the Hyper-Distribution using Internal Data**

First, compute the average of the likelihood of internal data over the remaining parameter space S.

Next, determine average likelihood from simulated data: draw a distribution from S, then draw *n*-values (*n,* the internal data sample size) from this distribution. For this sample compute the average likelihood over S. Determine whether the internal data average is greater than this value.

Repeat the process many times. If the percentage of times the sample average likelihood is greater than the simulated value is too low, then there is evidence that internal data does not validate the model and/or expert opinion.

- **Computing Regulatory Capital**

A first approach is to derive point estimates for the loss distribution parameters θ . In practice, $\hat{\theta}$ is approximated by taking the sample mean or median over simulated parameters $\{\theta_s\}$ under constraints imposed by expert opinion. The regulatory capital may be computed based on $\hat{\theta}$ and will be denoted as VaR | $\hat{\theta}$. This approach places more emphasis on the loss distribution parameters instead of VaR.

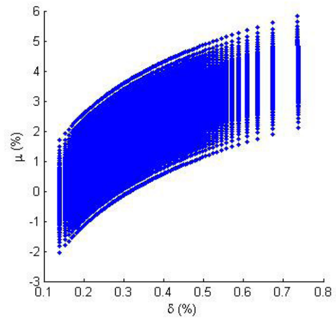
We propose estimating VaR directly by computing the 99.9% percentile for each set of parameters in $\{\theta_s\}$, and taking the mean or median of resulting $\{VaR_{.99.9\%, s}\}$. The value is denoted $\{V\hat{a}R_{99.9\%}\}$, Standard errors and credibility intervals may also be derived.

- **Example - Estimation**

The framework is applied in the context of business risk modeling. We have yearly financial statements for 99 financial institutions for up to 16 years. Revenues are adjusted in order to remove trading revenues and operational losses. Revenues are then divided by total assets to account for bank size resulting in return on assets (ROA). The objective is to estimate the hyper-distribution for return on assets.

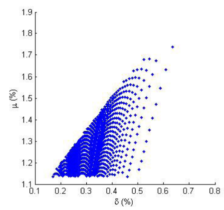
We assume return on assets follow a student-t distribution sharing the same degree of freedom ν but with different mean and scaling parameters μ and δ. Furthermore, we assume μ and lnδ follow a bivariate normal distribution.

A scatter plot of the resulting parameter space (with $\hat{\nu} = 3.8835$):

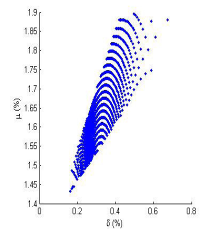


- **Example – Two Runs**

Pr[ROA<1.8983] > 50%	Y	Est-VaR -1,14%
Pr[ROA<1.1349]> 50%	N	Std. Err. 0,49%
Pr[ROA<1.3106]>25%	Y	Cred. Int. 95%
Pr[ROA<0.8910]>10%	Y	-0,37% -2,28%



Pr[ROA<1.8983] > 50%	Y	Est-VaR -0,82%
Pr[ROA<1.1349]> 50%	N	Std. Err 0,48%
Pr[ROA<1.3106]>25%	N	Cred. Int. 95%
Pr[ROA<1.2373]>10%	Y	-0,01% -1,89%



Overview Model III

- Recap: external data is used to demarcate a wide parameter space within a family of distributions. Scenarios in the form of a series of questions suggested by the sample space are used to filter the space.
- Internal data could be used in the original estimation process. Here the data is used to validate using a statistical procedure.

- Our goal has been to illustrate that subjective probabilities can be used in different ways to different ends.
- The organizational application of any of the Models illustrated in this talk would certainly serve to develop more refined risk sensibilities among the managers involved.
- We think that scenarios can be afforded a more interesting role than a perfunctory *ex post* validation of statistical estimates of VaR quantiles.
- We think that scenarios can facilitate the transfer of risk management techniques developed in the banking/financial sector to the non-financial sector.