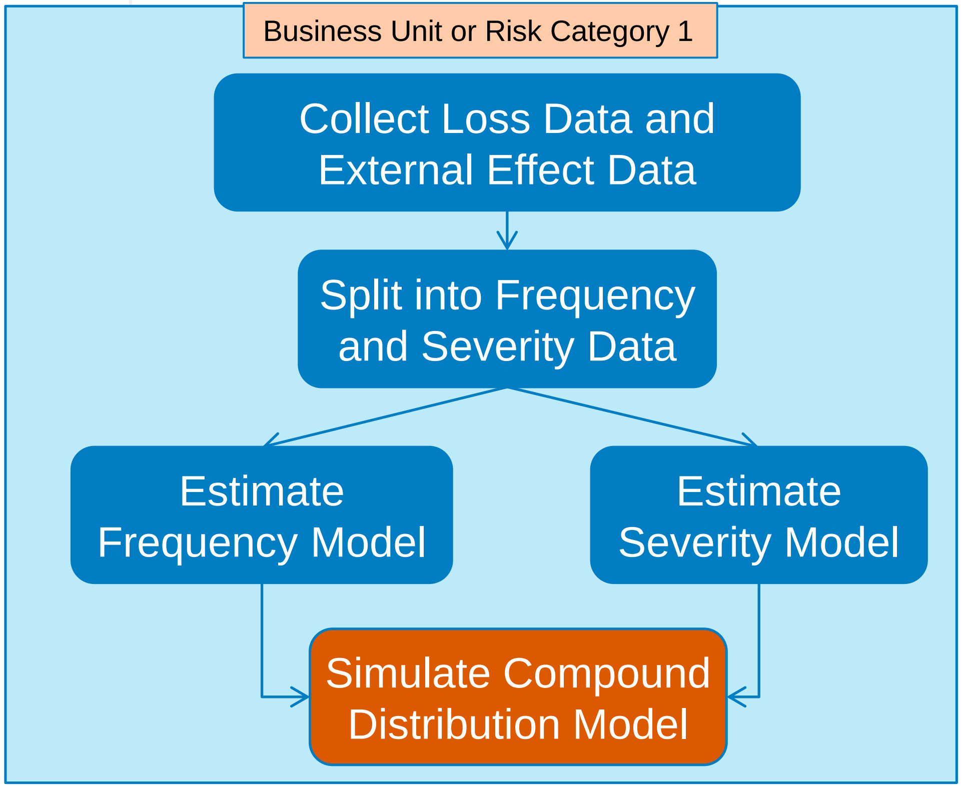
METHODS OF COMPUTING A LARGE NUMBER OF QUANTILES FROM AN AGGREGATE LOSS DISTRIBUTION

MAHESH V. JOSHI, Ph.D.
ADVANCED ANALYTICS R&D
SAS INSTITUTE INC.

LOSS DISTRIBUTION APPROACH

PROCESS



COMPOUND DISTRIBUTION MODEL (CDM)

- Collective risk model
  - $\{X_i\}$: iid random variables for severity
  - N: frequency random variable (independent of all $\{X_i\}$)
  - Aggregate loss is a random variable $S = \sum_{i=1}^{N} X_i$
- What is the probability distribution of S? The cumulative distribution function (CDF) of S is

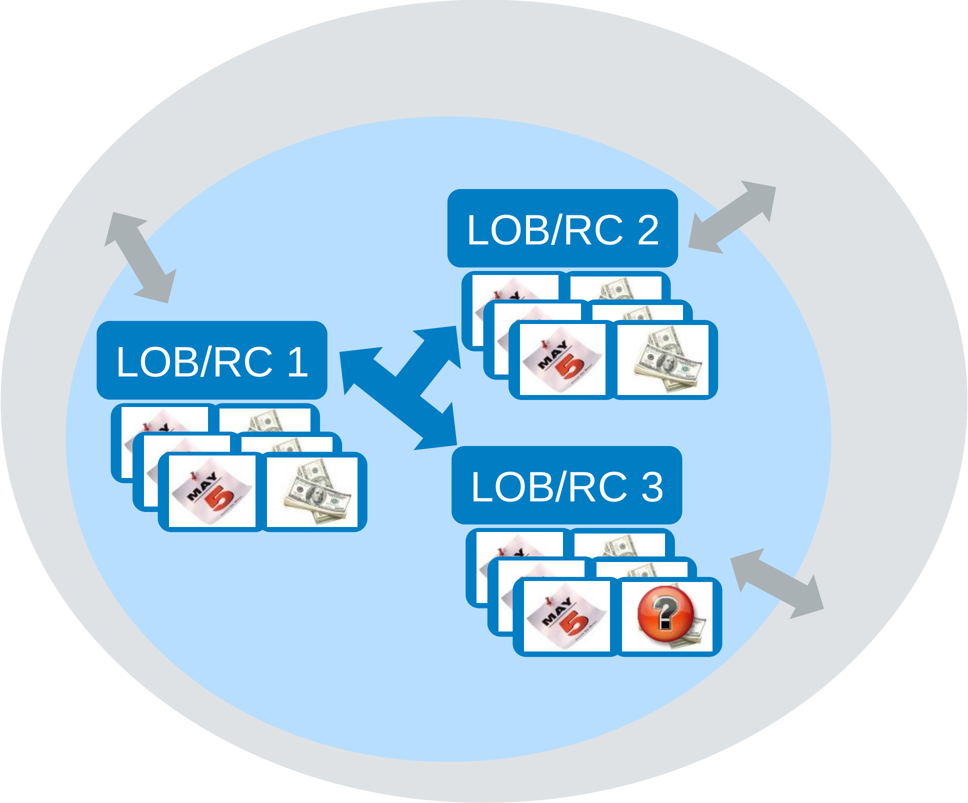
$$F_S(s) = \sum_{n=0}^{\infty} \Pr(N = n).\, F_X^{*n}(x)$$

- Closed form solution is rarely available; hence, simulation method is used

ECONOMIC CAPITAL MODELING

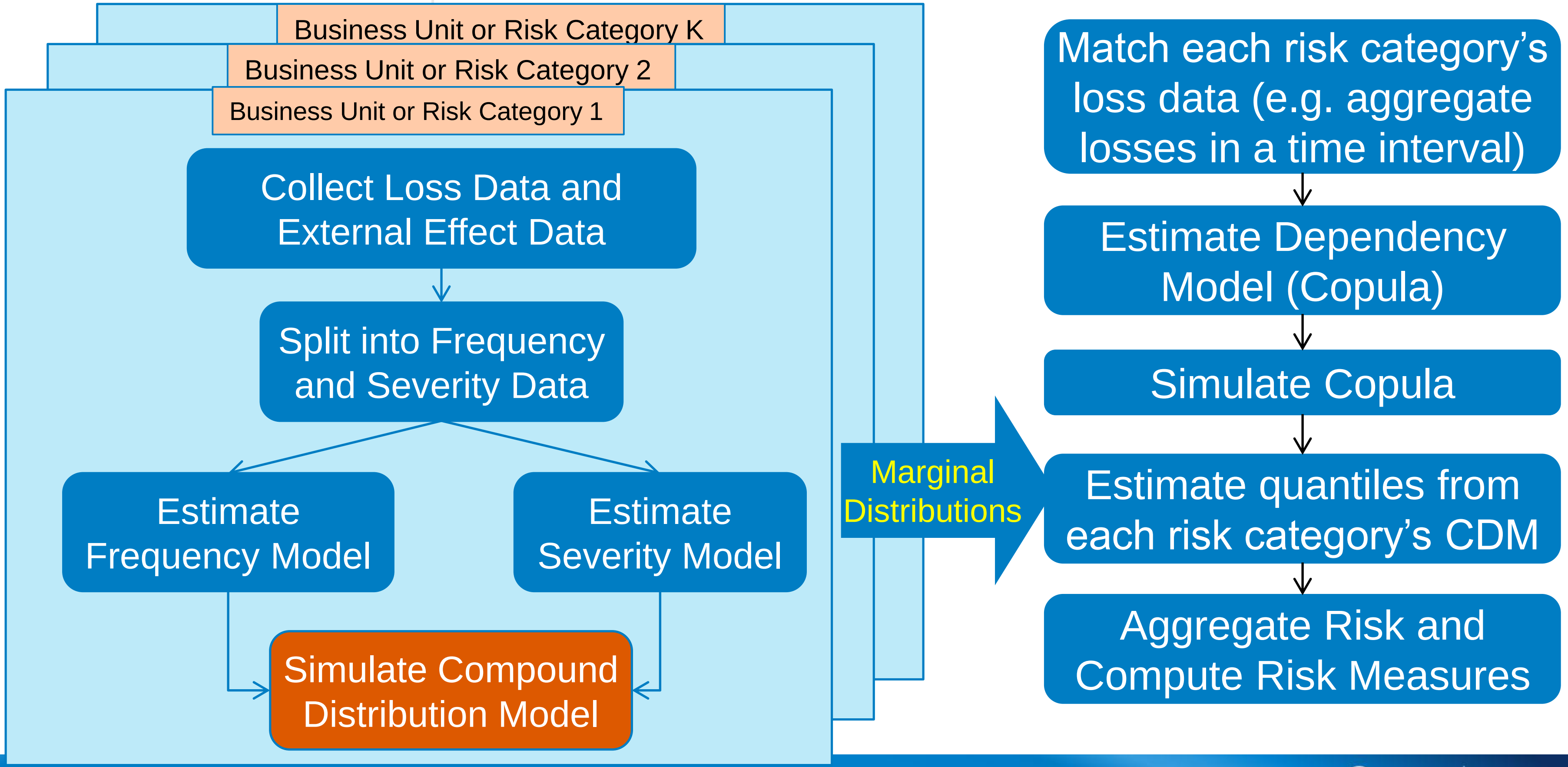
ENTERPRISE-WIDE AGGREGATE LOSS

- Need to account for correlation between different lines of business or risk categories
- Copulas help identify the dependence structure



ECONOMIC CAPITAL MODELING

AGGREGATING LOSSES FROM DIFFERENT UNITS



ECONOMIC CAPITAL MODELING

COMBINING COPULA AND CDM SIMULATIONS

RC1	RC2	RC3
0.3	0.6	0.2
0.5	0.25	0.7
…	…	…
0.75	0.5	0.6
0.2	0.9	0.3

N

→

RC1	RC2	RC3
$F_1^{-1}(0.3)$	$F_2^{-1}(0.6)$	$F_3^{-1}(0.2)$
$F_1^{-1}(0.5)$	$F_2^{-1}(0.25)$	$F_3^{-1}(0.7)$
…	…	…
$F_1^{-1}(0.75)$	$F_2^{-1}(0.5)$	$F_3^{-1}(0.6)$
$F_1^{-1}(0.2)$	$F_2^{-1}(0.9)$	$F_3^{-1}(0.3)$

→

Total Loss
$\sum_{j=1}^{3} F_j^{-1}(u_{1,j})$
$\sum_{j=1}^{3} F_j^{-1}(u_{2,j})$
…
$\sum_{j=1}^{3} F_j^{-1}(u_{N-1,j})$
$\sum_{j=1}^{3} F_j^{-1}(u_{N,j})$

- For each risk category (RC), F_j is the CDF of CDM
- N is typically in millions

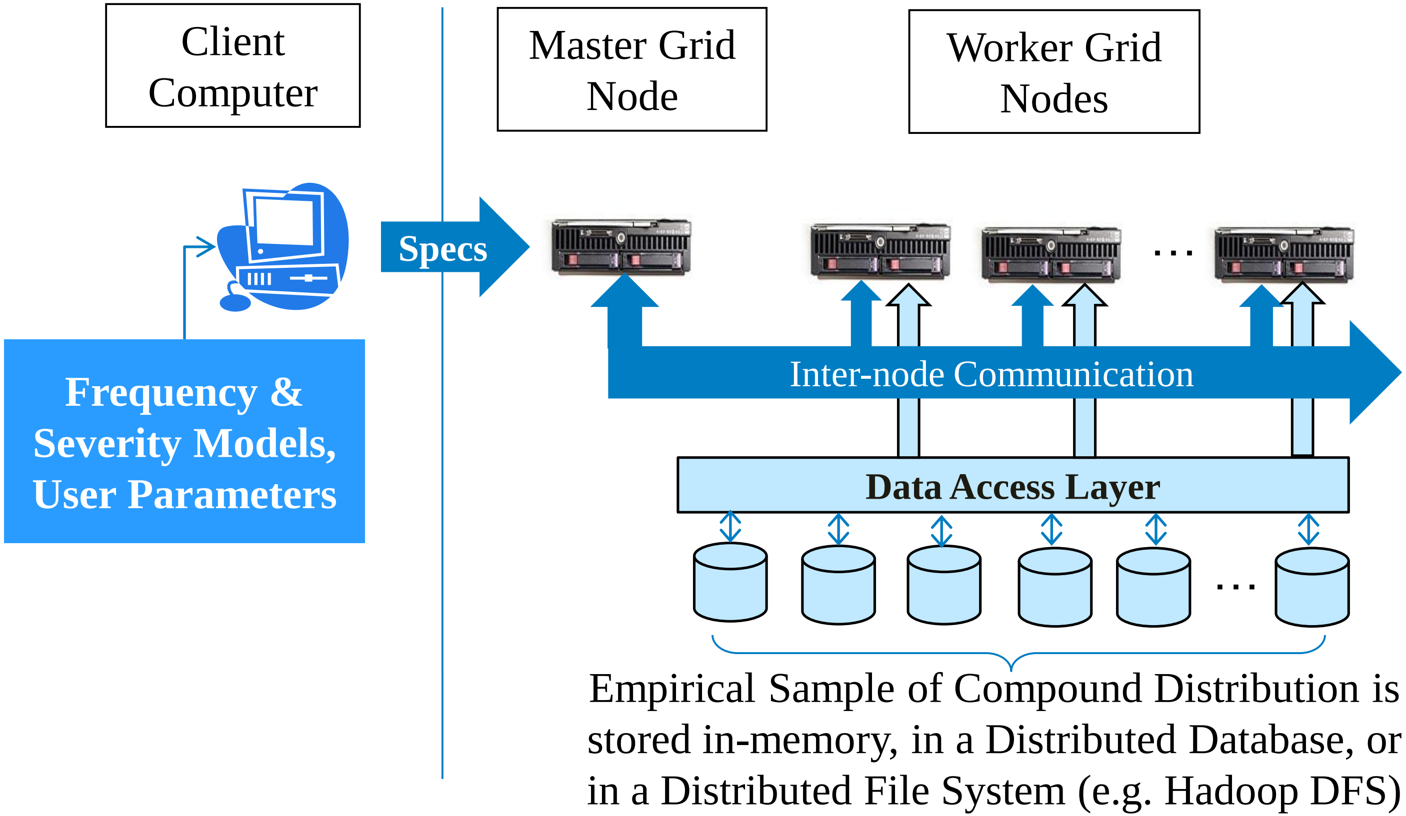
COMPUTING CDM PERCENTILES

CHALLENGES

- Need to compute a large number of percentiles from a large empirical sample of CDM; there are multiple such CDMs (one for each RC)
- The empirical sample might be stored in a distributed fashion on multiple computers if simulation was performed on multiple computers
- CDM sample of each RC might need to be stored for future use; storing multiple large, distributed samples might be expensive

COMPUTING CDM PERCENTILES

PARALLEL AND DISTRIBUTED CDM SIMULATION



COMPUTING CDM PERCENTILES

AN EMPIRICAL APPROACH

- Compute the EDF of CDM sample and store it along with the aggregate loss values. Then, sort the required percentiles in ascending order and lookup the desired percentiles in the EDF data structure
- If the CDM sample is distributed across multiple computers
  - Bring the sample on one machine and follow first bullet’s method, or
  - Employ a sophisticated distributed percentile computation algorithm that does not require bringing the CDM sample on one node

COMPUTING CDM PERCENTILES

A PARAMETRIC APPROXIMATION APPROACH

- Fit a parametric probability distribution to CDM’s empirical sample
- For more accurate percentile computations, fit the parametric distribution by using a minimum distance estimator (Cramér-von Mises)
  - Attempt to minimize distance between EDF (nonparametric) and CDF (parametric)

CDM'S PARAMETRIC APPROXIMATION

APPROXIMATING DISTRIBUTIONS TO TRY

- Mixture distribution might be more appropriate
  - Body-tail mixture
  - A finite mixture of multiple components, each with a distribution from the same of different parametric families

$$f(x;\Theta) = \sum_i p_i g_i(x;\Theta_\mathrm{i}) \qquad F(x;\Theta) = \sum_i p_i G_i(x;\Theta_\mathrm{i}) \qquad \sum_i p_i = 1$$

  - Zero-inflated family (mixture of a Bernoulli distribution for 0 and any parametric family for the non-0 values), because CDM sample typically contains lots of 0s

$$f(0;\Theta) = \phi + (1-\phi)h(0;\Theta) \qquad F(0;\Theta) = \phi + (1-\phi)H(0;\Theta)$$

$$f(x;\Theta) = (1-\phi)h(x;\Theta) \qquad F(x;\Theta) = \phi + (1-\phi)H(x;\Theta)$$

EXPERIMENTS | PARAMETRIC APPROXIMATION

- Case 1: Compounding of Poisson frequency and gamma severity
- Case 2: Compounding of negative binomial frequency and lognormal severity
- Tools used: SAS/ETS® and SAS® High Performance Econometrics
  - PROC COUNTREG: fits frequency models
  - PROC SEVERITY: fits any continuous distribution models for severity while accounting for censoring, truncation, and regression effects.
    - PROC HPSEVERITY: High performance version that can use a grid of multiple computers to speed up estimation, and can work on distributed data
  - PROC HPCDM: estimates compound distribution model by potentially using a grid of multiple computers to generate large, distributed empirical sample

EXPERIMENTS | CASE 1 (POISSON FREQUENCY X GAMMA SEVERITY)

- The numbers show values of Cramer-von Mises objective function defined in PROC HPSEVERITY as

  cvmobj = (_EDF_(y) - _CDF_(y))**2
- Tweedie is the best among the several candidates
  - Fitted value of index parameter ‘p’ is 1.333; for 1 < p < 2, Tweedie is a compound Poisson distribution
- Zero-inflated distributions perform consistently and significantly better than their *base* counterparts

logngpd	224.93052	
lognmix2	339.27012	
lognmix3	214.21028	
lognmix4	205.99310	*
lognmix5	293.15687	
Burr	250.75467	
Exp	277.64059	
Gamma	257.24872	
Igauss	431.61909	
Logn	372.61157	
Pareto	277.85501	
Gpd	277.64060	
Weibull	250.62073	

zilognmix2	20.43533	
zilognmix3	23.32376	
zilognmix4	18.18846	
zilognmix5	0.04437	*
ziburr	0.31210	
ziexp	172.07706	
zigamma	1.28357	
ziigauss	39.04171	
zilogn	27.86033	
zipareto	174.22185	
zigpd	172.07706	
ziweibull	0.45926	

tweedie	0.04079

EXPERIMENTS

CASE 2 (NEGATIVE BINOMIAL FREQUENCY X LOGNORMAL SEVERITY)

- Zero-inflated mixture of four lognormal distributions is the best among several candidates
- Again, zero-inflated distributions perform consistently and significantly better than their *base* counterparts

logngpd	32375	
lognmix2	32667	
lognmix3	32376	
lognmix4	32372	*
lognmix5	32372	
Burr	32391	
Exp	32933	
Gamma	32379	
Igauss	32429	
Logn	32397	
Pareto	32412	
Gpd	32412	
Weibull	34028	

zilognmix2	0.00425	
zilognmix3	0.14113	
zilognmix4	0.00381	*
zilognmix5	0.17993	
ziburr	0.21136	
ziexp	1.19346	
zigamma	1.07195	
ziigauss	1.24199	
zilogn	0.22983	
zipareto	0.61064	
zigpd	0.61064	
ziweibull	0.98002	

tweedie	1.14085

COMPARE APPROACHES

EMPIRICAL METHOD

- Pros:
  - Relatively easy to implement if entire sample is brought on one computer
  - Might be faster with sorted traversal of the EDF data structure
  - Always applicable!
- Cons:
  - Need to store the entire sample for future use (non-parsimony)
  - Might not be faster if the sample is distributed and it is prohibitively expensive to bring it all on one computer

COMPARE APPROACHES

PARAMETRIC APPROXIMATION

- Pros:
  - Parsimony: compresses the empirical sample into a few set of numbers (parameters)
  - Might be faster if approximating distribution can be found relatively quickly
  - Parallel nonlinear optimization algorithms can be employed to make estimation quicker when the CDM sample is distributed on multiple computers
  - Cost of estimation can be amortized over large number of quantile computations if quantiles are computable relatively quickly
- Cons:
  - Might not be applicable if satisfactory approximating distribution cannot be found
  - Might not be faster if search for an accurate approximating distribution takes longer
  - Might not be faster if the quantiles are expensive to compute (for mixture distribution, quantile often needs to be computed by numeric inversion of CDF).

SUMMARY

- Presented the problem that requires computation of large number of quantiles from multiple aggregate loss distributions
- Presented empirical and parametric approximation methods for computing percentiles
- Each method is worth trying depending on the scale of the problem and the ease with which approximating distribution can be found

- contact: mahesh.joshi@sas.com