# Can High Performance Computing Practically Improve A Live Corporate Default Prediction Platform?

Jin-Chuan Duan
(段锦泉)
National University of Singapore
(September 2020)

Credit Research Initiative
Risk Management Institute, National University of Singapore

# Corporate Default Prediction Globally

- Corporates (with limited liabilities) face default/bankruptcy. When a default occurs, the obligor (i.e., a corporate) may NOT be able to honor its debt obligations in full.
- The probability of default (PD) is time-dependent, unique to an obligor, specific to a horizon.
- A corporate may disappear for a reason other than default/bankruptcy, for example, a merger/acquisition. So, the probability of other exits (POE) must also be factored in.
- The PD and POE naturally depend on economic environments (common drivers) and firm-specific characteristics (individual attributes).
- The variables (common drivers and individual attributes) are expected to be time series dependent and cross-sectionally correlated.

# Corporate Default Prediction Globally (continued)

- The variable dimension is expected to be extremely high; for example, 3 common drivers and 5 individual attributes for 10,000 corporates will result in an overall dimension equal to 50,003.
- The data set can be viewed as a large incomplete data panel where the $Y$ variable (indexed by firm-time) is categorical (0,1,2) and the $\boldsymbol{X}$ variables (indexed by firm-time-number) is a vector.
- The PD model can be expressed conceptually as ($Y = 1$ denotes a default)

$$Prob_t\left(Y_{i,t+\tau} = 1\right) = f\left(\boldsymbol{X}_{i,t}; \boldsymbol{\theta}\right) \quad \text{for firm } i \text{ at time } t \text{ over horizon } \tau$$

- The PD system must obey many constraints, for example, a term structure constraint: $Prob_t\left(Y_{i,t+\tau+k} = 1\right) \geq Prob_t\left(Y_{i,t+\tau} = 1\right)$

# The CRI Computation Tasks

- The Credit Research Initiative (CRI) live corporate default prediction platform handles a database on over 70,000 exchange-listed firms in 133 economies covering a time span of 30 years.
- The currently active firms are over 36,000 in 133 economies and their daily-updated PD term structures are generated in a 3-time zone operation.
- The CRI computing tasks involve periodic model calibrations, default prediction updates reflecting input value changes, and various aggregations into economies/sectors.
- The CRI computing tasks are carried out in four frequencies: daily, monthly, quarterly and yearly.

# The CRI Computation Tasks (continued)

## Daily tasks

- Individual firm PD (probability of default) and AS (actuarial spread) term structures: Very fast
- Portfolio default rate distributions: Time consuming

## Nature of the work

- Compute various portfolio default rate distributions for countries, regions and sectors formed out of 36,000 currently active exchange-traded firms
- Require a default correlation model with a dynamic factor structure to simulate future scenarios, calibrate to individual PD term structures, and perform conditional convolutions, etc

| | |
|---|---|
| **Computation system** | 20+ CRI PCs (2 types of PCs - 8 CPU cores (use 4) and 12 CPU cores (use 8)) |
| **Software** | Julia 1.0.3 |
| **Computation time** | The whole daily operation takes about 3.5 hours. The PD term structure calibration is most time consuming, which takes about 2.5 - 3 hours. We need to perform optimization on 36,000+ firms, which are totally independent and parallelable. We divide them into 100-firm slices with each slice taking about 7-10 minutes to complete. The task is not memory-hungry. |

# The CRI Computation Tasks (continued)

**Monthly task** (Calibration for 6 regional PD models)

- Monthly update: Time consuming
- Full sequential run: Very time consuming

**Nature of the work**

- Monthly update the point estimates of the PD model (about 50 parameters for each model): Perform sequential Monte Carlo (SMC) update with the new data accumulated over a month
- Perform full sequential runs to compute confidence intervals for the model parameters

| | |
|---|---|
| **Computation system** | K80 and P100 GPU computers perform the main task, which involves repeated likelihood evaluations of a large dataset for multiple prediction horizons |
| **Software** | Julia 1.0.3 |
| **Computation time** | • Each of the 6 calibration groups takes about 2-4 hours (due to different data sizes) to complete the monthly update, and the task is memory-intensive. The data is growing by the month. Currently, the calculation takes about 10 GB of memory and 5 GB of GPU Memory.<br>• The full sequential run for each group takes about 2-3 days. |

# The CRI Computation Tasks (continued)

## Quarterly task

- Distant-to-Default (DTD) model calibration: Extremely time consuming

## Nature of the work

- The DTD model is used daily to generate a key risk factor (volatility-adjusted leverage ratio) in the PD model (New methodology: Joint estimation of multiple firms via SMC). For a sector in an economy of, say, 100 firms, we need to perform an SMC maximum likelihood estimation of a model with 301 model parameters where 1 common parameter for all 100 firms and 3 individual parameters specific to each firm.

| | |
|---|---|
| **Computation system** | 20+ CRI PCs (2 types of PCs - 8 CPU cores (use 4) and 12 CPU cores (use 8)) |
| **Software** | Julia 1.0.3 |
| **Computation time** | The whole operation takes about 1-2 days for each month out of a total of over 300 months (30-year time span). Firms are divided into multiple groups based on economy/sector, and thus parallelable. To make it manageable, we take a short cut to skip recalibration for a particular month of an economy/sector group if there were limited data revisions over that month in the historical file. |

# The CRI Computation Tasks (continued)

## Yearly tasks

- Default correlation model recalibration: Extremely time consuming
- Re-ranking CriSIFI (CRI Systemically Important Financial Institutions) for 2,000+ banks and insurance companies worldwide: Very time consuming

## Nature of the work

- The default correlation model is a low-rank factor model where factors are some pre-specified macro risk drivers resulting from a variable selection and sparse residual correlations are also allowed. The common factors follow the vector autoregressive time-series model.
- The CriSIF model hinges upon the default correlation model and relies on a constructed financial network that captures the notion of too-big-to-fail and too-connected-to-fail. The connections in the network are partial default correlations obtained by imposing regularization.

| | |
|---|---|
| **Computation system** | 20+ CRI PCs (2 types of PCs - 8 CPU cores (use 4) and 12 CPU cores (use 8)) |
| **Software** | Julia 1.0.3 |
| **Computation time** | The default correlation model recalibration takes about 1 month. And re-ranking CriSIFI takes about 1 week. Firms are divided into multiple groups based on economy/sector, and thus parallelable. |

# An Experimental Run on the NSCC Facilities

**The experimental run**

- Take 21 out of 300+ daily tasks for the PD term structure calibration (2.5-3 hours) to run our Julia code on multiple NSCC nodes (each node with 24 CPU cores).
- Each one of the 21 tasks took 4-6 minutes to complete (exclusive of the queuing time). An NSCC technician assisted us to submit the job for this experimental run, but we still experienced non-trivial variable queuing times (10-30 minutes).
- On the per-task basis, it ran marginally faster than using our own computers (7-10 minutes) even though each NSCC note has 24 CPU cores vs 8-12 cores in our computers. (I assume that the clock speed of a 24-core computer is set lower to avoid the heat problem.)
- Given 300+ tasks in total, we can expect to complete the daily task in 15 minutes if we can access 300+ nodes with a minimal queuing time.

In summary, NSCC may present a realistic alternative for the CRI's computational needs.