Machine Learning in Finance Workshop 2021

The Use of Synthetic Data to Determine Capital Adequacy

John C. Hull

University of Toronto

Columbia University Data Science Institute

Columbia Engineering
The Fu Foundation School of Engineering and Applied Science

Bloomberg

Background

- This is joint work with Jay Cao, Jacky Chen, Zissis Poulos, and Dorothy Zhang carried out at FinHub, a research Center at the Rotman School of Management
- The paper “Synthetic Data: A New Regulatory Tool” can be downloaded from my website:

  www-2.rotman.utoronto.ca/~hull/downloadablepublications

  or from ssrn 3908626

Synthetic Data

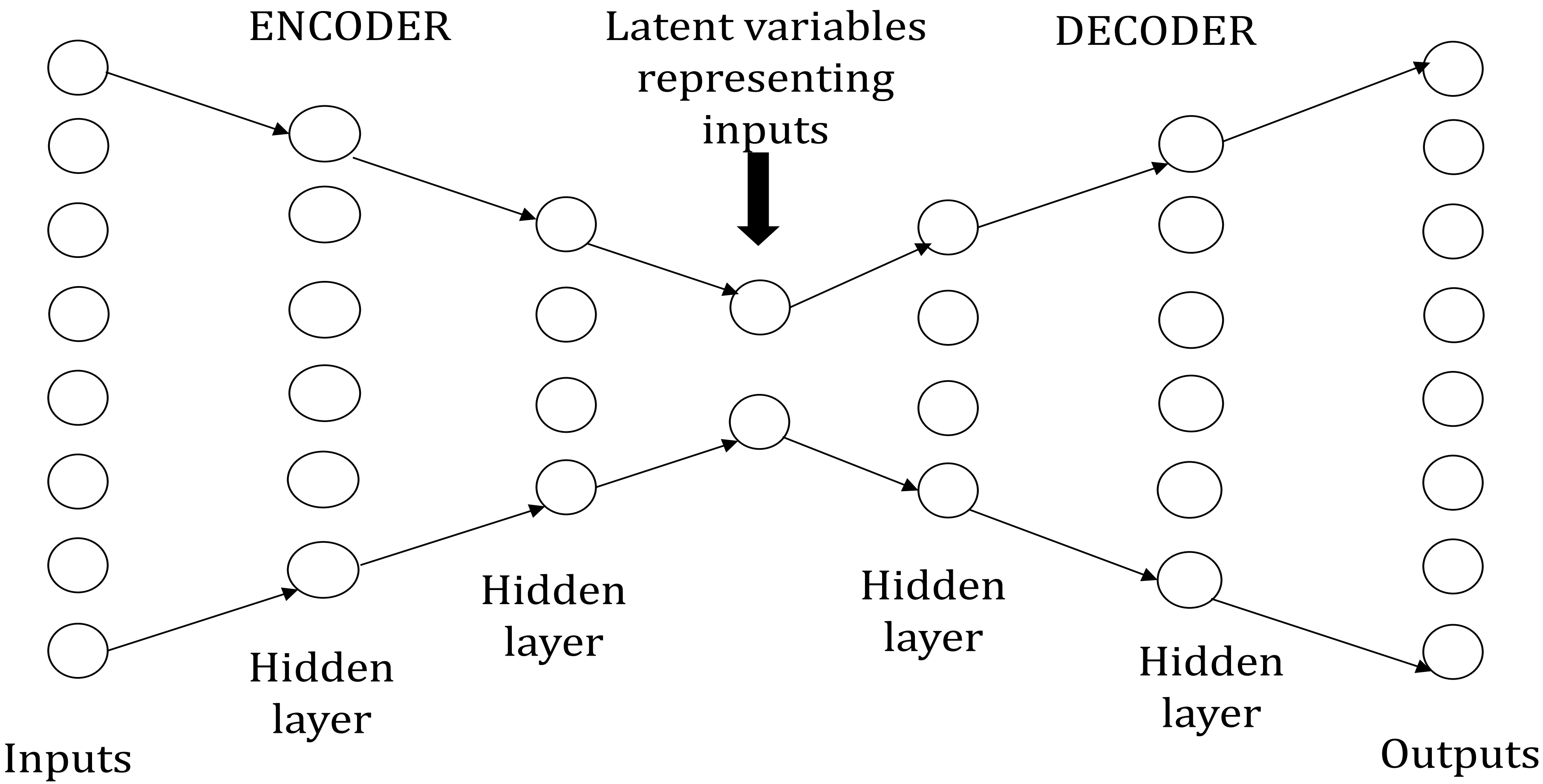
- Machine learning tools have been used to generate synthetic data that is indistinguishable from historical data
- In this research we investigate whether synthetic data can be used to provide reliable risk measures when confidence levels are high
- We think this is potentially useful to regulators who want banks to keep enough capital to withstand adverse outcomes with 99.9% confidence. It is also potentially useful to financial institutions themselves

Methods for producing synthetic data

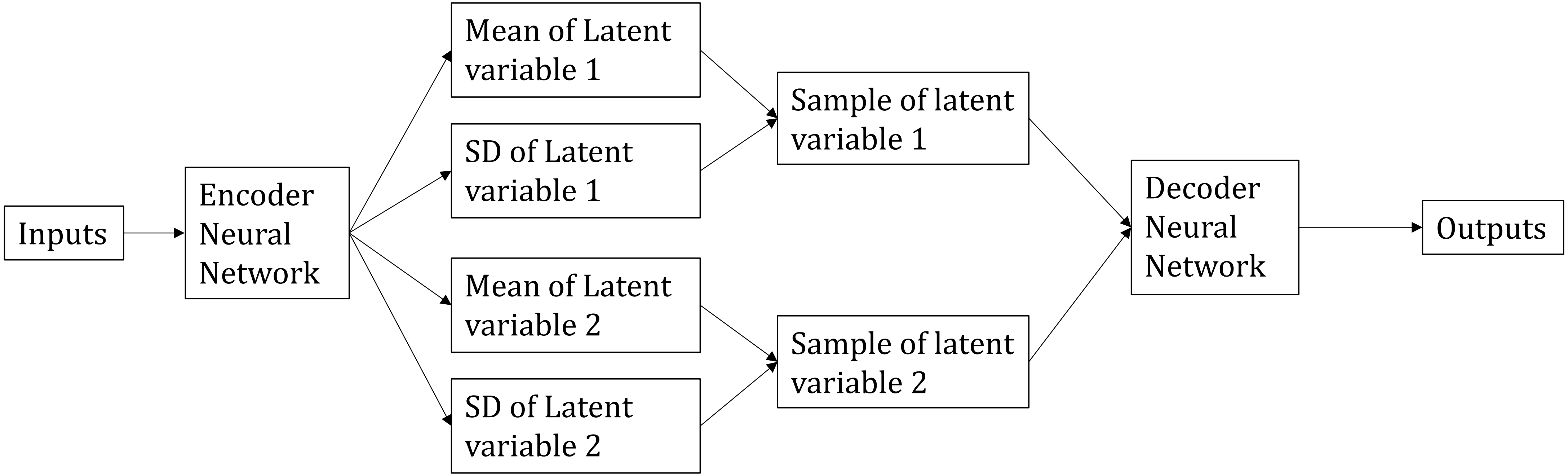
- Variational autoencoder (VAE)
- Generative adversarial network (GAN)
- Restricted Boltzman machine (RBM)
- Temporal Convolutional Network (TCN)

We chose VAE because it has been shown to work well with financial data. Furthermore it gives good results when a relatively small amount of historical data is available.

Autoencoder



Variational Autoencoder



The Objectives

- Make outputs as close as possible to inputs (the reconstruction error, RE)
- Make the unconditional distribution of each latent variable as close as possible to normal with mean zero and standard deviation 1 (achieved by minimizing Kullback-Leibler divergence, KL)

We minimize:

$$RE + \beta \times KL$$

C-VAE

- Conditional VAE works is an extension of VAE (see Sohn et al, 2015)
- In the encoder and decoder we condition on some attribute of the data
- The samples are taken from conditional distributions

Obtaining the synthetic data

- We sample from the latent variables
- This can give more extreme values than those in the historical data

To produce a 10-day 99.9% VaR: sample 50,000 ten-day results. The VaR is the 50th worst one

To produce a 10-day 99.9% ES: average the 50 worst 10-day movements

The Stylized facts

- Excess kurtosis
- Stochastic volatility
- Volatility clustering
- Autocorrelation of squared returns
- Correlation of return and volatility

Research shows that VAE handles the stylized facts well (Buehler et al, 2020 and Zhang, 2021) but some conditioning on attributes of the most recent data such as volatility

Jumps

- Before we backtested the use of synthetic data for stress testing we decided to test how well synthetic data handles jumps

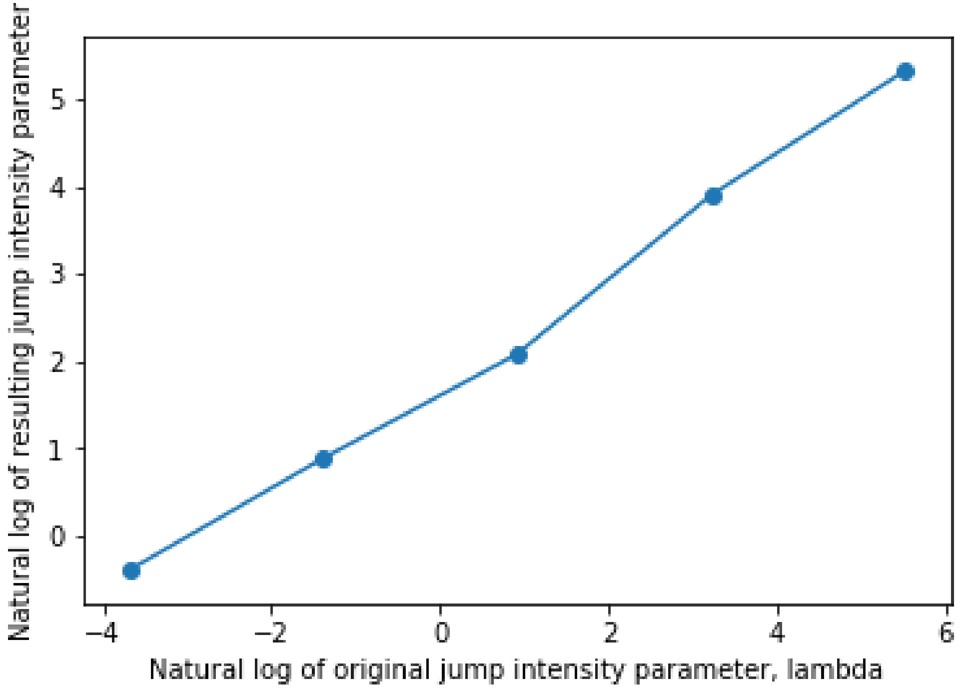
Our approach:

- Generate “historical” data using Merton’s mixed jump diffusion model with varying jump frequency
- Observe the jump frequency in the generated synthetic data

Detecting jumps in the synthetic data

- We used an approach suggested by Lee and Mykland (2008)
- This compares the daily change with a volatility measure calculated from the previous 16 days

Summary of results



Backtesting portfolios (4000 days of data)

- S&P 500
- USD/CAD exchange rate
- Spot price of gold
- ETF on oil
- ETF on VIX
- Three portfolio involving S&P 500 ETF and the sale of call options CC(0.1), CC(0.15), and CC(0.2)

Results

- We produced VaR and ES for the loss over the next 10 days using 99.9% and 99% confidence levels and the most recent 250 days of data
- When we used pure VAE there were too many exceptions
- C-VAE, conditioning on the volatility calculated over the previous 10 days, gave much better results

VaR with 99% confidence levels (Expected number of exception is 40)

Dataset	Number of Exceptions	p-value
S&P 500	44	0.525
USD/CAD	30	0.112
Gold	51	0.080
Oil	33	0.266
Vix	29	0.080
CC(0.1)	43	0.634
CC(0.15)	39	0.874
CC(0.2)	34	0.340
All	**303**	**0.340**

VaR with 99.9% confidence levels (Expected number of exception is 4)

Dataset	Number of Exceptions	p-value
S&P 500	5	0.609
USD/CAD	2	0.453
Gold	4	1.000
Oil	3	1.000
Vix	1	0.202
CC(0.1)	3	1.000
CC(0.15)	1	0.202
CC(0.2)	2	0.453
All	**21**	**0.051**

Expected shortfall

- For expected shortfall we use the Acerbi and Szekely z-statistic which compares the loss levels that give rise to exceptions with the expected shortfall estimate.
- A positive number indicates that loss levels are on average less than the expected shortfall estimate
- A negative number indicates that loss levels are on average greatere than the expected shortfall estimate

Expected shortfall results.

Dataset	Acerbi/Szekely z-statistic (99.9% confidence)	Acerbi/Szekely z-statistic (99% confidence)
S&P 500	0.41	0.43
USD/CAD	0.37	0.33
Gold	0.34	0.35
Oil	0.26	0.39
Vix	-0.57	0.09
CC(0.1)	0.30	0.30
CC(0.15)	0.38	0.28
CC(0.2)	0.05	0.26

Conclusions

- VAE can handle jumps reasonably well
- C-VAE seems to produce good estimates for VaR and ES for our data
- The results are on the conservative side (i.e. VaR and ES tend to be too high)
- Synthetic data is potentially useful for regulators. They can get 10-day risk measures with a high confidence level from the most recent 250 days of data.
- It is also a potentially useful stress testing tool for financial institutions