Stock Market Simulator using HMM and WGAN-GP[1]

Riasat Ali ISTIAQUE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Physical and Mathematical Sciences
College of Science

Joint work with PUN Chi Seng and YONG Yung Sin

SIAM Conference on Financial Mathematics and Engineering

6-9 June 2023

[1]Supported by Ministry of Education, Singapore (MOE-T2EP20220-0013)

Motivation

The need for synthetic financial data [Ass20] is many-fold:

- Lack of historical data.
- Synthesize new samples that are related to but cannot be mapped back to the real data.
- Machine learning models need vast amounts of training data.

This problem has been attacked in many ways:

- Simple statistical models [Bol86].
- Agent-based models [LeB06].
- Deep learning models
  - Generative adversarial models [TCT19].
  - Signature-based models [Bue+20].

Objectives

We want a data-driven approach to construct portfolios that:

- consists of simulated multivariate financial data, that
- fulfils well known empirical statistical properties, and
- showed superior risk measurement

over traditional model-based approaches, when applied to Value-at-Risk (VaR) calculations.

Assumptions

- There are 6 markets that are governed by 4 regimes, which changes daily.
- While the regimes are hidden, we can observe the prices of the markets.
- The distribution of the stock market specifically, is dependent on the regime prevalent on the day, as well as the day before.
- Thus there are 4 distributions, 1 for each regime, to learn, over the choice of time interval.
- Once learned, we should be able to use this tool to generate authentic synthetic stock market data, over the choice of time interval.

Overview of HMM-WGAN

We needed to model which days are (probably) favourable for stocks w.r.t. other asset classes
Classification Problem: no 'true' label for learning and time-series not necessarily independent.

- Use a Hidden Markov Model (HMM) to classify the days and learn the parameters of the model.

We specify 4 classes, called 'Market Painters' (MP), which we assume generates the stock returns that we observe daily.

- Use WGAN-GP to learn the distribution of each MP for each day, then use the sequence of generators to generate sample paths for the stocks.

Phase 1: Hidden Markov Model (HMM)

Intractability of computations means that we have to make simplifying assumptions like the Markov property:

$$\mathbb{P}(x_t|x_{t-1},...,x_0) = \mathbb{P}(x_t|x_{t-1}) \tag{1}$$

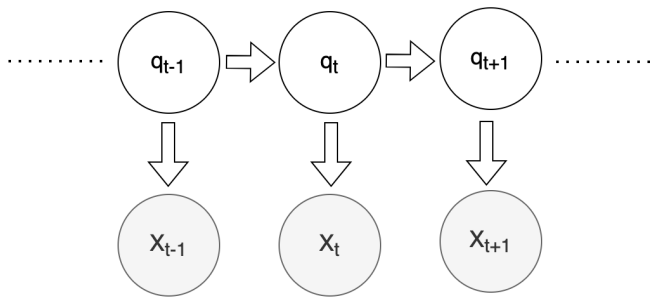


Figure: The latent variables, q_t (MP states), are dependent, not the observations, $\mathbf{X}_t$ (asset prices), themselves.

We apply the Baum-Welch [BP66] (EM) algorithm to obtain the parameters that best fits the in-sample observations and use those parameters to classify out-of-sample observations.

Data Preparation for HMM

The MPs are identified through the use of HMM based on a set of observable exogenous features.

Asset Class	**Representative Index**
Equities	S&P 500 TR Index
Commodities	DBIQ Optimum Yield Diversified Commodity Index TR
Corporate Credit	Bloomberg US Corporate TR Value Unhedged USD
Emerging Market (EM) Credit	J.P. Morgan Emerging Markets Bond Index Global Core
Nominal Bonds	ICE U.S. Treasury 20+ Year TR Index
Inflation-linked (IL) Bonds	Bloomberg US Treasury Inflation Notes TR Index Value Unhedged USD

Table: Indices chosen to represent the asset classes, from Jan '05 - Dec '21; training data used from Jan '05 - Jan '10 $\sim$ 32% of the total data.

HMM Outcome

		Market Painter at $t+1$, q_{t+1}			
		1	**2**	**3**	**4**
Market	1	96.1	3.9	0.0	0.0
Painter	2	19.0	79.7	1.3	0.0
at t, q_t	3	0.0	1.2	94.7	4.2
	4	0.0	0.0	15.9	84.1

Table: The transition probability matrix from the trained HMM (%).

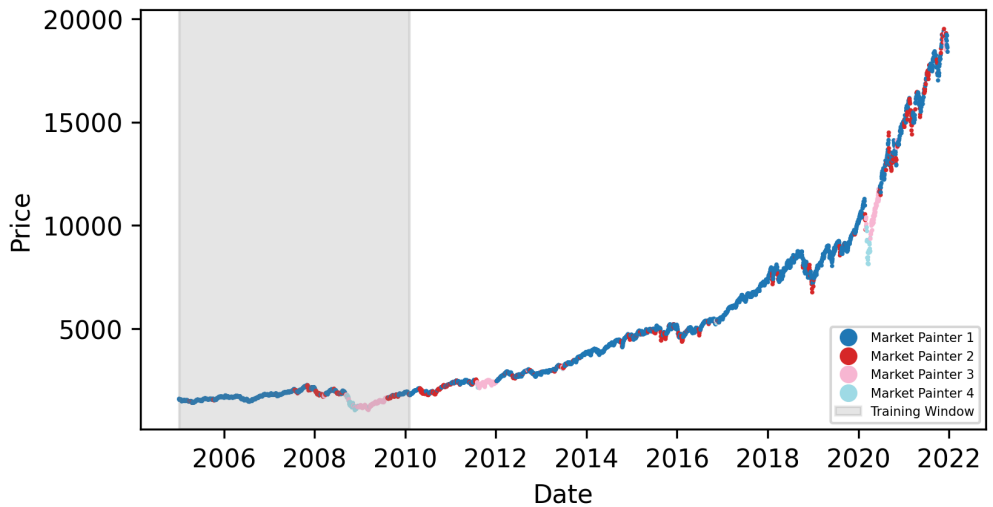


Figure: The daily MPs from Jan '05 - Dec '21 for XNDX.

Phase 2: Wasserstein GAN with Gradient Penalty (WGAN-GP)

The original GAN [Goo+14] introduced was found to be based on the KL-Divergence:

$$\min_G \max_D \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[\log D(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[\log(1 - D(\tilde{\boldsymbol{x}}))] \tag{2}$$

A statistical distance that was proposed is the Wasserstein metric which has a dual [ACB17]:

$$\min_G \max_C \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[C(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[C(\tilde{\boldsymbol{x}})] \tag{3}$$

A further improvement made is an addition of the gradient penalty (GP) component [Gul+17]:

$$\min_G \max_C \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_r}[C(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_g}[C(\tilde{\boldsymbol{x}})] + \gamma \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}}\sim\mathbb{P}_{\hat{\boldsymbol{x}}}}[(||\nabla_{\hat{\boldsymbol{x}}} C(\hat{\boldsymbol{x}})||_2 - 1)^2] \tag{4}$$

Data Preparation for WGAN-GP

- For the second dataset we randomly extracted four sets of datasets comprising of 8, 16, 32 and 64 companies within the NASDAQ 100 Total Return Index (XNDX).
- Once the daily market painter, q_t, has been identified, it is then paired up with the corresponding daily financial data forming $(q_t, \mathbf{X}_t)$ pairs, which forms the dataset that is to be used to train the WGAN-GP models.

WGAN-GP Outcome (Loss Profile)

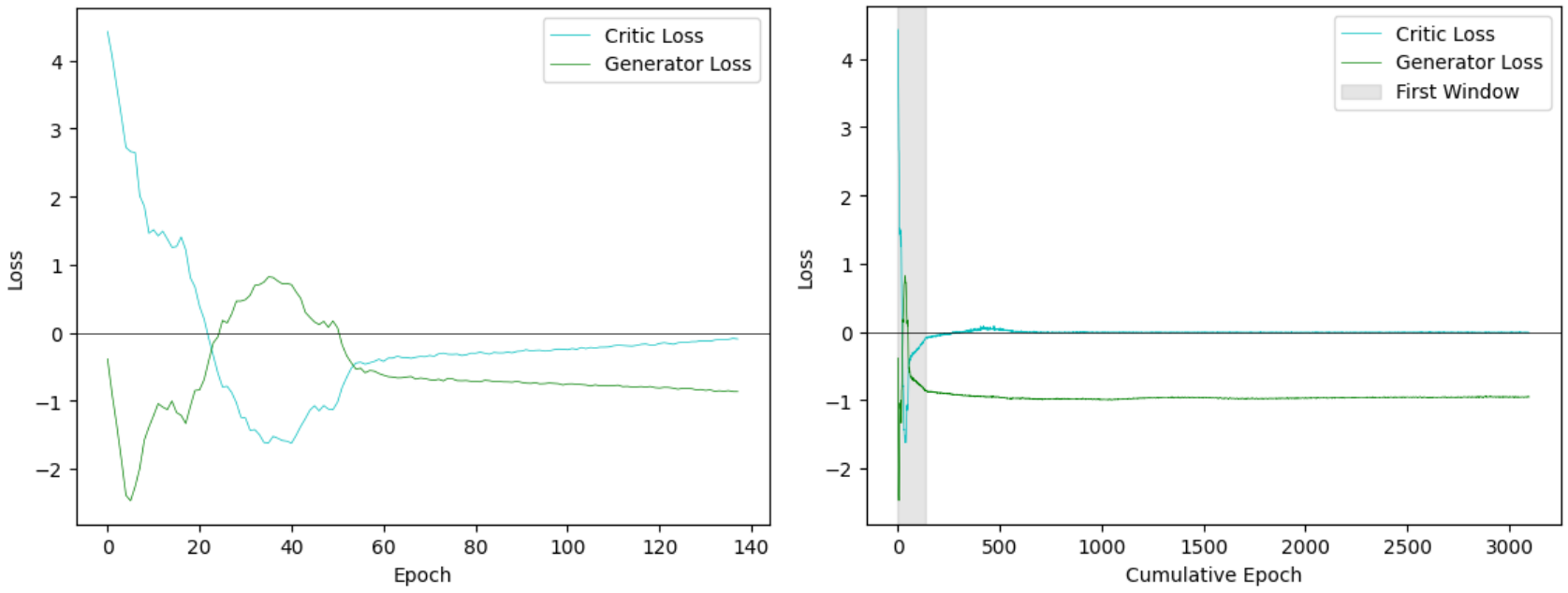


Figure: First window (L) and, cumulative window losses (R) for MP2.

WGAN-GP Outcome (Simulated Returns)

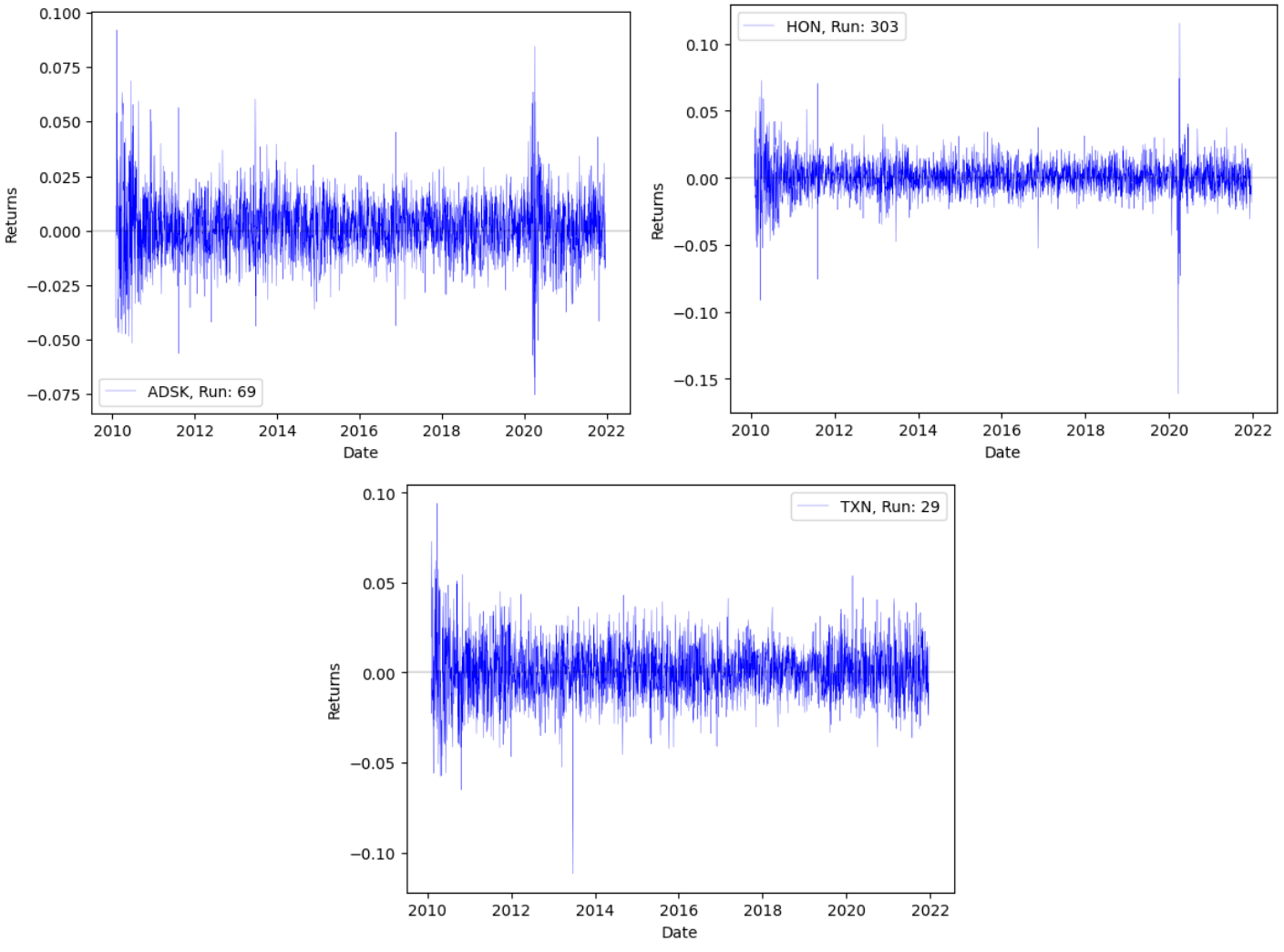


Figure: Out-of-sample simulated returns for 3 different stocks.

Empirical Properties

- We use eight statistical properties that can help to validate the simulated financial data.
- The stock market is shown to exhibit univariate statistical properties as studied by [Con01; JPR07; Cha+11] and multivariate statistical properties as consolidated by [Mar20].
- Before the daily simulated financial data is used for any financial application, the simulated financial data are validated to ensure that it fulfils such statistical properties.

Property 1: Linear Unpredictability

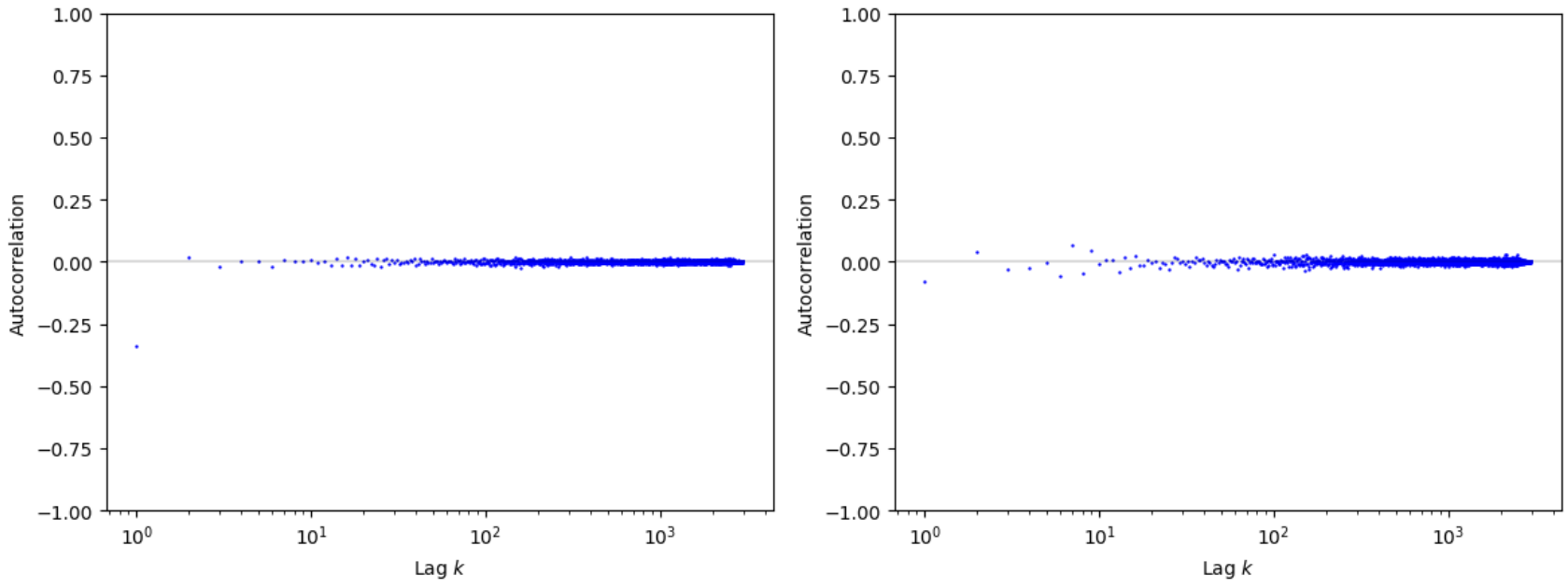


Figure: Training data (L), generated data (R). Widely known observation [Fam70; Pag96; Tan03; Cha+11] where it can be safely assumed that the autocorrelations to be zero for time lag $\geq$ 15 minutes [CPB97].

Property 2: Fat-Tailed Distribution

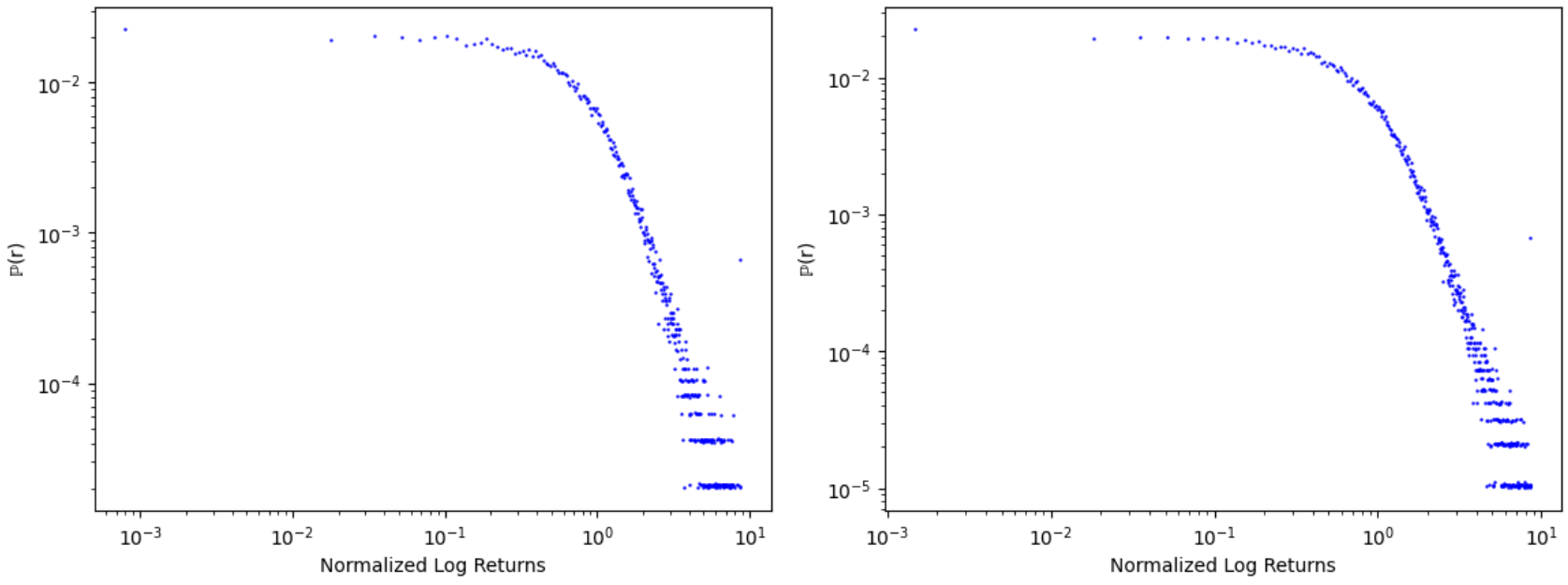


Figure: Training data (L), generated data (R). The probability distribution $P(r)$ is observed to consistently follow a power-law decay in the tails, where the value of α is typically between 3 to 5 [TCT19].

Property 3: Volatility Clustering

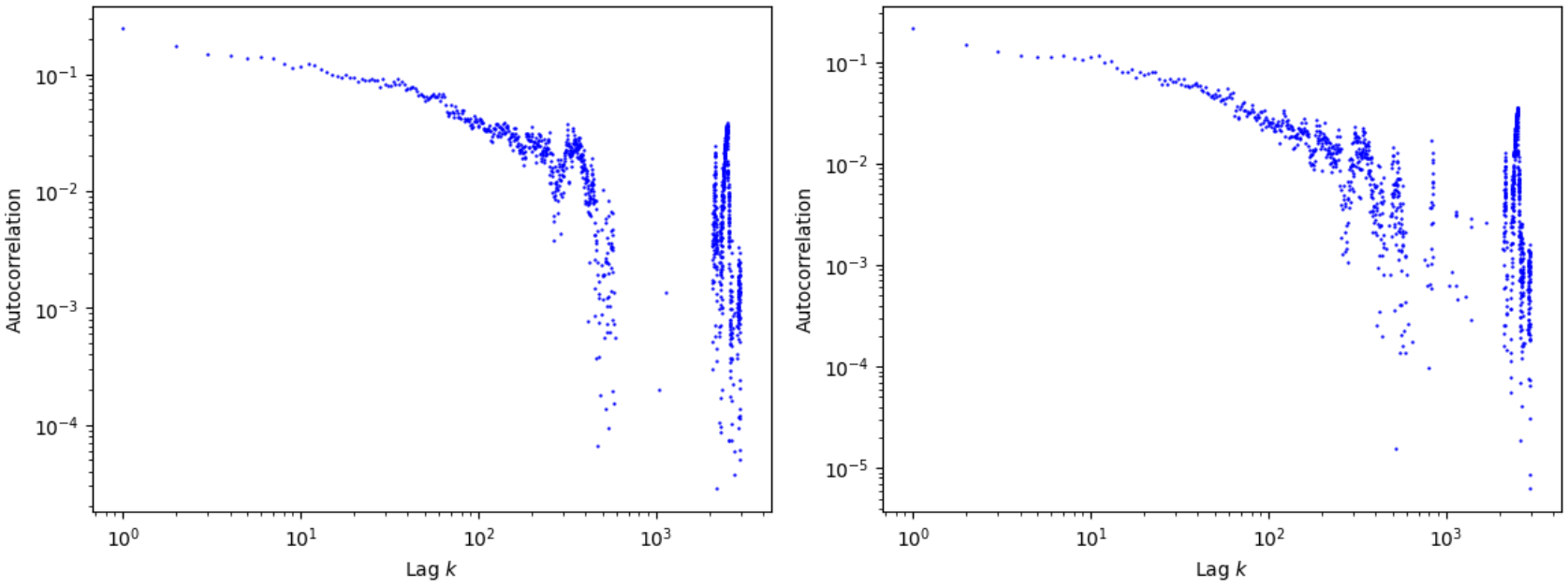


Figure: Training data (L), generated data (R). This autocorrelation slowly decays and remains significantly positive over a long period of time [Con01].

Property 4: Leverage Effect

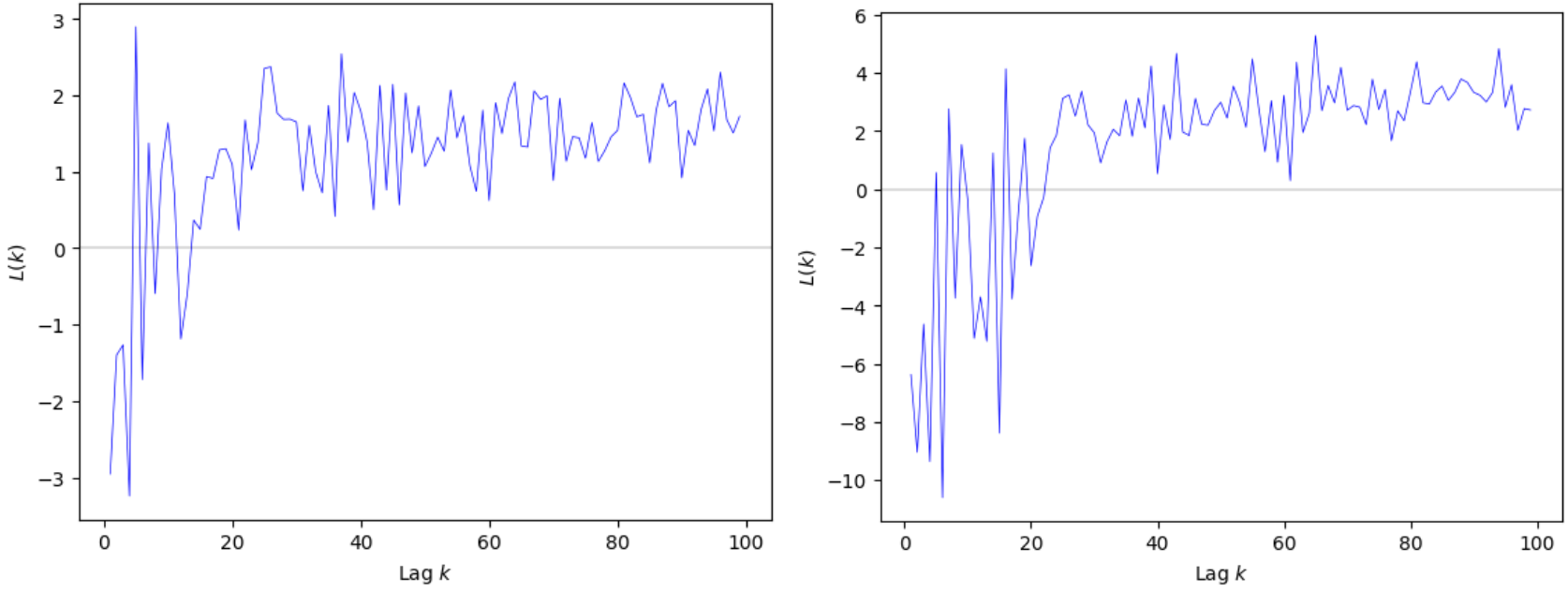


Figure: Training data (L), generated data (R). Negative returns induces higher volatility while positive returns may lead to stable stock prices [BMP01; PM03]. In contrast to other statistical properties, this property is market dependent [Qiu+06], and has a negative value for $1 \leq k \leq 10$, followed by an exponential decay [BMP01].

Property 5: Gain-Loss Asymmetry

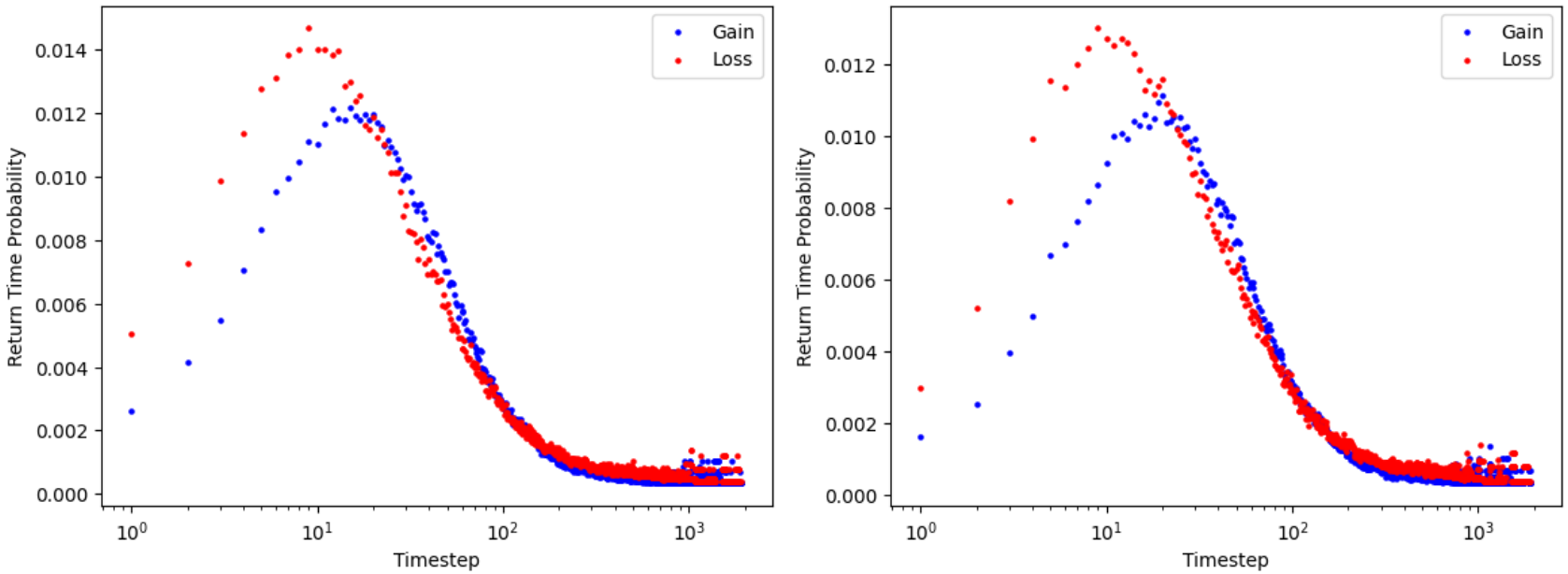


Figure: Training data (L), generated data (R). The peak of positive returns comes after the peak of negative returns, indicating the presence of asymmetry in price change [TCT19].

Property 6: Marchenko-Pastur Distribution

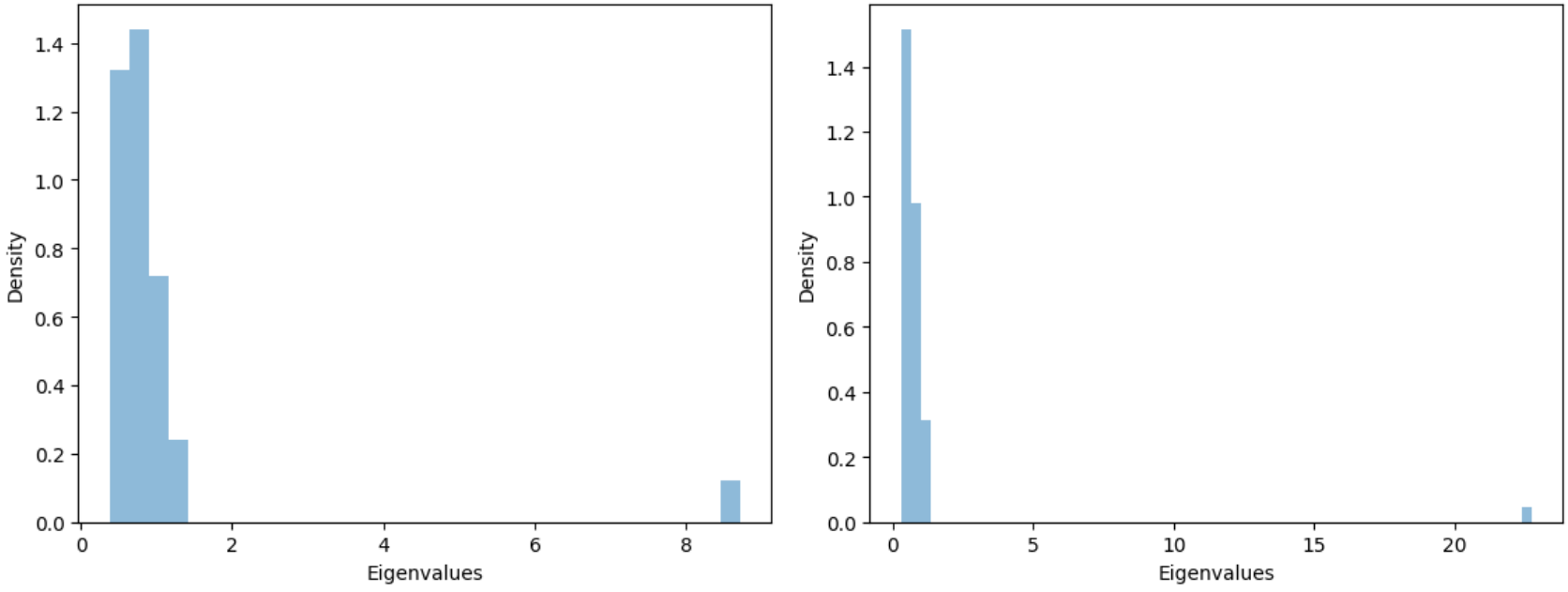


Figure: Training data (L), generated data (R). Investigation into both random matrices and empirical financial correlation matrices revealed that both have a very similar structure where most eigenvalues are low in value and an eigenvalue which is observed to be an outlier [Lal+00].

Property 7: Hierarchical Structure of Correlations

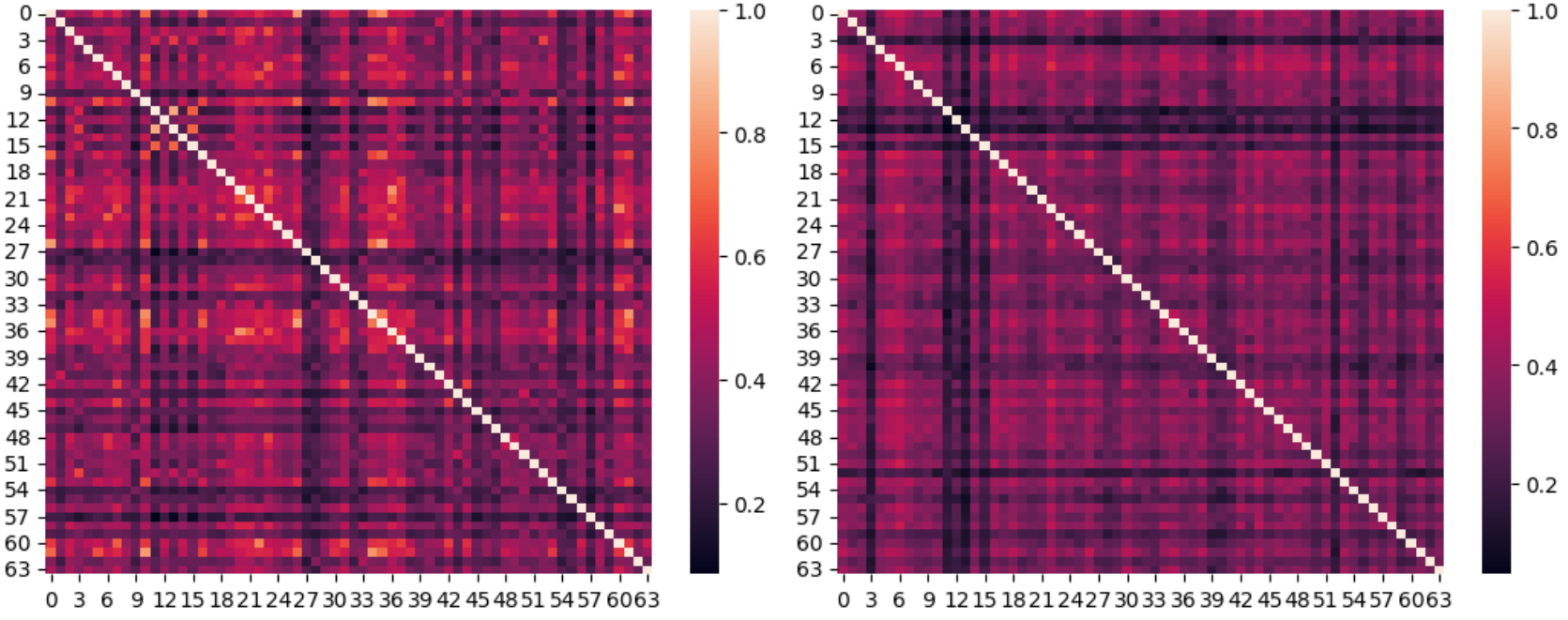


Figure: Training data (L), generated data (R). The correlation matrix was shown to have observable hierarchical structure [Man99]. Despite having random properties individually, the correlations in the generated time-series' gives it some form of meaningful structure, as corroborated with empirical time-series.

Property 8: Scale-Free Property of the MST

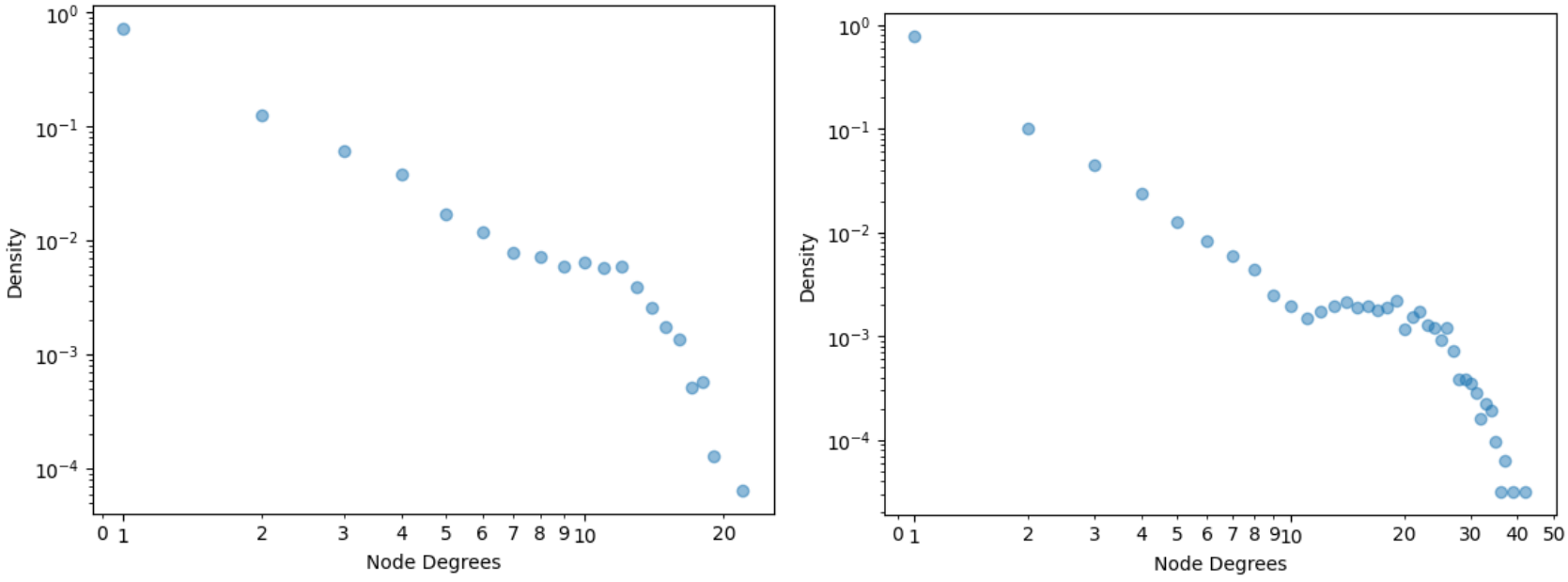


Figure: Training data (L), generated data (R). By looking at the networks formed by prices correlations matrices, the empirical datasets can be measured so as to validate proposed models using the distance metric in [Cal+04]. The nearest-neighbour single-linkage cluster algorithm method is employed. The corresponding empirical MSTs were shown to have a few nodes with high degrees that follow a power law - i.e. a scale-free property, before breaking down.

Optimal Market Exposure

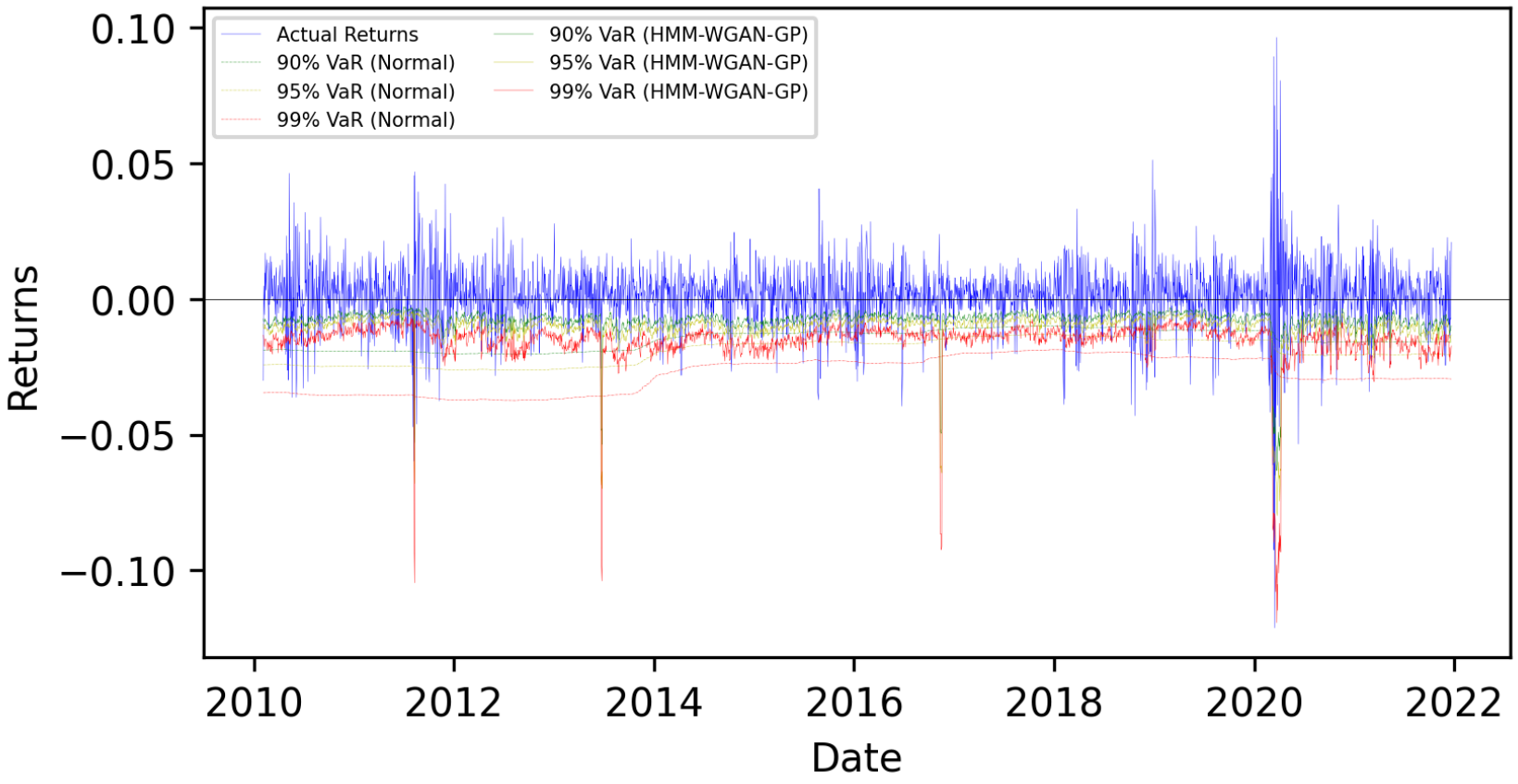
Financial institutions are interested in minimizing both market overexposure and opportunity costs in deploying capital.

- Value-at-Risk (VaR) computed at the 90/95/99th% for four (8, 16, 32, 64 assets) equally-weighted (EW) portfolios [DGU07] of simulated stocks.
- VaR_α^{MC} is the value at the $(1-\alpha)$th-percentile of the 500 MC simulations of the daily simulated EW portfolios.
- VaR_α^{normal} based on a normal distribution assumption on the actual historical data, with the moving window approach:

$$VaR_\alpha^{normal} = -(\mathbf{W}^T \boldsymbol{\mu} + Z_\alpha \sqrt{\mathbf{W}^T \boldsymbol{\Sigma} \mathbf{W}}) \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^n$ (vector of stock weights), $\boldsymbol{\mu} \in \mathbb{R}^n$ (average return of the individual stocks in the past $\varphi = 5$ years), $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ (covariance matrix of the stocks), and Z_α (Z-score).

Portfolio of 64 Simulated Stocks



Ideal Scenarios

The most ideal scenario would be that the VaR_α^{MC} is computed to be between the VaR_α^{normal} level and the actual returns that panned out over time.

α	8 Stocks	16 Stocks	32 Stocks	64 Stocks
99%	84.13%	86.39%	89.30%	91.70%
95%	79.24%	79.61%	84.07%	85.75%
90%	74.71%	75.02%	78.83%	81.84%

Table: Data reported is the proportion of windows our model's VaR level stayed within the bounds.

	8 Stocks		16 Stocks		32 Stocks		64 Stocks	
α	VaR^{MC}	VaR^{N}	VaR^{MC}	VaR^{N}	VaR^{MC}	VaR^{N}	VaR^{MC}	VaR^{N}
99%	2.21%	3.41%	1.46%	2.92%	1.82%	3.03%	1.82%	2.95%
95%	1.70%	2.50%	1.19%	2.14%	1.41%	2.23%	1.40%	2.17%
90%	1.47%	2.04%	1.06%	1.76%	1.24%	1.82%	1.21%	1.76%

Table: Data reported is the average delta of the daily returns between actual returns and VaR level.

Non-Ideal Scenarios

We now look at non-ideal scenarios where the the proportion of windows where the actual returns breached VaR_α^i levels for both our model and normal were computed.

	8 Stocks		16 Stocks		32 Stocks		64 Stocks	
α	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N
99%	6.79%	2.13%	11.55%	1.79%	7.23%	2.13%	6.96%	2.26%
95%	12.96%	4.25%	18.74%	4.15%	13.13%	4.46%	13.00%	4.32%
90%	17.49%	6.85%	3.40%	6.96%	18.00%	6.99%	16.78%	6.96%

Table: Data reported is the proportion of windows where VaR^{MC} was breached vs. the proportion of windows where VaR^N was breached.

	8 Stocks		16 Stocks		32 Stocks		64 Stocks	
α	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N	VaR^{MC}	VaR^N
99%	0.78%	1.14%	0.71%	1.25%	0.81%	1.21%	0.89%	1.11%
95%	0.83%	1.14%	0.73%	1.01%	0.83%	1.10%	0.83%	1.12%
90%	0.86%	1.09%	0.74%	0.92%	0.81%	1.02%	0.85%	1.01%

Table: Data reported is the average delta of the daily returns between actual returns and VaR level.

In Conclusion

- The normal distribution assumption is consistent and does not change significantly, our framework is shown to be able to react more quickly to volatile periods.
- We were able to show that we can generate synthetic multivariate data that fulfils empirical qualities.
- However recent studies show that from a theory perspective, when we employ WGAN-GP, we are indeed only ensuring that the target distribution's first moment is being 'matched' [Bińkowski+18].
- Thus we can think of amending the critic in such a way as to match all the infinite (sample) moments of the target distribution - Maximum Mean Discrepancy [Gretton+12]:

$$MMD(\mathbb{P}_r, \mathbb{P}_g; \mathcal{H}) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_r} [f(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g} [f(\tilde{\boldsymbol{x}})] \quad (6)$$

Bibliography

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[Ass20] Samuel Assefa. "Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls". In: *SSRN Electronic Journal* (Oct. 2020).

[Bińk+18] Mikołaj Bińkowski et al. "Demystifying MMD GANs". In: *International Conference on Learning Representations*. 2018.

[BMP01] J P Bouchaud, A Matacz, and M Potters. "Leverage effect in financial markets: the retarded volatility model". en. In: *Physical Review Letters* 87.22 (Nov. 2001), p. 228701.

Bibliography

[Bol86] Tim Bollerslev. "Generalized autoregressive conditional heteroskedasticity". en. In: *Journal of Econometrics* 31.3 (Apr. 1986), pp. 307–327.

[BP66] Leonard E. Baum and Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". In: *The Annals of Mathematical Statistics* 37.6 (Dec. 1966), pp. 1554–1563.

[Bue+20] Hans Buehler et al. "A Data-Driven Market Simulator for Small Data Environments". In: *SSRN Electronic Journal* (2020).

[Cal+04] Guido Caldarelli et al. "Emergence of Complexity in Financial Networks". In: *Complex Networks*. Springer Berlin Heidelberg, Aug. 2004, pp. 399–423.

Bibliography

[Cha+11] Anirban Chakraborti et al. "Econophysics review: I. Empirical facts". en. In: *Quant. Finance* 11.7 (July 2011), pp. 991–1012.

[Con01] R Cont. "Empirical properties of asset returns: stylized facts and statistical issues". In: *Quant. Finance* 1.2 (Feb. 2001), pp. 223–236.

[CPB97] Rama Cont, Marc Potters, and Jean-Philippe Bouchaud. "Scaling in Stock Market Data: Stable Laws and Beyond". In: *Scale Invariance and Beyond*. Springer Berlin Heidelberg, 1997, pp. 75–85.

[DGU07] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. "Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?" In: *Review of Financial Studies* 22.5 (Dec. 2007), pp. 1915–1953.

Bibliography

[Fam70] Eugene F. Fama. "Efficient Capital Markets: A Review of Theory and Empirical Work". In: *The Journal of Finance* 25.2 (May 1970), p. 383.

[Goo+14] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014.

[Gre+12] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773.

[Gul+17] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

Bibliography

[JPR07] Eric Jondeau, Ser-Huang Poon, and Michael Rockinger. "Statistical Properties of Financial Market Data". In: *Financial Modeling Under Non-Gaussian Distributions*. London: Springer London, 2007, pp. 7–32. ISBN: 978-1-84628-696-4.

[Lal+00] Laurent Laloux et al. "Random Matrix Theory And Financial Correlations". In: *International Journal of Theoretical and Applied Finance* 03.03 (July 2000), pp. 391–397.

[LeB06] Blake LeBaron. "Agent-based Computational Finance". In: *Handbook of Computational Economics*. Ed. by Leigh Tesfatsion and Kenneth L. Judd. Vol. 2. Handbook of Computational Economics. Elsevier, 2006. Chap. 24, pp. 1187–1233.

Bibliography

[Man99] R.N. Mantegna. "Hierarchical structure in financial markets". In: *The European Physical Journal B* 11.1 (Sept. 1999), pp. 193–197.

[Mar20] Gautier Marti. "CORRGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020.

[Pag96] Adrian Pagan. "The econometrics of financial markets". In: *Journal of Empirical Finance* 3.1 (May 1996), pp. 15–102.

[PM03] Josep Perelló and Jaume Masoliver. "Random diffusion and leverage effect in financial markets". en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67.3 Pt 2 (Mar. 2003), p. 037102.

Bibliography

[Qiu+06] T Qiu et al. "Return-volatility correlation in financial dynamics". en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 73.6 Pt 2 (June 2006), p. 065103.

[Tan03] Peter Tankov. *Financial Modelling with Jump Processes*. Chapman and Hall/CRC, Dec. 2003.

[TCT19] Shuntaro Takahashi, Yu Chen, and Kumiko Tanaka-Ishii. "Modeling financial time-series with generative adversarial networks". en. In: *Physica A: Statistical Mechanics and its Applications* 527.121261 (Aug. 2019), p. 121261.