FINX Lab Winter Internship

Tail GAN
Learning to simulate
Tail risk scenario

장윤수

Department of Industrial Engineering
Hanyang University

2025.02.26

FINANCIAL INNOVATION
& ANALYTICS LAB.

CONTENTS

1. Introduction

Regulations

- Risk estimation has become increasingly important in finance.
  - FRTB(Fundamental Review of the Trading Book) regulates the amount of capital banks ought to hold against market risk exposures.
  - FRTB particularly revisits and emphasize the use of VaR and ES as a measure of under stress.

Limitations

- AIQN(Autoregressive Implicit Quantile Network)
  - The idea of incorporating quantile properties into simulation model[1].
  - Quantile divergence adopted in AIQN is an average performance which provides no guarantees for the tail risks.

- GANs using EVT(Extreme Value Theory)
  - The idea of using GANs conditioned on the statistics of extreme events to generate samples using EVT[2].
  - Tail GAN does not rely on parametrization of tail probabilities.

2. Score Function & Data

Joint Elicitability

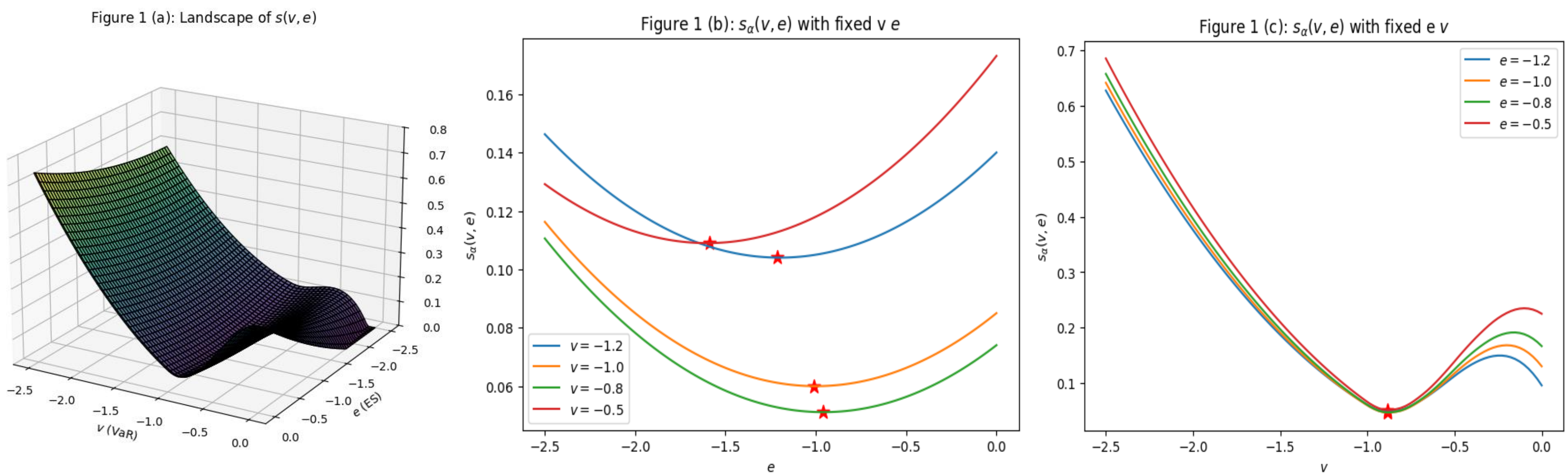
- Joint Elicitability of VaR and ES
  - Whereas ES is not elicitable, VaR at level a ∈ $(0,1)$ is elicitable for random variables with a unique a-quantile.
  - However, ES is elicitable in the sense of that the pair (VaR, ES) is jointly elicitable.

- Score function
  - In Tail GAN, this paper use a specific form of the score function[3].

$$S_\alpha(v,e,x) = \frac{W_\alpha}{2}(\mathbb{1}_{\{x\le v\}} - \alpha)(x^2 - v^2) + \mathbb{1}_{\{x\le v\}} e(v - x) + \alpha e\left(\frac{e}{2} - v\right), \text{ with } \frac{\mathrm{ES}_\alpha(\mu)}{\mathrm{VaR}_\alpha(\mu)} \ge W_\alpha \ge 1.$$

$$H_1(v) = -\frac{W_\alpha}{2}v^2,\ H_2(e) = \frac{\alpha}{2}e^2, \quad \text{with} \quad \frac{\mathrm{ES}_\alpha(\mu)}{\mathrm{VaR}_\alpha(\mu)} \ge W_\alpha \ge 1.$$

Landscape of Score Function

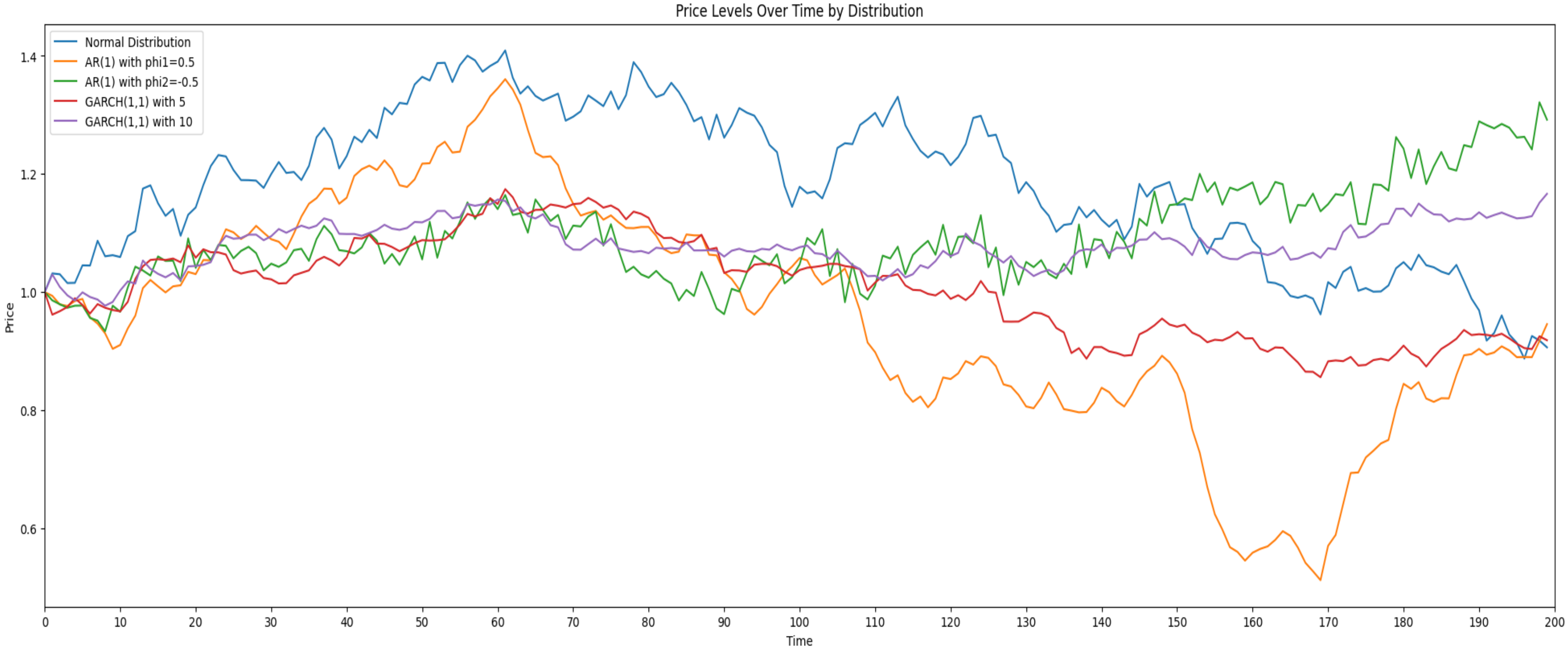


Synthetic data

- Gaussian Distribution
  - $\Delta p_{\{1,t\}} = \mu_{\{1,t\}}$
- AR(1) with ϕ_1=0.5
  - $\Delta p_{\{2,t\}} = \phi_1 \Delta p_{\{2,t-1\}} + \mu_{\{2,t\}}$
- AR(1) with ϕ_2=-0.5
  - $\Delta p_{\{3,t\}} = \phi_2 \Delta p_{\{3,t-1\}} + \mu_{\{3,t\}}$
- GARCH(1,1) with ν_1=5
  - $\Delta p_{\{4,t\}} = \epsilon_{\{4,t\}} = \sigma_{\{4,t\}} \eta_{\{1,t\}}$ $\quad \sigma_{4,t}^2 = \gamma_4 + \kappa_4 \varepsilon_{4,t-1}^2 + \beta_4 \sigma_{4,t-1}^2, \eta_{1,t} = \frac{u_{4,t}}{\sqrt{v_{1,t}/\nu_1}}$
- GARCH(1,1) with ν_2 =10
  - $\Delta p_{\{5,t\}} = \epsilon_{\{5,t\}} = \sigma_{\{5,t\}} \eta_{\{2,t\}}$ $\quad \sigma_{5,t}^2 = \gamma_5 + \kappa_5 \varepsilon_{5,t-1}^2 + \beta_5 \sigma_{5,t-1}^2, \eta_{2,t} = \frac{u_{5,t}}{\sqrt{v_{2,t}/\nu_2}}$

Synthetic data

- Convert return to price



Strategy

- Buy and Hold(55 PnLs)
  - Buy-and-hold strategy with 50 static portfolios extracted using a random weight matrix: 50 PnLs.
  - Buy-and-hold strategy with 5 assets: 5 PnLs.

- Mean Reversion(5 PnLs)
  - Mean reversion strategy with a rolling window of 10 periods.
  - Long signal when the current price < 10-period moving average × 0.95
  - Short signal when the current price > 10-period moving average × 1.05

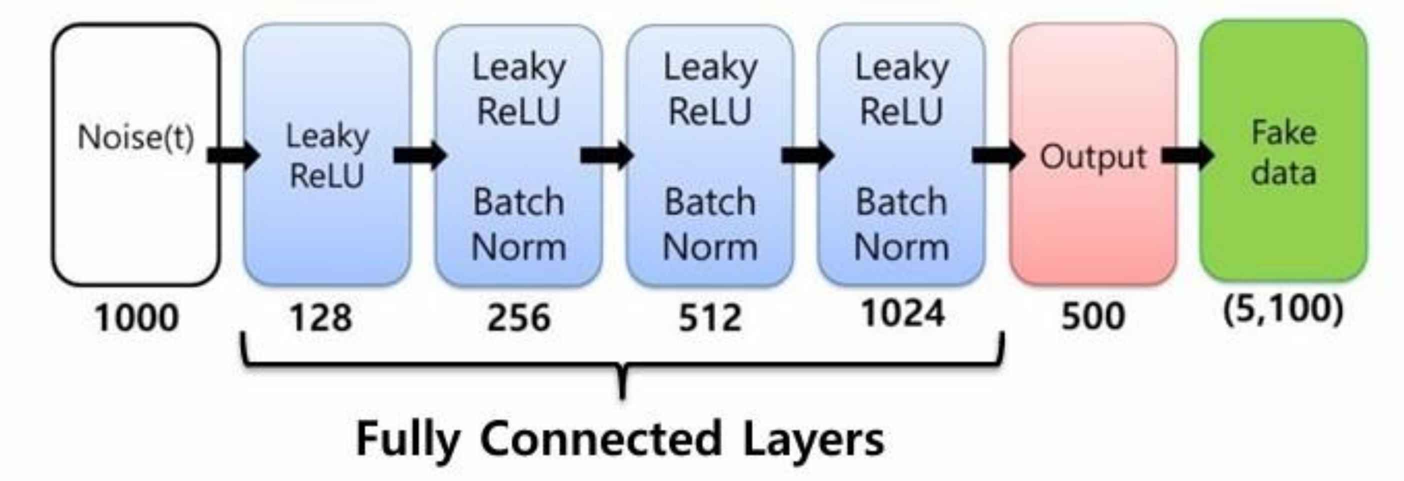
- Trend Following(5 PnLs)
  - Trend following strategy with short window of 5 periods and long window of 10 periods.
  - Long signal when short moving average > long moving average × 1.05
  - Short signal when short moving average < long moving average × 0.95

3. Architecture & Training

Generator

- Input data
  - To model extreme market fluctuations, t-distribution-based sampling is adopted as input for the generator.

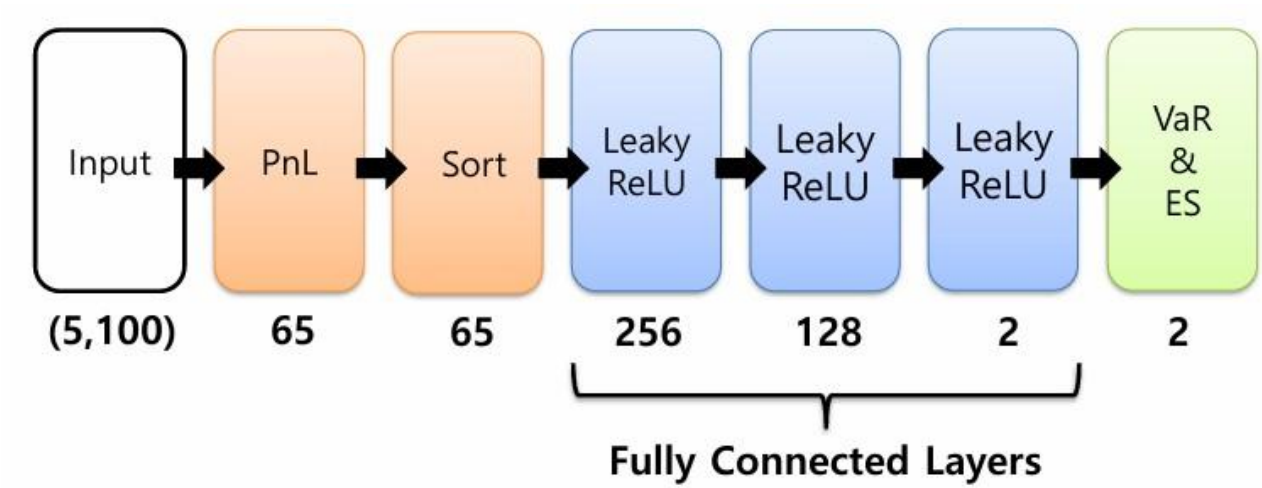
- Layers
  - Batch Normalization (momentum=0.8) and Leaky ReLU (α=0.2) were applied to stabilize training and preserve tail risk characteristics.



Discriminator

- Input data
  - Portfolio returns are soft-sorted using Neural Sort, allowing for differentiable ranking while preserving order information. The sorted returns are then passed through the layers to evaluate VaR and ES.

- Layers
  - The network consists of three fully connected layers with Leaky ReLU (α=0.2) activation, ensuring stable learning of tail risk patterns.



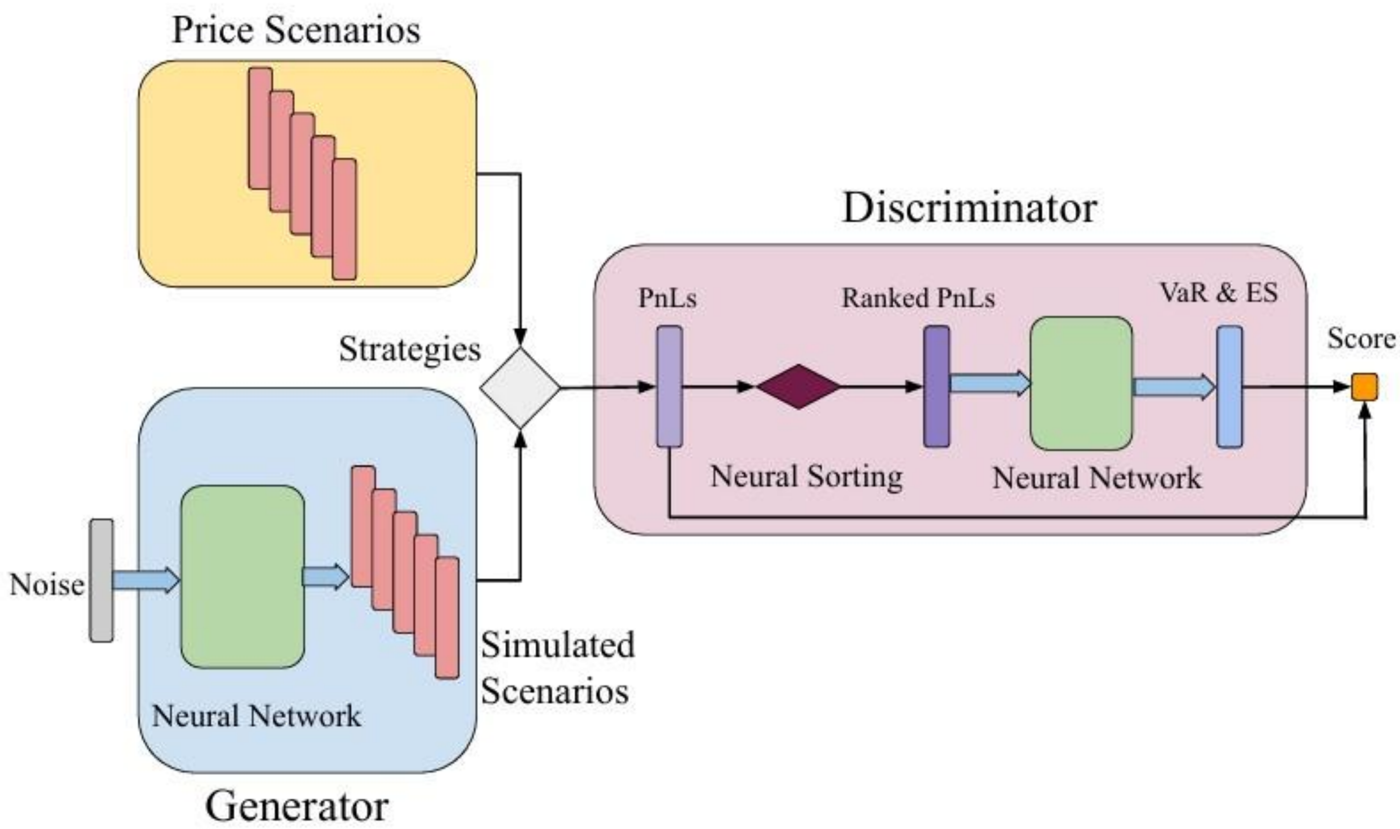
HANYANG UNIVERSITY

Tail GAN Training

- Loss function
  - The generator's loss is computed in the score function and is minimized to improve the generator's ability to produce data that the discriminator evaluates as realistic.
  - The discriminator loss is computed as the difference between the scores of real and generated data, aiming to maximize it.

- Optimizer
  - The Adam optimizer is employed to minimize the generator's loss while maximizing the discriminator's loss.
  - This optimization method is chosen for its adaptive learning rate and momentum-based updates, which help stabilize the adversarial training process.
  - Learning rate for discriminator: 1e-7, Learning rate for generator: 1e-6

Tail GAN Training

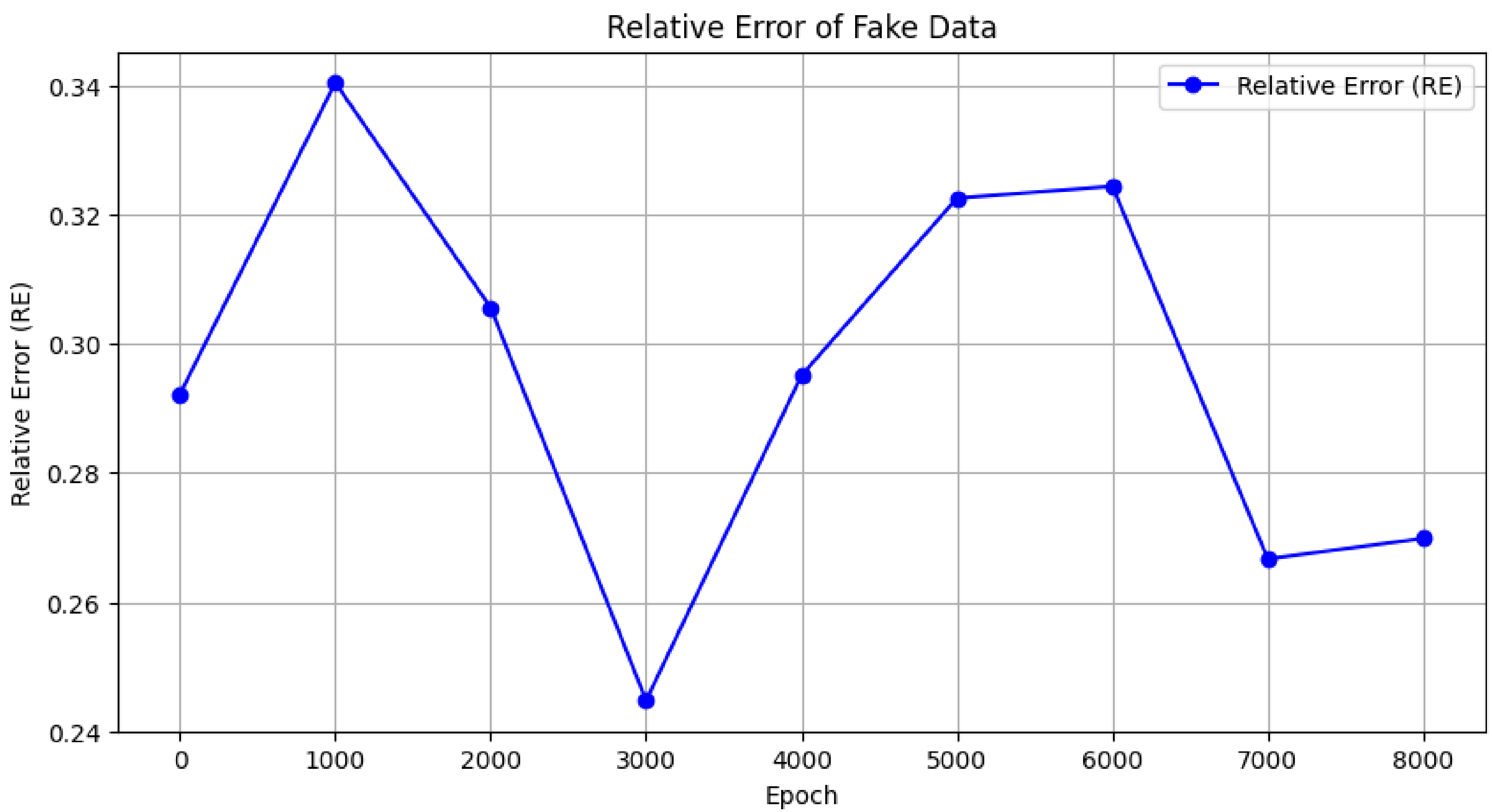


[4]

4. Result & Discussions

Relative Error Across Training Epochs

- Fake data selection
  - The RE reached its lowest value at epoch 3000, after which signs of overfitting became evident.



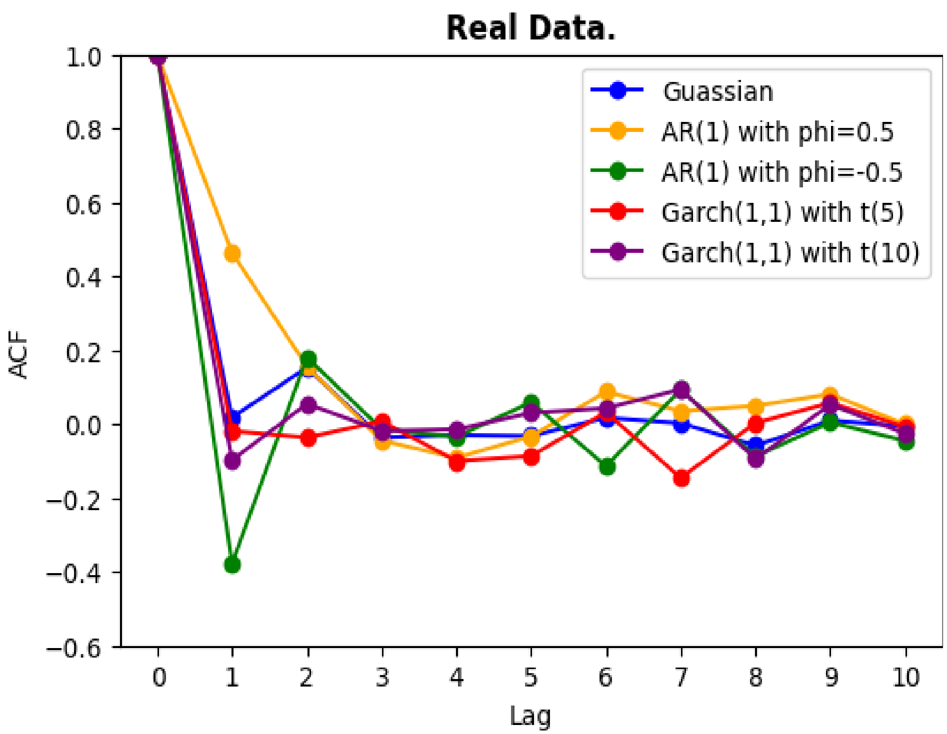
Performance Evaluation

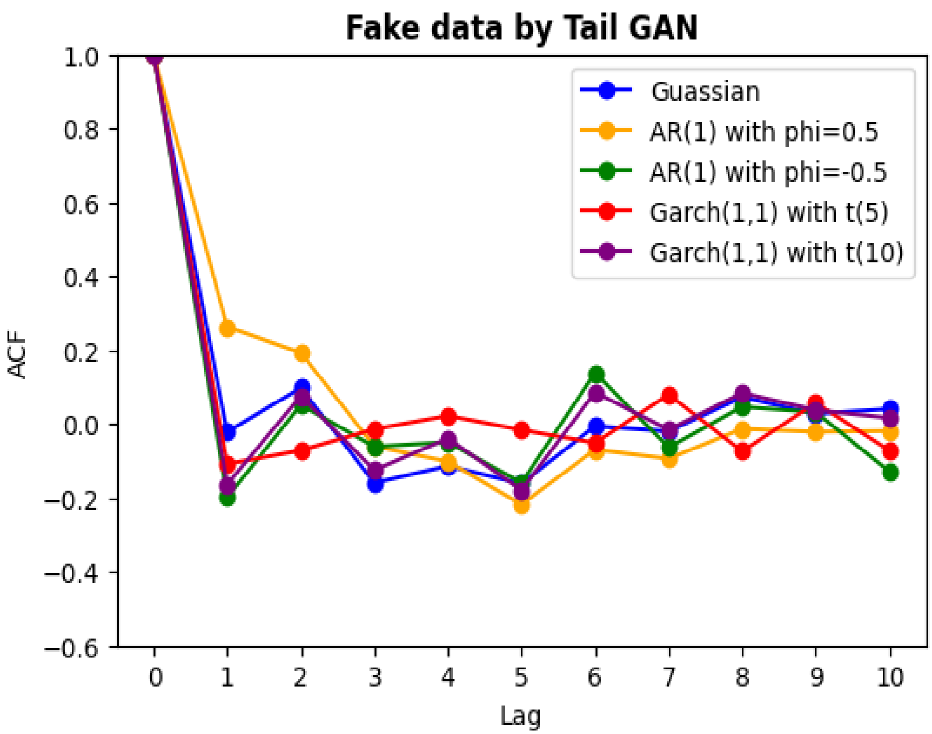
- The following tables show the errors observed in In-Sample and Out-Of-Sample.
  - The relatively higher RE of the fake data is likely attributed to generalization error caused by the reduction in the number of data samples.

	Sample Error(SE)	Relative Error(RE) between Fake data and Real data
Mean	1.3253%	24.4770%
Std Dec	0.8382%	6.3378%

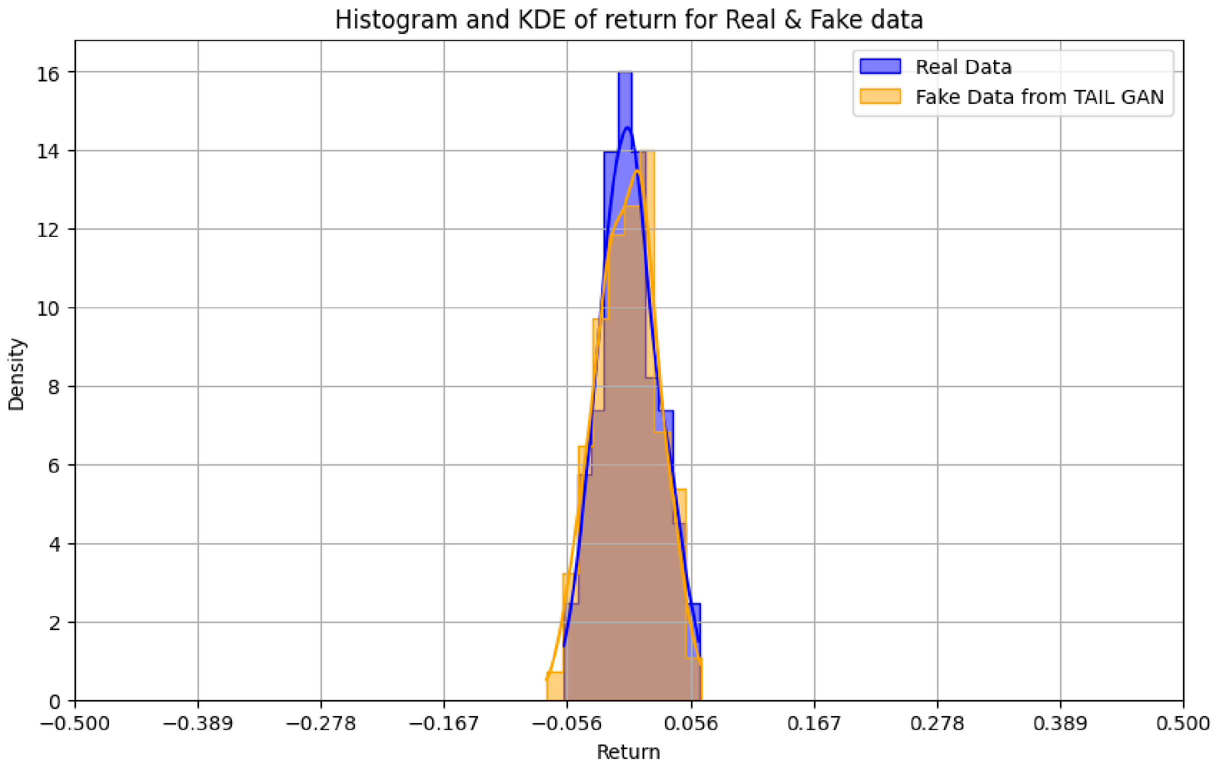
	Relative Error(RE) between Real data and OOS	Relative Error(RE) between Fake data and OOS
Mean	3.1142%	23.6349%
Std Dec	1.3254%	6.2649%

ACF of Real data and Fake data





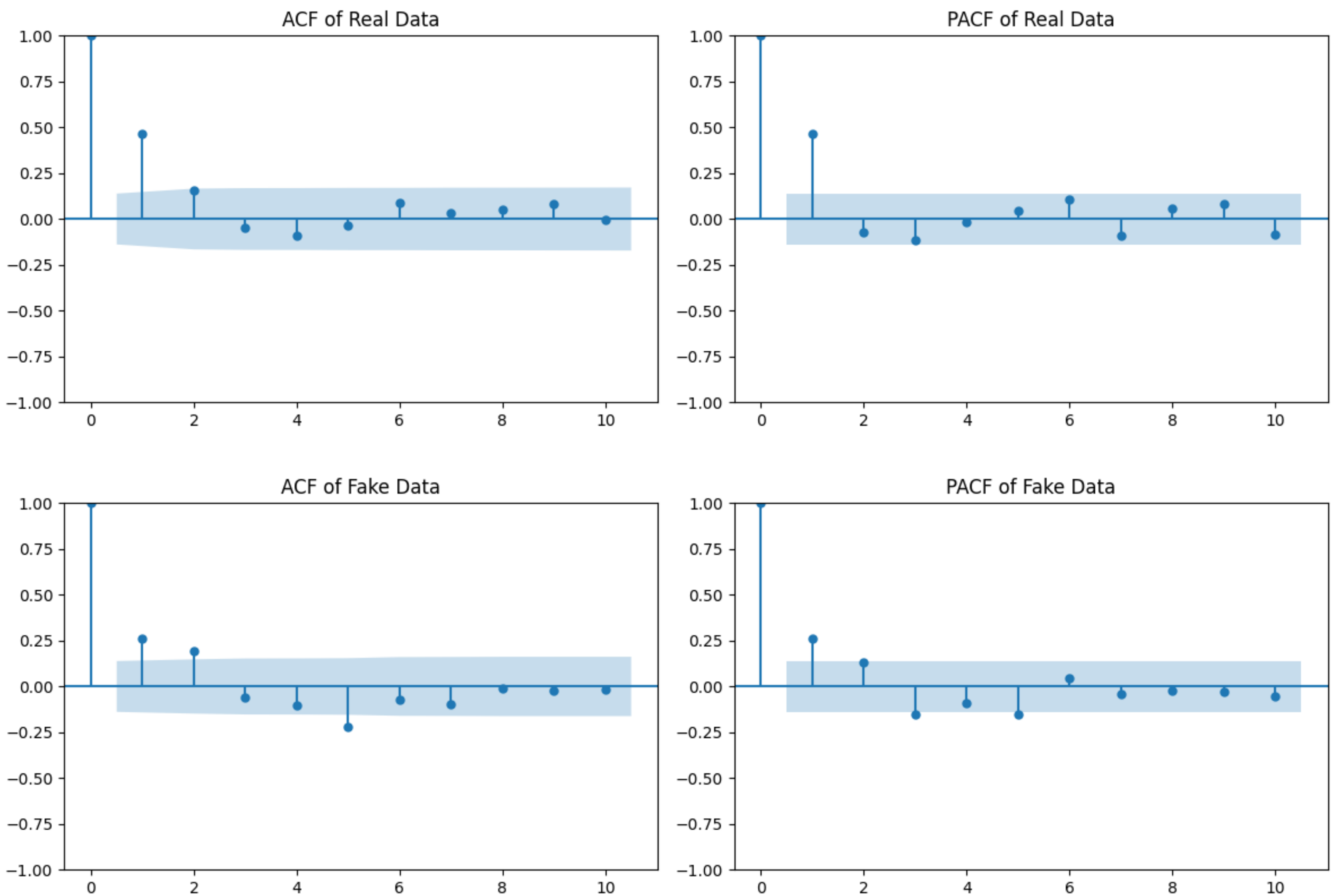
Gaussian Distribution



Data Type	Shapiro-Wilk Statistic	p-value	Normality Assumption
Real Data	0.9917	0.3088	Fail to reject H_0 (Normal)
Fake Data	0.9951	0.7590	Fail to reject H_0 (Normal)

AR(1) with phi= 0.5

- ACF and PACF of real and fake data



AR(1) with phi= 0.5

- ADF Test for real and fake data
  - An ADF test was performed to assess stationarity, and the results confirmed that both datasets are completely stationary.

Data Type	ADF Statistic	p-value	Critical Value(1%)	Critical Value(5%)	Critical Value(10%)	Stationary Assumption
Real Data	-8.4730	0.0000	-3.4636	-2.8762	-2.5746	Reject H_0(Stationary)
Fake Data	-7.4172	0.0000	-3.4638	-2.8763	-2.5746	Reject H_0(Stationary)

AR(1) with phi= 0.5

- Model fit and information criteria
  - The AR(1) model outperformed the AR(2) model in terms of Log-Likelihood, AIC, BIC, and HQIC, indicating a better model fit.

Model Information	Value
Model	AR(1)
No. Observations	200
Log Likelihood	381.111
AIC	-756.222
BIC	-746.342
HQIC	-752.224

Model Information	Value
Model	AR(2)
No. Observations	200
Log Likelihood	380.480
AIC	-752.959
BIC	-739.806
HQIC	-747.635

AR(1) with phi= 0.5

- Statistical significance of model parameters
  - Since the second lag in AR(2) is not statistically significant (p=0.059), and AR(1) provides a strong autoregressive structure with a significant coefficient (p=0.000), AR(1) is considered the more appropriate model.

Parameter	Coefficient	Std Err	z-value	p-value	95% Confidence interval
Constant	0.0017	0.003	0.652	0.514	[-0.003, 0.007]
y.L1	0.2625	0.068	3.843	0.000	[0.129, 0.396]

Parameter	Coefficient	Std Err	z-value	p-value	95% Confidence interval
Constant	0.0014	0.003	0.572	0.567	[-0.004, 0.006]
y. L1	0.2278	0.070	3.233	0.001	[0.090, 0.366]
y. L2	0.1333	0.070	1.892	0.059	[-0.005, 0.271]

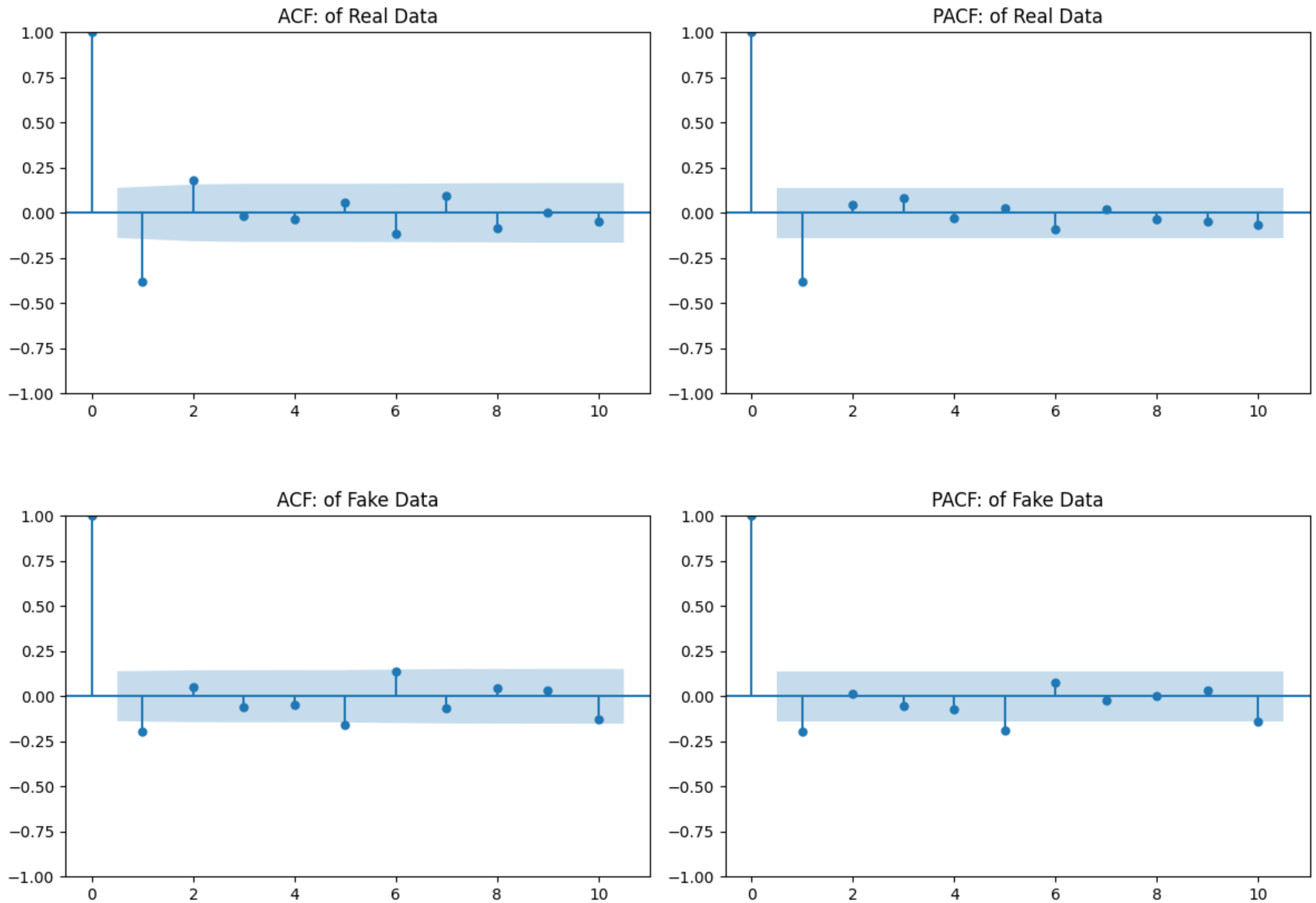
AR(1) with phi= 0.5

- Residual diagnostics: Ljung-Box Test
  - The Ljung-Box test results indicate that most p-values remain above 0.05, suggesting that the residuals are largely uncorrelated.

Lag	LB Statistic	P-value
1	0.2402	0.6241
2	5.6291	0.0600
3	7.3641	0.0612
4	7.7367	0.1017
5	15.5719	0.0082
6	15.5863	0.0162
7	16.7046	0.0194
8	16.7194	0.0332
9	16.8096	0.0518
10	16.8190	0.0785

AR(1) with phi= -0.5

- ACF and PACF of real and fake data



AR(1) with phi= -0.5

- ADF Test for real and fake data
  - An ADF test was performed to assess stationarity, and the results confirmed that both datasets are completely stationary.

Data Type	ADF Statistic	p-value	Critical Value(1%)	Critical Value(5%)	Critical Value(10%)	Stationary Assumption
Real Data	-20.8845	0.0000	-3.4636	-2.8762	-2.5746	Reject H_0(Stationary)
Fake Data	-4.3082	0.0004	-3.4654	-2.8770	-2.5750	Reject H_0(Stationary)

AR(1) with phi= -0.5

- Model fit and information criteria
  - The AR(1) model outperformed the AR(2) model in terms of AIC, BIC, and HQIC, indicating a better model fit.

Model Information	Value
Model	AR(1)
No. Observations	200
Log Likelihood	445.220
AIC	-904.441
BIC	-894.561
HQIC	-900.442

Model Information	Value
Model	AR(2)
No. Observations	200
Log Likelihood	452.749
AIC	-897.498
BIC	-884.345
HQIC	-892.174

AR(1) with phi= -0.5

- Statistical significance of model parameters
  - Since the second lag in AR(2) is not statistically significant (p=0.876), and AR(1) provides a strong autoregressive structure with a significant coefficient (p=0.004), AR(1) is considered the more appropriate model.

Parameter	Coefficient	Std Err	z-value	p-value	95% Confidence interval
Constant	0.0015	0.002	0.856	0.392	[-0.002, 0.005]
y.L1	-0.2000	0.069	-2.887	0.004	[-0.336, -0.064]

Parameter	Coefficient	Std Err	z-value	p-value	95% Confidence interval
Constant	0.0014	0.002	0.789	0.430	[-0.002, 0.005]
y. L1	-0.2038	0.071	-2.859	0.004	[-0.343, -0.064]
y. L2	0.0110	0.071	0.155	0.876	[-0.128, 0.150]

AR(1) with phi= -0.5

- Residual diagnostics: Ljung-Box Test
  - The Ljung-Box test results indicate that all p-values remain above 0.05, suggesting that the residuals are largely uncorrelated.

Lag	LB Statistic	P-value
1	0.0029	0.9569
2	0.0088	0.9956
3	1.1790	0.7581
4	3.9035	0.4192
5	7.4749	0.1876
6	9.4892	0.1479
7	9.6469	0.2095
8	10.1920	0.2518
9	10.2021	0.3344
10	13.2521	0.2099

GARCH(1,1) with t(5) Distribution

- Real data
  - Residual analysis of real data

Test	Result	Interpretation
Estimated Degrees of Freedom	4.91	Indicate the presence of fat tails.
K-S Test(t-distribution fit)	Stat= 0.029, p-value=0.99	The residuals closely follow a t-distribution (good fit).
Shapiro-Wilk Test	Stat= 0.97, p-value=0.00	The residuals do not follow a normal distribution.
Ljung-Box Q-Statistic (lag=10)	All p-values > 0.05	No significant autocorrelation detected up to lag 10.

  - Squared residual analysis of real data

Metric	Estimated Value	Interpretation
Omega (ω)	5.7796e-05	Represents the baseline level of volatility.
Alpha (α_1) (ARCH Effect)	0.0100	The impact of past shocks (ARCH effect) is minimal.
Beta (β_1) (GARCH Effect)	0.8900	Volatility exhibits strong persistence over time.
ARCH Test	Stat = 4.2993, p-value = 0.9328	No significant ARCH effect detected.
Ljung-Box Q-statistic (lag=10)	All p-values > 0.05	No significant autocorrelation in squared residuals.

GARCH(1,1) with t(5) Distribution

- Fake data generated by Tail GAN
  - Residual analysis of fake data

Test	Result	Interpretation
Estimated Degrees of Freedom	27.21	Indicate the presence lighter tails compared to real data.
K-S Test(t-distribution fit)	Stat= 0.035, p-value=0.96	The residuals closely follow a t-distribution.
Shapiro-Wilk Test	Stat= 0.99, p-value=0.64	The residuals do not significantly deviate from a normal distribution.
Ljung-Box Q-Statistic (lag=10)	All p-values > 0.05	No significant autocorrelation detected up to lag 10.

  - Squared residual analysis of fake data

Metric	Estimated Value	Interpretation
Omega (ω)	4.1528e-05	Represents the baseline level of volatility.
Alpha (α_1) (ARCH Effect)	0.0282	The impact of past shocks (ARCH effect) is minimal.
Beta (β_1) (GARCH Effect)	0.9187	Volatility exhibits strong persistence over time.
ARCH Test	stat=9.2075, p-value=0.5125	No significant ARCH effect detected.
Ljung-Box Q-statistic (lag=10)	All p-values > 0.05	No significant autocorrelation in squared residuals.

GARCH(1,1) with t(5) Distribution

- Fake data generated by Tail GAN
  - Among the different GARCH models evaluated on fake data, GARCH(1,1) with a normal distribution demonstrates the best fit based on AIC and BIC, making it the most appropriate model for capturing the volatility structure of the dataset.

Model	Log-Likelihood	AIC	BIC
GARCH(1,1)	468.1790	-926.3580	-909.8664
GARCH(1,1) with Normal distribution	467.4710	-926.9430	-913.7500
GARCH(1,2)	468.1789	-924.3578	-904.5679
GARCH(1,2) with Normal distribution	467.4972	-924.9945	-908.5029
GARCH(2,1)	467.1723	-922.3446	-902.5547
GARCH(2,1) with Normal distribution	467.6058	-925.2116	-908.7200

GARCH(1,1) with t(10) Distribution

- Real data
  - Residual analysis of real data

Test	Result	Interpretation
Estimated Degrees of Freedom	7.54	Indicate the presence of fat tails.
K-S Test(t-distribution fit)	stat=0.0382, p-value=0.9217	The residuals closely follow a t-distribution (good fit).
Shapiro-Wilk Test	stat=0.9833, p-value=0.0177	The residuals do not follow a normal distribution.
Ljung-Box Q-Statistic (lag=10)	All p-values > 0.05	No significant autocorrelation detected up to lag 10.

  - Squared residual analysis of real data

Metric	Estimated Value	Interpretation
Omega (ω)	6.3014e-05	Represents the baseline level of volatility.
Alpha (α_1) (ARCH Effect)	0.1000	The impact of past shocks (ARCH effect) is minimal.
Beta (β_1) (GARCH Effect)	0.4000	Volatility exhibits strong persistence over time.
ARCH Test	stat=6.7842, p-value=0.7457	No significant ARCH effect detected.
Ljung-Box Q-statistic (lag=10)	All p-values > 0.05	No significant autocorrelation in squared residuals.

GARCH(1,1) with t(10) Distribution

- Fake data generated by Tail GAN
  - Residual analysis of fake data

Test	Result	Interpretation
Estimated Degrees of Freedom	12	Indicate the presence of fat tails.
K-S Test(t-distribution fit)	stat=0.0664, p-value=0.3299	The residuals closely follow a t-distribution (good fit).
Shapiro-Wilk Test	stat=0.9821, p-value=0.0121	The residuals do not follow a normal distribution.
Ljung-Box Q-Statistic (lag=10)	All p-values > 0.05	No significant autocorrelation in squared residuals.

  - Squared residual analysis of fake data

Metric	Estimated Value	Interpretation
Omega (ω)	4.18e-05	Represents the baseline level of volatility.
Alpha (α_1) (ARCH Effect)	1.00e-02	The impact of past shocks (ARCH effect) is minimal.
Beta (β_1) (GARCH Effect)	0.8900	Volatility exhibits strong persistence over time.
ARCH Test	stat=9.4918, p-value=0.4862	No significant ARCH effect detected.
Ljung-Box Q-statistic (lag=10)	All p-values > 0.05	No significant autocorrelation in squared residuals.

GARCH(1,1) with t(10) Distribution

- Fake data generated by Tail GAN
  - All normal-distribution-based GARCH models resulted in **β = 1**, indicating a lack of stationarity.
  - Despite slightly higher AIC and BIC values, GARCH(1,1) with a t-distribution was chosen as the most appropriate model, as it provides a more stable and realistic representation of volatility dynamics.

Model	Log-Likelihood	AIC	BIC
GARCH(1,1)	487.9308	-965.8617	949.3951
GARCH(1,1) with Normal distribution	492.1200	-976.2400	-963.0670
GARCH(1,2)	487.9646	-963.9291	-944.1693
GARCH(1,2) with Normal distribution	492.1494	-974.2987	-957.8322
GARCH(2,1)	488.0090	-964.0179	-944.2581
GARCH(2,1) with Normal distribution	492.1002	-974.2004	-957.7339

5. Limitations & Future Improvements

Limitations

- Inconsistency across generated datasets
  - While the generated fake data follows the structure of (1000,5,200), not all 1000 datasets exhibit the same distribution as the real data.
  - Some datasets deviate from the expected distribution, potentially due to instability in GAN training or localized overfitting.

- Approximation rather than perfect replication
  - The residual analysis and GARCH model fitting confirm that the fake data approximates the volatility characteristics of real financial data.

Cause Analysis

- Differences in dataset
  - The original study utilized a (50,000, 5, 100) dataset with a batch size of 1,000, trained over 50 iterations. In contrast, this study used a (10,000, 5, 200) dataset, maintaining the same batch size but reducing training to 10 iterations, which may have limited the model's ability to fully capture the underlying distribution.

- Differences in training process
  - The original setup included a reset mechanism to stabilize training when the loss exceed ed a threshold. Due to computational constraints, this mechanism was omitted, potentially leading to higher variance and deviations in the generated data distributions.

Future Directions

- Comparative analysis with alternative models
  - Evaluating Tail GAN against other generative models, as done in the original study, would offer a more comprehensive assessment of its ability to capture financial time series vola tility.

- Appliance to real financial data
  - Beyond synthetic data, applying the model to real-world financial datasets would help as sess its practical applicability in realistic market conditions.

REFERENCES

- [1] Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. In International Conference on Machine Learning, pages 3936–3945. PMLR, 2018.
- [2] Siddharth Bhatia, Arjit Jain, and Bryan Hooi. ExGAN: Adversarial Generation of Extreme Samples. arXiv preprint arXiv:2009.08454, 2020.
- [3] Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. Journal of Banking & Finance, 26(7):1505–1518, 2002.
- [4] R. Cont, M. Cucuringu, R. Xu, and C. Zhang, "Tail-GAN: Learning to simulate tail risk scenarios," *arXiv preprint arXiv:2203.01781*, 2022. Figure 2, p. 12.

FINANCIAL INNOVATION
& ANALYTICS LAB.

Q&A