Scoping Note

Synthetic Data Augmentation for Trading Strategy Calibration and Risk Forecasting

Alaa Bouattour
Mahdi Ben Ayed

May 22, 2025

1. Background

The increasing reliance on machine learning models in financial forecasting tasks is often limited by the amount and diversity of available data. Synthetic data generation offers a promising solution, particularly in the context of financial time series where regime shifts and structural breaks may impair model generalization. This project explores whether realistic, regime-aware synthetic return series can improve both trading strategy calibration and tail-risk forecasting performance.

2. Objectives

The main objectives of the project are to:

- Design a regime-aware synthetic data generation process using Variational Autoencoders (VAEs).
- Evaluate the impact of synthetic data on the calibration and out-of-sample performance of a cross-sectional momentum strategy.
- Assess the influence of synthetic data on the calibration and statistical consistency of Value-at-Risk (VaR) forecasts.

3. Scope of Work

1. **Data Cleaning:** Handle missing data, validate time consistency, and normalize returns.
2. **Regime Detection:** (Handled separately by co-author; excluded from this document)
3. **Synthetic Data Generation:** Use a regime-specific VAE to generate realistic log-returns per market regime.
4. **Trading Strategy Calibration:** Optimize a long-short cross-sectional momentum strategy via grid search on real vs augmented data.
5. **Backtesting:** Compare in-sample and out-of-sample performance using cumulative return, Sharpe ratio, and max drawdown.
6. **VaR Forecasting:** Model 1-month VaR using Student-t historical simulation, and assess exceedance frequency using the Kupiec test.
7. **Comparative Analysis:** Contrast the performance and risk profiles of models trained on real versus augmented datasets.

4. Deliverables

- Cleaned and interpolated dataset of asset returns.
- Regime-tagged synthetic return panel.
- Grid-search calibration results for real and real+synthetic data.
- Full backtesting plots and metrics (cumulative return, drawdown, Sharpe ratio).
- VaR forecast plots and Kupiec test outputs.
- A structured LaTeX report including all code, results, and visualizations.

5. Constraints and Assumptions

- Synthetic data is assumed to be statistically similar to real data in each regime.
- Strategy hyperparameters are optimized only on the training set to preserve out-of-sample integrity.
- The same test set is used across all model evaluations for comparability.