2025-2 IMEN891M – Financial BigData Analysis

Proposal Presentation

오승준, 최우혁, Isac Johnsson

2025.09.30.

1 Introduction

POSTECH

Background

- High dimensional data analysis is essential, yet poses significant challenges in modern econometrics.
- Classic mean-variance **Markowitz portfolio theory** often fails into higher dimension of data due to unstable covariance estimation, omitted network dependencies and the curse of dimensionality.
- To address these challenges, various methodologies has before been proposed, a few important ones are:

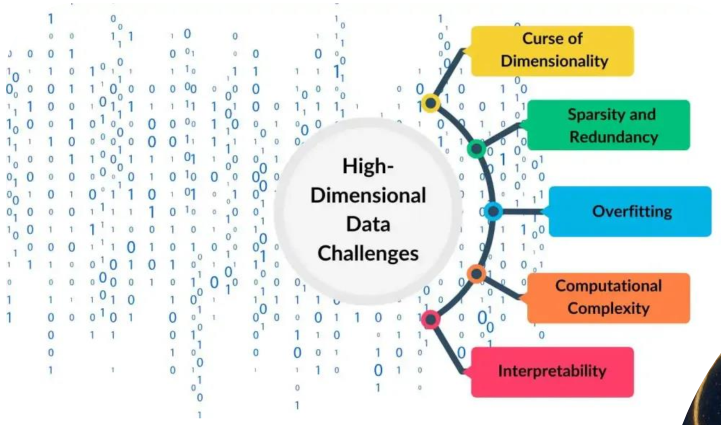

Image Source: Van Ottten, Neri; Spotintelligence.com

POET - High dimensional covariance estimation (Fan, Liao & Mincheva, 2013)

- By applying **factor models** and **sparsity** in the residual covariance matrix, it efficiently estimates **high-dimensional covariance matrices**.

SAR - Spatial Autoregression (Cliff & Ord, 1981; Baltegi et al., 2014)

- Model that **captures dependencies** structures across assets and markets.

LASSO - Regularized regression (Tibshirani, 1996)

- Imposes **sparsity on portfolio weights**, which improves stability in portfolio optimization in high dimensions.

Image Source: OpenAI

Therefore we integrate these High-Dimensional Methods for Robust Portfolio Construction

2 Methodological Framework

POSTECH

Step 1 : Macro-Factor Selection via LASSO

- We consider a large set of potential macroeconomic and market predictors (e.g., interest rates, exchange rates, volatility indices, credit spreads, commodity benchmarks).

LASSO - Regularized regression (Tibshirani, 1996)

- To avoid overfitting and spurious correlations, we apply LASSO regression : $\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$ where y is the asset return or factor proxy, and X is the predictor matrix.

- LASSO selects a sparse subject of variables, forming the observed macro-finance block f_t^{Macro}.

Step 2 : Covariance Estimation via POET

- To obtain a stable and interpretable covariance matrix in high dimensions, we use POET estimator.

POET - High dimensional covariance estimation (Fan, Liao & Mincheva, 2013)

The p-dimensional excess returns R_t are modeled as :

$$R_t = Bf_t + u_t$$

where $f_t = (f_t^{Macro}, f_t^{Latent})$ includes both selected macro factors and latent statistical factors.

The covariance decomposes as :

$$\Sigma = B\Sigma_f B^T + \Sigma_u$$

We estimate this via the **POET estimator** :

$$\hat{\Sigma}_{POET} = \hat{\Sigma}_{Factor} + \mathcal{T}_\omega(\hat{\Sigma}_u)$$

where $\hat{\Sigma}_{Factor}$ is obtained by PCA on residual returns (after macro-factors) and $\mathcal{T}_\omega$ applies adaptive thresholding to enforce sparsity in $\hat{\Sigma}_u$.

2 Methodological Framework

POSTECH

Step 3 : Residual Network Dependence via Spatial AR

- While POET assumes sparsity in the idiosyncratic covariance Σ_u, empirical evidence suggests that **idiosyncratic shocks may propagate via network spillovers** (e.g. financial contagion, sectoral interdependence).

SAR - Spatial Autoregression (Cliff & Ord, 1981; Baltegi et al., 2014)

We model this via a **Spatial AR** process :

$$u_t = \rho W u_t + \epsilon_t$$

where W is a pre-specified or estimated spatial weight matrix (e.g., correlation-based, geographic, or sectoral linkage).

This step allows us to go beyond unstructured sparsity and capture structured dependencies among residuals.

Step 4 : Portfolio Optimization

- With the integrated covariance estimator $\hat{\Sigma}_{Integrated}$ combining POET and SAR adjustments, and with expected returns $\hat{\mu}$ informed by LASSO-selected macro predictors, we construct portfolios:
- Global Minimum Variance (GMV) : $w_{GMV} = \frac{\hat{\Sigma}^{-1}1}{1^T\,\hat{\Sigma}^{-1}1}$
- Mean-Variance Optimal Portfolio : $w_{MV} = \frac{\hat{\Sigma}^{-1}\,\hat{\mu}}{1^T\,\hat{\Sigma}^{-1}\,\hat{\mu}}$
- Risk Budgeting Portfolio : Equalizing marginal contributions to risk subject to l_1 gross-exposure constraints.

3 Data and Implementation

POSTECH

DATA SELECTION

- We employ a cross-asset dataset (2015.01.01 - 2024.12.31) spanning approximately 70~80 series. (Data Source : investing.com , Yahoo Finance, etc.)

Equities

- Global indices (S&P 500 and it's sectoral indices, MSCI World, Major regional stock indexes like Nikkei, KOSPI etc…)

Fixed Income

- U.S. Treasury yields, corporate bond spreads

Commodities

- Gold, Oil(Brent, WTI), industrial metals, agricultural futures

Currencies

- USDKRW, JPYKRW, EURKRW, CNYKRW

Others

- Bitcoin (Cryptocurrencies)
- VIX(Volatility Measures)
- Macro indicator proxies (ex. CPI)

DATA PREPROCESSING

Frequency Harmonization

↓

Return Computation

↓

Stationarity Testing

↓

Standardization

3 Data and Implementation

POSTECH

IMPLEMENTATION METHOD

Step 1. Data Preprocessing

↓

Step 2. Macro factor Selection with LASSO

↓

Step 3. Combine macro factors with PCA-based latent factors in POET

↓

Step 4. Modeling about residuals with Spartial AR (using correlation-based network matrices)

↓

Step 5. Optimize Portfolio with Model

4 Expected Contributions

POSTECH

Methodological Integration

- Integrating LASSO, SAR and POET into one **unified framework.**
- Accounts for **factor structure**, **network dependence** and **predictor selection** in high-dimensional financial data.

Pedagogical Value

- Directly connects central concepts in Financial Big Data like **high-dimensional regression**, **factor models**, **sparsity** and **network econometrics**.

Financial Implications

- The framework is expected to give **Portfolio Stability** through more stable asset allocations under various gross-exposure constraints.
- Enhances **Tail-Risk Management**, by improving VaR and Expected Shortfall estimations during crises.
- Focuses on **Cross-Asset diversification**, by capturing contagion and cross market linkages.