Project 2 Proposal:
Stable Portfolio Compression with Trees and Local Regressions

Angus Cheung

November 14, 2025

Summary. Track a target index using a small, stable portfolio of stocks chosen via a Minimum Spanning Tree (MST) on correlations and reconstructed with local ridge regressions. Applicable to index-tracking passive funds.

Finance background.

- Portfolio: a weighted mix of stocks; tracking means matching a target (e.g., S&P 500) as closely as possible.
- Tracking error: the volatility of the portfolio's difference from the target; lower means better tracking.
- Why compress? Fewer stocks are cheaper and easier to manage while still approximating the market.

Core idea. Build an MST from denoised return correlations, pick k "basis" stocks for diversity, and express other stocks as short ridge regressions on their nearest basis neighbors; map the index's weights onto these k names to form a compact tracker.

Minimum viable product (MVP) features.

- Data loader for historical daily returns of S&P 500 constituents (fixed universe snapshot).
- Compute correlation matrix with simple shrinkage; convert to distances $d_{ij} = \sqrt{2(1-\rho_{ij})}$; build MST. Standardized return vectors have correlation $\rho_{ij} = \cos\theta$, so their Euclidean distance is $\|r_i - r_j\|_2 = \sqrt{2 - 2\cos\theta} = \sqrt{2(1-\rho_{ij})}$.
- Basis selection: max-spread on the tree to choose k diverse stocks.
- Local ridge reconstruction with $q \in \{1, 2\}$ nearest basis neighbors; form sparse mapping A.
- Construct compressed portfolio by mapping index weights w_{SPX} to basis weights $w_B = A^\top w_{\text{SPX}}$.
- Backtest with monthly rebalancing on a rolling window; report out-of-sample tracking error and turnover.

Stretch features (beyond MVP).

- Stability constraint on basis membership to reduce turnover; threshold for "worthwhile" changes.
- Small constrained refinement (nonnegative weights, sector caps) via least-squares on basis returns.
- Comparative baselines: PCA-to-names compression; LASSO subset selection; random k-subset.
- Basic transaction-cost model to convert turnover into expected drag.

Description of complexity.

- Algorithmic: building an MST from dense distances; efficient nearest-basis lookup on a tree.
- Statistical: denoising correlations; choosing k and ridge λ to balance bias/variance.
- Systems: rolling backtest with changing constituents/weights; ensuring reproducible pipelines.
- Evaluation: fair out-of-sample splits; comparing trackers across regimes (calm vs. crisis).