

Louvain School of Management

How many principal components for optimal portfolio selection?

Author: Léa Therasse
Supervisor(s): Nathan Lassance
Academic year 2025-2026
Master's degree in Business engineering

STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	Your answer
The content of my/our master's thesis is my/our original <u>output</u> , even if I/we used generative AI to help prepare it.	☒ Yes ☐ No
I/we master the theories, tools and methods that I/we use for the thesis.	☒ Yes ☐ No
I/we verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☒ Yes ☐ No
I/we acknowledge that I/we master the final output and that I am/we are able to explain it and answer questions about it.	☒ Yes ☐ No
I/we made sure not to provide generative AI tools with access to confidential data or information	☒ Yes ☐ No

SIGNATURE

Sign your declaration:

Author last name, first name(s)	Date	Signature
1 Léa Therasse	19/05/26	

Contents

1	Introduction	1
2	Literature review	3
2.1	Portfolio optimization and estimation error	3
2.2	High-dimensional covariance matrix estimation	6
2.3	Principal component analysis and factor models	8
2.4	PCA based portfolio construction	9
2.5	Selecting the number of principal components	10
2.5.1	Explained variance criteria	11
2.5.2	High-dimensional consistent estimator of Bai and Ng (2002)	12
2.5.3	Random matrix theory approaches	13
2.6	Summary and research gaps	13
3	Methodology	15
3.1	Data description	15
3.2	Portfolio construction	16
3.3	Proposed portfolio-driven approach for selecting the number of principal components	17
3.4	Out-of-sample performance evaluation	18
3.5	Benchmark portfolio strategies	19
4	Empirical results and analysis	21
4.1	The structural failure of the Tangency Portfolio	21
4.2	Impact of the number of principal components on out-of-sample performance . .	21
4.3	Optimal number of components and comparison to usual selection rules	23
4.3.1	Robustness check: sensitivity to the estimation window	25
4.4	Performance of PC-optimized portfolios relative to benchmark strategies	26
4.4.1	Robustness check: performance during market crises	27
4.5	Portfolio stability and the impact of transaction costs	29
4.5.1	Robustness check: the impact of transaction costs	31
5	Conclusion	33
5.1	Findings	33
5.2	Limitations and directions for future research	34
5.3	Closing remarks	35
6	Appendix	36

1 Introduction

Since the foundational work of Markowitz (1952), mean-variance optimization has remained the theoretical cornerstone of modern asset allocation. In practice, however, its implementation is notoriously fragile. As the true expected returns and covariance matrix are unobservable, investors must rely on sample estimates. In high-dimensional financial markets where the number of assets (N) rivals or exceeds the number of historical observations (T), the traditional sample covariance matrix becomes ill-conditioned. Consequently, mean-variance optimizers act as "error-maximizing" machines (Michaud, 1989), generating unstable out-of-sample portfolios that frequently fail to outperform a naive $1/N$ equally weighted allocation (DeMiguel, Garlappi, & Uppal, 2009). Given that mean-variance frameworks underpin a significant share of institutional asset management, the practical cost of this estimation failure is substantial: even marginal improvements in covariance estimation translate directly into more stable portfolio weights and material gains in risk-adjusted performance at scale.

To mitigate the effects of the curse of dimensionality, practitioners often rely on Principal Component Analysis (PCA) to stabilize the covariance matrix by filtering out idiosyncratic noise. Yet, applying PCA introduces a critical hyperparameter which is often arbitrarily resolved : the number of principal components (K) to retain. Standard approaches use statistical heuristics, such as variance thresholds or Random Matrix Theory, that are calibrated to minimize data reconstruction error. However, this thesis challenges that methodology, highlighting a fundamental issue: statistical optimality does not necessarily maximize an investor's out-of-sample risk-adjusted performance. To address this gap, this research proposes that K should be dynamically optimized via a portfolio-driven objective function. Specifically, this research introduces a purged cross-validation methodology (PCA_{cv}) that selects K by directly targeting out-of-sample portfolio volatility and Sharpe ratio. This methodological shift is driven by the central research question of this thesis: *What is the gain in out-of-sample portfolio performance from selecting the number of principal components optimally using a portfolio motivated objective function?*

To evaluate the economic value added by this approach, a comprehensive empirical analysis is conducted over a 35-year out-of-sample period (1990-2025) across three distinct asset universes: 48 Industry Portfolios, 100 Fama-French Portfolios, and a cross-section of 286 S&P 500 equities. Under both the Global Minimum Variance (GMV) and Tangency portfolio frameworks, PCA_{cv} is benchmarked against naive $1/N$ allocation, the full sample covariance matrix, Ledoit-Wolf shrinkage, and standard static PCA rules. The empirical results reveal several key findings. First, the relationship between retained components and out-of-sample risk is non-monotonic; retaining too many components reintroduces estimation noise, severely degrading performance. Second, the efficacy of dimensionality reduction is fundamentally dictated by the concentration ratio N/T : while shrinkage estimators provide superior downside protection in lower-dimensional spaces ($N = 48, 100$), the cross-validated PCA demonstrates deep structural resilience in highly over-parameterized environments such as the S&P 500. Third, while the portfolio-driven objective yields significant gross outperformance over static statistical rules, this advantage is largely consumed by transaction costs: the constant recalibration of K induces subspace instability

and portfolio turnover that make the strategy economically unviable under standard trading frictions. Ultimately, this thesis demonstrates that while dynamic PCA successfully solves the mathematical problem of out-of-sample covariance estimation, it introduces a critical structural challenge that future research must address by explicitly integrating turnover penalties into the selection objective.

Two methodological challenges shape the design of this study. The primary difficulty lies in avoiding look-ahead bias when evaluating a dynamic, data-driven selection rule on financial time-series, which motivates the strict chronological purging embedded in the cross-validation procedure. A secondary challenge concerns the survivorship bias inherent in the S&P 500 sample, which restricts the universe to continuously quoted equities and may affect the generalization of results to the broader index. Following the introduction, Section 2 reviews the relevant literature on portfolio optimization and PCA-based dimensionality reduction. Section 3 describes the data, portfolio construction, and the proposed cross-validation framework. Section 4 presents the empirical findings and robustness checks. Finally, section 5 concludes and outlines directions for future research.

2 Literature review

2.1 Portfolio optimization and estimation error

The conceptual pillars of contemporary asset allocation were defined by Markowitz in 1952, who introduced the mean-variance optimization framework to reconcile expected returns with portfolio risk. In the Markowitz (1952) framework, an investor allocates wealth across a finite set of risky assets in order to minimize portfolio variance for a given level of expected return, or equivalently, to maximize expected return for a given level of risk.

Let the components of the Markowitz framework be defined as follows:

- $R \in \mathbb{R}^N$: The vector of asset returns.
- $\mu = \mathbb{E}[R]$: The vector of expected asset returns.
- $\Sigma = \mathbb{E}[(R - \mu)(R - \mu)^\top]$: The $N \times N$ covariance matrix capturing the variances and cross-correlations of the assets.
- $w \in \mathbb{R}^N$: The vector of portfolio weights allocated to each asset.
- $w^\top \mathbf{1} = 1$: The budget constraint, ensuring the investor's wealth is fully allocated (where $\mathbf{1}$ is a vector of ones).
- $R_p = w^\top R$: The realized portfolio return.
- $\mathbb{E}[R_p] = w^\top \mu$: The expected portfolio return.
- $\text{Var}(R_p) = w^\top \Sigma w$: The portfolio variance, representing the total risk of the allocation.

With these variables defined, the classical mean-variance optimization problem, as formalized in subsequent utility theory, building on Markowitz's foundation, can be defined in its risk-aversion formulation, maximizing quadratic utility subject to a budget constraint, which can be written as:

$$\max_w w^\top \mu - \frac{\gamma}{2} w^\top \Sigma w \quad \text{s.t.} \quad w^\top \mathbf{1} = 1. \quad (1)$$

To understand the mathematical vulnerability of this framework, it is useful to examine its closed-form solution. For instance, as the investor's risk aversion approaches infinity (i.e., $\gamma \rightarrow \infty$), the objective reduces to pure risk minimization, yielding the Global Minimum Variance (GMV) portfolio. Assuming the covariance matrix Σ is invertible, a strong assumption that does not always hold in practice, its optimal weights are strictly determined by:

$$w^* = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}. \quad (2)$$

The Markowitz model relies on a set of powerful assumptions. Asset returns are assumed to be fully characterized by their first two moments; it is worth noting that returns can be non-normal, provided that the investor exhibits a quadratic utility function. The investment horizon is static, transaction costs are ignored, and the parameters μ and Σ are treated as known. Under these

presumptions, diversification arises naturally as a rational response to imperfect correlation among asset returns, and the efficient frontier summarizes the set of optimal portfolios.

In practice, the population moments μ and Σ are unobservable and must be estimated from historical data. This introduces estimation error, defined as the difference between the true, unknown parameters and their sample-based estimates. Replacing the true parameters with their sample estimates yields the empirical optimal weights. For an investor seeking to maximize risk-adjusted returns, the theoretical ideal is the Tangency portfolio, whose estimated weights are given by:

$$\hat{w}_{tan} = \frac{\hat{\Sigma}^{-1}\hat{\mu}}{\mathbf{1}^\top \hat{\Sigma}^{-1}\hat{\mu}}. \quad (3)$$

However, sample means ($\hat{\mu}$) are notoriously difficult to estimate accurately and are subject to much larger estimation errors than covariances. Because the Tangency portfolio relies directly on these highly noisy expected returns, it frequently suffers from extreme weight instability and severely degraded out-of-sample performance. Consequently, some of the empirical literature, and a major focus of this thesis, also isolates the objective of minimizing pure risk by evaluating the GMV portfolio. By completely omitting μ from the optimization process, the GMV avoids this massive source of estimation error, generally yielding much more robust and stable performance out-of-sample. The empirical optimal weights for the GMV portfolio are thus given by:

$$\hat{w}_{gmv} = \frac{\hat{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^\top \hat{\Sigma}^{-1}\mathbf{1}}. \quad (4)$$

Estimation error poses a particularly significant problem in portfolio optimization because these optimal weights depend on the inverse of the sample covariance matrix, $\hat{\Sigma}^{-1}$. Because the mathematical inversion process is highly sensitive to the smallest eigenvalues of the sample covariance matrix, even minor sampling variations in $\hat{\Sigma}$ are dramatically amplified. This leads to extreme, unstable portfolio weights that fluctuate wildly over time. This phenomenon led Michaud (1989) to famously describe mean-variance optimizers as "estimation-error maximizers". This problem becomes acute in high-dimensional settings, where the number of assets N is large relative to the available sample size T . As the ratio N/T approaches unity, the sample covariance matrix is noisy and ill-conditioned, and its inverse is highly sensitive to sampling variation.

The substantial economic costs associated with estimation error are rigorously quantified by DeMiguel, Garlappi, and Uppal (2009). The authors evaluate a wide range of optimized portfolio strategies across multiple datasets and benchmark them against the naive equally weighted diversification strategy, which is mathematically defined as $w_{ew} = \frac{1}{N}\mathbf{1}$. Using the out-of-sample Sharpe ratio to evaluate performance, they demonstrate that none of the considered strategies outperforms the simple $1/N$ portfolio consistently in out-of-sample terms. Their results suggest that, for realistic sample sizes, the theoretical gains promised by mean-variance optimization are outweighed by the negative impact of estimation error. This finding challenges the practical relevance of optimal diversification and highlights the vulnerability of portfolio optimization when parameters are estimated rather than known.

Recent evidence reinforces these conclusions. Gelmini and Uberti (2024) replicate the analysis of DeMiguel, Garlappi, and Uppal (2009) using updated datasets that include periods of pronounced market stress, such as the global financial crisis and the COVID-19 pandemic. Despite richer data and alternative portfolio strategies, their results confirm that the equally weighted portfolio remains a remarkably robust benchmark. These findings emphasize that estimation error is not a transitory phenomenon but rather a structural feature of portfolio selection in the face of uncertainty.

From a conceptual perspective, estimation error can be interpreted as a bias–variance trade-off in portfolio construction. Highly flexible models that attempt to exploit all available cross-sectional information tend to fit in-sample data well but suffer from poor out-of-sample performance due to overfitting. Conversely, simple allocation rules such as the $1/N$ portfolio exhibit higher bias but substantially lower variance, which can result in superior out-of-sample performance. This observation provides an intuitive explanation for the empirical success of naive diversification.

However, while the empirical success of the $1/N$ rule is well documented, it is important to recognize its theoretical limitations with regard to true risk distribution. As Meucci (2009) demonstrates, allocating equal capital across assets does not mean to an equal allocation of risk. Because financial assets are often highly correlated, a seemingly diversified equally weighted portfolio may actually be concentrated in a single dominant risk factor, creating an illusion of diversification.

A significant contribution to the solution of this high-dimensional problem is provided by Kozak et al. (2020), who developed a robust framework for estimating the Stochastic Discount Factor (SDF). Their methodology achieves superior out-of-sample performance by imposing a prior motivated by economic considerations that shrinks the SDF coefficients of low-variance principal components towards zero. Mathematically, shrinkage techniques mitigate the inversion explosion by implicitly adding a penalty term, such as a ridge penalty, to the sample covariance matrix before inversion. This yields the inverse of the matrix, which is written as $(\hat{\Sigma} + \lambda I)^{-1}$. By penalizing exposure to low-variance components, which are often dominated by noise, the low signal-to-noise ratio that typically plague high-dimensional datasets are effectively filtered out. However, while this approach prioritizes dominant systematic variation for SDF construction, it remains to be seen whether a strictly portfolio-motivated objective function might identify latent value in components that are usually discarded by standard statistical selection rules.

Taken together, these contributions highlight a key tension in portfolio optimization. Although the Markowitz framework provides a solid theoretical foundation, its practical use is heavily limited by estimation errors. To address this issue, a structure must be imposed on the estimation problem by shrinking parameters, constraining portfolio weights or reducing dimensionality through factor models. In this context, PCA emerges as a useful tool for balancing model complexity and robustness. However, as the subsequent sections will argue, the effectiveness of PCA-based approaches critically depends on how the number of retained components is chosen, a question that remains unresolved from a portfolio-centric perspective.

2.2 High-dimensional covariance matrix estimation

The implementation of mean-variance optimization is fundamentally hampered by the "curse of dimensionality" when the number of assets N is large relative to the sample size T . As highlighted by Ledoit and Wolf (2022), the sample covariance matrix becomes increasingly ill-conditioned as the concentration ratio N/T rises. In this case, the sample eigenvalues are systematically distorted: small eigenvalues, which correspond to low-variance directions, tend to be artificially underestimated, while large eigenvalues are overestimated. A mean-variance optimizer, which seeks to minimize risk, naturally overweights the directions associated with these artificially small eigenvalues. This results in portfolios that are highly unstable and dominated by estimation noise rather than true diversification opportunities.

To reduce this instability, the literature has proposed structural estimators that impose parsimony on the covariance matrix. A widely adopted approach relies on factor models, which decompose the $N \times 1$ vector of asset returns R into a common systematic component and an idiosyncratic residual:

$$R = BF + \epsilon \quad (5)$$

where the components of the factor model are defined as:

- B : $N \times K$ matrix of factor loadings, capturing the sensitivity of each asset to the latent factors.
- F : $K \times 1$ vector of latent factor returns, representing the common systematic risks.
- ϵ : $N \times 1$ vector of idiosyncratic errors, representing the asset-specific risk that is uncorrelated with the common factors.

Under a factor structure, the $N \times N$ covariance matrix can be expressed as:

$$\Sigma = B\Omega B^\top + \Delta \quad (6)$$

where the components of the factor covariance structure are defined as:

- Ω : $K \times K$ covariance matrix of the latent factors, capturing the variances and cross-correlations among the systematic factors.
- Δ : $N \times N$ covariance matrix of the residuals, which is typically assumed to be strictly diagonal, reflecting the assumption that idiosyncratic risks are uncorrelated across assets.

The exact properties of Δ depend on whether the model assumes a strict or an approximate factor structure:

- **Strict factor structure:** Idiosyncratic risks are assumed to be entirely uncorrelated across assets. Under this strong assumption, Δ is a strictly diagonal matrix.
- **Approximate factor structure:** Introduced by Chamberlain and Rothschild (1983), this relaxed framework allows idiosyncratic errors to exhibit weak cross-sectional correlations. Consequently, Δ is not required to be strictly diagonal. Instead, its maximum eigenvalue is assumed to be bounded as the number of assets N approaches infinity. This ensures that

the pervasive cross-sectional correlation is driven exclusively by the K common factors, rather than the residuals.

By restricting $K \ll N$, this framework reduces the dimensionality of the estimation problem from order N^2 to order NK , thereby stabilizing both the covariance estimator and its inverse, which is required for portfolio weight computation (Connor & Korajczyk, 1988).

A critical challenge in this framework is determining the appropriate number of latent factors K . The standard econometric approach treats this as a model selection problem. Bai and Ng (2002) propose a set of information criteria (IC_p) to consistently estimate the number of factors by balancing the residual variance of the model against a penalty for complexity. The general form of this criterion is:

$$IC(k) = \ln(V(k)) + k \cdot g(N, T) \quad (7)$$

with

- $V(k) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \hat{\epsilon}_{it}^2$: the average residual variance when k factors are estimated, with $\hat{\epsilon}_{it}$ representing the estimated idiosyncratic errors from the k -factor model.
- $g(N, T)$: a penalty function that scales with the cross-sectional (N) and time-series (T) dimensions to prevent overfitting. To ensure consistent estimation of the true number of factors, $g(N, T)$ must approach zero as $N, T \rightarrow \infty$, but at a slower rate than $1/\min(N, T)$. Among the specifications proposed by the authors, the most widely adopted penalty (IC_{p1}) is defined as:

$$g(N, T) = \left(\frac{N+T}{NT} \right) \ln \left(\frac{NT}{N+T} \right) \quad (8)$$

Similarly, in the context of high-dimensional continuous-time settings, Ait-Sahalia and Xiu (2017) show that principal component analysis (PCA) can recover the latent factor space when both N and T grow, provided that the number of retained components adequately captures the systematic variation.

These contributions primarily focus on statistical consistency and convergence towards the actual data generation process. However, from a portfolio perspective, statistical optimality does not necessarily imply economic optimality. The number of factors that best approximates the population covariance structure may differ from the number that which maximizes out-of-sample portfolio performance. In particular, including weak factors may reduce statistical bias (lowering $V(k)$) but substantially increase estimation variance, thereby harming portfolio stability.

This distinction motivates a portfolio-centric approach to dimensionality selection. The optimal number of factors for an investor is not the true statistical value K , but the number that achieves the best compromise between the bias caused by excluding relevant risks and the variance introduced by estimating unstable or low-signal components. For this reason, this study builds on this insight by explicitly evaluating the out-of-sample performance implications of alternative factor selection rules.

2.3 Principal component analysis and factor models

PCA serves as the canonical statistical technique for dimensionality reduction in high-dimensional datasets. As highlighted by Jolliffe and Cadima (2016), PCA transforms a set of correlated variables into a new set of uncorrelated variables, the principal components (PCs), which are ordered so that the first few retain most of the variation present in all of the original variables.

To understand this mechanically, let R be the $T \times N$ matrix of historical asset returns. The foundation of PCA relies on the spectral decomposition of the sample covariance matrix $\hat{\Sigma}$:

$$\hat{\Sigma} = V \hat{\Lambda} V^\top \quad (9)$$

where the components of this decomposition are defined as:

- V : $N \times N$ orthogonal matrix whose columns contain the eigenvectors of $\hat{\Sigma}$.
- $\hat{\Lambda}$: $N \times N$ diagonal matrix containing the corresponding sample eigenvalues, ordered from largest to smallest ($\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_N$).

PCA seeks the linear combinations of the original returns that maximize variance subject to orthogonality constraints. These principal components, denoted by the $T \times N$ matrix F , are obtained by projecting the returns onto the eigenvectors:

$$F = RV \quad (10)$$

In practice, rather than retaining all N components, only the first K components associated with the largest eigenvalues are kept (the first K columns of V). This provides the best rank- K approximation of the data in a least-squares sense, capturing the dominant modes of co-movement among assets.

In the context of financial econometrics, PCA can be interpreted as an estimator for latent factor models. Bai and Ng (2002) establish the theoretical link between the statistical output of PCA and the structural parameters of an approximate factor model. They demonstrate that, under certain assumptions and when both N and T are large, the principal components provide a consistent estimate of the space spanned by the true unobservable factors driving asset returns. Under this interpretation, the eigenvalues of the covariance matrix represent the magnitude of systematic risk associated with each factor, while the eigenvectors capture the factor loadings, or sensitivities, of each asset.

While PCA is optimal for minimizing the reconstruction error of the return covariance matrix, its objective function is purely statistical. The algorithm selects components based solely on the magnitude of their variance, implicitly assuming that components with larger variance (signal) are more relevant than those with smaller variance (noise). However, as noted in previous sections, this variance-maximization criterion does not incorporate any information regarding the pricing of risk, defined as the expected excess return required by investors as compensation for holding one unit of exposure to a specific systematic risk factor.

Consequently, a purely statistical approach might prioritize high-variance components that do not

carry any risk premium in the market, while prematurely dismissing lower-variance components that are economically highly rewarded. This gap means that even though the leading principal components dominate statistically, they are not necessarily the most efficient factors for building a portfolio that performs well out-of-sample. This limitation highlights the need to evaluate factor selection rules not just by their statistical fit, as in Bai and Ng (2002), but by their direct impact on the investor's economic objective and portfolio utility.

2.4 PCA based portfolio construction

Moving beyond the econometric estimation of latent risk factors, PCA has become a foundational tool for the practical construction of robust investment strategies. By projecting asset returns onto their principal components, investors can form uncorrelated eigen portfolios with weights determined by the eigenvectors of the covariance matrix. This process of orthogonalization allows portfolio managers to identify and address specific sources of systematic risk, independent of cross-asset correlations.

Formally, the returns of these eigen portfolios, denoted by the vector P , are obtained through the linear transformation:

$$P = RV \quad (11)$$

Importantly, the eigenportfolio returns P are mathematically identical to the principal components F . Because the eigenvector matrix V is orthogonal, meaning its columns are mutually perpendicular unit vectors such that $V^\top V = I$, the covariance matrix of these new eigenportfolios is strictly diagonal. The elements on this diagonal correspond exactly to the eigenvalues ($\hat{\Lambda}$) of the original sample covariance matrix. This mathematical property guarantees that the extracted risk factors are perfectly uncorrelated, forming the basis of what Meucci (2009) formally identifies as principal portfolios for managing diversification.

A prominent application of this methodology is provided by Kozak et al. (2020), who utilize PCA to construct a robust SDF. This concept is mathematically dual to the mean variance efficient portfolio. Facing a high-dimensional cross section of anomaly-based stock predictors, the authors apply PCA to orthogonalize the candidate factors. They demonstrate that instead of relying on a sparse set of heuristic factors, investors achieve superior out-of-sample performance by investing across the principal components and applying an economically motivated shrinkage penalty. Specifically, rather than discretely dropping components, they apply a ridge penalty to the portfolio weights w . This introduces an L_2 penalty term governed by a tuning parameter α :

$$\text{Penalty} = \alpha \sum_{i=1}^N w_i^2 \quad (12)$$

This approach operationalizes the insight that the leading principal components contain the most robust pricing signals. Meanwhile, lower-variance components are disproportionately contaminated by estimation noise and naturally receive weights shrunk toward zero.

Beyond traditional mean-variance optimization, PCA is also crucial for implementing advanced portfolio strategies that account for non-normal return distributions. Lassance and Vrms

(2021) highlight the severe curse of dimensionality that plagues the estimation of higher order comoments, such as coskewness and cokurtosis, in large investment universes. To illustrate this comoments bottleneck, consider the number of parameters required for estimation:

- Covariance matrix (second moment): Requires estimating on the order of $O(N^2)$ parameters.
- Coskewness tensor (third moment): Requires estimating on the order of $O(N^3)$ parameters.
- Cokurtosis tensor (fourth moment): Requires estimating on the order of $O(N^4)$ parameters.

By using PCA to project the data onto a much smaller set of K driving factors, they reduce the estimation burden from $O(N^4)$ down to $O(K^4)$. While they ultimately advocate for Independent Component Analysis (ICA) to achieve full statistical independence, their framework is directly based on the variance maximizing properties of PCA. By reducing the data to a small, parsimonious set of driving factors, they demonstrate that complex portfolio strategies can survive estimation error and perform well out-of-sample.

Despite these significant advances, a fundamental question remains largely unaddressed: What is the strictly optimal truncation point for the number of principal components from a portfolio allocation perspective? The literature has thoroughly established that PCA improves portfolio robustness, yet the exact number of retained factors K is typically determined via statistical heuristics or continuous penalty functions. For instance, Kozak et al. (2020) rely on a cross-validated ridge penalty rather than a discrete truncation, while Lassance and Vrms (2021) focus primarily on the distributional transformation of the components. Consequently, the existing literature leaves the purely portfolio-driven determination of K as an open empirical challenge. Identifying the exact number of principal components that maximizes out-of-sample risk-adjusted returns for a mean-variance investor forms the central focus of the subsequent analysis.

2.5 Selecting the number of principal components

Having established the theoretical and practical advantages of PCA for dimensionality reduction in portfolio construction, the critical implementation step is determining the optimal truncation point K , where $1 \leq K \leq N$. This decision fundamentally represents a bias variance trade-off.

Mathematically, when an investor truncates the eigenvalue decomposition at K components to construct an estimated covariance matrix $\hat{\Sigma}_K$, defined as $\hat{\Sigma}_K = \sum_{i=1}^K \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ (where λ_i represents the i -th largest eigenvalue and $\mathbf{v}_i$ its corresponding eigenvector), the resulting out-of-sample portfolio variance differs from the theoretical minimum. This error can be conceptually decomposed into two competing forces:

- **Bias (Underfitting):** Retaining too few components (K is too small) discards valuable information about latent systematic risk factors. The model implicitly assumes zero correlation where true systematic correlations exist, leaving the portfolio exposed to unhedged market dynamics.
- **Variance (Overfitting):** Retaining too many components (K approaches N) incorporates idiosyncratic noise. Because mean-variance optimizers mathematically invert the covariance

matrix, they aggressively exploit the artificially small eigenvalues associated with this noise. This amplifies estimation variance, leading to unstable weights and poor out-of-sample performance.

Consequently, substantial methodologies have been developed across the fields of statistics, econometrics and quantitative finance to systematically separate the "signal" (true systematic factors) from the "noise" (spurious sample correlations). These approaches range from intuitive graphical heuristics to rigorous asymptotic theories. The following subsections review three distinct paradigms for determining the number of principal components, in order to contextualize the factor selection rules evaluated in this study: traditional explained variance thresholds, the high-dimensional econometric information criteria pioneered by Bai and Ng (2002), and eigenvalue filtering techniques derived from Random Matrix Theory.

2.5.1 Explained variance criteria

The most obvious criterion for choosing the number of retained components relies on the distribution of the covariance or correlation matrix eigenvalues, which directly represent the variance captured by each component. As detailed by Jolliffe and Cadima (2016), these methods aim to retain enough components to account for a desired cumulative percentage of total variation, while discarding the remaining low-variance components.

To formalize these heuristics, let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ represent the sorted eigenvalues of the sample covariance matrix. According to the comprehensive framework established by Jolliffe (2002), the three most widely used variance based rules are:

- **Cumulative percentage of total variation:** The researcher selects a desired cumulative percentage of total variation and chooses the smallest number of components K for which this percentage is exceeded. Mathematically, this target τ is evaluated as:

$$K = \min \left\{ k : 100 \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^N \lambda_j} \geq \tau \right\} \quad (13)$$

- **Kaiser's rule:** Also known as the eigenvalue greater than one rule, this criterion suggests retaining only those principal components whose eigenvalues exceed 1, specifically when the analysis is performed on a correlation matrix. The underlying rationale is that any retained principal component should account for more variation than an average original variable.
- **Scree plot:** This represents a well-known graphical rule where the eigenvalues λ_i are plotted against their index i . The researcher visually looks for a sharp break or "elbow" in the plot. The principal components that appear after this break are considered a scree of noise and are discarded. In this context, it is important to note that when PCA is performed on a correlation matrix, the average of the eigenvalues is exactly equal to 1. This property leads to a widely used supplementary rule known as the Kaiser criterion, which suggests retaining only the principal components with eigenvalues strictly greater than 1.

While these variance based criteria are computationally straightforward, both Jolliffe (2002) and Jolliffe and Cadima (2016) explicitly note that these rules are somewhat ad hoc and frequently yield divergent estimates for K . Furthermore, from a portfolio optimization perspective, these rules are entirely agnostic to the pricing of risk. Retaining components simply because they explain a large portion of historical variance does not guarantee that those components carry positive risk premia or improve out-of-sample portfolio utility. This limitation highlights the need for more formal econometric and portfolio driven approaches.

2.5.2 High-dimensional consistent estimator of Bai and Ng (2002)

To move beyond the arbitrary nature of explained variance heuristics, the econometric literature has developed formal statistical procedures to determine the number of latent factors. The foundational benchmark in this area is Bai and Ng (2002), who provide a rigorous framework for estimating the dimensionality of large approximate factor models. Unlike traditional statistical tests that assume a fixed number of assets N as the time-series dimension T grows, their framework accommodates the reality of modern panel data where both N and T approach infinity.

They propose a set of information criteria, previously introduced in subsection 2.2, that mathematically formalize the bias-variance trade-off. Under this framework, the estimated number of factors $\hat{K}$ is chosen to minimize the penalized residual variance $IC(k)$:

$$\hat{K} = \arg \min_{0 \leq k \leq k_{max}} IC(k) \quad (14)$$

As established earlier, this objective minimizes the residual variance $V(k)$ while applying the penalty $g(N, T)$ to prevent overfitting as the panel dimensions grow. As long as the penalty function satisfies specific asymptotic conditions, this approach consistently estimates the true number of underlying statistical factors.

This econometric logic has been subsequently extended to handle the complexities of modern, high-dimensional trading environments. Notably, Ait-Sahalia and Xiu (2017) adapt the principles of consistent factor selection to continuous-time models using high-frequency data. They demonstrate that PCA can successfully recover the latent factor space even when the sampling frequency and the number of assets increase simultaneously. Furthermore, they explicitly bridge the gap between statistical estimation and financial application by showing that their PCA-based covariance estimator significantly outperforms the standard sample covariance matrix in out-of-sample portfolio allocation.

While these econometric advances successfully recover the actual data-generating process, they remain subject to the fundamental limitation mentioned previously: they optimize for statistical fit rather than investor utility. Because these criteria are designed to minimize residual variance and are indifferent to the pricing of risk, they may obscure low-variance components that are nevertheless valuable for robust out-of-sample portfolio construction.

2.5.3 Random matrix theory approaches

A third major approach to identifying systematic risk factors is based on mathematical theory, specifically Random Matrix Theory (RMT). RMT provides a framework for distinguishing between eigenvalues that represent true underlying structure (signal) and those that arise purely from random sampling (noise). The core idea is to compare the empirical distribution of the sample eigenvalues against a theoretical distribution, most notably the Marcenko Pastur law, which describes the eigenvalues of a purely random matrix with no true correlations.

To formalize this mathematically, let $c = N/T$ denote the concentration ratio of the dataset. Under the Marcenko Pastur law, the eigenvalues of a random noise matrix with uniform variance σ^2 will asymptotically, in the limit as $N, T \rightarrow \infty$ such that $N/T \rightarrow c$, fall within a strict bounded interval. The upper bound of this noise distribution is defined as:

$$\lambda_{max} = \sigma^2(1 + \sqrt{c})^2 \quad (15)$$

In a standard RMT application, only the principal components with empirical eigenvalues strictly greater than this theoretical upper bound ($\lambda_i > \lambda_{max}$) are retained as carrying genuine economic signal; the rest are discarded as spurious.

The conceptual foundations of RMT have deeply influenced modern portfolio theory, particularly through advanced covariance matrix estimation. As extensively detailed by Ledoit and Wolf (2022), RMT provides the analytical machinery required to understand exactly how sample eigenvalues become distorted in high-dimensional settings. Building on this, they develop a nonlinear shrinkage estimator that systematically corrects these distorted sample eigenvalues, pushing extreme values toward the center. While Ledoit and Wolf (2022) use RMT to continuously shrink the eigenvalues rather than impose a hard discrete truncation, the underlying objective remains the same: filtering out the idiosyncratic noise that sabotages mean-variance optimizers.

Despite its mathematical elegance in cleaning the covariance matrix, RMT shares the same fundamental limitation as the econometric criteria: it is purely a statistical filter. By classifying components strictly based on structural signal versus random noise, it ignores the pricing of risk and the investor's ultimate goal of maximizing out-of-sample risk-adjusted returns. This persistent gap across explained variance heuristics, econometric criteria, and random matrix approaches, where statistical validity does not guarantee economic optimality, forms the precise motivation for the empirical analysis that follows.

2.6 Summary and research gaps

The previous sections highlight a fundamental tension in modern asset allocation. Although mean-variance optimization is theoretically compelling, its usefulness in the real world is frequently limited by the estimation errors it relies on. As the cross-section of assets grows large relative to the available time-series, the sample covariance matrix becomes increasingly ill-conditioned. To mitigate this issue, the literature has increasingly turned to PCA to reduce dimensionality, filter out idiosyncratic noise, and stabilize portfolio weights.

However, a critical review of the literature reveals a persistent structural gap. Established methodologies for determining the optimal truncation point K , whether based on explained variance heuristics, high-dimensional econometric criteria of Bai and Ng (2002), or the eigenvalue filtering of RMT, are exclusively anchored in statistical objectives. These approaches are designed to minimize reconstruction error, achieve asymptotic convergence to the data-generating process, or isolate non-random signals. They are entirely agnostic to pricing risk, transaction costs and the investor's ultimate goal of achieving out-of-sample, risk-adjusted performance.

This disconnect implies that the statistically true number of factors is not necessarily the economically optimal number of factors for portfolio construction. An active investor might benefit from retaining a statistically insignificant component if it provides a valuable, uncorrelated risk premium, or from excluding a statistically valid component if its inclusion merely amplifies estimation variance and portfolio turnover.

This gap forms the explicit motivation for this study. By shifting the focus from statistical factor selection to portfolio driven factor selection, this paper seeks to answer the following primary research question:

What is the gain in out-of-sample portfolio performance from selecting the number of principal components optimally using a portfolio motivated objective function?

To systematically address this central problem, the subsequent empirical analysis is guided by three specific sub-questions:

1. **How does the out-of-sample portfolio performance evolve as the number of PCs varies?** Before establishing a selection rule, it is necessary to empirically map the economic bias-variance trade-off. This involves analyzing how the out-of-sample Sharpe ratio, volatility, and turnover behave as the portfolio moves from a highly restricted factor model (high bias) to a fully parameterized sample covariance matrix (high variance).
2. **How sensitive is the optimal number of PCs to market conditions and sample size?** The impact of estimation error is heavily dependent on the concentration ratio (N/T) and the prevailing macroeconomic regime. This question explores whether the optimal portfolio maximizing K is a static parameter or a dynamic feature that shifts during periods of severe market stress or varying estimation window lengths.
3. **Does selecting the number of principal components using a portfolio driven objective function lead to superior out-of-sample performance compared to standard selection rules?** This question directly tests the practical value of the proposed framework. It benchmarks the out-of-sample performance of the portfolio driven optimal K against the traditional statistical criteria reviewed in this chapter, quantifying the specific economic cost of relying on purely statistical factor selection.

By answering these questions, this study aims to bridge the gap between econometric factor modeling and practical portfolio optimization, offering a new perspective on dimensionality reduction in asset allocation.

3 Methodology

3.1 Data description

In order to empirically investigate the optimal number of principal components for portfolio selection, this study uses three distinct datasets of daily asset returns. The selected data include both individual equities and standard characteristic-sorted portfolios, providing a diverse testing ground that reflects both high-dimensional idiosyncratic noise and established systematic risk factors.

The sample period for all datasets starts from January 1990 to September 2025. This extended horizon is highly relevant for out-of-sample evaluation, as it incorporates multiple distinct macroeconomic regimes and periods of severe market stress, including the Dot-com bubble, the 2008 global financial crisis, the COVID-19 pandemic, and the subsequent inflationary environment. It is crucial to note that utilizing daily data over a 35-year period provides a significant time-series dimension (T). This large T mathematically justifies the inclusion of a large cross-sectional dimension (N), as it ensures that the sample covariance matrix has sufficient degrees of freedom to be estimated robustly before any PCA truncation is applied.

The datasets are structured as follows:

- **S&P 500 individual equities:** The primary dataset consists of daily returns for 286 individual stocks that are, or have been, constituents of the S&P 500 index. To maintain a balanced panel required for consistent covariance matrix estimation over the 35-year horizon, the sample is restricted to those equities with continuous daily return data available from 1990 to 2025. While restricting the universe to continuously quoted assets introduces a degree of survival bias, it provides the stable, high-dimensional environment ($N = 286$) needed to isolate and test the variance-filtering mechanics of PCA without the confounding effects of dynamic asset insertion and removal. We acknowledge that this survival bias naturally increases the expected returns of benchmark strategies. However, this effectively creates a more conservative testing environment, which raises the bar for our proposed PCA_{cv} method to demonstrate a reduction in relative risk.
- **100 Portfolios formed on size and book-to-market:** Sourced from the Kenneth R. French Data Library, this dataset contains the daily returns of 100 portfolios formed on the intersections of ten size (market capitalization) deciles and ten book-to-market equity deciles. These portfolios serve as a standard benchmark in empirical finance. Because they are pre-sorted on known pricing anomalies, they possess a stronger, built-in factor structure compared to individual equities, allowing us to test the PCA selection rules on an asset universe where the "true" systematic signals are theoretically lower-dimensional.
- **48 Industry Portfolios:** Also sourced from the Kenneth R. French Data Library, this dataset comprises daily returns for 48 industry-sorted portfolios. Unlike the size and value portfolios, industry portfolios capture distinct macroeconomic sensitivities and sector-specific shocks. Including this dataset allows us to evaluate the robustness of the PC-optimized portfolios across assets clustered by economic function rather than purely

statistical characteristics.

Dataset	Dimensionality (N)	Frequency	Source
S&P 500 Equities	286	Daily	Yahoo Finance
Size/BTM Portfolios	100	Daily	Kenneth R. French Data Library
Industry Portfolios	48	Daily	Kenneth R. French Data Library

Table 1: Summary of Datasets

3.2 Portfolio construction

To evaluate the impact of dimensionality reduction on asset allocation, we construct and rebalance portfolios dynamically over time. At each rebalancing step t , the investor uses a historical rolling-window of $T = 504$ trading days (approximately two years) to estimate the covariance matrix $\hat{\Sigma}_t$ and the expected returns $\hat{\mu}_t$. Expected returns are estimated using the simple historical sample mean of daily returns over this rolling 504-day window. This approach allows the portfolio to adapt to time-varying market conditions, generating a series of out-of-sample returns to evaluate performance.

Building on the theoretical framework outlined in the literature review, we now adapt the classical Markowitz formulations to our specific empirical design. While subsection 2.5 defined the statically truncated covariance matrix as $\hat{\Sigma}_K$, our backtesting methodology applies this reduction dynamically over a rolling-window. Instead of relying on static, full-rank sample estimators, the optimization at time t depends on a filtered covariance matrix, $\hat{\Sigma}_{K,t}$, reconstructed from K principal components, alongside conditional expected returns $\hat{\mu}_t$. To explore the bias-variance trade-off comprehensively, we evaluate two distinct optimization strategies:

1. The Global Minimum Variance (GMV) Portfolio:

As established in the literature, isolating the risk-minimization objective mitigates the severe estimation errors associated with sample means. The GMV optimization problem at time t , utilizing a truncated covariance matrix $\hat{\Sigma}_{K,t}$ reconstructed from K principal components, is formulated as:

$$\min_{w_t} w_t^\top \hat{\Sigma}_{K,t} w_t \quad \text{s.t.} \quad w_t^\top \mathbf{1} = 1. \quad (16)$$

Provided this filtered matrix is adequately regularized to guarantee full rank and invertibility, the optimal weights admit a straightforward closed-form solution:

$$w_{gmv,t} = \frac{\hat{\Sigma}_{K,t}^{-1} \mathbf{1}}{\mathbf{1}^\top \hat{\Sigma}_{K,t}^{-1} \mathbf{1}}. \quad (17)$$

2. The Tangency (Maximum Sharpe Ratio) Portfolio:

While the GMV portfolio tends to be empirically more robust out-of-sample, the Tangency portfolio remains the theoretical optimum for a mean-variance investor seeking to maximize risk-adjusted returns. Including this portfolio allows us to investigate whether PCA-based dimensionality reduction can adequately stabilize weights when expected returns are explicitly

targeted. The objective of maximizing the Sharpe ratio yields the following optimal weights:

$$w_{tan,t} = \frac{\hat{\Sigma}_{K,t}^{-1} \hat{\mu}_t}{\mathbf{1}^\top \hat{\Sigma}_{K,t}^{-1} \hat{\mu}_t}. \quad (18)$$

3.3 Proposed portfolio-driven approach for selecting the number of principal components

For the traditional statistical benchmarks, the number of retained components K at each time t is determined by static rules (e.g., explained variance thresholds or RMT bounds). However, these rules optimize for statistical fit rather than economic utility.

To address this limitation, we propose a dynamic, "portfolio-driven" approach. The core logic is to let the data determine the optimal number of components by evaluating how different values of K actually perform in terms of risk and return. To achieve this without introducing look-ahead bias, we implement a purged rolling-window cross-validation procedure exclusively within the available historical data.

Standard cross-validation techniques (e.g., K-fold) randomly partition data, which violates the chronological order of financial time-series and leads to data leakage. Furthermore, because daily financial returns often exhibit serial correlation, simply splitting past data chronologically is insufficient. Therefore, we rely on a "purged" cross-validation framework. This technique introduces a strict purging margin: an excluded buffer period of several days, between the training data and the validation data to ensure the "safe" validation zone is completely independent of the training period.

Mechanically, at each rebalancing step t , the algorithm executes the following step to find the optimal $\hat{K}_t$:

1. **Grid search definition:**

We define a grid of candidate values for $K \in \{1, 2, \dots, K_{max}\}$. To balance computational efficiency with economic relevance, the upper bound K_{max} is dynamically set to $\min(N, 30)$. This truncation is motivated by two primary considerations. First, the empirical factor pricing literature consistently demonstrates that a small number of pervasive principal components (typically fewer than ten) accounts for the dominant share of cross-sectional covariance and expected returns (Bai & Ng, 2002; Kozak et al., 2020). Second, retaining 30 as a ceiling provides a sufficiently wide buffer to capture dataset-specific structural nuances while preventing the algorithm from needlessly evaluating the deep idiosyncratic noise known to trigger the estimation penalty.

2. **Inner data splitting and purging (expanding window):**

Instead of a single chronological split, we apply a rigorous time-series cross-validation over the historical $T = 504$ window, divided into 4 sequential folds. For each fold, the inner validation set consists of the final 42 trading days, while the inner training set is expanding, encompassing all historical observations prior to the validation block. To enforce strict independence and neutralize short-term autocorrelation, a strict purging

margin of 5 trading days is systematically excluded from the boundary between the training and validation sets.

3. Parameter estimation:

For each candidate K and across each of the 4 folds, we perform PCA exclusively on the inner training set. We extract the K components, reconstruct the truncated covariance matrix $\hat{\Sigma}_K$ (and estimate $\hat{\mu}$ for the Tangency portfolio), and compute the optimal portfolio weights.

4. Validation testing:

These optimized weights are then applied to the respective 42-day "safe" inner validation set to generate simulated out-of-sample portfolio returns.

5. Objective optimization:

We calculate the performance metrics exclusively on this consolidated time-series. Drawing on the cross-validation principles established for portfolio calibration by DeMiguel, Garlappi, Nogales, and Uppal (2009) computing the metrics over the merged out-of-sample returns rather than averaging them across individual folds avoids evaluating performance on drastically reduced sample sizes. This significantly mitigates the small-sample noise that can otherwise distort volatility and Sharpe ratio estimations. For the GMV framework, we select the K that minimizes the realized volatility of this aggregated series. For the Tangency framework, we select the K that maximizes its Sharpe ratio. This winning value becomes our dynamically selected $\hat{K}_t$.

Once the optimal $\hat{K}_t$ is identified through this rigorous historical simulation, the subsets are eliminated. We apply PCA to the *entire* $T = 504$ rolling-window, extract exactly $\hat{K}_t$ components, and calculate the final portfolio weights. These are the weights deployed in the true, unobserved out-of-sample period at $t + 1$.

3.4 Out-of-sample performance evaluation

Let T_{oos} denote the total number of out-of-sample periods. At each rebalancing step t , the estimated portfolio weights $\hat{w}_t$ are held for one period, generating an out-of-sample portfolio return at $t + 1$:

$$R_{p,t+1} = \hat{w}_t^\top R_{t+1} \quad (19)$$

where R_{t+1} is the vector of realized asset returns at time $t + 1$.

The strategies are evaluated based on three primary out-of-sample performance metrics, chosen to reflect the dual objectives of risk minimization and utility maximization, as well as their practical feasibility:

- **Out-of-sample volatility ($\hat{\sigma}_{oos}$):**

As the primary objective of the GMV framework is strictly risk minimization, we measure

its success using the standard deviation of the out-of-sample portfolio returns:

$$\hat{\sigma}_{oos} = \sqrt{\frac{1}{T_{oos} - 1} \sum_{t=1}^{T_{oos}} (R_{p,t} - \hat{\mu}_{oos})^2} \quad (20)$$

where $\hat{\mu}_{oos}$ is the sample mean of the out-of-sample returns.

- **Out-of-sample Sharpe Ratio (SR_{oos}):**

To evaluate risk-adjusted performance, which is the explicit target of the Tangency portfolio framework, we calculate the out-of-sample Sharpe ratio. Assuming a risk-free rate of zero for daily returns, it is defined as:

$$SR_{oos} = \frac{\hat{\mu}_{oos}}{\hat{\sigma}_{oos}} \quad (21)$$

- **Portfolio Turnover (TO):**

Because the mathematical inversion of ill-conditioned covariance matrices amplifies estimation error, it frequently leads to extreme and unstable portfolio weights. We quantify this instability by measuring the average absolute change in portfolio weights at each rebalancing step. Crucially, as asset prices evolve during the holding period, the actual portfolio weights drift from their initial target allocation. Let $\hat{w}_{i,t+}$ represent the drifted weight of asset i just before the rebalancing at time $t + 1$. The average turnover across all periods is calculated as:

$$TO = \frac{1}{T_{oos}} \sum_{t=1}^{T_{oos}} \sum_{i=1}^N |\hat{w}_{i,t+1} - \hat{w}_{i,t+}| \quad (22)$$

Accounting for this price drift ($\hat{w}_{i,t+} \neq \hat{w}_{i,t}$) is essential, as it accurately reflects that even passive baseline strategies, such as the equally weighted $1/N$ portfolio, require active trading to restore their target weights and therefore exhibit strictly non-zero turnover.

3.5 Benchmark portfolio strategies

To isolate and quantify the economic value added by the proposed portfolio-driven cross-validation approach, the dynamically optimized PCA portfolios are evaluated against six established allocation strategies. These benchmarks encompass both heuristic diversification and traditional statistical dimensionality reduction techniques. For each of the PCA-based benchmarks, both the GMV and Tangency frameworks are implemented.

1. **Naive Diversification ($1/N$):**

The equally weighted portfolio, defined mathematically as $w_{ew} = \frac{1}{N} \mathbf{1}$. As demonstrated by DeMiguel, Garlappi, and Uppal (2009), this simple heuristic completely ignores all sample moments, thereby avoiding estimation error entirely. It is a notoriously difficult out-of-sample benchmark and provides a baseline against which to assess whether optimization adds any real economic value.

2. **Sample Covariance (Full N components):**

The standard Markowitz optimization utilizing the full sample covariance matrix $\hat{\Sigma}$ (and

the full sample mean vector $\hat{\mu}$ for the Tangency portfolio) without any PCA truncation. This serves as the baseline to measure the full, unmitigated impact of estimation error and the curse of dimensionality.

3. PCA (Explained Variance Threshold):

A standard statistical heuristic where the number of retained components K is chosen to explain a fixed cumulative percentage of the total variation in the data. Following standard industry practice, this threshold is typically set to capture the majority of the variance (e.g., 90% or 95%), discarding the remaining low-variance components as noise.

4. PCA (Random Matrix Theory):

This benchmark relies on the Marchenko-Pastur upper eigenvalue bound to filter out noise aggressively (see subsection 2.5.3). Because our PCA operates on the asset *correlation* matrix, where variables are standardized to unit variance ($\sigma^2 = 1$), the general bound $\lambda_{max} = \sigma^2(1 + \sqrt{N/T})^2$ simplifies. We therefore only retain those components where $\lambda_i > (1 + \sqrt{N/T})^2$. While mathematically rigorous, RMT assumes stationary and independent data. In financial markets characterized by volatility clustering, this heuristic proves highly conservative, often discarding weaker but economically meaningful risk factors.

5. PCA (Information Criterion):

A high-dimensional consistent estimator framework. We select the optimal K using the IC_{p1} criterion proposed by Bai and Ng (2002). This method balances residual variance against a structural penalty function $g(N, T)$ to asymptotically recover the "true" statistical number of latent factors.

6. Ledoit-Wolf Shrinkage:

A robust continuous regularization technique proposed by Ledoit and Wolf (2022). Rather than discarding data through hard thresholding, this method mitigates estimation error by pulling the noisy sample covariance matrix towards a highly structured, low-variance target (specifically, a diagonal matrix) via a convex linear combination. It serves as a critical benchmark for evaluating the structural trade-off between the discrete eigenvalue truncation of PCA and continuous matrix shrinkage in high-dimensional settings

By benchmarking our dynamically selected $\hat{K}_t$ against these static, purely statistical criteria, we can answer the central research question directly: Does aligning the dimensionality reduction process with a portfolio-motivated objective function yield superior out-of-sample performance compared to traditional statistical factor selection?

4 Empirical results and analysis

After establishing the theoretical mechanics of our cross-validated PCA methodology, this section focuses on its empirical outcomes. By implementing these strategies across asset universes with varying degrees of intrinsic factor structure, the subsequent analysis is structured to progressively unpack the economic value of dimensionality reduction. We begin by examining the sensitivity of out-of-sample performance to a static truncation of the covariance matrix, seeking to empirically identify the threshold where signal extraction gives way to the overfitting of idiosyncratic noise. Building on these initial observations, we evaluate the proposed dynamic factor selection rules, comparing our portfolio-driven objective against standard statistical heuristics while testing its robustness to the length of the estimation window. The analysis then broadens to benchmark these PC-optimized allocations against traditional investment strategies, including a stress-test of their resilience during severe market crises. Ultimately, to bridge the gap between theoretical econometrics and applied financial engineering, the final part of this section focuses on structural stability, evaluating whether the theoretical advantages of dimensionality reduction survive the practical friction of proportional transaction costs.

4.1 The structural failure of the Tangency Portfolio

Before examining the volatility dynamics, it is essential to address the behavior of the Tangency portfolios. Across all three datasets, the out-of-sample Sharpe ratio for the Tangency model proved to be highly erratic and predominantly negative. For instance, dropping to -0.37 for $K = 10$ within the S&P 500 equities sample. This instability aligns closely with the "error maximization" property of classical mean-variance optimization. Mechanically, within our specific setup, estimating the expected return vector ($\hat{\mu}$) using daily data over a 2-year rolling-window ($T = 504$) produces a highly noisy and statistically insignificant signal. While a 504-day sample provides sufficient observations to estimate a robust covariance matrix, the sample mean of daily returns is heavily dominated by short-term variance and lacks long-term predictive power. The optimizer inadvertently treats these transient historical anomalies as certain future trends, magnifying the estimation errors into extreme and unstable portfolio weights. Consequently, to cleanly evaluate the impact of dimensionality reduction, the remainder of this analysis focuses exclusively on the risk-reduction capabilities of the GMV portfolio, which relies solely on the more structurally robust estimation of the covariance matrix.

4.2 Impact of the number of principal components on out-of-sample performance

The initial objective of our empirical analysis is to assess the sensitivity of out-of-sample portfolio performance to the static selection of the number of retained principal components (K). To isolate this effect, we constrain the PCA model to reconstruct the covariance matrix using a fixed number of factors, ranging from $K = 1$ to $K = 10$, and observe the resulting annualized volatility and Sharpe ratio over the 35-year testing period.

Figure 1 illustrates the impact of adding principal components on the out-of-sample annualized

volatility of the GMV portfolios across our three datasets. The exact numerical values are detailed in Table A.1 in the Appendix.

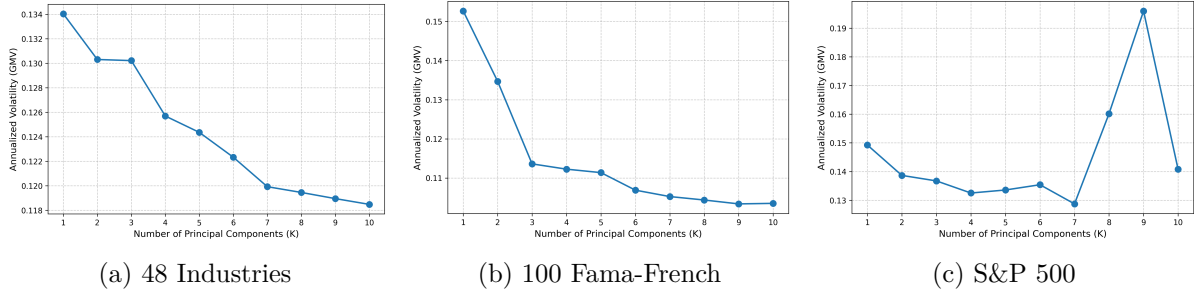


Figure 1: Impact of the number of principal components (K) on out-of-sample annualized GMV volatility.

Conceptually, the volatility trajectories presented in these graphs serve as an economic adaptation of the statistical scree plot discussed in the literature review. While a traditional scree plot graphs eigenvalues to visually locate an elbow, "beyond which components are discarded as a scree" of noise, Figure 1 maps this diagnostic logic directly onto out-of-sample portfolio risk. By substituting statistical variance with realized volatility, the analytical objective remains the identification of an empirical "elbow" in the curve. In this applied context, however, this break represents the optimal number of components K where the marginal benefit of capturing additional systematic factors is ultimately offset by the estimation penalty of overfitting unstable, low-variance directions.

Observing these volatility trajectories reveals three key dynamics relating to dimensionality reduction in portfolio construction.

1. The dominance of the initial market factors ($K = 1$ to $K = 3$).

Across all datasets, we observe an abrupt and pronounced initial decline in volatility when introducing the first few principal components. For the 100 Fama-French portfolios, for instance, the annualized volatility drops significantly from 15.26% at $K = 1$ to 11.36% at $K = 3$. This steep initial slope suggests that the primary eigenvector captures broad systematic market risk, while the immediate subsequent components likely reflect prevalent cross-sectional risk premia such as the size and value factors.

2. Diminishing marginal returns and stepwise reductions.

For datasets of moderate dimensionality, the volatility reduction follows distinct trajectories, though these diminish over time. The 100 Fama-French portfolios exhibit a classic "L-shape" or empirical elbow pattern, where volatility drops sharply before hitting a structural floor at $K = 3$, after which the curve flattens considerably. In contrast, the 48 Industry portfolios display a more staggered, stepwise reduction: the volatility initially decreases, forms a temporary plateau between $K = 2$ and $K = 3$, and then declines again at $K = 4$.

Despite these different path shapes, both datasets ultimately demonstrate diminishing marginal returns. By the time the model reaches $K = 7$, the addition of a component offers only a

negligible reduction in risk (decreasing from 12.23% to 11.99% for the 48 industries, and from 10.69% to 10.52% for the 100 Fama-French). This suggests that once a dataset-specific threshold is reached, most of the variance structure relevant to diversification has already been extracted.

3. The curse of dimensionality and overfitting and the estimation penalty (The S&P 500 Case)

Perhaps the most revealing dynamic occurs within our highest-dimensional sample, the S&P 500 equities ($N = 286$). While the volatility decreases steadily to reach a minimum at $K = 7$ (12.87%), it then begins to deteriorate, rising to 16.02% with the introduction of an eighth component. This trajectory serves as a practical illustration of the estimation penalty discussed previously, highlighting the danger of overfitting in large asset universes. When N is large relative to the time-series depth, the retention of lower-order components risks incorporating idiosyncratic, asset-specific noise rather than generalized systematic risk. The optimizer treats this noise as an exploitable signal, which ultimately degrades the out-of-sample portfolio weights.

It is particularly important to highlight the pronounced spike in out-of-sample volatility that occurs precisely at $K = 9$, where it reaches an extreme of 19.60%. A plausible hypothesis for this sudden deterioration is that the ninth principal component captures a highly localized, sample-specific noise pattern rather than a persistent systematic risk factor. When the GMV optimization algorithm incorporates this unstable direction, it severely exacerbates estimation error, leading to fragile portfolio weights that perform exceptionally poorly out-of-sample. The subsequent sharp drop in volatility at $K = 10$ (14.08%) suggests that the tenth component may act as an empirical counterweight or regularizer, partially offsetting the specific instability introduced by $K = 9$. Ultimately, this erratic trajectory serves as a cautionary boundary for static selection rules in high-dimensional equity universes.

4.3 Optimal number of components and comparison to usual selection rules

Building on the evidence from the previous section that a static K is suboptimal, the present section focuses on the evaluation of the behavior of dynamic factor selection methods over the 35-year rolling out-of-sample period. We compare our proposed Purged Cross-Validation methodology that optimizes both the GMV and Tangency objectives, against two standard heuristics: the 90% cumulative explained variance threshold (PCA_{var}) and the Random Matrix Theory eigenvalue bound (PCA_{rmt}).

Table 2 summarizes the average number of principal components retained by each algorithm across the entire backtesting period.

Table 2: Average number of principal components (K) selected dynamically (1990–2025)

Selection Rule	48 Industries	100 Fama-French	S&P 500 Equities
Explained Variance 90% (PCA_{var})	18.78	20.82	87.15
Random Matrix Theory (PCA_{rmt})	3.39	2.94	10.29
Purged CV - GMV Objective ($\text{PCA}_{\text{cv-gmv}}$)	7.86	11.15	16.73
Purged CV - Tangency Objective ($\text{PCA}_{\text{cv-tan}}$)	12.79	9.92	9.89

Note: This table reports the average number of principal components retained by the different selection rules across the three asset universes over the out-of-sample period.

The empirical averages highlight marked conceptual disparities between the methodologies, particularly as the dimensionality (N) of the asset universe increases.

The 90% explained variance heuristic (PCA_{var}) appears to systematically overestimate the true number of systematic factors. This tendency is clearly amplified in high-dimensional spaces; for the S&P 500 dataset, it retains an average of 87 components. By forcing the algorithm to capture 90% of the total variance, the model likely incorporates a significant amount of idiosyncratic, asset-specific noise. This inclusion has the effect of reducing the precision of the covariance estimate, thus potentially offsetting the primary benefits of dimensionality reduction.

By contrast, the RMT estimator (PCA_{rmt}) imposes much stricter dimensionality penalty. The methodology is rooted in asymptotic thresholding, which is an aggressive filtering process that, on average, retains only 3.39 and 2.94 factors for the 48 Industry and 100 Fama-French datasets, respectively. Although this approach provides a robust protection against overfitting, there is a risk of eliminating lower-variance principal components that could still contain economically significant risk premia, especially those whose eigenvalues lie just below the theoretical RMT bound.

The Purged Cross-Validation approach (PCA_{cv}) seems to navigate the trade-off between these two extremes. Under the GMV objective, the algorithm dynamically converges to a moderate retention rate (approximately 8 to 11 factors for the smaller dataset, and 17 for the S&P 500). By optimizing the empirical out-of-sample risk directly rather than relying on rigid statistical boundaries, the cross-validated model adapts more naturally to the intrinsic complexity of the dataset.

Switching to the Tangency objective reveals a highly distinct, dataset-dependent behavior, highlighting the significant difficulty of forecasting expected returns (μ). Most notably, within the high-dimensional S&P 500 sample, the Tangency framework imposes a much stricter dimensionality bottleneck, retaining an average of only 9.89 components compared to 16.73 for the GMV. As individual equities exhibit a notoriously low signal-to-noise ratio, the cross-validation algorithm learns to aggressively penalize lower-order principal components. It correctly identifies that attempting to extract risk premium signals from these volatile dimensions exacerbates the estimation penalty, ultimately degrading out-of-sample performance.

Conversely, the behavior across the pre-sorted portfolios offers additional economic nuances. For the 100 Fama-French portfolios, the selected dimensionality for both the GMV and Tangency objectives converges almost perfectly (11.15 and 9.92, respectively). This suggests, although

not conclusively, that among the assets ranked according to their anomalies, the pervasive eigenvectors driving covariance risk are essentially the same as those driving return premia. Interestingly, the 48 Industry portfolios exhibit an inverted dynamic: the Tangency objective favors a higher retention rate (12.79) than the GMV (7.86). A possible explanation lies in the nature of sector returns; generating risk-adjusted performance across macroeconomic industries likely requires capturing more granular, sector-specific shocks. While removing these secondary components might optimally minimize pure variance, it eliminates the cross-sectional signals necessary for the Tangency optimizer to function effectively.

4.3.1 Robustness check: sensitivity to the estimation window

In order to verify that the CV methodology is not spuriously dependent on the chosen 2-year rolling-window ($T = 504$ days), a sensitivity analysis was conducted. This involves recalculating the GMV out-of-sample volatility using a shorter 1-year window ($T = 252$) and a longer 4-year window ($T = 1008$).

The results, summarized in Table A.3, reveal a divergence based on the nature of the asset universe. For the pre-sorted portfolios, the 2-year baseline window remains the optimal spot, balancing the need for sufficient data points with the risk of obsolete information. For the 48 Industry portfolios, the baseline volatility of 12.36% (at $T = 504$) rises slightly to 12.37% for $T = 252$, and climbs to 13.23% for $T = 1008$. Similarly, the minimum volatility for the 100 Fama-French portfolios is achieved at $T = 504$ (10.52%), deteriorating if the window is extended to four years (11.96%).

However, the dynamic shifts drastically for the S&P 500 sample. Within this universe of individual equities, the sensitivity to the estimation window is particularly evident, with an absolute gap of 2.21 percentage points between the shortest and longest windows. Furthermore, in contrast to the diversified portfolios, the S&P 500 achieves its absolute minimum volatility using the shortest memory window ($T = 252$, yielding 12.65%).

This implies that while macroeconomic sector and factor correlations change progressively (favoring a 2-year memory), the underlying correlation structures of individual stocks are highly non-stationary and subject to rapid regime changes. For large, noisy equity universes, older data rapidly becomes toxic. By relying on a shorter 1-year window, the PCA model adapts more swiftly, dynamically discarding outdated covariance structures that would otherwise severely degrade out-of-sample performance.

Table 3: Sensitivity to estimation window length (T)

Asset Universe	PCA-CV GMV Volatility (Annualized)		
	$T = 252$ (1 Year)	$T = 504$ (2 Years, Baseline)	$T = 1008$ (4 Years)
48 Industry Portfolios	0.1237	0.1236	0.1323
100 Fama-French Portfolios	0.1077	0.1052	0.1196
S&P 500 Equities	0.1265	0.1378	0.1486

4.4 Performance of PC-optimized portfolios relative to benchmark strategies

To assess the economic value of our portfolio-driven component selection, we compare the out-of-sample performance of the cross-validated PCA strategies (PCA_{cv}) against baseline allocations and standard heuristic rules. Table 4 presents these results, allowing us to observe how out-of-sample performance evolves when moving from naive diversification to full covariance estimation, and finally to dimensionally reduced portfolios.

Table 4: Out-of-Sample Performance Across Datasets and Optimization Frameworks

Strategy	Global Minimum Variance		Tangency Portfolio	
	Volatility (Ann.)	Sharpe Ratio	Volatility (Ann.)	Sharpe Ratio
Panel A: 48 Industry Portfolios				
Naive ($1/N$)	0.1803	0.6814	0.1803	0.6814
Sample Cov Matrix ($K=N$)	0.1130	0.7102	6.5741	-0.1027
Ledoit-Wolf Shrinkage	0.1121	0.7198	96.5350	0.1636
PCA (90% Var)	0.1196	0.6618	40.0397	-0.0385
PCA (RMT)	0.1299	0.4982	18.2336	-0.1080
PCA (Cross-Validated)	0.1232	0.5021	64.5887	0.0874
Panel B: 100 Fama-French Portfolios				
Naive ($1/N$)	0.1989	0.6456	0.1989	0.6456
Sample Cov Matrix ($K=N$)	0.1029	1.5891	68.7740	-0.0903
Ledoit-Wolf Shrinkage	0.1169	-0.3856	7.5964	-0.1454
PCA (90% Var)	0.1069	1.5306	20.7451	-0.1065
PCA (RMT)	0.1178	1.2178	16.9666	0.1579
PCA (Cross-Validated)	0.1047	1.7113	26.6864	-0.1137
Panel C: S&P 500 Equities				
Naive ($1/N$)	0.1908	0.6132	0.1908	0.6132
Sample Cov Matrix ($K=N$)	0.1622	0.3190	205.1937	-0.2304
Ledoit-Wolf Shrinkage	0.1290	0.4862	47.4312	0.1319
PCA (90% Var)	0.1511	0.5438	270.4616	0.1964
PCA (RMT)	0.1391	0.2628	32.2014	0.1511
PCA (Cross-Validated)	0.1350	0.4478	265.3359	0.1336

Note: Best performance metrics within each panel and optimization framework are highlighted in bold.

Firstly, examining the GMV framework reveals that the effectiveness of dimensionality reduction is highly sensitive to the underlying eigenvalue structure of the asset universe. For the 48 industry portfolios (Panel A), however, the picture is somewhat counterintuitive. The unconstrained sample covariance matrix ($K = N$) and Ledoit-Wolf shrinkage dominate, while the PCA_{cv} approach yields a significantly lower Sharpe ratio of 0.5021. A possible explanation lies in the relatively low dimensionality of this specific cross-section. When N is small, the estimation risk embedded in the sample covariance matrix is often manageable over standard rolling-windows. Truncating the eigenvalue structure via PCA in this environment may inadvertently discard lower-order principal components containing subtle but valuable idiosyncratic correlation structures. This highlights a delicate trade-off: forcing dimensionality reduction when N is small degrades out-of-sample risk-adjusted returns by prioritizing variance reduction over the preservation of useful diversification information.

When expanding the analysis to the 100 Fama-French portfolios (Panel B), the value of a portfolio-driven selection objective becomes much clearer. In this dataset, the PCA_{cv} strategy achieves an outstanding Sharpe ratio of 1.7113, comfortably outperforming both the full sample matrix and standard heuristic rules like 90% variance capture. Because the Fama-French sorts are heavily driven by latent pricing factors (e.g., size and value), the cross-validated objective function appears to successfully isolate these true common factors while aggressively filtering out the structural noise that limits other estimators. It is worth noting the anomalous negative Sharpe ratio for the Ledoit-Wolf shrinkage in this panel. This reminds us that out-of-sample risk-adjusted returns can be highly sample-dependent and sensitive to specific realized return paths, even for structurally robust estimators.

The limits of covariance estimation are most visibly tested in the S&P 500 universe (Panel C). With $N = 286$, the estimation error in the full sample covariance matrix is severe. Here, continuous regularization via Ledoit-Wolf performs best in stabilizing out-of-sample volatility (0.1290). The PCA_{cv} strategy performs respectably in minimizing volatility (0.1350) but fails to beat the naive $1/N$ portfolio on a risk-adjusted basis. Individual equities contain massive amounts of idiosyncratic noise compared to pre-sorted portfolios. These results suggest, although not conclusively, that identifying the "optimal" number of components is significantly harder when the signal-to-noise ratio of the underlying assets is low.

However, the destructive impact of estimation error is, starkly revealed when shifting to the Tangency portfolio. Since average returns are notoriously difficult to estimate out-of-sample, integrating them into the optimization process rapidly destabilizes the portfolio weights. The Sharpe ratio (0.0874) for the 48 industry portfolios suggests that the objective function could occasionally filter out noise in the expected return better than standard rules. Yet, it completely breaks down in Panels B and C. Ultimately, naive diversification systematically dominates the Tangency framework. This indicates that while cross-validating the number of principal components can yield superior performance in structurally dense, covariance-only applications, it cannot rescue mean-variance optimization from the toxic estimation risk inherent in the expected return vector.

4.4.1 Robustness check: performance during market crises

To verify whether these out-of-sample characteristics persist under severe market stress, we perform a robustness test by evaluating the strategies across three major downturns: the Dot-com bubble burst (2000-2002), the 2008 Global Financial Crisis, and the COVID-19 pandemic (2020). Analyzing behavior during periods of high-volatility is crucial, as historical correlations tend to break down and asset returns cluster together, which severely tests the stability of the inverted covariance matrix.

The empirical results reveal an interesting structural divergence depending on the dimensionality of the investment universe. While Table A.2 in the Appendix details the numerical performance of all allocation methods, including the unfiltered sample covariance, the visual analysis in Figures 2 through 4 focuses exclusively on three benchmarks: the naive $1/N$ allocation, the Ledoit-Wolf shrinkage, and our proposed PCA_{cv} strategy.

During the Dot-com bubble, the PCA_{cv} strategy exhibits strong downside protection across all datasets. In the larger S&P 500 set, it delivers a solid 50.09% total return, outperforming the full-covariance strategy's 41.10%, Ledoit-Wolf shrinkage 42.18%, and significantly exceeding the slight negative return of the naive portfolio -1.64%. Similarly dominant performance occurs in the 100 Fama-French portfolios 68.15%. However, in the narrow 48-industry universe, PCA_{cv} (16.15%) slightly lags behind both the full sample (18.40 and Ledoit-Wolf 17.76% approaches, hinting at the limitations of dimensionality reduction in smaller cross-sections.

The 2008 and 2020 crises confirm this limitation and reveal a stark trade-off between estimation error and information loss. As illustrated in Figures 2 and 3, for the smaller $N = 48$ and $N = 100$ portfolios, traditional estimators actually provide better downside protection than PCA_{cv} . For instance, in the 100-portfolio universe during 2008 (Figure 3), the PCA_{cv} strategy returned -31.98%, performing worse than the unfiltered sample covariance (-25.00%) and Ledoit-Wolf shrinkage (-19.56%). These results suggest, although not conclusively, that when N is small relative to the estimation window T , retaining the full correlation structure might provide better hedging against systemic shocks than a truncated approximation.

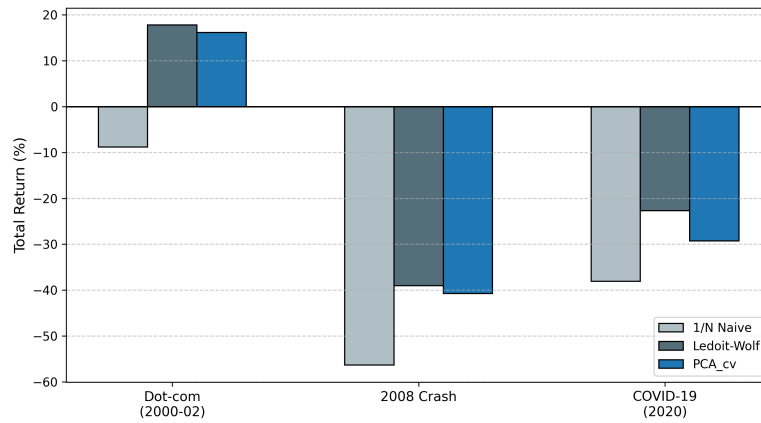


Figure 2: Total return of the competing GMV strategies during major market drawdowns: 48 Industry Portfolios.

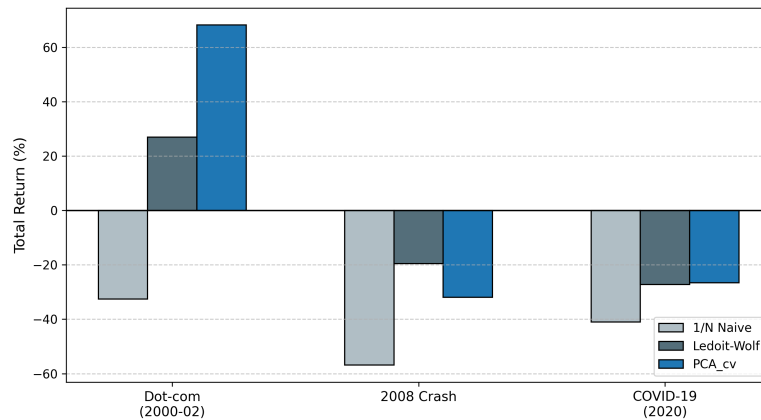


Figure 3: Total return during major market drawdowns: 100 Fama-French Portfolios.

The true economic value of the PCA-optimized approach is most evident in the high-dimensional S&P 500. As the ratio of assets to observations increases, the sample covariance matrix becomes increasingly ill-conditioned, magnifying estimation error and pushing the optimizer toward extreme, unstable weights which is illustrated in Figure 4. During the 2008 crash, the naive $1/N$ portfolio and the full sample covariance strategy experienced devastating drawdowns of -56.63% and -52.21%, respectively. In striking contrast, the PCA_{cv} strategy absorbed the shock remarkably well, limiting losses to -21.98% and easily outperforming Ledoit-Wolf (-42.56%). We observe a similar structural resilience during the COVID-19 sell-off, where PCA_{cv} (-21.67%) again outperformed both Ledoit-Wolf (-23.57%) and the full sample estimator (-30.53%).

A possible explanation lies in the change in the structure of asset return eigenvalues during market panics. In times of crisis, systemic risk increases, which is why the first principal components account for almost all of the cross-sectional variance. By dynamically selecting K via a portfolio-driven target, the PCA_{cv} strategy adaptively reduces the cross-section. It relies heavily on these dominant eigenvectors while eliminating the "tail" of lower-variance components, which largely become pure estimation noise during panics. This prevents the optimizer from adapting to illusory diversification opportunities, thus mitigating the significant estimation risk that affects unfiltered estimators on large datasets.

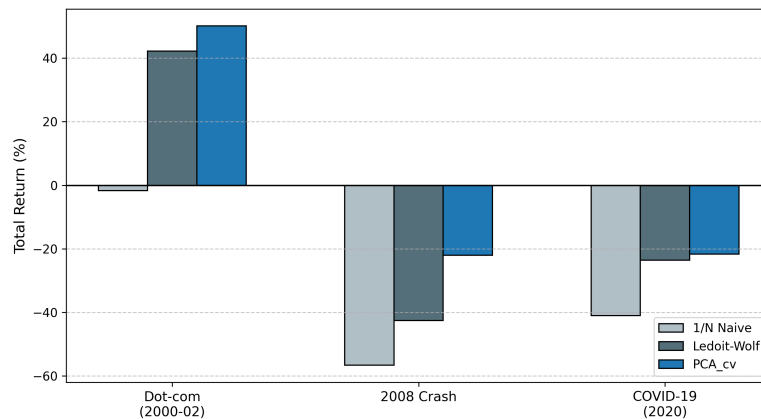


Figure 4: Total return during major market drawdowns: S&P 500 Equities ($N = 286$). The bar chart illustrates the severe performance degradation of the unfiltered sample estimator during market panics, contrasting with the structural resilience of the dimensionality-reduced PCA_{cv} allocation.

4.5 Portfolio stability and the impact of transaction costs

While our previous analysis highlighted the risk-reduction capabilities of dynamic dimensionality reduction, assessing the viability of an investment strategy requires looking beyond gross theoretical performance. Frequent rebalancing and extreme leverage can easily consume any statistical edge. To evaluate the practical implementability of these strategies, we analyze their daily turnover, maximum gross exposures, and net performance after fees, assuming proportional transaction costs of 10 basis points (0.10%) per trade.

Before analyzing the specific stability metrics detailed in Table 5, it is necessary to briefly address

the Tangency framework. As documented in Table A.3 in the Appendix, the unconstrained Tangency portfolios exhibit catastrophic structural instability across all dimensionalities. Highly sensitive to estimation errors in expected returns (μ), the optimization generates extreme, unrealistic long and short positions. This erratic behavior leads to a massive daily turnover that renders the strategy economically uninvestable, a degradation that persists even when Ledoit-Wolf shrinkage or cross-validated PCA method is applied. This confirms that dimension reduction on the covariance matrix alone is insufficient to stabilize mean-variance optimization when the mean vector is left unconstrained. Consequently, the remainder of our practical implementation analysis is strictly limited to the GMV portfolios.

Table 5: GMV Portfolio Stability: Turnover and Extreme Weights

Dataset	Strategy	Daily Turnover	Max Weight	Min Weight
48 Industries	Naive ($1/N$)	0.0073	0.02	0.02
	Full Sample	0.1087	0.88	-0.50
	Ledoit-Wolf Shrinkage	0.0923	0.76	-0.39
	PCA (Cross-Validated)	0.3956	0.78	-0.51
100 Fama-French	Naive ($1/N$)	0.0053	0.01	0.01
	Full Sample	0.2921	1.03	-0.43
	Ledoit-Wolf Shrinkage	0.2312	0.89	-0.30
	PCA (Cross-Validated)	0.6495	0.92	-0.40
S&P 500 Equities	Naive ($1/N$)	0.0123	0.00	0.00
	Full Sample	0.5511	0.32	-0.27
	Ledoit-Wolf Shrinkage	0.1208	0.17	-0.11
	PCA (Cross-Validated)	0.3761	0.13	-0.13

Note: The complete stability metrics across all PCA specifications, see Appendix Table A.3.

Focusing on the more robust GMV framework reveals a complex dynamic regarding turnover. In the lower-dimensional universes (48 and 100 portfolios), the full sample covariance matrix and Ledoit-Wolf shrinkage actually exhibit lower daily turnover than the PCA_{cv} approach. A plausible explanation lies in the nature of our dynamic methodology. Because the cross-validation algorithm continually adjusts the optimal number of components (K) in order to adapt to market fluctuations, the underlying subspace shifts frequently. This constant recalibration forces the optimizer to modify the portfolio weights, driving the turnover up to 0.6495 in the 100-portfolio dataset.

Interestingly, this relationship is reversed in the high-dimensional S&P 500 dataset. In this case, the traditional sample covariance matrix becomes so ill-conditioned and dominated by noise that its weights fluctuate wildly, resulting in a high turnover of 0.5511. The PCA_{cv} method successfully stabilizes the inverse covariance matrix and reduces the turnover to 0.3761 by forcefully truncating the noisy lower-order components. This suggests that in highly over-parameterized settings, dimensionality reduction acts as an implicit smoothing mechanism.

Nevertheless, we must acknowledge the solid benchmark set by the Ledoit-Wolf estimator, which achieves an impressively low turnover of 0.1208 in the S&P 500 universe, alongside highly constrained maximum and minimum weights. This argues that, although dynamic PCA is effective in mitigating extreme estimation error in large cross-sections, analytical shrinkage

towards a highly structured target provides superior weight stability. Ultimately, these results highlight a fundamental trade-off in portfolio engineering: the statistical optimality of dynamically selecting K to minimize out-of-sample variance must be carefully balanced against the economic friction induced by subspace instability.

4.5.1 Robustness check: the impact of transaction costs

Despite the theoretical risk-reduction benefits observed throughout this study, the introduction of a standard 10 bps transaction cost ¹ serves as a harsh reality check. As detailed in Table 6, the friction generated by rebalancing largely obliterates the out-of-sample edge of the optimized strategies.

Table 6: Impact of Transaction Costs on GMV Out-of-Sample Performance

Dataset	Strategy	Gross Performance		Net Performance (10 bps)	
		Volatility	Sharpe	Net Volatility	Net Sharpe
48IND	Naive (1/N)	0.1803	0.6814	0.1803	0.6712
	Sample Cov ($K = N$)	0.1130	0.7102	0.1130	0.4675
	Ledoit-Wolf	0.1121	0.7198	0.1121	0.5120
	PCA (CV)	0.1232	0.5021	0.1237	-0.3056
100FF	Naive (1/N)	0.1989	0.6456	0.1989	0.6389
	Sample Cov ($K = N$)	0.1029	1.5891	0.1032	0.8721
	Ledoit-Wolf	0.1169	-0.3856	0.1167	-0.8855
	PCA (CV)	0.1047	1.7113	0.1058	0.1462
S&P500	Naive (1/N)	0.1908	0.6132	0.1908	0.5970
	Sample Cov ($K = N$)	0.1622	0.3190	0.1623	-0.5368
	Ledoit-Wolf	0.1290	0.4862	0.1290	0.2500
	PCA (CV)	0.1350	0.4478	0.1354	-0.2533

While the Naive $1/N$ portfolio demonstrates resilience across all dimensionalities due to its low trading requirements (maintaining net Sharpe ratio above 0.59 across all datasets), the PCA_{CV} strategy is severely penalized. Its Sharpe ratio drops into negative territory for both the 48 Industry (-0.3056) and S&P 500 (-0.2533) datasets. In the context of the 100 Fama-French universe, where PCA_{CV} achieved a positive gross Sharpe ratio of 1.7113, the net Sharpe collapses to a marginal 0.1462 after accounting for trading friction.

These results unequivocally highlight a fundamental trade-off in modern quantitative asset allocation. The optimization of the number of principal components via a portfolio-driven objective successfully cleans the correlation matrix and mitigates out-of-sample volatility. However, the resulting dynamic rotation of the factor subspace induces a level of turnover that is economically unviable under standard trading frictions. For such dynamic dimensionality reduction techniques to be practically deployed, future iterations must explicitly integrate turnover penalties directly

¹Trading 286 individual equities is practically much more expensive than trading 48 highly liquid industry portfolios, meaning our 10 bps assumption is likely conservative for the S&P 500, making the turnover problem even more severe than reported here.

into the objective function, forcing the algorithm to balance statistical precision against the friction of execution.

5 Conclusion

This thesis set out to answer a deceptively simple question: How many principal components should an investor retain when constructing a mean-variance portfolio? The answer, as the empirical evidence makes clear, is that there is no single correct number, and that the criteria most commonly used to find one are asking the wrong question entirely. Standard selection rules based on explained variance thresholds, Random Matrix Theory, or the information criteria of Bai and Ng (2002) share a common objective: minimizing reconstruction error and achieving statistical consistency with the data-generating process. This thesis has argued, and empirically demonstrated across three datasets and a 35-year out-of-sample horizon, that statistical optimality and economic optimality are not the same thing, and that the gap between them has real financial consequences.

5.1 Findings

The empirical analysis yields four main findings, each of which contributes a distinct dimension to this central argument.

The first finding relates to the shape of the bias-variance trade-off in portfolio space. Across all three asset universes, the relationship between retained components and out-of-sample volatility is non-monotonic. The first few components significantly reduce risk by capturing broad market and cross-sectional risk premia; however, beyond this threshold, marginal benefits diminish rapidly. In high-dimensional settings, retaining further components actively degrades performance, since the optimizer treats idiosyncratic noise as an exploitable signal. This non-linearity is precisely what renders static selection rules structurally ill-equipped to navigate the bias-variance trade-off in portfolio space.

The second finding establishes that the economic value of portfolio-driven dimensionality reduction depends heavily on the concentration ratio N/T . In lower-dimensional settings, the estimation risk inherent in the full sample covariance matrix is manageable; however, truncating the eigenvalue structure may result in the discarding of genuinely valuable idiosyncratic correlation structures for diversification purposes. Conversely, in the highly over-parameterized S&P 500 universe, the cross-validated PCA demonstrates deep structural resilience during systemic shocks, limiting drawdowns to levels well below those of both the unfiltered estimator and Ledoit-Wolf shrinkage. The concentration ratio N/T appears to be a key predictor of whether portfolio-driven dimensionality reduction adds economic value, and practitioners should condition their methodological choice accordingly.

The third finding speaks to the relative merits of competing selection rules in structurally rich environments. Among the three datasets tested, the 100 Fama-French universe provides the clearest evidence in favor of a portfolio-driven objective. Because these portfolios are pre-sorted on known pricing anomalies, their latent factor structure is well-identified, and the cross-validated PCA objective successfully isolates these true common factors while filtering out structural noise. These two benchmarks illustrate the opposite failure modes of static selection rules. The RMT approach, while mathematically rigorous, proves consistently too conservative, retaining

an average of fewer than three components in lower-dimensional datasets and eliminating economically meaningful risk premia whose eigenvalues fall just below the Marchenko-Pastur bound. On the contrary, the 90% variance heuristic systematically over-includes components, largely defeating the purpose of dimensionality reduction. It is worth noting that in the S&P 500 universe, even the cross-validated PCA strategy fails to improve upon the naive $1/N$ benchmark on a risk-adjusted basis. This reinforces the second finding: when individual equities dominate the cross-section and the signal-to-noise ratio is low, continuous regularization provides superior weight stability over discrete eigenvalue truncation.

The fourth finding is the most consequential, and in many respects the central result of this thesis: the gross performance advantage of the cross-validated strategy is largely, and in some cases entirely, consumed by transaction costs. Assuming a standard proportional friction of 10 basis points per trade, the net Sharpe ratio collapses into negative territory for two of the three datasets, falling well below the naive $1/N$ benchmark. The structural cause of this erosion is the subspace instability inherent in dynamic factor selection: as the cross-validation algorithm continually recalibrates K in response to shifting market conditions, the retained eigenspace rotates frequently, forcing substantial revisions to portfolio weights. This finding does not merely qualify the thesis's main result, it reformulates it. The central challenge in portfolio-driven PCA is not identifying the right number of components in-sample; that problem is largely solved by the proposed methodology. The central challenge is stabilizing the resulting subspace sufficiently to make that precision economically worthwhile after deducting execution costs.

The Tangency portfolio analysis supports this conclusion from a different perspective: reducing dimensionality solely by applying it to the covariance matrix cannot compensate for the toxic estimation error embedded in the expected return vector. This makes unconstrained Tangency strategies economically unfeasible at daily frequencies across all tested dimensionality regimes.

5.2 Limitations and directions for future research

Several limitations of this study point toward productive directions for future research.

The most immediate and practically important extension concerns the integration of turnover penalties into the cross-validation objective function itself. The present framework selects K by minimizing out-of-sample volatility or maximizing the Sharpe ratio, with no direct penalty for the trading costs induced by subspace instability. A natural generalization would augment the objective with a regularization term penalizing changes in portfolio weights between rebalancing periods, forcing the algorithm to explicitly trade-off statistical precision against execution friction. Such a framework would target not the statistically optimal K , but the economically optimal one, a distinction this thesis has argued is fundamental.

A second extension concerns the estimation of expected returns. This thesis has focused specifically on the GMV framework in order to isolate the effect of covariance estimation. Future work could investigate whether combining portfolio-driven PCA with robust mean estimation techniques, such as shrinkage estimators for the return vector, momentum signals, or factor-based forecasts, could rehabilitate the Tangency objective within a dimensionality-reduced

framework. The pre-sorted structure of the 100 Fama-French portfolios, which embeds known return anomalies, makes this a suitable testing ground for such a hybrid approach.

Finally, two structural extensions merit further exploration. Firstly, the finding that the optimal K contracts sharply during market crises suggests the need for a regime-aware variant of the selection rule. By conditioning the algorithm on observable stress indicators such as the VIX or realized correlation, the model could predict changes in eigenvalue dispersion rather than merely adapting to them afterwards. Secondly, the present results suggest that the concentration ratio N/T acts as a first-order threshold below which portfolio-driven dimensionality reduction adds net economic value. However, this threshold is only approximately identified, somewhere between the $N = 48$ and $N = 100$ universes. A more systematic exploration across a wider range of asset counts and estimation window lengths would allow practitioners to tailor their methodological choice on the specific characteristics of their investment universe with greater precision.

5.3 Closing remarks

Taken together, the results of this thesis offer a nuanced answer to the question posed in its title. There is no single optimal number of principal components. The economically appropriate truncation point is a dynamic quantity that depends on the dimensionality of the investment universe, the prevailing market regime, the length of the estimation window, and critically, the transaction cost environment in which the strategy must operate. What this thesis establishes is that aligning the selection criterion with the investor's portfolio objective rather than a purely statistical measure of fit, produces meaningfully different, and in high-dimensional settings, empirically superior gross outcomes. The practical challenge that remains is not conceptual but structural: constructing a dynamic selection framework that is as sensitive to the investor's execution constraints as it is to the statistical properties of the return covariance matrix. Closing the gap between these two dimensions of optimality is, ultimately, what separates a theoretically compelling approach from a practically deployable one.

6 Appendix

Table A.1: Out-of-sample performance across fixed numbers of principal components (K).

Dataset	K	GMV Volatility	Tangency Sharpe
48 Industry Portfolios	1	0.1340	0.0207
	2	0.1303	-0.2550
	3	0.1302	-0.0626
	4	0.1257	-0.3679
	5	0.1244	-0.0124
	6	0.1223	-0.1381
	7	0.1199	0.0798
	8	0.1195	-0.0352
	9	0.1189	0.1414
	10	0.1185	0.0056
100 Fama-French Portfolios	1	0.1527	0.5202
	2	0.1347	0.1074
	3	0.1136	-0.1604
	4	0.1122	0.0252
	5	0.1114	-0.0754
	6	0.1069	0.3070
	7	0.1052	-0.0653
	8	0.1044	0.2056
	9	0.1034	-0.0331
	10	0.1035	0.3796
S&P 500 Equities	1	0.1493	-0.1522
	2	0.1387	0.0946
	3	0.1367	-0.1487
	4	0.1325	-0.0486
	5	0.1336	-0.3195
	6	0.1354	0.1844
	7	0.1287	0.1183
	8	0.1602	0.0336
	9	0.1960	-0.1400
	10	0.1408	-0.3701

Table A.2: Robustness checks: Total returns during historical market crashes.

Dataset	Market Crisis	Total Return (%)			
		1/N Naive	Sample Cov	Ledoit-Wolf	PCA _{cv} (GMV)
48 Industry	Dot-com (2000-2002)	-8.81%	+18.40%	+17.76%	+16.15%
	2008 Crash	-56.36%	-39.29%	-39.01%	-40.75%
	COVID-19 (2020)	-38.11%	-20.92%	-22.72%	-29.31%
100 Fama-French	Dot-com (2000-2002)	-32.62%	+35.50%	+26.99%	+68.15%
	2008 Crash	-56.89%	-25.00%	-19.56%	-31.98%
	COVID-19 (2020)	-40.98%	-26.24%	-27.26%	-26.56%
S&P 500	Dot-com (2000-2002)	-1.64%	+41.10%	+42.18%	+50.09%
	2008 Crash	-56.63%	-52.21%	-42.56%	-21.98%
	COVID-19 (2020)	-40.99%	-30.53%	-23.57%	-21.67%

Table A.3: Portfolio stability (Daily Turnover and Extreme Weights). The table reports turnover and max/min weight allocations for both the GMV and Tangency frameworks across the three primary datasets. Bold values indicate the lowest turnover among the active optimization strategies within the GMV framework.

Dataset	Strategy	GMV Framework			Tangency Framework		
		Turnover	Max W.	Min W.	Turnover	Max W.	Min W.
48 Industry	1/N Naive	0.0073	0.02	0.02	0.0073	0.02	0.02
	Full Sample	0.1087	0.88	-0.50	40.8164	347.30	-347.34
	Ledoit-Wolf	0.0923	0.76	-0.39	60.0748	9413.53	-6964.50
	PCA _{90%}	0.1663	0.80	-0.47	201.8739	4068.82	-4308.99
	PCA _{rmt}	0.0697	0.68	-0.30	128.1547	3314.32	-2758.08
	PCA _{cv}	0.3956	0.78	-0.51	143.4129	8747.42	-4686.84
100 Fama-French	1/N Naive	0.0053	0.01	0.01	0.0053	0.01	0.01
	Full Sample	0.2921	1.03	-0.43	82.0418	14574.35	-13888.92
	Ledoit-Wolf	0.2312	0.89	-0.30	28.6077	516.25	-568.30
	PCA _{90%}	0.2689	0.92	-0.29	89.0059	3754.44	-2059.59
	PCA _{rmt}	0.0932	0.90	-0.20	99.1580	2188.24	-1397.76
	PCA _{cv}	0.6495	0.92	-0.40	104.1649	1547.96	-2767.35
S&P 500	1/N Naive	0.0123	0.00	0.00	0.0123	0.00	0.00
	Full Sample	0.5511	0.32	-0.27	2051.8001	15432.80	-15009.52
	Ledoit-Wolf	0.1208	0.17	-0.11	162.9860	1598.31	-1387.34
	PCA _{90%}	0.2780	0.14	-0.14	238.1313	3131.94	-3279.53
	PCA _{rmt}	0.0978	0.12	-0.13	164.4332	3786.23	-3181.69
	PCA _{cv}	0.3761	0.13	-0.13	279.0127	17863.97	-13749.04

References

- Ait-Sahalia, Y., & Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, 201(2), 384–399.
- Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.
- Chamberlain, G., & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5), 1281–1304.
- Connor, G., & Korajczyk, R. A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics*, 21(2), 255–289.
- DeMiguel, V., Garlappi, L., Nogales, F. J., & Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5), 798–812.
- DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5), 1915–1953.
- Gelmini, M., & Uberti, P. (2024). The equally weighted portfolio still remains a challenging benchmark. *International Economics*, 179, 100525.
- Jolliffe, I. T. (2002). *Principal component analysis*. Springer-Verlag.
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.
- Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2), 271–292.
- Lassance, N., & Vrms, F. (2021). Portfolio selection with parsimonious higher comoments estimation. *Journal of Banking & Finance*, 126, 106115.
- Ledoit, O., & Wolf, M. (2022). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*, 20(1), 187–218.
- Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1), 77–91.
- Meucci, A. (2009). Managing diversification. *Risk*, 22(5), 74–79.
- Michaud, R. O. (1989). The Markowitz optimization enigma: Is ‘optimized’ optimal? *Financial Analysts Journal*, 45(1), 31–42.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm