UCLouvain

LOUVAIN
School of Management

Louvain School of Management

# Double Descent and Benign Overfitting in Portfolio Optimization

Author(s): Grégoire Vandooren
Supervisor(s): Nathan Lassance
Academic year 2025-2026
Dissertation for the master of Business Engineering
Master subject and focus Financial Engineering
Daytime schedule

## AI DECLARATION

This thesis has been prepared in accordance with the Louvain School of Management (LSM) guidelines regarding the responsible use of generative Artificial Intelligence.

In the interest of full transparency, I declare the specific ways in which generative AI tools were used in this work.

**DeepL / DeepL Write** was used strictly as a linguistic assistant: translating specific terms, improving phrasing, and correcting grammar and spelling throughout the manuscript.

**Claude and Gemini** were used as brainstorming tools to help structure the narrative, refine certain theoretical explanations, and format the LaTeX document. They also served as coding assistants to debug and optimize the R code used in the empirical analysis.

All suggestions were systematically reviewed and adapted to ensure full understanding and scientific accuracy.

| **Statements on responsible use of generative AI** | **Your answer** |
|---|---|
| The content of my master's thesis is my original output, even if I used generative AI to help prepare it. | ☒ Yes ☐ No |
| I master the theories, tools and methods that I use for the thesis. | ☒ Yes ☐ No |
| I 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 acknowledge that I master the final output and that I am able to explain it and answer questions about it. | ☒ Yes ☐ No |
| I made sure not to provide generative AI tools with access to confidential data or information. | ☒ Yes ☐ No |

**Sign your declaration:**

| **Author last name, first name(s)** | **Date** | **Signature** |
|---|---|---|
| Vandooren, Grégoire | 15/05/2026 |  |

## ACKNOWLEDGMENTS

I would like to start by thanking my supervisor, Professor Nathan Lassance. From our first call to discuss the topic, the data he provided for the empirical analysis and his feedback throughout the writing process, he was responsive at every stage. I am grateful for the time he invested in this work.

I would also like to thank my parents and my two sisters. Throughout my entire studies, they gave me the best possible environment to work and grow, and their support never wavered. None of this would have happened without them.

To my friends at LSM, thank you for being there during these five years, especially during the more stressful moments of this final stretch.

Finally, I want to mention my Erasmus semester at the University of Victoria in Canada. Discovering a new continent, new cultures and a completely different way of life was one of the most enriching experiences of my studies.

## Contents

## List of Figures

## List of Tables

# INTRODUCTION

Modern portfolio theory rests on a clear premise: given enough data, one can estimate the statistical structure of asset returns and use it to construct efficient allocations. In practice, this breaks down as soon as $N$ (the number of assets) grows large relative to $T$ (the number of observations): the sample covariance matrix becomes unreliable, and its inversion amplifies estimation noise rather than filtering it, producing portfolios that collapse out-of-sample. With monthly return data and realistic investment universes, the ratio $\rho = N/T$ routinely exceeds one, placing classical Markowitz optimization in a regime for which it was never designed.

The conventional response has been to treat high dimensionality as a pathology: shrink the covariance matrix, impose norm constraints, or reduce the asset universe. These approaches share a common principle: more complexity means more risk. Recent developments in statistical learning theory challenge this directly. Belkin et al. (2019) documented the *double descent* phenomenon, showing that prediction error can decrease again after a sharp peak at the interpolation threshold, precisely in the over-parameterized regime that classical theory condemns. Kelly et al. (2021) brought this insight into finance, showing that the Sharpe ratio of a ridgeless portfolio can exhibit a double ascent, continuing to rise once the interpolation threshold is crossed. The same pattern has since been documented empirically by Hellum et al. (2026), across factor portfolios and individual stocks.

To evaluate this new paradigm empirically, this thesis uses monthly returns of 440 S&P 500 constituents over the period January 2000 to September 2025. This dataset places us naturally in the over-parameterized regime with $\rho = 1.42$. We first examine the spectral structure of S&P 500 returns to assess whether the conditions for benign overfitting are met. We then trace the *double descent* in out-of-sample variance using the Global Minimum Variance portfolio via the Moore-Penrose pseudoinverse when $N \geq T$, and explore how Ridge regularization shapes the Sharpe ratio trajectory of the mean-variance portfolio. Finally, the $R^2$ paradox is examined empirically to determine whether strong risk-adjusted returns can coexist with negative out-of-sample predictive accuracy.

### Research Questions

The central research question of this thesis is the following:

> *In a high-dimensional regime ($N > T$), to what extent is the spectral structure of S&P 500 returns compatible with the theoretical framework of benign overfitting, and what are the implications of this framework for out-of-sample portfolio performance?*

This question is decomposed into four sub-research questions (SRQ):

- **SRQ1:** What type of overfitting regime characterizes the spectral structure of S&P 500 returns, and is it compatible with the conditions for benign overfitting derived by Bartlett et al. (2020)?
- **SRQ2:** Once the interpolation threshold is crossed, to what level does out-of-sample variance decrease in S&P 500 portfolios, and is this consistent with the underlying overfitting regime identified in SRQ1?
- **SRQ3:** How does adjusting the Ridge ($z$) parameter affect the Sharpe ratio of S&P 500 portfolios? Does it allow us to observe a "permanent ascent" trajectory?
- **SRQ4:** Does the $R^2$ paradox apply to the S&P 500? Why does low predictive accuracy not hinder good economic performance in high dimensions?

The thesis is organized as follows. Chapter I establishes the limitations of classical portfolio optimization. Chapter II develops the theoretical framework: the double descent curve, the taxonomy of overfitting regimes, the conditions for benign overfitting due to Bartlett et al. (2020), the double ascent curve of Kelly et al. (2021), and the $R^2$ paradox. Chapter III translates this framework into concrete estimation tools, covering Random Matrix Theory as a diagnostic instrument, the Moore-Penrose pseudoinverse for the ridgeless regime, and the Ridge portfolio estimator with its hold-out calibration procedure. Chapter IV presents the empirical results on S&P 500 data, organized around the four sub-research questions.

# Chapter I

# LIMITS OF CLASSICAL PORTFOLIO OPTIMIZATION

This chapter examines the fundamental limitations of traditional portfolio optimization. By analyzing how estimation errors and high dimensions lead Markowitz models to fail, we identify the structural causes that motivate the shift to the new paradigm presented in the rest of this thesis.

## 1.1 The Mean-Variance Framework and the MSR Portfolio

In his seminal article, Markowitz (1952) demonstrates that assets should not be evaluated in isolation but rather in terms of their effect on the overall risk of a portfolio. Therefore, the choice of a portfolio is based on the principle that accepting more risk only makes sense if it offers the possibility of a higher return.

On Markowitz's efficient frontier, the Maximum Sharpe Ratio (MSR) portfolio plays a special role. It corresponds to the risky portfolio that offers the best compromise between return and risk. In other words, it is the one that maximizes the expected return per unit of volatility (Sharpe, 1966). Assuming that short selling [(1)] is permitted, this is obtained by solving the following optimization problem:

$$\max_{\mathbf{w}} \frac{\mathbf{w}'\boldsymbol{\mu}}{\sqrt{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}} \quad \text{subject to} \quad \mathbf{1}'\mathbf{w} = 1 \tag{I.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean returns and covariance matrix of returns, respectively. $\mathbf{w}$ represents the portfolio weights for each possible asset.

The analytical solution to this problem yields the following weight vector:

$$\mathbf{w}_{MSR} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \tag{I.2}$$

[(1)] Short selling enables investors to sell assets they do not own (by borrowing them) to speculate on falling prices. Mathematically, this means that portfolio weights $w_i$ can be negative. This allows the algorithm to use the sale of overvalued assets to finance the purchase of better-performing ones.

In theory, this model serves as a benchmark for mean-variance efficiency. In practice, however, it is extremely fragile because it is very sensitive to estimation errors, particularly those in the expected return vector ($\boldsymbol{\mu}$). As Merton (1980) shows, estimating expected returns is inherently imprecise because the variance of the mean estimator decreases slowly with sample size. Due to the small magnitude of expected returns relative to their volatility, a reliable estimate would require an extremely long data horizon.

This inaccuracy has major consequences. As demonstrated by Chopra and Ziemba (1993), errors in averages are approximately ten times more damaging to performance than errors in variances or covariances. Compounding this issue, the weight formula in Equation (I.2) requires multiplying $\boldsymbol{\mu}$ by $\boldsymbol{\Sigma}^{-1}$, which is itself poorly estimated whenever the sample size $T$ is not large relative to the number of assets $N$. The two sources of error therefore interact and reinforce each other, making the MSR portfolio doubly exposed to estimation noise. The conditions under which $\boldsymbol{\Sigma}^{-1}$ becomes unreliable are examined in detail in Section 1.3. This instability significantly degrades out-of-sample performance and makes the MSR portfolio difficult to use in practice.

## 1.2 The Global Minimum Variance (GMV) portfolio

A well-established approach to address the instability of expected return estimates is to focus exclusively on the risk structure of assets. The GMV portfolio is the point located at the left end of Markowitz's efficient frontier. Its sole objective is to minimize the overall variance of the portfolio, without seeking to achieve a target return. Mathematically, it solves the following problem:

$$\min_{\mathbf{w}} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \quad \text{subject to} \quad \mathbf{1}'\mathbf{w} = 1 \tag{I.3}$$

The analytical solution of this problem depends only on the covariance matrix ($\boldsymbol{\Sigma}$):

$$\mathbf{w}_{GMV} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} \tag{I.4}$$

The adoption of this model is widely justified in academic literature. As Jagannathan and Ma (2003, p. 1652) point out, "the estimation error in the sample mean is so large that nothing much is lost in ignoring the mean altogether." Furthermore, DeMiguel et al. (2009) empirically demonstrated that the GMV portfolio generally outperforms all other out-of-sample mean-variance portfolios, even when using performance measures that take into account both return and risk.

By eliminating $\boldsymbol{\mu}$ from the objective, GMV removes the dominant source of estimation error in mean-variance optimization, which is precisely why Bodnar et al. (2018) rely on it

to study high-dimensional portfolios. The simplification is real, but it is not a full solution: GMV remains entirely dependent on how well the covariance matrix is estimated.

## 1.3 Estimation risk and instability of the covariance matrix

Portfolio optimization relies on inverting the sample covariance matrix $\hat{\mathbf{\Sigma}}$. However, the reliability of this operation depends on a delicate balance between the number of assets ($N$) and the sample size ($T$). When $N$ increases relative to $T$, the estimation error embedded in $\hat{\mathbf{\Sigma}}^{-1}$ becomes non-negligible and strongly affects portfolio construction.

### *1.3.1 The concentration ratio ($\rho = N/T$)*

The concentration ratio $\rho = N/T$ formalizes this relationship and serves as the central diagnostic parameter throughout this thesis. Three regimes can be distinguished.

When $\rho \to 0$, observations vastly outnumber assets. The sample covariance matrix $\hat{\mathbf{\Sigma}}$ is well-conditioned and provides a reliable approximation of the true covariance structure.

As $\rho \to 1$, the number of parameters to be estimated grows toward the amount of available information. Estimation risk begins to dominate the benefits of diversification, and the matrix becomes increasingly ill-conditioned[(2)].

When $\rho > 1$, the number of assets exceeds the number of observations entirely. This is the high-dimensional regime that this thesis specifically investigates, and its consequences for matrix estimation are examined in detail in the following section.

### *1.3.2 Marčenko-Pastur's law*

To understand how estimation noise distorts the spectrum of the sample covariance matrix $\hat{\mathbf{\Sigma}}$, it is useful to start from a controlled setting. The Marčenko-Pastur law provides precisely this benchmark: by characterizing the distribution of eigenvalues that arises in a world of pure noise, it gives us a reference point against which to distinguish genuine signals from sampling noise. The result applies under the following conditions:

- The true population covariance matrix $\mathbf{\Sigma}$ is equal to the identity ($\mathbf{\Sigma} = \mathbf{I}$): all assets are uncorrelated and have unit variance. This matrix is fixed and deterministic, it is not subject to any noise.
- Asset returns are independent and identically distributed across observations, drawn from a distribution with finite fourth moment.[(3)]

[(2)] A matrix is said to be ill-conditioned when small perturbations in the input data produce disproportionately large changes in the output. Formally, this is measured by the condition number $\kappa(\hat{\mathbf{\Sigma}}) = \lambda_{\max}/\lambda_{\min}$: as the smallest eigenvalue $\lambda_{\min}$ approaches zero, $\kappa$ diverges and the inversion becomes numerically unstable.

[(3)] The finite fourth moment condition ensures that returns do not have excessively heavy tails, which is required for the eigenvalues to converge to the Marčenko-Pastur bounds.

- Both $N$ and $T$ grow large simultaneously, while their ratio $\rho = N/T$ remains fixed.

Under these conditions, even though the population matrix $\mathbf{\Sigma}$ is perfectly well-behaved, the eigenvalues of the sample covariance matrix $\hat{\mathbf{\Sigma}}$ are no longer concentrated at one. Instead, they spread across a continuous interval $[\lambda_-, \lambda_+]$ defined by:

$$\lambda_\pm = (1 \pm \sqrt{\rho})^2 \qquad \text{Marčenko and Pastur (1967)} \tag{I.5}$$

This dispersion is driven entirely by finite sample noise: while the population matrix $\mathbf{\Sigma}$ remains unchanged, the sample matrix $\hat{\mathbf{\Sigma}}$ picks up spurious variation simply because the data are finite. As $\rho$ increases, the interval widens, meaning that estimation noise becomes more severe as the number of assets grows relative to the available observations. In the critical case where $\rho$ approaches one, the lower bound $\lambda_-$ tends toward zero.

This has direct consequences for portfolio optimization. Since the GMV portfolio weights depend on $\hat{\mathbf{\Sigma}}^{-1}$, and since matrix inversion amounts to dividing by each eigenvalue, eigenvalues close to zero generate weights that are arbitrarily large and unstable. Estimation noise is therefore not merely preserved through the optimization process, it is actively amplified, leading to erratic allocations and poor out-of-sample performance. When $N > T$, the situation becomes even more severe: $\hat{\mathbf{\Sigma}}$ loses full rank and the classical inverse no longer exists.

## 1.4 Structural Inefficiency: The failure to separate signal from noise

Markowitz's classical framework is based on the idea that diversification reduces risk. However, in real markets, variance is dominated by a few major factors (signals), while most of the data reflects noise. A key limitation of the classical optimization framework is its inability to filter out this noise, which obscures exposure to real signals.

As Michaud (1989, p. 33) demonstrated, the mean-variance optimizer acts as an "estimation-error maximizer". When faced with imperfect estimates, the algorithm pursues mathematical extremes, confusing statistical noise with real opportunities for risk reduction. According to Michaud (1989, p. 34), this process systematically "overweights those securities that have large estimated returns, negative correlations, and small variances." These extreme values are generally nothing more than sampling errors that the optimizer amplifies rather than neutralizing.

The instability of traditional portfolios is due to the fact that not all components of the covariance matrix are estimated with the same precision. Zhao et al. (2018) demonstrate that estimation error is asymmetric; large eigenvalues (the signal) are captured relatively well, whereas small eigenvalues (the noise) exhibit significant relative error. Formally, let

$\lambda_i$ and $\hat{\lambda}_i$ represent the $i$-th largest eigenvalues of the true population covariance matrix $\mathbf{\Sigma}$ and of the sample covariance matrix $\hat{\mathbf{\Sigma}}$, respectively. Before stating the inequality, it is necessary to define the operator norm $\|\cdot\|_{op}$. For any matrix $\mathbf{A}$, it is defined as:

$$\|\mathbf{A}\|_{op} = \sup_{\|x\|=1} \|\mathbf{A}x\| \tag{I.6}$$

Intuitively, multiplying a vector $x$ by a matrix $\mathbf{A}$ transforms it, generally changing both its direction and its magnitude. The operator norm captures the worst case of this transformation: it is the largest factor by which $\mathbf{A}$ can stretch a unit vector. Concretely, if $\mathbf{A}$ is a symmetric positive semi-definite matrix such as a covariance matrix, $\|\mathbf{A}\|_{op}$ is simply equal to its largest eigenvalue. Applied to the estimation error matrix $\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}$, the quantity $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}$ therefore measures the maximum distortion that the estimation error can inflict on any portfolio direction. It acts as a single, global upper bound on how far the estimated covariance can deviate from the true one. With this notation, the relative estimation error on the $i$-th eigenvalue satisfies:

$$\frac{|\lambda_i - \hat{\lambda}_i|}{\lambda_i} \leq \frac{\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}}{\lambda_i} \tag{I.7}$$

This inequality reveals a fundamental asymmetry: since the upper bound $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{op}$ is common to all eigenvalues, dividing by a small $\lambda_i$ yields a much larger relative error for the small eigenvalues than for the large ones. In other words, the signal components of the covariance matrix are estimated with reasonable precision, while the noise components are disproportionately amplified.

To analyze this amplification mechanism, Zhao et al. (2018) introduce a split index $k$, which partitions the eigenvectors of $\hat{\mathbf{\Sigma}}$ into two groups. The first $k$ eigenvectors span the signal space $\mathcal{S}$, corresponding to the well-estimated market factors. The remaining $N - k$ eigenvectors span the noise space $\mathcal{N}$, corresponding to poorly estimated directions.

To formalize the amplification, two quantities must be clearly distinguished. For any portfolio with weight vector $\mathbf{w}$, the Expected Variance (EV) is defined as the in-sample variance computed using the estimated covariance matrix:

$$EV(\mathbf{w}) = \mathbf{w}'\hat{\mathbf{\Sigma}}\mathbf{w} \tag{I.8}$$

It represents what the optimizer believes the portfolio's variance to be. The Realized Variance (RV), by contrast, is the out-of-sample variance computed using the true population covariance matrix:

$$RV(\mathbf{w}) = \mathbf{w}'\mathbf{\Sigma}\mathbf{w} \tag{I.9}$$

It represents the true variance the investor will actually bear, and serves as the benchmark

against which in-sample estimates are evaluated.

Using Lemma 3.3 from Zhao et al. (2018), for any separation $(\mathcal{S}, \mathcal{N})$, the optimal portfolio $\mathbf{w}^*$ can be decomposed as a combination of the signal portfolio $\mathbf{w}^*_{\mathcal{S}}$ and the noise portfolio $\mathbf{w}^*_{\mathcal{N}}$:

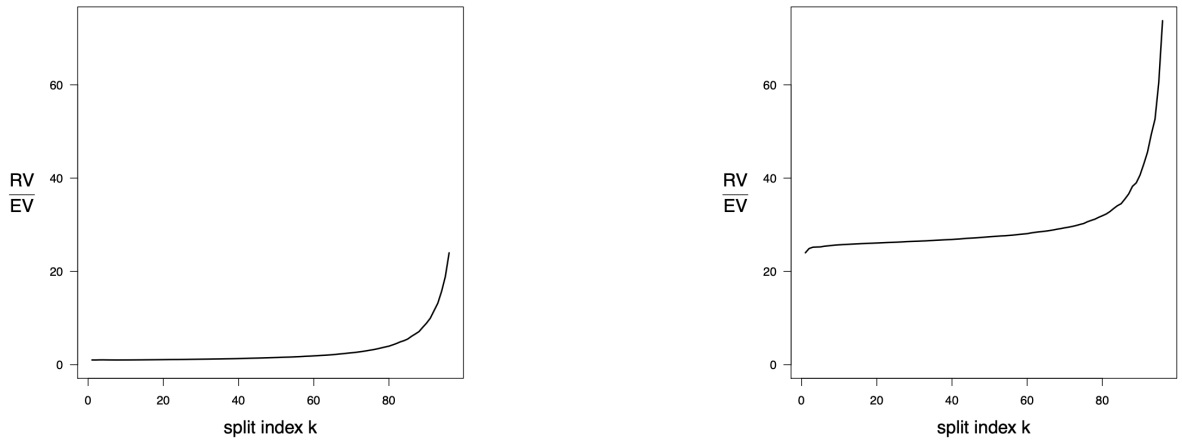
$$\mathbf{w}^* = \alpha \mathbf{w}^*_{\mathcal{S}} + (1 - \alpha)\mathbf{w}^*_{\mathcal{N}} \tag{I.10}$$

The weight $\alpha$ assigned to the signal theoretically depends on the Realized Variance ($RV$):

$$\alpha = \frac{1/RV(\mathbf{w}^*_{\mathcal{S}})}{1/RV(\mathbf{w}^*_{\mathcal{S}}) + 1/RV(\mathbf{w}^*_{\mathcal{N}})} \tag{I.11}$$

In a perfect world, the weight given to each component would depend on its actual risk contribution, as captured by $RV$. However, the optimizer only has access to $EV$. This discrepancy is precisely where the error amplification occurs:

- For the signal (the first $k$ components), the estimate is accurate ($EV \approx RV$), so the optimizer behaves correctly.
- For noise (components beyond $k$), the optimizer is subject to a systematic illusion: it perceives very low Expected Variance ($EV$) while the actual Realized Variance ($RV$) is much larger.




(a) The top-$k$ eigenvector portfolio $\mathbf{w}^*_{\mathcal{S}}$

(b) The remaining $(N-k)$ eigenvector portfolio $\mathbf{w}^*_{\mathcal{N}}$

Figure I.1: The Ratio between RV and EV. Reproduced from Zhao et al. (2018, p. 10, Fig. 2).

As illustrated in Figure I.1, the $RV/EV$ ratio for the noise portfolio $\mathbf{w}^*_{\mathcal{N}}$ is at least 20 times higher than that of the signal portfolio $\mathbf{w}^*_{\mathcal{S}}$ for virtually any choice of split index

$k$. Believing it is minimizing risk, the optimizer places excessive confidence in the noise components.

Throughout this chapter, we have seen that classical portfolio optimization faces a fundamental structural problem. The sample covariance matrix becomes increasingly unreliable as the number of assets $N$ grows relative to the number of observations $T$. Small eigenvalues, which are the most poorly estimated, are paradoxically the ones that drive the composition of the minimum variance portfolio through matrix inversion. The result is a framework that systematically mistakes noise for signal and produces portfolios that collapse out-of-sample.

This raises a natural question: what if high dimensionality, rather than being a problem to correct, could be turned into an advantage? Chapter II explores this possibility.

# Chapter II

# THE THEORY OF BENIGN OVERFITTING

This chapter introduces the concept of *benign overfitting* and the phenomenon of *double descent*, in which complexity is viewed as an asset rather than a risk. It then develops the *double ascent curve*, showing that increasing model complexity can improve risk-adjusted performance even when statistical accuracy, such as out-of-sample $R^2$, appears low.

## 2.1 From Classical Bias-Variance Trade-off to Double Descent

### *2.1.1 The Classical Paradigm: The U-Shape Risk Curve*

Before exploring the "double descent" revolution, it helps to recall the framework that has long dominated statistics and machine learning. The bias-variance trade-off, formalized by Geman et al. (1992) and later synthesized in the textbook of Hastie et al. (2009), is based on the idea that there is an "optimal complexity" for any model beyond which performance inevitably deteriorates.

Consider a data-generating process of the form $Y = f(X) + \epsilon$, where $f$ is the true unknown function and $\epsilon$ is a random noise term with $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$, independent of $X$. Given an estimator $\hat{f}$ trained on a finite sample, the expected prediction error at a point $x$ can be written as:

$$\text{Err}(x) = \mathbb{E}\left[(Y - \hat{f}(x))^2\right] \tag{II.1}$$

A standard algebraic decomposition yields:

$$\text{Err}(x) = \underbrace{(f(x) - \mathbb{E}[\hat{f}(x)])^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible error}} \tag{II.2}$$

where the three terms are defined as follows:

- The Bias$^2$, defined as $(f(x) - \mathbb{E}[\hat{f}(x)])^2$, measures the systematic error introduced by the gap between the true function $f$ and the average prediction of the estimator.

High bias indicates an "underfitting" model that is too simple to capture the actual structure of the data.

- The Variance, defined as $\mathbb{E}\left[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2\right]$, measures the sensitivity of the estimator to the particular sample on which it was trained. A high variance indicates that the model captures idiosyncratic fluctuations of the training data rather than the underlying signal, a situation known as overfitting.
- The Irreducible error $\sigma^2 = \text{Var}(\epsilon)$ is the variance of the noise term $\epsilon$ in the data-generating process. It represents the fundamental uncertainty in $Y$ that cannot be eliminated by any estimator, regardless of its complexity, because it is intrinsic to the phenomenon being modeled.

The traditional view posits that bias and variance evolve in opposite directions as model complexity (i.e., the number of parameters or assets used) increases. As parameters are added, bias automatically decreases because the model becomes more flexible. Conversely, variance increases because the model has "too much" freedom and begins to memorize sampling errors. The sum of these two forces creates a U-shaped total error curve, as illustrated in Figure II.1 from Fortmann-Roe (2012).

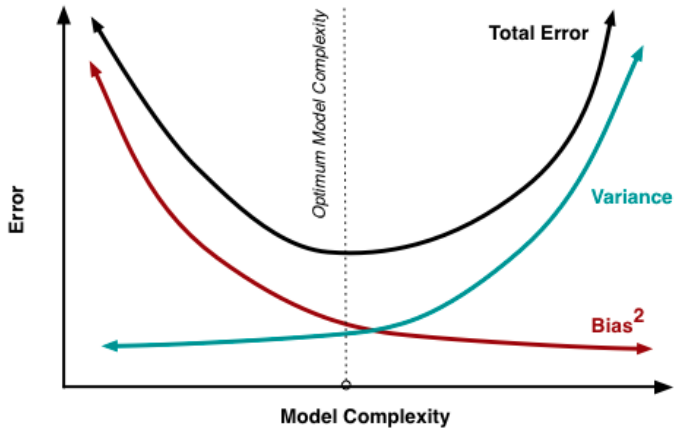



Figure II.1: The Bias-Variance Trade-off. Source: Reproduced from Fortmann-Roe (2012, Section 4.4, Fig. 6).

In this paradigm, the statistician's goal is to find the "sweet spot": the level of complexity that minimizes total error. According to this logic, exceeding this point constitutes a methodological error because the increase in variance ultimately outweighs the reduction in bias.

### *2.1.2 The Double Descent Curve*

Recent studies have revealed a more surprising reality: increasing model complexity beyond the number of observations can actually reduce prediction error. Belkin et al. (2019) dubbed this phenomenon "double descent". It describes learning regimes that are not captured by the classical bias-variance trade-off, and is particularly evident in highly overfitted models such as neural networks or financial models based on large asset universes.

The double descent curve (Figure II.2) revolves around a critical point called the interpolation threshold. This threshold depends on the capacity of the hypothesis space $\mathcal{H}$, which represents the richness or flexibility of the set of models that the algorithm can explore. In this context, capacity is directly related to the complexity of the model, often measured by its number of free parameters. This threshold is reached when the model's capacity is sufficient to perfectly fit the training data, thereby eliminating empirical risk. In portfolio optimization, this configuration occurs when the number of assets ($N$) is equal to the number of observations ($T$). At this stage, the model attempts to interpolate the noise in the data without sufficient degrees of freedom to smooth its estimates. This leads to a peak in out-of-sample risk and a poorly conditioned covariance matrix.

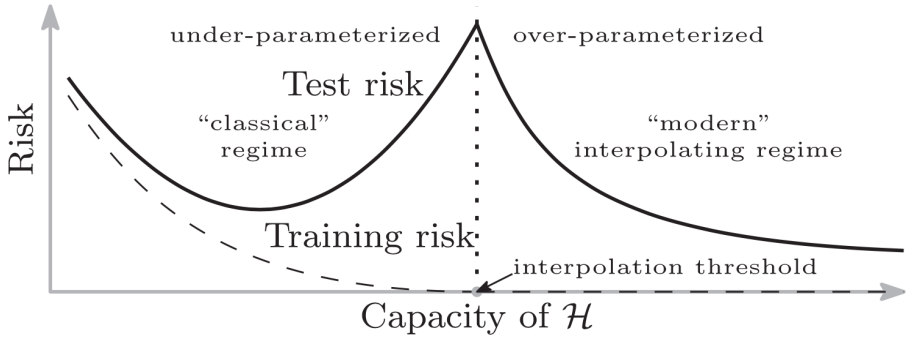



Figure II.2: The double descent curve showing the classical U-shape and the modern over-parameterized regime. Source: Reproduced from Belkin et al. (2019, p. 15850, Fig. 1).

However, once the threshold is crossed ($N > T$), the dynamics change substantially. As established in Section 1.3, $\hat{\boldsymbol{\Sigma}}$ loses full rank and the classical inverse no longer exists. We therefore turn to the Moore-Penrose pseudoinverse $\hat{\boldsymbol{\Sigma}}^{+}$, formally defined in Section 3.2, which selects the solution with the smallest Euclidean norm $\|w\|_2$ among the infinite number of solutions that perfectly fit the training data. Rather than concentrating portfolio weights on a few directions, it spreads them across all available assets, diluting estimation error across many securities. As Lu et al. (2024) point out, this is a form of implicit regularization: test error begins to decrease as complexity increases further beyond the threshold, precisely because no single noisy direction dominates the allocation.

What varies is not whether the error descends, but where it lands. With the pseudoinverse, crossing the interpolation threshold always brings the error back down from its peak. The question is how far, and whether it keeps falling. For overfitting to be truly *benign*, the error should recover below the under-parameterized floor and continue declining as $\rho$ grows. When conditions are not met, it either stabilizes at a comparable level or remains persistently high. Section 2.2 defines these regimes and the conditions that separate them.

## 2.2 Benign Overfitting

### *2.2.1 Types of Overfitting*

#### The Excess Risk

The starting point is the concept of excess risk. In a regression problem, the goal is to minimize the expected risk $R_T$[(1)], which measures the average prediction error on new data (Mallinar et al., 2022). However, there is an irreducible risk called the Bayes risk $R^*$, which is the minimum error that no model can surpass, even with perfect knowledge of the data distribution.

Excess risk is defined as the difference between the current risk of the model and the Bayes risk:

$$\bar{R} = R_T - R^* \tag{II.3}$$

Mallinar et al. (2022) classify overfitting into three regimes based on the asymptotic behavior of excess risk as $T \to \infty$, with $N$ held fixed (see Figure II.3):

- **A. Benign Overfitting:** Excess risk approaches zero as the sample size increases ($\lim_{T\to\infty} R_T = R^*$). The predictor adjusts globally to the actual function while fitting the noise very locally (in the form of narrow "peaks"), achieving statistical optimality.
- **B. Tempered Overfitting:** The most common regime for real models. Excess risk converges to a strictly positive constant ($\lim_{T\to\infty} R_T \in (R^*, \infty)$). The model is penalized by noise but the error remains bounded.
- **C. Catastrophic Overfitting:** The regime predicted by classical statistical theory for interpolating models. Excess risk becomes arbitrarily large ($\lim_{T\to\infty} R_T = \infty$), as the model interprets noise as genuine structure.

Table II.1 summarizes how excess risk mathematically defines each regime for regression problems:

(1) The expected risk $R_T$ corresponds to the Mean Squared Error (MSE), defined as $R_T = \mathbb{E}[(y - \hat{f}(x))^2]$, where $\hat{f}$ is the estimator constructed on $T$ observations.

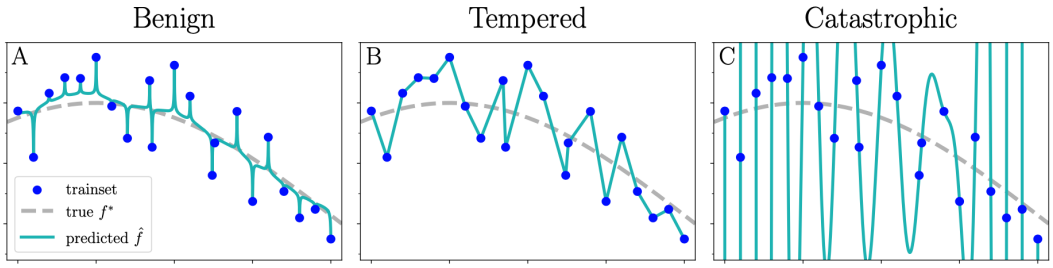



Figure II.3: Evolution of the predicted function $\hat{f}$ across different overfitting regimes. Reproduced from Mallinar et al. (2022, p. 2, Fig. 1).

| **Type of overfitting** | **Asymptotic behavior of risk ($R_T$)** |
|---|---|
| Benign | $\lim_{T\to\infty} R_T = R^*$ |
| Tempered | $\lim_{T\to\infty} R_T \in (R^*, \infty)$ |
| Catastrophic | $\lim_{T\to\infty} R_T = \infty$ |

Table II.1: Taxonomy of overfitting based on excess risk behavior

While the taxonomy of Mallinar et al. (2022) defines overfitting regimes by their results, it is necessary to analyze the structure of the covariance matrix $\mathbf{\Sigma}$ to understand their origin. Bartlett et al. (2020) demonstrate that this distinction depends on the distribution of variance in the parameter space. Their analysis is based on two notions of effective rank, denoted $r_k(\mathbf{\Sigma})$ and $R_k(\mathbf{\Sigma})$, which quantify the dimensionality of noise.

### *2.2.2 The Effective Rank Framework*

We denote the eigenvalues of the covariance matrix $\mathbf{\Sigma}$ as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. The parameter $k$ denotes the number of dominant principal components carrying the signal. The directions associated with the eigenvalues $\lambda_{k+1}, \ldots, \lambda_N$ are treated as noise.

In this context, Bartlett et al. (2020) define the first effective rank:

$$r_k(\mathbf{\Sigma}) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}} \tag{II.4}$$

which measures the total amount of variance contained in the secondary directions relative to the most important one. A high rank $r_k(\mathbf{\Sigma})$ indicates that the noise occupies a high-dimensional space and is not concentrated in a single unstable direction.
They also introduce the second effective rank:

$$R_k(\mathbf{\Sigma}) = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2} \tag{II.5}$$

which characterizes how residual variance is distributed among minor directions. This quantity is maximal when variance is distributed uniformly across secondary directions and minimal when it is concentrated on a few dominant axes. Well-distributed residual variance allows the model to absorb noise stably without disturbing the main signal.

### *2.2.3 Conditions for Benign Overfitting in Linear Regression*[2]

1. **Massive over-parameterization:** The set of assets must be large enough relative to the number of observations. Assets that are not very informative on their own provide additional degrees of freedom. This allows the model to absorb noise diffusely without degrading the main signal.

2. **Parsimony of the signal and the trace:** Building on the split index $k$ introduced in Section 2.2.2, Bartlett et al. (2020) define a data-driven threshold $k^*$ that formally identifies the boundary between signal and noise in the covariance structure:

$$k^* = \min\{k \geq 0 : r_k(\mathbf{\Sigma}) \geq bT\} \tag{II.6}$$

where $b > 1$ is a universal constant and $T$ denotes the sample size. Intuitively, $k^*$ represents the optimal number of principal components that concentrate the genuine predictive content of the data, the "useful signal." Beyond this threshold, the remaining directions form a noise space whose dimensionality is large relative to the sample size.

For benign overfitting to hold, two joint conditions must then be satisfied. First, the ratio between the total variance (the trace of the covariance) and the number of observations must tend toward zero as $T$ increases:

$$\lim_{T\to\infty} \frac{r_0(\mathbf{\Sigma})}{T} = 0 \tag{II.7}$$

Second, the number of signal directions $k^*$ must itself remain negligible relative to $T$:

$$\lim_{T\to\infty} \frac{k^*}{T} = 0 \tag{II.8}$$

Together, these two conditions express a form of parsimony: the true signal is confined to a low-dimensional subspace, while the remaining variance falls outside it. How that residual variance is distributed across noise directions is precisely what Conditions 3 and 4 examine.

[2] Bartlett et al. (2020) criteria apply to portfolio optimization because it shares the same mathematical structure as linear regression (under-determined linear systems and minimum norm solution).

3. **Noise dilution and isotropy:** The most critical condition is the model's ability to render noise harmless by dispersing it across many dimensions. Specifically, $R_{k^*}(\mathbf{\Sigma})$ must grow much faster than the number of observations ($T$):

$$\lim_{T\to\infty} \frac{T}{R_{k^*}(\mathbf{\Sigma})} = 0 \tag{II.9}$$

   This configuration allows the estimation error to be diluted across a multitude of secondary assets, a process favored by a covariance structure with an isotropic component[(3)].

4. **Structure of the "tail":** The eigenvalues associated with noise ($\lambda_i$ for $i > k^*$) must decrease, but slowly enough. If these eigenvalues fall too quickly (exponential decay), the noise concentrates on too few directions and becomes "catastrophic." Conversely, a "heavy tail"[(4)] ensures that the noise is evenly distributed.

## 2.3 The Double Ascent Curve

The double ascent curve, highlighted by Kelly et al. (2021) and Kelly and Xiu (2023), describes an improvement in expected performance as the complexity of the model increases. This approach marks a shift away from a purely statistical perspective: while double descent targets prediction error, the relevant criterion for investors is risk-adjusted performance, measured by the Sharpe ratio.

This performance gain rests on a realistic assumption: the market is so complex that no model can describe it perfectly. In this context, known as "mis-specification," investors seek an increasingly accurate approximation of economic reality rather than the exact model. Each new asset added to the portfolio is an additional source of information that enables the model to better understand the market's structure and improve the quality of the allocation.

### *2.3.1 The Ridge Estimator*

Kelly et al. (2021) exploit this complexity using the Ridge estimator (regularized least squares). In the context of return prediction, this estimator calculates the optimal regression coefficients ($\boldsymbol{\beta}$) by stabilizing the inverse of the moment matrix. These coefficients are then used to construct the portfolio. The simplified formula for this estimator is as follows:

$$\hat{\boldsymbol{\beta}}(z) = (z\mathbf{I} + \text{moment matrix})^{-1} \times \text{Covariance(signals, returns)} \tag{II.10}$$

[(3)] A component is said to be isotropic when it has the same variance in all directions, i.e. the noise is uniformly distributed and no direction is favored.

[(4)] the decay of eigenvalues $\lambda_i$ for $i > k^*$ is slow.

Where:

- **The moment matrix:** Captures the average cross-sectional interactions between all pairs of signals across time[(5)]. It plays a role analogous to the covariance matrix in standard regression: it summarises how much information each signal contributes and how redundant the signals are with one another. When $N$ approaches $T$, this matrix becomes ill-conditioned and its inversion unreliable, which is precisely where the Ridge parameter $z$ intervenes.
- **The parameter $z$ (shrinkage):** This value is added to the diagonal of the matrix to enable its inversion. Without this parameter ($z = 0$), we are in the "ridgeless" regime, and the classic calculation becomes impossible as $N$ approaches $T$. In this limiting case, we use the Moore-Penrose pseudoinverse (see Section 3.2).
- **The covariance vector (signals-returns):** This term represents the interaction between signals (asset characteristics) and future returns ($R_{t+1}$)[(6)]. Specifically, each element of this vector measures how strongly a given signal has historically co-moved with future returns.

Expanding the investment universe ($N$) sharpens this tension between signal richness and estimation noise, and the Sharpe ratio[(7)] is ultimately what judges the outcome. The Ridge estimator introduces $z$ as a stabilizer: by moderating the variance of portfolio weights, it prevents the model from overreacting to noise.

### *2.3.2 The Role of Shrinkage (z): From "Double Ascent" to "Permanent Ascent"*

Kelly et al. (2021) compare Sharpe ratio curves as a function of the concentration ratio $\rho = N/T$ to illustrate the impact of $z$ (see Figure II.4). This analysis reveals three distinct regimes:

- **A. The regime without regularization ($z \to 0$):** The ridgeless estimator corresponds to the Moore-Penrose pseudoinverse. When $\rho \geq 1$, the least-squares problem admits infinitely many solutions, and the pseudoinverse selects the one with minimum norm, implicitly spreading weights across assets and diluting estimation error. The resulting Sharpe ratio follows a double ascent trajectory: it rises in the

(5) Formally, following Kelly et al. (2021), the moment matrix is defined as $\frac{1}{T}\sum_t \mathbf{S}_t\mathbf{S}_t'$ and the covariance vector as $\frac{1}{T}\sum_t \mathbf{S}_t R_{t+1}$.

(6) In this context, the future return (or excess return) $R_{t+1}$ is modeled as a linear function of the signals observed at time $t$, augmented by an error term:

$$R_{t+1} = \mathbf{S}_t'\boldsymbol{\beta} + \epsilon_{t+1} \tag{II.11}$$

(7) Following Kelly et al. (2021), we use an unconditional definition: $SR = E[R^{\pi}_{t+1}]/\sqrt{E[(R^{\pi}_{t+1})^2]}$. This version uses the second moment in the denominator, allowing for a better mathematical characterization of dynamic strategies.

under-parameterized regime, collapses sharply at $\rho = 1$, then recovers and continues rising beyond the interpolation threshold.

- **B. The effect of shrinkage ($z$ fixed):** With $z > 0$, the "performance gap" at $N = T$ begins to close. Adding a constant to the diagonal prevents eigenvalues from approaching zero, resulting in a more stable transition.
- **C. Permanent ascent (optimal $z$):** One of the key conclusions of Kelly et al. (2021) is the existence of an optimal $z$. If regularization is precisely adjusted, the drop at $\rho = 1$ disappears completely, leading to a monotonically increasing curve known as "permanent ascent."

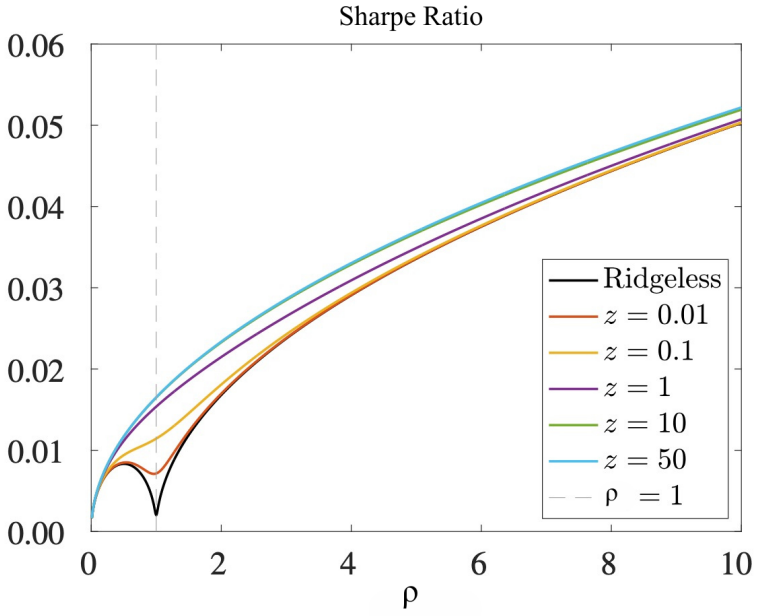



Figure II.4: The Double Ascent Curve for the Sharpe ratio. Adapted from Kelly et al. (2021, p. 39, Fig. 6).

## 2.4 The $R^2$ Paradox

Traditionally, we use the out-of-sample $R^2$ to assess model quality:

$$R^2_{OOS} = 1 - \frac{\sum(R_{t+1} - \hat{R}_{t+1})^2}{\sum(R_{t+1} - \bar{R}_{t+1})^2} \tag{II.12}$$

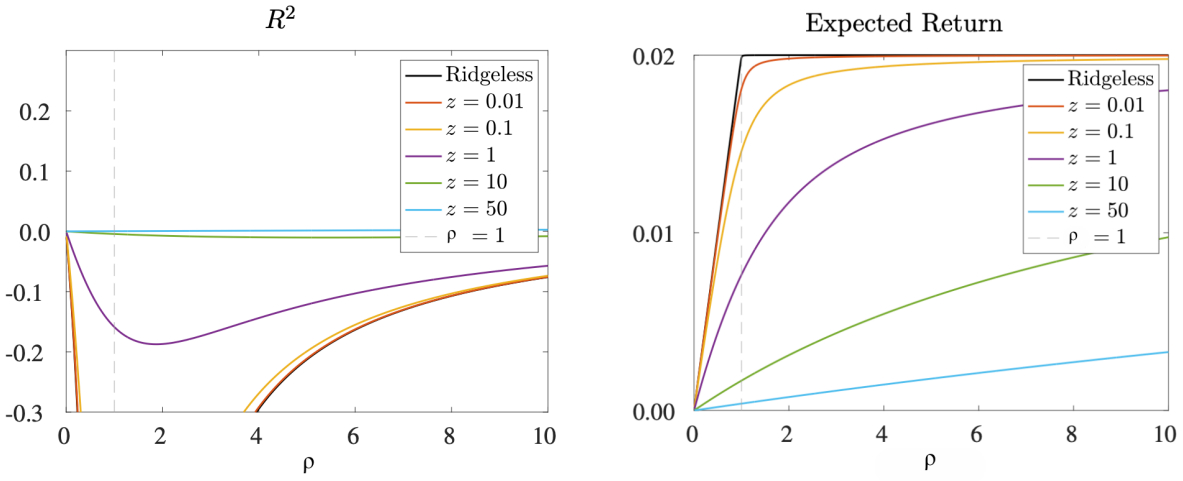
Where:

- $R_{t+1}$: denotes the realized return at time $t + 1$.
- $\hat{R}_{t+1}$: is the return predicted by the model using information available up to time $t$.
- $\bar{R}_{t+1}$: is the recursive mean of past realized portfolio returns, defined as $\bar{R}_{t+1} =$

$\frac{1}{t-1}\sum_{s=1}^{t-1} R_s^{\pi}$.[8]

- $\sum(R_{t+1} - \hat{R}_{t+1})^2$: represents the numerator, which measures the total prediction error of the model.
- $\sum(R_{t+1} - \bar{R}_{t+1})^2$: represents the denominator, which serves as a benchmark measuring the error one would obtain by simply predicting the average of all portfolio returns observed up to that point at each period $t$.

$R^2_{OOS}$ equals zero when the model performs no better than this naive benchmark, and turns negative when it performs worse. The results of Kelly et al. (2021) show that $R^2$ is nearly always zero or negative (see Figure II.5), regardless of the level of regularization $z$. At the threshold $N = T$, estimation error increases dramatically, pushing $R^2$ into negative values.




(a) Out-of-sample $R^2$ as a function of $\rho$

(b) Expected return as a function of $\rho$.

Figure II.5: The $R^2$ Paradox. Adapted from Kelly et al. (2021, p. 37, Fig. 4 & Fig. 5).

However, a major paradox emerges when we examine the expected return of the strategy: it increases with the number of assets, even as $R^2$ turns negative. This reveals that statistical accuracy and economic performance are distinct objectives. $R^2$ penalizes errors in the predicted *level* of returns: a forecast of +0.8% against a realization of +3% counts as failure, regardless of whether the portfolio was oriented in the right direction. Portfolio performance, by contrast, only requires that assets be ranked correctly on average.

[8] $R_s^{\pi}$ denotes the realized return of the Ridge portfolio at period $s$. This expanding-window average ensures the benchmark only uses information available at time $t$, avoiding look-ahead bias.

# Chapter III

# FROM NOISE DIAGNOSIS TO PORTFOLIO CONSTRUCTION

The preceding chapters have established the core tension of this thesis. Chapter I showed that the sample covariance matrix $\hat{\mathbf{\Sigma}}$ becomes ill-conditioned as the ratio $\rho = N/T$ grows: its small eigenvalues are overwhelmed by sampling noise, and their inversion in the Markowitz framework amplifies that noise rather than filtering it. Chapter II then proposed a conceptual shift, arguing that high dimensionality can become a structural advantage when estimation errors are sufficiently diffuse to be diluted across many directions. The present chapter translates this shift into concrete estimation tools.

Section 3.1 uses Random Matrix Theory as a diagnostic instrument: by characterising the spectral distribution that arises from pure noise, it identifies which eigenvalues of the sample covariance matrix $\hat{\mathbf{\Sigma}}$ carry genuine market information and which ones do not. This spectral reading of the data directly serves SRQ1, which asks whether the structure of S&P 500 returns is compatible with the conditions for benign overfitting. Section 3.2 introduces the Moore-Penrose pseudoinverse $\hat{\mathbf{\Sigma}}^{+}$ as the natural estimator for the ridgeless regime, the configuration in which the double descent phenomenon examined in SRQ2 is most clearly visible. Section 3.3 then defines the Ridge portfolio estimator and its cross-validated calibration, which is a practical tool for converting the instability observed at the interpolation threshold into the sustained performance improvement examined in SRQ3.

## 3.1 Random Matrix Theory (RMT)

As we established in Section 1.4, Markowitz's optimizer often acts as an "estimation error maximizer" due to its inability to distinguish between real correlations and random fluctuations. Laloux et al. (1999) show that most eigenvalues of the empirical covariance matrix are consistent with the hypothesis of a purely random matrix. Random Matrix Theory (RMT) provides a rigorous framework for identifying and cleaning this noise before optimization.

### *3.1.1 The "Bulk"*

The application of RMT rests on one key assumption: for a matrix whose components are purely random and independent, the distribution of eigenvalues falls within a bounded interval called the "bulk" (or body of the distribution).

Laloux et al. (1999)[(1)] derived the theoretical spectral density of the Marčenko-Pastur law, $\rho_C(\lambda)$, for a Wishart matrix[(2)] using the concentration ratio notation, $\rho = N/T$, in the limit where $N, T \to \infty$ with $\rho$ fixed:

$$\rho_C(\lambda) = \frac{1}{2\pi\rho\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{III.1}$$

Where the upper and lower bounds of the bulk are given by:

$$\lambda_\pm = \sigma^2(1 \pm \sqrt{\rho})^2 \tag{III.2}$$

In these expressions, $\sigma^2$ strictly denotes the variance of the purely random noise components. Since this true noise variance is unobservable in empirical financial data, $\sigma^2$ acts as a critical parameter that must be estimated.

According to RMT, when the eigenvalues of an observed matrix lie entirely within the bulk, they essentially reflect white noise. In this case, they cannot be interpreted as carrying reliable predictive information. Conversely, eigenvalues that clearly emerge from the bulk (spiking eigenvalues) are distinct from statistical noise and are associated with structural components of the market, which may contain a genuine signal.

### *3.1.2 The Eigenvalue Clipping*

The challenge of optimization lies in the small eigenvalues of the bulk. Laloux et al. (1999) show that eigenvalues closest to $\lambda_-$ are the most sensitive to noise. Yet they are precisely the ones that drive the composition of the GMV portfolio. To correct this instability, the method of Eigenvalue Clipping involves constructing a cleaned estimator $\mathbf{\Xi}_{clip}$.

Let $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ denote the spectral decomposition of the sample covariance matrix, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_N)$ contains the eigenvalues sorted in decreasing order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq$

(1) Although Laloux et al. (1999) initially developed these formulas for the classical regime where $\rho \leq 1$, the Marčenko-Pastur law remains mathematically valid in the high-dimensional regime ($N > T$). In this case, the matrix is singular and has $N - T$ zero eigenvalues. These zeros are excluded from the fit, and the continuous bulk of the remaining $T - 1$ positive eigenvalues is governed by the parameter $c = T/N$ rather than $\rho = N/T$. The bound $\lambda_+ = \sigma^2(1 + \sqrt{c})^2$ remains the universal indicator for separating signal from noise.

(2) A Wishart matrix is the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ of a sample drawn from a multivariate Gaussian distribution. It serves as a "null model" for isolating the spectral structure arising exclusively from sampling noise.

$\lambda_N \geq 0$) and $\mathbf{u}_i \in \mathbb{R}^N$ denotes the $i$-th column of $\mathbf{U}$, i.e. the unit-norm eigenvector of $\hat{\mathbf{\Sigma}}$ associated with $\lambda_i$, satisfying $\hat{\mathbf{\Sigma}}\,\mathbf{u}_i = \lambda_i \mathbf{u}_i$. The cleaned estimator is then defined as:

$$\mathbf{\Xi}_{clip} := \sum_{i=1}^{N} \xi_i^{clip}\, \mathbf{u}_i \mathbf{u}_i^T \tag{III.3}$$

where the cleaned eigenvalues $\xi_i^{clip}$ are adjusted according to the following rule:

$$\xi_i^{clip} = \begin{cases} \lambda_i & \text{if } \lambda_i > \lambda_+ \\ \bar{\lambda} & \text{otherwise} \end{cases} \tag{III.4}$$

In this expression, $\bar{\lambda}$ is the average of the eigenvalues located inside or below the bulk. It is chosen so that the trace of the matrix is preserved ($\mathrm{Tr}(\mathbf{\Xi}_{clip}) = \mathrm{Tr}(\hat{\mathbf{\Sigma}})$), ensuring that the total variance of the system remains constant:

$$\bar{\lambda} = \frac{1}{N-k} \sum_{i=k+1}^{N} \lambda_i \tag{III.5}$$

Crucially, the eigenvectors $\mathbf{u}_i$ remain unchanged throughout this procedure. Eigenvalue clipping acts solely on the magnitude of each component, not on the directions of variation identified in the data. The cleaned matrix $\mathbf{\Xi}_{clip}$ therefore retains the same principal axes as $\hat{\mathbf{\Sigma}}$, while dampening the influence of directions whose variance falls within the noise bulk.

### *3.1.3 Limits of Eigenvalue Clipping and the Effective Ratio ($\rho_{eff}$)*

Although influential, this historical approach has structural limitations. Bun et al. (2017) point out that the theoretical ratio $\rho = N/T$ is often insufficient for perfectly adjusting actual financial data. In practice, there is a discrepancy because this ratio is based on the assumption that asset returns are independent and identically distributed (i.i.d.) which is rarely verified in finance due to temporal autocorrelation.

To overcome this problem, an effective concentration ratio ($\rho_{eff}$) must be introduced. This parameter is calibrated so that the theoretical bound $\lambda_+ = \sigma^2(1 + \sqrt{\rho_{eff}})^2$ corresponds to the end of the bulk observed in the actual data. This calibration accounts for deviations from i.i.d. assumptions.

Finally, Bun et al. (2017) criticize the binary nature of clipping. The method assumes that signal eigenvalues ($\lambda_i > \lambda_+$) are perfectly estimated and need no correction. In reality, noise has a subtler effect beyond forming the bulk: it mechanically inflates the signal eigenvalues themselves. A factor is observed above its true value, pushed further outside

the bulk by sampling noise. Clipping detects it correctly as signal but retains its inflated value, leading to an overestimation of the economic intensity of the underlying factor.

## 3.2 Pseudoinverse Estimation When N > T

The diagnosis established in Section 3.1 reveals a key limitation of classical spectral correction. Eigenvalue clipping aims to restore the covariance matrix to a cleaner approximation of $\mathbf{\Sigma}$ by replacing noisy eigenvalues. But as soon as $N$ exceeds $T$, $\hat{\mathbf{\Sigma}}$ loses full rank: the $N - T$ zero eigenvalues are a structural feature of the data. Replacing them via clipping would restore invertibility, but artificially so, by assigning variance to directions that contain no information.

Since Markowitz optimization is mathematically equivalent to dividing by each eigenvalue $(1/\lambda)$, these zero eigenvalues make the calculation undefined. This leads to unstable weights and makes portfolio construction impossible. The interpolation threshold $(N = T)$ therefore calls for a different estimator.

### *3.2.1 Mathematical Definition and Mechanism*

In accordance with the estimation strategy of Lu et al. (2024), we use the Moore-Penrose pseudoinverse[(3)] of the sample covariance matrix, denoted $\hat{\mathbf{\Sigma}}^+$. Considering the eigenvalue decomposition of the sampled covariance matrix:

$$\hat{\mathbf{\Sigma}} = \sum_{i=1}^{N} \omega_i \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \tag{III.6}$$

where the eigenvalues $\omega_i$ are sorted in decreasing order. The rank of $\hat{\mathbf{\Sigma}}$ is at most $T - 1$, meaning only $K = \min(N, T - 1)$ eigenvalues are strictly positive. The Moore-Penrose pseudoinverse is then defined as:

$$\hat{\mathbf{\Sigma}}^+ = \sum_{i=1}^{K} \frac{1}{\omega_i} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T, \quad \text{with } K = \min(N, T - 1) \tag{III.7}$$

Unlike the classic inverse, which attempts to divide by all $N$ eigenvalues (including zero or nearly zero eigenvalues when $N \geq T$), the pseudo-inverse only divides by $K$ non-zero eigenvalues. For all other dimensions, where information is absent, it simply sets the value to zero.

[(3)] The pseudoinverse was first introduced by Moore (1920) and later formalized axiomatically by Penrose (1955).

### *3.2.2 Justification for the Estimator Selection*

1. **Validity in the overfitting regime:** The pseudoinverse is mathematically well-defined even when $\hat{\boldsymbol{\Sigma}}$ is singular, making portfolio construction possible beyond the interpolation threshold.

2. **Study of the limit regime and "benign overfitting":** A key motivation for using the pseudoinverse is its direct mathematical connection to the Ridge estimator. Specifically, $\hat{\boldsymbol{\Sigma}}^{+}$ corresponds to the limiting case of the Ridge-regularised inverse as the shrinkage parameter $z$ vanishes:

$$\hat{\boldsymbol{\Sigma}}^{+} = \lim_{z \to 0^{+}} (\hat{\boldsymbol{\Sigma}} + z\mathbf{I}_N)^{-1} \tag{III.8}$$

   This establishes the pseudoinverse as the natural reference point of the ridgeless regime. As Kelly et al. (2021) show, non-monotonic phenomena are most visible precisely in this limit, making it the ideal baseline from which to observe the performance collapse at the interpolation threshold ($N = T$). It then becomes straightforward to measure how reintroducing a positive regularisation parameter ($z > 0$) progressively compensates for this instability.

3. **Minimum Norm Property and Implicit Regularization:** As established in Section 2.1.2, the over-parameterized regime admits an infinite number of solutions that perfectly interpolate the training data. Among these, the pseudoinverse systematically selects the one that minimizes the Euclidean norm of the weights $\|\mathbf{w}\|_2$. By favoring the solution with the smallest norm, the pseudoinverse implicitly penalizes large and concentrated positions, producing an effect similar to that of an explicit regularization penalty, without requiring any additional parameter to be specified or tuned by the user.

In summary, the pseudoinverse serves as our reference point for analyzing portfolio behavior in the ridgeless regime, but its implicit regularization remains unstable near the interpolation threshold. Section 3.3 now introduces explicit shrinkage to address this limitation.

## 3.3 Ridge Portfolio Construction and Calibration

Section 3.2 showed that the pseudoinverse resolves the singularity problem when $N > T$. The resulting decline in out-of-sample variance already produces a "double ascent" in the Sharpe ratio, even in a risk-only GMV setting. Yet the GMV portfolio is, by construction, indifferent to expected returns. Moving to a full mean-variance objective removes that restriction: once $\hat{\boldsymbol{\mu}}$ re-enters the optimisation, the portfolio can exploit the return premium

available across the asset universe. The Sharpe ratio trajectory remains volatile without regularisation, however. Turning it into a monotonically increasing "permanent ascent" requires explicit shrinkage. We define the Mean-Variance Ridge estimator below and describe the procedure for selecting the shrinkage parameter $z$.

### *3.3.1 The Ridge Portfolio Estimator (Mean-Variance)*

In order to empirically validate the "permanent ascent" documented by Kelly et al. (2021), we must translate their predictive approach into a concrete allocation strategy. While Kelly et al. focus on estimating regression coefficients $\boldsymbol{\beta}$ to predict returns, our objective is to determine the optimal portfolio weights $\mathbf{w}$.

To bridge the gap between machine learning and portfolio optimization, we draw on the work of DeMiguel et al. (2009). Although earlier, their work establishes a fundamental framework by demonstrating that constraining the $L_2$ norm of portfolio weights ($||\mathbf{w}||_2^2 \leq \delta$) is mathematically equivalent to applying Ridge-type shrinkage to the covariance matrix within a minimum-variance setting. Inspired by this principle and extending it by analogy to the mean-variance optimization problem, we define our "Ridge" portfolio estimator as follows:

$$\mathbf{w}_{Ridge}(z) = \frac{(\hat{\boldsymbol{\Sigma}} + z\mathbf{I}_N)^{-1}\hat{\boldsymbol{\mu}}}{\mathbf{1}^\top(\hat{\boldsymbol{\Sigma}} + z\mathbf{I}_N)^{-1}\hat{\boldsymbol{\mu}}} \tag{III.9}$$

Where:

- $\hat{\boldsymbol{\Sigma}}$ is the covariance matrix of the sample and $\mathbf{I}_N$ is the identity matrix.
- $\hat{\boldsymbol{\mu}}$ is the vector of expected estimated returns.
- $z \in [0, \infty]$ is the regularization parameter that controls the intensity of the norm constraint.
- $(\hat{\boldsymbol{\Sigma}}+z\mathbf{I}_N)^{-1}$ stabilizes the inversion by "inflating" small eigenvalues (noise), preventing the weights from exploding.

The effectiveness of this estimator depends on the setting of parameter $z$, which balances noise reduction and signal preservation. To identify the optimal value $z^*$ that maximizes the Sharpe ratio, we favor an empirical approach that can adapt to real market conditions. The following section details the cross-validation procedure used to dynamically calibrate this parameter.

### *3.3.2 The Cross-Validation Approach*

The regularization parameter $z$ is a hyperparameter[(4)] that cannot be chosen by directly maximizing the in-sample Sharpe ratio: doing so would always yield $z = 0$, since the unregularized model always fits the training data best. As Arlot and Celisse (2010) point out, evaluating an algorithm's performance on the same data used to train it yields an overoptimistic result that fails to reflect true generalization capacity.

To circumvent this bias, we must evaluate candidate values of $z$ on data that were not used for training. Of the procedures identified by Arlot and Celisse (2010), we select the hold-out method. Unlike permutation approaches, such as $V$-fold, which would disrupt the temporal structure, the Hold-Out method allows for a strict chronological division. Training is done exclusively on past data, and validation is done on immediate future data, thus respecting the prohibition on using future data (no look-ahead bias).

#### The Hold-Out Principle

All available historical data $\mathcal{D}_n$ is divided into two disjoint subsets:

1. **Training Sample ($I_t$):** This covers the first $T_{\text{train}}$ observations. It is used to calculate fundamental statistics, such as the covariance matrix and expected returns, and to build portfolios for different values of $z$.
2. **Validation Sample ($I_v$):** This includes the most recent data immediately prior to the decision-making moment ($t$). It is used exclusively to evaluate the "out-of-sample" performance of the portfolios constructed in the previous step, thereby simulating what would have occurred in the recent past.

#### Selection Criterion

The objective is to identify the parameter $z$ that maximizes risk-adjusted return. Specifically, at each rebalancing date $t$ (i.e., each time the portfolio needs to be updated), we test a grid of candidates $z \in \mathcal{Z}$.

The optimal parameter $z^*$ is the one that generated the highest Sharpe Ratio on the validation sample $I_v$. It is this parameter $z^*$, empirically validated over the period immediately preceding $t$, that is then selected to construct the final portfolio held until the next period $(t + 1)$.

[(4)] A model parameter set by the user prior to estimation, rather than learned directly from the data.

# Chapter IV

# EMPIRICAL TEST (S&P 500)

## 4.1 Data Description

The empirical analysis is conducted on a dataset provided by WRDS (Wharton Research Data Services, 2025). It covers the monthly returns of 440 individual stocks drawn from the S&P 500 index over the period January 2000 to September 2025, yielding $T = 309$ monthly observations per asset. Among all $1,070$ stocks that held membership in the index at some point during the sample window, only those with an uninterrupted listing throughout the entire period were retained [(1)]. This filtering produces a balanced panel with a concentration ratio of $\rho = N/T = 1.42$, placing the dataset naturally in the over-parameterized regime.

## 4.2 Spectral Diagnosis (SRQ1)

*SRQ1: What type of overfitting regime characterizes the spectral structure of S&P 500 returns, and is it compatible with the conditions for benign overfitting derived by Bartlett et al. (2020)?*

The first step of the empirical analysis is to examine the eigenvalue structure of the sample covariance matrix. It pursues two objectives: assessing the fit of the empirical eigenvalue distribution to the theoretical Marčenko-Pastur density $\rho_C(\lambda)$, and testing whether the conditions of Bartlett et al. (2020) for benign overfitting are satisfied in practice.

### *4.2.1 Spectral structure and the Marčenko-Pastur fit*

The sample covariance matrix $\hat{\mathbf{\Sigma}}$ is computed from $N = 440$ assets and $T = 309$ monthly observations. Since $N > T$, the matrix is rank-deficient. It admits at most $T - 1 = 308$ strictly positive eigenvalues, with the remaining 132 equal to zero by construction. These structural zeros are excluded from the spectral analysis as they carry no information about the underlying return distribution.

[(1)] This selection criterion introduces a survivorship bias. However, this is not of primary concern since our focus is on the spectral and statistical properties of the return matrix rather than live investment conclusions.

Following the discussion in Section 3.1, the Marčenko-Pastur bounds are computed on the 308 non-zero eigenvalues using the ratio $c = T/N = 0.702$ rather than $\rho = N/T$ (see Section 3.1.1, footnote (1)). To obtain a reliable estimate of the noise variance $\sigma^2$, we adopt a two-pass calibration procedure. In the first pass, the top ten eigenvalues are excluded as obvious market factors and the average of the remaining eigenvalues provides an initial estimate of $\sigma^2$. This estimate is used to compute a preliminary $\lambda^+$ which identifies a first set of signal components. In the second pass, $\sigma^2$ is re-estimated using only the eigenvalues classified as noise in the first pass, yielding a refined and self-consistent estimate of the bulk variance.[(2)] The resulting bounds are:

$$\lambda^- = 0.000201, \quad \lambda^+ = 0.0259$$

Figure IV.1 plots the empirical density of the 259 eigenvalues within the bulk $[\lambda^-, \lambda^+]$ against the theoretical Marčenko-Pastur density. The red curve tracks the bulk shape reasonably well, with some local discrepancies near the lower bound. These 259 eigenvalues represent 88.9% of the non-zero spectrum, confirming that estimation noise dominates the covariance structure. Forty-nine eigenvalues exceed $\lambda^+ = 0.0259$, representing 11.1% of all eigenvalues. The largest, at 4.51, almost certainly corresponds to the broad market factor. The remaining 48 likely reflect sector- and industry-level correlations.

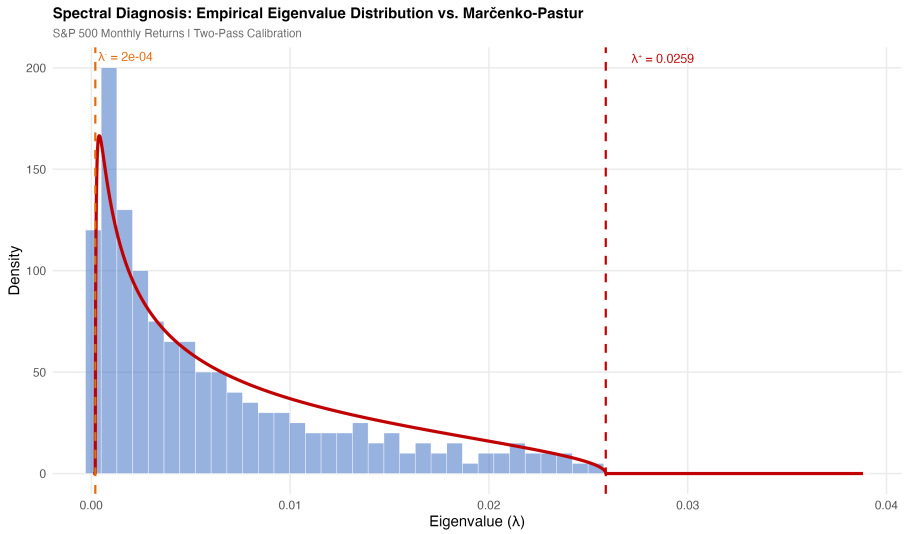



Figure IV.1: Empirical eigenvalue distribution of the S&P 500 sample covariance matrix (blue histogram) compared to the theoretical Marčenko-Pastur density (red curve). Only the 259 eigenvalues within $[\lambda^-, \lambda^+]$ are shown.

A one-sample Kolmogorov-Smirnov test assesses this visual fit. It computes $D$, the

[(2)] We purposely avoid using the simple average $\mathrm{Tr}(\hat{\mathbf{\Sigma}})/N$ as our starting point for the noise variance. Since this average is heavily skewed by the few dominant eigenvalues carrying the market signal, it would mechanically inflate our $\sigma^2$ estimate and end up pushing the theoretical Marčenko-Pastur boundary ($\lambda^+$) much higher than it should be.

maximum vertical distance between the empirical and theoretical cumulative distributions: the larger $D$, the more the two diverge. The result is $D = 0.134$ ($p = 0.0002$), formally rejecting the null hypothesis that the bulk eigenvalues follow the MP law exactly. This rejection should not be over-interpreted: with 259 eigenvalues the test detects even minor deviations, and some deviation is expected given that the MP law assumes i.i.d. returns.

### *4.2.2 Bartlett Conditions for Benign Overfitting*

The four conditions of Bartlett et al. (2020) are evaluated on the noise tail $\lambda_{50}, \ldots, \lambda_{308}$, with $k^* = 49$ signal components identified via the Marčenko-Pastur upper bound $\lambda_+$. This differs from their original criterion, which relies on a universal constant $b > 1$ to determine $k^*$ (Equation II.6). On financial data, this constant is difficult to pin down: choosing $b = 1.1$ or $b = 1.5$ can shift $k^*$ substantially. Using $\lambda_+$ as the separation threshold avoids this ambiguity entirely, as it requires no external constant. Table IV.1 summarizes the results.

Table IV.1: Empirical test of Bartlett et al. (2020) conditions for benign overfitting on the S&P 500 dataset.

| **Cond.** | **Description** | **Formula** | **Value** | **Threshold** | **Satisfied?** |
|---|---|---|---|---|---|
| 1 | Massive over-parameterization | $\rho = N/T$ | 1.424 | $> 1$ | Yes |
| 2a | Signal parsimony | $k^*/T$ | 0.159 | $\to 0$ | Borderline |
| 2b | Trace parsimony | $r_0(\mathbf{\Sigma})/T$ | 0.0088 | $\to 0$ | Yes |
| 3 | Noise isotropy | $R_{k^*}(\mathbf{\Sigma})$ | 124.1 | $\gg T = 309$ | No |
| 4 | Tail decay (exp. fit $R^2$) | — | 0.977 | $\ll 1$ | No |

Condition 1 is satisfied by construction. With $\rho = N/T = 1.424$, the dataset lies in the over-parameterized regime. This provides the degrees of freedom needed for noise diffusion.

Condition 2a is borderline as the ratio $k^*/T = 0.159$ is not strictly negligible. While 49 signal components might not reflect a low-dimensional structure in a strict asymptotic sense, 88.9% of the eigenvalues still reside in the noise subspace. Condition 2b is satisfied. Total variance is not dominated by any single direction ($r_0(\mathbf{\Sigma})/T = 0.0088 \approx 0$).

For Condition 3, the second effective rank $R_{k^*} = 124.1$ falls well below $T = 309$. Residual noise variance is not sufficiently diffuse across the 259 noise directions. A few noise eigenvalues absorb a disproportionate share of the variance. This fails the isotropy condition required for benign overfitting.

For Condition 4, Figure IV.2 plots the noise eigenvalues in log scale against their rank in the noise subspace. A straight downward line in this representation signals exponential decay, the unfavorable configuration described by Bartlett et al. (2020). A linear regression on the log-transformed eigenvalues yields $R^2 = 0.977$. This high value confirms that the

decay is nearly perfectly exponential. The noise tail of S&P 500 returns fails the heavy-tail condition.

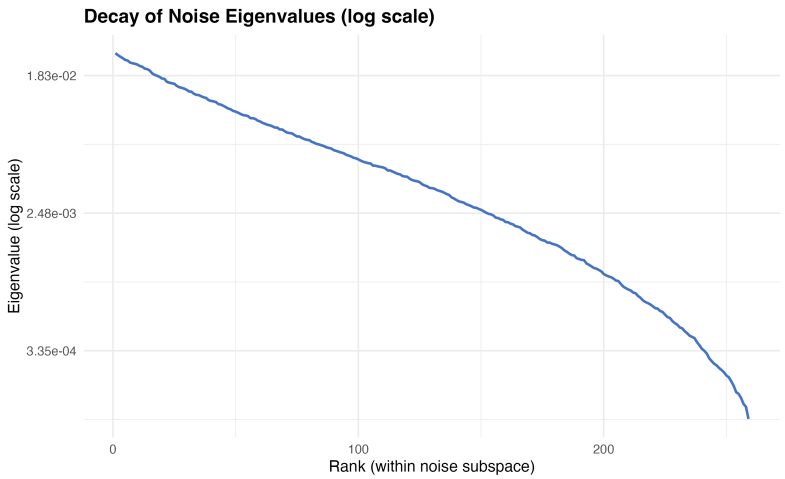



Figure IV.2: Log-scale decay of noise eigenvalues.

**Conclusion on SRQ1**

The spectral analysis leads to a nuanced answer to SRQ1. Conditions 1 and 2b are satisfied. The dataset is over-parameterized by construction ($\rho = 1.42$) and total variance is not dominated by any single direction. Condition 2a is borderline: the ratio $k^*/T = 0.159$ is notably far from zero. Conditions 3 and 4 are not met. The noise tail shows near-exponential decay and variance concentrates on too few directions rather than diffusing across the noise subspace.

Following the taxonomy of Mallinar et al. (2022) introduced in Section 2.2.1, the S&P 500 falls in the *tempered overfitting* regime. Estimation errors remain bounded and do not diverge. Noise is not fully neutralized in the way benign overfitting would require.

## 4.3 Double Descent in Risk (SRQ2)

*SRQ2: Once the interpolation threshold is crossed, to what level does out-of-sample variance decrease in S&P 500 portfolios, and is this consistent with the underlying overfitting regime identified in SRQ1?*

### *4.3.1 Experimental design*

To isolate the effect of the concentration ratio $\rho = N/T$, the estimation window is fixed at $T_{\text{train}} = 120$ months. This ten-year horizon is a standard in the empirical portfolio literature (DeMiguel et al., 2009). $N$ is then varied across a grid of 36 values covering $\rho \in [0.08, 2.92]$, constructed from three sequences: a step size of 10 in the under-parameterized zone, a finer step of 3 in the critical interval $\rho \in [0.875, 1.125]$ to capture the sharp dynamics near

the interpolation threshold, and a step size of 15 in the over-parameterized zone where the curve is expected to flatten. Although 440 stocks are available, the grid is capped at $N = 350$. Beyond $\rho \approx 2.9$, the curve has already stabilized and additional points would add no information about the shape of the descent.

To address the composition noise that arises when each value of $N$ draws a completely independent set of assets, a nested-set averaging procedure is adopted. One hundred independent random orderings of the full asset universe are generated; for each ordering, the portfolio for a given $N$ uses the first $N$ assets in that order. For each ordering and each $N$, out-of-sample variance is computed over $n_{roll} = 188$ rolling windows. The reported variance is then the average of these 100 estimates. At each step $t$, GMV weights are computed using the standard inverse when $N < T_{train}$ and the Moore-Penrose pseudoinverse $\hat{\mathbf{\Sigma}}^{+}$ when $N \geq T_{train}$ (Section 3.2). A LOESS[(3)] smoother (span = 0.2) is then applied to reduce residual noise without attenuating the spike at $\rho = 1$.

### *4.3.2 A three-regime risk structure*

Figure IV.3 is organized around three regimes.

**Under-parameterized regime ($\rho < 1$):** Out-of-sample variance remains low and stable averaging 0.0203 with a mild upward drift as $\rho$ approaches one.

**Interpolation threshold ($\rho = 1$):** Out-of-sample variance spikes to 0.343 (the red dot), corresponding to a monthly portfolio volatility of $\sqrt{0.343} \approx 58.6\%$.

**Over-parameterized regime ($\rho > 1$):** Once the threshold is crossed, variance drops sharply, falling back to an average of 0.0135. The descent is steepest in $\rho \in (1, 1.1)$ and flattens thereafter.

**On the shape of the curve**

Unlike the full double descent often illustrated in theoretical frameworks (see Section 2.1.2, Figure II.2), our empirical curve for the S&P 500 takes the shape of a single spike. This divergence in shape highlights the impact of the asset universe choice. For instance, Hellum et al. (2026) observe a proper U-shape because they use anomaly factor portfolios. These factors have low correlations, so adding more assets sharply reduces portfolio risk. In contrast, we use individual S&P 500 stocks. Since these large-cap equities are heavily tied to the same dominant market factor, a small portfolio of $N = 10$ assets ($\rho = 0.08$)

[(3)] LOESS (Locally Estimated Scatterplot Smoothing) is a non-parametric smoothing method that fits a weighted local regression at each point using only the neighboring observations. The `span` parameter controls the fraction of data used in each local fit: a smaller span preserves sharp local features such as the spike at $\rho = 1$, while a larger span produces a smoother but less responsive curve.

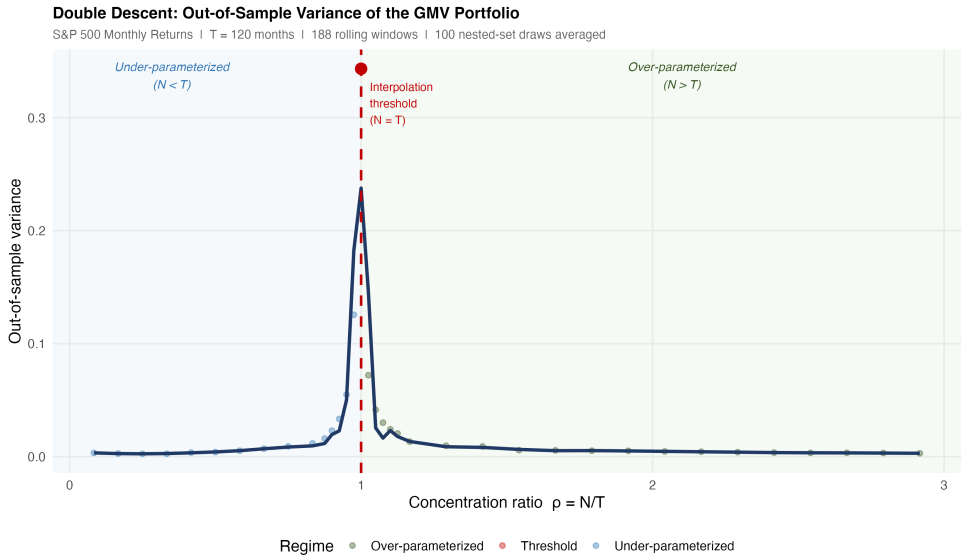



Figure IV.3: Out-of-sample variance of the GMV portfolio as a function of $\rho = N/T$. Estimation window $T_{\text{train}} = 120$ months, 188 rolling windows, averaged over 100 nested-set orderings. The solid line is a LOESS fit (span = 0.2).

already captures most of the available diversification. The marginal gain from adding more stocks is therefore negligible in this zone, so variance stays flat rather than descending first.

### Conclusion on SRQ2

Once the interpolation threshold is crossed, out-of-sample variance falls to an average of 0.0135, modestly below the under-parameterized level of 0.0203. The pseudoinverse makes this regime viable through implicit regularization: it spreads portfolio weights across all assets, diluting estimation error.

However, this result does not signal a benign regime. In a truly benign setting, variance would continue declining as $\rho$ grows rather than stabilizing. Here, the curve drops steeply for $\rho \in (1, 1.1)$ before stabilizing around $\rho \approx 1.5$.

This is consistent with the tempered classification from SRQ1. The noise tail is not isotropic enough to drive sustained improvement beyond the interpolation threshold, but diffuse enough to bring variance back to a comparable level. Both the level and the shape of the descent confirm that overfitting in S&P 500 returns is bounded rather than benign.

## 4.4 Performance Analysis (SRQ3)

*SRQ3: How does adjusting the Ridge parameter z affect the Sharpe ratio of S&P 500 portfolios? Does it allow us to observe a "permanent ascent" trajectory?*

### *4.4.1 Experimental design*

The analysis extends the SRQ2 framework by replacing the GMV objective with the mean-variance Ridge estimator defined in Section 3.3:

$$\mathbf{w}_{\text{Ridge}}(z) = \frac{(\hat{\boldsymbol{\Sigma}} + z\mathbf{I}_N)^{-1}\hat{\boldsymbol{\mu}}}{\mathbf{1}^\top(\hat{\boldsymbol{\Sigma}} + z\mathbf{I}_N)^{-1}\hat{\boldsymbol{\mu}}} \tag{IV.1}$$

where $\hat{\boldsymbol{\mu}}$ is the sample mean return vector and $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix, both estimated on the training window. The same grid of 36 values of $N$ is used ($\rho \in [0.08, 2.92]$), with $T_{\text{train}} = 120$ months. Performance is evaluated via annualized Sharpe ratio over $n_{\text{roll}} = 152$ rolling windows.[4] To isolate the effect of $\rho$ from composition noise, a single nested-set ordering is applied (instead of 100 in Section 4.3). This compromise is necessary because the cross-validation loop makes a large number of orderings computationally prohibitive.

Two experiments are conducted over the same regularization grid $z \in \{0, 0.01, 0.1, 1, 10, 50, 100, 500\}$. In the first, $z$ is fixed across all rolling windows at each grid value in turn. The case $z = 0$ corresponds to the ridgeless pseudoinverse. In the second, $z$ is selected at each rolling step via hold-out cross-validation. The estimation window is split into a training period ($T_{\text{train}} = 120$ months) and a validation period ($T_{\text{val}} = 36$ months). The value $z^*$ that maximizes the Sharpe ratio on the validation set is then retained to construct the out-of-sample portfolio.

### *4.4.2 Effect of fixed regularization*

Figure IV.4[5] illustrates how regularization reshapes the Sharpe ratio curve. Three patterns stand out.

The ridgeless estimator ($z = 0$) performs poorly throughout. Its Sharpe ratio oscillates erratically across all values of $\rho$. Relative to the GMV case in SRQ2, the damage is more severe. The mean-variance optimizer depends on $\hat{\boldsymbol{\mu}}$ in addition to $\hat{\boldsymbol{\Sigma}}$. This introduces a second source of estimation noise that compounds the instability at the interpolation threshold.

Moderate regularization ($z = 0.1$) brings a meaningful improvement. The curve is considerably smoother and no longer collapses at $\rho = 1$. The overall Sharpe ratio nonetheless remains low across all regimes. This level of shrinkage is insufficient to stabilize the inversion of a near-singular covariance matrix.

[4] The reduction from 188 to 152 windows relative to SRQ2 reflects the allocation of $T_{\text{val}} = 36$ months to the hold-out validation set used for $z^*$ selection, as described in Section 3.3.2.

[5] For clarity, the figure displays only six representative values omitting $z = 0.01$ and $z = 50$ which lie between already-shown curves.

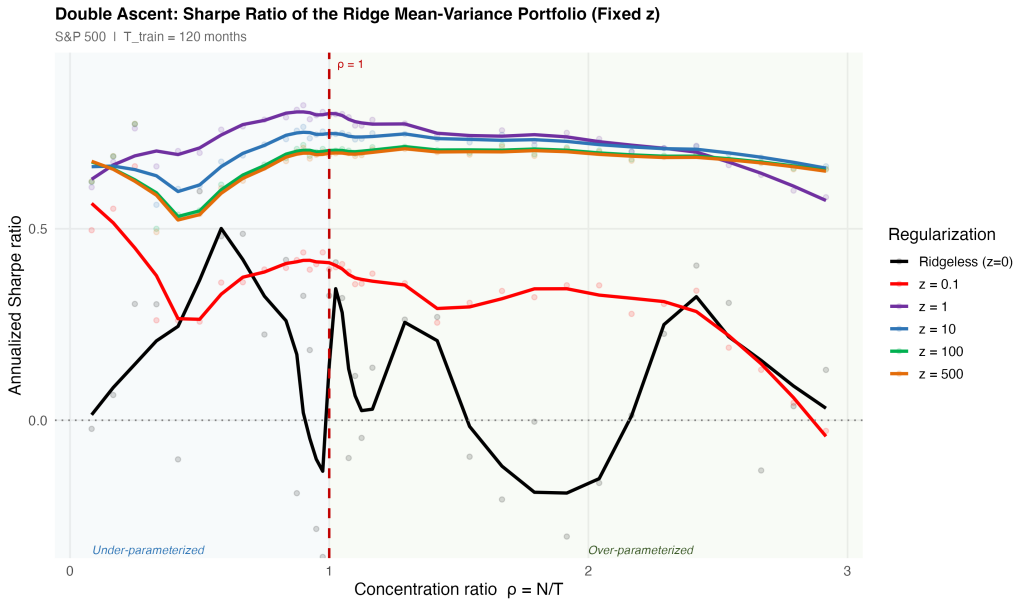



Figure IV.4: Annualized Sharpe ratio of the Ridge Mean-Variance portfolio as a function of the concentration ratio $\rho = N/T$ for several fixed values of $z$. The solid lines are LOESS fits (span = 0.25). The ridgeless case ($z = 0$) uses the Moore-Penrose pseudoinverse for $N \geq T_{\text{train}}$.

Strong regularization ($z \geq 1$) tells a different story. The curves for $z = 1$, $z = 10$, $z = 100$ and $z = 500$ are nearly superimposed in the over-parameterized regime. All maintain an annualized Sharpe ratio comfortably above 0.5. They also show a much flatter trend across $\rho$, erasing the drop at $\rho = 1$.[(6)]

### *4.4.3 The Cross-Validated optimal $z^*$*

Table IV.2 reports how often each $z^*$ was selected across all $N$ values and all 152 rolling windows, for a total of 5,472 decisions. The most frequent choice is $z^* = 500$, selected in 2,848 cases (52%). This dominance does not mean strong shrinkage is clearly optimal. The validation Sharpe is nearly flat for $z \geq 10$. Differences across that range are too small to reflect a genuine signal on a 36-month window. $z = 500$ simply accumulates wins by noise.

Values $z^* \in \{10, 50, 100\}$ collect only 363 selections combined. They are systematically beaten within that plateau. The remaining 2,261 selections fall on $z^* \leq 1$, reflecting windows where low shrinkage genuinely dominates.

The distribution is polarized. Hold-out cross-validation picks a side: strong regularization or weak. Rarely anything in between.

[(6)] The same analysis was replicated using a Ridge GMV portfolio. Results are reported in Appendix A.

Table IV.2: Frequency of the optimal regularization parameter $z^*$ selected by the 36-month hold-out cross-validation across all $N$ and all 152 rolling windows.

| **Shrinkage parameter ($z^*$)** | **0** | **0.01** | **0.1** | **1** | **10** | **50** | **100** | **500** | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| **Selection frequency** | 712 | 500 | 439 | 610 | 266 | 82 | 15 | 2 848 | **5 472** |

### *4.4.4 Towards Permanent Ascent*

Figure IV.5 presents the Sharpe ratio obtained when $z^*$ is selected dynamically at each rolling step. Performance averages are quite similar on both sides of the interpolation threshold. The Sharpe ratio is 0.331 in the under-parameterized regime and 0.325 in the over-parameterized regime. However, the trough at $\rho = 1$ is not eliminated and the "permanent ascent" of Kelly et al. (2021) is therefore not fully achieved.

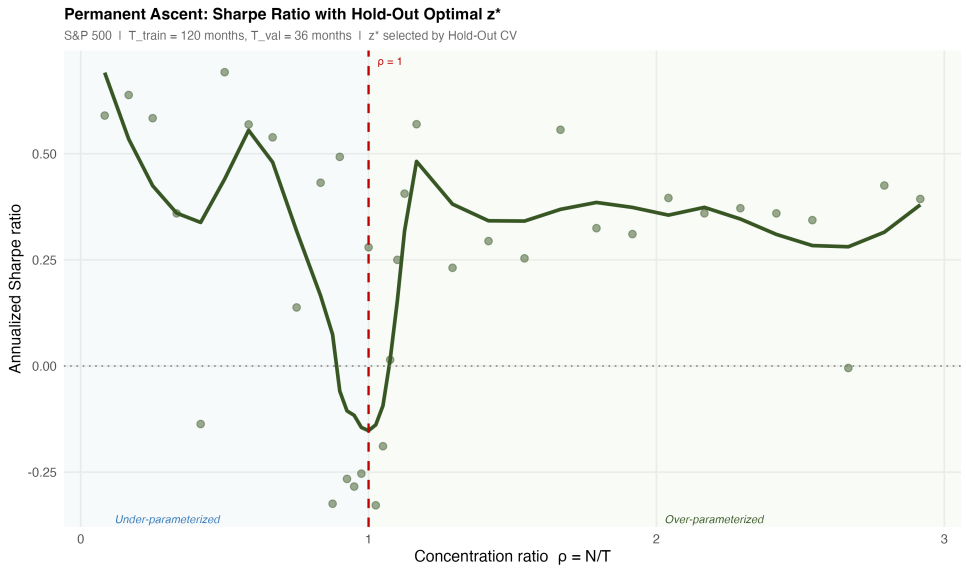



Figure IV.5: Annualized Sharpe ratio with hold-out cross-validated $z^*$ as a function of $\rho = N/T$. The solid line is a LOESS fit (span = 0.30).

This outcome is rooted in a statistical property of the selection procedure itself. Table IV.3 shows that low-shrinkage values produce both the weakest average validation Sharpe ratios and the highest volatility, at 0.632 for $z = 0$ compared to 0.350 for $z = 500$. Yet Table IV.2 shows they are selected in 2,261 out of 5,472 decisions. This is not a contradiction. Over the validation window, a high-variance candidate occasionally lands above the stable plateau of $z \geq 10$, not because it is genuinely better but because its wider distribution gives it more chances to win any given draw. The hold-out procedure therefore ends up selecting values that are poor on average but lucky on that particular window.

Table IV.3: Overall validation Sharpe ratio distribution statistics across all rolling windows

| **Shrinkage ($z$)** | **Mean** | **Volatility** | **Absolute Max** |
|---|---|---|---|
| 0 | 0.052 | 0.632 | 2.04 |
| 0.01 | 0.057 | 0.582 | 1.86 |
| 0.1 | 0.323 | 0.514 | 1.98 |
| 1 | 0.658 | 0.354 | 1.95 |
| 10 | 0.728 | 0.341 | 2.03 |
| 50 | 0.730 | 0.348 | 2.03 |
| 100 | 0.730 | 0.349 | 2.03 |
| 500 | 0.730 | 0.350 | 2.04 |

**Conclusion on SRQ3**

Adjusting $z$ has a clear effect on the Sharpe ratio curve. Beyond a threshold level of regularization ($z \geq 1$), the collapse at $\rho = 1$ is largely suppressed and performance remains stable across all values of $\rho$, effectively erasing the interpolation threshold as a source of instability.

Cross-validated $z^*$ produces the opposite result. Instead of smoothing the threshold, it reinstates a sharp trough at $\rho = 1$, where Sharpe ratios turn negative. Average performance across regimes is nearly identical (0.331 under-parameterized vs. 0.325 over-parameterized), but this masks severe instability at the threshold itself.

The failure of cross-validated $z^*$ is not accidental. Low-shrinkage values carry the highest volatility of validation Sharpe, which causes the selection procedure to pick them more often than their true average performance warrants. This connects back to SRQ1: in a tempered overfitting environment, no calibration procedure can compensate for noise that is too concentrated to diffuse harmlessly. Heavy unconditional shrinkage does better precisely because it stops trying. Permanent ascent is not achieved here, but the result is informative nonetheless.

## 4.5 Empirical Evidence of the $R^2$ Paradox (SRQ4)

*SRQ4: Does the $R^2$ paradox apply to the S&P 500? Why does low predictive accuracy not hinder good economic performance in high dimensions?*

Using the same rolling framework, the portfolio-level out-of-sample $R^2$ defined in Equation (II.12) is computed at each step $t$. The Ridge return forecast is $\hat{R}_{t+1} = \mathbf{w}_t^\top \hat{\boldsymbol{\mu}}_t$[(7)] and the benchmark $\bar{R}_{t+1}$ is the recursive mean of past realized portfolio returns, which avoids

[(7)]This is defined at the portfolio level, which differs from the individual stock-level definition used by Kelly et al. (2021).

any look-ahead bias. We fix $z = 1$ for this analysis, as it preserves enough signal in the forecast while controlling estimation noise.

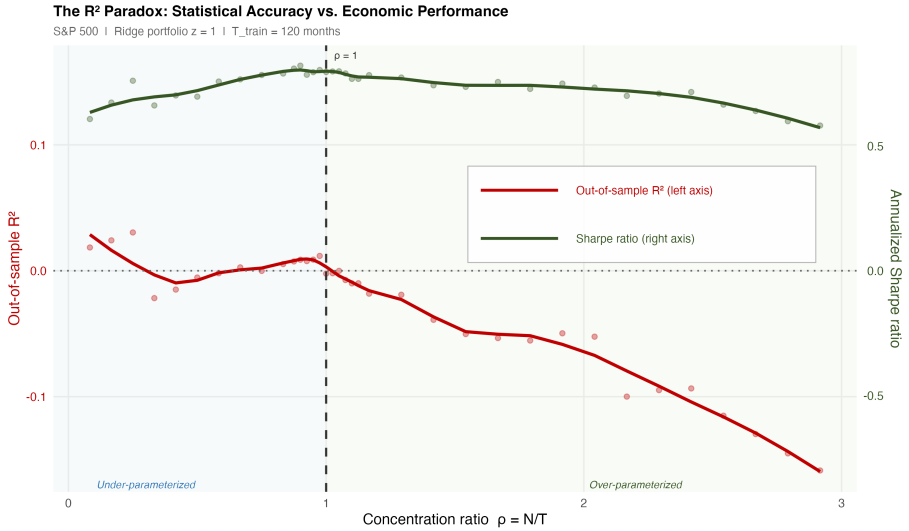



Figure IV.6: Out-of-sample $R^2$ and annualized Sharpe ratio as a function of $\rho = N/T$. Ridge portfolio with $z = 1$, $T_{\text{train}} = 120$ months, 152 rolling windows.

Figure IV.6 confirms the paradox. In the under-parameterized regime, $R^2$ hovers near zero and crosses into negative territory at the interpolation threshold. It then deteriorates steadily as $\rho$ grows, turning negative for every rolling window in the over-parameterized regime. The Sharpe ratio, by contrast, stays positive and roughly stable across all values of $\rho$. As established in Section 2.4, these two metrics measure different things: $R^2$ penalizes errors in the predicted level of returns, while portfolio performance only requires that assets be ranked correctly on average.

### Conclusion on SRQ4

The $R^2$ paradox is confirmed empirically on S&P 500 data. Predictive accuracy deteriorates as the model moves into the over-parameterized regime, with every rolling window yielding a negative $R^2$. Yet the Sharpe ratio holds steady above 0.50. Portfolio success depends on ranking assets correctly, not on predicting their exact returns.

# CONCLUSION

## Summary of Findings

This thesis examined to what extent the spectral structure of S&P 500 returns is compatible with the theoretical framework of benign overfitting, and what this implies for out-of-sample portfolio performance in a high-dimensional regime. The answer is nuanced.

The spectral diagnosis of SRQ1 established the baseline. S&P 500 returns fall in a *tempered overfitting* regime following the taxonomy of Mallinar et al. (2022). The over-parameterization condition is met but the noise tail is not diffuse enough to satisfy the isotropy requirements of Bartlett et al. (2020). The S&P 500 sits in the middle: not chaotic enough for catastrophic failure, not well-structured enough to make overfitting genuinely benign.

SRQ2 confirmed the double descent in variance. At the interpolation threshold, out-of-sample variance spikes sharply before falling back once $\rho > 1$. The over-parameterized average lands modestly below the under-parameterized floor. The descent is clear but it plateaus around $\rho \approx 1.5$ rather than continuing to decline with complexity. The pseudoinverse stabilizes the portfolio, yet the trajectory is more consistent with a tempered regime than a benign one.

SRQ3 showed that regularization is what actually matters. Fixed strong shrinkage suppresses the interpolation threshold and sustains performance across all regimes. Adaptive calibration via hold-out cross-validation fails: low-shrinkage values carry the highest volatility of validation Sharpe, causing the selection procedure to pick them more often than their true average performance warrants. Permanent ascent is not achieved.

SRQ4 confirmed the $R^2$ paradox. In the over-parameterized regime, predictive accuracy collapses while risk-adjusted performance holds. Forecasting expected returns correctly and building a profitable portfolio are simply different skills.

## Limitations

First, the discrete structure of the regularization grid $z \in \{0, 0.01, 0.1, 1, 10, 50, 100, 500\}$ provides insufficient coverage of the region $z \in (0, 10)$. Beyond $z = 10$, the validation

Sharpe is nearly flat, causing the procedure to select $z^* = 500$ by noise rather than by signal. The real blind spot is precisely this intermediate range, where regularization begins to stabilize the portfolio without over-shrinking the weights.

Second, the mean-variance Ridge estimator requires expected return estimates $\hat{\boldsymbol{\mu}}$, calculated here using the historical sample mean. This is a known weak point in portfolio optimisation (Chopra & Ziemba, 1993; Merton, 1980). The noise introduced by $\hat{\boldsymbol{\mu}}$ compounds the instability at the interpolation threshold and contributes to the difficulty of achieving permanent ascent in SRQ3.

Finally, all empirical results depend on the specific dataset used: 440 S&P 500 constituents with monthly returns over January 2000 to September 2025, yielding $\rho = 1.42$. A different universe, a shorter sample, or weekly data would shift $\rho$ and could change the regime entirely. The tempered classification, the variance spike at the interpolation threshold and the Sharpe ratio levels are empirical findings tied to large-cap US equities and should not be extrapolated without caution.

### Future Research

This thesis fixes one regularization parameter at a time, which is a structural constraint. With a single $z$, the penalty applied to each principal component follows mechanically from the formula $\lambda_i/(\lambda_i + z)$: every component is treated according to the same rule, leaving no room to handle intermediate eigenvalues differently. Kelly et al. (2023) propose a way around this with the Universal Portfolio Shrinkage Approximator (UPSA). Rather than selecting one $z$, UPSA builds a weighted combination of Ridge portfolios across a grid of penalties. When the validation Sharpe is nearly flat and no single $z$ clearly dominates, selecting one by force is arbitrary. An ensemble avoids that choice entirely. It also allows for non-linear spectral shrinkage: large noisy eigenvalues can be heavily penalized while smaller components that carry genuine signal are left room to contribute. In periods of financial stress, where correlation spikes abruptly reshape the eigenvalue spectrum, this flexibility matters even more. A natural extension would be to apply UPSA to the same S&P 500 dataset and test whether it achieves the permanent ascent that hold-out cross-validation failed to produce, by avoiding the need to select a single $z$ altogether.

## BIBLIOGRAPHY


Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070. https://doi.org/10.1073/pnas.1907378117

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Bodnar, T., Parolya, N., & Schmid, W. (2018). Estimation of the global minimum variance portfolio in high dimensions. *European Journal of Operational Research*, *266*(1), 371–390. https://doi.org/10.1016/j.ejor.2017.09.028

Bun, J., Bouchaud, J.-P., & Potters, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, *666*, 1–109. https://doi.org/10.1016/j.physrep.2016.10.005

Chopra, V. K., & Ziemba, W. T. (1993). The effect of errors in means, variances and covariances on optimal portfolio choice. *Journal of Portfolio Management*, *19*(2), 6–11. https://doi.org/10.3905/jpm.1993.409440

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. https://doi.org/10.1287/mnsc.1080.0986

Fortmann-Roe, S. (2012). *Understanding the bias-variance tradeoff* [Blog post]. Retrieved January 6, 2026, from http://scott.fortmann-roe.com/docs/BiasVariance.html

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hellum, O., Jensen, T. I., Kelly, B. T., & Malamud, S. (2026). Complex modern portfolio theory. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.6443319

Jagannathan, R., & Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, *58*(4), 1651–1683. https://doi.org/10.1111/1540-6261.00580

Kelly, B. T., Malamud, S., Pourmohammadi, M., & Trojani, F. (2023). *Universal portfolio shrinkage* (Working Paper No. 32004) (Revised April 2025). National Bureau of Economic Research. https://www.nber.org/papers/w32004

Kelly, B. T., Malamud, S., & Zhou, K. (2021). *The virtue of complexity in return prediction* (Research Paper No. 21-90) (Available at SSRN 3984925). Swiss Finance Institute. https://doi.org/10.2139/ssrn.3984925

Kelly, B. T., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*. https://doi.org/10.2139/ssrn.4501707

Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. (1999). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, *3*(3), 391–397. https://doi.org/10.1142/S0219024900000255

Lu, Y., et al. (2024). Double descent in portfolio optimization: Dance between theoretical sharpe ratio and estimation accuracy. *arXiv preprint arXiv:2411.18830*. https://doi.org/10.48550/arXiv.2411.18830

Mallinar, N., et al. (2022). Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569v3*. https://doi.org/10.48550/arXiv.2207.06569

Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, *1*(4), 457–483. https://doi.org/10.1070/SM1967v001n04ABEH001994

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. https://doi.org/10.2307/2975974

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, *8*(4), 323–361. https://doi.org/10.1016/0304-405X(80)90007-0

Michaud, R. O. (1989). The markowitz optimization enigma: Is optimized optimal? *Financial Analysts Journal*, *45*(1), 31–42. https://doi.org/10.2469/faj.v45.n1.31

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, *26*, 394–395.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51*(3), 406–413. https://doi.org/10.1017/S0305004100030401

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, *39*, 119–138. https://doi.org/10.1086/294846

Wharton Research Data Services. (2025). S&P 500 Index Data. https://wrds-www.wharton.upenn.edu/pages/classroom/the-sp-500/

Zhao, L., Chakrabarti, D., & Muthuraman, K. (2018). *Portfolio construction by mitigating error amplification: The bounded-noise portfolio* (Working Paper No. 2999407). SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2999407

# Chapter A

# RIDGE GMV PORTFOLIO ANALYSIS

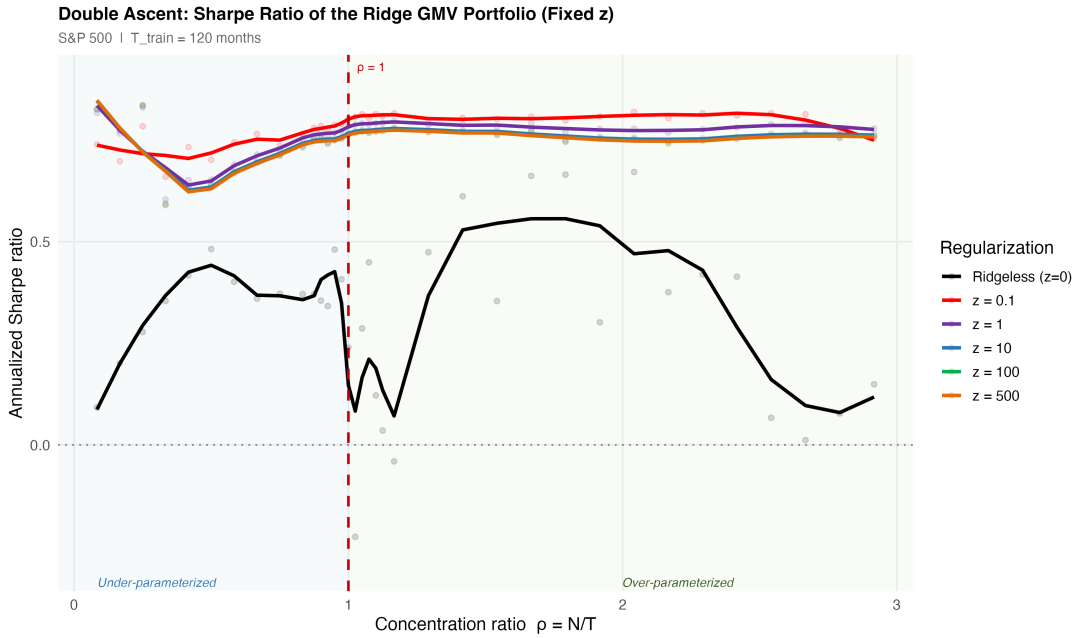



Figure A.1: Annualized Sharpe ratio of the Ridge GMV portfolio as a function of $\rho = N/T$ for several fixed values of $z$. The solid lines are LOESS fits (span $= 0.25$). The ridgeless case ($z = 0$) uses the Moore-Penrose pseudoinverse for $N \geq T_{\text{train}}$.

This figure replicates the analysis of Section 4.4.2 using a Ridge GMV portfolio, which excludes $\hat{\boldsymbol{\mu}}$ entirely. Two differences stand out relative to Figure IV.4. First, moderate regularisation ($z = 0.1$) now joins the cluster of strongly regularised curves and even produces the highest Sharpe ratios across all regimes. In the mean-variance case, $z = 0.1$ remained clearly below all curves with $z \geq 1$. Second, the ridgeless estimator ($z = 0$) is less erratic here than in the mean-variance case, though it remains unstable. Both observations point in the same direction: removing $\hat{\boldsymbol{\mu}}$ reduces the overall noise level, allowing lighter regularisation to achieve what stronger shrinkage was needed for in the mean-variance framework.

**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**