The Estimation–Efficiency Frontier in Sparse Portfolios

Jiaqin Chen*, Geng Deng*, Alberto Quaini† and Ming Yuan‡

*Wells Fargo, †Erasmus University and ‡Columbia University

January 14, 2026

Abstract

Sparse portfolios are commonly justified by trading frictions or managerial constraints. We show that, even in frictionless markets, sparsity is an economically principled form of complexity control in high-dimensional portfolio choice. Increasing the number of active holdings improves diversification but amplifies finite-sample estimation error. We formalize this tradeoff through an estimation–efficiency frontier that decomposes Sharpe-ratio losses into efficiency losses from restricting the asset set and estimation losses from learning portfolio weights. Under an approximate factor structure, these forces generate a sharp interior optimum for portfolio sparsity, yielding testable predictions for optimal portfolio complexity.

Keywords: Portfolio choice; large asset markets; sparse portfolios; mean–variance optimization; estimation risk; out-of-sample performance.

JEL classification: G11, G12, C58, C61, C55.

[0] Address for Correspondence: Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027.

1 Introduction

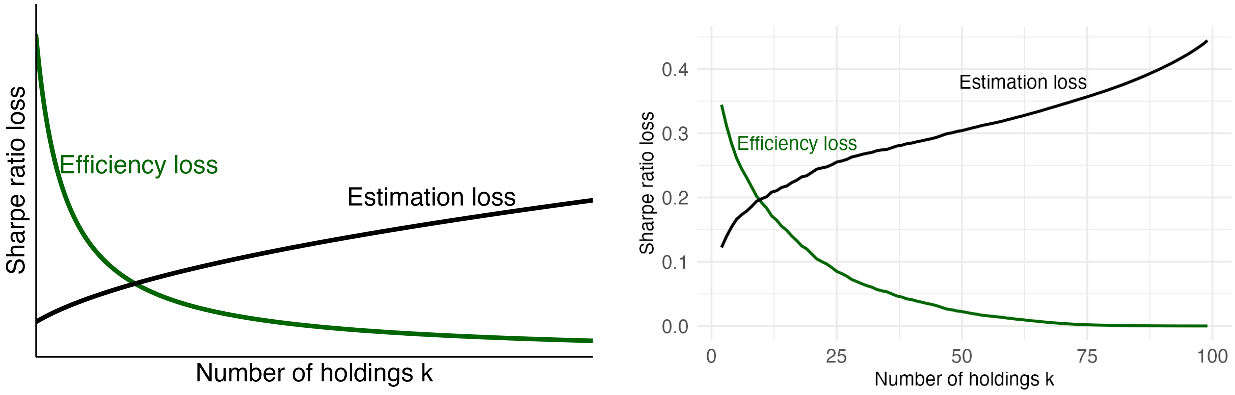
When the investment universe is large relative to the available return history, portfolio choice is not only about *which* risks to hold, but also about *how much complexity* the data can support. In such environments, the number of degrees of freedom embedded in a portfolio rule becomes an economic decision rather than a technical detail. A natural and transparent measure of this complexity is the number of active positions, or portfolio cardinality k. This paper asks a simple question: *How complex should a portfolio rule be in modern high-dimensional markets?* Equivalently, what is the economically optimal number of active positions given the sample size T, the universe size N, and the amount of redundancy in returns implied by factor structure?

The idea that investors should hold well-diversified portfolios is one of the most enduring lessons of modern finance. Markowitz (1952)'s mean–variance framework formalized this principle by demonstrating that diversification allows investors to achieve higher expected returns for a given level of risk. Yet, despite the appeal of the theory, fully diversified mean–variance portfolios have long been difficult to implement in practice. When N is large relative to T, estimates of expected returns and covariances become unreliable; see, e.g., Jobson and Korkie (1989) and Britten-Jones (1999). As a result, the classical mean–variance portfolio often performs poorly out-of-sample, producing extreme weights and unstable allocations. This fragility is not a second-order issue: efficient weights can be highly sensitive to small perturbations in the inputs (Best and Grauer, 1991; Chopra et al., 1993). Michaud (1989) emphasizes that naive plug-in efficient frontiers can be misleading and proposes resampling as a stabilization device. The core problem is that density is not free: allowing many active positions expands the parameter space that must be estimated, and the out-of-sample costs can dominate any in-sample efficiency gains.

A variety of remedies have been proposed to address this estimation risk, including shrinkage estimators for means and covariances (e.g., Ledoit and Wolf, 2003, 2004, 2017), Bayesian approaches (e.g., Jorion, 1986; Tu and Zhou, 2010; Johannes et al., 2014), and combination strategies that blend multiple portfolio rules (e.g., Tu and Zhou, 2011). Robust and ambiguity-aware formulations instead seek portfolios that remain attractive when moments are misspecified (e.g., Goldfarb and Iyengar, 2003; Tütüncü and Koenig, 2004; Garlappi et al., 2007). Another influential line of work shows that imposing seemingly "wrong" constraints can in fact improve performance. Jagannathan and Ma (2003), for example, demonstrate that prohibiting short sales, despite limiting the feasible set, reduces estimation noise sufficiently to yield better realized Sharpe ratios. In this spirit, researchers have increasingly turned to sparse portfolio selection (e.g., Chang et al., 2000; Brodie et al., 2009; Gao and Li, 2013), which restricts the number of assets that may be held. Sparse portfolios are appealing because they are simple to manage and often robust out-of-sample. The common justification, however, is practical rather than theoretical: sparsity is viewed as a response to market frictions such as transaction costs, liquidity constraints, or managerial complexity.

This paper shows that sparsity has a deeper justification. Even in the absence of any market frictions, controlling portfolio complexity through sparsity can substantially reduce estimation risk. The key insight is that limiting the number of active positions constrains the dimensionality of the estimation problem, producing a favorable bias–variance tradeoff. Complexity helps because holding more assets improves spanning and diversification; complexity hurts because it forces the investor to estimate more high-dimensional objects from finite data. Sparsity is the decision variable that implements this complexity control.

Preview of results. Our theory delivers a portfolio complexity frontier that formalizes this tradeoff and yields sharp comparative statics for the optimal complexity k^*. We proceed in two steps. First, we characterize the Sharpe-ratio loss induced by estimating a k–sparse mean–variance portfolio from T observations in an N-asset universe. We decompose this loss into selection risk (mistakes in *what* the investor holds) and allocation risk (mistakes in *how much* the investor holds conditional on the selected set), and show that the total estimation-driven Sharpe loss grows proportionally to $\sqrt{k \log N/T}$, capturing both allocation error and the statistical cost of searching over a large universe. Second, we study the opportunity cost of complexity control: how much mean–variance efficiency is lost by restricting attention to k assets even when moments are known. Under an approximate factor model (Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1993), we show that this efficiency loss declines rapidly with k and depends on factor strength. Under strong factors it scales at most as $1/k$, while under weaker factors it scales as $N^{1-\zeta}/k$, where $\zeta \in (0, 1]$ measures factor strength.



(a) Conceptual illustration.

(b) Simulation evidence.

Figure 1: **Estimation–efficiency tradeoff in sparse portfolios.** Panel (a) illustrates efficiency loss ($1/k$) and estimation risk ($\sqrt{k \log N/T}$). Panel (b) is obtained from the simulation design in Section 5, with $N = 100$ assets generated from a three-factor model.

The economic force governing the optimal sparsity level k^* is redundancy: strong factor structure implies that many assets span similar risks, so excluding most of them entails little efficiency loss. Weak factor structure, by contrast, requires holding more assets to achieve comparable diversification.

Combining the two forces yields an interior sweet spot for sparsity. In particular, sparse portfolios can be nearly mean–variance efficient when k grows faster than $N^{1-\zeta}$ yet remains much smaller than $T/\log N$. Moreover, balancing the estimation-loss term $\sqrt{k \log N/T}$ against the efficiency-loss term $N^{1-\zeta}/k$ yields the scaling law

$$k^* \asymp \left(N^{1-\zeta}\sqrt{T/\log N}\right)^{2/3},$$

which is the paper's central economic prediction about optimal portfolio complexity: the optimal number of active positions increases with sample size T, increases as redundancy declines (weaker factors, smaller ζ), and varies systematically with the universe size N. Figure 1 summarizes the implied tradeoff between efficiency loss and estimation risk.

Implementing an exact cardinality constraint is computationally demanding in realistic universes (Chang et al., 2000; Gao and Li, 2013; Bertsimas and Cory-Wright, 2022). We therefore also study a tractable relaxation based on ℓ_1 regularization (Brodie et al., 2009), which is closely related to the lasso (Tibshirani, 1996). This relaxation is also closely connected to gross-exposure and ℓ_1-type stabilization in portfolio choice (Fan et al., 2012). We show that the ℓ_1-regularized portfolio inherits the same economic forces: it trades off approximation error from restricting the investment universe against statistical error from estimating high-dimensional moments. Under suitable scaling of the penalty and regularity conditions, the resulting portfolios attain the same asymptotic efficiency guarantees as the cardinality-constrained rule in the high-dimensional regime. Moreover, the approach remains stable when the factor structure is only approximate—so expected returns include nontrivial idiosyncratic components rather than satisfying exact factor pricing—making it well suited for empirical implementation.

The sweet-spot prediction is visible in the data. Figure 2 plots the realized out-of-sample Sharpe ratio of the ℓ_1-regularized mean–variance strategy as a function of the target sparsity level k in a large monthly managed-portfolio universe. The Sharpe ratio profile is hump-shaped: performance improves sharply as k increases from very small values, but eventually deteriorates as the portfolio becomes dense. This pattern mirrors the complexity frontier: the opportunity cost of excluding assets falls quickly with k under a factor structure, while estimation risk rises with the dimensionality of the inputs learned from a finite sample. See Section 6 for further discussion.

Our work complements a growing literature that links portfolio optimization with high-dimensional statistics. Shrinkage estimators for covariance matrices (Ledoit and Wolf, 2003, 2004, 2017), high-dimensional regularization (Fan et al., 2012; Ao et al., 2019), and algorithmic sparse selection (Bertsimas and Cory-Wright, 2022) all attempt to stabilize estimation. In operations research, sparse mean–variance selection has long been studied through heuristics and mixed-integer formulations (Chang et al., 2000; Gao and Li, 2013). Related econometric approaches use structured ℓ_1 penalties to induce sparsity while retaining convexity (Kremer et al., 2020) and propose generalizations of sparse portfolio rules under alternative regularizers (Fastrich et al., 2015; Dai and Wen, 2018). By offering an explicit decomposition of estimation risk and showing that sparsity achieves optimal risk reduction, we provide a theoretical foundation that unifies these approaches with the economic logic of complexity control in portfolio choice.

Moreover, while our results are closely related to the high-dimensional statistics literature,

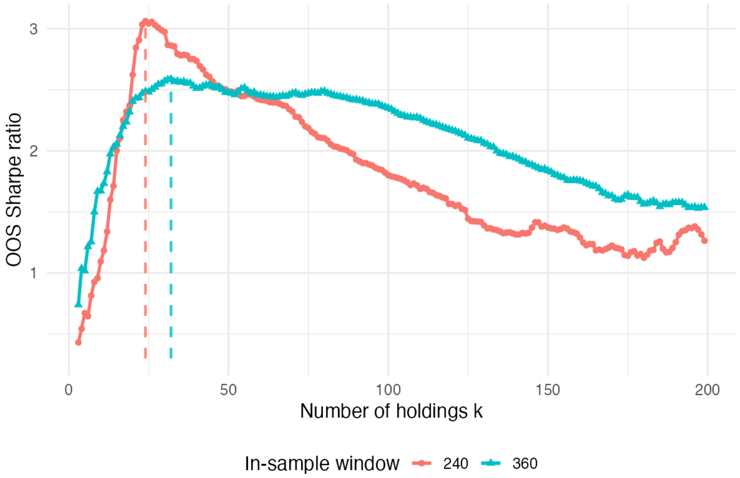


Figure 2: Out-of-sample Sharpe ratio as a function of the target cardinality k in the monthly managed-portfolio universe ($N = 200$). The vertical dashed lines indicate the value of k that maximizes the realized out-of-sample Sharpe ratio for each in-sample estimation window length ($T \in \{240, 360\}$ months). Portfolios are computed from the ℓ_1 relaxation and evaluated one month ahead ($h = 1$) using a rolling-window out-of-sample procedure.

there are fundamental differences that materially affect both interpretation and applicability. A central assumption in high-dimensional statistics is that the parameter of interest, though high dimensional, is sparse or approximately sparse; see, for example, Bühlmann and Van De Geer (2011); Hastie et al. (2015). This premise is ill-suited for portfolio choice. In empirically relevant environments, most notably when asset returns admit a factor structure, the true (but infeasible) mean–variance efficient portfolio necessarily assigns nonzero weights to essentially all assets. Importantly, our results demonstrate that sparsity in portfolio implementation does not require sparsity of the population optimum. Even when the efficient portfolio is fully dense, sparse portfolio selection, obtained either via cardinality constraints or ℓ_1 penalization, remains capable of achieving near-efficiency. With a suitable choice of the sparsity level, these procedures deliver feasible portfolios that are asymptotically mean–variance efficient, thereby reconciling sparsity-based methods with economically realistic return structures. A second key distinction concerns regularity conditions on the design matrix. Much of the high-dimensional statistics literature relies on restricted eigenvalue–type conditions imposed on the matrix of historical returns; see, for example, van de Geer and Bühlmann (2009). These assumptions are not only technical and unverifiable in practice (Bandeira et al., 2013), but are also systematically violated in financial data, particularly in the presence of common risk factors. By contrast, our analysis does not rely on such conditions and is therefore well suited to empirically relevant asset return models.

Taken together, our results establish sparsity as a principled form of risk control in portfolio choice. By limiting the complexity of the estimated model, sparsity mitigates the impact of sampling noise on portfolio performance. Under broad conditions, this reduction in estimation risk outweighs the loss from excluding some assets, leading to portfolios that are both stable and nearly efficient. In large markets, the opportunity cost of imposing sparsity becomes vanishingly small. Hence, even in a frictionless setting, a rational investor may prefer a sparse portfolio not because trading all assets is costly, but because density is statistically costly: it enlarges the estimation problem and can expose her to substantial estimation uncertainty.

This perspective reframes the role of constraints in portfolio theory. Constraints such as no-short-sale (Jagannathan and Ma, 2003), norm limits (e.g., DeMiguel et al., 2009), robust-optimization uncertainty sets (e.g., Goldfarb and Iyengar, 2003; Tütüncü and Koenig, 2004), or cardinality restrictions are often viewed as distortions introduced for practical reasons. Our findings suggest instead that such constraints can serve as statistical regularizers, improving decision quality in the presence of uncertain inputs. From this viewpoint, the optimal degree of sparsity depends not on trading costs but on the relative magnitudes of estimation noise and diversification benefits. This reinterpretation connects classical portfolio theory with modern statistical learning and offers a unified framework for understanding when and why simpler portfolios perform better. Sparse portfolios also provide transparent allocations that are easier to explain and govern—an increasingly important consideration for institutional investors operating under regulatory and fiduciary scrutiny.

The remainder of the paper is organized as follows. Section 2 introduces the mean–variance framework, and defines the sparse portfolio rules. Section 3 develops the central tradeoff between estimation risk and the opportunity cost of sparsity, and derives conditions under which sparse portfolios are asymptotically efficient under approximate factor structure. Section 4 discusses the ℓ_1 relaxation as a practical sparse portfolio implementation. Section 5 presents simulation evidence, and Section 6 reports empirical results. Section 7 concludes.

2 Framework and Sparse Portfolio Rules

This section sets up the framework for studying the *portfolio complexity frontier*. When the investment universe is large and the available return history is limited, portfolio choice is not only about which risks to load on; it is also about how many degrees of freedom the investor can reliably estimate from finite data. We measure portfolio complexity by the number of active positions, that is, the portfolio cardinality k. The analysis below formalizes a tradeoff: increasing k can improve spanning and diversification, but it also expands the parameter space that must be estimated and can amplify estimation risk when N is large relative to T. Sparsity is the decision variable that implements this complexity control.

We begin from the mean–variance formulation of portfolio choice, which provides a benchmark notion of efficiency through the population Sharpe ratio. We then define the sample mean–variance portfolio and two sparse portfolio rules: a cardinality-constrained strategy and an ℓ_1-regularized strategy. We view both rules as statistical regularizers that move the investor along the complexity frontier in large universes. Notation is defined as it is introduced, and Appendix A collects all notation used throughout the theory.

2.1 Framework

Consider an investment universe of N risky assets with excess returns collected in the random vector $r \in \mathbb{R}^N$. An investor chooses portfolio weights $w \in \mathbb{R}^N$ and forms the portfolio return

$$r_{p,w} = w^\top r = w_1 r_1 + \ldots + w_N r_N.$$

Let $\mu := \mathbb{E}[r]$ and $\Sigma := \mathbb{V}[r]$ denote the mean and covariance matrix of r.

We evaluate portfolios by their Sharpe ratios. The Sharpe ratio of portfolio w is

$$\theta(w) := \frac{\mathbb{E}(r_{p,w})}{\sqrt{\mathbb{V}(r_{p,w})}} = \frac{w^\top \mu}{\sqrt{w^\top \Sigma w}}.$$

The population-efficient Sharpe ratio is

$$\theta^* := \max_{w \in \mathbb{R}^N} \theta(w).$$

When Σ is of full rank and $\mu \neq 0$, this maximum is attained and satisfies

$$\theta^* = \sqrt{\mu^\top \Sigma^{-1} \mu}.$$

We adopt the classical mean–variance objective

$$\mathbb{E}[r_{p,w}] - \frac{\gamma}{2}\mathbb{V}(r_{p,w}) = w^\top \mu - \frac{\gamma}{2} w^\top \Sigma w,$$

where $\gamma > 0$ measures risk aversion. The maximizer is

$$w_* := \underset{w \in \mathbb{R}^N}{\operatorname{argmax}} \Big\{ w^\top \mu - \frac{\gamma}{2} w^\top \Sigma w \Big\} = \frac{1}{\gamma} \Sigma^{-1} \mu.$$

This portfolio also attains the population-efficient Sharpe ratio:

$$\theta(w) \leq \theta(w_*) = \theta^*,$$

and equality holds if and only if $w \propto w_*$, provided that $\theta^* > 0$. Because $\theta(\cdot)$ is scale invariant, we interpret w_* as the Sharpe-optimal direction (the tangency portfolio) up to proportionality; the parameter γ only scales leverage. We therefore use w_* as a benchmark for measuring the performance of portfolio rules. In particular, we quantify the suboptimality of any portfolio weight w by

$$\theta^* - \theta(w).$$

In practice, the population moments μ and Σ are unknown and must be estimated from a sample $(r_1, \ldots, r_T)$. A natural empirical analogue of the mean–variance objective is

$$\widehat{U}_T(w) := w^\top \widehat{\mu} - \frac{\gamma}{2} w^\top \widehat{\Sigma} w,$$

where

$$\widehat{\mu} := \frac{1}{T} \sum_{t=1}^{T} r_t, \qquad \widehat{\Sigma} := \frac{1}{T} \sum_{t=1}^{T} (r_t - \widehat{\mu})(r_t - \widehat{\mu})^\top. \tag{1}$$

The corresponding plug-in mean–variance portfolio is

$$\widehat{w} := \underset{w \in \mathbb{R}^N}{\operatorname{argmax}}\, \widehat{U}_T(w) = \frac{1}{\gamma} \widehat{\Sigma}^{-1} \widehat{\mu},$$

provided that $\widehat{\Sigma}$ is nonsingular.

Although intuitive, this estimator is notoriously fragile when the cross-sectional dimension N is not small relative to the sample size T (Jobson and Korkie, 1980; Britten-Jones, 1999; Best and Grauer, 1991; Chopra et al., 1993; Michaud, 1989; Ledoit and Wolf, 2004; Kan and Smith, 2008; El Karoui, 2010). Sampling error in $\widehat{\mu}$ and $\widehat{\Sigma}$ can produce unstable weights and large out-of-sample losses. When N is comparable to or exceeds T, the covariance estimate may even be singular, rendering $\widehat{w}$ undefined. This fragility illustrates the economic content of the complexity choice: allowing many degrees of freedom forces the investor to learn high-dimensional objects from finite data, and small input errors can translate into large swings in estimated optimal weights.

2.2 Sparse Portfolio Rules

A natural way to control estimation risk is to control portfolio complexity. Instead of estimating all N weights freely, the investor chooses an effective dimension $k \ll N$. Increasing k expands the feasible span and improves diversification, but it also raises the statistical burden of estimating the inputs and can lead to overfitting when T is limited. Even in frictionless markets, this tradeoff is central: the investor may prefer to exclude some assets not because trading is costly, but because doing so reduces the sensitivity of the portfolio rule to estimation noise.

Sample cardinality-constrained strategy. For an integer $N \geq 1$, write $[N] := \{1, \ldots, N\}$. Let $\operatorname{supp}(w) := \{i \in [N] : w_i \neq 0\}$ denote the support (index set of nonzero elements) of a weight vector $w \in \mathbb{R}^N$, and let $\|w\|_0 := |\operatorname{supp}(w)|$ denote the number of active positions. A k–sparse portfolio satisfies $\|w\|_0 \leq k$, meaning that the investor takes positions in at most $k \geq 0$ assets. The sample cardinality-constrained portfolio selection problem is

$$\widehat{w}_k := \underset{w \in \mathbb{R}^N:\ \|w\|_0 \leq k}{\operatorname{argmax}} \quad \widehat{U}_T(w). \tag{2}$$

The tuning parameter k governs complexity. When $k = N$, (2) reduces to the unrestricted mean–variance case; as k decreases, the problem imposes stronger regularization by restricting the number of degrees of freedom that the data must pin down. Although (2) is nonconvex and generally NP-hard (Gao and Li, 2013), it provides a useful benchmark for understanding how limiting the number of active positions affects out-of-sample performance.[1]

ℓ_1-regularized strategy. A computationally tractable alternative replaces the hard cardinality constraint with a continuous penalty that encourages sparsity by penalizing gross exposure,

$$\widehat{w}_{\ell_1,\lambda} := \underset{w \in \mathbb{R}^N}{\operatorname{argmax}} \left\{ \widehat{U}_T(w) - \lambda \|w\|_1 \right\}, \tag{3}$$

where $\|w\|_1 := \sum_{i=1}^N |w_i|$ and $\lambda \geq 0$ controls the degree of shrinkage; see, e.g., Brodie et al. (2009). When $\lambda = 0$, (3) reduces to the unrestricted mean–variance portfolio; as λ increases, small positions are pushed toward zero, yielding parsimonious allocations. Economically, $\|w\|_1$ is gross exposure in a long–short book, so penalizing it directly limits the optimizer's ability to turn small estimation errors into large offsetting positions. In this sense, ℓ_1 regularization controls complexity even when exact cardinality constraints are computationally infeasible.

The key point is that the ℓ_1 penalty acts like a linear holding cost per unit of gross position. An asset enters the portfolio only if its marginal contribution to the sample objective exceeds this cost. Because the cost has a kink at zero (the marginal penalty is discontinuous when a weight moves away from zero), the optimizer typically satisfies a thresholding rule: many assets have insufficient estimated risk-adjusted benefit to justify a nonzero position and therefore remain exactly at $w_i = 0$. As a result, ℓ_1 regularization delivers exact sparsity while remaining convex and computationally efficient, making it a natural proxy for cardinality constraints in large universes and a practical way to move along the portfolio complexity frontier. More details are given in Section 4.

[1] NP-hard means that, in general, the computational time required to find the exact global optimum can grow exponentially with the number of assets N, so exact solution methods may become infeasible in large universes.

2.3 Sharpe Ratio Losses

Our objective $\widehat{U}_T(w)$ is a convenient device for constructing portfolios from data, but performance is naturally assessed out of sample using the population Sharpe ratio $\theta(\cdot)$ defined above. Because $\theta(\cdot)$ is scale invariant, all comparisons below are understood up to proportionality of portfolio weights.

Complexity frontier and efficiency loss. For any index set $A \subset [N]$, let r_A denote the subvector of returns with indices in A, and let (μ_A, Σ_A) be the corresponding population moments. Define the best attainable Sharpe ratio when trading only assets in A as

$$\theta_A := \max_{\operatorname{supp}(w) \subset A} \theta(w) = \sqrt{\mu_A^\top \Sigma_A^{-1} \mu_A}, \tag{4}$$

where the closed form holds whenever Σ_A is full rank and $\mu_A \neq 0$. Restricting attention to at most k active positions yields the population *portfolio complexity frontier*

$$\theta_k^* := \max_{\|w\|_0 \leq k} \theta(w) = \max_{A \subset [N]:\, |A| \leq k} \theta_A. \tag{5}$$

The associated *efficiency loss* (or opportunity cost) of restricting the investable span to k assets is

$$\Delta_k := \theta^* - \theta_k^* \geq 0, \tag{6}$$

which is purely a population object: it captures the economic cost of limiting complexity under perfect knowledge of (μ, Σ).

Estimation loss and its selection–allocation decomposition. Let $\widehat{w}_k$ be the sample k–sparse portfolio in (2) with selected support (holdings) given by

$$\widehat{A}_k := \operatorname{supp}(\widehat{w}_k) = \{i \in [N] : \ \widehat{w}_{k,i} \neq 0\}. \tag{7}$$

Because $\widehat{w}_k$ is constructed from noisy moments, its population Sharpe ratio $\theta(\widehat{w}_k)$ typically falls short of θ_k^* even when k is fixed. We refer to the gap

$$\theta_k^* - \theta(\widehat{w}_k) \tag{8}$$

as the *(pointwise) estimation loss* at complexity level k, and we study its expectation over samples in the next section.

A key advantage of sparse portfolios is that their estimation loss admits a sharp decomposition that separates mistakes in *what* is held from mistakes in *how much* is held. Let $\theta_{\widehat{A}_k}$ denote the *oracle* Sharpe ratio on the selected support, obtained by re-optimizing with population moments while keeping holdings fixed:

$$\theta_{\widehat{A}_k} := \max_{\operatorname{supp}(w)\subset \widehat{A}_k} \theta(w). \tag{9}$$

Then we have the identity

$$\underbrace{\theta_k^* - \theta(\widehat{w}_k)}_{\text{Estimation loss}} = \underbrace{\left[\theta_k^* - \theta_{\widehat{A}_k}\right]}_{\text{Selection loss}} + \underbrace{\left[\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\right]}_{\text{Allocation loss}}. \tag{10}$$

The first term measures the cost of selecting an inferior subset of assets relative to the best k-asset subset; it isolates mistakes in holdings and is therefore a natural measure of *selection risk*. The second term measures the loss from estimating weights conditional on the selected holdings; it isolates errors in allocations given the chosen asset set and is therefore a measure of *allocation risk*.

Total Sharpe-ratio loss. Combining (6) and (10) yields an economically transparent decomposition of the realized suboptimality of sparse mean–variance portfolios:

$$\underbrace{\theta^* - \theta(\widehat{w}_k)}_{\text{Total loss}} = \underbrace{\Delta_k}_{\text{Efficiency loss}} + \underbrace{\left[\theta_k^* - \theta_{\widehat{A}_k}\right]}_{\text{Selection loss}} + \underbrace{\left[\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\right]}_{\text{Allocation loss}}. \tag{11}$$

The first term decreases with k by construction, reflecting improved spanning as the portfolio becomes more complex. The latter two terms capture estimation risk and typically increase with complexity, because more holdings require learning more inputs and create more scope for overfitting.

Decomposition (11) applies to $\widehat{w}_{\ell_1,\lambda}$ in (3) as well, by replacing $\widehat{w}_k$ with $\widehat{w}_{\ell_1,\lambda}$ and interpreting $\operatorname{supp}(\widehat{w}_{\ell_1,\lambda})$ as the λ–induced active set. In empirical work, it is often convenient to index λ by the induced cardinality $\widehat{k}(\lambda) := \|\widehat{w}_{\ell_1,\lambda}\|_0$, so that the ℓ_1 path can be compared directly to the k–sparse frontier.

3 Estimation and Efficiency Losses

This section quantifies the magnitude of the components in the Sharpe-ratio loss decomposition (10) and clarifies how they scale with (N, T, k) under economically standard distributional and factor-structure conditions. The goal is to formalize the portfolio complexity frontier by quantifying how the Sharpe-ratio loss varies with the number of active positions k in a large investment universe. At a fixed complexity level k, the investor faces a purely statistical problem: how much Sharpe-ratio performance is lost because the portfolio is constructed from estimated moments rather than from (μ, Σ)? At the same time, restricting complexity has an opportunity cost: even with perfect knowledge of (μ, Σ), limiting the investable span to k assets can reduce the attainable Sharpe ratio, with the magnitude of this loss governed by redundancy in returns and the strength of factor structure.

We proceed in two parts. First, we provide high-dimensional bounds for the estimation loss $\theta_k^* - \theta(\widehat{w}_k)$ and its selection–allocation decomposition. Second, we characterize the rate at which the efficiency loss Δ_k declines with k when returns admit an approximate factor pricing structure.

As a global summary, define the *estimation risk* at complexity level k by

$$\theta_k^* - \theta(\widehat{w}_k) = \left[\theta_k^* - \theta_{\widehat{A}_k}\right] + \left[\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\right].$$

This object isolates the expected Sharpe-ratio loss due solely to statistical uncertainty at a fixed portfolio complexity. Economically, it is the price of asking the data to identify a k-dimensional portfolio rule in an N-asset universe from only T observations. The next subsections bound the two components, the selection risk and allocation risk, and establish their scaling in k, N, and T.

3.1 Selection Risk

To study selection risk, it is useful to make explicit how the cardinality-constrained estimator chooses its holdings. For any subset $A \subset [N]$, write $\widehat{\mu}_A$ and $\widehat{\Sigma}_A$ for the subvector and principal submatrix of $(\widehat{\mu}, \widehat{\Sigma})$ indexed by A, and define the plug-in Sharpe ratio statistic

$$\widehat{\theta}_A := \max_{\mathrm{supp}(w)\subset A} \frac{w^\top \widehat{\mu}}{\sqrt{w^\top \widehat{\Sigma} w}} = \sqrt{\widehat{\mu}_A^\top \widehat{\Sigma}_A^{-1} \widehat{\mu}_A},$$

where the closed form holds whenever $\widehat{\Sigma}_A$ is full rank. Under mean–variance utility, the maximal in-sample value achievable on a fixed subset A is a strictly increasing function of $\widehat{\theta}_A^2 := (\widehat{\theta}_A)^2$. Hence, the cardinality-constrained estimator selects the support by solving the equivalent best-subset problem:[2]

$$\widehat{A}_k \in \operatorname*{argmax}_{A\subset[N]:\ |A|\le k} \widehat{\theta}_A^2.$$

This formulation clarifies that selection risk is fundamentally an extremum-estimation problem over a large collection of candidate holdings. The challenge is not estimating θ_A for a single subset A, but ranking many subsets using noisy performance estimates. The relevant object is thus the uniform fluctuation $\sup_{|A|\le k} |\widehat{\theta}_A^2 - \theta_A^2|$: we can only assert that a subset of assets is preferable to another if its estimated Sharpe ratio is sufficiently larger, and controlling this winner's curse effect determines the rate of selection risk.

Proposition 1 (Bound on selection risk)**.** *Suppose $r_t \sim \mathcal{N}(\mu, \Sigma)$ independently over t, and Σ is invertible so that $\theta^* = (\mu^\top \Sigma^{-1} \mu)^{1/2} < \infty$. Then for any $k < N$,*

$$\theta_k^* - \theta_{\widehat{A}_k} = O_p\left(\sqrt{\frac{\log \binom{N}{k}}{T}}\right), \qquad \textit{as } T \to \infty.$$

Economically, Proposition 1 says that the cost of choosing the wrong holdings falls with the length of the estimation window, provided the investor does not attempt to learn an overly complex portfolio from limited data. The key mechanism is selection under uncertainty. For a fixed subset A with $|A| \le k$, the estimation error in $\widehat{\theta}_A^2$ is of order $\sqrt{|A|/T}$. Selection, however, depends on the *largest* such error over all admissible subsets, because the chosen support is the maximizer of a noisy performance statistic.

The $\log \binom{N}{k}$ factor is the statistical cost of searching across a broad universe (Bühlmann and Van De Geer, 2011; Hastie et al., 2015). Note that both θ_A and $\widehat{\theta}_A$ are monotone

[2] see Lemma 1 in Appendix D.

in set inclusion: if $A \supseteq B$ then $\theta_A \geq \theta_B$ and $\widehat{\theta}_A \geq \widehat{\theta}_B$. Hence, whenever $k < N$ and $\theta^* > 0$, the optimizer uses the full budget, and it suffices to restrict attention to supports with $|A| = k$; equivalently, $|\widehat{A}_k| = k$. The class of admissible supports then has size $\binom{N}{k}$, which grows rapidly with N even for moderate k. Controlling the maximum deviation $\sup_{|A|=k} |\widehat{\theta}_A^2 - \theta_A^2|$ therefore requires paying a complexity term of order $\log \binom{N}{k}$. Using the standard bound $\binom{N}{k} \leq (eN/k)^k$, this logarithm scales like $k \log N$ (up to lower-order terms), yielding the uniform fluctuation rate $\sqrt{(k \log N)/T}$. In financial terms, as the investment universe expands, it becomes increasingly likely that at least one subset appears to deliver an exceptionally high in-sample Sharpe ratio purely by chance, and selection risk quantifies the performance penalty of acting on that spurious ranking.

Remark 1 (Unbiased estimate of θ_A^2)**.** *The naive plug-in statistic $\widehat{\theta}_A^2$ is known to be upwardly biased. Motivated by the exact Gaussian finite-sample analysis in Kan and Zhou (2007) (see also Kan and Smith (2008); Kan et al. (2024)), it is convenient to also consider the associated bias-adjusted estimator of the squared population Sharpe ratio,*

$$\tilde{\theta}_A^2 := \frac{T - |A| - 2}{T} \widehat{\theta}_A^2 - \frac{|A|}{T}, \qquad \tilde{\theta}_A := \sqrt{\max\{\tilde{\theta}_A^2, 0\}}.$$

Following the same argument as in Proposition 1, we can show that the selection risk incurred by maximizing $\tilde{\theta}_A$ is also on the order of $\sqrt{(k \log N)/T}$.

Remark 2 (Gaussianity, tails, and time dependence)**.** *The Gaussian i.i.d. assumption is adopted for analytical transparency. We do not interpret it as a literal model for asset returns, which are typically heavy-tailed and serially dependent. The $\sqrt{k \log N/T}$ scaling, however, is a high-dimensional benchmark rather than a Gaussian artifact. Similar rates can be obtained under substantially weaker conditions, provided that for each fixed A the moment estimators admit sub-Gaussian concentration and that dependence is sufficiently weak (e.g., mixing) so that an effective sample size remains proportional to T. Appendix E derives the same selection and allocation risk bounds as in Propositions 1 and 2 under weak serial dependence.*

Under heavier tails, the same scaling can generally be recovered by using robust or truncated estimators of (μ, Σ) on each subset; without such robustification, extreme observations can dominate estimated Sharpe ratios and amplify the winner's curse effect induced by searching over many supports.

3.2 Allocation Risk

Selection is only the first step. Conditional on a selected subset A, the investor still forms weights using estimated moments. For any subset $A \subset [N]$, let $\widehat{w}_A$ denote the mean–variance portfolio computed from $(\widehat{\mu}_A, \widehat{\Sigma}_A)$ and embedded into $\mathbb{R}^N$ by setting weights outside A to zero, that is,

$$\widehat{w}_A := \underset{w \in \mathbb{R}^N:\ \mathrm{supp}(w) \subset A}{\operatorname{argmax}} \widehat{U}_T(w).$$

When $\widehat{\Sigma}_A$ is full rank, the active weights are given by

$$\frac{1}{\gamma} \widehat{\Sigma}_A^{-1} \widehat{\mu}_A.$$

The population benchmark on A attains Sharpe ratio θ_A, whereas the implementable portfolio $\widehat{w}_A$ uses estimated moments and therefore incurs weight-estimation error.

For a fixed subset A with $|A| \leq k$, standard mean–variance arguments imply that the resulting Sharpe shortfall $\theta_A - \theta(\widehat{w}_A)$ is of order $\sqrt{|A|/T}$, reflecting that one is estimating a $|A|$-dimensional set of optimality conditions. The sparse rule, however, does not condition on a fixed A: it selects $\widehat{A}_k$ by searching over $\sum_{j \leq k} \binom{N}{j}$ candidates and then estimates weights on the chosen holdings. Hence the relevant object is allocation error evaluated on a data-driven support, which requires controlling weight-estimation error uniformly over the same class of candidate subsets.

Proposition 2 (Bound on allocation risk)**.** *Under the assumptions of Proposition 1,*

$$\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}}\right), \qquad \text{as } T \to \infty.$$

A few remarks are in order. First, sparsity keeps the intrinsic dimension of the allocation problem at k, so the baseline fixed-support error behaves like $\sqrt{k/T}$. Second, because the holdings are chosen endogenously through an extremum over many subsets, the moments on the selected support are subject to the same winner's curse forces as in selection: among many candidates, some subsets look unusually favorable in sample. Bounding the induced weight-estimation error uniformly over candidate subsets therefore introduces the same $\log N$ search penalty. It is worth noting that when $k \ll N$, the allocation bound has the same scaling as the selection bound. Economically, allocation risk measures the sensitivity of mean–variance exposures to small perturbations in estimated moments; sparsity limits this sensitivity by restricting dimensionality, while data-driven selection requires that this sensitivity be controlled uniformly across the broad set of candidate holdings.

Combining Propositions 1 and 2 with the decomposition (10) yields an overall characterization of estimation risk.

Theorem 1 (Estimation risk of sparse portfolios)**.** *Suppose $r_t \sim \mathcal{N}(\mu, \Sigma)$ independently over t, and Σ is invertible so that $\theta^* = (\mu^\top \Sigma^{-1}\mu)^{1/2} < \infty$. Then*

$$\theta_k^* - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}}\right), \qquad \textit{as } T \to \infty.$$

Theorem 1 formalizes the central statistical force behind the portfolio complexity frontier. Holding T fixed, allowing more active positions increases estimation risk at rate $\sqrt{k}$ (up to $\log N$), because the investor must estimate and combine more unknown moments and because holdings are chosen by searching over many candidate portfolios. This increasing-in-k estimation loss corresponds to the "estimation risk" curve in Figure 1. It is the statistical counterpart of the well-known instability of plug-in mean–variance weights in large universes.

The rate also delivers a transparent feasibility condition for reliable optimization: for estimation loss to be asymptotically negligible it is sufficient that $k \log N = o(T)$, so that portfolio complexity grows slowly relative to available data. When $k \log N$ is not small relative to T, even mild overfitting in estimated means and covariances can translate into economically meaningful Sharpe-ratio losses. In the next subsection, we complement this increasing-in-k estimation cost with a decreasing-in-k efficiency cost, and we show how factor strength and redundancy determine how quickly the opportunity cost of sparsity disappears as k grows.

Remark 3 (Magnitude in financial applications)**.** *Theorem 1 is an order statement, so it should be read as a diagnostic for when estimation risk is likely to be economically first order rather than as a tight quantitative bound. A useful benchmark is the unitless complexity index $\sqrt{k \log N/T}$ that governs the slope of the "estimation risk" curve in Figure 1. For monthly managed-portfolio sorts with $N = 50$ portfolios and a typical in-sample window* *of $T = 60$ months (five years), this index is about 0.57 for $k = 5$ and 0.81 for $k = 10$. For a higher-dimensional monthly universe with $N = 300$ portfolios and the same $T = 60$, it rises to about 0.69 ($k = 5$), 0.98 ($k = 10$), and 2.18 ($k = 50$), suggesting that dense allocations are particularly prone to overfitting at standard estimation horizons. In daily individual-asset applications with $N = 3000$ and $T \approx 1260$ trading days (five years), the index is substantially smaller for small k—about 0.18 ($k = 5$) and 0.25 ($k = 10$)—but it is still non-negligible for moderately complex portfolios (about 0.56 for $k = 50$). With a longer daily window of $T \approx 7560$ (thirty years), these values fall further to about 0.07 ($k = 5$), 0.10 ($k = 10$), and 0.23 ($k = 50$), consistent with the idea that richer time-series information can support higher portfolio complexity.*

Two caveats are important for interpretation. First, the constant hidden in the $O_p(\cdot)$ term depends on signal strength and conditioning of moments, so these numbers should be viewed as relative indicators of difficulty across settings. Second, the i.i.d. Gaussian benchmark is optimistic for financial returns: serial dependence and heavy tails reduce the effective sample size, steepening the estimation-risk curve in Figure 1 and pushing the economically optimal complexity k^* toward smaller values in practice.

3.3 Efficiency Loss: The Opportunity Cost of Sparsity

Estimation risk accounts for one side of the portfolio complexity frontier. The other side is a purely economic opportunity cost. Even with perfect knowledge of (μ, Σ), restricting attention to portfolios with $\|w\|_0 \leq k$ can reduce the best attainable Sharpe ratio from the unconstrained benchmark θ^* to the k–sparse frontier θ_k^* in (5). The resulting *efficiency loss*,

$$\theta^* - \theta_k^*,$$

measures the cost of limiting the investable span to k assets. This object isolates the benefit of complexity on the economic side: increasing k expands spanning possibilities and improves diversification even when moments are known.

To characterize $\theta^* - \theta_k^*$ one must restrict the return distribution. Absent structure, the mean–variance efficient direction $w_* \propto \Sigma^{-1}\mu$ may load on many weakly informative directions, and there is no disciplined way to pin down how quickly the k–sparse frontier approaches θ^* as k grows. We therefore study the efficiency loss under factor structures for returns. This assumption isolates the economically relevant case in which the mean–variance frontier is largely spanned by a low-dimensional set of systematic risks. Under such structure, a sparse portfolio can approximate the priced component of returns with a limited number of holdings; the remaining gains from increasing k come primarily from diversifying idiosyncratic risk. Redundancy in returns—captured by factor pervasiveness and, more generally, factor strength—therefore governs how costly it is to limit portfolio complexity.

3.3.1 One-factor benchmark

We begin with a simple one-factor environment that makes the mechanics transparent. Suppose

$$\mu = \mu_m \beta, \qquad \Sigma = \sigma_m^2 \beta\beta^\top + \sigma_0^2 I_N, \tag{12}$$

where $\mu_m \in \mathbb{R}$ and $\sigma_m^2 > 0$ are the mean and variance of the factor return, $\beta \in \mathbb{R}^N$ collects factor loadings, and $\sigma_0^2 > 0$ is idiosyncratic variance. Under (12), the efficient direction $w_* \propto \Sigma^{-1}\mu$ is fully diversified and assigns positive weights to all assets, with relative weights governed by factor exposures. In this benchmark, the economic role of diversification is clean: expected returns are entirely systematic, so the unconstrained optimum uses breadth to diversify residual risk while preserving exposure to the factor premium.

For any subset $A \subset [N]$, the best Sharpe ratio attainable using only assets in A admits a closed form. Writing β_A for the restriction of β to A,

$$\theta_A = \left(\frac{\mu_m^2 \beta_A^\top \beta_A}{\sigma_0^2 + \sigma_m^2 \beta_A^\top \beta_A} \right)^{1/2}, \qquad \theta^* = \left(\frac{\mu_m^2 \beta^\top \beta}{\sigma_0^2 + \sigma_m^2 \beta^\top \beta} \right)^{1/2}. \tag{13}$$

Thus, in a one-factor model, the loss from sparsity is governed by how much aggregate factor exposure $\beta_A^\top \beta_A$ the subset can retain and how quickly the residual variance diversifies as more assets are included. This isolates the opportunity-cost channel: limiting k matters insofar as it limits the portfolio's ability to keep systematic exposure while shedding idiosyncratic risk.

A natural benchmark for θ_k^* is obtained by choosing the k assets with the largest contribution to $\beta_A^\top \beta_A$. When factor loadings are not too heterogeneous, adding assets increases $\beta_A^\top \beta_A$ roughly linearly in k, while the idiosyncratic component of risk diversifies at the same rate. This implies that incremental efficiency gains from expanding the support diminish at rate $1/k$: once the portfolio has captured the systematic component, marginal improvements come primarily from finer residual-risk diversification.

Proposition 3 (Efficiency loss under a one-factor structure)**.** *Under* (12)*, if* $\liminf_{N\to\infty} \beta^\top \beta / N > 0$*, then*

$$\theta^* - \theta_k^* = O\left(\frac{1}{k}\right).$$

Proposition 3 has a clear economic interpretation. The condition $\liminf_{N\to\infty} \beta^\top \beta / N > 0$ is a pervasiveness requirement: average squared factor exposure does not vanish as the investment universe expands, so the systematic component remains economically material even in large markets. Under this condition, sparsity primarily limits the ability to diversify idiosyncratic risk, and the resulting opportunity cost declines quickly as k grows.

3.3.2 Approximate factor models and factor strength

The one-factor logic extends to a general approximate factor structure with pricing errors. We adopt the standard specification used in Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1993:

Assumption 1 (Approximate factor structure with pricing errors)**.**

$$\mu = \alpha + B\mu_f, \qquad \Sigma = B\Sigma_f B^\top + \Sigma_0, \tag{14}$$

where $B \in \mathbb{R}^{N\times L}$ *collects factor loadings,* $\mu_f \in \mathbb{R}^L$ *and* $\Sigma_f \in \mathbb{R}^{L\times L}$ *are factor premia and covariance,* $\Sigma_0 \in \mathbb{R}^{N\times N}$ *is idiosyncratic risk, and* $\alpha \in \mathbb{R}^N$ *captures pricing errors.*

The key economic question is how quickly a k-asset subset can replicate the relevant factor span while diversifying residual risk, and how much additional Sharpe ratio can be generated by exploiting pricing errors once trading is restricted to a sparse subset.

We also impose mild regularity on factor and idiosyncratic risk, which imply invertibility:

Assumption 2 (Bounded eigenvalues)**.** *There exist constants* $0 < \underline{c} < \overline{c} < \infty$*, independent of* N*, such that*

$$\underline{c} \leq \lambda_{\min}(\Sigma_0) \leq \lambda_{\max}(\Sigma_0) \leq \overline{c}, \qquad \underline{c} \leq \lambda_{\min}(\Sigma_f) \leq \lambda_{\max}(\Sigma_f) \leq \overline{c}. \tag{15}$$

To quantify the strength of the factor component relative to idiosyncratic risk, assume:

Assumption 3 (Factor strength)**.**

$$\frac{1}{N^{\zeta}} B^{\top} \Sigma_0^{-1} B \to \Sigma_B \qquad \text{as } N \to \infty, \tag{16}$$

for some $\zeta \in (0, 1]$ *and some positive definite matrix* Σ_B*.*

Finally, we restrict the economically relevant component of pricing errors:

Assumption 4 (Vanishing standardized pricing errors)**.** *The standardized aggregate magnitude of pricing errors vanishes at rate* $1/N$*:*

$$\|\Sigma_0^{-1}\alpha\|_1 = O(1/N)\,, \qquad \text{as } N \to \infty.$$

Assumption 1 allows expected returns to have a systematic component priced by a small number of factors and a residual component of pricing errors, while covariances decompose into factor risk plus idiosyncratic risk. Assumption 2 ensures idiosyncratic risk is well behaved and diversifiable and factor risk is nondegenerate. Assumption 3 captures how redundant the cross section is: with strong factors ($\zeta = 1$), systematic risk can be replicated with relatively few names, whereas weaker factors ($\zeta < 1$) require larger portfolios to span the same systematic opportunities. Finally, Assumption 4 is a deliberately strong restriction on pricing errors. It imposes that pricing errors are not economically scalable through idiosyncratic hedging: $\|\Sigma_0^{-1}\alpha\|_1 = O(1/N)$ rules out sequences of economies in which standardized alphas accumulate in the cross section. Under this strict condition, residual mispricing cannot overturn the spanning logic implied by factor redundancy, so the opportunity cost of sparsity remains governed by the factor component even when exact pricing fails.

Under these conditions, sparsity entails an opportunity cost because a k-asset subset may not fully span the systematic component of the mean–variance optimum. The next result quantifies how this cost declines with portfolio size and factor strength.

Theorem 2 (Efficiency loss under approximate factor models)**.** *Let Assumptions 1, 2, 3, and 4 hold, with factor-strength parameter* $\zeta \in (0, 1]$*. Then for any* $k \geq L$*,*

$$\theta^* - \theta_k^* = O\big(N^{1-\zeta}/k\big)\,, \qquad \text{as } N \to \infty.$$

In particular, when factors are strong ($\zeta = 1$)*,*

$$\theta^* - \theta_k^* = O(1/k)\,.$$

This declining opportunity cost corresponds to the "efficiency loss" curve in Figure 1, and aligns with the evidence that marginal diversification benefits become small beyond moderate portfolio sizes (Elton and Gruber, 1977; Statman, 1987).

Remark 4 (Magnitude in financial applications)**.** *A useful yardstick for the factor-driven component is the unitless index* $N^{1-\zeta}/k$*. Under strong factors* ($\zeta = 1$)*, it reduces to* $1/k$*, which is* 0.20 *for* $k = 5$*,* 0.10 *for* $k = 10$*, and* 0.02 *for* $k = 50$*. For weaker factors, the same index declines more slowly with* k*. For example, with* $N = 300$ *and* $\zeta = 1/2$*,* $N^{1-\zeta}/k \approx 17.3/k$*, yielding* 3.46 *(*$k = 5$*),* 1.73 *(*$k = 10$*), and* 0.35 *(*$k = 50$*). With* $N = 3000$ *and* $\zeta = 1/2$*,* $N^{1-\zeta}/k \approx 54.8/k$*, yielding* 11.0 *(*$k = 5$*),* 5.5 *(*$k = 10$*), and* 1.1 *(*$k = 50$*). Pricing errors contribute through* $\|\Sigma_0^{-1}\alpha\|_1$*, which is* $o(1)$ *under Assumption 4.*

The same caveat as in Remark 3 applies: constants hidden in the $O(\cdot)$ *terms depend on the distribution of loadings and the conditioning of* (Σ_f, Σ_0)*, so these indices should be viewed as comparative indicators across environments.*

3.4 Optimal Portfolio Complexity and Comparative Statics

We now combine the two sides of the portfolio complexity frontier and extract its economic implications. Efficiency loss captures the benefit of complexity: under approximate factor structure with pricing errors, the opportunity cost of restricting to k assets declines like $N^{1-\zeta}/k$ (Theorem 2). Estimation risk captures the statistical cost of complexity: Theorem 1 implies that the out-of-sample loss from learning a k–sparse rule from T observations grows like $\sqrt{k \log N/T}$. These forces move in opposite directions as k increases. For small k, the investor sacrifices too much investable span and forgoes systematic opportunities; for large k, the investor asks the data to estimate and combine too many inputs and incurs substantial overfitting. This tradeoff implies that realized Sharpe ratios are naturally expected to be hump-shaped in k, with an interior complexity level balancing diversification against estimation noise.

A concise asymptotic statement follows immediately.

Theorem 3 (Asymptotic efficiency)**.** *Under the assumptions of Theorem 1 and Theorem 2,*

$$\theta^* - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}} + \frac{N^{1-\zeta}}{k}\right), \qquad \text{as } T \to \infty.$$

In particular, if $N^{1-\zeta} \ll k \ll T/\log N$, then $\theta(\widehat{w}_k) \xrightarrow{p} \theta^*$.

Theorem 3 clarifies when sparse mean–variance rules retain essentially full efficiency in large markets. The lower bound $k \gg N^{1-\zeta}$ ensures that the selected subset can approximate the systematic span implied by the factor structure, so the opportunity cost of limiting complexity is small. The upper bound $k \ll T/\log N$ ensures that estimation noise remains controlled, so the statistical cost of complexity is small.

The result also delivers sharp comparative statics. Holding (N, ζ) fixed, a larger estimation window T relaxes the statistical upper bound on k and therefore supports higher portfolio complexity. Holding (T, ζ) fixed, a larger universe N increases the statistical cost through the $\log N$ search penalty, and it also changes redundancy through the factor-strength scaling $N^{1-\zeta}$. Factor strength, summarized by ζ, is the key economic primitive shifting the frontier. Under strong factors ($\zeta = 1$), the opportunity cost term is $O(1/k)$ and thus vanishes quickly, so the dominant constraint on complexity is statistical rather than economic. When factors are weaker ($\zeta < 1$), redundancy is lower and the opportunity cost declines more slowly through $N^{1-\zeta}/k$, so achieving near-efficiency requires larger k.

Remark 5 (Balancing rates)**.** *The bound in Theorem 3 suggests a benchmark scaling for the complexity level that balances the two leading terms:*

$$\sqrt{\frac{k \log N}{T}} \approx \frac{N^{1-\zeta}}{k} \qquad \Longrightarrow \qquad k \asymp \left(N^{1-\zeta}\sqrt{\frac{T}{\log N}}\right)^{2/3}.$$

This is not a prescription—constant factors depend on unknown features of (μ, Σ)*—but it makes explicit how the economically relevant portfolio complexity increases with available data and decreases with factor strength.*

3.5 Implications for ℓ_1 Regularization

An analogous economic tradeoff applies to the ℓ_1-regularized rule. Whereas the cardinality constraint makes portfolio complexity explicit through k, the ℓ_1 penalty implements complexity control in a convex and computationally tractable way: it shrinks weights and sets some positions exactly to zero, limiting the degrees of freedom that must be learned from finite data. At the same time, under the approximate factor structure in Assumption 1, the opportunity cost of excluding assets is limited because systematic risks can be spanned with relatively few holdings.

For ℓ_1 regularization we allow pricing errors to persist with N, but restrict their standardized aggregate magnitude. Relative to Assumption 4, which imposes a $1/k$-type restriction to obtain a $1/k$ efficiency-loss rate under strong factors, we only require that aggregate standardized mispricing does not diverge:

Assumption 5 (Bounded standardized pricing errors)**.** *The standardized aggregate magnitude of pricing errors remains bounded:*

$$\limsup_{N\to\infty} \|\Sigma_0^{-1}\alpha\|_1 < \infty.$$

This condition rules out sequences of economies in which standardized pricing errors grow without bound and, in turn, limits the scope for large cross-sectional arbitrage opportunities that would require increasingly aggressive idiosyncratic hedging. We also maintain the factor-strength and regularity conditions from Assumptions 2 and 3.

Theorem 4 (Efficiency of ℓ_1-regularized portfolios under approximate factor models)**.** *Suppose Assumptions 1, 2, 3, and 5 hold with factor-strength parameter* $\zeta \in (0, 1]$*. Then there exists a numerical constant* $C_0 > 0$ *such that, for any* $\lambda \geq C_0\sqrt{\log N/T}$*,*

$$\theta^* - \theta(\widehat{w}_{\ell_1,\lambda}) = O_p\big(N^{1-\zeta}\lambda\big)\,.$$

In particular, if $T \gg N^{2(1-\zeta)} \log N$ *and* $\lambda = C\sqrt{\log N/T}$ *for a sufficiently large constant* C*, then* $\theta(\widehat{w}_{\ell_1,\lambda}) \xrightarrow{p} \theta^*$*.*

Theorem 4 shows that ℓ_1 regularization can deliver asymptotic mean–variance efficiency under the same factor-strength primitive as in Theorem 2, while remaining robust to persistent pricing errors. Interpreted through the portfolio complexity frontier, λ plays the role of a continuous complexity control: it is chosen large enough to dominate uniform sampling noise, but not so large as to materially truncate exposure to systematic premia. When factors are strong ($\zeta = 1$), choosing λ of order $\sqrt{\log N/T}$ requires only $T \gg \log N$ for the gap $\theta^* - \theta(\widehat{w}_{\ell_1,\lambda})$ to vanish. For weaker factors ($\zeta < 1$), the required sample size is larger, reflecting that redundancy is lower and the opportunity cost of excluding assets declines more slowly with effective complexity.

The requirement $\lambda \geq C_0\sqrt{\log N/T}$ is the standard high-dimensional choice that makes the penalty dominate uniform sampling noise in estimated moments. In portfolio terms, it limits extreme positions driven by estimation error while still allowing factor-related signals to enter the solution. Larger values of λ preserve the convergence result but induce additional shrinkage and hence lower effective portfolio complexity.

The next section turns to implementation. We show how the ℓ_1-regularized rule in (3) provides a computationally tractable surrogate for cardinality constraints, and we discuss how λ maps into an effective sparsity level in empirically relevant investment universes.

4 Computational Implementation: ℓ_1 Relaxation

The cardinality constraint in (2) is useful for theory because it isolates the economic tradeoff between spanning and estimation noise. In practice, however, computing the exact solution quickly becomes infeasible once N is moderately large, so we implement sparse selection through the ℓ_1-regularized rule in (3).

4.1 Why Exact Subset Selection is Infeasible at Scale

Solving (2) exactly requires choosing an optimal support among all subsets $A \subset [N]$ with $|A| \leq k$ and then optimizing weights conditional on that support. The combinatorial component alone is prohibitive: the number of candidate supports is $\sum_{j=1}^{k} \binom{N}{j}$, which for economically relevant k grows essentially exponentially in N.

A concrete illustration makes the scale stark. With $N = 300$ and $k = 150$,

$$\binom{300}{150} \approx 9.4 \times 10^{88},$$

which already exceeds the often-quoted $\sim 10^{80}$ atoms in the observable universe. Even an idealized exhaustive search that could evaluate 10^9 candidate supports per second would require on the order of 10^{72} years. This is why exact k–sparse mean–variance optimization is intractable in large universes and why practical procedures must avoid discrete enumeration.

Exact methods do exist for moderate N, and they are useful both as benchmarks and as diagnostic tools. A leading example is a mixed-integer quadratic program (MIQP), which enforces sparsity by introducing binary inclusion variables and linking constraints that set $w_i = 0$ whenever an asset is excluded. Modern MIQP solvers can deliver high-quality solutions at moderate scale, but the underlying branch-and-bound search has worst-case complexity exponential in N. For large universes, MIQP is therefore typically run with time limits and prescribed optimality-gap tolerances, and should be viewed as a controlled approximation rather than a guaranteed global optimum; see Appendix C. Beyond MIQP, a growing literature develops scalable algorithms that interpolate between continuous relaxations and combinatorial search; see, e.g., Bertsimas and Cory-Wright (2022).

In our simulation and empirical analysis below, ℓ_1 regularization is the default implementation because it scales reliably to large universes and yields a full regularization path, while MIQP is used as a benchmark/robustness check when computation permits.

4.2 ℓ_1 Regularization as a Convex Relaxation

We therefore implement sparse portfolios through the convex relaxation (3), which replaces the discrete constraint $\|w\|_0 \leq k$ by a continuous penalty on gross exposure $\|w\|_1$. The appeal is not only computational. Economically, $\|w\|_1$ is leverage in a long–short book, and penalizing it directly limits the optimizer's ability to translate small errors in $(\widehat{\mu}, \widehat{\Sigma})$ into large offsetting positions.

The mechanism generating exact sparsity is transparent from the first-order optimality conditions. Let $\widehat{w}_{\ell_1,\lambda}$ solve (3). Since $\widehat{U}_T(\cdot)$ is concave and $\|\cdot\|_1$ is convex, the objective in (3) is concave, and $\widehat{w}_{\ell_1,\lambda}$ is characterized by the KKT conditions

$$\widehat{\mu} - \gamma\widehat{\Sigma}\widehat{w}_{\ell_1,\lambda} \in \lambda\partial\|\widehat{w}_{\ell_1,\lambda}\|_1,$$

where $\partial\|w\|_1$ denotes the subdifferential of the ℓ_1 norm, i.e.,

$$\partial\|w\|_1 := \left\{u \in \mathbb{R}^N : \; u_i = \mathrm{sign}(w_i) \text{ if } w_i \neq 0, \text{ and } u_i \in [-1,1] \text{ if } w_i = 0\right\},$$

and sign : $\mathbb{R} \to \{-1, 0, 1\}$ is defined by

$$\mathrm{sign}(x) := \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

Therefore, for each asset i,

$$\begin{aligned} \widehat{w}_{\ell_1,\lambda,i} \neq 0 \quad &\Rightarrow \quad \left(\widehat{\mu} - \gamma\widehat{\Sigma}\widehat{w}_{\ell_1,\lambda}\right)_i = \lambda\mathrm{sign}(\widehat{w}_{\ell_1,\lambda,i}), \\ \widehat{w}_{\ell_1,\lambda,i} = 0 \quad &\Rightarrow \quad \left(\widehat{\mu} - \gamma\widehat{\Sigma}\widehat{w}_{\ell_1,\lambda}\right)_i \in [-\lambda, \lambda]. \end{aligned}$$

Thus, an asset is excluded when its marginal contribution to the mean–variance objective, net of covariance interactions with the rest of the portfolio, falls below the penalty threshold. The kink of $|w_i|$ at zero is what produces exact zeros, rather than merely small weights.

Moreover, ℓ_1 is a sharp convex proxy for cardinality on economically relevant domains. Under natural position limits (e.g., $\|w\|_\infty \leq 1$), the convex hull of k–sparse portfolios coincides with an ℓ_1 ball,[3]

$$\operatorname{conv}\{w : \|w\|_0 \le k,\ \|w\|_\infty \le 1\} = \{w : \|w\|_1 \le k,\ \|w\|_\infty \le 1\},$$

so $\|w\|_1$ provides the tightest convex relaxation of sparsity once scale is controlled; see Argyriou et al. (2012).

4.3 Computation of the Solution Path

Problem (3) is a concave quadratic program with an ℓ_1 penalty, and can be rewritten in a standard lasso form; see, e.g., Li (2015). Up to an additive constant, maximizing (3) is equivalent to minimizing

$$\frac{1}{2} w^\top \widehat{M} w - \widehat{\mu}^\top w + \lambda \|w\|_1, \qquad \widehat{M} := \gamma \widehat{\Sigma}.$$

When $\widehat{\Sigma}$ is singular (e.g., $N \ge T$), it can be replaced by a stabilized version $\widehat{\Sigma}_\varepsilon := \widehat{\Sigma} + \varepsilon I_N$ for a small $\varepsilon > 0$ (equivalently, an ℓ_2 stabilization), which makes $\widehat{M}$ positive definite and does not affect the economic content of the sparse regularization.

Let $\widehat{M} = C^\top C$ be a Cholesky factorization. Then

$$\frac{1}{2} w^\top \widehat{M} w - \widehat{\mu}^\top w = \frac{1}{2}\|Cw - y\|_2^2 - \frac{1}{2}\|y\|_2^2, \qquad y := C^{-\top}\widehat{\mu},$$

so (3) is equivalent to the lasso problem

$$\min_{w \in \mathbb{R}^N} \frac{1}{2}\|Cw - y\|_2^2 + \lambda \|w\|_1.$$

This representation enables the use of mature high-dimensional solvers (coordinate descent, proximal-gradient methods, and homotopy/LARS-type path algorithms) that compute either a single solution at a given λ or the entire regularization path over a grid of λ values. Warm starts make the path computation particularly efficient: the solution at λ_{m+1} is close to the solution at λ_m. Implementation details are found in Appendix C.

[3] For a set $S \subset \mathbb{R}^N$ the convex hull is the smallest convex set that contains all its points:

$$\operatorname{conv}(S) := \left\{ \sum_{j=1}^m \pi_j x_j :\ m \ge 1,\ x_j \in S,\ \pi_j \ge 0,\ \sum_{j=1}^m \pi_j = 1 \right\}.$$

4.4 From λ to an Implementable Sparsity Level

In empirical work, it is often convenient to interpret λ through the induced support size $\widehat{k}(\lambda) := \|\widehat{w}_{\ell_1,\lambda}\|_0$ and to compare ℓ_1 portfolios to the k–sparse frontier by matching $\widehat{k}(\lambda)$ to a target k. Along the lasso path, $\widehat{k}(\lambda)$ is typically nonincreasing in λ, though it need not decrease by one at a time and ties can occur when multiple assets enter or exit together.

Because the Sharpe ratio is scale invariant, the optimization delivers a direction that can be normalized after estimation to satisfy the implementation convention of interest (e.g., a unit budget constraint $1^\top w = 1$, a target volatility, or a target gross exposure $\|w\|_1 \leq \tau$). In our empirical analysis we therefore compute $\widehat{w}_{\ell_1,\lambda}$ from (3) and then rescale it to the desired trading scale without changing its population Sharpe ratio ranking.

Finally, the theory suggests a natural order for the penalty level: taking λ on the order $\sqrt{\log N/T}$ is sufficient to dominate coordinatewise sampling noise in high dimensions and prevents the optimizer from chasing spurious sample Sharpe opportunities. In practice, we implement this by evaluating a grid of λ values that spans from the smallest penalty yielding an essentially dense solution to the largest penalty yielding the zero portfolio, and we select a point on the path by targeting a desired $\widehat{k}(\lambda)$.

5 Simulation Analysis

This section provides a controlled numerical illustration of the tradeoff highlighted by our theory: sparsity reduces estimation risk (Theorem 1) but may entail an opportunity cost from excluding assets (Theorem 2). We simulate returns from a Fama–French three-factor structure calibrated to U.S. equity markets. This calibration corresponds to the strong-factor case in our approximate factor framework (i.e., $\zeta = 1$), under which the efficiency loss of restricting to k assets decays at the rate $O(1/k)$, while estimation risk scales as $O\big(\sqrt{k \log N/T}\big)$.

In this simulation DGP, $\theta^* = \sqrt{\mu^\top \Sigma^{-1} \mu}$ is available in closed form, and for any fixed subset A, the population Sharpe ratio $\theta_A = \sqrt{\mu_A^\top \Sigma_A^{-1} \mu_A}$ is explicit. However, the population-optimal k–sparse Sharpe ratio $\theta_k^* = \max_{|A|=k} \theta_A$ is the solution to a combinatorial maximization over $\binom{N}{k}$ subsets and does not admit a tractable closed form under heterogeneous (α_i, β_i). Likewise, the estimation loss of the empirically selected optimizer involves the joint distribution of the maximizer over subsets and is not analytically available beyond bounds. Accordingly, we quantify the efficiency cost of sparsity numerically and retain Monte Carlo evaluation for estimation losses.

5.1 Data Generating Process

We work with a three-factor structure in the spirit of the Fama–French model. For each asset $i = 1, \ldots, N$ and month $t = 1, \ldots, T$, excess returns are generated according to

$$r_{it} = \alpha_i + \beta_i^\top f_t + \varepsilon_{it},$$

where $f_t = (f_{M,t}, f_{S,t}, f_{V,t})^\top$ collects the market (MKT), size (SMB), and value (HML) factors, $\beta_i = (\beta_{iM}, \beta_{iS}, \beta_{iV})^\top$ is the vector of factor loadings for asset i, and ε_{it} is idiosyncratic noise. This induces the approximate factor model

$$\mu = \alpha + B\mu_f, \qquad \Sigma = B\Sigma_f B^\top + \sigma_\varepsilon^2 I_N,$$

where μ and Σ are the population mean and covariance of $r_t = (r_{1t}, \ldots, r_{Nt})^\top$, $\alpha = (\alpha_1, \ldots, \alpha_N)^\top$, and B is the $N \times 3$ loading matrix with ith row $\beta_i^\top$.

We simulate $\{f_t\}_{t=1}^T$ at the monthly frequency as i.i.d. Gaussian:

$$f_t = \mu_f + u_t, \qquad u_t \sim \mathcal{N}(0, \Sigma_f),$$

independently over t. The factor premia and volatilities are calibrated to reflect typical U.S. equity market characteristics,

$$\mu_f = (0.08, 0.03, 0.04)^\top/12, \qquad \Sigma_f = \operatorname{diag}\left(\left(\frac{0.16}{\sqrt{12}}\right)^2, \left(\frac{0.14}{\sqrt{12}}\right)^2, \left(\frac{0.14}{\sqrt{12}}\right)^2\right).$$

Idiosyncratic shocks are homoskedastic Gaussian,

$$\varepsilon_{it} = \sigma_\varepsilon \eta_{it}, \qquad \eta_{it} \sim \mathcal{N}(0, 1), \qquad \sigma_\varepsilon = \frac{0.20}{\sqrt{12}},$$

independently across i and t.

For each Monte Carlo replication and each cross-sectional size N, we draw intercepts and factor loadings once and keep them fixed over time. The intercepts are

$$\alpha_i = \frac{0.02}{12} + \frac{0.01}{\sqrt{12}} Z_i^{(\alpha)}, \qquad Z_i^{(\alpha)} \sim \mathcal{N}(0,1),$$

and loadings are generated as

$$\beta_{ij} = \frac{\beta_{\text{mean},j}}{12} + \frac{0.3}{\sqrt{12}} Z_{ij}^{(\beta)}, \qquad Z_{ij}^{(\beta)} \sim \mathcal{N}(0,1), \quad j = 1,2,3,$$

with mean exposure vector $\beta_{\text{mean}} = (1,0,0)^\top$.

5.2 Simulation Design and Portfolio Construction

We work with an investment universe of $N = 100$ assets and sample sizes $T \in \{120, 240, 480, 960\}$ (monthly observations). For each (N,T) pair we run 10,000 independent Monte Carlo replications. In each replication, we simulate $\{r_t\}_{t=1}^T$ from the data-generating process in Section 5.1 and form the plug-in estimators $(\widehat{\mu}, \widehat{\Sigma})$ as in (1). These are the inputs a mean–variance investor would construct from a historical window before choosing portfolio weights.

Our goal is to quantify the tradeoff between the efficiency cost of sparsity and the finite-sample cost of estimating moments. For each cardinality level $k \in [N]$ we benchmark the sample portfolio $\widehat{w}_k$ against the unrestricted population optimum $\theta^* = \max_{w \in \mathbb{R}^N} \theta(w)$ via

$$\theta^* - \theta(\widehat{w}_k) = \underbrace{\left[\theta^* - \theta_k^*\right]}_{\text{Efficiency loss}} + \underbrace{\left[\theta_k^* - \theta(\widehat{w}_k)\right]}_{\text{Estimation loss}}. \tag{17}$$

The first term is the opportunity cost of imposing $\|w\|_0 \le k$ even under perfect knowledge of (μ, Σ); the second term captures the incremental performance loss induced by using estimated moments.

Population-efficient k–sparse Sharpe ratio and efficiency loss. We compute the population benchmark at sparsity level k,

$$\theta_k^* := \max_{\|w\|_0 \le k} \theta(w), \qquad \theta(w) = \frac{w^\top \mu}{\sqrt{w^\top \Sigma w}},$$

as in (5). In each replication, θ_k^* is computed from the true moments (μ, Σ) using the same numerical approach used for the sample problem. This delivers the replication-specific efficiency-loss curve $\Delta_k := \theta^* - \theta_k^*$.

Sample k–sparse portfolio and estimation loss. Given $(\widehat{\mu}, \widehat{\Sigma})$, we solve the sample cardinality-constrained problem (2) and obtain the optimizer $\widehat{w}_k$. Its out-of-sample performance is summarized by the population Sharpe ratio $\theta(\widehat{w}_k)$ computed under the true (μ, Σ). The estimation loss is therefore $\theta_k^* - \theta(\widehat{w}_k)$.

To disentangle how estimation loss arises, we evaluate the decomposition in (10),

$$\underbrace{\theta_k^* - \theta(\widehat{w}_k)}_{\text{Estimation loss}} = \underbrace{\left[\theta_k^* - \theta_{\widehat{A}_k}\right]}_{\text{Selection loss}} + \underbrace{\left[\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\right]}_{\text{Allocation loss}}.$$

The two components correspond to subset selection error and weight-estimation error on the selected subset.

Oracle Sharpe ratio on the selected support. Let $\widehat{A}_k := \{i \in [N] : \widehat{w}_{k,i} \neq 0\}$ denote the support selected by the sample procedure, using (7). To remove weight-estimation error while keeping the selected subset fixed, we compute

$$\theta_{\widehat{A}_k} := \max_{\operatorname{supp}(w) \subset \widehat{A}_k} \theta(w),$$

as in (9). Because $\theta_{\widehat{A}_k}$ is defined in terms of the true moments (μ, Σ), it isolates the contribution of selection to estimation loss.

In all cases, performance is measured by the population Sharpe ratio $\theta(\cdot)$ computed under the true (μ, Σ), and we report Monte Carlo averages across the 10,000 replications for each (T, k).

Because the exact cardinality problem is NP-hard, we compute $\widehat{w}_k$ using the numerical approximations described in Appendix C.

5.3 Simulation Results

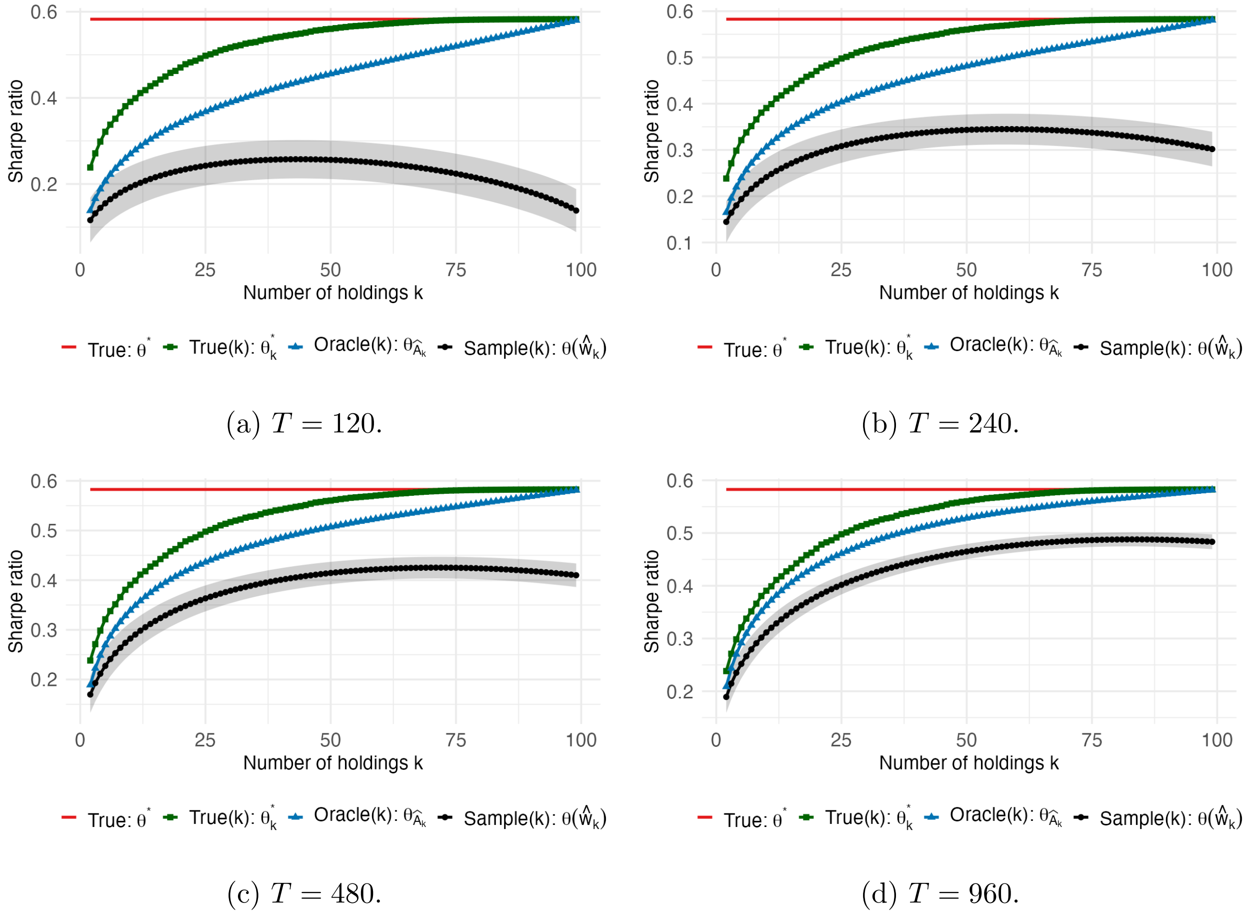
We illustrate the estimation–efficiency tradeoff in a controlled environment using Monte Carlo experiments under the Fama–French three-factor calibration. The simulations trace, as a function of the cardinality constraint k, the population-efficient frontier θ_k^*, the population performance of the corresponding sample optimizer $\theta(\widehat{w}_k)$, and the intermediate oracle benchmark $\theta_{\widehat{A}_k}$ that holds the selected support fixed. This structure allows us to quantify not only the total Sharpe-ratio loss $\theta^* - \theta(\widehat{w}_k)$, but also its decomposition into efficiency versus estimation components and, within estimation, selection versus allocation channels. We vary the in-sample length T, holding N fixed, and report averages across Monte Carlo replications to isolate the systematic dependence of out-of-sample performance on portfolio complexity.

5.3.1 Population Sharpe ratio profiles

Figure 3 summarizes the population Sharpe ratio profiles implied by our Fama–French three-factor calibration. Each panel reports four benchmarks as a function of the cardinality level $k \in [N]$: the unrestricted population optimum θ^*, the population-efficient k–sparse frontier θ_k^*, the oracle Sharpe ratio on the sample-selected support $\theta_{\widehat{A}_k}$, and the out-of-sample Sharpe ratio of the sample optimizer $\theta(\widehat{w}_k)$. Two features are immediate.

First, θ^* is constant in k by definition: it is the Sharpe ratio attainable when the investor may use all N assets. In contrast, both θ_k^* and $\theta_{\widehat{A}_k}$ are increasing in k and exhibit diminishing marginal gains, until they match θ^* at $k = N$. This reflects a simple economic mechanism: enlarging the admissible set of holdings can only expand the span of systematic risks that can be exploited and diversified, but in a strong-factor environment the incremental benefit of adding assets saturates quickly. For any $k < N$, $\theta_k^* \geq \theta_{\widehat{A}_k}$ because the oracle is restricted to the particular subset selected by the sample procedure. As T increases, $\theta_{\widehat{A}_k}$ shifts upward toward θ_k^*, consistent with the probability of selecting a near-efficient subset increasing in larger samples.

Second, only the sample curve $\theta(\widehat{w}_k)$ is non-monotone. For small k, allowing additional positions increases $\theta(\widehat{w}_k)$ because the efficiency gain from diversification dominates sampling noise. As k grows large relative to T, the optimizer becomes increasingly sensitive to estimation error in $(\widehat{\mu}, \widehat{\Sigma})$, and the population performance of plug-in weights deteriorates. This hump-shaped pattern and its dependence on T are precisely the comparative statics predicted by Theorem 1.



(a) $T = 120$.

(b) $T = 240$.

(c) $T = 480$.

(d) $T = 960$.

Figure 3: **Sharpe ratio profiles of sparse portfolios.** Population Sharpe ratios as a function of the cardinality k for $N = 100$ assets under the Fama–French 3-factor calibration, for each in-sample length T. We have, from highest to lowest, the unrestricted true optimum θ^*, the k–sparse true frontier θ^*_k, the average oracle frontier $\theta_{\widehat{A}_k}$ under the optimal sample selection, and the average frontier $\theta(\widehat{w}_k)$ of the sample optimal k–sparse portfolio. For the latter frontier, the gray ribbon shows ±1 s.e. around its average. Results are based on 10,000 Monte Carlo replications.

5.3.2 Estimation–efficiency losses

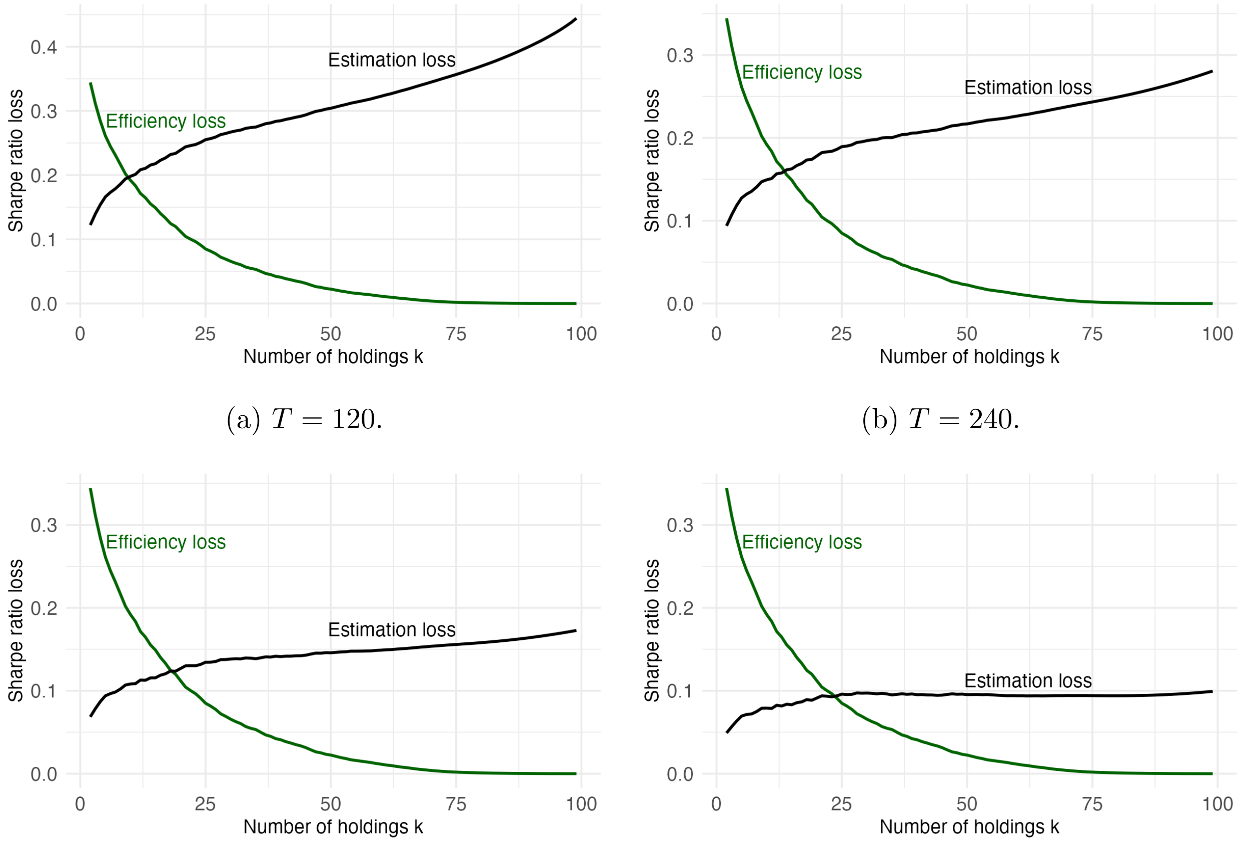
Figure 4 makes the underlying tradeoff explicit by plotting the two components of the total loss decomposition (17). The efficiency loss $\theta^* - \theta_k^*$ is a population object and therefore does not depend on T. Under approximate factor structure, Theorem 2 predicts that this opportunity cost declines at the rate $N^{1-\zeta}/k$, and in particular at rate $1/k$ under strong factors ($\zeta = 1$). Consistent with this prediction, the efficiency-loss curve falls rapidly in k and becomes nearly flat for moderate portfolio sizes, reflecting that a small number of holdings is sufficient to span the relevant systematic risks.

In contrast, the estimation loss $\theta_k^* - \theta(\widehat{w}_k)$ increases with k because higher-dimensional optimization amplifies moment-estimation noise. Theorem 1 implies the rate $\theta_k^* - \theta(\widehat{w}_k) = O_p\big(\sqrt{k \log N/T}\big)$, so, holding k fixed, the curve shifts downward with T at the usual $T^{-1/2}$ speed, while holding T fixed, it grows like $\sqrt{k}$ (up to a $\log N$ factor). The intersection of a T-invariant efficiency loss of order $1/k$ and a T-dependent estimation loss of order $\sqrt{k \log N/T}$ rationalizes the rightward movement of the ex-post optimal k^* as T increases, matching the movement of the hump in Figure 3. This is the finite-sample manifestation of the complexity frontier summarized by Theorem 3, which combines both forces through the bound $\theta^* - \theta(\widehat{w}_k) = O_p\big(\sqrt{k \log N/T} + N^{1-\zeta}/k\big)$.

5.3.3 Allocation–selection decomposition

Figure 5 decomposes the estimation loss into selection and allocation components (10). Our theory highlights a key asymmetry: selection risk is an extremum-estimation problem over $\binom{N}{k}$ candidate supports, so Proposition 1 yields the complexity term $O_p\Big(\sqrt{\log\binom{N}{k}/T}\Big)$, whereas Proposition 2 gives $O_p\big(\sqrt{k \log N/T}\big)$ for weight estimation on the selected support.

This distinction explains why the selection component is not monotone in k. For small k, $\log\binom{N}{k}$ grows quickly (roughly like $k \log(N/k)$), so the statistical difficulty of ranking supports increases as the constraint is relaxed: the optimizer searches over a rapidly expanding class of candidate holdings, making it more likely that some subset looks spuriously attractive in sample. As k becomes large, however, selection becomes progressively less consequential:



(a) $T = 120$.

(b) $T = 240$.

(c) $T = 480$.

(d) $T = 960$.

Figure 4: **Efficiency and estimation losses under sparsity.** Decomposition of the average total loss $\theta^* - \theta(\widehat{w}_k)$ into the efficiency loss $\theta^* - \theta^*_k$ and the estimation loss $\theta^*_k - \theta(\widehat{w}_k)$ as a function of the cardinality k, for $N = 100$ assets under the Fama–French 3-factor calibration and for each sample size T. The estimation loss is averaged over 10,000 Monte Carlo replications.

(i) many size-k supports overlap heavily, so even an imperfect maximizer $\widehat{A}_k$ typically shares most holdings with a population-efficient subset, shrinking $\theta^*_k - \theta_{\widehat{A}_k}$; (ii) under factor structure the population frontier θ^*_k is relatively flat at high k (Theorem 2), so perturbations of the selected set translate into small Sharpe-ratio differences; and (iii) as $k \to N$ the constraint becomes slack and selection is mechanically irrelevant. Taken together, these forces generate a hump-shaped selection loss, rising at low-to-moderate k as the search space expands, then falling for large k as the opportunity set becomes redundant and the constraint fades.

In contrast, allocation loss increases with k because it is driven by the instability of plug-in mean–variance weights in higher dimension: estimating and inverting larger moment submatrices and combining more assets amplifies sampling error. Overall, the figure illustrates that selection is the dominant channel at low sparsity levels, while allocation becomes the primary channel at higher k.

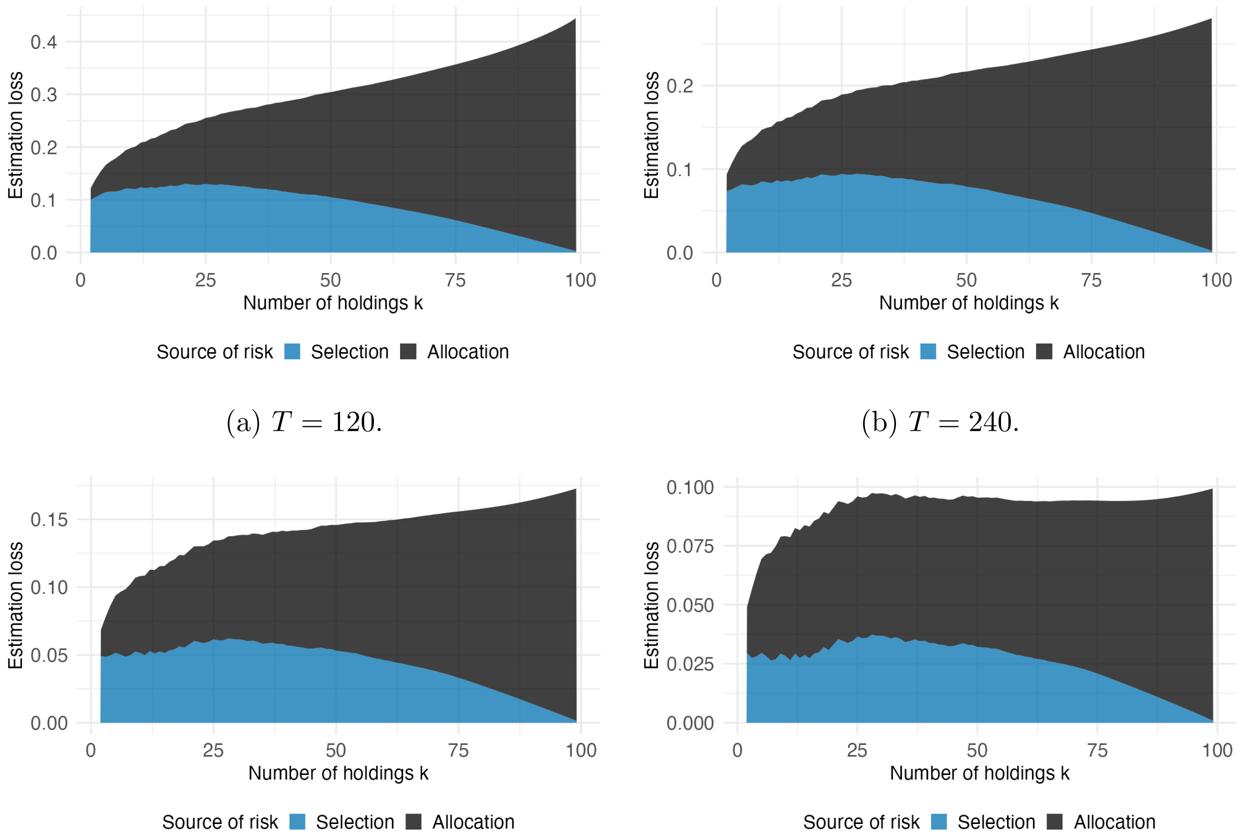
5.3.4 Optimal portfolio sparsity and feasible tuning rules

In each Monte Carlo replication, we can define the ex-post optimal sparsity level

$$k^* \in \underset{k\in[N]}{\operatorname{argmax}} \; \theta(\widehat{w}_k),$$

namely the number of holdings that maximizes the population performance of the sample optimal k–sparse portfolio.

Figure 6 reports the distribution of k^* across replications for different in-sample window lengths $T \in \{120, 240, 480, 960\}$ under the Fama–French three-factor calibration (with $N = 100$). The boxplots shift monotonically to the right as T increases: the typical optimal sparsity rises from about 44 holdings at $T = 120$ (with the middle 50% of replications spanning roughly 34 to 54) to about 58 holdings at $T = 240$ (interquartile range roughly 49 to 66), to about 72 holdings at $T = 480$ (interquartile range roughly 65 to 78), and to about 83 holdings at $T = 960$ (interquartile range roughly 78 to 88). Dispersion falls markedly with sample size: the standard deviation declines from about 14.6 at $T = 120$ to about 7.7 at $T = 960$, consistent with a sharper concentration of k^* on higher-cardinality portfolios as estimation error subsides.



(a) $T = 120$.

(b) $T = 240$.

(c) $T = 480$.

(d) $T = 960$.

Figure 5: **Decomposition of estimation loss.** Decomposition of the average estimation loss $\theta_k^* - \theta(\widehat{w}_k)$ into the selection loss $\theta_k^* - \theta_{\widehat{A}_k}$ and the allocation loss $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)$ as a function of the cardinality k, for $N = 100$ assets under the Fama–French 3-factor calibration and for each sample size T. Curves are averaged over 10,000 Monte Carlo replications.

These comparative statics align with the rate balance in Remark 5. The estimation term grows like $\sqrt{k \log N/T}$ while the efficiency term decays like $N^{1-\zeta}/k$, so equating them implies

$$k \asymp \left(N^{1-\zeta} \sqrt{\frac{T}{\log N}} \right)^{2/3} \propto T^{1/3} \quad \text{(for fixed } N, \zeta).$$

Thus k^* should increase with T, matching the rightward shift of the boxplots in Figure 6; the reduced dispersion reflects lower sampling noise at longer windows.

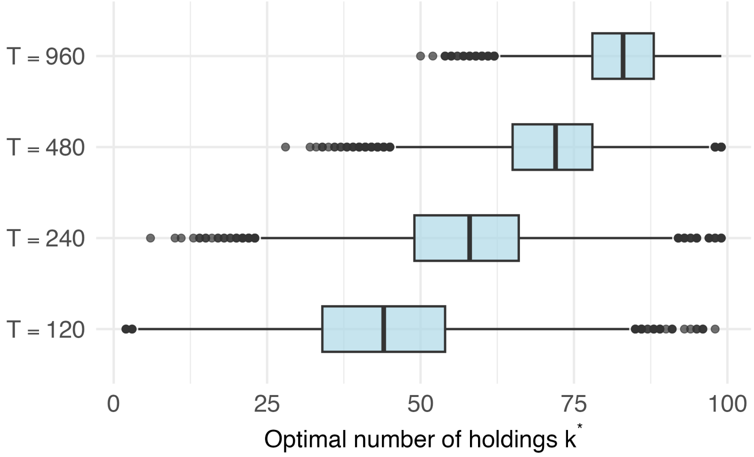


Figure 6: **Ex-post optimal sparsity k^*.** Boxplots across Monte Carlo replications under the Fama–French 3-factor calibration with $N = 100$, for each sample size T.

While informative, k^* is an *infeasible* benchmark: it depends on the population Sharpe ratio of the sample optimizer, which is not observable to the investor. We therefore propose and assess simple, fully data-driven tuning rules that select k without access to $\theta(\widehat{w}_k)$, while still delivering a population Sharpe ratio close to the best-in-hindsight value $\max_k \theta(\widehat{w}_k)$.

We consider two feasible procedures.

- *Leave-one-out (LOO).* The LOO rule selects k by constructing, for each candidate cardinality, a pseudo out-of-sample return series that mimics real-time implementation. Each period's portfolio is estimated without using that period's return, and the resulting weight vector is then applied to the omitted observation. Repeating this procedure for every $t \in [T]$ yields a sequence of LOO portfolio realizations for each k, whose

Sharpe ratio serves as a feasible score. The selected $\widehat{k}_{\text{LOO}}$ maximizes this score over the grid.

Formally, fix $k \in [N]$. For each $t \in [T]$, let $(\widehat{\mu}^{(-t)}, \widehat{\Sigma}^{(-t)})$ denote the sample moments computed from the leave-one-out sample $\{r_s : s \in [T] \setminus \{t\}\}$, and let $\widehat{w}_k^{(-t)}$ be the corresponding k–sparse optimizer (computed using the same estimation routine as in the baseline). Define the LOO portfolio return at time t as

$$\widehat{r}_t^{\text{LOO}}(k) := r_t^\top \widehat{w}_k^{(-t)}, \qquad t = 1, \ldots, T,$$

and collect these pseudo out-of-sample realizations in

$$\widehat{r}^{\text{LOO}}(k) := (\widehat{r}_1^{\text{LOO}}(k), \ldots, \widehat{r}_T^{\text{LOO}}(k))^\top.$$

We evaluate each k by the Sharpe ratio of the LOO return series,

$$\widehat{\theta}_{\text{LOO}}(k) := \frac{\overline{r}^{\text{LOO}}(k)}{\sqrt{\widehat{v}^{\text{LOO}}(k)}}, \quad \overline{r}^{\text{LOO}}(k) := \frac{1}{T}\sum_{t=1}^{T} \widehat{r}_t^{\text{LOO}}(k),$$
$$\widehat{v}^{\text{LOO}}(k) := \frac{1}{T}\sum_{t=1}^{T}\Big(\widehat{r}_t^{\text{LOO}}(k) - \overline{r}^{\text{LOO}}(k)\Big)^2,$$

and select

$$\widehat{k}_{\text{LOO}} \in \operatorname*{argmax}_{k \in [N]} \ \widehat{\theta}_{\text{LOO}}(k).$$

By construction, $\widehat{\theta}_{\text{LOO}}(k)$ evaluates each candidate sparsity level using weights estimated without observing the period-t return used for evaluation.

- *Unbiased estimator (UE).* For each k, let $\widehat{A}_k$ be the support selected by the sample k–sparse procedure and compute the bias-adjusted squared Sharpe statistic $\tilde{\theta}^2_{\widehat{A}_k}$ from Remark 1.[4] We then select

$$\widehat{k}_{\text{UE}} \in \operatorname*{argmax}_{k \in [N]} \ \tilde{\theta}^2_{\widehat{A}_k}.$$

In both cases, after choosing $\widehat{k}$ we refit the sample optimizer at that sparsity level and evaluate its population performance $\theta(\widehat{w}_{\widehat{k}})$ in the Monte Carlo design.

[4] The UE rule is not "unbiased" for the ex-post optimal sparsity level k^*. Rather, it relies on an unbiased (or bias-adjusted) estimator of the squared Sharpe ratio conditional on a fixed selected set A, and applies it to the random support $\widehat{A}_k$ produced by the data-driven selection step. As a result, UE corrects the in-sample objective for within-support finite-sample bias, but it does not account for selection risk.

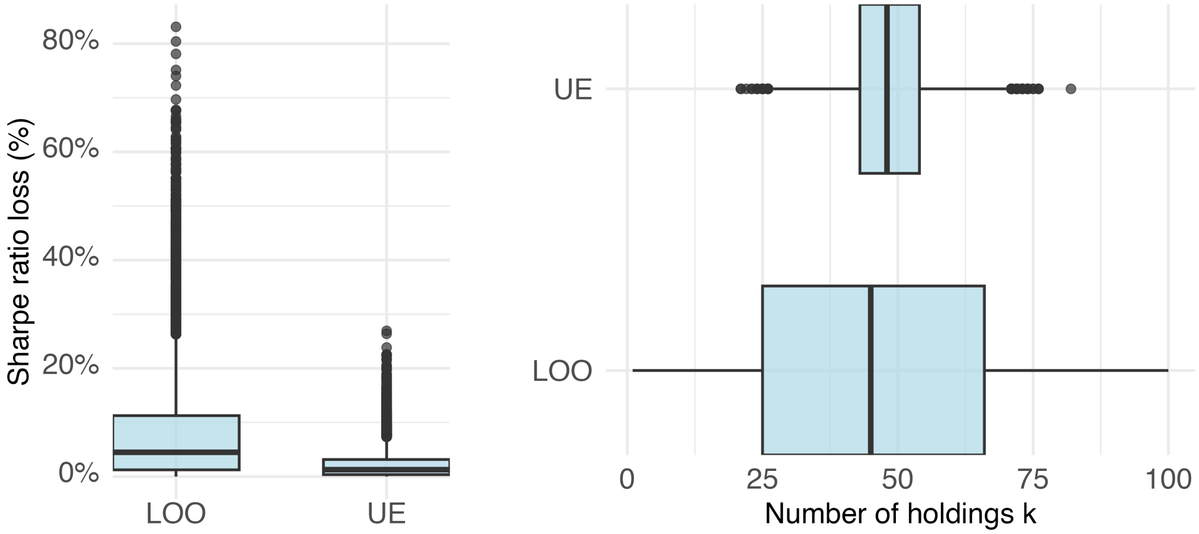
Figure 7 reports the distribution across 10,000 replications (for $N = 100$ and $T = 240$) of the percentage population Sharpe ratio loss relative to the infeasible benchmark,

$$\frac{\max_k \theta(\widehat{w}_k) - \theta(\widehat{w}_{\widehat{k}})}{\max_k \theta(\widehat{w}_k)},$$

together with the distribution of selected cardinalities $\widehat{k}$. Both feasible rules deliver small relative losses, but UE is markedly tighter and closer to the benchmark: the average percentage loss is 2.28% (median 1.28%) under UE versus 8.15% (median 4.48%) under LOO. The boxplots also reveal a pronounced reduction in dispersion under UE: its interquartile range is $[0.36\%, 3.15\%]$, compared to $[1.24\%, 11.26\%]$ under LOO. The right panel shows that both rules favor interior sparsity levels, but UE concentrates sharply on the middle of the k–grid (median 48, interquartile range $[43, 54]$), whereas LOO exhibits substantial dispersion (median 45, interquartile range $[25, 66]$). Taken together, these results indicate that simple bias correction of the in-sample Sharpe objective can provide a stable, fully feasible tuning rule that closely tracks the population-optimal sparsity level in this strong-factor calibration.

6 Empirical Analysis

This section studies whether the empirical performance of sparse portfolios reflects the trade-off formalized by our theory. Theorem 1 predicts that, for a given sample size, increasing the number of active positions amplifies estimation risk, while Theorem 2 predicts that the opportunity cost of excluding assets declines rapidly as k increases under a factor structure. Taken together, these results suggest an interior optimal sparsity level: as k rises from very small values, out-of-sample Sharpe ratios should initially improve as diversification opportunities expand, but eventually deteriorate as estimation error dominates. We examine this prediction by estimating k–sparse portfolios over rolling windows and evaluating their realized out-of-sample Sharpe ratios as a function of k.



(a) Percentage population Sharpe ratio loss relative to $\max_k \theta(\widehat{w}_k)$.

(b) Selected cardinality $\widehat{k}$.

Figure 7: **Feasible tuning rules for sparsity.** Boxplots across 10,000 Monte Carlo replications under the Fama–French 3-factor calibration with $N = 100$ and $T = 240$. The left panel reports the percentage population Sharpe ratio loss $\big(\max_k \theta(\widehat{w}_k) - \theta(\widehat{w}_{\widehat{k}})\big) / \max_k \theta(\widehat{w}_k)$, where $\widehat{k}$ is selected by leave-one-out cross-validation (LOO) or by maximizing the unbiased squared Sharpe statistic (UE). The right panel reports the distribution of selected portfolio cardinalities $\widehat{k}$ for the same two rules.

6.1 Estimation and Evaluation

Let $\{r_t\}_{t=1}^{T_0}$ denote a time series of N managed-portfolio excess returns, with $r_t \in \mathbb{R}^N$. For each dataset, we consider a set of in-sample window lengths $\mathcal{T}$ and write $T_{\max} := \max \mathcal{T}$. To maintain comparability across window lengths, all out-of-sample summaries reported below are computed over a common evaluation period of length $T_0 - T_{\max}$.

Fix $T \in \mathcal{T}$. For each rebalancing date $t \in \{T, \dots, T_0 - 1\}$, we compute the plug-in moments

$$\widehat{\mu}_{T,t} := \frac{1}{T} \sum_{s=t-T+1}^{t} r_s, \qquad \widehat{\Sigma}_{T,t} := \frac{1}{T} \sum_{s=t-T+1}^{t} r_s r_s^\top - \widehat{\mu}_{T,t} \widehat{\mu}_{T,t}^\top.$$

Given $(\widehat{\mu}_{T,t}, \widehat{\Sigma}_{T,t})$, we estimate a k–sparse portfolio $\widehat{w}_{T,k,t}$ by solving the ℓ_1-relaxation (3) of the cardinality-constrained problem (2) with the window-(T,t) moments. By construction, $\widehat{w}_{T,k,t}$ is formed at the end of period t and held over period $t+1$. As a robustness check, we also consider a standard gross-exposure normalization in which the estimated weights are rescaled to satisfy $\|\widehat{w}_{T,k,t}\|_1 \leq 1$. This constraint caps the overall leverage of the long–short positions and facilitates comparisons of portfolio magnitudes across window lengths T and rebalancing dates.

6.1.1 Out-of-sample Sharpe ratio

We evaluate performance out-of-sample at horizon $h = 1$ (rebalancing at each period). To maintain comparability across in-sample window lengths, we compute all out-of-sample summaries over a common evaluation period. Let $\mathcal{T}$ denote the set of in-sample window lengths considered for a given dataset and define $T_{\max} := \max \mathcal{T}$. We evaluate each (T,k) strategy over rebalancing dates $t = T_{\max}, \dots, T_0 - 1$, so that the out-of-sample length is the same $T_0 - T_{\max}$ for every $T \in \mathcal{T}$.[5] This ensures that differences across T reflect in-sample estimation error rather than differences in out-of-sample evaluation noise.

[5] For example, Dataset 1 is monthly from October 1971 through October 2025, and we use $\mathcal{T} = \{240, 360, 480\}$, so all reported out-of-sample quantities are computed over the last $T_0 - 480$ months.

The realized out-of-sample portfolio return associated with (T, k) at date $t+1$ is

$$r_{t+1}^{\mathrm{oos}}(T, k) := \widehat{w}_{T,k,t}^{\top} r_{t+1}, \qquad t = T_{\max}, \ldots, T_0 - 1.$$

Let $\{r_{t+1}^{\mathrm{oos}}(T, k)\}_{t=T_{\max}}^{T_0-1}$ denote the resulting out-of-sample return sequence. We summarize its risk-adjusted performance using the realized Sharpe ratio

$$\widehat{\theta}^{\mathrm{oos}}(T, k) := \frac{\overline{r}^{\mathrm{oos}}(T, k)}{s^{\mathrm{oos}}(T, k)},$$

where

$$\overline{r}^{\mathrm{oos}}(T, k) := \frac{1}{T_0 - T_{\max}} \sum_{t=T_{\max}}^{T_0-1} r_{t+1}^{\mathrm{oos}}(T, k),$$

$$s^{\mathrm{oos}}(T, k) := \sqrt{\frac{1}{T_0 - T_{\max} - 1} \sum_{t=T_{\max}}^{T_0-1} \left(r_{t+1}^{\mathrm{oos}}(T, k) - \overline{r}^{\mathrm{oos}}(T, k)\right)^2}.$$

When reporting results, we annualize Sharpe ratios by multiplying $\widehat{\theta}^{\mathrm{oos}}(T, k)$ by $\sqrt{12}$ for monthly data and by $\sqrt{252}$ for daily data. The empirical object of interest is the k–profile $\{\widehat{\theta}^{\mathrm{oos}}(T, k)\}_{k=1}^{N}$ for each T.

To complement the out-of-sample Sharpe ratio evidence, we study the time-series stability of the estimated sparse portfolios. In our framework, the out-of-sample performance shortfall relative to the population k-sparse frontier reflects two conceptually distinct components: selection risk, which operates through changes in the selected support, and allocation risk, which operates through noisy weights conditional on a given support. The next two diagnostics do not estimate these loss components directly, but they provide empirical proxies for each source of instability. Finally, we report turnover to connect our statistical motivation for sparsity to the complementary frictions-based view in which sparse portfolios are attractive because they reduce trading intensity and transaction-cost exposure.

6.1.2 Selection Instability

Because our portfolios are explicitly sparse, it is useful to separate instability in *which* assets are held from instability in *how* weights are assigned given the selected set. Let

$$S_{t,(k)}(T) := \mathrm{supp}\big(\widehat{w}_{T,k,t}\big) = \{i \in [N] : \ \widehat{w}_{T,k,t,i} \neq 0\}$$

denote the selected support at date t under window length T. We quantify the similarity of consecutive supports using the Jaccard index (Jaccard, 1901),

$$J_t(T,k) := \frac{|S_{t,(k)}(T) \cap S_{t-1,(k)}(T)|}{|S_{t,(k)}(T) \cup S_{t-1,(k)}(T)|}, \qquad t = T_{\max}+1,\dots,T_0-1,$$

and define selection instability as the associated Jaccard distance,

$$D_t(T,k) := 1 - J_t(T,k).$$

This distance equals zero when the selected asset set is unchanged and approaches one when consecutive supports have little overlap. We summarize selection instability by its time-series mean,

$$\bar{D}(T,k) := \frac{1}{T_0 - T_{\max} - 1} \sum_{t=T_{\max}+1}^{T_0-1} D_t(T,k),$$

which captures the average rate at which the strategy rotates across holdings as new data arrive.[6] High $\bar{D}(T,k)$ is a direct manifestation of selection risk and indicates that small changes in the estimation window induce frequent changes in the selected support.

6.1.3 Weight Instability

Selection instability isolates changes in the set of active positions. Weight instability instead measures how sensitive the recommended allocations are across adjacent estimation windows, regardless of whether the support changes. This is a direct diagnostic of estimation risk: when moments are noisy, small updates to the in-sample window can lead to large changes in the target weights.

For each (T,k), we measure changes in consecutive implemented target weights via

$$\Delta_t^{(1)}(T,k) := \left\| \widehat{w}_{T,k,t} - \widehat{w}_{T,k,t-1} \right\|_1, \qquad t = T_{\max}+1,\dots,T_0-1.$$

[6] In practice, rebalancing rules are often implemented with buffers: a currently held asset may be retained even if it is no longer among the top k names under the updated estimate, provided it remains close to the cutoff (e.g., among the best candidates right after the k-selected ones). Such buffering reduces unnecessary turnover at the cost of occasionally keeping a slightly suboptimal constituent.

We summarize instability across time by the corresponding medians,[7]

$$\widehat{\Delta}^{(1)}(T,k) := \text{median}\Big\{\Delta_t^{(1)}(T,k) : t = T_{\max}+1,\dots,T_0-1\Big\}.$$

Low values of $\widehat{\Delta}^{(1)}(T,k)$ indicate that the strategy delivers stable exposures across rebalancing dates, whereas high values are symptomatic of noise amplification in the mean–variance inputs.

6.1.4 Turnover

Selection instability and weight instability are statistical diagnostics. To connect them to implementability, we compute portfolio turnover in Appendix B.1, which measures the trading required to rebalance from the drifted holdings to the newly recommended target weights.

Let r_t^f denote the risk-free rate and define total simple returns on the managed portfolios as $R_{t+1} := r_{t+1} + r_{t+1}^f \iota$, where ι is a conformable vector of ones, so that gross return multipliers are $G_{t+1} := \iota + R_{t+1}$. Given the portfolio $\widehat{w}_{T,k,t}$ formed at the end of period t and held over period $t+1$, the *pre-trade* weights at the end of period $t+1$ (after returns and before rebalancing) are

$$\widehat{w}_{T,k,t+1}^{\text{pre}} := \frac{\widehat{w}_{T,k,t} \odot G_{t+1}}{\iota^\top\big(\widehat{w}_{T,k,t} \odot G_{t+1}\big)},$$

where $\odot$ denotes elementwise multiplication. One-way turnover at the rebalance date $t+1$ is

$$\text{TO}_{t+1}(T,k) := \big\|\widehat{w}_{T,k,t+1} - \widehat{w}_{T,k,t+1}^{\text{pre}}\big\|_1 = \sum_{i=1}^{N}\big|\widehat{w}_{T,k,t+1,i} - \widehat{w}_{T,k,t+1,i}^{\text{pre}}\big|.$$

Under proportional transaction costs, if c denotes a one-way cost rate, then the period-$t+1$ trading cost is approximately $c\,\text{TO}_{t+1}(T,k)$, so turnover governs the wedge between gross and net performance.

Because turnover can occasionally spike, we summarize trading intensity via the median one-way turnover,

$$\widehat{\text{TO}}(T,k) := \text{median}\Big\{\text{TO}_{t+1}(T,k) : t = T_{\max},\dots,T_0-2\Big\}.$$

[7] We use medians because the instability measures can exhibit occasional spikes (e.g., around turbulent periods or abrupt re-optimizations). The sample median has a 50% breakdown point and is therefore a robust measure of location; see, e.g., Hampel (2001).

Turnover results are reported and discussed in Appendix B.1.

6.2 Data

In our empirical study, we consider the following datasets of test assets of varying dimensions and frequencies.

1. **Dataset 1: 100 size–value portfolios (N=100, monthly).** Our baseline dataset consists of excess returns on 100 characteristic-managed portfolios formed by double-sorting U.S. equities on size and book-to-market, observed at the monthly frequency. The sample runs from October 1971 through October 2025.

2. **Dataset 2: Dataset 1 + 49 industry portfolios (N=149, monthly).** To enrich the cross-section with industry tilts, we augment Dataset 1 with excess returns on 49 U.S. industry portfolios observed over the same period.

3. **Dataset 3: Broad characteristic-sorted universe + 49 industry portfolios (N=274, daily).** To assess whether the sparsity tradeoff is present at higher frequency, we construct a daily managed-portfolio universe of several sets of 25 double-sorted portfolio excess returns formed on various characteristics (size, book-to-market, operating profitability, investment, short-term reversal, momentum, and long-term reversal), and add the 49 daily industry portfolios. The resulting investment universe has $9 \times 25 + 49 = 274$ daily return series. The daily sample runs from January 2, 1991, to December 31, 2024.

4. **Dataset 4: International size–value and size–profitability portfolios (N=200, monthly).** To study whether the sparsity tradeoff carries over to non-U.S. equity markets, we collect international double-sorted managed-portfolios formed on size and book-to-market and on size and operating profitability in each of four regions: North America, Europe, Japan, and Asia Pacific (ex Japan). Each region contributes 25 portfolios per sort, so the combined universe has $4 \times (25 + 25) = 200$ monthly excess return series, observed between November 1990 and October 2025.

5. **Dataset 5: Broad characteristic-sorted universe + 17 industry portfolios (N=352, monthly).** Our largest dataset expands the managed-portfolio universe to 335 excess-return series formed by double-sorting on various pairs of characteristics (including size, book-to-market, operating profitability, accruals, investment, long-term reversal, market beta, net share issue, variance, and residual variance) and adds the 17 industry portfolios. The sample runs from October 1971 through December 2023.

All return series are obtained from the Kenneth R. French Data Library. Excess returns are computed relative to the risk-free rate (one-month Treasury bill rate). In addition to the test-asset universes above, we repeat the analysis after augmenting each dataset with traded Fama–French factor-mimicking portfolios. For the three monthly U.S. datasets (Datasets 1, 2, and 5), we append the full Fama–French five-factor set (MKT, SMB, HML, RMW, CMA). For the daily managed-portfolio universe (Dataset 3), we append the Fama–French three factors (MKT, SMB, HML). For the international universe (Dataset 4), we append region-specific Fama–French three-factor returns for North America, Europe, Japan, and Asia Pacific (ex Japan), so that the augmented design includes four copies of (MKT, SMB, HML), one for each region. This augmentation allows us to assess the extent to which including traded factor portfolios directly affects the sparsity–estimation tradeoff.

6.3 Results

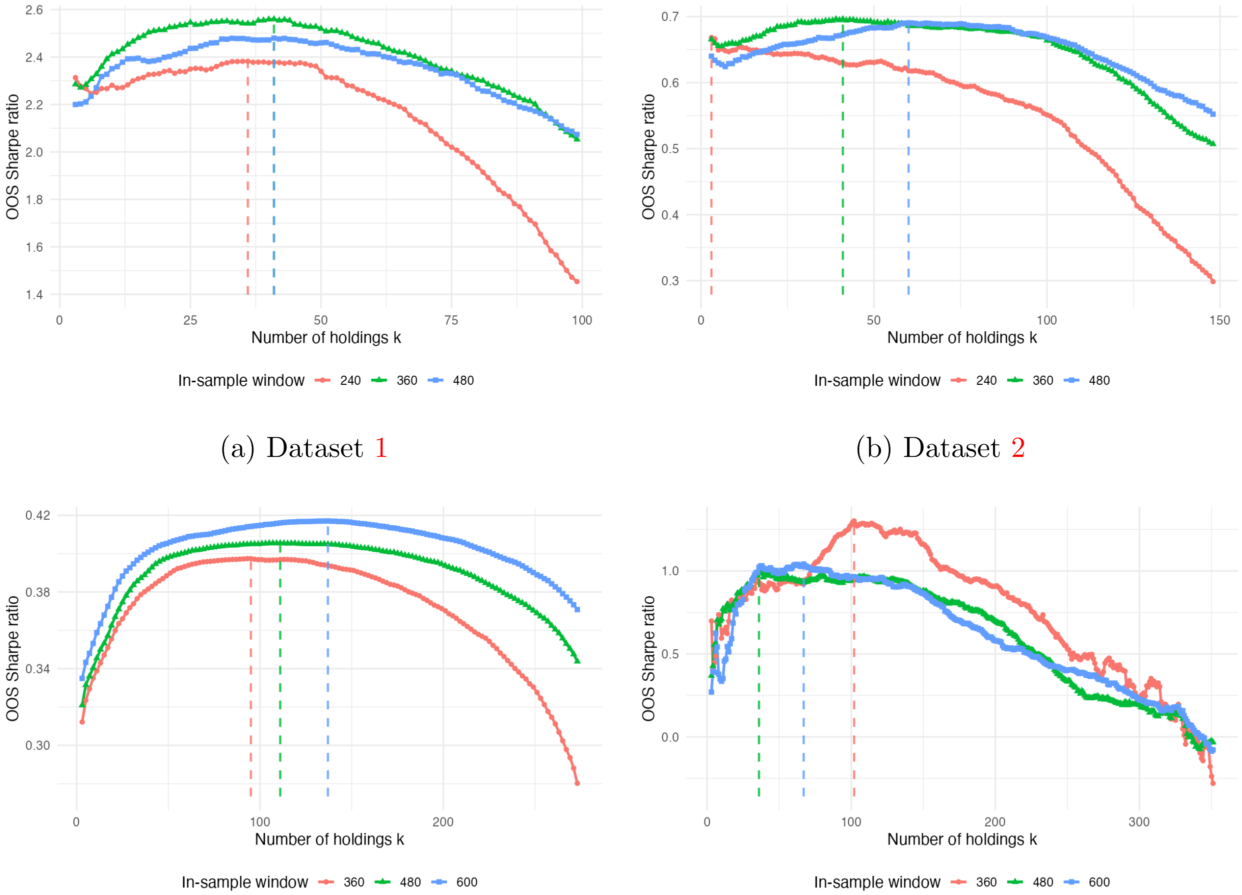
We now summarize the empirical performance of k–sparse portfolios across the test-asset universes described above. Our primary evaluation metric is the realized out-of-sample Sharpe ratio, computed from a rolling procedure and reported as a function of the target sparsity level k for several in-sample window lengths T_{in}. To connect performance to the theory's decomposition of estimation risk, we then report two complementary stability diagnostics: selection instability, which captures how frequently the set of active holdings changes over time, and weight instability, which captures how volatile the fitted weights are conditional on the selected support. Together, these figures characterize how the benefits of expanding the support (spanning and diversification) trade off against the rising difficulty of estimating a higher-dimensional mean–variance structure. Finally, we turn from diagnostic k–profiles to implementable model selection that can be used by an investor who must choose k in real time. We therefore evaluate two simple feasible tuning rules that update the target sparsity infrequently (once per year in monthly data), and compare their realized out-of-sample Sharpe ratios to the infeasible best-in-hindsight benchmark that selects k ex-post to maximize realized out-of-sample performance.

6.3.1 Out-of-sample Sharpe Ratio

Figures 8 and 2 (Dataset 4) report realized out-of-sample Sharpe ratios of the ℓ_1 portfolio rule as a function of the target sparsity level k, for four monthly test-asset universes and one daily universe. The holding period is fixed at $h = 1$, and each curve corresponds to a different in-sample window length T_{in}. In the daily universe, T_{in} is measured in trading days and $h = 1$ denotes a one-day holding period.

Across datasets and window lengths, the Sharpe ratio is systematically non-monotone in k (except for just one case where it is monotonically decreasing in k): it rises at low cardinalities and eventually declines once the portfolio becomes sufficiently dense. This hump-shaped profile is the empirical counterpart of the central tradeoff emphasized by the theory. Increasing k expands the span of attainable payoffs and improves diversification, but it also enlarges the parameter space in which means and covariances must be estimated, thereby amplifying sampling error in both selection and allocation.

The peak location shifts with T_{in} in a way that is broadly consistent with lower estimation noise in longer windows. In Dataset 1 (100 assets), the Sharpe-maximizing sparsity is around $k^* \approx 35$ for $T_{\text{in}} = 240$ and $k^* \approx 45$ for $T_{\text{in}} = 360, 480$. In Dataset 2 (149 assets), the peak is around $k^* \approx 40$ for $T_{\text{in}} = 360$ and $k^* \approx 60$ for $T_{\text{in}} = 480$; the only monotonic decline in k appears here, at $T_{\text{in}} = 240$. Note that, when we augment Dataset 2 with the Fama–French factor-mimicking portfolios, all Sharpe-ratio profiles become hump-shaped across in-sample windows, restoring an interior optimum k^* for each T_{in}; see Figure 17 in Appendix B.2. In the international monthly universe (Dataset 4, 200 assets), the peak occurs near $k^* \approx 25$ for $T_{\text{in}} = 240$ and $k^* \approx 30$ for $T_{\text{in}} = 360$. In the U.S. monthly universe (Dataset 5, 352 assets), k^* is about 100 for $T_{\text{in}} = 360$, then moves back to roughly 40 for $T_{\text{in}} = 480$ and about 70 for



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 5

Figure 8: Out-of-sample Sharpe ratio of the k–sparse ℓ_1 portfolio rule, plotted as a function of the target sparsity level k for the monthly and daily test-asset universes. The holding period is $h = 1$. All OOS quantities are computed over a common evaluation window of length $T_0 - T_{\max}$ within each dataset.

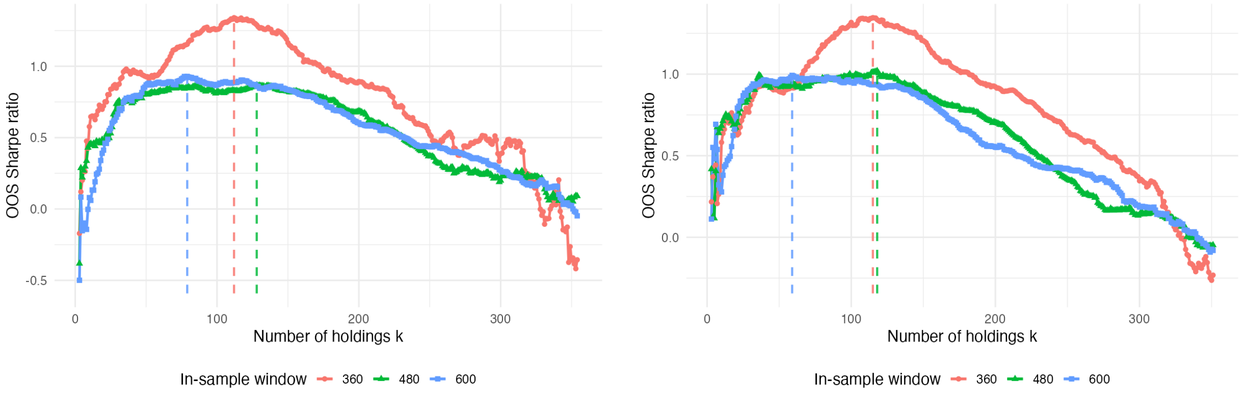
$T_{\text{in}} = 600$. Finally, in the U.S. daily universe (Dataset 3, 274 assets), the peak lies around $k^* \approx 95$–110 for $T_{\text{in}} = 360$–480 and rises to about 140 for $T_{\text{in}} = 600$. Overall, the average selected share k^*/N is about 30% and it stays approximately within 50% of the investable universe across these dataset–window combinations. More informatively than pinpointing the single maximizing k^*, the figures reveal a broad high-performance region in which the out-of-sample Sharpe ratio remains close to its peak; in all but one case this region occurs at intermediate sparsity levels, with the lone exception favoring extremely sparse portfolios.

Cross-dataset differences are informative about the effective dimension of the return space and about how quickly marginal spanning benefits diminish. In the size–value benchmark (Dataset 1), relatively small supports already deliver near-maximum Sharpe ratios, consistent with strong common variation in returns. Adding industry portfolios (Dataset 2) increases the value of additional holdings and makes the Sharpe-maximizing cardinality more sensitive to T_{in}, suggesting that sector-level risks are only partially captured by the baseline size and value spreads and become easier to exploit as estimation noise declines. The managed-portfolio universes display peaks at substantially larger k, indicating that diversifying idiosyncratic components and capturing heterogeneous strategy exposures requires a broader support, even though the Sharpe profile remains hump-shaped. The non-monotonic shift in the maximizing k for very long windows in the U.S. monthly managed data is also consistent with the idea that, beyond sampling variance, slow-moving nonstationarities and regime changes can make very long estimation windows less effective for identifying a stable high-dimensional mean–variance structure.

Overall, the Sharpe-ratio evidence supports an interior optimal sparsity level that depends jointly on the signal-to-noise ratio (as proxied by T_{in}) and the structure of the test-asset universe, providing direct empirical content to the estimation–efficiency tradeoff highlighted by the theory.

6.3.2 Out-of-sample Sharpe Ratio: Robustness Checks

Two features could potentially influence the hump-shaped Sharpe profiles in Figures 8: (i) the omission of standard systematic payoffs from the investment universe and (ii) variation in the scale of portfolio positions along the regularization path and across rolling windows. We address both concerns through targeted robustness checks in the two managed-portfolio universes, Datasets 3 and 5, reported in Figures 9 and 10; corresponding results for the remaining datasets are deferred to the appendix.



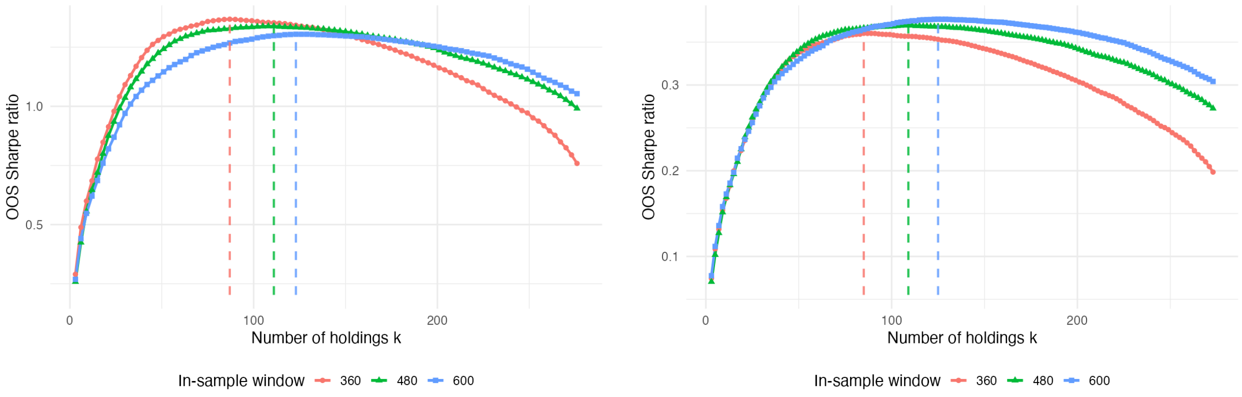
(a) Factor-augmented dataset

(b) Normalized weights: $\|w\|_1 = 1$

Figure 9: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule in the U.S. monthly managed-portfolio universe (Dataset 5), under the two robustness checks. The holding period is $h = 1$.

First, we augment the investable universe with Fama–French factor-mimicking portfolios and rerun the rolling out-of-sample analysis on the expanded cross-section. In particular, we append the full Fama–French five-factor set (MKT, SMB, HML, RMW, CMA) to the U.S. monthly managed-portfolio universe (Dataset 5), and the Fama–French three factors (MKT, SMB, HML) to the U.S. daily managed-portfolio universe (Dataset 3). This check serves two purposes. It verifies that our findings are not an artifact of excluding widely used benchmark factor exposures from the opportunity set, and it probes the economic mechanism behind sparsity by introducing additional tradable payoffs that proxy for well-documented systematic risks.

Empirically, the Sharpe profiles remain qualitatively hump-shaped in k in both managed universes. Relative to the baseline, the maximizing sparsity shifts modestly. In the daily universe, the peak moves down to $k^* \approx 85$ for $T_{\text{in}} = 360$ and to roughly $k^* \approx 110$–120 for $T_{\text{in}} = 480$–600, compared with $k^* \approx 95$–110 for $T_{\text{in}} = 360$–480 and about 140 for $T_{\text{in}} = 600$ without factor augmentation. In the monthly universe, the peak shifts slightly upward at intermediate windows, to about $k^* \approx 110$ for $T_{\text{in}} = 360$ and $k^* \approx 130$ for $T_{\text{in}} = 480$, while



(a) Factor-augment dataset

(b) Normalized weights: $\|w\|_1 = 1$

Figure 10: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule in the U.S. daily managed-portfolio universe (Dataset 3), under the two robustness checks. The holding period is $h = 1$ (one trading day).

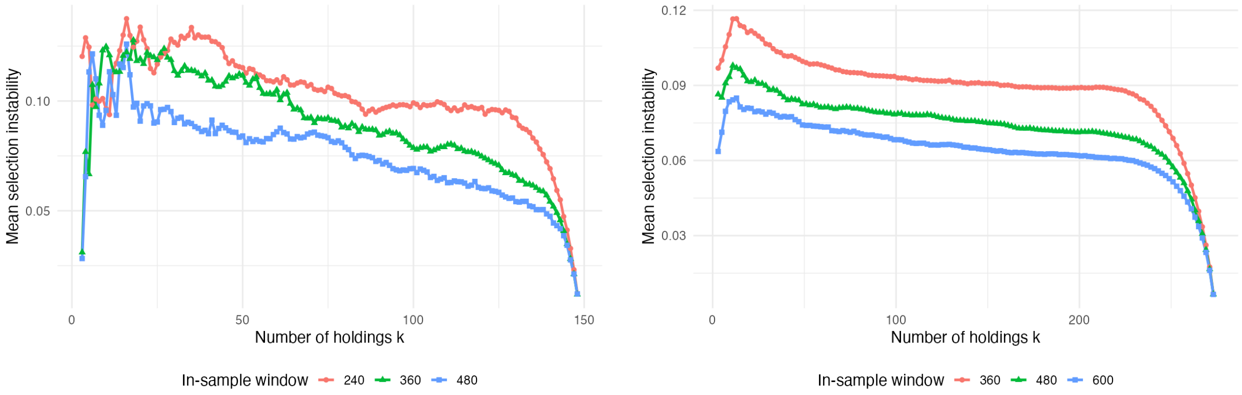
remaining around $k^* \approx 70$ for $T_{\text{in}} = 600$ (versus about 100, 40, and 70 in the baseline for $T_{\text{in}} = 360$, 480, and 600, respectively). Importantly, these shifts are local: the range of sparsity levels delivering Sharpe ratios close to the maximum remains largely unchanged, indicating that the main conclusions are robust to making standard factor payoffs explicitly tradable.

Second, we return to the baseline asset universes without factor augmentation and normalize the realized portfolio weights to satisfy $\|w\|_1 = 1$. The goal is to maintain comparability across rolling windows by imposing a common gross-exposure scale. This normalization does not conflict with the use of the ℓ_1 penalty in (3): the penalty is used to trace a sparse selection path (and hence to target a given sparsity index k), whereas the ex-post unit-ℓ_1 normalization fixes the scale of the position. Varying k while keeping $\|w\|_1 = 1$ is therefore a closer proxy to how sparse portfolios are implemented in practice, where the investor allocates a fixed amount of wealth (or gross exposure) and the number of active holdings is the object of choice rather than a determinant of leverage.

Imposing this gross-exposure normalization leaves the Sharpe profiles virtually unchanged: the curves remain hump-shaped, the location of the high-performance middle-k region is stable, and any changes in the maximizing k^* are minor. This indicates that the patterns documented in the baseline figures are not driven by incidental variation in portfolio scale across the regularization path or across rolling estimation windows.

6.3.3 Selection and Weight Instability

Figures 11 and 12 report two diagnostics that speak directly to the two components of estimation risk emphasized by the theory. Selection instability measures how much the selected support varies across rebalancing dates, while weight instability measures how much the refitted weights vary conditional on the selected support. These objects do not estimate Sharpe-ratio losses mechanically, but they provide an interpretable map from the data to the two channels of estimation error: unstable supports point to selection risk, and unstable within-support weights point to allocation risk. For brevity, we display the results for the size–value–industry universe (Dataset 2) and for the U.S. daily managed-portfolio universe (Dataset 3); the corresponding figures for the remaining datasets are reported in the appendix.

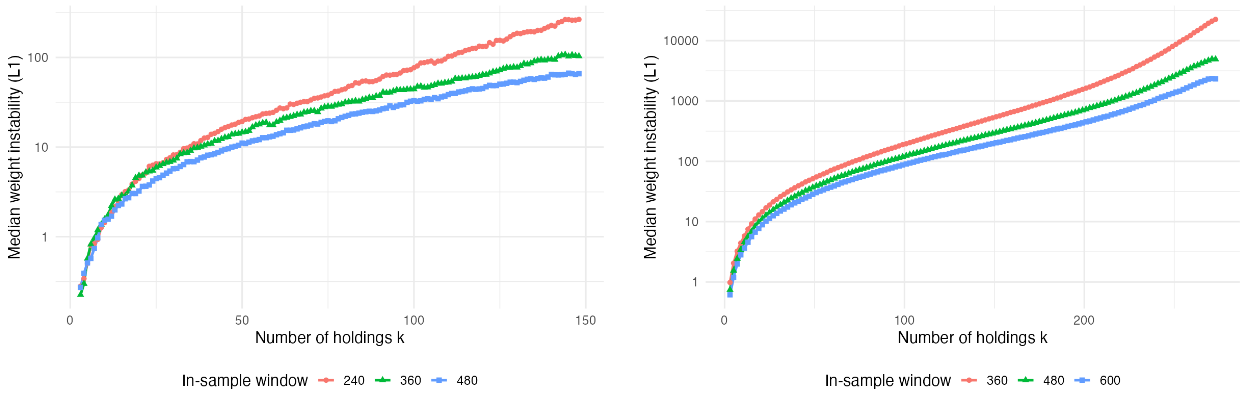


(a) Dataset 2

(b) Dataset 3

Figure 11: Selection instability of the ℓ_1 portfolio rule as a function of the target sparsity level k for Dataset 2 (monthly) and Dataset 3 (daily). The holding period is $h = 1$ (one month for monthly data and one trading day for daily data).

The selection-instability curves exhibit a robust and economically intuitive pattern. Selection instability is highest for very sparse portfolios, where the support is chosen by ranking a large number of candidate subsets using noisy Sharpe-ratio estimates. When k is small,



(a) Dataset 2

(b) Dataset 3

Figure 12: Weight instability (conditional on the selected support) of the ℓ_1 portfolio rule as a function of k for Dataset 2 (monthly) and Dataset 3 (daily). Reported on a log scale and summarized by the time-series median. The holding period is $h = 1$.

the search space expands quickly as k increases—there are many more admissible supports and many of them deliver nearly indistinguishable in-sample fit—so small perturbations in estimated moments can reshuffle the top-ranked set and generate low overlap across rebalancing dates. As k grows further, two forces stabilize the selected holdings. First, supports of size k overlap more by construction, so even when the optimizer substitutes some names, most holdings remain the same. Second, in these datasets many assets are close substitutes (e.g., within style or industry groups), so enlarging the support allows the optimizer to accommodate sampling noise by adding marginal holdings rather than replacing core ones. As a result, instability declines at higher k and becomes especially small as k approaches N, where the feasible supports are necessarily very similar and selection becomes nearly mechanical. This pattern is clear-cut in the daily managed-portfolio universe (Dataset 3) and is also evident in the size–value–industry universe (Dataset 2); we obtain qualitatively similar results for the remaining datasets, as reported in the appendix.

Weight instability, reported on a log scale and summarized by the time-series median, moves in the opposite direction. Across all datasets and window lengths, it increases with k, with the steepest rise occurring at low cardinalities. This pattern is consistent with allocation risk increasing with the dimension of the refitted mean–variance problem. When k is small, the refit operates in a low-dimensional subspace and the resulting weights are comparatively stable across rebalancing dates. As k grows, more weights must be estimated from the same window of data, and the within-support optimizer becomes increasingly sensitive to noise in the estimated covariance matrix and mean vector, leading to larger fluctuations in the refitted portfolio even when the support itself is relatively stable. The log scaling highlights that the increase is not merely incremental: moving away from very sparse portfolios can quickly amplify the variability of the fitted weights.

Taken together, these diagnostics complement the Sharpe-ratio evidence. The hump-shaped out-of-sample Sharpe profiles arise in a region where the marginal benefit of expanding the support—lower selection risk and improved spanning—is progressively offset by higher allocation risk. The empirical fact that selection instability falls with k while weight instability rises with k provides direct qualitative support for the model's decomposition of estimation risk into selection and allocation components.

Turnover and implementation. Selection instability and weight instability are informative because they map naturally into portfolio turnover. When the selected support changes from one rebalancing date to the next, the investor must unwind positions that exit the support and establish positions that enter, generating trading even if the within-support weights were otherwise stable. Conditional on a stable support, fluctuations in refitted weights also generate trading through rebalancing. Turnover therefore connects the statistical instability diagnostics to an economically interpretable object measured in dollars traded. This link is useful for interpreting the hump-shaped out-of-sample Sharpe profiles: as k increases, the selected set tends to stabilize, but the refitted weights become increasingly sensitive to estimation noise, so trading induced by rebalancing can rise even when selection risk declines. Our main focus is the intrinsic estimation cost of sparsity in frictionless markets, but a complementary motivation for sparse portfolio rules in the applied literature is precisely to reduce turnover and trading costs. In Appendix B.1, we therefore report turnover profiles as an implementation diagnostic for the ℓ_1 rule.

6.3.4 Feasible tuning rules

The Sharpe-ratio profiles in Figures 2 and 8 are computed for each fixed target sparsity $k \in [N]$. In practice, however, the investor must choose k in a fully feasible way, without access to future out-of-sample returns. We thus conclude our empirical investigation by evaluating two simple, implementable tuning rules that select k using only information available at the tuning date, and comparing their realized out-of-sample Sharpe ratios to an infeasible best-in-hindsight benchmark.

For a fixed window length $T \in \mathcal{T}$, define the realized out-of-sample Sharpe ratio $\widehat{\theta}^{\text{oos}}(T,k)$, as in Section 6.1.1. The *ex-post* benchmark selects the best-performing cardinality in hindsight,

$$\widehat{\theta}^{\text{oos}}_{\max}(T) \;:=\; \max_{k\in[N]} \; \widehat{\theta}^{\text{oos}}(T,k).$$

This benchmark is not implementable because it uses the same out-of-sample returns that it evaluates.

We implement feasible tuning for the monthly datasets by updating k once per calendar year and holding it fixed within the year, while continuing to rebalance the portfolio each month. Let $\{\tau_j\}_{j=1}^{J}$ denote the sequence of annual tuning dates within the common evaluation period (e.g., the end of December of each year), and let

$$\mathcal{I}_j \;:=\; \{t: \; \tau_j \le t < \tau_{j+1}\}$$

denote the set of monthly rebalancing dates whose out-of-sample returns fall in year j (with the obvious convention for the last block). For each T and each tuning date τ_j, we choose a cardinality $\widehat{k}_j(T)$ using only the information available up to τ_j (i.e., the most recent T monthly observations ending at τ_j), and then implement the strategy over $t \in \mathcal{I}_j$ using the rolling-window weights $\widehat{w}_{T,\widehat{k}_j(T),t}$.

Concretely, the realized out-of-sample return series under a tuned rule "rule"$\in \{\text{LOO}, \text{UE}\}$ is

$$r^{\text{oos}}_{t+1}\big(T, \widehat{k}^{\text{rule}}(T)\big) \;:=\; \widehat{w}^{\top}_{T,\widehat{k}^{\text{rule}}_j(T),t} r_{t+1}, \qquad t \in \mathcal{I}_j,$$

and its realized Sharpe ratio $\widehat{\theta}^{\text{oos}}_{\text{rule}}(T)$ is computed from the concatenated sequence over the common evaluation period, annualized as in Section 6.1.1.

At each tuning date τ_j and for each candidate k, we compute two feasible scores (the analogues of the Monte Carlo rules in Section 5.3.4):

- *Leave-one-out (LOO).* Using the most recent in-sample window of length T ending at τ_j, we construct a leave-one-out pseudo out-of-sample return series for each k and compute its Sharpe ratio $\widehat{\theta}_{\mathrm{LOO},\tau_j}(T,k)$. We then set

$$\widehat{k}_j^{\mathrm{LOO}}(T) \in \underset{k\in[N]}{\operatorname{argmax}}\ \widehat{\theta}_{\mathrm{LOO},\tau_j}(T,k).$$

- *Unbiased-estimator (UE).* Using the same in-sample window, we compute for each k the bias-adjusted squared Sharpe statistic $\widetilde{\theta}^2_{\widehat{A}_{T,k,\tau_j}}$ from Remark 1, where $\widehat{A}_{T,k,\tau_j}$ is the support selected by the k–sparse procedure at τ_j. We then set

$$\widehat{k}_j^{\mathrm{UE}}(T) \in \underset{k\in[N]}{\operatorname{argmax}}\ \widetilde{\theta}^2_{\widehat{A}_{T,k,\tau_j}}.$$

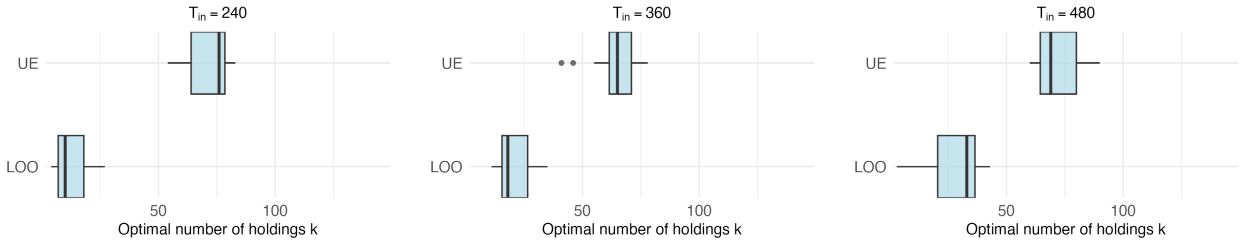
Both rules are fully real-time implementable: they use only the T returns available at the tuning date and do not condition on future out-of-sample performance.

Table 1 reports, for two representative monthly datasets, the infeasible ex-post benchmark $\widehat{\theta}^{\mathrm{oos}}_{\max}(T)$ and the realized Sharpe ratios under annual LOO and UE tuning, $\widehat{\theta}^{\mathrm{oos}}_{\mathrm{LOO}}(T)$ and $\widehat{\theta}^{\mathrm{oos}}_{\mathrm{UE}}(T)$. These results are for the baseline design without factor augmentation and without ex-post ℓ_1 normalization; the corresponding robustness results are reported in Appendix B.3.1.

Comparing Table 1 to the fixed-k Sharpe profiles in Figures 2 and 8 shows that feasible tuning can recover a substantial share of the attainable out-of-sample performance. In both datasets, the annually tuned rules deliver realized Sharpe ratios that remain close to the best-in-hindsight benchmark $\widehat{\theta}^{\mathrm{oos}}_{\max}(T)$, especially once the in-sample window is sufficiently long. This pattern is most pronounced for the UE rule: for the size–value–industry universe (Dataset 2), UE essentially matches the ex-post benchmark at $T_{\mathrm{in}} = 360$ (0.69 versus 0.70) and remains close at $T_{\mathrm{in}} = 480$ (0.67 versus 0.69), while it performs less well at the shortest window $T_{\mathrm{in}} = 240$. For the international universe (Dataset 4), UE also improves markedly with T_{in} and becomes competitive at $T_{\mathrm{in}} = 360$, where it attains 2.30 versus the benchmark 2.59. Overall, the evidence is consistent with the interpretation that UE benefits from lower estimation noise: when the window is long enough for the bias-adjusted Sharpe objective to be informative, it provides a stable and effective mechanism for selecting an interior sparsity level without look-ahead.

The tuning behavior of the two rules differs systematically. Figures 13 and 14 show that LOO tends to select substantially smaller supports than UE, i.e., it favors much sparser portfolios (lower k) across tuning years and across in-sample window lengths. This tendency is consistent with LOO penalizing model complexity more aggressively in short samples: evaluating each candidate k via leave-one-out pseudo out-of-sample returns can amplify the impact of estimation noise and thus tilt the selected k toward very sparse allocations. By contrast, UE typically selects less sparse portfolios. This is consistent with its role as a within-support bias correction of the in-sample Sharpe objective: it primarily adjusts for the finite-sample upward bias in the plug-in Sharpe ratio conditional on a given support. In particular, UE does not directly account for selection risk, the additional overfitting that arises because the support itself is chosen adaptively from many competing candidate sets. As a result, relative to LOO (which evaluates each k through pseudo out-of-sample performance and thus implicitly penalizes selection instability), UE tends to put less weight on the instability of the selected set and correspondingly favors larger, less sparse portfolios.

Taken together, the table and figure evidence indicates that feasible annual tuning rules can deliver high out-of-sample Sharpe ratios, with UE performing particularly well once T_{in} is large enough for the bias adjustment to reliably discriminate among candidate sparsity levels.



(a) $T_{\text{in}} = 240$ (b) $T_{\text{in}} = 360$ (c) $T_{\text{in}} = 480$

Figure 13: **Chosen sparsity under annual tuning: Dataset 2.** Boxplots of the annually selected cardinalities $\widehat{k}_j^{\text{LOO}}(T)$ and $\widehat{k}_j^{\text{UE}}(T)$ across tuning years, for each T_{in}.

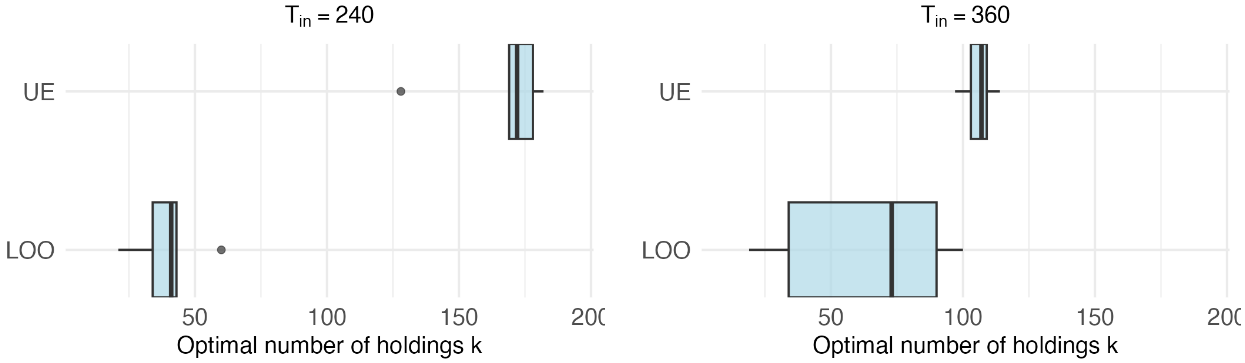
Table 1: **Feasible tuning rules (annual tuning, monthly data).** For each in-sample window length T_{in}, the table reports the infeasible best-in-hindsight Sharpe ratio $\widehat{\theta}^{\text{oos}}_{\max}(T)$ (max over k), and the realized out-of-sample Sharpe ratios under annual tuning of k via leave-one-out (LOO) and the unbiased-estimator rule (UE). All Sharpe ratios are annualized. Baseline design: no factor augmentation and no ex-post gross-exposure normalization.

(a) Dataset 2

T_{in}	$\widehat{\theta}^{\text{oos}}_{\max}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	0.6682	0.6356	0.5965
360	0.6961	0.6750	0.6896
480	0.6905	0.6549	0.6743

(b) Dataset 4

T_{in}	$\widehat{\theta}^{\text{oos}}_{\max}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	3.0631	2.2499	1.1421
360	2.5874	2.0981	2.3037



(a) $T_{\text{in}} = 240$

(b) $T_{\text{in}} = 360$

Figure 14: **Chosen sparsity under annual tuning: Dataset 4.** Boxplots of the annually selected cardinalities $\widehat{k}^{\text{LOO}}_j(T)$ and $\widehat{k}^{\text{UE}}_j(T)$ across tuning years, for each T_{in}.

7 Concluding Remarks

This paper reexamines the role of sparsity in portfolio selection and provides a new rationale for why investors may optimally prefer to hold only a small number of assets, even in the absence of market frictions. The classical mean–variance paradigm prescribes broad diversification, but its empirical implementation suffers from substantial estimation risk when the number of assets is large relative to available data. We show that constraining the number of holdings can substantially mitigate this estimation risk, and that, under realistic return-generating processes, the associated loss in mean–variance efficiency is small for empirically relevant sparsity levels.

Our theoretical analysis disentangles two components of estimation risk: selection risk, associated with identifying which assets to include, and allocation risk, associated with estimating their optimal weights. We prove that both components can be tightly controlled when the number of holdings k is modest relative to the sample size T, with total estimation risk of order $\sqrt{k \log N/T}$. Under approximate factor structures, the opportunity cost of sparsity—the loss of efficiency from excluding some assets—declines as $1/k$ under strong factors and as $N^{1-\zeta}/k$ when factors are weak (where $\zeta \in (0,1]$ indexes factor strength: $\zeta = 1$ for strong/pervasive factors, smaller ζ for weaker factors). Together, these results imply that sparse portfolios can approach mean–variance efficiency as long as $N^{1-\zeta} \ll k \ll T/\log N$. In large markets, this condition can be easily satisfied, suggesting that sparsity is not merely a practical constraint but a theoretically justified form of regularization.

We further demonstrate that the ℓ_1-regularized portfolio, a continuous relaxation of the cardinality constraint, inherits these desirable properties. It achieves asymptotic efficiency under similar scaling conditions and remains robust in the presence of pricing errors or model misspecification. From a computational standpoint, the ℓ_1 approach is tractable even for thousands of assets and produces stable, interpretable portfolios. From an economic standpoint, it offers a disciplined way to balance diversification benefits against estimation uncertainty.

Our simulation and empirical analyses confirm these theoretical predictions. In simulated markets calibrated to the Fama–French factor structure, sparse portfolios achieve dramatic improvements in out-of-sample performance relative to the fully diversified plug-in mean–variance portfolio, particularly when the estimation window is short. In actual returns of the Fama–French 100 portfolios, sparse portfolios constructed with only five to eight holdings attain nearly the same Sharpe ratios as their unconstrained counterparts. The ℓ_1-regularized rules yield smooth, stable performance, and the latter offers additional robustness to data variation. These results reinforce the interpretation of sparsity as an economically rational, data-driven constraint that enhances decision quality under uncertainty.

More broadly, our findings suggest that traditional portfolio constraints, often viewed as ad hoc or friction-motivated, can be understood as *statistical regularizers* that improve out-of-sample performance by limiting portfolio complexity. This perspective connects portfolio theory with modern developments in high-dimensional statistics and machine learning. It reframes diversification not as a binary choice between concentrated and fully diversified portfolios, but as a continuum of information-efficient allocations determined by the available data and estimation precision.

The analysis also has practical implications for portfolio management. In environments where data are limited or asset universes are large, investors may rationally prefer sparse allocations even in the absence of trading frictions. Sparse and ℓ_1-regularized portfolios deliver robustness and interpretability—features increasingly valued in risk management and quantitative investing. Incorporating sparsity into portfolio construction can thus improve both performance and transparency.

Future research may extend our framework in several directions. First, refining dynamic sparsity, where the sparsity level k and the set of active assets adjust gradually over time, could link our results to optimal rebalancing under learning. Second, exploring conditional information structures could further clarify how sparsity interacts with time-varying estimation risk. Finally, empirical studies that apply sparse optimization to real institutional portfolios could test whether investors naturally exhibit the degree of sparsity predicted by our theory.

In sum, this paper provides a unified explanation for the success of sparse portfolio strategies. Even in frictionless markets, sparsity improves decision quality by managing estimation uncertainty. Under realistic factor structures, such portfolios can achieve near mean–variance efficiency with only a small number of holdings. Simplicity, in this context, is not a limitation, it is an economically rational response to the complexity of estimation itself.

References

Mengmeng Ao, Li Yingying, and Xinghua Zheng. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019.

Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems*, 25, 2012.

Afonso S Bandeira, Edgar Dobriban, Dustin G Mixon, and William F Sawin. Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6): 3448–3450, 2013.

Dimitris Bertsimas and Ryan Cory-Wright. A scalable algorithm for sparse portfolio selection. *INFORMS Journal on Computing*, 34(3):1489–1511, 2022.

Michael J Best and Robert R Grauer. On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *The Review of Financial Studies*, 4(2):315–342, 1991.

Mark Britten-Jones. The sampling error in estimates of mean-variance efficient portfolio weights. *The Journal of Finance*, 54(2):655–671, 1999.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica: Journal of the Econometric Society*, pages 1281–1304, 1983.

T-J Chang, Nigel Meade, John E Beasley, and Yazid M Sharaiha. Heuristics for cardinality constrained portfolio optimisation. *Computers & Operations Research*, 27(13):1271–1302, 2000.

Vijay K Chopra, William T Ziemba, et al. The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

Gregory Connor and Robert A Korajczyk. A test for the number of factors in an approximate factor model. *the Journal of Finance*, 48(4):1263–1291, 1993.

Zhifeng Dai and Fenghua Wen. A generalized approach to sparse and stable portfolio optimization problem. *Journal of Industrial and Management Optimization*, 14(4):1651–1666, 2018.

F.R. De Hoog and R.M.M. Mattheij. Subset selection for matrices. *Linear Algebra and its Applications*, 422(2-3):349–359, 2007.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Noureddine El Karoui. High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *Annals of statistics*, 38 (6):3487–3566, 2010.

Edwin J Elton and Martin J Gruber. Risk reduction and portfolio size: An analytical solution. *The Journal of Business*, 50(4):415–437, 1977.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012.

Björn Fastrich, Sandra Paterlini, and Peter Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, 12(3):417–434, 2015.

Jianjun Gao and Duan Li. Optimal cardinality constrained portfolio selection. *Operations Research*, 61(3):745–761, 2013.

Lorenzo Garlappi, Raman Uppal, and Tan Wang. Portfolio selection with parameter and model uncertainty: A multi-prior approach. *The Review of Financial Studies*, 20(1):41–81, 2007.

Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

Frank R Hampel. Robust statistics: A brief introduction and overview. In *Research Report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*, volume 94. Seminar für Statistik, Eidgenössische Technische Hochschule, 2001.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

J David Jobson and Bob Korkie. Estimation for markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554, 1980.

John D Jobson and Bob Korkie. A performance interpretation of multivariate tests of asset set intersection, spanning, and mean-variance efficiency. *Journal of Financial and Quantitative Analysis*, 24(2):185–204, 1989.

Michael Johannes, Arthur Korteweg, and Nicholas Polson. Sequential learning, predictability, and optimal portfolio returns. *The Journal of Finance*, 69(2):611–644, 2014.

Philippe Jorion. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative analysis*, 21(3):279–292, 1986.

Raymond Kan and Daniel R Smith. The distribution of the sample minimum-variance frontier. *Management Science*, 54(7):1364–1380, 2008.

Raymond Kan and Guofu Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656, 2007.

Raymond Kan, Xiaolu Wang, and Xinghua Zheng. In-sample and out-of-sample sharpe ratios of multi-factor asset pricing models. *Journal of Financial Economics*, 155:103837, 2024.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Philipp J Kremer, Sangkyun Lee, Małgorzata Bogdan, and Sandra Paterlini. Sparse portfolio selection via the sorted ℓ_1-norm. *Journal of Banking & Finance*, 110:105687, 2020.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5): 603–621, 2003.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.

Jiahan Li. Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business & Economic Statistics*, 33(3):381–392, 2015.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952.

Richard O Michaud. The markowitz optimization enigma: Is 'optimized'optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

Meir Statman. How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, 22(3):353–363, 1987.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Jun Tu and Guofu Zhou. Incorporating economic objectives into bayesian priors: Portfolio choice under parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 45 (4):959–986, 2010.

Jun Tu and Guofu Zhou. Markowitz meets talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1):204–215, 2011.

Reha H Tütüncü and Mark Koenig. Robust asset allocation. *Annals of Operations Research*, 132(1):157–187, 2004.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

A Notation

Let $N \geq 1$ denote the number of risky assets and let $T \geq 1$ denote the sample size. Write $[N] := \{1, \dots, N\}$. Vectors are column vectors; for $x \in \mathbb{R}^N$, its ith component is x_i and its transpose is $x^\top$. For a subset $A \subset [N]$, its cardinality is $|A|$. For $k \in [N]$, $\binom{N}{k}$ denotes the number of subsets $A \subset [N]$ with $|A| = k$. For $x \in \mathbb{R}^N$ and $A \subset [N]$, the subvector $x_A \in \mathbb{R}^{|A|}$ collects the components $(x_i)_{i \in A}$ in the natural order. For a matrix $M \in \mathbb{R}^{N \times N}$ and $A \subset [N]$, the principal submatrix $M_A \in \mathbb{R}^{|A| \times |A|}$ is obtained by restricting both rows and columns of M to indices in A. The $N \times N$ identity matrix is I_N.

The support of a vector $w \in \mathbb{R}^N$ is $\operatorname{supp}(w) := \{i \in [N] : w_i \neq 0\}$. The ℓ_0 "norm" is $\|w\|_0 := |\operatorname{supp}(w)|$. The ℓ_1, ℓ_2, and ℓ_∞ norms are

$$\|w\|_1 := \sum_{i=1}^{N} |w_i|, \qquad \|w\|_2 := \Big(\sum_{i=1}^{N} w_i^2\Big)^{1/2}, \qquad \|w\|_\infty := \max_{1 \leq i \leq N} |w_i|.$$

For a set $\mathcal{W} \subset \mathbb{R}^N$, $\operatorname{conv}(\mathcal{W})$ denotes its convex hull. For $z \in \mathbb{R}$, the sign function is $\operatorname{sign}(z) = 1$ if $z > 0$, $\operatorname{sign}(z) = -1$ if $z < 0$, and $\operatorname{sign}(0) = 0$. For $w \in \mathbb{R}^N$, the subdifferential of $\|w\|_1$ is the set $\partial \|w\|_1$ of vectors $u \in \mathbb{R}^N$ satisfying $u_i = \operatorname{sign}(w_i)$ when $w_i \neq 0$ and $u_i \in [-1, 1]$ when $w_i = 0$.

For symmetric matrices M, $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the smallest and largest eigenvalue of M. A Cholesky factorization of a positive definite matrix M is $M = C^\top C$, where C is upper triangular with positive diagonal entries.

Asymptotic notation is as follows. For a real sequence (a_n),

$$\liminf_{n \to \infty} a_n := \lim_{n \to \infty} \inf_{m \geq n} a_m, \qquad \limsup_{n \to \infty} a_n := \lim_{n \to \infty} \sup_{m \geq n} a_m.$$

For deterministic sequences (a_n) and (b_n) with $b_n > 0$, $a_n = O(b_n)$ means $\sup_n |a_n|/b_n < \infty$, and $a_n = o(b_n)$ means $|a_n|/b_n \to 0$. For random sequences (X_n) and deterministic (b_n) with $b_n > 0$, $X_n = O_p(b_n)$ means X_n/b_n is tight; that is, for every $\varepsilon > 0$ there exists $M < \infty$ such that

$$\sup_{n \geq 1} \mathbb{P}\left(|X_n/b_n| > M\right) \leq \varepsilon,$$

and $X_n = o_p(b_n)$ means $X_n/b_n \xrightarrow{p} 0$. The notation $a_n \ll b_n$ means $a_n/b_n \to 0$ (hence $a_n \ll b_n$ is synonymous with $a_n = o(b_n)$), and $a_n \asymp b_n$ means both $a_n = O(b_n)$ and $b_n = O(a_n)$.

Convergence in probability is denoted by $\xrightarrow{p}$. The distributional notation $r_t \sim \mathcal{N}(\mu, \Sigma)$ means that r_t is Gaussian with mean μ and covariance Σ.

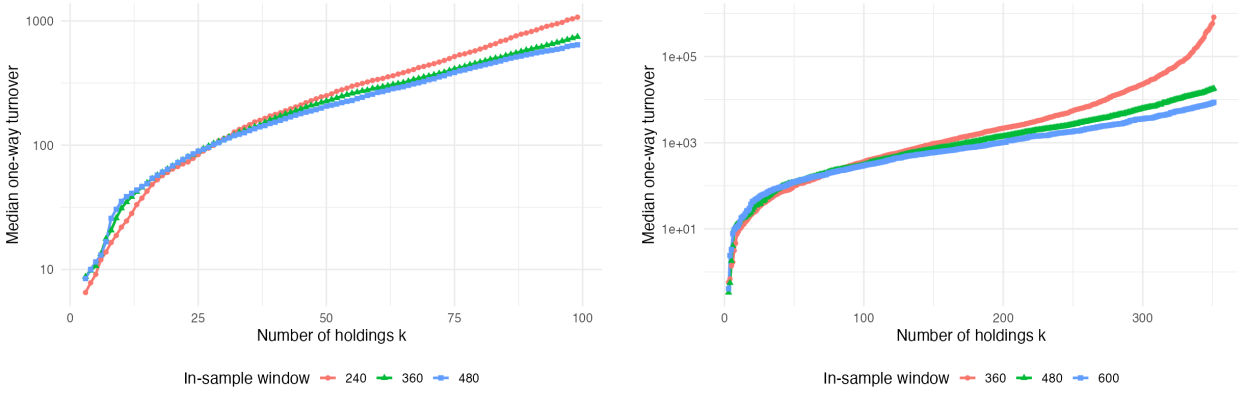
B Further Empirical Results

This appendix collects additional empirical results that complement the main text. We first report turnover profiles of the k–sparse portfolio rule (2) (implemented via the ℓ_1 relaxation (3)) as a function of the target sparsity level k. We then provide the remaining robustness check figures referenced in the main text: out-of-sample Sharpe ratios obtained (i) after augmenting the investable universe with Fama–French factor-mimicking portfolios and (ii) after normalizing realized portfolio weights to satisfy $\|w\|_1 = 1$. We conclude by reporting selection- and weight-instability diagnostics for the datasets omitted from the main text.

B.1 Turnover

We report turnover of sparse portfolios as a function of the target sparsity level k. One-way turnover is measured at rebalancing date t as the total absolute change in portfolio weights between the pre-trade and post-trade portfolios and is summarized by the time-series median; see Section 6.1.4. The resulting profiles provide a direct implementation-oriented complement to the instability diagnostics in the main text, translating selection and weight fluctuations into a dollar-trading metric.

For brevity, we display turnover results for two representative monthly universes—the size–value benchmark (Dataset 1) and the broad U.S. monthly managed-portfolio universe (Dataset 5)—and for the daily managed-portfolio universe (Dataset 3). Figure 15 reports turnover for the two monthly datasets, while Figure 16 reports turnover for the daily dataset. Across these datasets and in-sample window lengths, median one-way turnover increases with k. This pattern is consistent with the behavior of weight instability: as the target support expands, a larger set of active positions must be re-optimized and rebalanced, making the portfolio more sensitive to sampling variation in estimated moments and increasing the amount of trading required to track the rule. At the same time, because selection instability typically declines with k, the composition of the traded positions shifts from being driven by entry/exit of assets at low k toward rebalancing within a larger support at higher k.



(a) Dataset 1

(b) Dataset 5

Figure 15: Median one-way turnover of the ℓ_1 portfolio rule as a function of the target sparsity level k for two monthly test-asset universes. The holding period is $h = 1$.

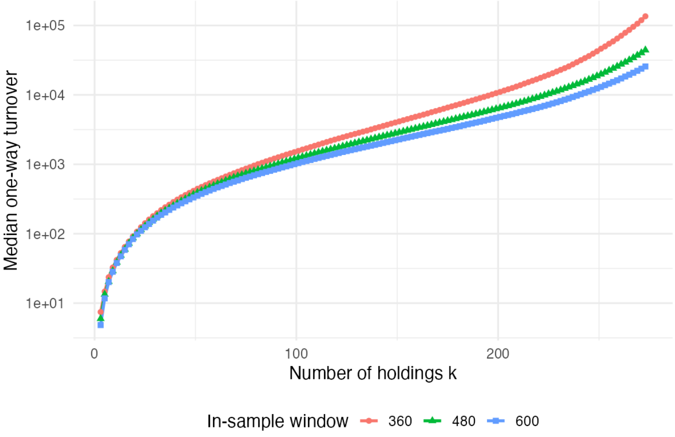
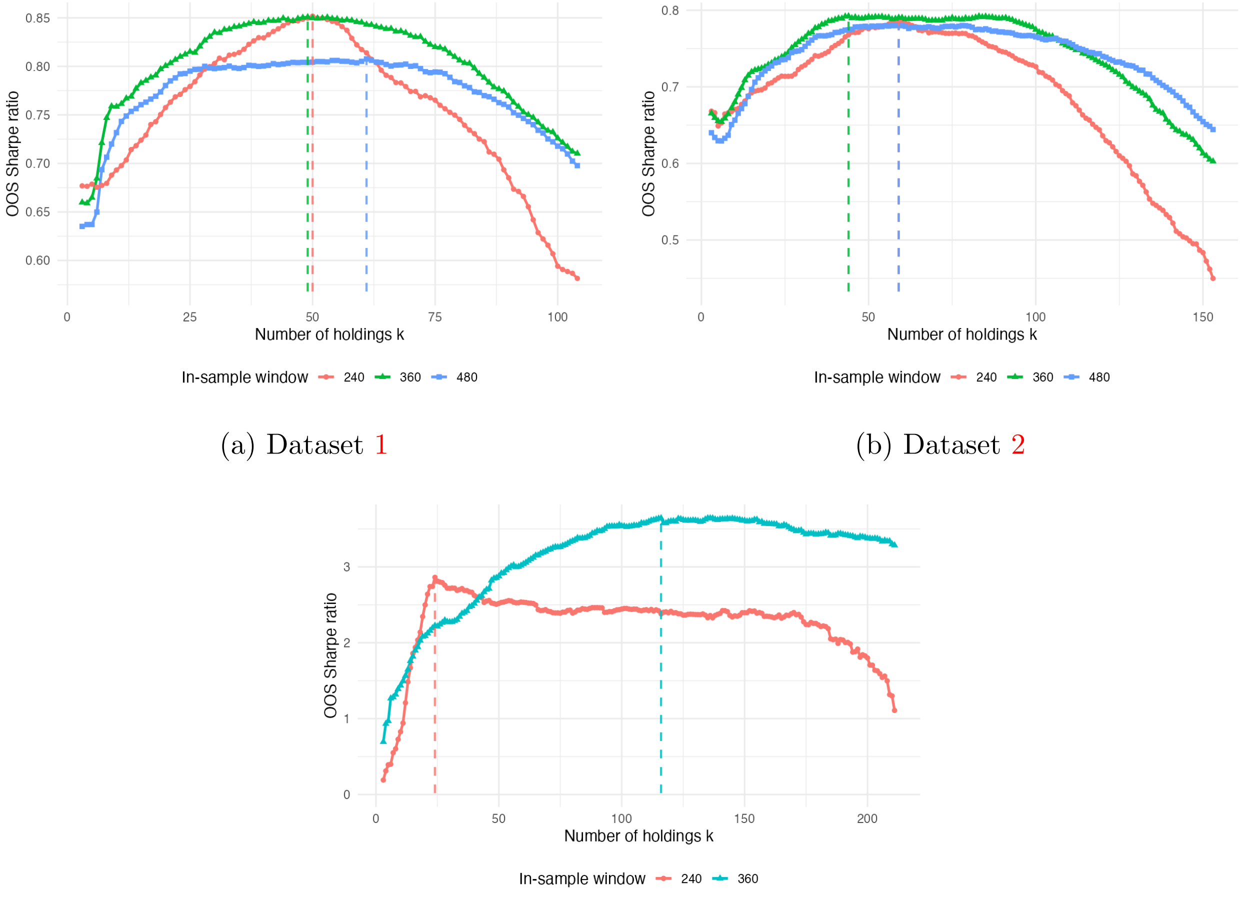


Figure 16: Median one-way turnover of the ℓ_1 portfolio rule as a function of the target sparsity level k for Dataset 3. The holding period is $h = 1$.

B.2 Robustness Checks

This subsection reports the robustness checks of out-of-sample Sharpe-ratio profiles for the datasets omitted from the main text robustness discussion (cf. Section 6.3.2). Specifically, we display results for the size–value benchmark (Dataset 1), the size–value–industry universe (Dataset 2), and the international monthly managed-portfolio universe (Dataset 4).

Figure 17 augments each universe with the corresponding Fama–French factor-mimicking portfolios, while Figure 18 reports the same analysis for the original datasets after normalizing realized portfolio weights to satisfy $\|w\|_1 = 1$. In both cases, the qualitative patterns mirror those in the main text.

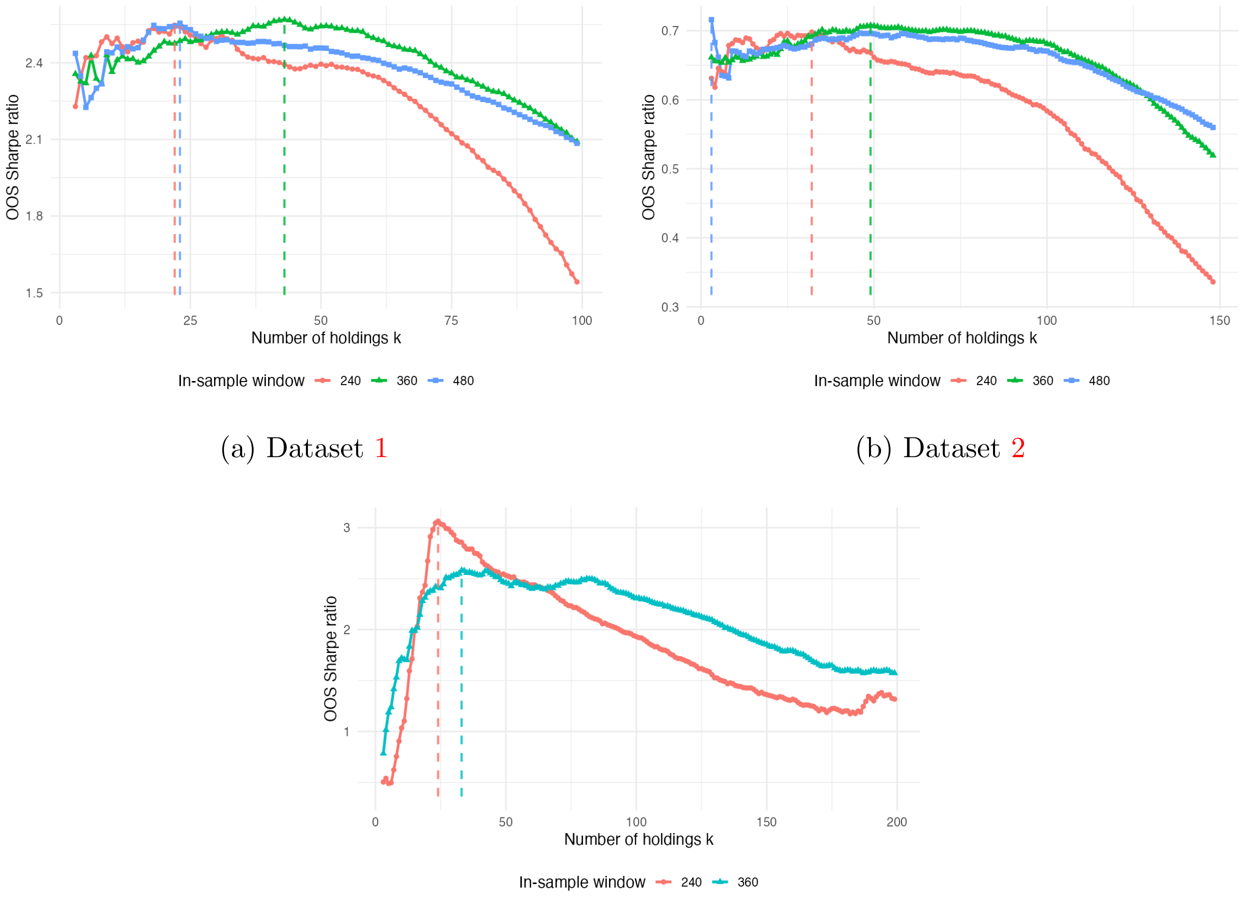


(a) Dataset 1

(b) Dataset 2

(c) Dataset 4

Figure 17: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule, plotted as a function of the target sparsity level k for the monthly datasets augmented with Fama–French factor-mimicking portfolios. The holding period is $h = 1$.



(a) Dataset 1

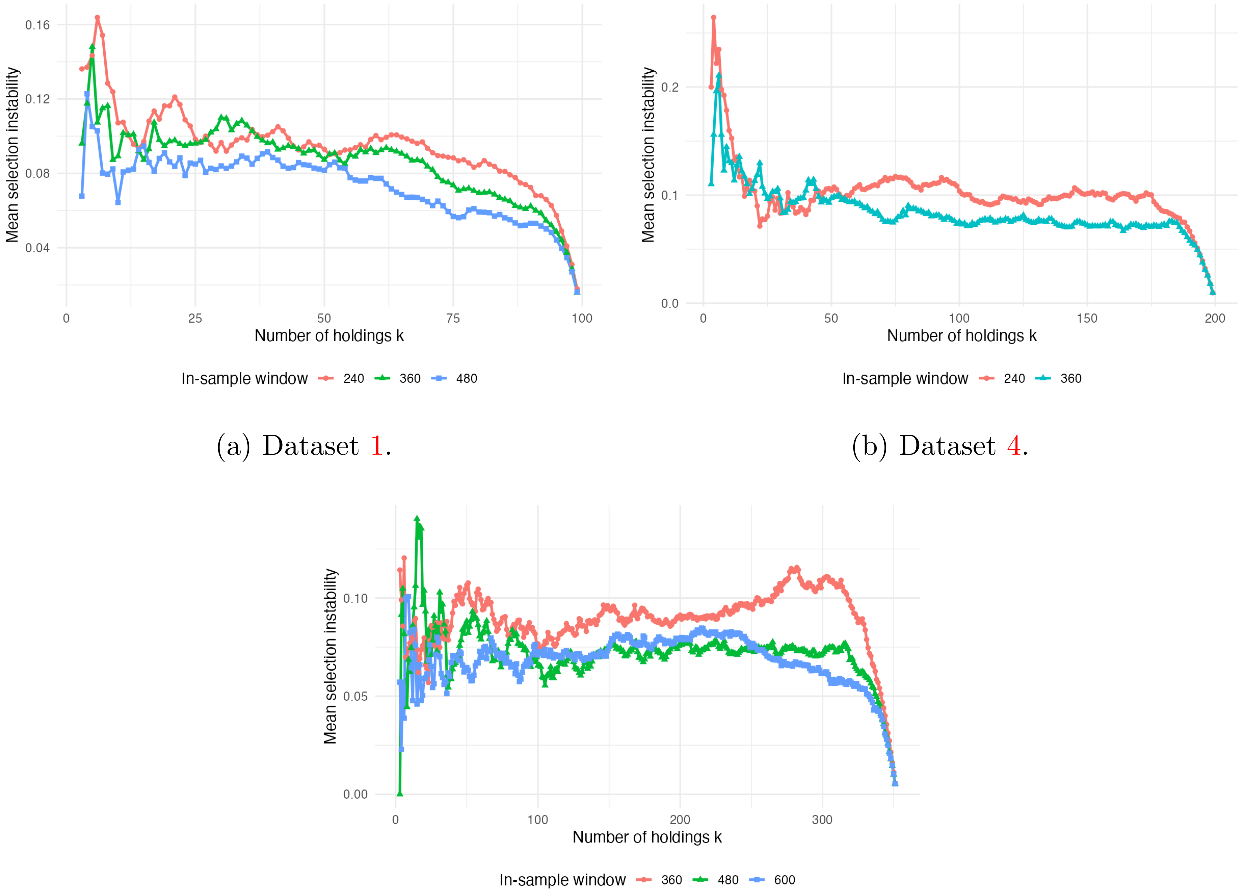
(b) Dataset 2

(c) Dataset 4

Figure 18: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule, plotted as a function of the target sparsity level k for the monthly datasets, with realized portfolio weights normalized so that $\|w\|_1 = 1$. The holding period is $h = 1$.

B.3 Selection and Weight Instability for the Remaining Datasets

We now report the selection- and weight-instability diagnostics for the datasets omitted from the main text presentation in Section 6.3.3. In particular, we display results for the size–value benchmark (Dataset 1), the international monthly managed-portfolio universe (Dataset 4), and the U.S. monthly managed-portfolio universe (Dataset 5). The qualitative patterns mirror those in the main text.

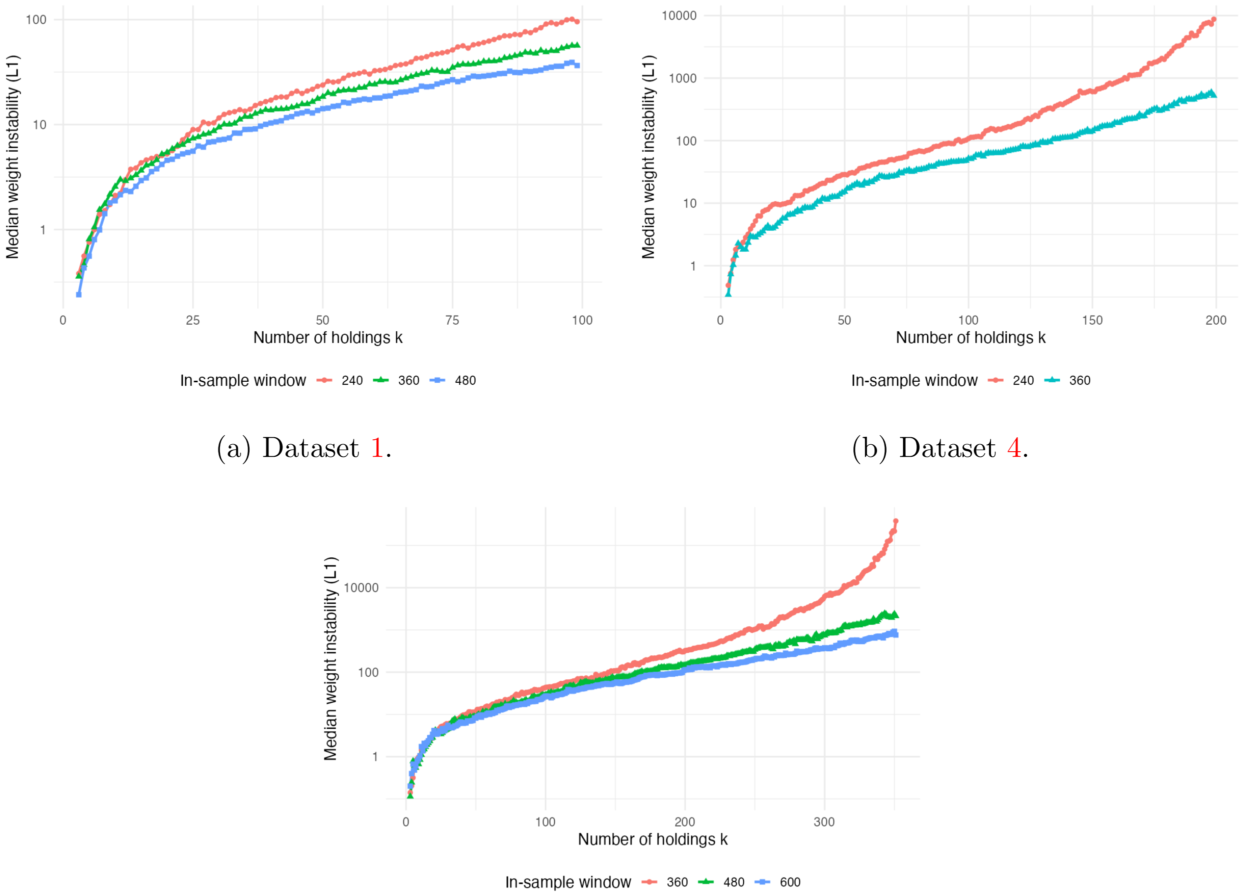


(a) Dataset 1.

(b) Dataset 4.

(c) Dataset 5.

Figure 19: Selection instability of the ℓ_1 portfolio rule as a function of the target sparsity level k for the monthly datasets omitted from the instability figures in the main text. The holding period is $h = 1$.



(a) Dataset 1.

(b) Dataset 4.

(c) Dataset 5.

Figure 20: Weight instability (conditional on the selected support) of the ℓ_1 portfolio rule as a function of k for the monthly datasets omitted from the main-text instability figures. Reported on a log scale and summarized by the time-series median. The holding period is $h = 1$.

B.3.1 Feasible Tuning Rules: Robustness Checks

We complement Section 6.3.4 by reporting the remaining robustness checks for feasible tuning. Specifically, we repeat the annual-tuning exercise (i) after augmenting the investable universe with Fama–French factor-mimicking portfolios and (ii) after normalizing realized portfolio weights to satisfy $\|w\|_1 = 1$ (gross-exposure normalization). We report results for the size–value–industry universe (Dataset 2) and the international universe (Dataset 4), matching the datasets used in the main text.

Table 2 shows that the main conclusions are robust. First, making standard factor payoffs explicitly tradable typically raises the level of attainable out-of-sample performance and leaves the relative ranking of tuning rules largely unchanged: both LOO and UE continue to deliver Sharpe ratios close to the infeasible benchmark $\widehat{\theta}^{\text{oos}}_{\max}(T)$, with UE often narrowing the gap at longer windows. Second, imposing $\|w\|_1 = 1$ does not materially alter the overall message that feasible tuning can achieve high out-of-sample Sharpe ratios. In particular, the normalization changes the scale of the implemented weights but preserves the ability of both rules to select an interior cardinality that performs well out-of-sample.

Figures 21 and 22 report the corresponding boxplots of annually selected sparsity levels. As in the baseline results, LOO tends to select substantially sparser portfolios than UE across years and window lengths. This pattern persists under both robustness designs (factor augmentation and ℓ_1 normalization), indicating that it reflects a systematic difference in how the two criteria trade off within-window fit against complexity and instability, rather than an artifact of the particular portfolio scale or the omission of benchmark factor payoffs.

C Numerical Algorithms

This appendix summarizes the numerical procedures used to approximate the k–sparse sample optimizer in (2). Because (2) is nonconvex and NP–hard, we rely on two complementary approaches. Our default method is the ℓ_1 relaxation (3), which scales reliably and delivers a full solution path. When problem sizes permit, we also solve a mixed-integer quadratic

Table 2: **Feasible tuning rules: robustness checks (annual tuning, monthly data).** For each in-sample window length T_{in}, the table reports the infeasible best-in-hindsight Sharpe ratio $\widehat{\theta}^{\text{oos}}_{\text{max}}(T)$ (max over k), and the realized out-of-sample Sharpe ratios under annual tuning of k via leave-one-out (LOO) and the unbiased-estimator rule (UE). All Sharpe ratios are annualized. Panels vary whether factor-mimicking portfolios are included and whether realized weights are normalized to satisfy $\|w\|_1 = 1$.

(a) Dataset 2: $\|w\|_1 = 1$, no factors

T_{in}	$\widehat{\theta}^{\text{oos}}_{\text{max}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	0.6980	0.5863	0.6368
360	0.7075	0.6372	0.7055
480	0.7160	0.6616	0.7049

(b) Dataset 2: factors, no normalization

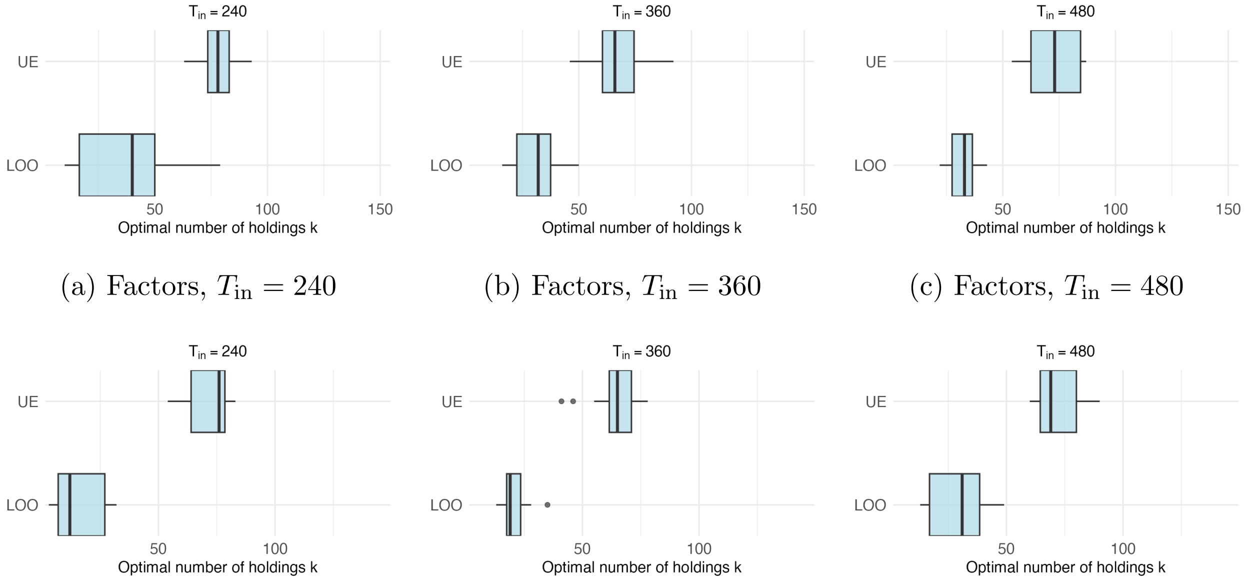
T_{in}	$\widehat{\theta}^{\text{oos}}_{\text{max}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	0.7852	0.7278	0.7617
360	0.7922	0.7491	0.7744
480	0.7801	0.7443	0.7573

(c) Dataset 4: factors, no normalization

T_{in}	$\widehat{\theta}^{\text{oos}}_{\text{max}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	2.8618	1.4821	1.6794
360	3.6421	3.4332	3.2483

(d) Dataset 4: $\|w\|_1 = 1$, no factors

T_{in}	$\widehat{\theta}^{\text{oos}}_{\text{max}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	3.0642	2.6346	1.1360
360	2.5808	2.2653	2.2284



(a) Factors, $T_{\text{in}} = 240$ (b) Factors, $T_{\text{in}} = 360$ (c) Factors, $T_{\text{in}} = 480$

(d) $\|w\|_1 = 1$, $T_{\text{in}} = 240$ (e) $\|w\|_1 = 1$, $T_{\text{in}} = 360$ (f) $\|w\|_1 = 1$, $T_{\text{in}} = 480$

Figure 21: **Chosen sparsity under annual tuning (robustness): Dataset 2.** Boxplots of annually selected cardinalities under factor augmentation (top row) and under ℓ_1 gross-exposure normalization (bottom row).

program (MIQP) that enforces the cardinality constraint directly and serves as a benchmark.

C.1 Lasso Relaxation

Section 4.3 shows that (3) can be written as a lasso problem. Here we record the specific reparametrization used in our code and how we map the regularization path to a target sparsity level.

Let

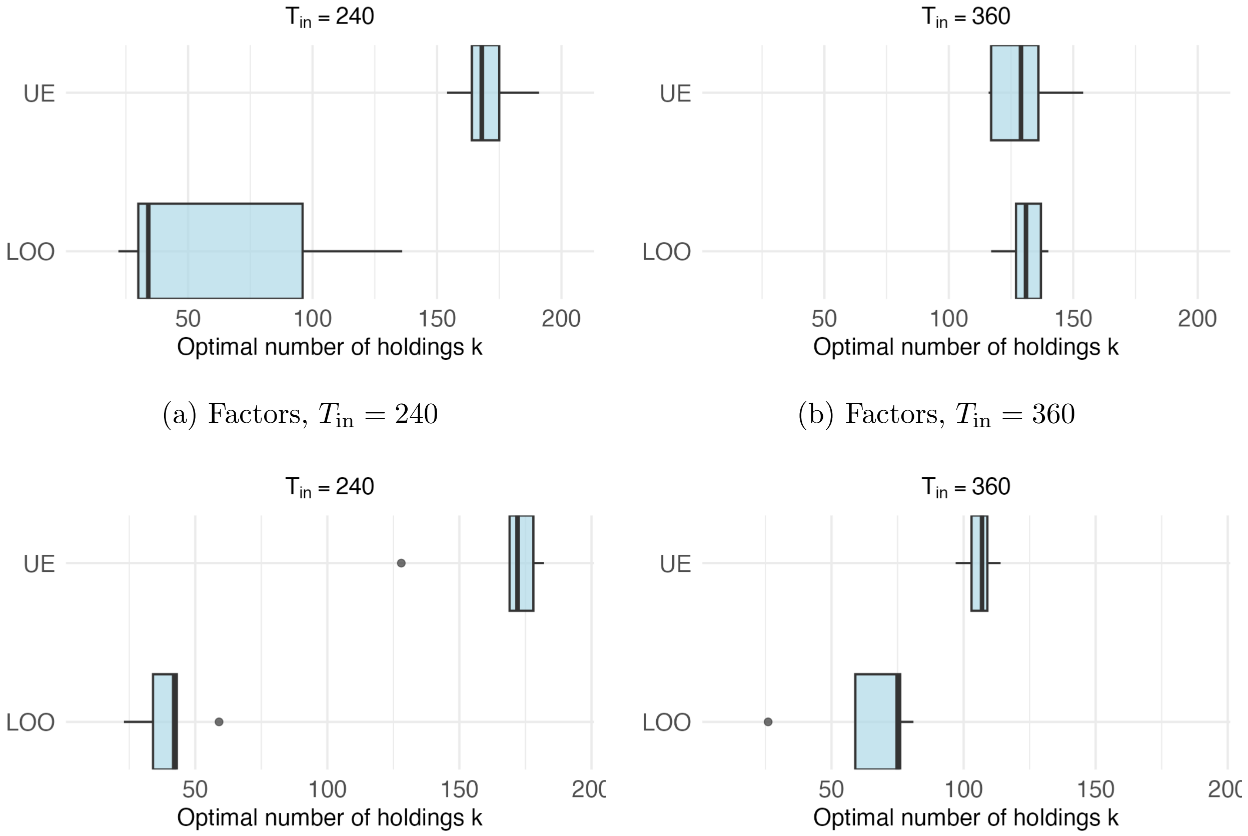
$$\widehat{U}_T(w) = w^\top \widehat{\mu} - \frac{\gamma}{2} w^\top \widehat{\Sigma} w, \qquad \widehat{M} := \gamma \widehat{\Sigma}.$$

Let C be the upper-triangular Cholesky factor of $\widehat{M}$, so that

$$C^\top C = \widehat{M}.$$

Define

$$X_T := \sqrt{T}\, C, \qquad y_T := \sqrt{T}\, C^{-\top} \widehat{\mu},$$



(a) Factors, $T_{\text{in}} = 240$

(b) Factors, $T_{\text{in}} = 360$

(c) $\|w\|_1 = 1$, $T_{\text{in}} = 240$

(d) $\|w\|_1 = 1$, $T_{\text{in}} = 360$

Figure 22: **Chosen sparsity under annual tuning (robustness): Dataset 4.** Boxplots of annually selected cardinalities under factor augmentation (top row) and under ℓ_1 gross-exposure normalization (bottom row).

where $C^{-\top}\widehat{\mu}$ is computed via a triangular solve. Then, up to an additive constant independent of w,

$$\widehat{U}_T(w) = -\frac{1}{2T}\|y_T - X_T w\|_2^2 + \text{const.}$$

Therefore, computing $\widehat{w}_{\ell_1,\lambda}$ in (3) is equivalent to solving the penalized least-squares problem

$$\operatorname*{argmin}_{w\in\mathbb{R}^N} \frac{1}{2T}\|y_T - X_T w\|_2^2 + \lambda\|w\|_1.$$

Path computation and selection of a target k. We compute the lasso path using the `lars` package in R (Efron et al., 2004). We pass (X_T, y_T) with no intercept and no additional standardization so that the algebraic equivalence above is preserved. The algorithm returns a sequence of knots at which the active set changes.

To select a portfolio with a target cardinality k, we scan the knots and compute the support size $\widehat{k}(\lambda) = \|\widehat{w}_{\ell_1,\lambda}\|_0$, where coefficients with $|w_i| \leq$ `tol` are treated as zero for a small numerical threshold `tol`. If the path contains a knot with $\widehat{k}(\lambda) = k$, we select the first such knot (largest penalty). If not, we select the knot with the largest $\widehat{k}(\lambda) \leq k$, breaking ties in favor of the later knot (which corresponds to a smaller penalty and hence less shrinkage). The resulting weight vector is then rescaled to match the desired implementation convention (e.g., $1^\top w = 1$ or a target gross exposure), as discussed in Section 4.4.

C.2 MIQP-based Approximation

As a benchmark, we also solve a mixed-integer quadratic program that enforces the cardinality constraint through binary inclusion variables. Using the same sample moments $(\widehat{\mu}, \widehat{\Sigma})$, we introduce portfolio weights $w \in \mathbb{R}^N$ and binary variables $v \in \{0,1\}^N$, where $v_i = 1$ indicates that asset i is active. We solve

$$\begin{aligned}\max_{w\in\mathbb{R}^N,\ v\in\{0,1\}^N} \quad & w^\top \widehat{\mu} - \frac{\gamma}{2} w^\top \widehat{\Sigma}\, w \\ \text{s.t.} \quad & \sum_{i=1}^N v_i \le k, \\ & f_{\min,i} v_i \le w_i \le f_{\max,i} v_i, \qquad i = 1, \dots, N, \\ & 1^\top w = 1.\end{aligned}$$

The linking constraints force $w_i = 0$ whenever $v_i = 0$ and impose position bounds $[f_{\min,i}, f_{\max,i}]$ whenever $v_i = 1$. We solve this MIQP using `Gurobi`. The solver is run with a prescribed relative optimality gap tolerance and a time limit that increases with problem size, so for larger instances the MIQP should be interpreted as a high-quality heuristic rather than a guaranteed global optimum.

Bound expansion and warm starts. To mitigate the possibility that the solution is driven by overly tight box constraints, we apply a simple bound-expansion heuristic. After an initial solve, we identify indices i for which the returned weight w_i is (numerically) at $f_{\min,i}$ or $f_{\max,i}$. We then expand the corresponding bound(s) by multiplying $f_{\min,i}$ and/or $f_{\max,i}$ by a factor larger than one and re-solve using the previous solution as a warm start. This expansion is iterated a small number of times or until no active bounds remain.

D Proofs

Lemma 1 (Selection rule)**.** *Let $\widehat{w}_k$ solve (2). Then*

$$\widehat{A}_k := \operatorname{supp}(\widehat{w}_k) \in \operatorname*{argmax}_{A\subset[N]:\ |A|\le k} \widehat{\theta}^2_A.$$

Proof of Lemma 1. For any subset $A \subset [N]$, define the coordinate subspace

$$\mathcal{W}(A) := \{w \in \mathbb{R}^N :\ \operatorname{supp}(w) \subset A\}.$$

Since every w with $\|w\|_0 \le k$ belongs to $\mathcal{W}(\operatorname{supp}(w))$ and $|\operatorname{supp}(w)| \le k$, we can write

$$\max_{\|w\|_0\le k} \widehat{U}_T(w) = \max_{A\subset[N]:\,|A|\le k} \ \max_{w\in\mathcal{W}(A)} \widehat{U}_T(w).$$

Fix $A \subset [N]$ with $|A| \le k$ and restrict $\widehat{U}_T$ to $\mathcal{W}(A)$. Writing $w = (w_A, 0)$, the objective becomes

$$\widehat{U}_T(w) = w_A^\top \widehat{\mu}_A - \frac{\gamma}{2} w_A^\top \widehat{\Sigma}_A w_A.$$

Whenever $\widehat{\Sigma}_A$ is nonsingular, this is a strictly concave quadratic function of w_A, with unique maximizer

$$\widehat{w}(A)_A = \frac{1}{\gamma} \widehat{\Sigma}_A^{-1} \widehat{\mu}_A, \qquad \widehat{w}(A)_{A^c} = 0,$$

and corresponding maximized value

$$\max_{w \in \mathcal{W}(A)} \widehat{U}_T(w) = \frac{1}{2\gamma}\, \widehat{\mu}_A^\top \widehat{\Sigma}_A^{-1} \widehat{\mu}_A = \frac{1}{2\gamma}\, \widehat{\theta}_A^2.$$

(The same conclusion holds if one restricts attention to those A for which $\widehat{\Sigma}_A$ is invertible; this is the case relevant for $\widehat{\theta}_A^2$ as defined above.) Therefore, maximizing $\widehat{U}_T$ over supports of size at most k is equivalent to maximizing $\widehat{\theta}_A^2$ over $A \subset [N]$ with $|A| \le k$.

Since $\widehat{w}_k$ attains the left-hand side and $\widehat{A}_k = \operatorname{supp}(\widehat{w}_k)$ is admissible, we must have

$$\widehat{A}_k \in \operatorname*{argmax}_{A \subset [N]:\, |A| \le k} \widehat{\theta}_A^2,$$

which is the claim. □

Proof of Proposition 1. Let

$$A_k^* \in \operatorname*{argmax}_{A \subset [N]:\, |A| \le k} \theta_A \qquad \text{so that} \qquad \theta_k^* = \theta_{A_k^*}.$$

Because $x \mapsto \sqrt{x}$ is strictly increasing on $[0, \infty)$, Lemma 1 implies that $\widehat{A}_k$ maximizes

$$\widehat{\theta}_A = \sqrt{\widehat{\mu}_A^\top \widehat{\Sigma}_A^{-1} \widehat{\mu}_A}$$

over $A \subset [N]$ with $|A| \le k$, whenever $\widehat{\Sigma}_A$ is invertible. Under $r_t \sim \mathcal{N}(\mu, \Sigma)$ i.i.d., $\widehat{\Sigma}_A$ is (a rescaled) Wishart matrix on $\mathbb{R}^m$ and is invertible almost surely as soon as $T - 1 \ge m$. In particular, for all large T with $T > k + 1$, $\widehat{\Sigma}_A$ is invertible almost surely for every A with $|A| \le k$. Hence $\widehat{\theta}_{A_k^*} \le \widehat{\theta}_{\widehat{A}_k}$. Together with $\theta_{A_k^*} \ge \theta_{\widehat{A}_k}$ this yields

$$\theta_k^* - \theta_{\widehat{A}_k} = \theta_{A_k^*} - \theta_{\widehat{A}_k} \le \left(\theta_{A_k^*} - \widehat{\theta}_{A_k^*}\right) + \left(\widehat{\theta}_{\widehat{A}_k} - \theta_{\widehat{A}_k}\right) \le 2 \sup_{A \subset [N]:\, |A| \le k} \left|\widehat{\theta}_A - \theta_A\right|.$$

It therefore suffices to show that the uniform deviation on the right-hand side is $O_p\big(\sqrt{k\log N/T}\big)$.

Note that

$$\widehat{\theta}_A^2 =_d \frac{U}{V}$$

for two independent random variable $U \sim \chi_k^2(T\theta_A^2)$ and $V \sim \chi_{T-k}^2$. We can write

$$U = U_1 + U_2 + T\theta_A^2$$

for some $U_1 \sim \chi_k^2$ and $U_2 \sim N(0, 4T\theta_A^2)$. Note that U_1 are U_2 are dependent. By χ^2 and Gaussian tail bounds, we have

$$\mathbb{P}[|U_1 - k| \geq xk] \leq 2\exp\big(-(x\wedge x^2)k/4\big)$$

and

$$\mathbb{P}[|U_2| \geq x] \leq 2\exp\big(-x^2/(8T\theta_A^2)\big)$$

for any $x \geq 0$. In other words, with probability at least $1-4e^{-x}$,

$$\left|\frac{U}{T} - \left(\theta_A^2 + \frac{k}{T}\right)\right| \leq \frac{k}{T}\left(\frac{4x}{k} \vee \sqrt{\frac{4x}{k}}\right) + \sqrt{\frac{8\theta_A^2 x}{T}}.$$

Similarly, with probability at least $1-2e^{-x}$,

$$\left|\frac{V}{T-k-2} - 1\right| \leq \frac{2}{T-k-2} + \frac{T-k}{T-k-2}\left(\frac{4x}{T-k} \vee \sqrt{\frac{4x}{T-k}}\right).$$

Taking $x = (1+\alpha)\log\binom{N}{k} < T$, we get

$$\begin{aligned}
|\widehat{\theta}_A^2 - \theta_A^2| &\leq \frac{4(1+\alpha)\log\binom{N}{k}}{T} + \sqrt{\frac{8(1+\alpha)\theta_A^2\log\binom{N}{k}}{T}} \\
&\quad +2\left(\theta_A^2 + \frac{k}{T}\right)\left(\frac{2}{T-k-2} + \frac{T-k}{T-k-2}\sqrt{\frac{4(1+\alpha)\log\binom{N}{k}}{T-k}}\right) \\
&\leq C(\theta_A^2+1)\sqrt{\frac{(1+\alpha)\log\binom{N}{k}}{T}} \\
&\leq C(\theta_{*,k}^2+1)\sqrt{\frac{(1+\alpha)\log\binom{N}{k}}{T}},
\end{aligned}$$

for some numerical constant $C > 0$, with probability at least $1 - 4\log\binom{N}{k}^{-(1+\alpha)}$. An application of union bounds to all A such that $|A| \le k$ yields

$$\max_{A\in[N]:|A|\le k} |\widehat{\theta}_A^2 - \theta_A^2| \le C(\theta_{*,k}^2 + 1)\sqrt{\frac{(1+\alpha)\log\binom{N}{k}}{T}}.$$

with probability at least

$$1 - 4\sum_{k\ge 1}\binom{N}{k}^{-\alpha}.$$

□

Proof of Proposition 2. Let $\widehat{A}_k := \operatorname{supp}(\widehat{w}_k)$. By optimality of $\widehat{w}_k$ in (2), conditional on the support $\widehat{A}_k$ the weights maximize $\widehat{U}_T$ over $\{w : \operatorname{supp}(w) \subset \widehat{A}_k\}$, hence coincide with the unconstrained mean–variance optimizer on that support, i.e.

$$\widehat{w}_k = \widehat{w}(\widehat{A}_k), \qquad \widehat{w}(A)_A = \frac{1}{\gamma}\widehat{\Sigma}_A^{-1}\widehat{\mu}_A, \quad \widehat{w}(A)_{A^c} = 0,$$

whenever $\widehat{\Sigma}_A$ is invertible. Under $r_t \sim \mathcal{N}(\mu, \Sigma)$ i.i.d., $\widehat{\Sigma}_A$ is invertible almost surely for all $|A| \le k$ once $T > k+1$, so the preceding display holds for all large T. Therefore

$$\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = \theta_{\widehat{A}_k} - \theta\big(\widehat{w}(\widehat{A}_k)\big) \le \sup_{A\subset[N]:\,|A|\le k}\Big\{\theta_A - \theta\big(\widehat{w}(A)\big)\Big\}.$$

For a fixed $A \subset [N]$ such that $|A| \le k$,

$$\theta_A^2 = \mu_A^\top \Sigma_A^{-1}\mu_A, \qquad \text{and} \qquad \theta^2(\widehat{w}_A) = \widehat{\mu}_A^\top \widehat{\Sigma}_A^{-1}\Sigma_A\widehat{\Sigma}_A^{-1}\widehat{\mu}_A.$$

Write $r_{t,A} = \mu_A + \Sigma_A^{1/2} Z_t$, then Z_ts are independent $N(0,1)$ random variables. Denote by $\bar{Z}$ and S the sample mean and covariance matrix of $Z_1, \ldots, Z_T$ respectively. Then

$$\widehat{\mu}_A = \mu_A + \Sigma_A^{1/2}\bar{Z}, \qquad \text{and} \qquad \widehat{\Sigma}_A = \Sigma_A^{1/2} S \Sigma_A^{1/2}.$$

Thus,

$$\theta^2(\widehat{w}_A) = \mu_A^\top \Sigma_A^{-1/2} S^{-2}\Sigma_A^{-1/2}\mu_A + 2\mu_A^\top\Sigma_A^{-1/2}S^{-2}\bar{Z} + \bar{Z}^\top S^{-2}\bar{Z}.$$

Observe that with probability at least $1 - N^{-(1+\alpha)k}$,

$$\|S - I\| \le C\sqrt{\frac{(1+\alpha)k\log N}{T}}.$$

for some constant $C > 0$ where $\|\cdot\|$ stands for the matrix spectral norm. See, e.g., Koltchinskii and Lounici (2017). Denote this event by $\mathcal{E}_1$. On the other hand, $T\bar{Z}^2 \sim \chi^2_T$. As before, with probability at least $1 - N^{-(1+\alpha)k}$,

$$\bar{Z}^2 \leq C\sqrt{\frac{(1+\alpha)k\log N}{T}}.$$

Denote this event $\mathcal{E}_2$. then under $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\theta_A - \theta(\widehat{w}_A) \leq C\sqrt{\frac{(1+\alpha)k\log N}{T}}.$$

An application of union bound then yields that

$$\max_{A\subset[N]:\,|A|\leq k}\big[\theta_A - \theta(\widehat{w}_A)\big] \leq C\sqrt{\frac{(1+\alpha)k\log N}{T}}.$$

with probability at least

$$1 - 2\sum_{k\geq 1}\binom{N}{k}N^{-(1+\alpha)k}.$$

□

Proof of Theorem 1. Let $\widehat{A}_k := \operatorname{supp}(\widehat{w}_k)$. By definition of θ^*_k and by $\theta_{\widehat{A}_k} = \max_{\operatorname{supp}(w)\subset\widehat{A}_k}\theta(w)$, we have the decomposition

$$\theta^*_k - \theta(\widehat{w}_k) = \big(\theta^*_k - \theta_{\widehat{A}_k}\big) + \big(\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\big).$$

The term $\theta^*_k - \theta_{\widehat{A}_k}$ is controlled by Proposition 1, which yields

$$\theta^*_k - \theta_{\widehat{A}_k} = O_p\left(\sqrt{\frac{k\log N}{T}}\right).$$

Under the same assumptions, the term $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)$ is controlled by Proposition 2, which yields

$$\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k\log N}{T}}\right).$$

Summing the two bounds and using that $O_p(a_T) + O_p(a_T) = O_p(a_T)$ with $a_T := \sqrt{k\log N/T}$ gives

$$\theta^*_k - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k\log N}{T}}\right),$$

which proves the theorem. □

Proof of Proposition 3. It follows immediately from Theorem 2. Under (12) we can write $\mu = B\mu_f$ and $\Sigma = B\Sigma_f B^\top + \Sigma_0$ with $L = 1$, $B = \beta \in \mathbb{R}^N$, $\mu_f = \mu_m$, $\Sigma_f = \sigma_m^2$, and $\Sigma_0 = \sigma_0^2 I_N$. Then

$$N^{-1}B^\top \Sigma_0^{-1} B = N^{-1}\beta^\top(\sigma_0^{-2} I_N)\beta = \sigma_0^{-2}\frac{\beta^\top\beta}{N} \ \to\ \Sigma_B > 0$$

by the assumption $\liminf_{N\to\infty} \beta^\top\beta/N > 0$, so the strong-factor condition holds with $\zeta = 1$. Moreover, Σ_f and Σ_0 have eigenvalues bounded away from 0 and ∞ by (12). The conclusion $\theta^* - \theta_k^* = O(1/k)$ follows from Theorem 2 with $\zeta = 1$. □

Proof of Theorem 2. Under Assumption 1, write

$$\mu = \mu^{(0)} + \alpha, \qquad \mu^{(0)} := B\mu_f, \qquad \Sigma = B\Sigma_f B^\top + \Sigma_0.$$

For any $A \subset [N]$, let $\mu_A, \alpha_A, B_A, \Sigma_A, \Sigma_{0,A}$ denote the corresponding subvectors and principal submatrices, and define

$$\theta_A := \sqrt{\mu_A^\top \Sigma_A^{-1}\mu_A}, \qquad \theta_A^{(0)} := \sqrt{(\mu_A^{(0)})^\top \Sigma_A^{-1}\mu_A^{(0)}}, \qquad \mu_A^{(0)} := B_A\mu_f.$$

Let $\theta^* = \theta_{[N]}$, $\theta_k^* = \max_{|A|\le k}\theta_A$, and similarly $\theta^{*,(0)} = \theta_{[N]}^{(0)}$ and $\theta_k^{*,(0)} = \max_{|A|\le k}\theta_A^{(0)}$.

Step 1 (perturbation in the mean). For any positive definite matrix M and vectors x, y, $\big|\|M^{1/2}x\|_2 - \|M^{1/2}y\|_2\big| \le \|M^{1/2}(x-y)\|_2$. Applying this with $M = \Sigma_A^{-1}$, $x = \mu_A$ and $y = \mu_A^{(0)}$ yields

$$|\theta_A - \theta_A^{(0)}| \le \sqrt{\alpha_A^\top \Sigma_A^{-1}\alpha_A}. \tag{18}$$

Since $\Sigma_A \succeq \Sigma_{0,A}$, inversion reverses the Loewner order and $\Sigma_A^{-1} \preceq \Sigma_{0,A}^{-1}$, hence

$$\alpha_A^\top\Sigma_A^{-1}\alpha_A \le \alpha_A^\top\Sigma_{0,A}^{-1}\alpha_A \le \alpha^\top\Sigma_0^{-1}\alpha.$$

Define $\Delta_N := \sqrt{\alpha^\top\Sigma_0^{-1}\alpha}$. Then (18) implies, uniformly over A,

$$\theta_A^{(0)} - \Delta_N \le \theta_A \le \theta_A^{(0)} + \Delta_N. \tag{19}$$

In particular,

$$\theta^* \le \theta^{*,(0)} + \Delta_N, \qquad \theta_k^* \ge \theta_k^{*,(0)} - \Delta_N, \qquad \Longrightarrow \qquad \theta^* - \theta_k^* \le \big(\theta^{*,(0)} - \theta_k^{*,(0)}\big) + 2\Delta_N. \tag{20}$$

Step 2 (the pricing-error term is negligible). By Assumption 2, $\|\Sigma_0\|_{\mathrm{op}} \le \bar{c}$, hence

$$\Delta_N^2 = (\Sigma_0^{-1}\alpha)^\top \Sigma_0 (\Sigma_0^{-1}\alpha) \le \|\Sigma_0\|_{\mathrm{op}} \, \|\Sigma_0^{-1}\alpha\|_2^2 \le \bar{c} \, \|\Sigma_0^{-1}\alpha\|_1^2.$$

Under Assumption 4, $\|\Sigma_0^{-1}\alpha\|_1 = O(1/N)$, so

$$\Delta_N = O(1/N). \tag{21}$$

Step 3 (efficiency loss for the factor-priced component). We now bound $\theta^{*,(0)} - \theta_k^{*,(0)}$ using the same argument as in the exact-factor case. Under Assumptions 1 and 2, for every A the matrix $\Sigma_A = B_A \Sigma_f B_A^\top + \Sigma_{0,A}$ is positive definite, so the Woodbury identity applies. For any $A \subset [N]$, define

$$H_A := B_A^\top \Sigma_{0,A}^{-1} B_A \in \mathbb{R}^{L \times L}, \qquad q := \mu_f^\top \Sigma_f^{-1} \mu_f.$$

The Woodbury identity yields

$$\Sigma_A^{-1} = \Sigma_{0,A}^{-1} - \Sigma_{0,A}^{-1} B_A \big(\Sigma_f^{-1} + H_A\big)^{-1} B_A^\top \Sigma_{0,A}^{-1},$$

and since $\mu_A^{(0)} = B_A \mu_f$,

$$\begin{aligned} (\theta_A^{(0)})^2 &= (\mu_A^{(0)})^\top \Sigma_A^{-1} \mu_A^{(0)} \\ &= \mu_f^\top \Big(H_A - H_A(\Sigma_f^{-1} + H_A)^{-1} H_A\Big) \mu_f = q - \mu_f^\top \Sigma_f^{-1} (\Sigma_f^{-1} + H_A)^{-1} \Sigma_f^{-1} \mu_f. \end{aligned}$$

Therefore $0 \le (\theta_A^{(0)})^2 \le q$, and since $\Sigma_f^{-1} + H_A \succeq H_A$,

$$(\Sigma_f^{-1} + H_A)^{-1} \preceq H_A^{-1},$$

whence

$$0 \le q - (\theta_A^{(0)})^2 \le \mu_f^\top \Sigma_f^{-1} H_A^{-1} \Sigma_f^{-1} \mu_f \le \frac{\|\Sigma_f^{-1}\mu_f\|_2^2}{\lambda_{\min}(H_A)}. \tag{22}$$

We first apply this to $A = [N]$. Writing $H := H_{[N]} = B^\top \Sigma_0^{-1} B$, Assumption 3 implies that $N^{-\zeta} H \to \Sigma_B \succ 0$, hence for all sufficiently large N,

$$\lambda_{\min}(H) \ge \tfrac{1}{2} \, \lambda_{\min}(\Sigma_B) \, N^{\zeta}. \tag{23}$$

Combining (22) with (23) yields

$$0 \le q - (\theta^{*,(0)})^2 = O(N^{-\zeta}). \tag{24}$$

Next, as in the exact-factor proof, fix $k \geq 1$ and choose $\widetilde{A} \subset [N]$ with $|\widetilde{A}| = k$ maximizing $\det(B_A^\top B_A)$ over all $|A| = k$. Theorem 1 in De Hoog and Mattheij (2007) implies that for $k \geq L$,

$$\lambda_{\min}\big(B_{\widetilde{A}}^\top B_{\widetilde{A}}\big) \geq \frac{k}{LN}\, \lambda_{\min}(B^\top B). \tag{25}$$

By Assumption 2, $\lambda_{\max}(\Sigma_{0,A}) \leq \lambda_{\max}(\Sigma_0)$ for all A, hence $\Sigma_{0,A}^{-1} \succeq \lambda_{\max}(\Sigma_0)^{-1} I_{|A|}$ and therefore

$$H_A = B_A^\top \Sigma_{0,A}^{-1} B_A \succeq \lambda_{\max}(\Sigma_0)^{-1} B_A^\top B_A.$$

It follows from (25) that

$$\lambda_{\min}(H_{\widetilde{A}}) \geq \frac{1}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}\big(B_{\widetilde{A}}^\top B_{\widetilde{A}}\big) \geq \frac{k}{LN}\, \frac{1}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}(B^\top B). \tag{26}$$

Moreover, $\Sigma_0^{-1} \preceq \lambda_{\min}(\Sigma_0)^{-1} I_N$ implies $H = B^\top \Sigma_0^{-1} B \preceq \lambda_{\min}(\Sigma_0)^{-1} B^\top B$, hence $\lambda_{\min}(B^\top B) \geq \lambda_{\min}(\Sigma_0)\lambda_{\min}(H)$. Substituting into (26) gives

$$\lambda_{\min}(H_{\widetilde{A}}) \geq \frac{k}{LN}\, \frac{\lambda_{\min}(\Sigma_0)}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}(H). \tag{27}$$

Combining (27) with (23) yields $\lambda_{\min}(H_{\widetilde{A}}) \geq c\, k\, N^{\zeta - 1}$ for all sufficiently large N, for a constant $c > 0$ depending only on L, Σ_B, and the eigenvalue bounds in Assumption 2. Plugging this into (22) gives

$$0 \leq q - (\theta_{\widetilde{A}}^{(0)})^2 = O\left(\frac{N^{1-\zeta}}{k}\right). \tag{28}$$

Since $\theta_k^{*,(0)} = \max_{|A| \leq k} \theta_A^{(0)} \geq \theta_{\widetilde{A}}^{(0)}$, we have $q - (\theta_k^{*,(0)})^2 \leq q - (\theta_{\widetilde{A}}^{(0)})^2$, hence

$$0 \leq q - (\theta_k^{*,(0)})^2 = O\left(\frac{N^{1-\zeta}}{k}\right).$$

Therefore,

$$(\theta^{*,(0)})^2 - (\theta_k^{*,(0)})^2 \leq q - (\theta_k^{*,(0)})^2 = O\left(\frac{N^{1-\zeta}}{k}\right).$$

If $q = 0$, then $\mu^{(0)} = 0$ and $\theta^{*,(0)} = \theta_k^{*,(0)} = 0$, so $\theta^{*,(0)} - \theta_k^{*,(0)} = 0$. Otherwise $q > 0$, and (24) implies that $\theta^{*,(0)}$ is bounded away from 0 for all sufficiently large N, hence

$$\theta^{*,(0)} - \theta_k^{*,(0)} = \frac{(\theta^{*,(0)})^2 - (\theta_k^{*,(0)})^2}{\theta^{*,(0)} + \theta_k^{*,(0)}} = O\left(\frac{N^{1-\zeta}}{k}\right). \tag{29}$$

Step 4 (combine the bounds). Substituting (29) and (21) into (20) gives

$$\theta^* - \theta_k^* = O\left(\frac{N^{1-\zeta}}{k}\right) + O(1/N).$$

Since $k \le N$ and $\zeta \in (0,1]$, we have $1/N \le N^{1-\zeta}/k$, so $O(1/N)$ is absorbed into $O(N^{1-\zeta}/k)$. This proves the theorem. In the strong-factor case $\zeta = 1$, the bound reduces to $O(1/k)$. □

Proof of Theorem 3. By the definition of θ_k^* and the estimator $\widehat{w}_k$, we have the identity

$$\theta^* - \theta(\widehat{w}_k) = \big(\theta^* - \theta_k^*\big) + \big(\theta_k^* - \theta(\widehat{w}_k)\big).$$

The first term is the efficiency loss $\Delta_k = \theta^* - \theta_k^*$, which is deterministic conditional on (μ, Σ) and is controlled by Theorem 2. Under the assumptions of that theorem (in particular Assumption 4 in the presence of pricing errors),

$$\theta^* - \theta_k^* = O\left(\frac{N^{1-\zeta}}{k}\right).$$

The second term is the estimation loss at sparsity level k, which is controlled by Theorem 1:

$$\theta_k^* - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}}\right).$$

Combining these bounds yields

$$\theta^* - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}} + \frac{N^{1-\zeta}}{k}\right).$$

For the final claim, assume $N^{1-\zeta} \ll k \ll T/\log N$. Then $N^{1-\zeta}/k \to 0$ and $k \log N/T \to 0$, hence both terms inside the $O_p(\cdot)$ rate converge to zero. Therefore $\theta^* - \theta(\widehat{w}_k) \xrightarrow{p} 0$, i.e. $\theta(\widehat{w}_k) \xrightarrow{p} \theta^*$. □

Proof of Theorem 4. Write $\widehat{w} := \widehat{w}_{\ell_1,\lambda}$ and let $w_* := (1/\gamma)\Sigma^{-1}\mu$, so that $\theta(w_*) = \theta^*$. Set $\Delta := \widehat{w} - w_*$.

Step 1: bounding $\|w_*\|_1$. Under Assumption 1, define

$$K := \Sigma_f^{-1} + B^\top \Sigma_0^{-1} B.$$

By the Woodbury identity,

$$\Sigma^{-1} = \Sigma_0^{-1} - \Sigma_0^{-1} B K^{-1} B^\top \Sigma_0^{-1},$$

and hence

$$\Sigma^{-1}\mu = \Big(I - \Sigma_0^{-1} B K^{-1} B^\top\Big)\Sigma_0^{-1}\alpha \;+\; \Sigma_0^{-1} B K^{-1} \Sigma_f^{-1}\mu_f. \tag{30}$$

Because L is fixed, Assumption 3 implies $\lambda_{\min}(B^\top \Sigma_0^{-1} B) \asymp N^\zeta$, while Assumption 2 implies $\|\Sigma_f^{-1}\|_{\mathrm{op}} = O(1)$, so

$$\|K^{-1}\|_{\mathrm{op}} = O(N^{-\zeta}), \qquad \|K^{-1}\|_{1\to 1} = O(N^{-\zeta}),$$

the second bound using equivalence of matrix norms in fixed dimension. Moreover, Assumption 2 gives $\|\Sigma_0^{-1}\|_{\mathrm{op}} = O(1)$ and, by Assumption 3, $\|B\|_{\mathrm{op}}^2 = \lambda_{\max}(B^\top B) = O(N^\zeta)$, hence

$$\|\Sigma_0^{-1} B\|_{\mathrm{op}} = O(N^{\zeta/2}), \qquad \|\Sigma_0^{-1} B\|_{1\to 1} \le \sqrt{N}\,\|\Sigma_0^{-1} B\|_{\mathrm{op}} = O(N^{(1+\zeta)/2}).$$

Finally, since L is fixed, we use the standard bounded-loading regularity $\|B^\top\|_{1\to 1} = O(1)$.[8]

For the first term in (30), let $H := \Sigma_0^{-1} B K^{-1} B^\top$. Then

$$\|(I - H)\Sigma_0^{-1}\alpha\|_1 \le (1 + \|H\|_{1\to 1})\,\|\Sigma_0^{-1}\alpha\|_1.$$

Using $\|H\|_{1\to 1} \le \|\Sigma_0^{-1} B\|_{1\to 1} \|K^{-1}\|_{1\to 1} \|B^\top\|_{1\to 1}$ yields

$$\|H\|_{1\to 1} = O\big(N^{(1+\zeta)/2} \cdot N^{-\zeta} \cdot 1\big) = O\big(N^{(1-\zeta)/2}\big),$$

and therefore, by Assumption 5,

$$\|(I - H)\Sigma_0^{-1}\alpha\|_1 = O\big(N^{(1-\zeta)/2}\big). \tag{31}$$

For the second term in (30), use $\|x\|_1 \le \sqrt{N}\|x\|_2$ and obtain

$$\|\Sigma_0^{-1} B K^{-1} \Sigma_f^{-1}\mu_f\|_1 \le \sqrt{N}\,\|\Sigma_0^{-1} B\|_{\mathrm{op}}\,\|K^{-1}\|_{\mathrm{op}}\,\|\Sigma_f^{-1}\mu_f\|_2 = O\big(N^{(1-\zeta)/2}\big),$$

since $\|\Sigma_f^{-1}\mu_f\|_2 = O(1)$ in fixed dimension. Combining with (31) gives

$$\|\Sigma^{-1}\mu\|_1 = O\big(N^{(1-\zeta)/2}\big),$$

hence

$$\|w_*\|_1 = \frac{1}{\gamma}\|\Sigma^{-1}\mu\|_1 = O\big(N^{(1-\zeta)/2}\big). \tag{32}$$

[8] Equivalently, $\max_{i\le N} \sum_{\ell=1}^L |B_{i\ell}| = O(1)$.

Step 2: basic inequality and ℓ_1 control. By optimality of $\widehat{w}$ for the ℓ_1-penalized sample objective,

$$\widehat{\mu}^\top \widehat{w} - \frac{\gamma}{2}\widehat{w}^\top \widehat{\Sigma}\widehat{w} - \lambda\|\widehat{w}\|_1 \geq \widehat{\mu}^\top w_* - \frac{\gamma}{2}w_*^\top \widehat{\Sigma} w_* - \lambda\|w_*\|_1.$$

Expanding the quadratic term yields

$$\frac{\gamma}{2}\Delta^\top \widehat{\Sigma}\Delta \leq \Delta^\top\big(\widehat{\mu} - \gamma\widehat{\Sigma}w_*\big) + \lambda(\|w_*\|_1 - \|\widehat{w}\|_1). \tag{33}$$

Let $S_T := \widehat{\mu} - \gamma\widehat{\Sigma}w_*$. Under Gaussian sampling and Assumption 2, Bernstein's inequality and a union bound imply

$$\|S_T\|_\infty = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{34}$$

Let $\mathcal{E}_T := \{\|S_T\|_\infty \leq \lambda/2\}$. By (34), there exists a numerical constant $C_0 > 0$ such that $\mathbb{P}(\mathcal{E}_T) \to 1$ whenever $\lambda \geq C_0\sqrt{\log N/T}$. On $\mathcal{E}_T$, Hölder's inequality in (33) gives

$$\frac{\gamma}{2}\Delta^\top \widehat{\Sigma}\Delta \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|w_*\|_1 - \|\widehat{w}\|_1).$$

Since $\Delta^\top \widehat{\Sigma}\Delta \geq 0$ and $\|\Delta\|_1 \leq \|\widehat{w}\|_1 + \|w_*\|_1$, we obtain

$$\|\widehat{w}\|_1 \leq 3\|w_*\|_1, \qquad \|\Delta\|_1 \leq 4\|w_*\|_1, \qquad \Delta^\top \widehat{\Sigma}\Delta \leq \frac{3\lambda}{\gamma}\|w_*\|_1. \tag{35}$$

Step 3: bounding $\Delta^\top\Sigma\Delta$. For any $x \in \mathbb{R}^N$, $|x^\top(\widehat{\Sigma} - \Sigma)x| \leq \|\widehat{\Sigma} - \Sigma\|_{\max}\|x\|_1^2$, hence

$$\Delta^\top\Sigma\Delta \leq \Delta^\top\widehat{\Sigma}\Delta + \|\widehat{\Sigma} - \Sigma\|_{\max}\|\Delta\|_1^2.$$

Under Gaussian sampling and Assumption 2, Bernstein's inequality and a union bound yield

$$\|\widehat{\Sigma} - \Sigma\|_{\max} = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{36}$$

On $\mathcal{E}_T$, combining (35)–(36) and using $\lambda \geq C_0\sqrt{\log N/T}$ gives

$$\Delta^\top\Sigma\Delta = O_p\Big(\lambda\|w_*\|_1 + \lambda\|w_*\|_1^2\Big) = O_p\big(\lambda\|w_*\|_1^2\big),$$

and together with (32),

$$\Delta^\top\Sigma\Delta = O_p\big(N^{1-\zeta}\lambda\big). \tag{37}$$

Step 4: translating to Sharpe-ratio loss. The identity

$$\theta^{*2} - \theta(w)^2 = \inf_{a \in \mathbb{R}} (aw - \Sigma^{-1}\mu)^\top \Sigma (aw - \Sigma^{-1}\mu)$$

implies, by choosing $a = \gamma$ and using $\Sigma^{-1}\mu = \gamma w_*$,

$$\theta^{*2} - \theta(\widehat{w})^2 \le (\gamma\widehat{w} - \Sigma^{-1}\mu)^\top \Sigma (\gamma\widehat{w} - \Sigma^{-1}\mu) = \gamma^2 \Delta^\top \Sigma \Delta.$$

Combining with (37) yields $\theta^{*2} - \theta(\widehat{w})^2 = O_p(N^{1-\zeta}\lambda)$. If $\theta^* = 0$ the claim is trivial. Otherwise, since $\theta(\widehat{w}) \le \theta^*$,

$$\theta^* - \theta(\widehat{w}) = \frac{\theta^{*2} - \theta(\widehat{w})^2}{\theta^* + \theta(\widehat{w})} \le \frac{\theta^{*2} - \theta(\widehat{w})^2}{\theta^*} = O_p\big(N^{1-\zeta}\lambda\big)\,,$$

which is the first statement.

For the final statement, take $\lambda = C\sqrt{\log N/T}$ with C sufficiently large. Then

$$\theta^* - \theta(\widehat{w}_{\ell_1,\lambda}) = O_p\left(N^{1-\zeta}\sqrt{\frac{\log N}{T}}\right)\,,$$

which converges to 0 when $T \gg N^{2(1-\zeta)}\log N$. Hence $\theta(\widehat{w}_{\ell_1,\lambda}) \xrightarrow{p} \theta^*$. □

E Estimation Loss with Time Dependence and Beyond Gaussian

This appendix extends Proposition 1, Proposition 2, and Theorem 1, which are stated under i.i.d. Gaussian sampling, to time-dependent and general sub-Gaussian returns. We show that, without the Gaussian assumption and under weak serial dependence formalized through α-mixing, the conclusion of Theorem 1 does not change: the Sharpe-ratio loss due to estimation risk remains of order $\sqrt{k\log N/T}$.

Recall that a random vector $r \in \mathbb{R}^N$ is sub-Gaussian if there exists $\sigma \in \mathbb{R}$ so that

$$\mathbb{E}\exp(v^\top(r - \mathbb{E}r)) \le \exp(\|v\|_2^2\sigma^2/2), \qquad \forall v \in \mathbb{R}^N.$$

Let $\mathcal{F}_s^t := \sigma(r_u : \; s \le u \le t)$ denote the σ-algebra generated by the return realizations $\{r_u : \; s \le u \le t\}$, and define the *$\alpha$-mixing* coefficients of the strictly stationary process $\{r_t\}_{t\in\mathbb{Z}}$ by

$$\alpha(\ell) := \sup_{t\in\mathbb{Z}} \sup_{A\in\mathcal{F}_{-\infty}^t,\; B\in\mathcal{F}_{t+\ell}^{\infty}} \big|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|, \qquad \ell \ge 0.$$

The process is α-mixing if $\alpha(\ell) \to 0$ as $\ell \to \infty$. Throughout this appendix we impose a mild quantitative version of this condition.

Assumption 6 (sub-Gaussianity and weak dependence)**.** *The return sequence $\{r_t\}_{t\in\mathbb{Z}}$ is strictly stationary with $\mathbb{E}[r_t] = \mu \in \mathbb{R}^N$ and $\mathbb{V}(r_t) = \Sigma \in \mathbb{R}^{N\times N}$ positive definite, such that $\Sigma^{-1/2}(r_t - \mu)$ is sub-Gaussian. Moreover, the α-mixing coefficients satisfy a geometric bound*

$$\alpha(\ell) \le C_\alpha \rho_\alpha^\ell \qquad \textit{for some } C_\alpha > 0,\; \rho_\alpha \in (0,1).$$

Under Assumption 6, the serial dependence is sufficiently weak that the effective sample size remains proportional to T.

E.1 Results under weak dependence

The selection argument in Proposition 1 relies on uniform control of $\widehat{\theta}_A - \theta_A$ over all $|A| \le k$. Under mixing, the same structure applies once fixed-support moment concentration is established under weak dependence.

Proposition 4 (Bound on selection risk under time dependence)**.** *Let Assumption 6 hold and $\theta^* = \sqrt{\mu^\top\Sigma^{-1}\mu} < \infty$. Then*

$$\theta_k^* - \theta_{\widehat{A}_k} = O_p\left(\sqrt{\frac{k\log N}{T}}\right), \qquad \textit{as } T \to \infty.$$

The allocation argument in Proposition 2 compares the population Sharpe ratio on a selected support to the population Sharpe ratio of the plug-in weights on that same support. Under mixing, the same deterministic inequalities apply and only the fixed-support concentration changes.

Proposition 5 (Bound on allocation risk under time dependence)**.** *Let Assumption 6 hold and $\theta^* = \sqrt{\mu^\top \Sigma^{-1}\mu} < \infty$. Then*

$$\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}}\right), \qquad \text{as } T \to \infty.$$

Combining the mixing analogues of the selection- and allocation-risk bounds with the decomposition (10) yields the time-dependent counterpart of Theorem 1.

Theorem 5 (Estimation risk of sparse portfolios under time dependence)**.** *Let Assumption 6 hold, and suppose $\theta^* = \sqrt{\mu^\top \Sigma^{-1}\mu} < \infty$. Then the k–sparse plug-in portfolio $\widehat{w}_k$ computed from $(\widehat{\mu}, \widehat{\Sigma})$ satisfies*

$$\theta_k^* - \theta(\widehat{w}_k) = O_p\left(\sqrt{\frac{k \log N}{T}}\right), \qquad \text{as } T \to \infty.$$

Theorem 5 has the same $\sqrt{k \log N/T}$ scaling as the i.i.d. benchmark, Theorem 1. Time dependence affects only the constants hidden in the $O_p(\cdot)$ term, through the strength and persistence of serial correlation summarized by the α-mixing rate. Economically, weak dependence reduces the amount of independent information in a fixed-length window, but under geometric mixing this reduction is bounded and does not change the rate at which estimation risk grows with portfolio complexity k and search breadth N.

E.2 Proofs

We first develop a fixed-support concentration result for $(\widehat{\mu}_A, \widehat{\Sigma}_A)$ under mixing, which replaces the Gaussian/Wishart bounds used in the i.i.d. proofs. We then prove the selection- and allocation-risk bounds under time dependence by repeating the i.i.d. arguments with this replacement. Finally, we conclude with the proof of Theorem 5.

We use Davydov's inequality. If U is $\mathcal{F}_{-\infty}^{t}$-measurable and V is $\mathcal{F}_{t+\ell}^{\infty}$-measurable, and if $\mathbb{E}[|U|^4] < \infty$ and $\mathbb{E}[|V|^4] < \infty$, then

$$|\mathrm{Cov}(U, V)| \leq C\, \alpha(\ell)^{1/2}\, \|U\|_4\, \|V\|_4, \qquad \ell \geq 0, \tag{38}$$

for a universal constant $C > 0$, where $\|W\|_4 := (\mathbb{E}[|W|^4])^{1/4}$. Under Assumption 6, sub-Gaussianity guarantees finite fourth moments for the linear and quadratic functionals we use below.

Lemma 2 (Moment concentration on a fixed support under mixing)**.** *Under Assumption 6, there exists a finite constant $C_{\mathrm{mix}} > 0$ depending only on (C_α, ρ_α) and on the conditioning of Σ such that the following holds. For any subset $A \subset [N]$ with $m := |A|$ and any $x \geq 0$, with probability at least $1 - 4e^{-x}$,*

$$\left\|\Sigma_A^{-1/2}(\widehat{\mu}_A - \mu_A)\right\|_2 \leq C_{\mathrm{mix}}\sqrt{\frac{m+x}{T}}, \tag{39}$$

$$\left\|\Sigma_A^{-1/2}\widehat{\Sigma}_A\Sigma_A^{-1/2} - I_m\right\|_{\mathrm{op}} \leq C_{\mathrm{mix}}\left(\sqrt{\frac{m+x}{T}} + \frac{m+x}{T}\right). \tag{40}$$

Proof. Fix $A \subset [N]$ with $m = |A|$ and define the standardized series

$$x_t := \Sigma_A^{-1/2}(r_{t,A} - \mu_A) \in \mathbb{R}^m, \qquad t = 1, \ldots, T.$$

Then $\mathbb{E}[x_t] = 0$ and $\mathbb{V}(x_t) = I_m$. Let $\bar{x} := T^{-1}\sum_{t=1}^T x_t$ and define

$$S := \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(x_t - \bar{x})^\top.$$

It is clear that $\widehat{\Sigma}_A = \Sigma_A^{1/2} S \Sigma_A^{1/2}$.

Step 1: bound the dependence strength in quadratic forms. For any unit vectors $u, v \in \mathbb{R}^m$ and any lag $\ell \geq 0$, $u^\top x_t$ is $\mathcal{F}_{-\infty}^t$-measurable and $v^\top x_{t+\ell}$ is $\mathcal{F}_{t+\ell}^\infty$-measurable, hence by (38),

$$\left|u^\top \mathbb{E}[x_t x_{t+\ell}^\top] v\right| = \left|\mathrm{Cov}(u^\top x_t, v^\top x_{t+\ell})\right| \leq C\,\alpha(\ell)^{1/2}\,\|u^\top x_t\|_4\,\|v^\top x_{t+\ell}\|_4.$$

Because $\mathbb{V}(u^\top x_t) = 1$ and $(u^\top x_t)$ is sub-Gaussian, $\|u^\top x_t\|_4$ is finite, and similarly for $v^\top x_{t+\ell}$. Absorbing constants yields

$$\|C_\ell\|_{\mathrm{op}} := \left\|\mathbb{E}[x_t x_{t+\ell}^\top]\right\|_{\mathrm{op}} \leq C\,\alpha(\ell)^{1/2}. \tag{41}$$

Define the finite dependence constant

$$M_\alpha := 1 + 2\sum_{\ell=1}^{\infty}\|C_\ell\|_{\mathrm{op}} \leq 1 + 2C\sum_{\ell=1}^{\infty}\alpha(\ell)^{1/2} < \infty,$$

where finiteness follows from geometric mixing in Assumption 6.

Step 2: sample mean concentration. Because $\{x_t\}$ is jointly Gaussian, $\sqrt{T}\,\bar{x}$ is Gaussian with covariance

$$\mathbb{V}(\sqrt{T}\,\bar{x}) = \sum_{\ell=-(T-1)}^{T-1} \left(1 - \frac{|\ell|}{T}\right) \mathbb{E}[x_1 x_{1+\ell}^\top] \preceq \left(1 + 2\sum_{\ell=1}^{T-1} \|C_\ell\|_{\mathrm{op}}\right) I_m \preceq M_\alpha I_m.$$

Hence $\sqrt{T}\,\bar{x}$ is sub-Gaussian with covariance proxy $M_\alpha I_m$, and standard sub-Gaussian norm concentration implies that for all $x \geq 0$,

$$\mathbb{P}\left(\|\bar{x}\|_2 \geq \sqrt{\frac{M_\alpha}{T}} \big(\sqrt{m} + \sqrt{2x}\big) \right) \leq e^{-cx},$$

for some constant $c > 0$. Absorbing numerical factors into the constant gives (39).

Step 3: second-moment concentration. Fix $u \in \mathbb{R}^m$ with $\|u\|_2 = 1$ and define $z_t := u^\top x_t$. Then $\{z_t\}_{t=1}^T$ is a mean-zero sub-Gaussian vector with $\mathbb{V}(z_t) = 1$. Let $z := (z_1, \ldots, z_T)^\top \in \mathbb{R}^T$ and let $R_u := \mathbb{E}[zz^\top]$ be its $T \times T$ covariance matrix. By (41), R_u is Toeplitz with 1 on the diagonal and $|(R_u)_{st}| = |\mathbb{E}[z_s z_t]| \leq \|C_{|s-t|}\|_{\mathrm{op}}$, which implies

$$\|R_u\|_{\mathrm{op}} \leq 1 + 2\sum_{\ell=1}^{T-1} \|C_\ell\|_{\mathrm{op}} \leq M_\alpha. \tag{42}$$

Moreover $\mathrm{trace}(R_u) = \sum_{t=1}^T \mathbb{V}(z_t) = T$.

Since $u^\top S u = T^{-1} \sum_{t=1}^T z_t^2 = T^{-1} z^\top z$, we control $|u^\top (S - I_m) u|$ via concentration of quadratic forms in sub-Gaussian vectors:

$$\mathbb{P}\left(\big|u^\top (S - I_m) u\big| \geq 2M_\alpha \sqrt{\frac{x}{T}} + 2M_\alpha \frac{x}{T} \right) \leq 2e^{-cx}, \tag{43}$$

for some constant $c > 0$.

Step 4: upgrade to operator norm via a net. Let $\mathcal{N}$ be a 1/4-net of the unit sphere in $\mathbb{R}^m$, so that $|\mathcal{N}| \leq 9^m$ and $\|S - I_m\|_{\mathrm{op}} \leq 2\sup_{u\in\mathcal{N}} |u^\top (S - I_m) u|$. Apply (43) with x replaced by $x + m\log 9$ and take a union bound over $u \in \mathcal{N}$ to obtain that with probability at least $1 - 2e^{-x}$,

$$\big\|\Sigma_A^{-1/2} \widehat{\Sigma}_A \Sigma_A^{-1/2} - I_m\big\|_{\mathrm{op}} = \|S - I_m\|_{\mathrm{op}} \leq C M_\alpha \left(\sqrt{\frac{m+x}{T}} + \frac{m+x}{T} \right)$$

for a constant $C > 0$. □

Lemma 2 replaces the fixed-support Gaussian/Wishart concentration used in the i.i.d. proofs of Propositions 1 and 2. Repeating those arguments verbatim with C_{mix} in place of the i.i.d. constant yields the same $\sqrt{k \log N/T}$ bounds for selection and allocation risk under Assumption 6. Combining them with the decomposition (10) proves Theorem 5.

Proof of Proposition 4. Repeat the proof of Proposition 1 up to the display

$$\theta_k^* - \theta_{\widehat{A}_k} \leq 2 \sup_{A \subset [N] \colon |A| \leq k} \big|\widehat{\theta}_A - \theta_A\big|.$$

Fix A with $m := |A| \leq k$. Let x_t, $\bar{x}$ and S be defined as in the proof of 2. Then

$$\widehat{\theta}_A^2 - \theta_A^2 = \mu_A^\top \Sigma_A^{-1/2}(S^{-1} - I_m)\Sigma_A^{-1/2}\mu_A + 2\mu_A^\top \Sigma_A^{-1/2} S^{-1}\bar{x} + \bar{x}^\top S^{-1}\bar{x}.$$

Lemma 2 implies that with probability at least $1 - 4e^{-x}$,

$$\big\|\bar{x} - \Sigma_A^{-1/2}\mu_A\big\|_2 \leq C_{\text{mix}}\sqrt{\frac{m+x}{T}}, \qquad \big\|S - I_m\big\|_{\text{op}} \leq C_{\text{mix}}\left(\sqrt{\frac{m+x}{T}} + \frac{m+x}{T}\right).$$

On the event in which the second bound is at most 1/2, the same Lipschitz argument for $M \mapsto M^{-1/2}$ yields

$$\big\|S^{-1/2} - I_m\big\|_{\text{op}} \leq C\,\big\|S - I_m\big\|_{\text{op}}/$$

Using the fact that $\theta_A = \|\mu_A^\top \Sigma_A^{-1}\mu_A\|_2 \leq \theta^* < \infty$, we obtain

$$|\widehat{\theta}_A - \theta_A| \leq C(1+\theta^*)\sqrt{\frac{m+x}{T}}$$

with probability at least $1 - 4e^{-x}$, where C depends only on C_{mix}.

Finally, union bound over $\{A : |A| \leq k\}$ exactly as in the i.i.d. proof. Choosing $x = (1+\alpha)m\log N$ and using $\binom{N}{m} \leq (eN/m)^m$ yields

$$\sup_{A \subset [N] \colon |A| \leq k} |\widehat{\theta}_A - \theta_A| = O_p\left(\sqrt{\frac{k \log N}{T}}\right),$$

and the claim follows. □

Proof of Proposition 5. Repeat the proof of Proposition 2 up to

$$\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) \leq \sup_{A \subset [N] \colon |A| \leq k} \Big\{\theta_A - \theta\big(\widehat{w}(A)\big)\Big\}.$$

Fix A with $m := |A| \le k$ and use the same notation $u_A := \Sigma_A^{-1/2}\mu_A$, $d_A := \Sigma_A^{-1/2}(\widehat{\mu}_A - \mu_A)$, and $G_A := \Sigma_A^{-1/2}\widehat{\Sigma}_A\Sigma_A^{-1/2}$. Define

$$z_A := \Sigma_A^{1/2}\widehat{\Sigma}_A^{-1}\widehat{\mu}_A,$$

so that $z_A = G_A^{-1}(u_A + d_A)$. The deterministic inequality $\theta_A - \theta(\widehat{w}(A)) \le 4\|z_A - u_A\|_2$ is unchanged.

On the event $\|G_A - I_m\|_{\mathrm{op}} \le 1/2$, the same bounds imply

$$\|z_A - u_A\|_2 \le 2\|d_A\|_2 + 2\|G_A - I_m\|_{\mathrm{op}}\,\|u_A\|_2.$$

Lemma 2 yields, with probability at least $1 - 4e^{-x}$,

$$\|d_A\|_2 \le C_{\mathrm{mix}}\sqrt{\frac{m+x}{T}}, \qquad \|G_A - I_m\|_{\mathrm{op}} \le C_{\mathrm{mix}}\left(\sqrt{\frac{m+x}{T}} + \frac{m+x}{T}\right).$$

Using $\|u_A\|_2 = \theta_A \le \theta^*$ and $\theta^* < \infty$, and taking T large so that the right-hand side is at most $1/2$, yields $\|G_A - I_m\|_{\mathrm{op}} \le 1/2$ on the same event and therefore

$$\theta_A - \theta\big(\widehat{w}(A)\big) \le C(1+\theta^*)\sqrt{\frac{m+x}{T}}$$

with probability at least $1 - 4e^{-x}$, where C depends only on C_{mix}.

Union bound over A with $|A| \le k$ proceeds exactly as in the i.i.d. case. Choosing $x = (1+\alpha)m\log N$ and using $\binom{N}{m} \le (eN/m)^m$ yields

$$\sup_{A\subset[N]:\,|A|\le k}\Big\{\theta_A - \theta\big(\widehat{w}(A)\big)\Big\} = O_p\left(\sqrt{\frac{k\log N}{T}}\right),$$

and the claim follows. □

With Propositions 4 and 5 in hand, the proof of Theorem 5 is identical to the i.i.d. argument.

Proof of Theorem 5. Let $\widehat{A}_k := \mathrm{supp}(\widehat{w}_k)$. As in the i.i.d. case,

$$\theta_k^* - \theta(\widehat{w}_k) = \big(\theta_k^* - \theta_{\widehat{A}_k}\big) + \big(\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)\big).$$

Under Assumption 6 and $\theta^* < \infty$, Proposition 4 gives $\theta_k^* - \theta_{\widehat{A}_k} = O_p(\sqrt{k\log N/T})$, and Proposition 5 gives $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = O_p(\sqrt{k\log N/T})$. Summing the two bounds yields the stated rate. □