The Estimation–Efficiency Frontier in Portfolio Complexity

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

*Wells Fargo, †Erasmus University and ‡Columbia University

May 27, 2026

Abstract

How many assets should a portfolio hold in large markets? We treat portfolio size as an endogenous choice in mean–variance allocation. Adding assets expands diversification, but also magnifies the cost of selecting holdings and learning weights from finite data. Active sets therefore balance marginal risk–return benefits against marginal estimation costs. Under factor structure, many securities are near-redundant claims on similar priced risks, so excluding assets can be inexpensive in population even without trading frictions. We formalize this tradeoff through an estimation–efficiency frontier that decomposes Sharpe-ratio losses into population spanning losses and finite-sample selection and allocation losses. The frontier implies diminishing returns to portfolio size and an interior optimum, yielding a scaling law for optimal cardinality in sample size, number of assets, and factor strength. The same frontier identifies an efficiency window in which sparse Markowitz plug-in portfolios are asymptotically efficient relative to the population mean–variance benchmark. Across managed portfolios, self-financing anomalies, and individual stocks, the estimation–efficiency tradeoff appears as hump-shaped out-of-sample Sharpe-ratio profiles in cardinality.

Address for correspondence: Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027.

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

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

1 Introduction

Modern portfolio theory (Markowitz, 1952) gives investors a simple and powerful prescription: diversify. When population moments are known, expanding the opportunity set can only improve the attainable mean–variance frontier. In empirical portfolio choice, however, diversification is not free. Each additional active security must be evaluated, selected, and weighted using a finite return history. When the investment universe is large relative to the available sample, a denser portfolio can therefore become a source of statistical risk rather than only a source of diversification. This paper studies the resulting problem of portfolio allocation under parameter uncertainty: how many securities should an investor actively use when each additional claim expands the payoff span but also increases the shadow cost of portfolio complexity?

We develop an estimation–efficiency frontier for active portfolio complexity. The frontier has two sides. The economic side is the gain from using more assets: a larger active set improves spanning and residual-risk diversification. The statistical side is the cost of learning a larger allocation rule from finite data: the investor must decide which assets to hold and how much to invest in each. The optimal number of active holdings is therefore endogenous. It is chosen inside the portfolio problem as a function of the sample size, the size of the investment universe, and the degree of redundancy in asset returns. The central implication is that the optimal portfolio is generally neither extremely concentrated nor fully dense. It holds enough assets to capture the relevant economic risks, but not so many that estimation error dominates out-of-sample performance.

The economic mechanism is cross-sectional redundancy. In a factor economy, many se-

curities are near-redundant claims on similar priced risks. The number of listed securities is therefore not the same as the number of economically distinct investment opportunities. This does not imply that the population-efficient portfolio is sparse. With known moments, the tangency portfolio is typically dense: broad holdings help fine-tune factor exposures and diversify idiosyncratic components. Sparsity is valuable for a different reason. In finite samples, a moderately small active set can preserve most of the population frontier while reducing the sampling error in means, covariances, and weights. The strength of this redundancy is summarized in our theory by a factor strength parameter ζ, which also has a natural economic interpretation as a market redundancy parameter: higher ζ means that systematic risks are more pervasively represented across securities, so fewer active claims are needed to span the relevant opportunities.

The empirical fragility of plug-in mean–variance optimization is well known. When N is large relative to T, estimates of expected returns and covariances become unreliable (e.g., Jobson and Korkie, 1989; Britten-Jones, 1999), and small perturbations in the inputs can generate large changes in efficient weights (Best and Grauer, 1991; Chopra et al., 1993). The resulting optimized frontiers can be misleading and unstable (Michaud, 1989). A large literature responds by regularizing either the inputs or the portfolio rule. Examples include shrinkage estimators for means and covariances (e.g., Ledoit and Wolf, 2003, 2004, 2017), Bayesian methods (e.g., Jorion, 1986; Tu and Zhou, 2010; Johannes et al., 2014), portfolio combinations (e.g., Tu and Zhou, 2011), and robust or ambiguity-aware formulations (e.g., Goldfarb and Iyengar, 2003; Tütüncü and Koenig, 2004; Garlappi et al., 2007). Another influential insight is that constraints that are "wrong" in population can improve realized performance when moments are estimated: for example, Jagannathan and Ma (2003) show that no-short-sale constraints can reduce estimation noise enough to improve out-of-sample outcomes.

Sparse portfolio selection is a natural member of this broader regularization literature (e.g., Chang et al., 2000; Brodie et al., 2009; Gao and Li, 2013; Ao et al., 2019). Sparse port-

folios are often justified by transaction costs, liquidity, governance, or managerial simplicity. Our contribution is to show that sparsity has a more basic economic interpretation. Even in frictionless markets, the investor may prefer a limited active set because an unrestricted allocation is statistically costly to estimate. The sparsity level is not imposed outside the model. It is chosen by comparing the marginal opportunity cost of excluding assets with the marginal estimation benefit of reducing the dimension of the allocation problem.

We study this tradeoff in the canonical plug-in Markowitz problem to strip away choices about priors, shrinkage intensities, factor estimators, robust moment estimators, and learning architectures. This allows us to isolate the economic mechanism that makes sparsity valuable: cross-sectional redundancy. In a factor economy, many securities are redundant claims on similar risks, so excluding most of them can have small population cost. The question is how much estimation risk is saved by doing so. The benchmark is not meant to rank all regularization methods; after deriving the main rates, we explain how the same frontier extends to more sophisticated sparse and dense rules with method-specific estimation costs. This also motivates our deliberately narrow notion of complexity. We measure complexity by cardinality, k, the number of active positions. Cardinality counts how many claims the data must distinguish, select, and weight. It is not the only form of complexity, but it is the one directly linked to both the population span and the dimension of the allocation rule.

1.1 Main Results and Economic Mechanism

We formalize this idea through an estimation–efficiency frontier that characterizes how the population Sharpe ratio of an estimated portfolio varies with cardinality k. The analysis delivers three main results.

First, we characterize the Sharpe-ratio loss from estimating a k-asset mean–variance portfolio using T observations from an N-asset universe. The estimation loss decomposes into two components. Selection risk is the cost of choosing the wrong assets. Allocation

risk is the cost of assigning the wrong weights conditional on the selected assets. For the benchmark cardinality-constrained plug-in rule, the total estimation-driven loss scales as

$$\sqrt{(k \log N)/T}.$$

The term k is the effective dimension of the allocation problem, while the $\log N$ term reflects the winner's-curse cost of searching over a large universe of possible supports.

Second, we characterize the population opportunity cost of limited complexity. Under an approximate factor structure, the efficiency loss from restricting the portfolio to k active securities declines with k at a rate governed by market redundancy. With strong factors, the loss is of order $1/k$. With weaker factors, it scales as

$$N^{1-\zeta}/k,$$

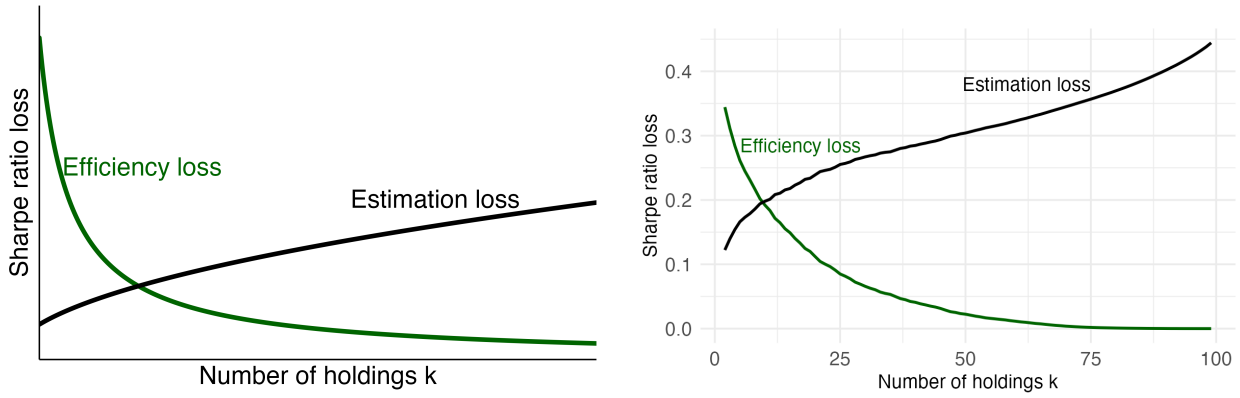
where $\zeta \in (0, 1]$ measures the pervasiveness of systematic risk in the cross section. Economically, the result says that sparse portfolios can be nearly efficient when many securities are near substitutes for the same priced risks. Once the active set spans the main systematic components, additional holdings primarily improve residual-risk diversification.

Third, balancing the statistical and economic sides of the frontier yields an interior optimum for endogenous portfolio complexity:

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

This scaling law is the paper's central economic prediction. Optimal cardinality rises with the amount of data, rises when the universe is less redundant, and depends on universe size through both the opportunity-cost term and the statistical search penalty. The same logic yields the efficiency window

$$N^{1-\zeta} \ll k \ll T/\log N.$$



(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 Appendix B, with $N = 100$ assets generated from a three-factor model.

The lower inequality says that the portfolio must be broad enough to approximate the systematic span. The upper inequality says that the active degrees of freedom must be small enough to be estimated reliably. Sparse portfolios can therefore be asymptotically efficient without assuming that the population-efficient portfolio is sparse.

Figure 1 summarizes the frontier: efficiency loss declines as the active set expands, while estimation risk rises with the dimension of the estimated allocation. Their intersection is the endogenous complexity choice. The marginal interpretation is straightforward. An additional asset should enter the portfolio only if its incremental contribution to the attainable risk–return frontier exceeds the additional estimation risk from selecting and weighting it. This condition is distinct from a trading-cost or liquidity argument. The investor may exclude an asset even when it is costless to trade, because using it requires the data to support another active allocation decision. In this sense, constraints such as cardinality limits, no-short-sale constraints, norm bounds, and robust-optimization uncertainty sets can be interpreted not only as frictions or institutional restrictions, but also as complexity controls that improve decision quality under parameter uncertainty.

1.2 Relation to Dense Complexity and Sparse Subset Recovery

Our notion of complexity differs from, and complements, recent work emphasizing the virtue of complexity in finance (Kelly et al., 2024; Didisheim et al., 2024) and machine-learning applications to asset pricing and portfolio choice (see, e.g., DeMiguel et al., 2023; Kaniel et al., 2023). That literature shows that richer signal spaces, risk models, and investment universes can improve performance when regularization is effective. Closest in spirit is the contemporaneous work of Hellum et al. (2026), who study dense Markowitz and minimum-variance portfolios as the asset-to-sample ratio $c = N/T$ varies. They show that, when the number of assets exceeds the number of observations, adding assets can improve dense ridgeless and ridge-regularized Markowitz and minimum-variance portfolios through implicit regularization. We study a different margin. Dense complexity asks whether expanding a full-universe regularized opportunity set improves performance; sparse complexity asks how many of those claims should receive nonzero weights. The two views are complementary: more assets, signals, or factors may improve the opportunity set, while the shadow cost of selecting and weighting many active claims can still make an interior cardinality optimal.

A closely related paper is Lassance et al. (2025). They derive a finite sample rule for the number of assets to use in several mean–variance combination strategies when moments are estimated, and then leave the choice of which assets to hold to a separate selection rule. Our object is different: we treat active cardinality as a complexity parameter of the portfolio rule itself, so that subset choice and weight estimation enter the same performance comparison. Their analysis provides a size calculation for particular portfolio rules; ours explains why an interior level of active complexity arises in large asset markets.

These distinctions also clarify how the frontier extends beyond the simple plug-in rule. The efficiency term is a population object: the best possible k-active rule pays the sparsity-induced frontier loss relative to the population tangency portfolio, and any implementable sparse rule can only add estimation or optimization loss on top of it. This population

term is unchanged whether the rule uses sample moments, shrinkage, Bayesian priors, factor estimates, robust covariance estimates, or norm constraints (e.g., Ledoit and Wolf, 2003, 2017; DeMiguel et al., 2009; Fan et al., 2012); only a fully dense rule avoids it. What changes across methods is the estimation term. The cardinality-constrained plug-in rule delivers the benchmark rate $\sqrt{(k \log N)/T}$. A more sophisticated sparse rule (e.g., employing regularized moment estimators) has its own rate, say $r_{N,T}$, and balances this rate against the population efficiency rate $N^{1-\zeta}/k$. A fully dense regularized rule has no support-selection loss and no sparsity-induced efficiency loss; its performance is governed by its own estimation or regularization error $r_{N,T}$. Unlike dense-complexity analyses, our frontier does not require fixing an asymptotic path for N/T or using random-matrix limits. We do not attempt a full analytical or empirical ranking of shrinkage, Bayesian, factor, robust, dense, and sparse rules. Instead, we show that even the simple sparse plug-in rule can be asymptotically efficient relative to the best possible benchmark, the population tangency portfolio.

The role of sparsity here also differs from its role in canonical high-dimensional sparse regression. In that literature, sparsity is typically an assumption about the data-generating parameter: only a small subset of regressors is relevant, and methods such as the lasso are analyzed as recovery or prediction tools under restricted-eigenvalue, compatibility, or related sparse-design conditions (Bühlmann and Van De Geer, 2011; Hastie et al., 2015). In portfolio choice, by contrast, the target is not support recovery. The object is a decision rule: a vector of portfolio weights evaluated by the out-of-sample Sharpe ratio it delivers. Under empirically relevant factor structures, the population-efficient portfolio is generally dense, because broad holdings help span systematic risks and diversify residual risk. Sparse allocation is therefore not a recovery device for a sparse truth; it is a finite-sample allocation rule that limits the number of active decisions the data must support. Moreover, common factors create the cross-sectional collinearity that makes sparse-regression conditions difficult to justify in financial returns, but the same dependence is economically useful: it makes many securities redundant claims, or near substitutes, for the same underlying risks. Sparsity exploits this

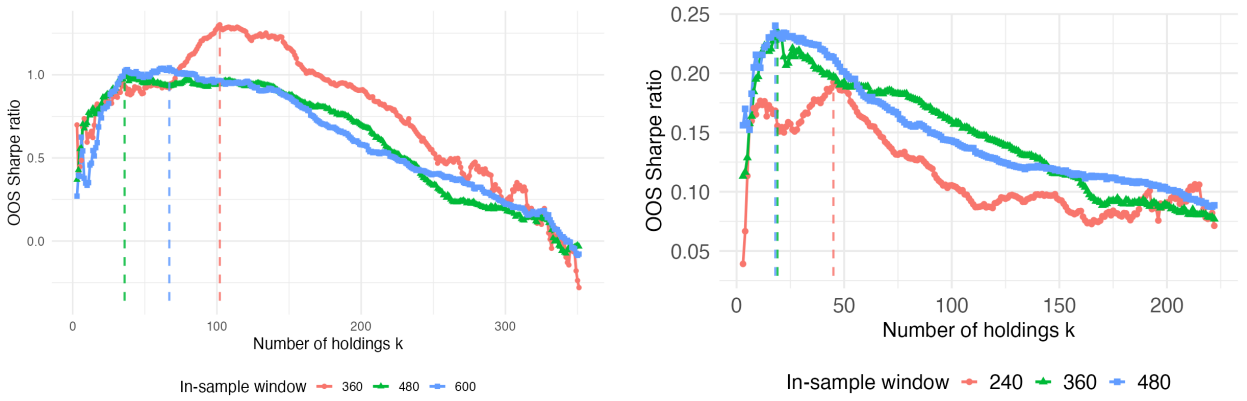
market redundancy rather than assuming that the population-efficient portfolio is sparse.

1.3 Implementation and Evidence

Exact cardinality-constrained mean–variance optimization is computationally hard in realistic universes (Chang et al., 2000; Gao and Li, 2013). We therefore also study a tractable implementation based on ℓ_1 regularization (Brodie et al., 2009), which is closely related to the lasso (Tibshirani, 1996) and to gross-exposure stabilization in portfolio choice (Fan et al., 2012). The ℓ_1 penalty provides a feasible convex way to move along the complexity frontier. It limits the optimizer's ability to convert small sampling errors into large offsetting long–short positions, while still allowing exposure to priced systematic risks. The population efficiency term is the same as for any sparse rule; the penalty changes how estimation error is controlled and provides a practical route to asymptotic efficiency under approximate factor structure.

The simulations illustrate the theory in a controlled environment calibrated to a Fama–French three-factor structure. Because the population moments are known in the simulation, we can trace the unrestricted efficient Sharpe ratio, the population k-sparse frontier, the oracle Sharpe ratio on the sample-selected support, and the realized population Sharpe ratio of the sample optimizer. The population sparse frontier rises and flattens with k, while the estimated portfolio is hump-shaped. The loss decomposition confirms the mechanism: efficiency loss declines with k, selection risk is most important when the support is very small, and allocation risk rises as the within-support mean–variance problem becomes higher dimensional. The optimal k shifts upward with the sample length T, consistent with the scaling law for k^*.

The empirical counterpart of the theory is a Sharpe-ratio profile in k. If the mechanism is correct, out-of-sample performance should rise when the investor moves away from extreme concentration, because additional holdings improve span and diversification. It should even-



(a) Broad U.S. monthly managed portfolios

(b) Individual stock returns

Figure 2: Out-of-sample Sharpe ratio as a function of the target cardinality k in two monthly universes. Panel (a) reports the broad U.S. monthly managed-portfolio universe ($N = 352$), with in-sample windows $T \in \{360, 480, 600\}$ months. Panel (b) reports the balanced CRSP individual-stock universe ($N = 223$), with in-sample windows $T \in \{240, 360, 480\}$ months. The vertical dashed lines indicate the value of k that maximizes the realized out-of-sample Sharpe ratio for each in-sample estimation window. Portfolios are computed from the ℓ_1 relaxation and evaluated one month ahead ($h = 1$) using a rolling-window out-of-sample procedure.

tually fall when the active set becomes too large for the available sample, because estimation error dominates. This is what we find. Figure 2 shows the characteristic hump-shaped relation between target cardinality and out-of-sample Sharpe ratio in two distinct settings: a broad U.S. monthly managed-portfolio universe and a balanced panel of individual U.S. stocks.

The empirical analysis is broader than these illustrations. We study several Kenneth French managed-portfolio universes, including size–value, industry, characteristic-sorted, daily, and international test assets. We also examine self-financing anomaly portfolios from Open Source Asset Pricing (Chen and Zimmermann, 2022) and a balanced panel of CRSP individual stocks. Across these settings, realized out-of-sample Sharpe ratios are typically non-monotone in k. Performance improves from very sparse portfolios, peaks at an intermediate active set, and then deteriorates as the portfolio becomes too dense. The high-performance region is sparse relative to the size of the investable universe, but it is not

extremely concentrated. This pattern is consistent with the frontier: the opportunity cost of excluding assets falls quickly under factor redundancy, while the statistical cost of estimating a larger allocation rises with k.

The channel evidence supports the decomposition. Selection instability is high for very sparse portfolios, where small changes in estimated moments can reshuffle the chosen support. It declines as k grows, because consecutive supports overlap more and many assets or strategies are close substitutes. Weight instability moves in the opposite direction: it increases with k, reflecting the rising sensitivity of within-support mean–variance weights to estimation error. The Sharpe-ratio peak occurs where the gains from improved span and lower selection instability are offset by higher allocation instability. Robustness checks show that the main profiles are stable to adding traded Fama–French factor-mimicking portfolios and to normalizing realized weights to a common gross-exposure scale. Feasible tuning rules based on leave-one-out performance or bias-adjusted Sharpe estimates recover a substantial share of the best-in-hindsight performance, indicating that the interior complexity choice is not merely an ex-post artifact.

Taken together, the results establish sparsity as a principled form of risk control in portfolio choice. Sparse portfolios do not claim that only a few assets matter in the population economy. They respond to the fact that the investor observes only a finite sample. In large asset or strategy universes with strong common variation, the marginal benefits of additional holdings can be smaller than the stability gained by reducing the number of active allocation decisions.

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 estimation–efficiency frontier, characterizes selection and allocation risk, and derives efficiency losses under approximate factor structure. Section 4 discusses the economic mechanism and statistical scope of the theory. Section 5 reports the empirical results, and the final section concludes. The appendices contain simulation evidence, additional empirical results, numer-

ical algorithms, supplementary theory, and proofs.

2 Framework and Sparse Portfolio Rules

This section defines the mean–variance benchmark and the sparse portfolio rules used in the theory. Portfolio complexity is measured by the number of active positions, or cardinality k. Appendix A provides a consolidated notation reference.

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 and the population-efficient Sharpe ratio are, respectively,[1]

$$\theta(w) := \frac{\mathbb{E}(r_{p,w})}{\sqrt{\mathbb{V}(r_{p,w})}} = \frac{w^\top \mu}{\sqrt{w^\top \Sigma w}}, \qquad \theta^* := \max_{w \in \mathbb{R}^N} \theta(w).$$

When Σ is of full rank, this maximum is attained and satisfies

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

We adopt the classical mean–variance objective

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

[1] We work with the convention $\theta(0) = 0$ for the Sharpe ratio of the zero portfolio $w = 0$.

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.

When N is not small relative to T, the plug-in rule is sensitive to sampling error in $\widehat{\mu}$ and $\widehat{\Sigma}$ and may be undefined if $\widehat{\Sigma}$ is singular (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). This motivates the sparse rules below.

2.2 Sparse Portfolio Rules

We study two sparse rules: a hard cardinality-constrained rule and its ℓ_1-regularized relaxation.

Sample cardinality-constrained strategy. Write $[N] := \{1, \dots, N\}$, let $\text{supp}(w) := \{i \in [N] : w_i \neq 0\}$ denote the support (index set of nonzero elements) of a weight vector $w \in \mathbb{R}^N$, and let $\|w\|_0 := |\text{supp}(w)|$ 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

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

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.

Problem (2) is nonconvex and computationally hard. Solving the problem exactly requires searching over candidate supports $A \subset [N]$ with $|A| \leq k$ and, for each support, optimizing the Markowitz objective using only the assets in A. The number of candidate supports grows extremely quickly. For example, with $N = 300$ and $k = 150$, a single cardinality layer contains

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

possible supports. Even at an idealized rate of 10^9 support evaluations per second, enumerating this layer alone would take roughly 3×10^{72} years. Thus, although (2) is generally nonconvex and NP-hard (Gao and Li, 2013), it is best viewed in large universes as a transparent benchmark for portfolio complexity, rather than as a scalable implementation strategy.[2]

[2]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.

ℓ_1-regularized strategy. A computationally tractable way to approximate the cardinality-constrained problem is to replace the hard support constraint with a convex penalty on gross exposure:

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

where $\|w\|_1 := \sum_{i=1}^N |w_i|$ and $\lambda \geq 0$ controls the strength of the penalty; see, e.g., Brodie et al. (2009). When $\lambda = 0$, (3) reduces to the unrestricted plug-in rule. For $\lambda > 0$, the penalty shrinks gross exposure and can set weights exactly to zero. The ℓ_1 norm has a kink at zero, so the first-order conditions imply a thresholding behavior: an asset receives a nonzero weight only when its marginal contribution to the sample mean–variance objective is large enough to overcome the penalty. Assets whose estimated risk-adjusted contribution is below this threshold remain exactly at zero. Thus, varying λ traces a regularization path of portfolios with different effective support sizes. We index the resulting path by its induced support size $\widehat{k}(\lambda) := \|\widehat{w}_{\ell_1,\lambda}\|_0$; Appendix E gives the computational details.

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

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

We refer to

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

as the (pointwise) *estimation loss* at complexity level k. It decomposes into support-selection and weight-allocation components. 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:

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

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 two terms are the selection risk and allocation risk, respectively.

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 is a population restriction cost reflecting improved spanning as the portfolio becomes more complex; the last two terms are sample-dependent and typically increase with complexity. The same decomposition applies to $\widehat{w}_{\ell_1,\lambda}$ by using its induced active set.

3 Estimation and Efficiency Losses

We construct bounds scaling with (N, T, k) for the two terms in (10) and then the population loss Δ_k in (6). The estimation bounds use high-dimensional concentration over candidate supports; the efficiency bound uses approximate factor structure.

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

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

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:[3]

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

The relevant object is therefore the uniform fluctuation $\sup_{|A|\le k} |\widehat{\theta}_A^2 - \theta_A^2|$ over the class of admissible supports.

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

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

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). Both θ_A and $\widehat{\theta}_A$ are monotone in set inclusion: if $A \supseteq B$, then $\theta_A \ge \theta_B$ and $\widehat{\theta}_A \ge \widehat{\theta}_B$. Hence, for the purpose of bounding the search problem, it suffices to consider candidate sets with $|A| = k$. The class of admissible supports then has size $\binom{N}{k}$, which grows rapidly with N even for moderate k. Controlling $\sup_{|A|=k} |\widehat{\theta}_A^2 - \theta_A^2|$ therefore requires a complexity term of order $\log\binom{N}{k}$. Since $\binom{N}{k} \le (eN/k)^k$, this term is bounded by $k\log(eN/k)$ and hence by order $k\log N$. In financial terms, as the investment universe expands, the winner's curse in asset selection becomes more severe: at least one subset is likely to display an exceptionally high in-sample Sharpe ratio by chance, and

[3] see Lemma F.1 in Appendix F.

selection risk is the performance penalty from 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 D 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 $|A| \le k$, 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 $\sum_{j \le k} \binom{N}{j}$ 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* $\theta^* < \infty$*. Then*

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

Theorem 1 formalizes the upward-sloping statistical side of the portfolio complexity frontier. At complexity level k, the Sharpe-ratio loss induced by estimating the portfolio from data is of order $\sqrt{\frac{k \log N}{T}}$. Hence, holding T and N fixed, the leading dependence on portfolio cardinality is $\sqrt{k}$. Economically, this loss reflects two related costs of complexity. First, a larger active set requires estimating and combining a higher-dimensional set of mean and covariance inputs. Second, because the active set is chosen by searching over many candidate supports, sampling noise is amplified by a high-dimensional selection effect. This increasing-in-k term corresponds to the estimation-risk curve in Figure 1, and is the sparse-portfolio 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

We now bound the population loss $\Delta_k = \theta^* - \theta_k^*$, which measures the cost of limiting the investable span to k assets even with perfect knowledge of (μ, Σ). Without structure, no general rate links θ_k^* to θ^* as k grows. The next results impose a factor structure to obtain

such a rate.

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 generically dense, 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 density 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 the factor structure in* (12)*, if* $\liminf_{N\to\infty} \beta^\top \beta/N > 0$*, then*

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

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 of 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 for a fixed number of factors L*,* $B \in \mathbb{R}^{N\times L}$ *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.*

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

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

of N, such that

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

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 $1/N$:*

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

Assumptions 1–4 decompose means and covariances into factor and idiosyncratic components, control conditioning, scale factor strength through ζ, and prevent pricing errors from dominating the factor-spanning channel. The next result gives the resulting efficiency-loss rate.

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 ($\zeta = 1$),

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

The bound identifies factor redundancy as the mechanism behind the declining opportunity cost of sparsity. A sparse portfolio need not reproduce the dense mean–variance allocation asset by asset; as k grows, it can approximate the priced factor span while diversifying residual risk. Greater redundancy, summarized by larger ζ, makes this approximation cheaper, yielding the benchmark $1/k$ loss under strong factors, whereas weaker redundancy inflates the loss by $N^{1-\zeta}$. This is the "efficiency loss" curve in Figure 1, and is consistent with 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

Combining Theorems 1 and 2 gives the following frontier.

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

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

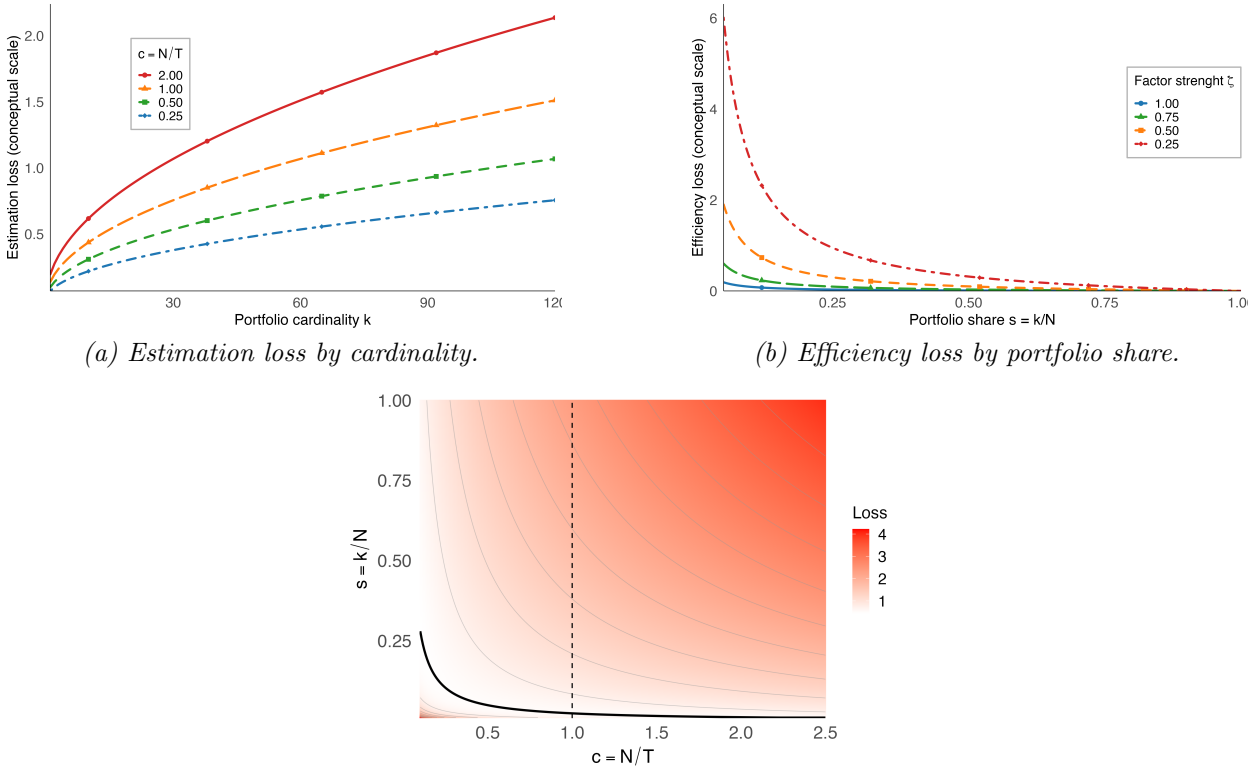
In particular, if $N^{1-\zeta} \ll k \ll T/\log N$, then $\theta(\widehat{w}_k) \xrightarrow{p} \theta^*$.

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.

Figure 3 gives a visual counterpart to the two terms in Theorem 3. Panel (a) shows that the estimation component rises with portfolio cardinality and is steeper when $c = N/T$ is larger, reflecting the greater statistical cost of estimating more active positions in a higher-dimensional universe. Panel (b) shows that the efficiency component declines with the portfolio share $s = k/N$ and with factor strength ζ, because stronger factor structure makes excluded assets more redundant. Panel (c) combines the two forces in share coordinates: for fixed constants and fixed ζ, larger c shifts the minimizing ridge toward lower s, so for a fixed universe the optimal cardinality decreases as the sample becomes short relative to the cross section.

Remark 5 (Balancing and admissible growth rates)**.** *The bound in Theorem 3 suggests a benchmark scaling for the complexity level obtained by balancing the estimation and efficiency*



(a) Estimation loss by cardinality.

(b) Efficiency loss by portfolio share.

(c) Loss surface and optimal share.

Figure 3: Conceptual comparative statics for Theorem 3. Top left: estimation loss in k for several values of $c = N/T$. Top right: efficiency loss in $s = k/N$ for several values of ζ. Bottom: loss surface in (c, s) with the minimizing sparsity share.

terms:

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

This rate is not a tuning prescription, as constant factors depend on unknown features of (μ, Σ)*, but it makes explicit the comparative statics of portfolio complexity: the economically relevant number of active positions increases with available data and decreases as factor strength increases. More generally, the theory does not require a fixed asymptotic relation between* N *and* T*. It allows* $N/T \to 0$*,* $N/T \to c \in (0, \infty)$*, and some regimes with* $N \gg T$*. The operative requirement is that portfolio complexity lie in the admissible window*

$$N^{1-\zeta} \ll k \ll T/\log N.$$

Equivalently, such a window exists when $N^{1-\zeta}\log N = o(T)$.

3.5 Implications for ℓ_1 Regularization

An analogous economic tradeoff applies to the ℓ_1-regularized rule. For ℓ_1 regularization we allow pricing errors to persist with N, but restrict their standardized aggregate magnitude. Relative to Assumption 4, which imposes the stronger $O(1/N)$ restriction used to preserve the factor-driven $1/k$ 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 returns are i.i.d. Gaussian, and Assumptions 1, 2, 3, and 5 hold. Then there exists a numerical constant* $C_0 > 0$ *such that, for any* $\lambda \geq C_0\sqrt{\log N/T}$*,*

$$\theta^* - \theta(\widehat{w}_{\ell_1,\lambda}) = O_p\big(N^{1-\zeta}\lambda\big), \qquad \text{as } N, T \to \infty.$$

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 per-

sistent 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.

Implementation details are collected in Appendix E, which explains how the ℓ_1-regularized rule in (3) is computed, how the lasso path is mapped into an induced support size, and how MIQP solutions are used as benchmarks when problem sizes permit.

4 Economic Mechanism and Scope

The preceding results should be read as a theory of portfolio allocation under parameter uncertainty. The claim is not that the population tangency portfolio is sparse, nor that sample-moment plug-in optimization is the preferred empirical implementation. The claim is narrower and more structural: when many assets provide overlapping access to the same priced risks, portfolio cardinality becomes an endogenous complexity choice. Adding an asset expands the attainable payoff span, but it also creates another active decision that must be selected and weighted from finite data.

4.1 Market Redundancy and the Shadow Cost of Portfolio Complexity

The economic primitive is market redundancy. In a factor economy, many listed securities are near-redundant claims on similar systematic risks. The number of traded assets is therefore not the same as the number of economically distinct opportunities. A sparse portfolio can often preserve the relevant factor exposure with a limited active set; after that span is largely captured, additional securities mainly provide finer residual-risk diversification.

Assumption 3 formalizes this idea. The matrix $B^\top \Sigma_0^{-1} B$ measures the aggregate standardized factor exposure available in the cross section. The exponent ζ governs how quickly this exposure grows with N. We therefore interpret ζ as a market redundancy parameter: when $\zeta = 1$, systematic risks are pervasively represented and many assets are close substitutes; when $\zeta < 1$, redundancy is weaker and a larger support is needed to approximate the same factor span. Theorem 2 gives the associated population opportunity cost, $\theta^* - \theta_k^* = O\big(N^{1-\zeta}/k\big)$, with the strong-factor benchmark $O(1/k)$ when $\zeta = 1$.

The opposing force is statistical. Estimation loss is the shadow cost of portfolio complexity: it is the Sharpe-ratio cost of asking the data to identify a k-dimensional portfolio rule in an N-asset universe from only T observations. This cost has two components: selection risk, from choosing the wrong active set, and allocation risk, from assigning noisy weights conditional on that set. For the benchmark sparse plug-in rule, Theorem 1 gives $\theta_k^* - \theta(\widehat{w}_k) = O_p\Big(\sqrt{(k \log N)/T}\Big)$. The factor k is the active dimension of the allocation problem, while $\log N$ is the search cost of ranking candidate supports in a large universe.

Combining the population and statistical terms yields the frontier in Theorem 3:

$$\underbrace{\theta^* - \theta(\widehat{w}_k)}_{\text{Total Sharpe-ratio loss}} = O_p\left(\sqrt{\frac{k \log N}{T}} + \frac{N^{1-\zeta}}{k}\right).$$

Thus sparsity is not imposed for its own sake. A security is excluded when its marginal con-

tribution to spanning and residual-risk diversification is smaller than the additional shadow cost of selecting and weighting it. The admissible efficiency window, $N^{1-\zeta} \ll k \ll T/\log N$, expresses the same logic: the support must be large enough to span priced risks, but small enough for the data to support the active decisions.

4.2 Not Sparse Recovery

This mechanism is distinct from sparse recovery in high-dimensional regression. In canonical sparse-regression problems, sparsity is an assumption on the unknown population coefficient vector, and the goal is often support recovery or prediction under restricted-eigenvalue, compatibility, or related sparse-design conditions (e.g., Bühlmann and Van De Geer, 2011; Hastie et al., 2015; Wainwright, 2019). Portfolio choice has a different object: a decision rule evaluated by its out-of-sample Sharpe ratio. Under the factor structures considered here, the population mean–variance direction $w_* \propto \Sigma^{-1}\mu$ is generically dense. Broad holdings help fine-tune systematic exposure and diversify residual risk when moments are known. Sparse allocation is therefore not a claim that w_* has few nonzero components. It is a finite-sample rule for deciding how many active positions the data can support. The condition $N^{1-\zeta} \ll k \ll T/\log N$ is correspondingly not a sparse-target condition. It is a feasibility condition for portfolio allocation under parameter uncertainty.

This distinction also explains why common-factor dependence is not a nuisance in the economic argument. In sparse regression, strong collinearity can undermine support recovery. Here the same dependence is useful: it turns many securities into near substitutes for the same underlying risks and makes sparse spanning economically cheap.

The interpretation of the ℓ_1 rule is the same. The penalty in Theorem 4 is not used to recover a truly sparse population portfolio. It is a continuous complexity control. Choosing $\lambda \asymp \sqrt{\log N/T}$ makes the penalty large enough to dominate high-dimensional sampling noise while preserving exposure to systematic premia. Under Assumptions 1, 2, 3, and 5, $\theta^* -$

$\theta(\widehat{w}_{\ell_1,\lambda}) = O_p\big(N^{1-\zeta}\lambda\big)$. The bounded standardized-pricing-error condition rules out scalable residual arbitrage through idiosyncratic hedging; it does not require the dense tangency portfolio to be approximately sparse.

4.3 Sparse Plug-In Rules, Shrinkage, and Dense Complexity

The sparse plug-in rule is an analytic benchmark. Its role is to make the shadow cost of portfolio complexity explicit by using sample moments, selecting a support, and estimating weights on that support. It should not be read as a claim that unregularized sample moments are the best empirical inputs.

The efficiency term is population-level and rule-independent. Any k-active rule pays the same frontier cost $\theta^* - \theta_k^*$ relative to the dense population tangency benchmark. What changes across implementations is the estimation term. If a sparse procedure based on shrinkage, robust moments, factor estimates, or another regularized input has estimation loss $r_{N,T}(k)$, its rate-level frontier is

$$r_{N,T}(k) + N^{1-\zeta}/k.$$

Better moment estimates lower $r_{N,T}(k)$ and can support a larger active set; noisier estimates shift the optimum toward smaller supports.

Dense shrinkage rules address the same parameter-uncertainty problem from the other side. They keep the full investable span and therefore avoid the support-restriction term, but they must control estimation error in a dense direction, through shrinkage of means, covariances, weights, factors, forecasts, or portfolio combinations (e.g., Jorion, 1986; Tu and Zhou, 2011; Ledoit and Wolf, 2003, 2017; DeMiguel et al., 2009; Kelly et al., 2024; Hellum et al., 2026). If a dense rule has Sharpe-ratio loss $s_{N,T}$, then dense shrinkage improves on

sparse allocation at the rate level when

$$s_{N,T} < \inf_{1 \le k \le N} \left\{ r_{N,T}(k) + N^{1-\zeta}/k \right\}.$$

Conversely, sparsity is attractive when the reduction in estimation loss from lowering the active dimension exceeds the population cost of omitting redundant claims. The paper does not rank all dense and sparse regularization designs. It identifies when cardinality itself is an economically meaningful control variable. In applications, shrinkage and sparsity can be combined: regularized moment estimators can be used inside a sparse rule, and penalties such as elastic-net formulations can jointly shrink and select.

4.4 Practical Guidance

The balancing rate in Remark 5, $k \asymp \left(N^{1-\zeta} \sqrt{T/\log N} \right)^{2/3}$, should be read as comparative statics, not as a calibrated tuning formula. It states that more data can support more active holdings, while greater market redundancy lowers the number of names needed for near-efficiency. The constants depend on signal strength, conditioning of (μ, Σ), tail behavior, serial dependence, the moment estimator, and the normalization used in implementation.

Empirically, the frontier is best used as a diagnostic. One should trace a sparse path – either directly in k or through the induced support size along an ℓ_1 path – and select complexity using only in-sample or real-time information, such as rolling validation, leave-one-out scores, bias-adjusted Sharpe estimates, or other pre-specified tuning rules. This preserves the economic interpretation: sparsity is a disciplined response to parameter uncertainty, not a substitute for validation.

5 Empirical Analysis

The central empirical question is whether the estimation–efficiency frontier is visible in realized sparse-portfolio performance. Theorem 1 predicts that, for a given sample size, increasing the number of active positions amplifies estimation risk, with leading dependence on cardinality of order $\sqrt{k}$. Theorem 2 predicts that, under a factor structure, the opportunity cost of excluding assets declines rapidly with k, with leading dependence of order $1/k$ in the strong-factor case. 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. Empirically, 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.[4]

5.1 Estimation and Evaluation

Let $\{r_t\}_{t=1}^{T_0}$ denote a time series of N test-asset excess returns, with $r_t \in \mathbb{R}^N$. For each dataset, we consider a set of in-sample window lengths $\mathcal{T} \subset \mathbb{N}$ and write $T_{\max} := \max \mathcal{T}$. To maintain comparability across window lengths, all out-of-sample summaries are computed over the common evaluation period $t = T_{\max}, \ldots, T_0 - 1$, so that every $T \in \mathcal{T}$ is evaluated on the same $T_0 - T_{\max}$ out-of-sample observations.[5] This convention ensures that differences across T reflect in-sample estimation error rather than differences in out-of-sample evaluation noise.

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

[4]The simulation analysis in Appendix B examines the same prediction in a controlled economic setting, using a calibrated strong-factor three-factor model, and provides the simulation evidence accompanying the theory and empirical analysis.

[5]For example, Dataset 1 is monthly from January 1971 through December 2023, and we use $\mathcal{T} = \{360, 480, 600\}$, so all reported out-of-sample quantities are computed over the last $T_0 - T_{\max}$ months, with $T_{\max} = 600$.

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 $\widehat{w}_{T,k,t}$ using the numerical implementation described in Appendix E, targeting cardinality k along the ℓ_1 regularization path, and 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 $\|\widehat{w}_{T,k,t}\|_1 = 1$. This normalization fixes the gross-exposure scale of the long–short positions and facilitates comparisons of portfolio magnitudes across window lengths T and rebalancing dates.

5.1.1 Out-of-sample Sharpe ratio

We evaluate performance out-of-sample at horizon $h = 1$, with rebalancing at each period. For each (T,k) strategy, the realized out-of-sample portfolio return 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 $\widehat{\theta}^{\text{oos}}(T,k)$ by $\sqrt{12}$ for monthly data and by $\sqrt{252}$ for daily data. The empirical object of interest is the k–profile $\{\widehat{\theta}^{\text{oos}}(T,k)\}_{k=1}^{N}$ for each T.

5.1.2 Selection Instability, Weight Instability, and Turnover

To complement the out-of-sample Sharpe-ratio evidence, we report three diagnostics of portfolio stability and implementation. The first two are motivated by the selection–allocation decomposition in (10). Selection instability captures changes in which assets are held, and is measured by the average Jaccard distance between consecutive selected supports. Weight instability captures changes in the recommended allocations, and is measured by the median ℓ_1 distance between consecutive target weight vectors. Finally, turnover translates these support and weight fluctuations into an implementation metric: the median one-way trading required to move from the post-return, pre-trade portfolio to the newly recommended target portfolio. Compactly, the three objects are:[6]

$$\begin{aligned}
\bar{D}(T,k) &= \text{mean}_t\{1 - J(S_{t,(k)}(T), S_{t-1,(k)}(T))\},\\
\widehat{\Delta}^{(1)}(T,k) &= \text{median}_t\big\|\widehat{w}_{T,k,t} - \widehat{w}_{T,k,t-1}\big\|_1,\\
\widehat{\text{TO}}(T,k) &= \text{median}_t\big\|\widehat{w}_{T,k,t+1} - \widehat{w}^{\text{pre}}_{T,k,t+1}\big\|_1.
\end{aligned}$$

Here $S_{t,(k)}(T) = \text{supp}(\widehat{w}_{T,k,t})$, $J(\cdot,\cdot)$ denotes the Jaccard similarity index, and $\widehat{w}^{\text{pre}}_{T,k,t+1}$ is the post-return portfolio before rebalancing. Appendix C.2 gives the formal definitions and implementation details.

[6] In our framework, the out-of-sample performance shortfall relative to the population k-sparse frontier reflects selection risk, which operates through changes in the selected support, and allocation risk, which operates through noisy weights conditional on a given support. The selection and weight instability diagnostics used in this paper do not estimate these loss components directly, but they provide empirical proxies for these sources of instability.

5.2 Data

To examine the sparsity tradeoff across distinct return spaces, we use seven test-asset universes that differ in construction, cross-sectional dimension, sampling frequency, and sample period. Here, we describe the four datasets used in the main text.

1. **Dataset 1: Broad characteristic-sorted universe + 17 industry portfolios (N=352, monthly).** Our largest monthly 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 January 1971 through December 2023.

2. **Dataset 2: 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 from nine sets of 25 double-sorted portfolio excess returns: size–book-to-market, size–operating profitability, size–investment, size–short-term reversal, size–momentum, size–long-term reversal, book-to-market–operating profitability, book-to-market–investment, and operating profitability–investment. We then 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.

3. **Dataset 3: Anomaly long–short portfolios (N=143, monthly).** To study sparsity in a broad panel of self-financing anomaly strategies, we use monthly long–short portfolio returns. The universe contains 143 anomaly portfolios spanning valuation, investment, profitability, issuance, momentum, reversal, liquidity, volatility, and other predictor-based strategies. The sample runs from January 1971 through December 2024.

4. **Dataset 4: Individual U.S. equity returns from CRSP (N=223, monthly).** To study the sparsity tradeoff in a panel of individual securities, we construct a balanced panel of U.S. common stocks from CRSP with a full monthly return history over the sample period considered, which leaves us with 223 assets. The sample runs from October 1971 through October 2025.

Appendix C.1 describes three supplementary monthly universes used only in the appendix: the U.S. 100 size–value portfolio universe, the same universe augmented with U.S. industry portfolios, and an international size–value/profitability universe.

Datasets 1 and 2 are obtained from the Kenneth R. French Data Library (French, 2026). Excess returns are computed relative to the risk-free rate from the same data library. Dataset 3 is obtained from Open Source Asset Pricing (Chen and Zimmermann, 2022). Because the anomaly portfolios are constructed as long–short portfolios, they are already self-financing excess-return series. Dataset 4 is sourced from CRSP, and we compute excess returns by subtracting the one-month Treasury bill rate from the monthly stock return for each security.

For the factor-augmentation robustness checks in the main text, we append the Fama–French three factors (MKT, SMB, HML) to the broad monthly U.S. managed-portfolio universe (Dataset 1) and to the daily managed-portfolio universe (Dataset 2). We leave Datasets 3 and 4 unaugmented because the goal there is to study sparsity across self-financing anomaly strategies and individual securities, rather than to combine those exercises with explicit factor-spanning tests.

5.3 Results

The empirical evidence is consistent with the complexity frontier across the return spaces studied in the main text and appendix. The out-of-sample Sharpe-ratio profiles are generally hump-shaped in k: performance improves as very sparse portfolios gain spanning and

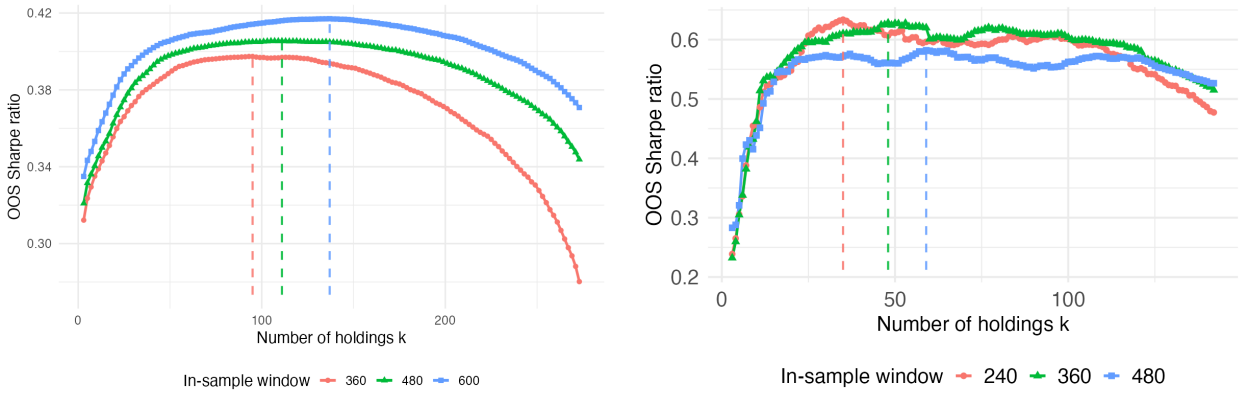
diversification benefits, and then declines as larger supports magnify estimation error. The ex-post high-performance region differs across datasets, reflecting differences in the effective dimension and redundancy of each return space.

The same qualitative pattern remains largely unchanged after augmenting the investable universe with Fama–French factor-mimicking portfolios and after normalizing realized weights to satisfy $\|w\|_1 = 1$. The stability diagnostics move in opposite directions: selection instability generally decreases with k, whereas weight instability generally increases with k. Finally, the leave-one-out (LOO) and bias-adjusted Sharpe estimator (UE) tuning rules typically select interior cardinalities and recover a large share of the best-in-hindsight out-of-sample performance without using future returns.

5.3.1 Out-of-sample Sharpe Ratio

Figure 2 in the introduction reports realized out-of-sample Sharpe ratios of the ℓ_1 portfolio rule for the broad U.S. monthly managed-portfolio universe (Dataset 1) and for the CRSP individual-stock universe (Dataset 4). Figure 4 reports the corresponding profiles for the U.S. daily managed-portfolio universe (Dataset 2) and for the anomaly long–short universe (Dataset 3). Since our objective is to study the sparsity tradeoff for each target cardinality $k \in [N]$, we index the ℓ_1 regularization path by cardinality: for each k, we choose the first penalty value λ along the path whose solution satisfies $\|\widehat{w}_{\ell_1,\lambda}\|_0 = k$, and otherwise use the closest admissible path point according to the implementation rule in Appendix E. 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 dominant empirical pattern is non-monotonicity in k: the Sharpe ratio typically 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.



(a) Dataset 2 (b) Dataset 3

Figure 4: Out-of-sample Sharpe ratio of the k–sparse ℓ_1 portfolio rule in the U.S. daily managed-portfolio universe (Dataset 2) and in the anomaly long–short universe (Dataset 3). The holding period is $h = 1$ (one trading day for the daily universe and one month for the anomaly universe), and all OOS quantities are computed over a common evaluation window of length $T_0 - T_{\max}$ within each dataset.

The peak location shifts with T_{in} in a way broadly consistent with lower estimation noise in longer windows. In the broad U.S. monthly managed-portfolio universe (Dataset 1), shown in Figure 2a, the Sharpe-maximizing sparsity is about 100 for $T_{\text{in}} = 360$, then moves back to roughly 40 for $T_{\text{in}} = 480$ and about 70 for $T_{\text{in}} = 600$. More important than the precise maximizing value is the broad high-performance region in which the out-of-sample Sharpe ratio remains close to its peak.

Dataset 4 offers a complementary asset-level perspective. In Figure 2b, the Sharpe profiles are more concentrated: the optimal sparsity level remains well below 50 throughout, and for $T_{\text{in}} = 360, 480$ it stays below about 25. This pattern suggests that, for individual assets, the marginal diversification gains from expanding the support are exhausted relatively quickly.

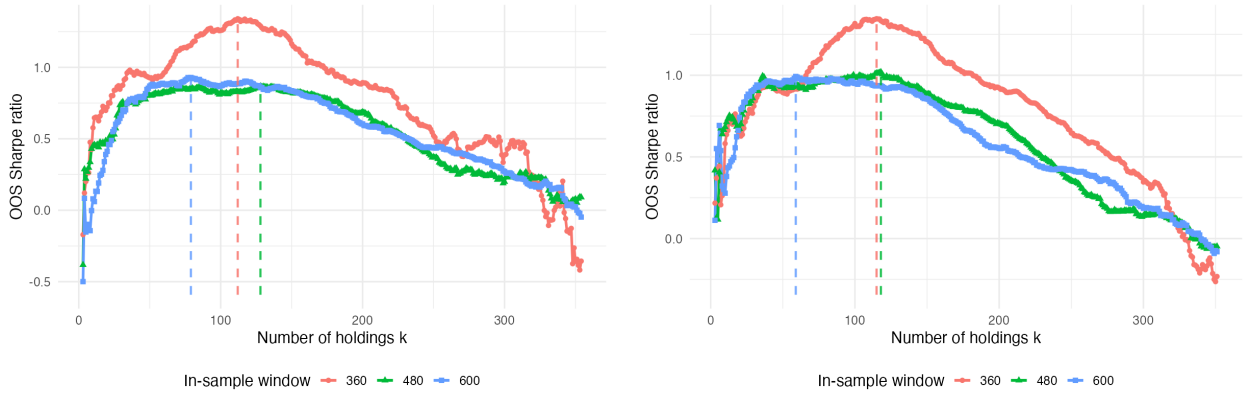
Figure 4 shows that the same non-monotone pattern also emerges in the daily managed-portfolio universe and in the anomaly universe. In the daily universe (Dataset 2, 274 assets), the peak lies around $k^* \approx 95$–110 for $T_{\text{in}} = 360$–480 and rises to about 140 for $T_{\text{in}} = 600$. In the anomaly universe, the Sharpe-maximizing sparsity levels are around $k^* \approx 35$, 50, and 60 for $T_{\text{in}} = 240$, 360, and 480, respectively. The out-of-sample Sharpe ratio rises steeply

toward its maximum to the left of $k \approx 35$ and then declines only gradually to the right of $k \approx 60$, indicating a fairly broad high-performance region. Because the anomaly assets are self-financing long–short strategy returns, this evidence shows that the sparsity tradeoff is not tied to long-only portfolio construction.

Overall, the main-text 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. The same qualitative patterns also appear in the remaining monthly datasets, reported in Appendix C.3.1.

5.3.2 Out-of-sample Sharpe Ratio: Robustness Checks

Two features could potentially influence the hump-shaped Sharpe profiles in Figures 2 and 4: (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. In the main text we report both robustness checks for the broad U.S. monthly managed-portfolio universe; the remaining robustness figures are deferred to Appendix C.3.2.



(a) Factor-augmented dataset

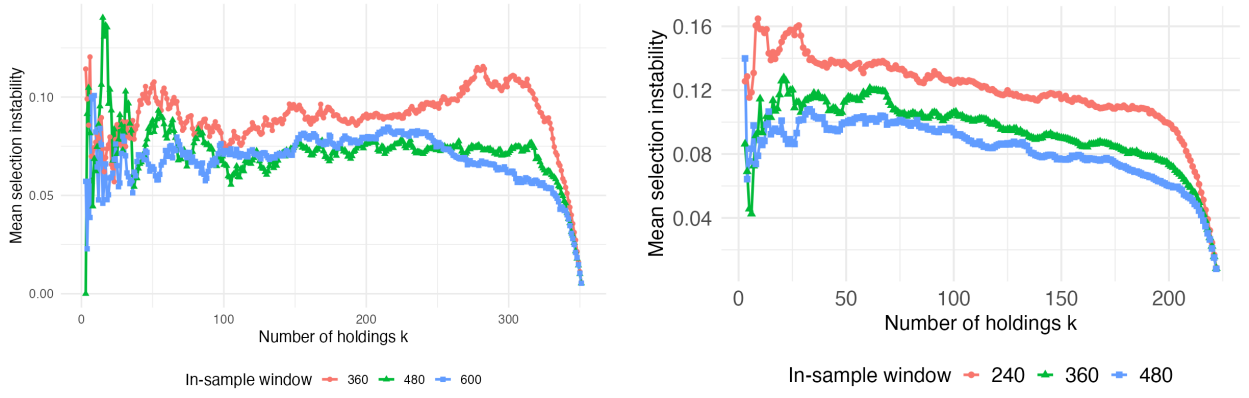
(b) Normalized weights: $\|w\|_1 = 1$

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

Figure 5 shows that the hump-shaped Sharpe-ratio profile survives both robustness checks. Making standard factor payoffs explicitly tradable shifts the high-performance region only modestly, while ex-post ℓ_1 normalization leaves the broad shape nearly unchanged. In both cases, performance still rises from very sparse supports, reaches an interior optimum, and then declines as the portfolio becomes too dense. The remaining robustness figures for the other datasets are reported in Appendix C.3.2.

5.3.3 Selection and Weight Instability

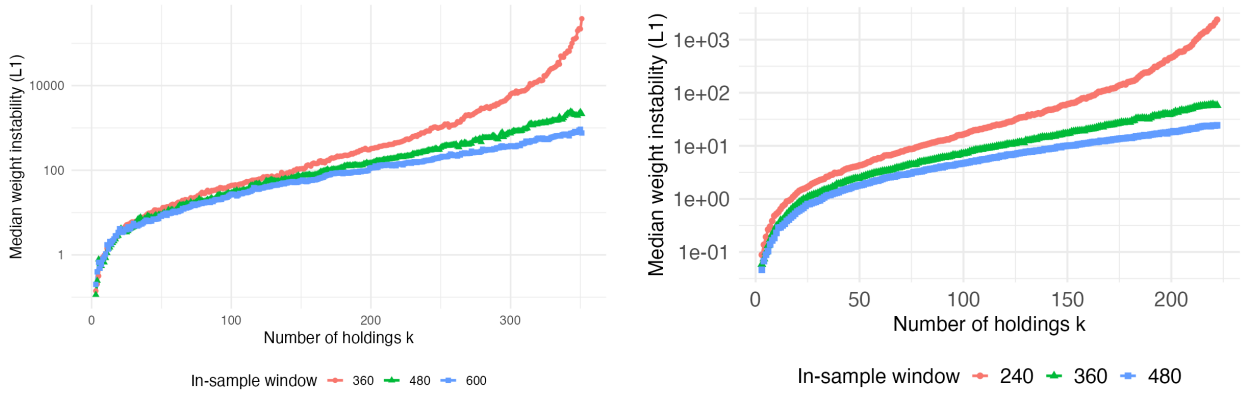
Figures 6 and 7 report diagnostics that speak directly to the two components of estimation risk emphasized by the theory. Selection instability measures how much the selected support varies across rebalancing dates, while weight instability measures how much the refitted weights vary conditional on the selected support. These objects do not estimate Sharpe-ratio losses directly, 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. In the main text we focus on the broad monthly managed universe and the individual-stock panel; the corresponding figures for the remaining datasets are reported in Appendix C.3.3.



(a) Dataset 1

(b) Dataset 4

Figure 6: Selection instability of the ℓ_1 portfolio rule as a function of the target sparsity level k for Dataset 1 and Dataset 4. The holding period is $h = 1$ month.



(a) Dataset 1

(b) Dataset 4

Figure 7: Weight instability (conditional on the selected support) of the ℓ_1 portfolio rule as a function of k for Dataset 1 and Dataset 4. Reported on a log scale and summarized by the time-series median. The holding period is $h = 1$ month.

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, 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 assets, most holdings remain the same. Second, when many assets are close substitutes (e.g., within style or industry groups), 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 both the broad monthly managed universe and the individual-stock panel, and it also appears in the remaining datasets reported in the appendix.

Weight instability, summarized by the time-series median, moves in the opposite direc-

tion. 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 and weight instability map naturally into portfolio turnover. When the selected support changes across rebalancing dates, the investor must unwind exiting positions and establish entering positions. Even when the support is stable, fluctuations in refitted weights generate trading through rebalancing. Turnover therefore translates the statistical instability diagnostics into a dollar trading metric.

Our main focus is the intrinsic estimation cost of sparsity in frictionless markets, but sparse portfolio rules are also attractive in applied settings because they can reduce trading intensity and transaction-cost exposure. The full turnover analysis is reported in Appendix C.3.4. Empirically, turnover increases with k across representative datasets.

5.3.4 Feasible Tuning Rules

The Sharpe-ratio profiles in Figures 2 and 4 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 therefore conclude the empirical analysis by evaluating two simple, implementable tuning rules. Both select k using only information available at the tuning date. We compare their realized out-of-sample Sharpe ratios to an infeasible best-in-hindsight benchmark.

The two criteria are the LOO and UE scores defined in Appendix B.4; the empirical adaptation is that they are recomputed at annual tuning dates using the most recent rolling in-sample window. LOO selects the k with the highest leave-one-out pseudo out-of-sample Sharpe ratio, while UE selects the k with the highest bias-adjusted squared Sharpe statistic for the support chosen at that tuning date. The selected cardinality is held fixed until the next annual tuning date, while the portfolio weights continue to be re-estimated at each monthly rebalancing date. This makes the tuning exercise fully feasible, since neither score uses future out-of-sample returns.

Appendix C.3.5 reports the empirical implementation, the baseline tuning results, the comparison with the ex-post benchmark, and the selected-cardinality boxplots. The main empirical message is that feasible annual tuning recovers a substantial share of the attainable out-of-sample Sharpe ratio. This is especially true for the UE rule once the in-sample window is sufficiently long, while LOO tends to choose more aggressively sparse portfolios. The difference between the two tuning rules is economically informative. LOO penalizes complexity more strongly and therefore favors smaller supports, while UE typically selects less sparse portfolios because it corrects the within-support Sharpe objective without directly penalizing support-selection instability. The corresponding robustness results are reported in Appendix C.3.6.

6 Concluding Remarks

This paper shows that sparsity can serve as optimal complexity control in high-dimensional portfolio choice, not because the population tangency portfolio is sparse, but because density is statistically costly. Under factor structure, when many assets span similar risks, a moderately sparse rule can approximate the population tangency benchmark while reducing the number of noisy inputs that must be learned from finite data.

The theory formalizes this logic through an estimation–efficiency frontier. For a k-sparse plug-in rule, estimation loss decomposes into selection and allocation components and is of order $\sqrt{k \log N/T}$. Under an approximate factor structure, the population efficiency loss from restricting the support declines as $1/k$ with strong factors and as $N^{1-\zeta}/k$ with weaker factors. Hence sparse portfolios can approach the population mean–variance benchmark whenever $N^{1-\zeta} \ll k \ll T/\log N$. The ℓ_1-regularized rule provides a tractable relaxation that preserves the same economic mechanism and remains robust to bounded pricing errors within the approximate-factor framework.

The simulation evidence in Appendix B and the empirical analysis support this interpretation. In controlled factor-model experiments, increasing the support first improves spanning and then raises estimation error, generating an interior optimum. In realized returns, ℓ_1-regularized portfolios display similar non-monotone Sharpe-ratio profiles across managed-portfolio universes, individual-stock returns, anomaly strategies, and supplementary robustness designs. Feasible leave-one-out and bias-adjusted Sharpe tuning rules also select non-dense supports using only information available at the tuning date.

These findings reinterpret portfolio constraints as statistical regularizers. Cardinality restrictions, norm limits, and related constraints need not be viewed only as responses to frictions or implementation costs; they can improve out-of-sample decision quality by limiting the complexity of the estimated rule. In large, redundant markets, sparse allocations can therefore be both interpretable and nearly efficient relative to the population benchmark.

Future research may extend the framework in several directions. First, developing dynamic sparsity models, in which both the sparsity level k and the active set adjust gradually over time, could connect the results to optimal rebalancing under learning. Second, exploring conditional information structures could clarify how sparsity interacts with time-varying estimation risk. Finally, empirical studies using institutional portfolio data could test whether investors naturally exhibit the degree of sparsity predicted by the theory.

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.

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.

Andrew Y. Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264, 2022.

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.

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.

Victor DeMiguel, Javier Gil-Bazo, Francisco J Nogales, and André AP Santos. Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3):103737, 2023.

Antoine Didisheim, Shikun Barry Ke, Bryan T Kelly, and Semyon Malamud. Apt or "aipt"? the surprising dominance of large factor models. Technical report, National Bureau of Economic Research, 2024.

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.

Kenneth R. French. Data library: Fama/french factors and portfolios. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, 2026. Accessed: May 2026.

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.

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

Oliver Hellum, Theis Ingerslev Jensen, Bryan T Kelly, and Semyon Malamud. Complex modern portfolio theory. *Available at SSRN 6443319*, 2026.

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.

Ron Kaniel, Zihan Lin, Markus Pelger, and Stijn Van Nieuwerburgh. Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 150(1):94–138, 2023.

Bryan Kelly, Semyon Malamud, and Kangying Zhou. The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503, 2024.

Nathan Lassance, Rodolphe Vanderveken, and Frédéric Vrins. Optimal portfolio size under parameter uncertainty. *Journal of Financial and Quantitative Analysis*, pages 1–67, 2025.

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.

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.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Appendix

A Notation

For $N \geq 1$, write $[N] := \{1, \ldots, N\}$. For an index set $A \subset [N]$, let $|A|$ denote its cardinality. For $k \in [N]$, $\binom{N}{k}$ denotes the number of subsets $A \subset [N]$ with $|A| = k$. Consider a generic vector $w \in \mathbb{R}^N$. $w_A \in \mathbb{R}^{|A|}$ denotes the subvector with indices in A, w_i its i-th component, $w^\top$ its transpose, and

$$\operatorname{supp}(w) := \{i \in [N] : w_i \neq 0\}$$

its support. 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 $w \in \mathbb{R}^N$, 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 $z \in \mathbb{R}$, the sign function is $\operatorname{sign}(z) = 1$ if $z > 0$, $\operatorname{sign}(z) = -1$ if $z < 0$, and $\operatorname{sign}(0) = 0$.

Consider a generic rectangular matrix $B \in \mathbb{R}^{N \times L}$ and a generic square matrix $M \in \mathbb{R}^{N \times N}$. $M_A \in \mathbb{R}^{|A| \times |A|}$ denotes the principal submatrix obtained by retaining rows and columns in A. $\lambda_i(M)$, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote, respectively, the i-th eigenvalue, the maximum and the minimum eigenvalue of M. $B_A \in \mathbb{R}^{|A| \times L}$ denotes the submatrix formed by retaining the rows indexed by A. The entrywise maximum norm and the matrix norm induced by the vector ℓ_1 norm of B are denoted, respectively,

$$\|B\|_{\max} := \max_{i,j} |B_{ij}|, \qquad \|B\|_{1 \to 1} := \sup_{\|x\|_1 = 1} \|Bx\|_1,$$

while its operator norm is

$$\|B\|_{\mathrm{op}} := \sup_{x \in \mathbb{R}^L \setminus \{0\}} \frac{\|Bx\|_2}{\|x\|_2} = \sup_{\|x\|_2 = 1} \|Bx\|_2 = \sqrt{\lambda_{\max}(B^\top B)}.$$

If M is symmetric, then $\|M\|_{\mathrm{op}} = \max_i |\lambda_i(M)|$. If M is positive semidefinite, we write $M \succeq 0$, and its operator norm simplifies to $\|M\|_{\mathrm{op}} = \lambda_{\max}(M)$; if it is positive definite, we write $M \succ 0$. For a symmetric matrix M' of the same dimension as M, $M \preceq M'$ means $M' - M$ is positive semidefinite, and $M \prec M'$ means $M' - M$ is positive definite. A Cholesky factorization of a positive definite matrix M is $M = C^\top C$, where C is upper triangular with positive diagonal entries. The identity matrix of dimension N is I_N.

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 $\limsup_{n\to\infty} |a_n|/b_n < \infty$, and $a_n = o(b_n)$ means $|a_n|/b_n \to 0$. Convergence in probability is denoted by $\xrightarrow{p}$. For random sequences (X_n) and deterministic (b_n) with $b_n > 0$, $X_n = O_p(b_n)$ means that X_n/b_n is bounded in probability; equivalently, for every $\varepsilon > 0$ there exists $M < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}(|X_n/b_n| > M) \leq \varepsilon.$$

$X_n = o_p(b_n)$ 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 $a_n = o(b_n)$. The notation $a_n \gg b_n$ means $b_n \ll a_n$, and $a_n \asymp b_n$ means both $a_n = O(b_n)$ and $b_n = O(a_n)$. The distributional notation $r_t \sim \mathcal{N}(\mu, \Sigma)$ means that r_t is Gaussian with mean μ and covariance Σ. We write $\mathbb{P}$, $\mathbb{E}$, and $\mathbb{V}$ for probability, expectation operator, and variance-covariance operator.

B Simulation Analysis

A calibrated three-factor Monte Carlo design allows us to measure the estimation–efficiency tradeoff directly. Returns are generated from a Fama–French-style model in a strong-factor environment, so the population moments are known and the opportunity cost of sparsity can be separated from the finite-sample cost of estimating sparse portfolio weights. We then trace how these two losses vary with the target cardinality k and the sample size T, and we study feasible rules for selecting k from the data.

B.1 Data-Generating Process and Calibration

We work with a three-factor structure in the spirit of the Fama–French model. For each asset $i = 1, \dots, N$ and month $t = 1, \dots, T$, 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 $r_t = (r_{1t}, \dots, r_{Nt})^\top$, $\alpha = (\alpha_1, \dots, \alpha_N)^\top$, 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 typical U.S. equity-market magnitudes:

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

Idiosyncratic shocks are homoskedastic Gaussian,

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

independently across i and t. For the cross-sectional size considered below, we draw intercepts and factor loadings once and keep them fixed over time and across Monte Carlo replications. 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 $\beta_{\text{mean}} = (1,0,0)^\top$. This design corresponds to the strong-factor case, so the theory predicts a rapidly declining efficiency loss $\theta^* - \theta_k^*$ and an increasing finite-sample estimation loss $\theta_k^* - \theta(\widehat{w}_k)$ as k rises.

B.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. Conditional on the population parameters described above, we run 10,000 independent Monte Carlo replications for each T, simulating $\{r_t\}_{t=1}^T$ from the factor model and then forming the plug-in estimators $\widehat{\mu}$ and $\widehat{\Sigma}$ from the simulated sample.

For each cardinality level $k \in [N]$, we benchmark the sample portfolio $\widehat{w}_k$ against the unrestricted population optimum $\theta^* = \max_{w \in \mathbb{R}^N} \theta(w)$ through the decomposition

$$\theta^* - \theta(\widehat{w}_k) = \underbrace{\left[\theta^* - \theta_k^*\right]}_{\text{Efficiency loss}} + \underbrace{\left[\theta_k^* - \theta(\widehat{w}_k)\right]}_{\text{Estimation loss}}. \tag{B.1}$$

The first term, $\theta^* - \theta_k^*$, is the opportunity cost of imposing $\|w\|_0 \leq k$ even under perfect knowledge of (μ, Σ). The second term, $\theta_k^* - \theta(\widehat{w}_k)$, is the incremental loss induced by using estimated moments.

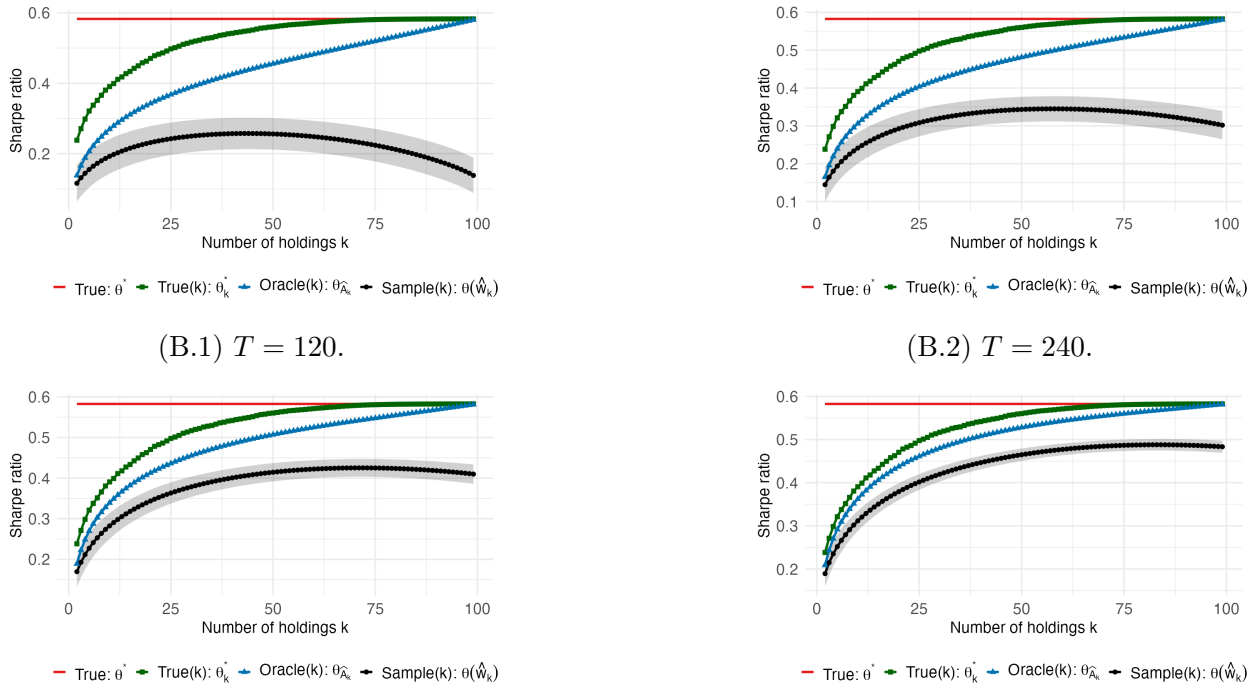
The population benchmark at sparsity level k is $\theta_k^* := \max_{\|w\|_0 \leq k} \theta(w)$, as in (5). In each replication, θ_k^* is computed from the true moments (μ, Σ) using the same numerical approach used for the sample problem. Given $(\widehat{\mu}, \widehat{\Sigma})$, we compute a sparse sample portfolio $\widehat{w}_k$ using the numerical implementation described in Appendix E, targeting cardinality k along the ℓ_1 regularization path. We evaluate its out-of-sample performance using the population Sharpe ratio $\theta(\widehat{w}_k)$ computed under the true (μ, Σ).

To separate the channels inside estimation loss, we also compute the selected support $\widehat{A}_k := \{i \in [N] : \widehat{w}_{k,i} \neq 0\}$ and the oracle Sharpe ratio on that support, defined in (9). This yields the decomposition into selection loss $\theta_k^* - \theta_{\widehat{A}_k}$ and allocation loss $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)$, as in (10). Because the exact cardinality problem is NP-hard, we compute $\widehat{w}_k$ using the numerical approximations described in Appendix E.

B.3 Simulation Results

The simulations confirm the main comparative statics of the theory. As k increases, the population-efficient sparse frontier rises and quickly approaches the unrestricted benchmark, reflecting diminishing marginal gains from adding assets in a factor-structured market. By contrast, the sample sparse portfolio has a hump-shaped Sharpe-ratio profile: performance initially improves with k and then deteriorates once estimation error begins to dominate.

Sharpe-ratio profiles. Figure B.1 reports the main Sharpe-ratio profiles. The unrestricted population optimum θ^* is constant in k, while the population-efficient sparse frontier θ_k^* increases with diminishing marginal gains. The oracle curve $\theta_{\widehat{A}_k}$ moves toward θ_k^* as T increases, indicating improved support selection in larger samples. By contrast, the sample curve $\theta(\widehat{w}_k)$ is non-monotone: it rises when the efficiency gain from additional holdings dominates, and falls when the estimation cost of learning a higher-dimensional portfolio becomes large. The location of the hump shifts to the right as T increases, consistent with larger samples supporting more complex portfolios.

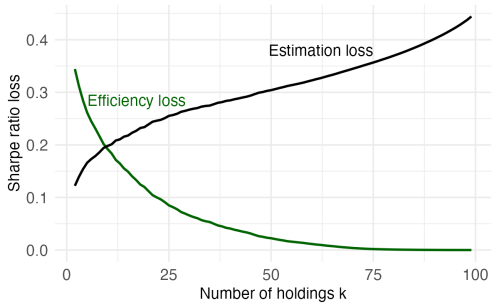


(B.1) $T = 120$. (B.2) $T = 240$.

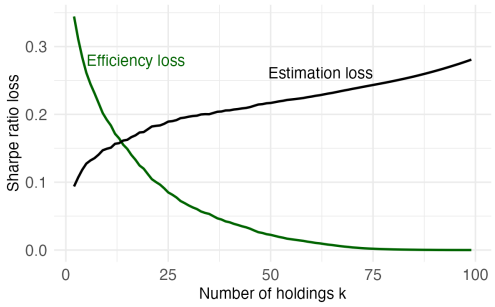
(B.3) $T = 480$. (B.4) $T = 960$.

Figure B.1: **Sharpe-ratio profiles of sparse portfolios.** Population Sharpe ratios as a function of cardinality k for $N = 100$ assets under the Fama–French three-factor calibration, for each in-sample length T. The curves report the unrestricted population optimum θ^*, the k-sparse population frontier θ_k^*, the average oracle frontier $\theta_{\widehat{A}_k}$ on the sample-selected support, and the average population Sharpe ratio $\theta(\widehat{w}_k)$ of the implemented sparse sample portfolio. For $\theta(\widehat{w}_k)$, the gray ribbon shows ± 1 standard error around the Monte Carlo average. Results are based on 10,000 replications.

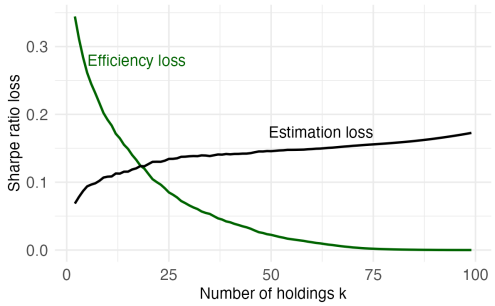
Efficiency and estimation losses. Figure B.2 plots the two components in (B.1). The efficiency loss $\theta^* - \theta_k^*$ is a population object and is therefore invariant to T; in the strong-factor calibration, it falls rapidly with k. The estimation loss $\theta_k^* - \theta(\widehat{w}_k)$ increases with k, because more active positions require estimating a larger mean–variance problem, and it shifts downward as T increases. The crossing of these two forces explains both the hump in Figure B.1 and the movement of the performance-maximizing cardinality toward larger k at longer estimation windows.



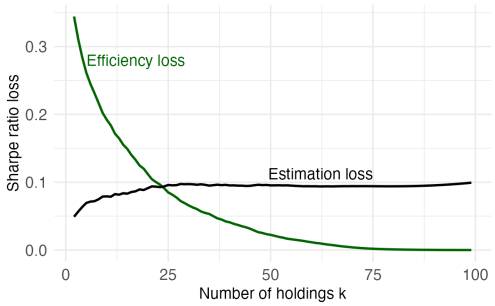
(B.1) $T = 120$.



(B.2) $T = 240$.



(B.3) $T = 480$.

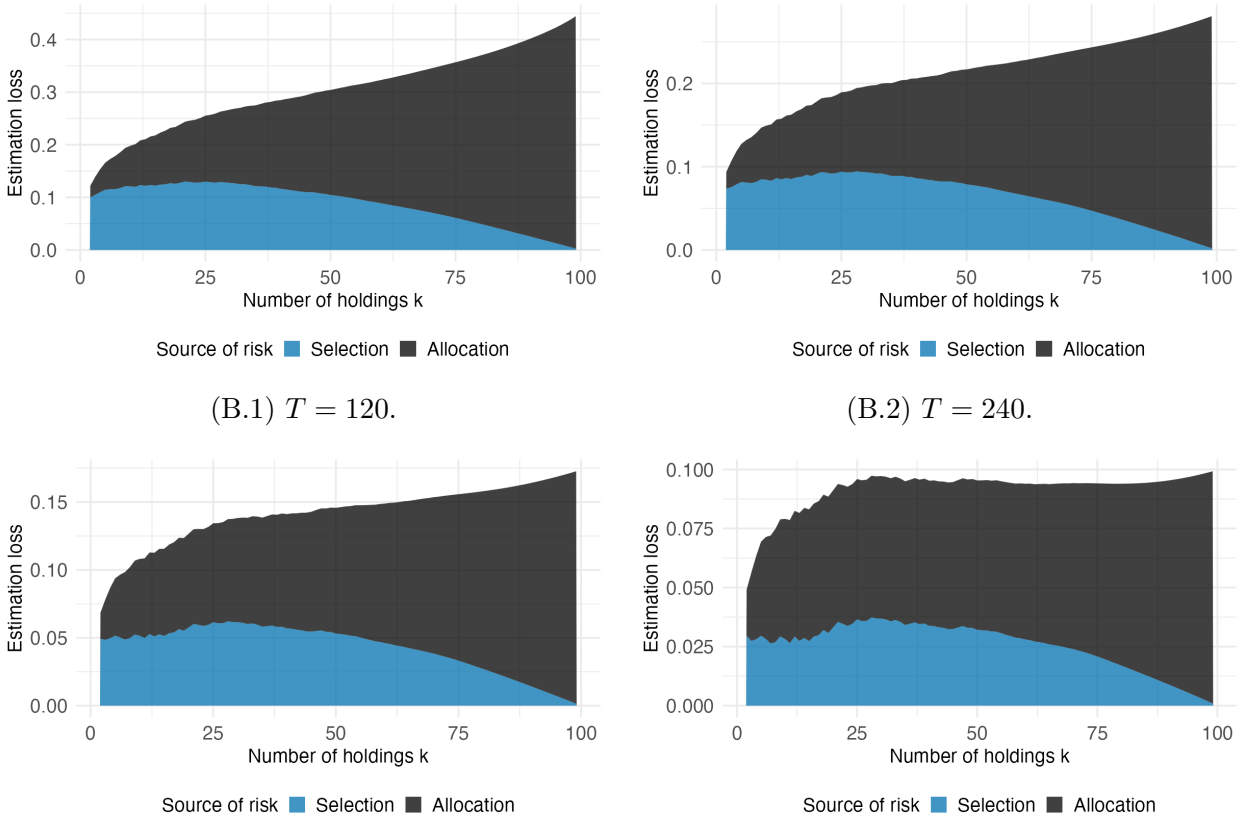


(B.4) $T = 960$.

Figure B.2: **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 three-factor calibration and for each sample size T. The estimation loss is averaged over 10,000 Monte Carlo replications.

Selection–allocation decomposition. Figure B.3 decomposes the estimation loss into selection and allocation components as in (10). The selection component $\theta_k^* - \theta_{\widehat{A}_k}$ is not monotone in k: it rises initially as the support search problem becomes harder, then falls once

supports overlap heavily and the population sparse frontier becomes flatter. By contrast, allocation loss $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k)$ increases with k, reflecting the greater sensitivity of higher-dimensional plug-in mean–variance weights to sampling noise. Overall, selection dominates at lower sparsity levels, while allocation becomes the main source of estimation loss at larger k.



(B.1) $T = 120$. (B.2) $T = 240$.

(B.3) $T = 480$. (B.4) $T = 960$.

Figure B.3: **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 three-factor calibration and for each sample size T. Curves are averaged over 10,000 Monte Carlo replications.

B.4 Optimal Sparsity and Feasible Tuning Rules

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

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

namely the number of holdings that maximizes the population performance of the implemented sparse sample portfolio. Figure B.4 reports the distribution of k^* across replications for different in-sample window lengths. The boxplots shift monotonically to the right as T increases, and dispersion falls with sample size. This matches the rate balance in Remark 5: longer samples support richer portfolios and sharpen the concentration of the optimal sparsity level.

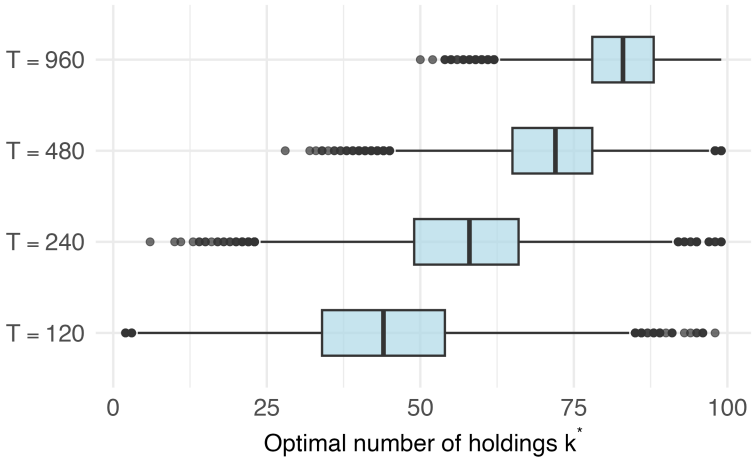


Figure B.4: **Ex-post optimal sparsity k^*.** Boxplots across Monte Carlo replications under the Fama–French three-factor calibration with $N = 100$, for each sample size T.

While informative, k^* is an *infeasible* benchmark: it depends on the population Sharpe ratio of the implemented 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 $\max_k \theta(\widehat{w}_k)$. LOO is a direct cross-validation criterion, while UE uses the finite-sample bias correction for squared Sharpe ratios motivated by the Gaussian mean–variance analysis

of Kan and Zhou (2007); Kan and Smith (2008).

- *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 within the in-sample window, 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 $\widehat{k}_{\mathrm{LOO}}$ 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

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

  and collect these pseudo out-of-sample realizations in

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

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

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

  and select

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

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

- *Bias-adjusted Sharpe 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; this correction follows the exact Gaussian finite-sample logic in Kan and Zhou (2007); Kan and Smith (2008).[7] We then select

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

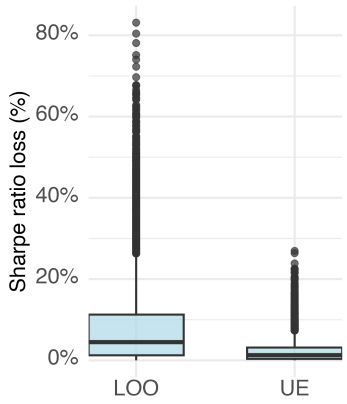
In both cases, after choosing $\widehat{k}$ we refit the sample optimizer at that sparsity level and evaluate its population performance $\theta(\widehat{w}_{\widehat{k}})$ in the Monte Carlo design.

Figure B.5 reports the distribution across the replications (for $T = 240$) of the percentage population Sharpe ratio loss relative to the infeasible benchmark,

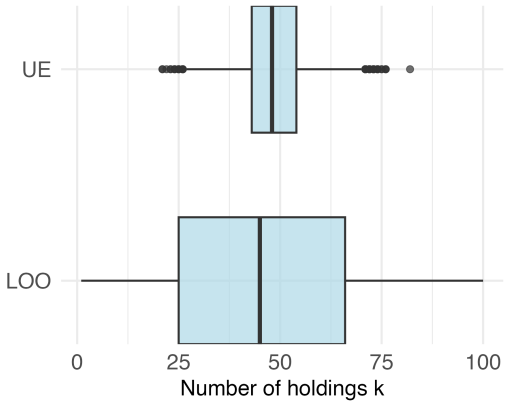
$$\frac{\max_k \theta(\widehat{w}_k) - \theta(\widehat{w}_{\widehat{k}})}{\max_k \theta(\widehat{w}_k)},$$

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 boxplots also show that both rules favor interior sparsity levels, with UE concentrating more tightly on the middle of the k-grid while LOO exhibits substantially more dispersion. These results suggest that, in this strong-factor calibration, the bias-adjusted Sharpe criterion can provide a stable feasible tuning rule that tracks the best-in-hindsight sparsity level closely.

[7] The UE rule is not "unbiased" for the ex-post optimal sparsity level k^*. Rather, it relies on a 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.



(B.1) Percentage population Sharpe ratio loss relative to $\max_k \theta(\widehat{w}_k)$.



(B.2) Selected cardinality $\widehat{k}$.

Figure B.5: **Feasible tuning rules for sparsity.** Boxplots across 10,000 Monte Carlo replications under the Fama–French three-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 bias-adjusted Sharpe estimator (UE). The right panel reports the distribution of selected portfolio cardinalities $\widehat{k}$ for the same two rules.

C Complementary Empirical Analysis

This appendix collects empirical material omitted from the main text. We first describe the supplementary datasets used only in the appendix. Table C.1 then reports the common out-of-sample evaluation window size for each dataset and in-sample window length. After that, we provide the formal definitions of the selection-instability, weight-instability, and turnover diagnostics. We then report additional out-of-sample Sharpe-ratio profiles, robustness checks based on factor augmentation and gross-exposure normalization, remaining instability diagnostics, turnover profiles, and feasible tuning-rule results.

C.1 Supplementary Data Used in the Appendix

The appendix reports additional empirical results for three monthly test-asset universes that are not used in the main-text figures. They are included to show that the sparsity tradeoff is stable across simpler size–value benchmarks, size–value universes augmented with industry

portfolios, and international managed-portfolio returns.

5. **Dataset 5: 100 size–value portfolios (N=100, monthly).** This baseline supplementary dataset consists of 100 U.S. characteristic-managed portfolios formed by double-sorting equities on size and book-to-market. The portfolio return series are obtained from the Kenneth R. French Data Library (French, 2026). We convert returns to excess returns by subtracting the one-month Treasury bill rate supplied by the same data library. The sample runs from January 1971 through October 2025.

6. **Dataset 6: Dataset 5 + 49 industry portfolios (N=149, monthly).** To enrich the size–value benchmark with industry tilts, we augment Dataset 5 with 49 U.S. industry portfolios from the Kenneth R. French Data Library. The industry portfolio returns are observed over the same period and are converted to excess returns using the same one-month Treasury bill rate. The resulting universe contains $100+49 = 149$ monthly excess-return series.

7. **Dataset 7: International size–value and size–profitability portfolios (N=200, monthly).** To study whether the sparsity tradeoff carries over to international 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 return series. The series are obtained from the Kenneth R. French Data Library and are converted to excess returns using the risk-free-rate convention supplied with the corresponding international factor data. The sample runs from November 1990 through October 2025.

For factor-augmentation robustness checks, we append traded Fama–French factor-mimicking portfolios to these supplementary universes. For Datasets 5 and 6, we append the U.S. Fama–French five-factor set (MKT, SMB, HML, RMW, CMA). The resulting augmented universes

have 105 and 154 assets, respectively. For Dataset 7, we append region-specific Fama–French three-factor returns for North America, Europe, Japan, and Asia Pacific (ex Japan). The augmented international design therefore includes four copies of (MKT, SMB, HML), one for each region, and has $200 + 4 \times 3 = 212$ assets. The factor-mimicking portfolios are already excess-return or self-financing factor returns, so they are appended without an additional risk-free-rate subtraction.

Table C.1: **Out-of-sample evaluation-window sizes.** For each dataset, all in-sample window lengths are evaluated over the same out-of-sample period, of length $T_0 - T_{\max}$, where $T_{\max} := \max \mathcal{T}$.

Dataset	In-sample window	Out-of-sample window
Dataset 1	360 months	36 months
	480 months	36 months
	600 months	36 months
Dataset 2	360 days	7,964 days
	480 days	7,964 days
	600 days	7,964 days
Dataset 3	240 months	168 months
	360 months	168 months
	480 months	168 months
Dataset 4	240 months	169 months
	360 months	169 months
	480 months	169 months
Dataset 5	240 months	178 months
	360 months	178 months
	480 months	178 months
Dataset 6	240 months	178 months
	360 months	178 months
	480 months	178 months
Dataset 7	240 months	60 months
	360 months	60 months

C.2 Selection Instability, Weight Instability, and Turnover

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

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

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

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

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.[8] 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.

[8] In practice, rebalancing rules are often implemented with buffers: a currently held asset may be retained even if it is no longer selected under the updated optimization, provided it remains close to the selection margin. Such buffering reduces unnecessary turnover at the cost of occasionally retaining a slightly suboptimal constituent.

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

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

We summarize instability across time by the corresponding medians,[9]

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

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.

Turnover. We report turnover of sparse portfolios as a function of the target sparsity level k. 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

[9] 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).

returns and before rebalancing) are:[10]

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

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

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

Under proportional transaction costs, if c denotes a one-way cost rate, then the period-$t+1$ trading cost is approximately $c\,\mathrm{TO}_{t+1}(T,k)$, 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,

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

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.

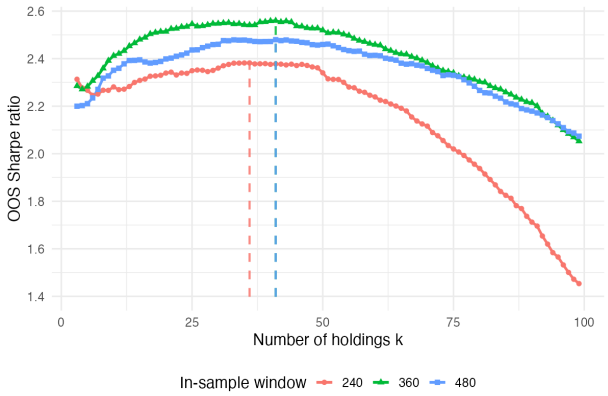
C.3 Complementary Empirical Evidence

C.3.1 Out-of-Sample Sharpe-Ratio Profiles

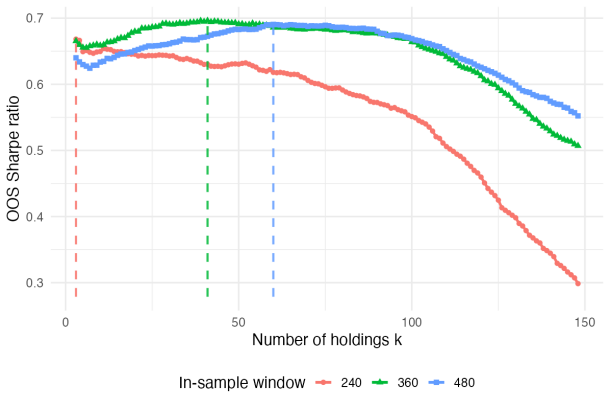
This subsection reports the baseline out-of-sample Sharpe-ratio profiles for the datasets omitted from the main-text discussion in Section 5.3.1. Specifically, we display the size–value benchmark (Dataset 5), the size–value–industry universe (Dataset 6), and the international monthly managed-portfolio universe (Dataset 7). The same qualitative non-monotonicity

[10] The drift formula applies directly to budget-normalized asset portfolios. For long–short or self-financing strategy portfolios, we interpret the resulting turnover measure as an implementation proxy after applying the portfolio's normalization convention; it should be read as trading intensity in strategy weights rather than physical share turnover.

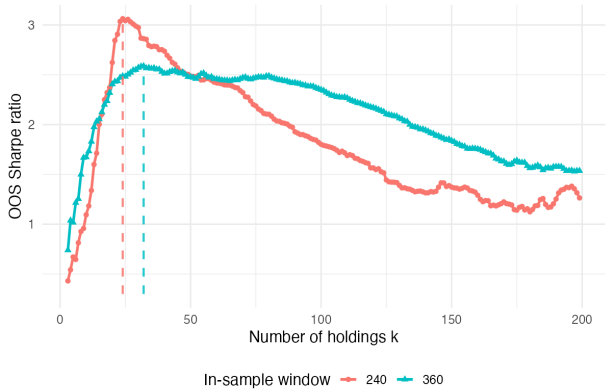
appears in each case: Sharpe ratios improve when moving away from very sparse supports and eventually decline once portfolio complexity becomes sufficiently high.



(C.1) Dataset 5



(C.2) Dataset 6



(C.3) Dataset 7

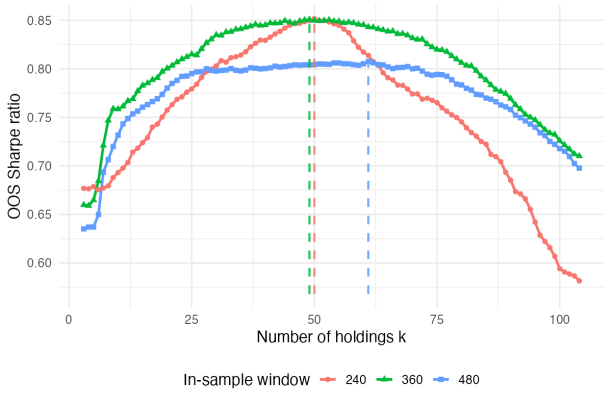
Figure C.1: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule, plotted as a function of the target sparsity level k for the monthly datasets omitted from the main-text Sharpe-ratio discussion. The holding period is $h = 1$, and all OOS quantities are computed over a common evaluation window within each dataset.

C.3.2 Robustness Checks

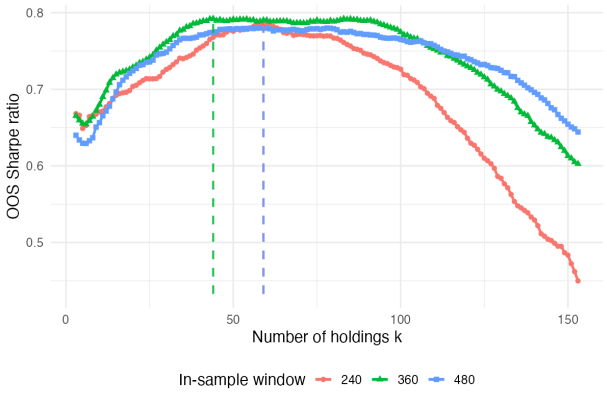
We report the remaining robustness checks for the out-of-sample Sharpe-ratio profiles discussed in Section 5.3.2. We display results for the size–value benchmark (Dataset 5), the size–value–industry universe (Dataset 6), the international monthly managed-portfolio universe (Dataset 7), and the U.S. daily managed-portfolio universe (Dataset 2). The anomaly

and individual-stock universes are left unaugmented because the goal in those exercises is to study sparsity across strategy payoffs and individual securities rather than explicit factor-spanning robustness.

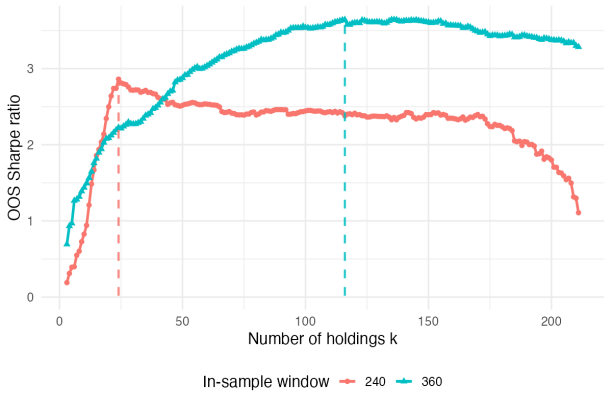
Figure C.2 augments the relevant universes with the corresponding Fama–French factor-mimicking portfolios, while Figure C.3 reports the same analysis after normalizing realized portfolio weights to satisfy $\|w\|_1 = 1$. In all cases, the qualitative patterns mirror those in the main text.



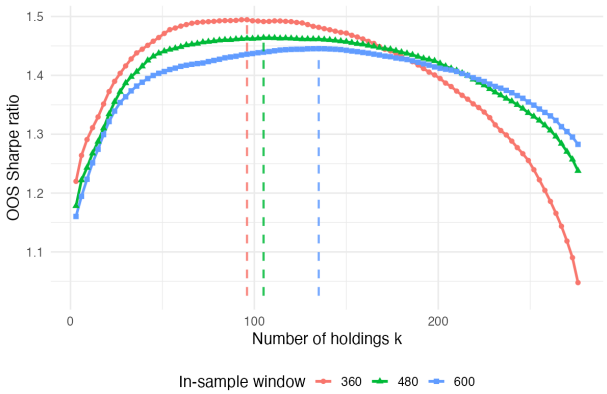
(C.1) Dataset 5



(C.2) Dataset 6

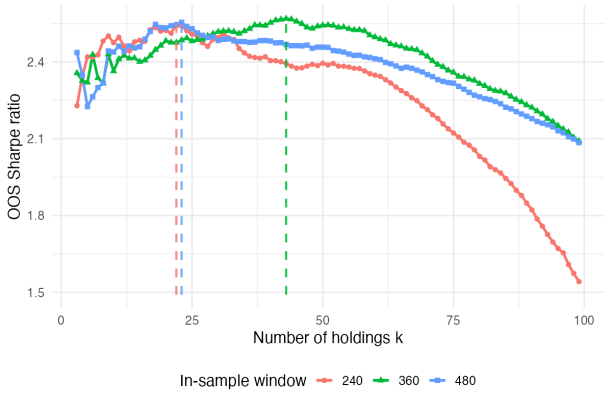


(C.3) Dataset 7

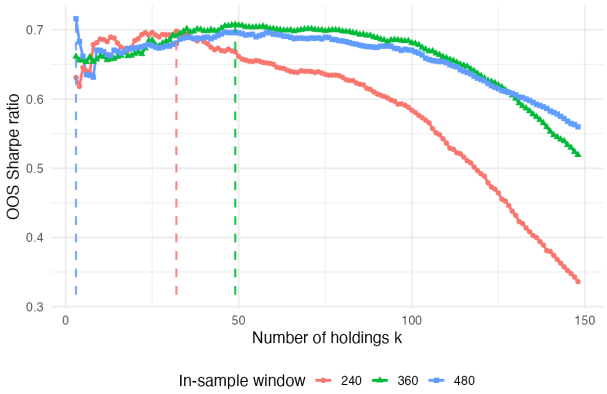


(C.4) Dataset 2

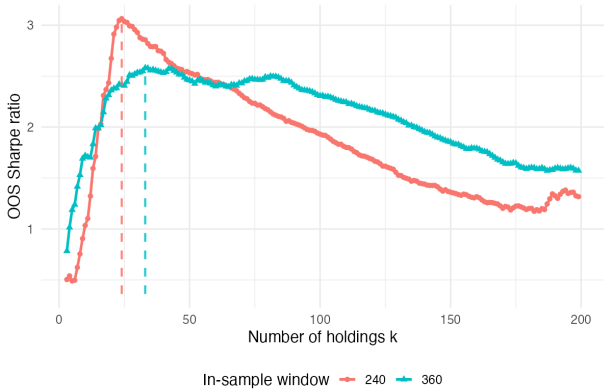
Figure C.2: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule, plotted as a function of the target sparsity level k for the datasets augmented with Fama–French factor-mimicking portfolios. The holding period is $h = 1$.



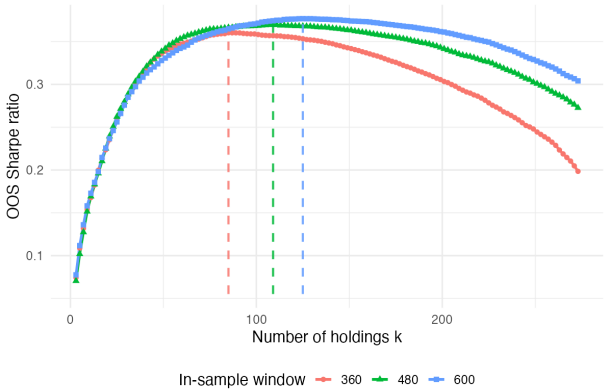
(C.1) Dataset 5



(C.2) Dataset 6



(C.3) Dataset 7



(C.4) Dataset 2

Figure C.3: Out-of-sample Sharpe ratio of the ℓ_1 portfolio rule, plotted as a function of the target sparsity level k, with realized portfolio weights normalized so that $\|w\|_1 = 1$. The holding period is $h = 1$.

C.3.3 Selection and Weight Instability

We now report the selection- and weight-instability diagnostics for the datasets omitted from the main text presentation in Section 5.3.3. The qualitative patterns mirror those in the main text, although the anomaly universe is the case in which selection instability remains relatively flat and erratic until large k.

C.3.4 Turnover

Figure C.6 reports turnover profiles for four representative datasets: the broad U.S. monthly managed-portfolio universe (Dataset 1), the U.S. daily managed-portfolio universe (Dataset 2), the anomaly long–short universe (Dataset 3), and the size–value benchmark (Dataset 5). Across the reported panels and in-sample window lengths, median one-way turnover increases with k. This is consistent with the allocation channel documented in the instability diagnostics: as the target support expands, more active weights must be re-estimated and rebalanced at each date, so small changes in estimated mean–variance inputs translate into more trading. In the anomaly universe, the same logic applies at the strategy level, where turnover reflects reallocations across selected long–short anomaly portfolios and changes in their relative weights.

C.3.5 Feasible Tuning Rules

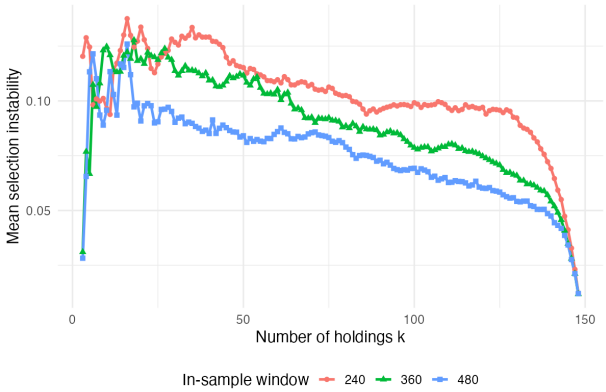
For a fixed window length $T \in \mathcal{T}$, define the realized out-of-sample Sharpe ratio $\widehat{\theta}^{\text{oos}}(T,k)$, as in Section 5.1.1. The *ex-post* benchmark selects the best-performing cardinality in hindsight,

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

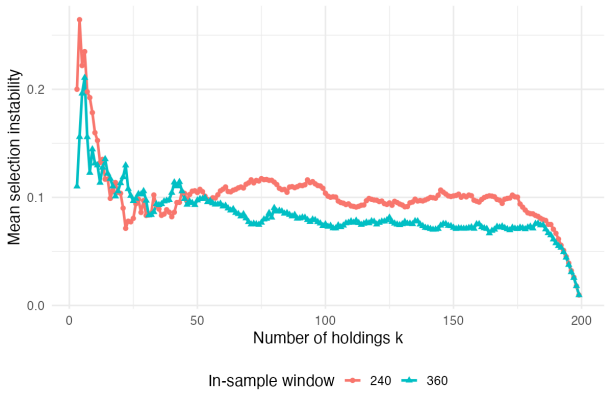
This benchmark is not implementable because it uses the same out-of-sample returns that it evaluates.

We implement feasible tuning for the monthly datasets by updating k once per calen-

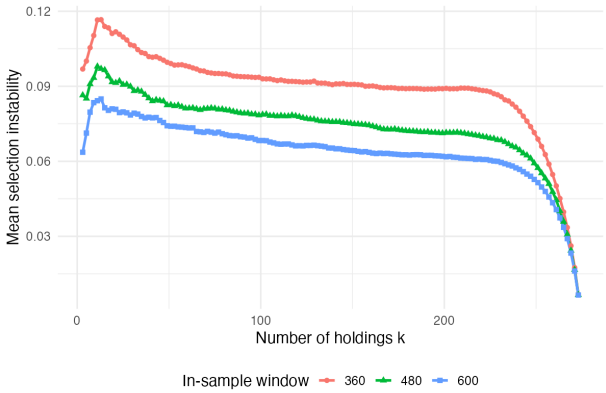
(C.1) Dataset 5.



(C.2) Dataset 6.



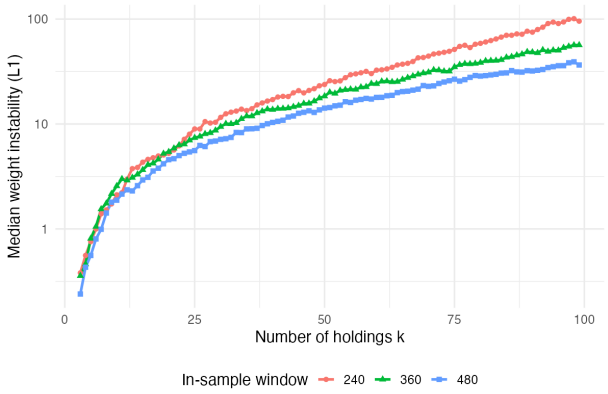
(C.3) Dataset 7.



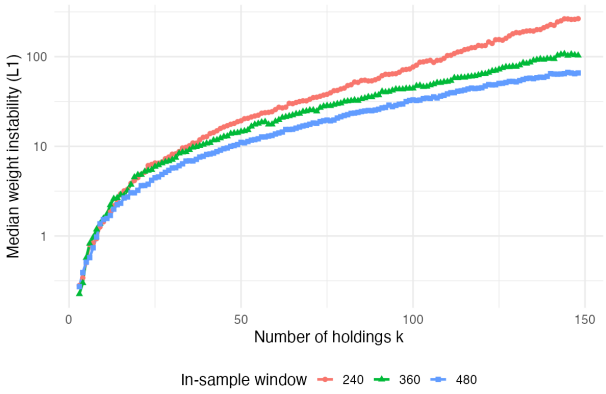
(C.4) Dataset 2.

(C.5) Dataset 3.

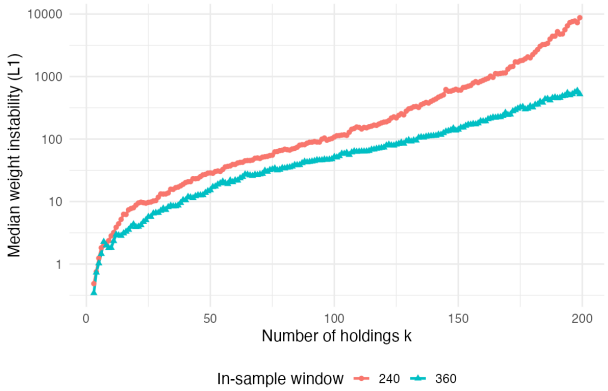
Figure C.4: Selection instability of the ℓ_1 portfolio rule as a function of the target sparsity level k for the datasets omitted from the main-text instability figures. The holding period is $h = 1$ (one trading day for Dataset 2 and one month for the other datasets).



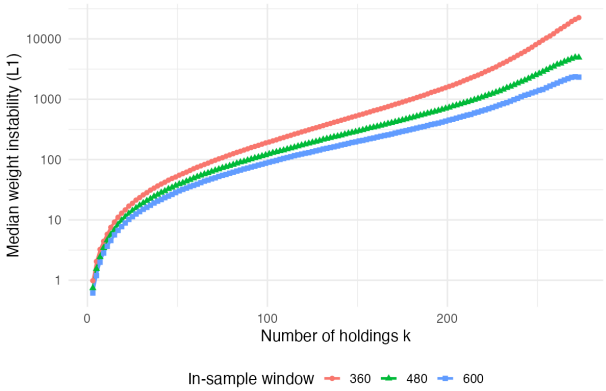
(C.1) Dataset 5.



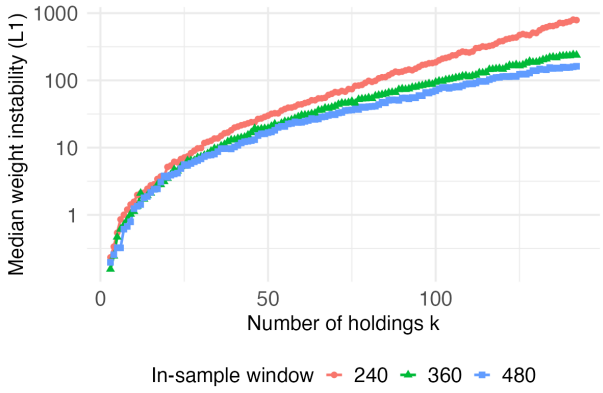
(C.2) Dataset 6.



(C.3) Dataset 7.

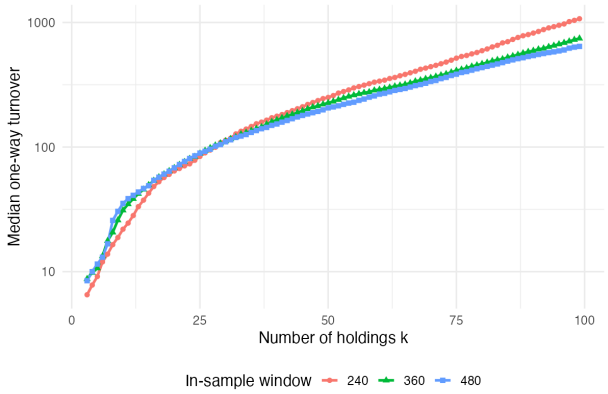


(C.4) Dataset 2.

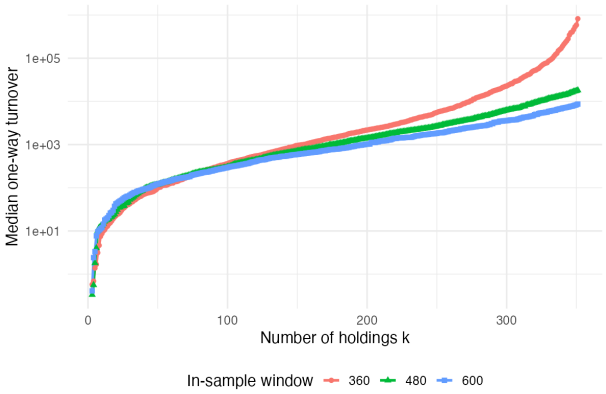


(C.5) Dataset 3.

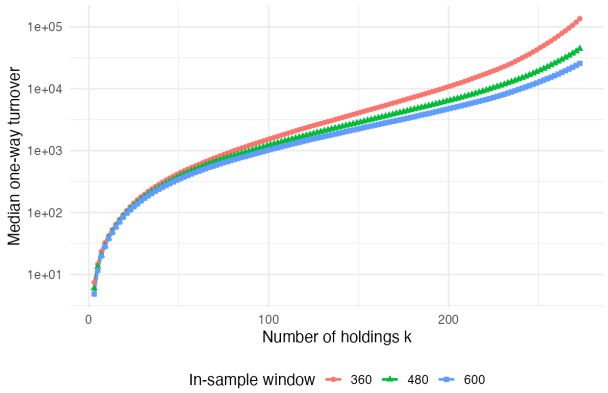
Figure C.5: Weight instability (conditional on the selected support) of the ℓ_1 portfolio rule as a function of k for the datasets omitted from the main-text instability figures. Reported on a log scale and summarized by the time-series median. The holding period is $h = 1$ (one trading day for Dataset 2 and one month for the other datasets).



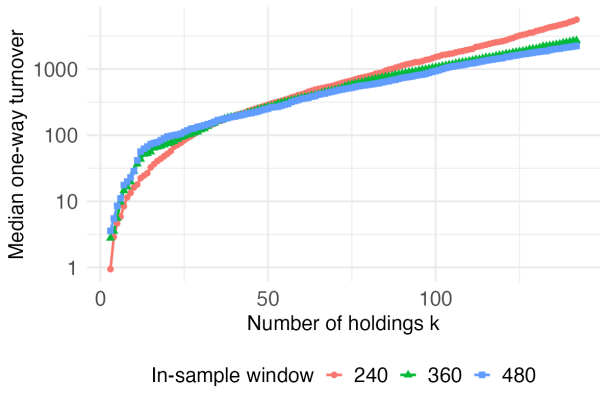
(C.1) Dataset 5



(C.2) Dataset 1



(C.3) Dataset 2



(C.4) Dataset 3

Figure C.6: Median one-way turnover of the ℓ_1 portfolio rule as a function of the target sparsity level k. The monthly panels use holding period $h = 1$ month; the daily panel uses holding period $h = 1$ trading day.

dar 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, for a tuning rule $q \in \{\mathrm{LOO}, \mathrm{UE}\}$, the realized out-of-sample return series is

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

and its realized Sharpe ratio $\widehat{\theta}^{\mathrm{oos}}_{q}(T)$ is computed from the concatenated sequence over the common evaluation period, annualized as in Section 5.1.1. At each tuning date τ_j and for each candidate k, we compute the same two feasible scores defined in Appendix B.4, using the most recent empirical in-sample window in place of the Monte Carlo sample. The LOO score is the leave-one-out pseudo out-of-sample Sharpe ratio computed within that window, and the UE score is the bias-adjusted squared Sharpe statistic on the support selected by the k–sparse procedure at τ_j. Equivalently,

$$\widehat{k}^{\mathrm{LOO}}_j(T) \in \operatorname*{argmax}_{k\in[N]} \widehat{\theta}_{\mathrm{LOO},\tau_j}(T,k), \qquad \widehat{k}^{\mathrm{UE}}_j(T) \in \operatorname*{argmax}_{k\in[N]} \tilde{\theta}^{2}_{\widehat{A}_{T,k,\tau_j}},$$

where the construction of $\widehat{\theta}_{\mathrm{LOO},\tau_j}(T,k)$ and $\tilde{\theta}^{2}_{\widehat{A}_{T,k,\tau_j}}$ follows Appendix B.4. 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 C.2 reports the infeasible ex-post benchmark and the realized Sharpe ratios under annual LOO and UE tuning for two supplementary monthly managed-portfolio datasets: Dataset 6 and Dataset 7. Figures C.7 and C.8 report the selected-cardinality behavior for these datasets and for the anomaly and individual-stock universes. 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 C.3.6.

Comparing Table C.2 to the corresponding fixed-k Sharpe-ratio profiles in Figure C.1 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 6), 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 7), 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 C.7 and C.8 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.

For Datasets 3 and 4, the selected-cardinality boxplots have complementary interpretations. In the anomaly universe, the tuning rules choose the number of long–short anomaly strategies to combine; in the individual-stock universe, they choose the number of individual stocks to hold. The qualitative pattern is similar across the two return spaces, although the individual-stock panel generally favors much sparser selected supports. In both cases, the rules select finite, non-dense portfolios, reinforcing the main implementation message that data-driven sparsity can control estimation risk without collapsing to a single-asset or single-strategy allocation.

Taken together, the table and figure results indicate 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.

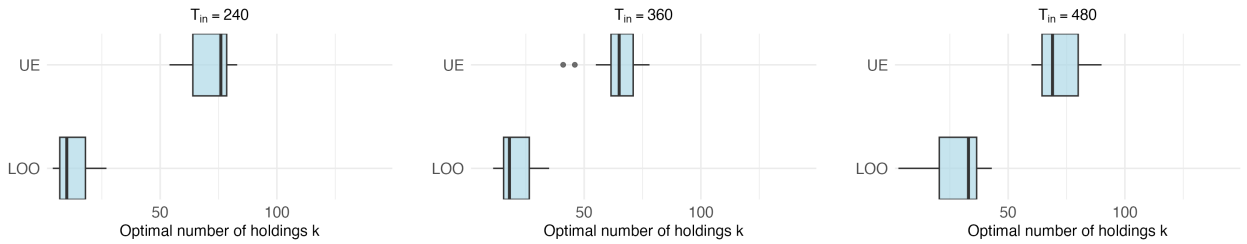
Table C.2: **Feasible tuning rules (annual tuning, monthly data).** For each in-sample window length T_{in}, the table reports the infeasible best-in-hindsight Sharpe ratio $\widehat{\theta}^{\text{oos}}_{\max}(T)$ (max over k), and the realized out-of-sample Sharpe ratios under annual tuning of k via leave-one-out (LOO) and the bias-adjusted Sharpe estimator (UE). All Sharpe ratios are annualized. Baseline design: no factor augmentation and no ex-post gross-exposure normalization.

(C.1) Dataset 6

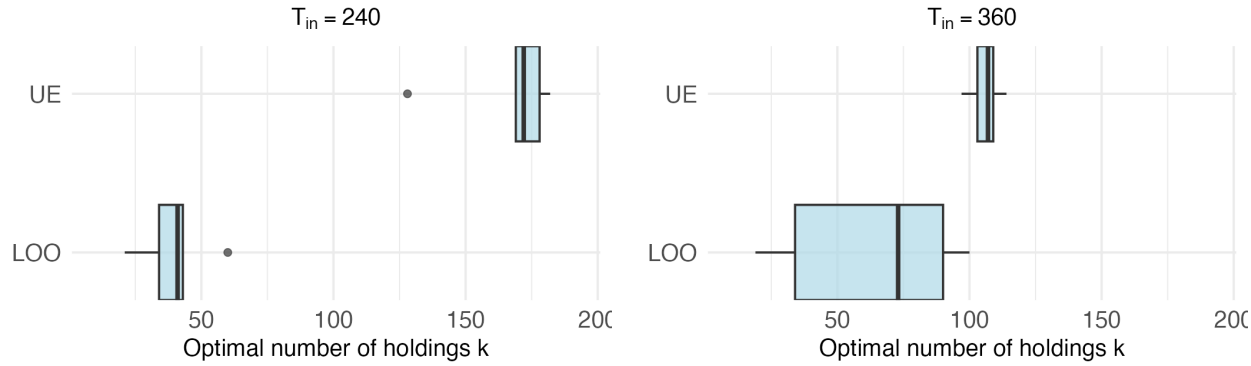
T_{in}	$\widehat{\theta}^{\text{oos}}_{\max}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	0.6682	0.6356	0.5965
360	0.6961	0.6750	0.6896
480	0.6905	0.6549	0.6743

(C.2) Dataset 7

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

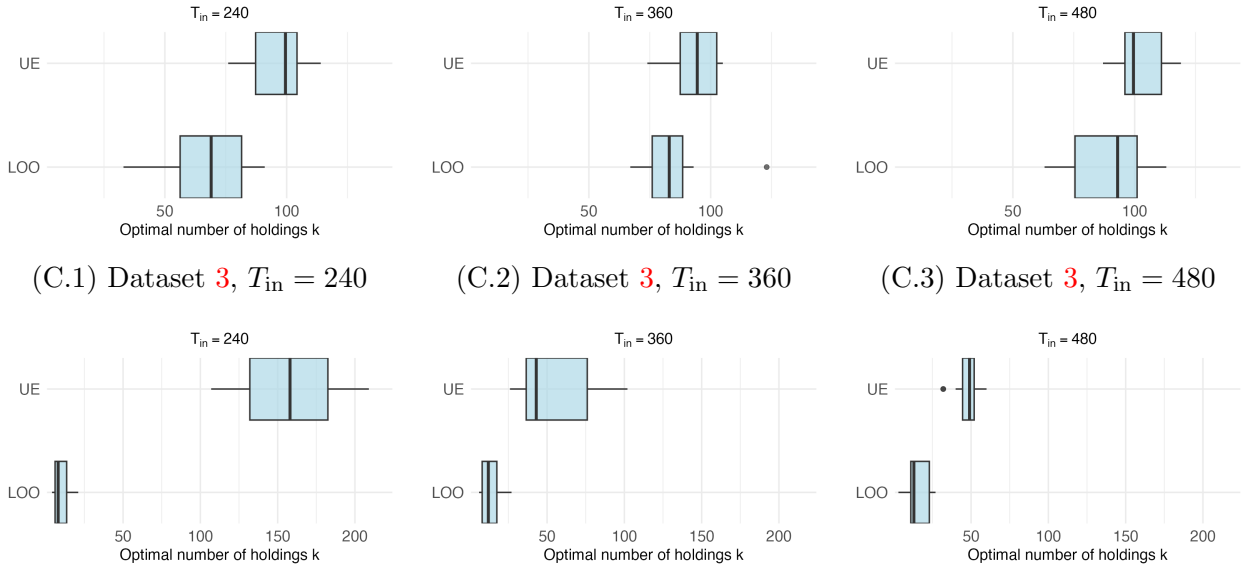


(C.1) Dataset 6, $T_{\text{in}} = 240$ (C.2) Dataset 6, $T_{\text{in}} = 360$ (C.3) Dataset 6, $T_{\text{in}} = 480$



(C.4) Dataset 7, $T_{\text{in}} = 240$ (C.5) Dataset 7, $T_{\text{in}} = 360$

Figure C.7: **Chosen sparsity under annual tuning: managed-portfolio universes.** Boxplots of the annually selected cardinalities $\widehat{k}_j^{\text{LOO}}(T)$ and $\widehat{k}_j^{\text{UE}}(T)$ across tuning years. The top row reports the size–value–industry universe, and the bottom row reports the international managed-portfolio universe.



(C.1) Dataset 3, $T_{\text{in}} = 240$ (C.2) Dataset 3, $T_{\text{in}} = 360$ (C.3) Dataset 3, $T_{\text{in}} = 480$

(C.4) Dataset 4, $T_{\text{in}} = 240$ (C.5) Dataset 4, $T_{\text{in}} = 360$ (C.6) Dataset 4, $T_{\text{in}} = 480$

Figure C.8: **Chosen sparsity under annual tuning: anomaly portfolios and individual stocks.** Boxplots of the annually selected cardinalities $\widehat{k}_j^{\text{LOO}}(T)$ and $\widehat{k}_j^{\text{UE}}(T)$ across tuning years. The top row reports the anomaly long–short universe, and the bottom row reports the individual-stock universe.

C.3.6 Feasible Tuning Rules: Robustness Checks

We complement Section 5.3.4 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 6) and the international universe (Dataset 7), matching the supplementary datasets used in the baseline feasible-tuning table.

Table C.3 shows that the baseline tuning conclusions are broadly robust to these two design changes. 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 C.9 and C.10 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.

Table C.3: **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 $\widehat{\theta}^{\text{oos}}_{\max}(T)$ (max over k), and the realized out-of-sample Sharpe ratios under annual tuning of k via leave-one-out (LOO) and the bias-adjusted Sharpe estimator (UE). All Sharpe ratios are annualized. Panels vary whether factor-mimicking portfolios are included and whether realized weights are normalized to satisfy $\|w\|_1 = 1$.

(C.1) Dataset 6: $\|w\|_1 = 1$, no factors

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

(C.2) Dataset 6: factors, no normalization

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

(C.3) Dataset 7: factors, no normalization

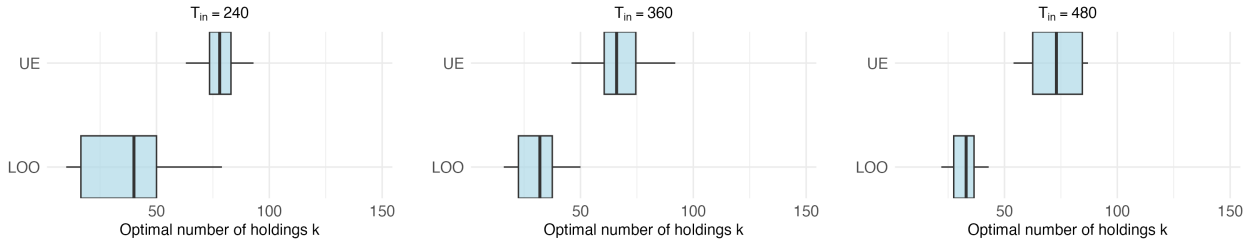
T_{in}	$\widehat{\theta}^{\text{oos}}_{\max}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	2.8618	1.4821	1.6794
360	3.6421	3.4332	3.2483

(C.4) Dataset 7: $\|w\|_1 = 1$, no factors

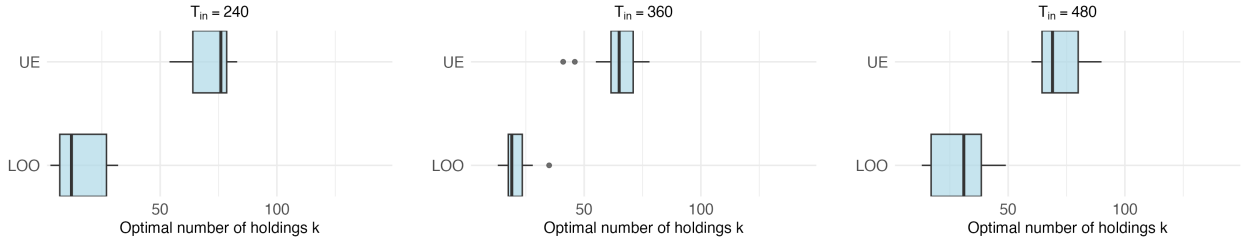
T_{in}	$\widehat{\theta}^{\text{oos}}_{\max}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{LOO}}(T)$	$\widehat{\theta}^{\text{oos}}_{\text{UE}}(T)$
240	3.0642	2.6346	1.1360
360	2.5808	2.2653	2.2284

D Complementary Theoretical Analysis

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



(C.1) Factors, $T_{\text{in}} = 240$ (C.2) Factors, $T_{\text{in}} = 360$ (C.3) Factors, $T_{\text{in}} = 480$



(C.4) $\|w\|_1 = 1$, $T_{\text{in}} = 240$ (C.5) $\|w\|_1 = 1$, $T_{\text{in}} = 360$ (C.6) $\|w\|_1 = 1$, $T_{\text{in}} = 480$

Figure C.9: **Chosen sparsity under annual tuning (robustness): Dataset 6.** Boxplots of annually selected cardinalities under factor augmentation (top row) and under ℓ_1 gross-exposure normalization (bottom row).

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 $\sigma > 0$ so that

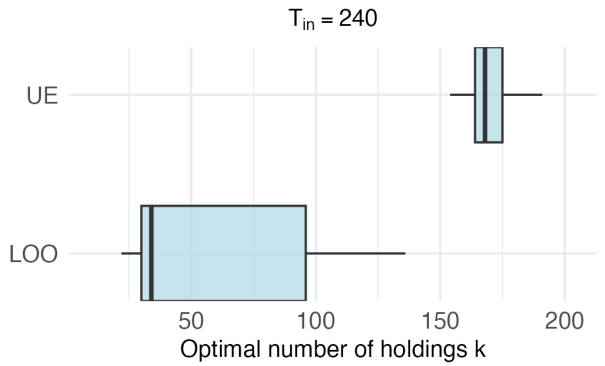
$$\mathbb{E} \exp(v^\top (r - \mathbb{E}r)) \leq \exp(\|v\|_2^2 \sigma^2/2), \qquad \forall v \in \mathbb{R}^N.$$

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

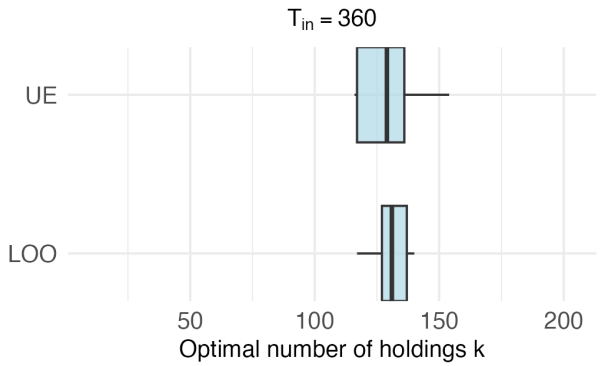
$$\alpha(\ell) := \sup_{t\in\mathbb{Z}} \sup_{A\in\mathcal{F}_{-\infty}^t,\ B\in\mathcal{F}_{t+\ell}^\infty} \big|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|, \qquad \ell \geq 0.$$

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.

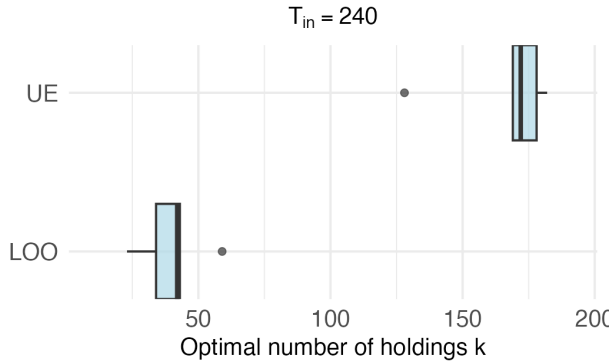
Assumption D.1 (sub-Gaussianity and weak dependence)**.** *The return sequence* $\{r_t\}_{t\in\mathbb{Z}}$ *is*



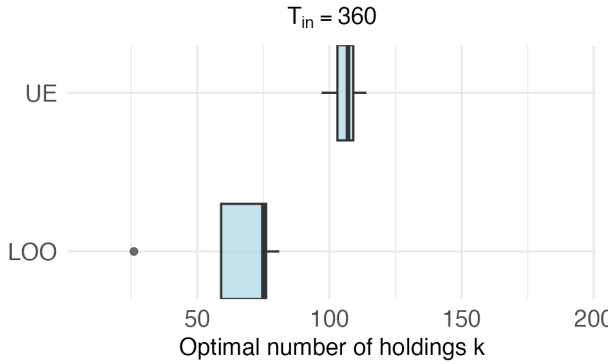
(C.1) Factors, $T_{\text{in}} = 240$



(C.2) Factors, $T_{\text{in}} = 360$



(C.3) $\|w\|_1 = 1$, $T_{\text{in}} = 240$



(C.4) $\|w\|_1 = 1$, $T_{\text{in}} = 360$

Figure C.10: **Chosen sparsity under annual tuning (robustness): Dataset 7.** Boxplots of annually selected cardinalities under factor augmentation (top row) and under ℓ_1 gross-exposure normalization (bottom row).

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*

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

Under Assumption D.1, the serial dependence is sufficiently weak that the effective sample size remains proportional to T.

D.1 Results Under Weak Dependence

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

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

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

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 D.2 (Bound on allocation risk under weak dependence)**.** *Let Assumption D.1 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 D.1 (Estimation risk of sparse portfolios under weak dependence)**.** *Let Assumption D.1 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 D.1 has the same $\sqrt{k \log N / T}$ scaling as the i.i.d. benchmark, Theorem 1. Weak 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 dimension N.

D.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 weak dependence by repeating the i.i.d. arguments with this replacement. Finally, we conclude with the proof of Theorem D.1.

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

$$|\mathrm{Cov}(U, V)| \le C\, \alpha(\ell)^{1/2}\, \|U\|_4\, \|V\|_4, \qquad \ell \ge 0, \tag{D.1}$$

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

Lemma D.1 (Moment concentration on a fixed support under mixing)**.** *Under Assumption D.1, 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 $1 - 4e^{-x}$,*

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

$$\left\|\Sigma_A^{-1/2}\widehat{\Sigma}_A\Sigma_A^{-1/2} - I_m\right\|_{\mathrm{op}} \leq C_{\mathrm{mix}}\left(\sqrt{\frac{m+x}{T}} + \frac{m+x}{T}\right). \tag{D.3}$$

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 (D.1),

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

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

$$\|C_\ell\|_{\mathrm{op}} := \left\|\mathbb{E}[x_t x_{t+\ell}^\top]\right\|_{\mathrm{op}} \leq C\,\alpha(\ell)^{1/2}. \tag{D.4}$$

Define the finite dependence constant

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

where finiteness follows from geometric mixing in Assumption D.1.

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 $x \ge 0$,

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

for some constant $c > 0$. Absorbing numerical factors into the constant gives (D.2).

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 (D.4), R_u is Toeplitz with 1 on the diagonal and $|(R_u)_{st}| = |\mathbb{E}[z_s z_t]| \le \|C_{|s-t|}\|_{\mathrm{op}}$, which implies

$$\|R_u\|_{\mathrm{op}} \le 1 + 2\sum_{\ell=1}^{T-1} \|C_\ell\|_{\mathrm{op}} \le M_\alpha. \tag{D.5}$$

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:

$$\mathbb{P}\left(\left|u^\top (S - I_m)u\right| \geq 2M_\alpha \sqrt{\frac{x}{T}} + 2M_\alpha \frac{x}{T}\right) \leq 2e^{-cx}, \tag{D.6}$$

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 (D.6) 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 $1 - 2e^{-x}$,

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

for a constant $C > 0$. □

Lemma D.1 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 D.1. Combining them with the decomposition (10) proves Theorem D.1.

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

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

Fix A with $m := |A| \leq k$. Let x_t, $\bar{x}$ and S be defined as in the proof of Lemma D.1. 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 D.1 implies that with probability at least $1 - 4e^{-x}$,

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

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

$$\left\|S^{-1/2} - I_m\right\|_{\mathrm{op}} \le C\left\|S - I_m\right\|_{\mathrm{op}}.$$

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

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

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

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

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

and the claim follows. □

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

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

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 D.1 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 $\theta^* < \infty$, take T large enough that the covariance-deviation bound is at most $1/2$. Then $\|G_A - I_m\|_{\mathrm{op}} \le 1/2$ 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 D.1 and D.2 in hand, the proof of Theorem D.1 is identical to the i.i.d. argument.

Proof of Theorem D.1. 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 D.1 and $\theta^* < \infty$, Proposition D.1 gives $\theta_k^* - \theta_{\widehat{A}_k} = O_p(\sqrt{k \log N/T})$, and Proposition D.2 gives $\theta_{\widehat{A}_k} - \theta(\widehat{w}_k) = O_p(\sqrt{k \log N/T})$. Summing the two bounds yields the stated rate. □

E Numerical Algorithms

This appendix collects the computational details for implementing sparse portfolio rules. The cardinality-constrained rule in (2) is useful for theory because it makes portfolio complexity explicit through k. In practice, however, solving the exact cardinality problem quickly becomes infeasible once N is moderately large. We therefore use the ℓ_1-regularized rule in (3) as the default implementation. The ℓ_1 relaxation scales reliably to large universes, delivers an entire regularization path, and provides a direct way to index portfolios by their induced support size. When problem sizes permit, we also solve a mixed-integer quadratic program (MIQP) that enforces the cardinality constraint directly and serves as a benchmark.

E.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 grows essentially exponentially in N when k is a non-negligible fraction of the universe size.

A concrete illustration conveys the scale. 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 not a viable large-scale implementation strategy.

Exact methods are nevertheless useful at moderate scale. A leading formulation is a mixed-integer quadratic program, which enforces sparsity by introducing binary inclusion variables and linking constraints that set $w_i = 0$ whenever asset i is excluded. Modern MIQP solvers can deliver high-quality solutions for moderate N, 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 also Bertsimas and Cory-Wright (2022).

E.2 ℓ_1 Regularization as a Convex Relaxation

We implement sparse portfolios through the convex relaxation (3), which replaces the discrete constraint $\|w\|_0 \le k$ with a continuous penalty on gross exposure $\|w\|_1$. The appeal is both computational and economic. Computationally, the problem is a concave quadratic program with an ℓ_1 penalty. Economically, $\|w\|_1$ is leverage in a long–short book, so penalizing it limits the optimizer's ability to transform 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 $\widehat{U}_T(\cdot) - \lambda\|\cdot\|_1$ is concave, $\widehat{w}_{\ell_1,\lambda}$ is characterized by the KKT condition

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

where $\partial\|w\|_1$ denotes the subdifferential of the ℓ_1 norm:

$$\partial\|w\|_1 := \Big\{u \in \mathbb{R}^N : u_i = \operatorname{sign}(w_i) \text{ if } w_i \neq 0, \quad u_i \in [-1,1] \text{ if } w_i = 0\Big\}.$$

Equivalently, for each asset i,

$$\widehat{w}_{\ell_1,\lambda,i} \neq 0 \quad \Longrightarrow \quad \left(\widehat{\mu} - \gamma\widehat{\Sigma}\widehat{w}_{\ell_1,\lambda}\right)_i = \lambda \operatorname{sign}(\widehat{w}_{\ell_1,\lambda,i}),$$
$$\widehat{w}_{\ell_1,\lambda,i} = 0 \quad \Longrightarrow \quad \left(\widehat{\mu} - \gamma\widehat{\Sigma}\widehat{w}_{\ell_1,\lambda}\right)_i \in [-\lambda, \lambda].$$

Thus, an asset is excluded when its marginal contribution to the sample 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.

The ℓ_1 norm is also a natural tight convex proxy for cardinality once scale is controlled. Under natural position limits, for example $\|w\|_\infty \leq 1$, the convex hull of k–sparse portfolios coincides with an ℓ_1 ball:[11]

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

Therefore, $\|w\|_1$ provides the tightest convex relaxation of sparsity on this bounded domain; see Argyriou et al. (2012).

E.3 From λ to an Implementable Sparsity Level

The cardinality rule is indexed directly by k, whereas the ℓ_1 rule is indexed by the penalty λ. To compare the two rules empirically, we use the induced support size

$$\widehat{k}(\lambda) := \|\widehat{w}_{\ell_1,\lambda}\|_0.$$

A target sparsity level k is implemented by choosing a value of λ along the lasso path whose solution has induced support size at or near k. Along the path, $\widehat{k}(\lambda)$ is a step function of

[11] We use the notation $\operatorname{conv}A$ for the convex hull of a set $A \subset \mathbb{R}^N$.

λ: increasing λ typically shrinks the support, but support sizes may be skipped and ties can occur when several assets enter or exit together.

Because the Sharpe ratio is scale invariant, the solution of (3) can be interpreted as a portfolio direction whenever it is nonzero. After estimation, this direction can be rescaled to the implementation convention of interest, such as a target volatility or target gross exposure. A unit-budget normalization $1^\top w = 1$ is used only when the estimated direction has nonzero budget exposure.

The theory suggests a natural order for the penalty level. Taking λ on the order of $\sqrt{\log N/T}$ is sufficient to dominate coordinatewise sampling noise in high dimensions and prevents the optimizer from chasing spurious sample Sharpe opportunities. In practice, we evaluate a grid of λ values spanning from a very small penalty, which yields an essentially dense solution, to a large penalty, which yields the zero portfolio, and then select a point on the path by targeting the desired $\widehat{k}(\lambda)$.

E.4 Numerical Algorithms: Lasso and MIQP Approximations

E.4.1 Lasso relaxation

Problem (3) can be written as a standard lasso problem; see, e.g., Li (2015). Let

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

When $\widehat{\Sigma}$ is singular or ill-conditioned, for example when $N \geq T$, we work with a stabilized covariance matrix

$$\widehat{\Sigma}_\varepsilon := \widehat{\Sigma} + \varepsilon I_N, \qquad \varepsilon > 0,$$

and set $\widehat{M} := \gamma \widehat{\Sigma}_\varepsilon$. This is equivalent to adding a small ℓ_2 stabilization and ensures that $\widehat{M}$ is positive definite.

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

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

Define

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

where $C^{-\top}\widehat{\mu}$ is computed by 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

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

We compute the lasso path using the `lars` package in R (Efron et al., 2004), using `intercept = FALSE` and `normalize = FALSE`. We pass (X_T, y_T) without 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 target cardinality k, we scan the knots and compute

$$\widehat{k}(\lambda) = \|\widehat{w}_{\ell_1,\lambda}\|_0,$$

where coefficients satisfying $|w_i| \le$ `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, corresponding to the largest penalty attaining that support size. If no knot has exactly k active positions, we select the knot with the largest $\widehat{k}(\lambda) \le k$, 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.4.2 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 $(\widehat{\mu}, \widehat{\Sigma})$, 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. A budget-normalized MIQP formulation is

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

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. For larger instances, the MIQP solution should therefore be interpreted as a high-quality approximation rather than a guaranteed global optimum.

To reduce the possibility that the solution is driven by overly tight box constraints, we apply a 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 move the corresponding bound outward: lower bounds are made more negative and upper bounds are made larger, using a fixed expansion factor and the previous solution as a warm start. This expansion is iterated a small number of times or until no active bounds remain.

F Proofs

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

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

Proof of Lemma F.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 $\widehat{A}_k = \mathrm{supp}(\widehat{w}_k)$ 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 F.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

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

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 variables $U \sim \chi_k^2(T\theta_A^2)$ and $V \sim \chi_{T-k}^2$. We can write

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

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

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

and

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

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 appli-

cation of union bounds to all A such that $|A| \le k$ yields

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

with probability at least

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

□

Proof of Proposition 2. Let $\widehat{A}_k := \mathrm{supp}(\widehat{w}_k)$. By optimality of $\widehat{w}_k$ in (2), conditional on the support $\widehat{A}_k$ the weights maximize $\widehat{U}_T$ over $\{w : \mathrm{supp}(w) \subset \widehat{A}_k\}$, 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_t's 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 $1 - N^{-(1+\alpha)k}$,

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

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

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

An application of union bound then yields that

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

with probability at least

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

□

Proof of Theorem 1. Let $\widehat{A}_k := \mathrm{supp}(\widehat{w}_k)$. By definition of θ_k^* and by $\theta_{\widehat{A}_k} = \max_{\mathrm{supp}(w) \subset \widehat{A}_k} \theta(w)$,

we have the decomposition

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

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

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

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 (F.1) implies, uniformly over A,

$$\theta_A^{(0)} - \Delta_N \le \theta_A \le \theta_A^{(0)} + \Delta_N. \tag{F.2}$$

In particular,

$$\theta^* \le \theta^{*,(0)} + \Delta_N, \qquad \theta_k^* \ge \theta_k^{*,(0)} - \Delta_N, \qquad \Longrightarrow \qquad \theta^* - \theta_k^* \le \big(\theta^{*,(0)} - \theta_k^{*,(0)}\big) + 2\Delta_N. \tag{F.3}$$

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

$$\Delta_N = O(1/N). \tag{F.4}$$

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

$$0 \le q - (\theta_A^{(0)})^2 \le \mu_f^\top \Sigma_f^{-1} H_A^{-1} \Sigma_f^{-1} \mu_f \le \frac{\|\Sigma_f^{-1}\mu_f\|_2^2}{\lambda_{\min}(H_A)}. \tag{F.5}$$

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,

$$\lambda_{\min}(H) \geq \tfrac{1}{2}\, \lambda_{\min}(\Sigma_B)\, N^{\zeta}. \tag{F.6}$$

Combining (F.5) with (F.6) yields

$$0 \leq q - (\theta^{*,(0)})^2 = O(N^{-\zeta}). \tag{F.7}$$

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 $k \geq L$,

$$\lambda_{\min}\big(B_{\widetilde{A}}^\top B_{\widetilde{A}}\big) \geq \frac{k}{LN}\, \lambda_{\min}(B^\top B). \tag{F.8}$$

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 (F.8) that

$$\lambda_{\min}(H_{\widetilde{A}}) \geq \frac{1}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}\big(B_{\widetilde{A}}^\top B_{\widetilde{A}}\big) \geq \frac{k}{LN}\, \frac{1}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}(B^\top B). \tag{F.9}$$

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 (F.9) gives

$$\lambda_{\min}(H_{\widetilde{A}}) \geq \frac{k}{LN}\, \frac{\lambda_{\min}(\Sigma_0)}{\lambda_{\max}(\Sigma_0)}\, \lambda_{\min}(H). \tag{F.10}$$

Combining (F.10) with (F.6) 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 (F.5) gives

$$0 \le q - (\theta^{(0)}_{\widetilde{A}})^2 = O\left(\frac{N^{1-\zeta}}{k}\right). \tag{F.11}$$

Since $\theta^{*,(0)}_k = \max_{|A|\le k} \theta^{(0)}_A \ge \theta^{(0)}_{\widetilde{A}}$, we have $q - (\theta^{*,(0)}_k)^2 \le q - (\theta^{(0)}_{\widetilde{A}})^2$, hence

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

Therefore,

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

If $q = 0$, then $\mu^{(0)} = 0$ and $\theta^{*,(0)} = \theta^{*,(0)}_k = 0$, so $\theta^{*,(0)} - \theta^{*,(0)}_k = 0$. Otherwise $q > 0$, and (F.7) implies that $\theta^{*,(0)}$ is bounded away from 0 for all sufficiently large N, hence

$$\theta^{*,(0)} - \theta^{*,(0)}_k = \frac{(\theta^{*,(0)})^2 - (\theta^{*,(0)}_k)^2}{\theta^{*,(0)} + \theta^{*,(0)}_k} = O\left(\frac{N^{1-\zeta}}{k}\right). \tag{F.12}$$

Step 4 (combine the bounds). Substituting (F.12) and (F.4) into (F.3) 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

$$\Sigma^{-1}\mu = \left(I - \Sigma_0^{-1} B K^{-1} B^\top\right)\Sigma_0^{-1}\alpha \; + \; \Sigma_0^{-1} B K^{-1} \Sigma_f^{-1} \mu_f. \tag{F.13}$$

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 $\|B^\top\|_{1\to 1} = O(1)$.[12]

For the first term in (F.13), 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,

$$\|(I - H)\Sigma_0^{-1}\alpha\|_1 = O\big(N^{(1-\zeta)/2}\big). \tag{F.14}$$

For the second term in (F.13), use $\|x\|_1 \le \sqrt{N}\|x\|_2$ and obtain

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

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

since $\|\Sigma_f^{-1}\mu_f\|_2 = O(1)$ in fixed dimension. Combining with (F.14) gives

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

hence

$$\|w_*\|_1 = \frac{1}{\gamma}\|\Sigma^{-1}\mu\|_1 = O\big(N^{(1-\zeta)/2}\big)\,. \tag{F.15}$$

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

$$\frac{\gamma}{2}\Delta^\top\widehat{\Sigma}\Delta \leq \Delta^\top\big(\widehat{\mu} - \gamma\widehat{\Sigma}w_*\big) + \lambda(\|w_*\|_1 - \|\widehat{w}\|_1). \tag{F.16}$$

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

$$\|S_T\|_\infty = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{F.17}$$

Let $\mathcal{E}_T := \{\|S_T\|_\infty \leq \lambda/2\}$. By (F.17), 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 (F.16) 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

$$\|\widehat{w}\|_1 \leq 3\|w_*\|_1, \qquad \|\Delta\|_1 \leq 4\|w_*\|_1, \qquad \Delta^\top\widehat{\Sigma}\Delta \leq \frac{3\lambda}{\gamma}\|w_*\|_1. \tag{F.18}$$

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

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

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

$$\|\widehat{\Sigma} - \Sigma\|_{\max} = O_p\left(\sqrt{\frac{\log N}{T}}\right). \tag{F.19}$$

On $\mathcal{E}_T$, combining (F.18)–(F.19) and using $\lambda \ge C_0\sqrt{\log N/T}$ 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 (F.15),

$$\Delta^\top \Sigma \Delta = O_p\big(N^{1-\zeta}\lambda\big). \tag{F.20}$$

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 (F.20) 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 $\theta(\widehat{w}_{\ell_1,\lambda}) \xrightarrow{p} \theta^*$. □

Appendix References

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

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

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

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

Kenneth R. French. Data library: Fama/french factors and portfolios. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html, 2026. Accessed: May 2026.

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.

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

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.

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

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