When Does Hierarchical Risk Parity Beat Markowitz? A Controlled Comparison Under Known Covariance

Eugen Soloviov*

Abstract

Hierarchical Risk Parity (HRP) replaces covariance-matrix inversion with hierarchical clustering and recursive inverse-variance allocation, and is widely promoted as the practical cure for the instability of Markowitz optimization. We give this claim a controlled test in which the *true* covariance matrix is known by construction, so the realized risk of any weight vector is computable exactly as $w^\top \Sigma w$ and the unconstrained minimum-variance oracle is an exact floor. Across 4,800 experiments (4,000 Gaussian, 800 multivariate Student-t) spanning four true-correlation families—one-factor, nested hierarchical, unstructured (normalized Wishart), and equicorrelation—with $N \in \{10, \ldots, 100\}$ assets and estimation windows of $T/N \in [0.5, 10]$, the verdict splits cleanly. (i) HRP's *insurance value against naive Markowitz is real*: at $T = N$, where the sample covariance is singular, sample minimum variance realizes a median 16.8× the oracle risk (above 100× in 9% of cases), while HRP realizes 2.3×, beats sample minimum variance in 87% of data-starved experiments ($T/N \leq 1$; every single one in the unstructured family at $T/N = 0.5$), and never suffers an inversion catastrophe. (ii) *Its advantage over shrinkage is mostly not*: Ledoit–Wolf minimum variance—equally inversion-proof, one line of scikit-learn—realizes 1.8× at $T = N$ and beats HRP at every T/N in the factor, hierarchical, and equicorrelation families; HRP's win rate against it is 0.34 at $T/N \leq 1$ and falls to 0.05 at $T/N \geq 5$, with HRP ahead only under unstructured correlation at small T (win rate ≈0.55). The 1.8-versus-2.3 comparison pits shorts-allowed LW against long-only HRP; on the like-for-like long-only comparison at $T = N$ the pooled medians are a near-tie (clipped LW 2.32× vs. HRP 2.29×), though LW remains ahead in paired comparisons. (iii) HRP is *not a consistent minimum-variance estimator*: by $T/N = 10$ sample and shrunk minimum variance converge to ≈1.10× the oracle while HRP plateaus at 1.40–3.20× depending on family. (iv) Much of HRP's behavior is the inverse-variance portfolio inside it: in the data-starved band its win rate against diagonal inverse variance exceeds 0.5 only when clusters genuinely differ in tightness (rising 0.35 → 0.56 across dispersion terciles of our hierarchical worlds), in the data-rich band it edges IV more broadly (win rates 0.46–0.63 at $T/N \geq 5$), and the much-debated linkage choice is immaterial (the three linkage medians agree to within 0.02–0.06). (v) Estimated *means*, not estimated covariances, remain the real catastrophe: tangency portfolios built from estimated means capture a median 10% of the oracle Sharpe ratio at $T/N \leq 1$ and are negative 28% of the time. The Student-t batch leaves every qualitative ranking unchanged. Practical reading: HRP is real insurance against inverting a noisy covariance matrix when the correlation structure is unknown and data are scarce—but if you can shrink, shrinking is better almost everywhere.

1 Introduction

Markowitz mean–variance optimization [20] is provably optimal with known inputs and notoriously fragile with estimated ones: the optimizer amplifies exactly the estimation errors that matter most

*Independent Researcher. ORCID: 0009-0006-3148-111X. Correspondence: suenot@gmail.com. Code to reproduce every number and figure: `https://github.com/suenot/hrp-validation`.

[1, 3, 21]. The most influential recent response in practitioner circles is Hierarchical Risk Parity [HRP; 18], which never inverts a matrix: it clusters assets by correlation distance, reorders the covariance into quasi-diagonal form, and allocates top-down by recursive inverse-variance bisection. HRP is now a default in several open-source portfolio libraries and is routinely recommended as the robust replacement for Markowitz.[1]

The evidence behind the recommendation is thinner than its adoption. The original paper demonstrates HRP on one numerical example and one Monte Carlo design, against unshrunk benchmarks; most follow-ups are empirical backtests in which the true covariance is unknown, so no one can say how far any allocator sits from the attainable optimum, or how much of its loss is estimation error versus specification error. Meanwhile the estimation-error literature has its own well-validated remedy—shrinkage [14–16]—which is almost never the benchmark HRP is compared against.

This paper supplies the missing controlled comparison. We generate the true covariance Σ from four structural families chosen to span the cases where HRP's clustering prior is right (nested hierarchical blocks), partially right (one common factor), and wrong (unstructured dense correlation; equicorrelation), hand every allocator the same finite sample of T observations, and score the resulting weights *on the truth*: realized risk $w^\top \Sigma w$ against the exact unconstrained minimum-variance oracle. There is no backtest noise and no data snooping; the only randomness is the sampling error every estimator faces. We sweep the ratio T/N from 0.5 (deeply singular) to 10 (data-rich), because the break-even results of DeMiguel et al. [7] identify exactly this ratio as the axis on which optimization beats naive rules.

Our findings are deliberately mixed, and more useful for it. HRP's headline story—"Markowitz explodes, HRP does not"—is confirmed, and quantified: at $T = N$ the sample minimum-variance portfolio realizes a median 16.8× the oracle risk, HRP 2.3×. But the comparison that practitioner writing rarely runs—HRP against Ledoit–Wolf shrinkage, which is equally inversion-proof and one line of scikit-learn—reverses the story almost everywhere: shrinkage wins at every T/N in three of our four families, and HRP's edge survives only where there is no structure for its dendrogram to find and very little data to estimate with. Linkage choice, the subject of a sizable applied debate, moves the median result by 0.02–0.06 depending on family; the dispersion of cluster tightness, which nobody debates, moves HRP's data-starved win rate against its own inverse-variance core from 0.35 to 0.56. And the classical disaster remains where DeMiguel et al. [7] located it: portfolios that estimate *means* capture a median 10% of the attainable Sharpe ratio in the data-starved regime and are negative 28% of the time.

Contributions.

1. A reproducible known-covariance testbed for portfolio allocators: four true-correlation families with every generator parameter drawn from one disclosed sampling table and recorded per experiment, Gaussian and multivariate Student-t returns, and exact evaluation of realized risk $w^\top \Sigma w$ against the exact minimum-variance oracle (Sections 3–5).
2. A quantification of HRP's insurance value against naive Markowitz in the data-starved regime, and of its failure to converge in the data-rich regime: HRP's realized-risk plateau at 1.40–3.20× the oracle isolates its specification gap, the loss it would retain even with the true covariance (Sections 6.1–6.2).

[1] Including in the author's own earlier practitioner writing, where the exact pipeline tested here—single-linkage clustering on correlation distance, quasi-diagonalization, recursive inverse-variance bisection—was recommended over Markowitz inversion for allocating capital across market-making strategies. This paper is therefore a controlled self-audit of advice the author has himself circulated, not an attack on a third party.

3. The comparison that matters and is rarely run: HRP versus Ledoit–Wolf shrinkage, stratified by structure family and T/N, with win rates and paired log-ratios. Shrinkage wins almost everywhere; we delineate the exception (unstructured correlation, $T/N \leq 1$) and show that even there plain inverse variance is better still (Section 6.3).

4. Evidence on *which* of HRP's ingredients matter: cluster-tightness dispersion does (data-starved win rate vs. inverse variance $0.35 \to 0.56$ across terciles), linkage does not (the three linkage medians agree to within 0.02–0.06, and the distance-matrix convention behind the linkage moves them by at most 0.02), and HRP's turnover and concentration profile is closer to inverse variance than to any optimizer (Sections 6.4–6.5).

5. A replication, inside the same harness, of the classical result that estimated means are the dominant failure mode: noisy-μ tangency directions capture a median 10% of the oracle Sharpe at $T/N \leq 1$, negative in 28% of experiments—below every μ-free rule in the same regime, from sample minimum variance (18% pooled) to the diffuse rules of Table 4 (26–70%) (Section 6.6).

2 Related work

Estimation error in mean–variance optimization. Mean–variance portfolio selection [20] is optimal only when its inputs are known; in practice they must be estimated, and the optimizer interacts badly with estimation noise. Jobson and Korkie [10] showed in early Monte Carlo experiments that sample-based efficient portfolios can be dramatically inferior to the population-optimal ones, and Jobson and Korkie [11] argued that naive alternatives are often preferable in realistic sample sizes. Michaud [21] popularized the diagnosis of mean–variance optimizers as "estimation-error maximizers" that load on the most extreme—and most mis-estimated—inputs. Subsequent work quantified the mechanism: optimal weights are extremely sensitive to small perturbations of expected returns [1], errors in means dominate but covariance errors matter increasingly at high risk tolerance [3], and even with optimal Bayesian use of sample information, parameter uncertainty erodes most of the theoretical gains from optimization [12]. The starkest statement of the problem is due to DeMiguel et al. [7]: across fourteen optimizing models and seven empirical datasets, none consistently outperforms the equally weighted $1/N$ portfolio out of sample, because the estimation window needed for sample-based mean–variance to win is far longer than anything available in practice. Their analysis ties the break-even point explicitly to the ratio of sample size to universe size—the same T/N axis we stratify on.

Covariance-focused remedies. One response is to abandon expected-return estimation and target the minimum-variance portfolio, which depends only on the covariance matrix and has performed well empirically [4, 5]. But the covariance matrix itself is noisy when N is large relative to T. Random matrix theory shows that the bulk of sample correlation eigenvalues from financial data is indistinguishable from noise [13, 25], motivating a large literature on eigenvalue cleaning [2]. Shrinkage estimators address the same ill-conditioning by pulling the sample covariance toward a structured target—a single-factor model [14], a scaled identity with asymptotic optimality guarantees [16], or the constant-correlation target of the widely used "Honey, I shrunk" estimator [15]. Jagannathan and Ma [9] showed that no-short-sale constraints act as an implicit shrinkage of extreme covariances, which explains why constrained sample-based minimum-variance portfolios perform respectably. This literature establishes that *how* the covariance is estimated matters; it says less about how the optimizer's sensitivity to the remaining error depends on the structure of

the true covariance.

Hierarchical portfolio construction. A parallel line, rooted in the observation that financial correlation matrices carry a hierarchical organization [19], sidesteps matrix inversion altogether. Hierarchical Risk Parity [18] clusters assets with single-linkage agglomerative clustering on a correlation distance, quasi-diagonalizes the covariance, and allocates by recursive inverse-variance bisection. Raffinot [26] proposed hierarchical clustering-based allocation with alternative linkages and equal weighting across clusters, later extended to hierarchical equal risk contribution in a working paper [27]. Further extensions incorporate tail-dependence-based distances in multi-asset multi-factor settings [17], constraints and a continuum between cluster-level and asset-level allocation [24, a working paper], and Monte-Carlo-based modifications for managed futures [22]. Jaeger et al. [8] use explainable-machine-learning tools to attribute when and why HRP-type strategies beat equal weight in block-bootstrapped multi-asset data. Most recently, Cotton [6] showed that HRP and minimum variance are endpoints of a single family of Schur-complement allocations, sharpening the question of where on that spectrum one should sit—a question whose answer must depend on estimation error.

The gap. The original evidence for HRP consists of one numerical example and one Monte Carlo experiment with a particular block-diagonal-plus-shocks data-generating process, evaluated against unshrunk inverse-variance benchmarks [18]. Follow-up studies are predominantly empirical backtests [8, 17, 26], where the true covariance is unknown, so realized out-of-sample variance confounds estimation error with model misspecification and nonstationarity, and no oracle is available to measure either. Meanwhile, the $1/N$ literature [7] and the shrinkage literature [15, 16] each demonstrate their preferred remedy on their own testbeds, with no common ground truth. What is missing is a controlled comparison in which the true covariance is known by construction and varied systematically across structural families—strict factor models, genuinely hierarchical block structures, unstructured dense matrices, and equicorrelation—while the sample length T and universe size N are swept over the T/N ratios that drive the DeMiguel-style break-even results. Against the known-covariance oracle, the excess realized risk of each portfolio rule decomposes cleanly into a specification term (what the rule would lose even with perfect inputs, relevant for HRP and $1/N$, which are not oracle-optimal) and an estimation term (what it loses to sampling noise). To our knowledge no published study reports this decomposition for HRP against sample and Ledoit–Wolf-shrunk minimum variance, inverse variance, and $1/N$ across structures and T/N regimes, nor does any systematically examine HRP's sensitivity to the linkage rule, despite the well-documented chaining behavior of single linkage [23] and Raffinot's evidence that linkage choice materially changes hierarchical allocations [26]. This paper fills that gap.

3 Allocators

Every allocator maps an estimation window $R \in \mathbb{R}^{T\times N}$ (or the sample covariance built from it) to full-investment weights, $\sum_i w_i = 1$; all are scored on the true Σ. The risk-only track contains ten methods that never see a mean estimate.

Equal weight (EW). $w_i = 1/N$, the DeMiguel et al. [7] benchmark.

Inverse variance (IV). $w_i \propto 1/\widehat{\Sigma}_{ii}$, the diagonal-only allocator. It is the degenerate HRP that ignores all correlation information, and therefore the cleanest control for how much HRP's

clustering machinery adds.

Sample minimum variance (MV). The closed-form unconstrained minimizer of $w^\top \widehat{\Sigma} w$ subject to $\mathbf{1}^\top w = 1$,

$$w_{\mathrm{MV}} \;=\; \frac{\widehat{\Sigma}^{+}\mathbf{1}}{\mathbf{1}^\top \widehat{\Sigma}^{+}\mathbf{1}}, \tag{1}$$

shorts allowed, with $\widehat{\Sigma}$ the sample covariance (DDOF=1). When $T \le N$ the sample covariance is singular; we use the Moore–Penrose pseudoinverse (the minimum-norm solution) and flag the record as rank-deficient rather than failing, since this is exactly the regime in which the HRP debate lives. The pseudoinverse truncates at NumPy's default relative cutoff (`rcond` $= 10^{-15}$); the 16.8× headline of Section 6.1 is insensitive to this choice: replaying every $T/N = 1$ experiment with cutoffs $10^{-15}/10^{-10}/10^{-6}/10^{-3}$ gives pooled medians 16.82/16.82/15.61/3.98, so only an absurdly aggressive cutoff—itself a crude regularizer—changes the picture.

Long-only clip (MV-LO). Negative weights clipped to zero, then renormalized. This is deliberately the crude practitioner heuristic, not the long-only quadratic program: it is what naive implementations do, and the implicit-shrinkage result of Jagannathan and Ma [9] suggests even this crude projection should help.

Ledoit–Wolf minimum variance (LW, LW-LO). Equation (1) applied to the shrinkage estimator of Ledoit and Wolf [16],

$$\widehat{\Sigma}_{\mathrm{LW}} \;=\; (1-\delta)\,\widehat{\Sigma}_{\mathrm{MLE}} \;+\; \delta\,\frac{\operatorname{tr}\widehat{\Sigma}_{\mathrm{MLE}}}{N}\,I, \tag{2}$$

with the analytically optimal intensity δ as computed by scikit-learn. The shrunk matrix is always well-conditioned, so LW is exactly as inversion-proof as HRP. LW-LO applies the same clip as MV-LO. In our experiments the fitted intensity behaves as theory predicts, with a median δ of 0.42 at $T/N = 0.5$ declining to 0.04 at $T/N = 10$.

Hierarchical Risk Parity (HRP). Following López de Prado [18], in four steps. (1) From the sample correlation $\hat{\rho}$, form the distance $d_{ij} = \sqrt{(1-\hat{\rho}_{ij})/2} \in [0,1]$. (2) Run agglomerative clustering; the original prescribes *single* linkage. Two conventions for the clustered distance coexist: our main implementation links directly on the pairwise distances d (a convention common among library implementations), whereas the original code listing passes the *square matrix* of d to the clustering routine, which therefore links on the Euclidean distance-of-distances $\tilde{d}_{ij} = \|d_{\cdot i} - d_{\cdot j}\|_2$. We run both; the stored $\tilde{d}$ variant is the faithfulness check of Section 6.5, where the two conventions prove interchangeable in aggregate. (3) Quasi-diagonalize: reorder assets by the dendrogram leaf order, which places mutually correlated assets adjacently. (4) Recursive bisection: split the ordered list into contiguous halves L and R (the paper bisects the ordered list, not the dendrogram), compute each half's variance under its inverse-variance-weighted sub-portfolio, $\tilde{V}_C = w_C^\top \widehat{\Sigma}_C w_C$ with $w_C \propto \operatorname{diag}(\widehat{\Sigma}_C)^{-1}$, and allocate the fraction $\alpha = 1 - \tilde{V}_L/(\tilde{V}_L + \tilde{V}_R)$ to the left half; recurse until singletons. All weights are positive and sum to one by construction; no matrix is ever inverted. On the condensed-d convention we run three linkage variants, SINGLE (the original prescription), AVERAGE, and WARD—the last formally defined only for Euclidean distances, so we report it as a sensitivity variant, not as HRP.

Noisy-μ Markowitz (MV-μ, MV-μ-LW). Recorded separately on the Sharpe track only: the tangency *direction* $\widehat{\Sigma}^{+}\widehat{\mu}$ (sample or LW covariance) with $\widehat{\mu}$ the sample mean vector. The Sharpe ratio is invariant to positive scaling, so the direction is evaluated unnormalized, and a wrong sign honestly shows up as a negative true Sharpe ratio.

4 Simulation framework

The design rests on one device: the true covariance is known, so estimation error—the entire argument for HRP—can be isolated and measured exactly. Each experiment draws a true correlation matrix C from one of four families, true volatilities, and a true mean vector; hands every allocator a finite sample; and scores the resulting weights on the truth.

4.1 Four true-correlation families

The families span the cases where HRP's clustering prior is right, partially right, and wrong (Figure 1a–d).

One-factor (CAPM-like; weak cluster structure). $C = \beta\beta^{\top} + \mathrm{diag}(1-\beta^2)$ with $\beta_i = \sqrt{R_i^2}$ and per-asset factor variance shares R_i^2 drawn uniformly in $[r^2-\Delta,\, r^2+\Delta]$ (clipped to $[0.01, 0.90]$), where the level r^2 and dispersion Δ are sampled per Table 1.

Nested hierarchical (the world HRP is built for). A global factor, B block factors, and nested sub-block factors with variance shares g, b, s:

$$C \;=\; AA^{\top} + \mathrm{diag}(1 - g_i - b_i - s_i), \tag{3}$$

where row i of A carries loadings $\sqrt{g_i}$, $\sqrt{b_i}$, $\sqrt{s_i}$ on its global, block, and sub-block factor. Expected correlations are $\approx g$ across blocks, $\approx g + b$ within a block, and $\approx g + b + s$ within a sub-block. The number of sub-clusters in a block is capped at $\min(S, \lfloor n_b/2 \rfloor)$, with S the sampled sub-clusters-per-block count and n_b the block size, so every sub-cluster contains at least two assets. Shares are jittered per asset, capped so idiosyncratic variance stays ≥ 0.10, and—crucially for Section 6.4—each block's and sub-block's share is scaled by a per-cluster multiplier drawn from $U(1-\texttt{block\_disp},\, 1+\texttt{block\_disp})$: at $\texttt{block\_disp} = 0$ all clusters are equally tight (inverse variance is then near-optimal by symmetry); at large values clusters differ strongly in tightness, which is exactly the asymmetry cluster-level allocation is meant to exploit. The asset order is randomly permuted so no method can read the structure off the input ordering (Figure 1f).

Unstructured (nothing for the dendrogram to find). A normalized Wishart draw: $C = \mathrm{corr}(X^{\top}X)$ with $X \in \mathbb{R}^{m \times N}$ i.i.d. standard normal and $m = \max(N{+}2, \texttt{round}(\texttt{wishart\_dof\_ratio} \cdot N))$, almost surely full rank; lower ratios give wilder random correlations.

Equicorrelation (structure with no hierarchy). $C = (1-\rho)I + \rho\,\mathbf{1}\mathbf{1}^{\top}$ with $\rho \sim U(0.05, 0.70)$, positive definite for $\rho > -1/(N-1)$.

Table 1: The complete sampling design. Each experiment draws one configuration; every sampled value is stored in the per-experiment record. Ranges (a, b) denote $U(a, b)$; braces denote uniform choice.

Parameter	Range / choices	Applies to
N (assets)	$\{10, 30, 50, 100\}$	all
T/N (window/universe)	$\{0.5, 1, 2, 5, 10\}$, $T = \max(5, \texttt{round}(T/N \cdot N))$	all
`vol_lo` (lowest vol/period)	$(0.005, 0.02)$	all
`vol_ratio` ($\sigma_{\max}/\sigma_{\min}$)	$(1.5, 6.0)$	all
`sharpe_mean` (per period)	$(0.0, 0.10)$	all
`sharpe_disp`	$(0.0, 0.05)$	all
t-dof	$(4.0, 10.0)$	Student-t batch
`factor_r2` (r^2)	$(0.10, 0.70)$	one-factor
`factor_r2_disp` (Δ)	$(0.0, 0.15)$	one-factor
B (blocks)	$\{2, 3, 4, 5\}$ (s.t. $B \leq N/4$)	hierarchical
sub-clusters per block	$\{1, 2, 3\}$	hierarchical
`global_share` (g)	$(0.05, 0.30)$	hierarchical
`block_share` (b)	$(0.10, 0.40)$	hierarchical
`sub_share` (s)	$(0.00, 0.30)$	hierarchical
total-share cap	0.90 (idiosyncratic ≥ 0.10)	hierarchical
`loading_jitter`	$(0.0, 0.30)$	hierarchical
`block_disp`	$(0.0, 0.8)$	hierarchical
`wishart_dof_ratio`	$(1.2, 5.0)$, $m = \max(N{+}2, \texttt{round}(\cdot N))$	unstructured
`equi_rho` (ρ)	$(0.05, 0.70)$	equicorrelation

4.2 Sampling design: every knob disclosed

Every scalar that influences the generated truth is either a sampled field of the experiment configuration, drawn from the explicit ranges in Table 1 and stored in every output record, or an asset-level draw made from the experiment RNG whose realized summary (mean off-diagonal correlation, volatility spread, minimum eigenvalue) is recorded per experiment. There are no other knobs. Universe size is drawn from $N \in \{10, 30, 50, 100\}$ and the window length is $T = \max(5, \texttt{round}(T/N \cdot N))$ with $T/N \in \{0.5, 1, 2, 5, 10\}$—the key axis. True volatilities are log-uniform, $\sigma_i = \texttt{vol\_lo} \cdot \texttt{vol\_ratio}^{U(0,1)}$, so $\Sigma = DCD$ with $D = \mathrm{diag}(\sigma)$; the true mean is $\mu_i = s_i \sigma_i$ with per-asset Sharpe $s_i \sim \mathcal{N}(\texttt{sharpe\_mean}, \texttt{sharpe\_disp}^2)$, giving the Sharpe track of Section 6.6 a meaningful oracle.

4.3 Return sampling and exact evaluation

The main batch draws T i.i.d. Gaussian observations $r_t \sim \mathcal{N}(\mu, \Sigma)$. The robustness batch draws a multivariate Student-t with $\nu \sim U(4, 10)$ degrees of freedom—a common chi-square mixing variable across the cross-section, rescaled by $\sqrt{(\nu - 2)/\nu}$ so the population covariance is exactly Σ: the same truth, fatter joint tails. Each experiment draws *two* independent windows from the same truth; the second is used only to measure estimation-noise turnover, $\frac{1}{2}\|w^{(1)} - w^{(2)}\|_1$, the weight movement caused by resampling the data with the truth fixed.

Evaluation is exact. The oracle is the unconstrained minimum-variance portfolio on the true covariance, $w^\star = \Sigma^{-1}\mathbf{1}/(\mathbf{1}^\top \Sigma^{-1}\mathbf{1})$, the exact lower bound on realized variance over all full-investment

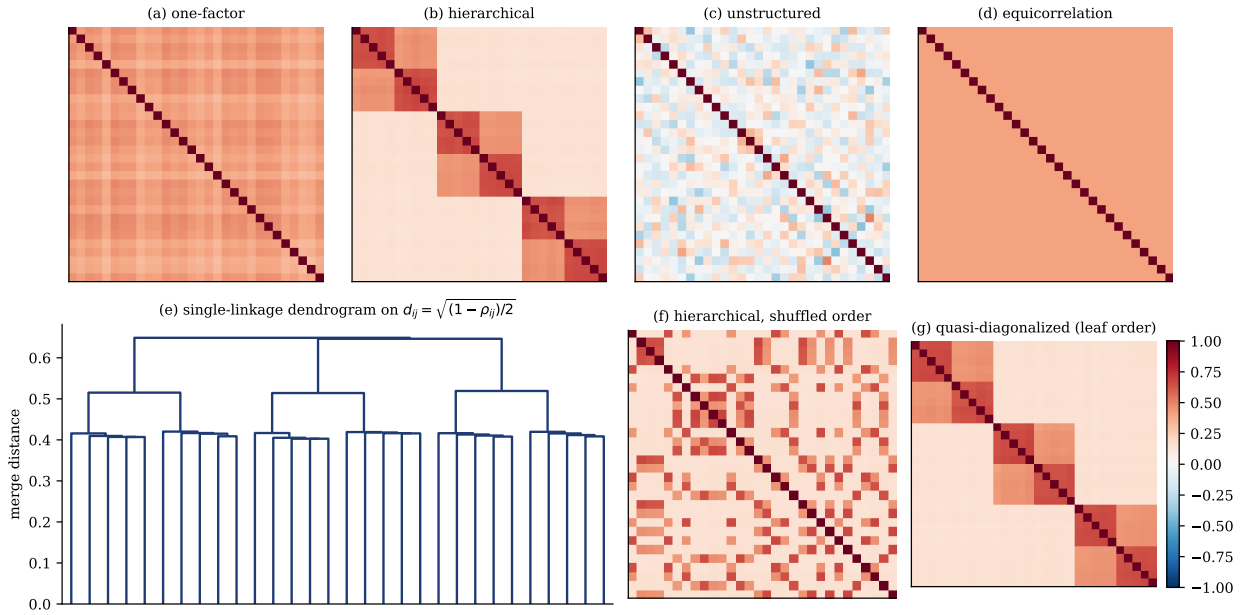

Figure 1: The ground truth and HRP's view of it. (a–d) True correlation matrices from the four families (illustrative fixed configurations, $N = 30$): one-factor, nested hierarchical (three blocks, two sub-clusters each; panel (b) shows it in its natural block order), unstructured normalized-Wishart, and equicorrelation. (e) Single-linkage dendrogram on the correlation distance $d_{ij} = \sqrt{(1-\rho_{ij})/2}$ of the hierarchical truth. (f) The same hierarchical correlation matrix in the randomly permuted asset order the allocators actually see. (g) The matrix reordered by the dendrogram leaf order: HRP's quasi-diagonalization recovers the block structure.

weight vectors, so the headline metric

$$\text{risk ratio}(w) \;=\; \frac{w^\top \Sigma w}{w^{\star\top} \Sigma\, w^\star} \;\geq\; 1 \tag{4}$$

holds by construction. On the Sharpe track each portfolio's true Sharpe ratio $w^\top \mu / \sqrt{w^\top \Sigma w}$ is reported as a fraction of the oracle tangency ceiling $\sqrt{\mu^\top \Sigma^{-1} \mu}$. Because long-only rules (EW, IV, HRP, the clipped variants) are not oracle-optimal even with perfect inputs, their risk ratio contains a specification term in addition to estimation error; the large-T limit of Section 6.2 reveals it.

5 Experimental setup

We run 4,800 experiments: 4,000 Gaussian (seed 101) and 800 Student-t (seed 707), each drawing its family uniformly at random—the Gaussian batch lands 1,007 one-factor, 1,018 hierarchical, 1,011 unstructured, and 964 equicorrelation experiments, with 179–222 Gaussian experiments per family $\times$ T/N cell. Each experiment runs all ten risk-track allocators and both noisy-μ directions on the same window. All results below are stratified by family and T/N; we never let an aggregate hide a family reversal, and aggregates are quoted only when the direction agrees across families. Runs are deterministic given the released seeds (Python 3.14.4, NumPy 2.4.6, SciPy 1.17.1, pandas 3.0.3, scikit-learn 1.9.0); the full per-experiment records ($4{,}800 \times 86$ fields, including every sampled configuration value) ship with the code.

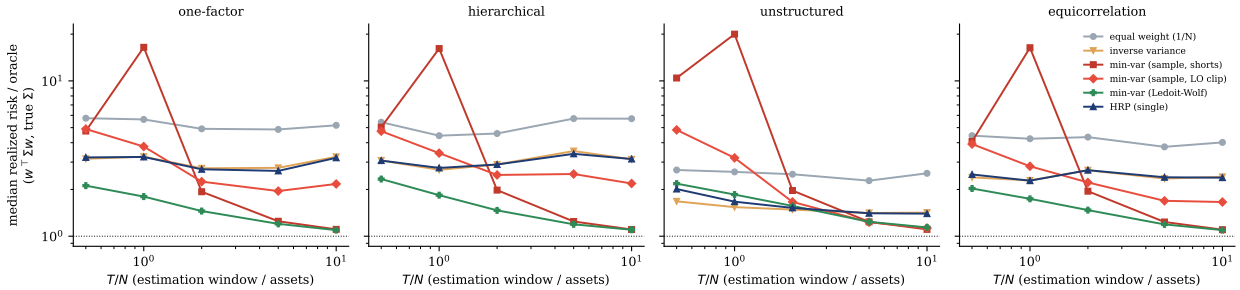

Figure 2: Median realized risk relative to the oracle (Eq. 4, log–log axes) versus T/N, per true-structure family (179–222 Gaussian experiments per point). Six of the ten risk-track methods are shown; average/ward/$\tilde{d}$ HRP and LW-LO are omitted for legibility and sit on top of their siblings (Sections 6.5 and 6.1). Sample minimum variance is catastrophic at $T \approx N$ and best-in-class by $T/N \geq 5$; Ledoit–Wolf tracks or beats HRP everywhere except the unstructured family at $T/N \leq 1$; HRP and inverse variance flatten out at a family-dependent plateau instead of converging to the oracle line at 1.

Table 2: Median realized risk / oracle (Eq. 4) by method and true-structure family, at the singular point $T/N = 1$ and in the data-rich regime $T/N = 10$ (Gaussian batch; 180–212 experiments per cell). Bold marks each column's best method.

	$T/N = 1$ (singular)				$T/N = 10$ (data-rich)			
Method	1-fac.	hier.	unstr.	equi.	1-fac.	hier.	unstr.	equi.
EW ($1/N$)	5.65	4.44	2.60	4.24	5.18	5.71	2.54	4.02
IV	3.25	2.68	**1.54**	2.29	3.24	3.13	1.42	2.41
MV (sample)	16.45	16.15	20.00	16.35	1.11	1.10	**1.11**	1.10
MV-LO (clip)	3.78	3.43	3.19	2.82	2.17	2.18	1.13	1.66
LW	**1.80**	**1.84**	1.85	**1.74**	**1.09**	**1.10**	1.14	**1.09**
LW-LO (clip)	2.88	2.59	1.85	2.31	2.20	2.24	1.15	1.74
HRP (single)	3.24	2.75	1.67	2.28	3.20	3.14	1.40	2.38

6 Results

6.1 Data-starved regime: the insurance value is real

At $T/N \leq 1$ the sample covariance is singular ($T - 1 < N$ in every such experiment), and Eq. (1) on the pseudoinverse is a disaster: pooled across families, sample MV realizes a median 16.8× the oracle risk at $T/N = 1$, exceeds 10× in 65% of $T/N = 1$ experiments and 100× in 9% of them. HRP realizes a median 2.3× at the same point, equal weight 3.8×, inverse variance 2.2× (Table 2, left panel). HRP beats sample MV in 87% of all $T/N \leq 1$ experiments—in the unstructured family at $T/N = 0.5$ it wins 1.00 of them (201 of 201)—and it cannot suffer an inversion catastrophe: its weights are positive by construction, so its worst cases are those of any long-only rule facing an oracle that uses shorts (it exceeds 10× the oracle in 8.8% of $T/N \leq 1$ experiments, of the same order as EW's 19.9% and IV's 8.6%, against 42.6% for sample MV). The practitioner folklore is, to this extent, confirmed and quantified.

Two refinements. First, the catastrophe peaks at the singularity threshold $T \approx N$, not at the smallest sample: at $T/N = 0.5$ sample MV's median is 4.1–10.4× by family, versus 16.2–20.0× at $T/N = 1$. With $T \ll N$ the pseudoinverse solves the problem in a much lower-dimensional

Table 3: Where HRP wins: fraction of experiments in which HRP (single linkage) realizes strictly lower true risk than each opponent, by family, pooled over the data-starved band ($T/N \leq 1$, $n = 378$–418 per family) and the data-rich band ($T/N \geq 5$, $n = 387$–423). Gaussian batch.

Band	Family	vs EW	vs IV	vs MV	vs MV-LO	vs LW	vs LW-LO
$T/N \leq 1$	one-factor	0.97	0.55	0.79	0.89	0.23	0.36
	hierarchical	0.95	0.41	0.86	0.90	0.28	0.37
	unstructured	0.76	0.16	1.00	0.95	0.55	0.55
	equicorr.	0.95	0.37	0.84	0.85	0.32	0.38
$T/N \geq 5$	one-factor	1.00	0.63	0.07	0.05	0.04	0.04
	hierarchical	1.00	0.60	0.04	0.03	0.01	0.02
	unstructured	0.95	0.58	0.08	0.06	0.06	0.06
	equicorr.	1.00	0.46	0.09	0.05	0.06	0.06

range space, an implicit regularization; at $T \approx N$ the matrix is invertible-but-barely, its smallest eigenvalues nearly zero, and the optimizer amplifies them. Second, the crude long-only clip already buys back most of the insurance: MV-LO's median at $T/N = 1$ is 2.8–3.8×, consistent with the constraints-as-shrinkage result of Jagannathan and Ma [9]—though it is still beaten by HRP in 85–95% of $T/N \leq 1$ experiments per family (Table 3).

6.2 Data-rich regime: HRP does not converge

The right panel of Table 2 shows the other side. By $T/N = 10$, sample MV and LW converge to ≈1.10× the oracle in every family (pooled medians 1.10 for both), and the estimation penalty is visibly vanishing along the MV and LW curves of Figure 2. HRP's curve is flat: 3.20 (one-factor), 3.14 (hierarchical), 1.40 (unstructured), 2.38 (equicorrelation) at $T/N = 10$—essentially its $T/N = 1$ values. HRP is not a consistent estimator of the minimum-variance portfolio: more data does not help it, because its loss is dominated by a *specification* gap, not estimation error. The $T/N = 10$ plateau is a direct estimate of that gap—what HRP would lose even given the true covariance—and it ranges from 40% extra risk (unstructured) to 220% (one-factor). The same plateau afflicts inverse variance and equal weight, which is the first hint of Section 6.4: HRP inherits its asymptotics from the inverse-variance rule inside it. Notably, even at $T/N = 10$ the *clipped* variants do not converge either (MV-LO 1.13–2.18): with an oracle that shorts, any long-only rule retains a specification gap; what distinguishes HRP is that it carries this gap while also being promoted as the remedy for estimation error.

6.3 The honest headline: shrinkage beats HRP almost everywhere

Table 2 already shows it in medians; Table 3 and Figure 3 show it experiment by experiment. Ledoit–Wolf minimum variance—which fixes the same singularity HRP fixes, with one line of scikit-learn—realizes lower true risk than HRP at *every* T/N in the one-factor, hierarchical, and equicorrelation families. Pooled across families, HRP's win rate against LW is 0.34 at $T/N \leq 1$ and falls to 0.05 at $T/N \geq 5$. The single exception is the unstructured family at small T: with nothing for the shrinkage target (or the dendrogram) to exploit, HRP edges LW with win rates of 0.54 at $T/N = 0.5$ and 0.57 at $T/N = 1$, and a median paired log-ratio near zero (-0.02 dex at $T/N = 1$).[2] Even this exception deflates on inspection: in the same cells, plain inverse variance is

[2]Even this corner is not uniform: stratified by universe size, HRP's win rate against LW in the unstructured $T/N \leq 1$ pool is 0.34 at $N = 10$, 0.59 at $N = 30$, 0.60 at $N = 50$, and 0.69 at $N = 100$. At $N = 10$ LW is ahead

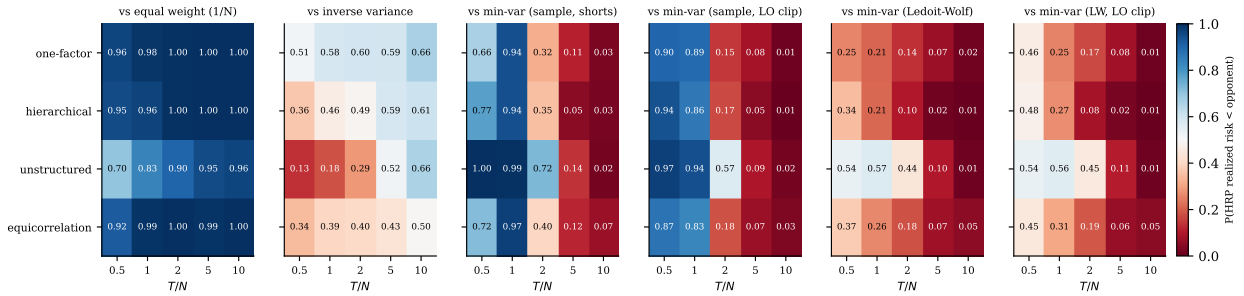

Figure 3: P(HRP realized risk $<$ opponent's) per family $\times$ T/N cell (Gaussian batch, 179–222 experiments per cell; blue = HRP better, red = opponent better). HRP dominates equal weight everywhere and naive minimum variance (with or without the long-only clip) wherever $T/N \leq 1$; against Ledoit–Wolf it is below 0.5 in every cell of the one-factor, hierarchical, and equicorrelation families, with the unstructured small-T corner (win rates 0.54–0.57) the lone exception. Against inverse variance HRP hovers near 0.5 except in the unstructured family at small T, where the clustering machinery actively hurts (0.13–0.29).

better still (median 1.54× vs. HRP's 1.67× at $T/N = 1$; HRP beats IV in only 0.16 of unstructured $T/N \leq 1$ experiments). Where HRP beats shrinkage, it does so not because clustering helps but because *correlations are pure noise there and HRP mostly ignores them*—and the allocator that ignores them entirely wins.

It is worth being precise about what HRP wins against, because the practitioner comparison is usually HRP versus naive Markowitz: against sample MV and the clipped MV-LO at $T/N \leq 1$, HRP wins 0.79–1.00 of experiments per family. The insurance is real; it is the choice of comparator that flatters HRP. Against the equally inversion-proof LW, the picture inverts almost everywhere.

6.4 HRP is mostly the inverse-variance portfolio inside it

Against diagonal inverse variance, HRP's win rates hover near 0.5 in most cells (Table 3, "vs IV"), and the median paired log-ratios are tiny (within ± 0.006 dex, i.e. $\pm 1.5\%$ in risk, in every hierarchical-family cell). The pattern splits by band: with ample data HRP is modestly ahead of IV in three of four families (win rates 0.46–0.63 at $T/N \geq 5$), while in the data-starved band the clustering machinery earns its keep only where the truth has the specific asymmetry it is built to exploit. Our hierarchical generator makes that asymmetry an explicit, sampled parameter: `block_disp` scales each cluster's variance share by $U(1-d, 1+d)$, so $d = 0$ gives equally tight clusters (inverse variance near-optimal by symmetry) and large d gives clusters of very different tightness. Within the hierarchical family at $T/N \leq 2$, HRP's win rate against IV rises monotonically across `block_disp` terciles: 0.35 (mean $d = 0.13$), 0.40 (mean $d = 0.38$), 0.56 (mean $d = 0.66$). The median paired advantage remains small even in the top tercile (-0.002 dex, $\approx 0.4\%$ less risk), so the dispersion effect is real but modest in magnitude—and note that against LW the same terciles read 0.28/0.19/0.19: cluster asymmetry helps HRP relative to its own diagonal core, not relative to shrinkage. In short: most of HRP's behavior is the IVP; the dendrogram adds a structure-conditional refinement, not a transformation.

even here; the exception is a large-N, small-T phenomenon.

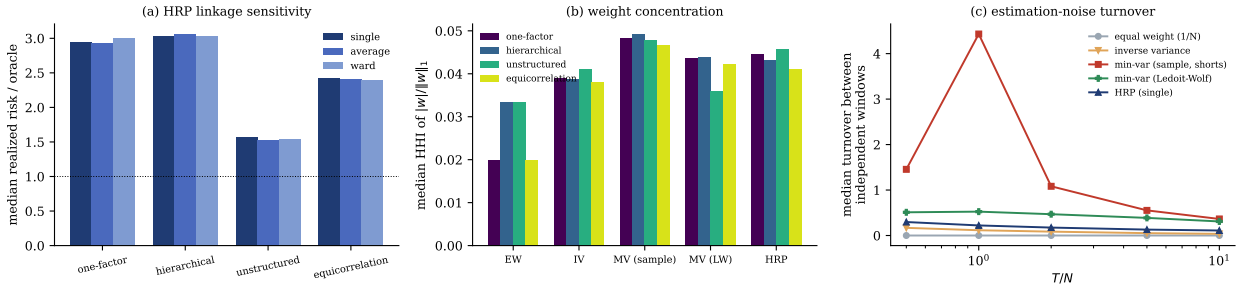

Figure 4: (a) Linkage sensitivity: median realized risk / oracle for single/average/ward HRP, by family (all T/N pooled); the spread within each family is 0.02–0.06. (b) Median concentration (Herfindahl index of $|w|/\|w\|_1$) by method and family: every method is diffuse in the median; EW's $1/N$ differs across families only through their N mixes. (c) Estimation-noise turnover: median $\frac{1}{2}\|w^{(1)} - w^{(2)}\|_1$ between two independent windows drawn from the same truth, by T/N (all families pooled). Sample minimum variance peaks at 4.43 at the singularity threshold $T/N = 1$; HRP stays at 0.11–0.30, below Ledoit–Wolf throughout.

6.5 Linkage does not matter (and turnover diagnostics)

Practitioner debate spends considerable energy on the linkage rule, citing single linkage's chaining pathology [23] and preferring average or Ward variants [26]. On the minimum-variance criterion the choice is immaterial in our testbed: the median realized-risk ratios of single/average/ward HRP per family agree to 2.95/2.93/3.00 (one-factor), 3.04/3.06/3.04 (hierarchical), 1.57/1.53/1.54 (unstructured), and 2.42/2.42/2.39 (equicorrelation)—the three linkage medians agree to within 0.02–0.06 (Figure 4a), a disagreement one to two orders of magnitude smaller than the gaps between allocator families in Table 2. No linkage is best in more than two families, and which one "wins" flips across families. Anyone tuning HRP should tune *whether* to cluster, not *how* to link.

The distance-matrix convention matters just as little, which doubles as a faithfulness check on our implementation (Section 3): recomputing single-linkage HRP with the linkage on the Euclidean distance-of-distances $\tilde{d}$—the matrix the original code listing hands to the clustering routine—moves the per-family median risk ratio by at most 0.02 (2.95 → 2.94 one-factor, 3.04 → 3.05 hierarchical, 1.57 → 1.58 unstructured, 2.42 → 2.44 equicorrelation), with a paired median log-ratio of exactly zero and a $\tilde{d}$ win rate of 0.44–0.48 per family. The per-experiment weights genuinely differ—the two conventions realize bit-identical risk in only 4–9% of experiments—but no aggregate conclusion in this paper changes under either convention.

The same figure reports two operational diagnostics. Concentration (Figure 4b): all methods are diffuse in median terms (gross-weight Herfindahl 0.02–0.08), with HRP (≈0.04) close to IV and LW; the optimizers are not visibly more concentrated in the median—their pathology is instability, not static concentration, and sample MV holds a median gross short position of 0.57 versus 0.25 for LW. Estimation-noise turnover (Figure 4c) makes the instability explicit: redrawing the estimation window from the same truth moves sample MV's weights by a median half-L_1 distance of 4.43 at $T/N = 1$—more than four times the entire book—versus 0.52 for LW, 0.22 for HRP, and 0.12 for IV. HRP's practical appeal of low, stable turnover is genuine; it is inherited from the near-diagonal allocation, and LW pays roughly twice HRP's turnover for its risk advantage.

Table 4: Sharpe track: median true Sharpe ratio as a fraction of the oracle tangency ceiling, by family and T/N band (Gaussian batch). MV-μ and MV-μ-LW are tangency directions built from estimated means with the sample and Ledoit–Wolf covariance; "neg." is the fraction of experiments with negative true Sharpe ratio for MV-μ; the last column is the *oracle* minimum-variance portfolio's own capture (the ceiling for the risk-only rules under this mean model).

Band	Family	EW	LW	HRP	MV-μ	MV-μ-LW	neg.	or. MV
$T/N \leq 1$	one-factor	0.37	0.26	0.36	0.11	0.22	0.28	0.20
	hierarchical	0.40	0.30	0.39	0.10	0.21	0.29	0.22
	unstructured	0.66	0.70	0.57	0.11	0.20	0.23	0.84
	equicorr.	0.40	0.28	0.39	0.08	0.19	0.33	0.23
$T/N \geq 5$	one-factor	0.34	0.23	0.36	0.55	0.56	0.05	0.24
	hierarchical	0.36	0.21	0.36	0.56	0.57	0.06	0.21
	unstructured	0.67	0.82	0.67	0.71	0.64	0.02	0.85
	equicorr.	0.44	0.26	0.44	0.48	0.50	0.06	0.26

6.6 Estimated means are the real catastrophe

Everything above concerns risk-only allocators, which is where the HRP-versus-shrinkage debate lives. For perspective, the same harness runs the classical Markowitz tangency direction with *estimated* means, $\widehat{\Sigma}^{+}\widehat{\mu}$, and scores its true Sharpe ratio against the oracle tangency ceiling $\sqrt{\mu^{\top}\Sigma^{-1}\mu}$ (Table 4). At $T/N \leq 1$ the noisy-μ direction captures a median 8–11% of the attainable Sharpe ratio by family (pooled 10%), and its true Sharpe ratio is *negative*—the estimated direction points the wrong way—in 28% of experiments. Shrinking the covariance inside the tangency formula helps (19–22% captured) but does not rescue it; the failure is in $\widehat{\mu}$, reproducing DeMiguel et al. [7] and the error-sensitivity hierarchy of Chopra and Ziemba [3] inside a fully controlled environment. By contrast, the three μ-free rules of Table 4 capture 26–70% in the same regime, and even the weakest μ-free allocator on this track, sample minimum variance, captures a pooled median 18% (per-family medians 0.15–0.25)—nearly twice the noisy-μ direction. Even with ten observations per asset, noisy-μ tangency captures only 48–71% and is still negative in 2–6% of experiments.

Two honest caveats from the same table. First, minimum variance is not the Sharpe-optimal objective: the *oracle* minimum-variance portfolio itself captures only 0.20–0.26 of the tangency Sharpe in the three structured families (it captures 0.84–0.85 in the unstructured family, where the tangency and minimum-variance portfolios nearly coincide under our mean model). Equal weight and HRP, whose Sharpe capture (0.34–0.44 in structured families) *exceeds* the oracle minimum-variance floor's, profit from a mean model in which all assets have similar expected Sharpe ratios—a regime that flatters diffuse portfolios on the Sharpe track, exactly as the $1/N$ literature would predict. Second, these Sharpe-track numbers therefore do not contradict the risk-track ranking; they answer a different question ("how much Sharpe, given this mean structure?" versus "how close to the minimum-variance floor?").

6.7 Student-t robustness

The 800-experiment multivariate Student-t batch ($\nu \sim U(4, 10)$, covariance rescaled to exactly Σ) reproduces every qualitative ranking (Table 5). Fat joint tails make the inversion-based estimators a little worse in every family (sample MV by +0.04 to +0.25 in the median, LW by +0.02 to +0.11); the diffuse rules move by up to 1.1 in the noisier 193–205-experiment t-cells without changing sign or order anywhere. HRP's win rate against LW by family is 0.13–0.38 under Student-t versus 0.14–

Table 5: Gaussian vs. Student-t sampling: median realized risk / oracle by family (all T/N pooled; 964–1,018 Gaussian and 193–205 Student-t experiments per family), and HRP's win rate against Ledoit–Wolf in the same pool.

Returns	Family	EW	MV	LW	HRP	win HRP vs LW
Gaussian	one-factor	5.24	1.90	1.41	2.95	0.14
	hierarchical	5.22	1.97	1.43	3.04	0.14
	unstructured	2.51	1.88	1.48	1.57	0.32
	equicorr.	4.14	1.82	1.34	2.42	0.19
Student-t	one-factor	5.14	2.05	1.48	3.19	0.13
	hierarchical	4.16	2.22	1.47	2.50	0.18
	unstructured	2.48	2.08	1.60	1.73	0.38
	equicorr.	4.47	1.86	1.36	2.65	0.19

0.32 under Gaussian sampling, with the unstructured family again the most HRP-friendly and the one-factor family the least. Fatter tails do not reorder the comparison.

7 Discussion

When should one actually use HRP? Our results support a narrow but real use case: *covariance structure unknown or plausibly unstructured, and* $T \lesssim N$. There HRP matches Ledoit–Wolf (win ≈0.55), never explodes, and turns over a twentieth as much as unconstrained sample MV (0.22 vs. 4.43 at $T/N = 1$) and a third as much as the clipped variant. If one believes the truth is dense random correlation, plain inverse variance is better still; HRP is the hedge for not knowing. In every structured world we tested—one common factor, genuine nested hierarchy (the case HRP was designed for!), constant correlation—Ledoit–Wolf shrinkage realized less true risk at every T/N, with win rates against HRP of 0.63–0.99 per cell. The deepest irony of the comparison is that HRP loses most clearly in the hierarchical family itself: knowing that the truth is hierarchical is covariance information, and the estimator that uses all of $\widehat{\Sigma}$ (gently shrunk) exploits it better than a greedy dendrogram that uses only the ordering it induces.

What the plateau means. Because the oracle is exact, the $T/N \to 10$ limit cleanly separates the two failure modes that backtests confound. The optimizers' excess risk is almost entirely estimation error (it vanishes: 1.10× at $T/N = 10$); HRP's is almost entirely specification error (it persists: 1.40–3.20×). The practical corollary: more history helps Markowitz-type rules and does essentially nothing for HRP, so the break-even T/N between them is not a constant of nature but falls out of the data: HRP's win rate against sample MV crosses 0.5 between $T/N = 1$ and 2 in the three structured families and between 2 and 5 under unstructured correlation. This locates HRP and minimum variance at the two ends of the Schur-complement continuum of Cotton [6] along an *estimation-error* axis: our results say to sit near the HRP end only briefly, and to slide toward the (shrunk) MV end as soon as T permits.

What HRP's machinery actually contributes. The decomposition across Sections 6.4–6.5 is unflattering to the celebrated parts of the algorithm. The dendrogram's linkage rule is irrelevant (median spreads of 0.02–0.06, and at most 0.02 for the distance-matrix convention); the quasi-diagonalization matters only through the bisection it enables; and in the data-starved band the bisection improves on diagonal inverse variance only when clusters differ materially in tightness

(`block_disp` terciles: 0.35 → 0.56 win rate vs IV), by a median margin under half a percent of risk—with ample data it edges IV more broadly (win rates 0.46–0.63 at $T/N \geq 5$), but those wins do not close the gap to shrinkage. HRP's robustness—the property that sells it—comes from the part nobody advertises: it is approximately an inverse-variance portfolio, and inverse variance is hard to break. Stated plainly: if you want HRP's safety, IV gives you most of it; if you want optimality, shrinkage gives you more of it; HRP's niche is the intersection where you want a long-only, inversion-free rule that still reacts a little to cluster structure.

Means remain the priority. Nothing in the covariance debate compares to the damage of estimating means: a median 90% of attainable Sharpe destroyed at $T/N \leq 1$, with the direction outright wrong in 28% of experiments. Practitioners arguing HRP versus shrinkage while feeding estimated means into anything have, in our measurements, chosen the wrong battle by an order of magnitude.

8 Limitations

Returns in our testbed are i.i.d. across time—Gaussian in the main batch, multivariate Student-t (common mixing variable, $\nu \in [4, 10]$) in the robustness batch—and the true covariance is constant over the estimation window. Real returns exhibit autocorrelation, volatility clustering, regime switches, and asymmetric tail dependence; all of these degrade every estimator, and dynamics that mimic block structure intermittently could shift the HRP-versus-shrinkage margin in either direction. Our evaluation is single-period with no transaction costs: we report estimation-noise turnover as a diagnostic but do not net it against risk performance, which would favor HRP and IV. The long-only variants use the crude clip-and-renormalize heuristic rather than the constrained quadratic program, deliberately (it is what naive implementations do), so our MV-LO and LW-LO numbers are lower bounds on what proper constrained optimization achieves. We likewise test a single shrinkage estimator—Ledoit–Wolf toward a scaled identity; the constant-correlation target [15], the factor target [14], and nonlinear shrinkage are untested, and the lone unstructured small-T corner where HRP edges LW might not survive a better-matched target. The oracle floor allows short positions, so long-only rules carry a structural specification gap in the risk ratio; this affects levels, not the paired comparisons between long-only methods. Our four families, though spanning factor, hierarchical, dense-random, and equicorrelated worlds, are all static linear structures—three with positive mean correlation, while the unstructured family's mean off-diagonal correlation is approximately zero—and the mean model ($\mu_i = s_i\sigma_i$ with similar per-asset Sharpe ratios) is one defensible choice among several; the Sharpe-track levels in Table 4 depend on it, though the risk-track results do not use means at all. Finally, we test HRP as specified in the original paper plus linkage variants; extensions (HERC, constrained or Schur-blended versions) may behave differently and can be dropped into the released harness.

9 Conclusion

We measured Hierarchical Risk Parity against Markowitz-type and shrinkage allocators in a controlled environment where the true covariance is known and every realized risk is exact. The popular story is half right. HRP's insurance value against naive sample minimum variance is real and large: at $T = N$ the optimizer realizes a median 16.8× the attainable risk floor and HRP 2.3×, winning 87% of data-starved comparisons. But against Ledoit–Wolf shrinkage—which cures the same singularity at the same computational cost—HRP loses at every T/N in every structured

family we tested, including the nested-hierarchical worlds it was designed for, winning only in the unstructured small-T corner where its own diagonal core (inverse variance) is better still. HRP does not converge to the optimum as data grows (plateau at 1.40–3.20×), its much-debated linkage choice moves family medians by 0.02–0.06, and its genuine advantages—no inversion, positive weights, low estimation-noise turnover—are inherited from the inverse-variance portfolio it wraps. Estimated means remain the catastrophe that dwarfs the covariance debate: noisy-μ tangency captures a median 10% of attainable Sharpe at $T/N \leq 1$ and points the wrong way 28% of the time. Use HRP when the structure is unknown and the window is short; shrink when you can; and treat any allocator that needs estimated means as the actual emergency.

Reproducibility. All experiments are deterministic given the released seeds (101 Gaussian, 707 Student-t); the script `scripts/run_all.py` regenerates every record, summary, and figure input (`results/results.json`, `results/records.csv`, 4,800 experiments × 86 recorded fields), `python -m hrp_experiments.figures` regenerates the four figures, and `scripts/check_paper_numbers.py` verifies that every number quoted in this paper matches the generated results. The covariance generator (Table 1 is its complete parameter space), allocators, simulation harness, and analysis are provided as an open-source package at https://github.com/suenot/hrp-validation.

References

[1] 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. doi: 10.1093/rfs/4.2.315.

[2] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017. doi: 10.1016/j.physrep.2016.10.005.

[3] Vijay K. Chopra and William T. Ziemba. The effect of errors in means, variances, and covariances on optimal portfolio choice. *The Journal of Portfolio Management*, 19(2):6–11, 1993. doi: 10.3905/jpm.1993.409440.

[4] Roger Clarke, Harindra de Silva, and Steven Thorley. Minimum-variance portfolio composition. *The Journal of Portfolio Management*, 37(2):31–45, 2011. doi: 10.3905/jpm.2011.37.2.031.

[5] Roger G. Clarke, Harindra de Silva, and Steven Thorley. Minimum-variance portfolios in the U.S. equity market. *The Journal of Portfolio Management*, 33(1):10–24, 2006. doi: 10.3905/jpm.2006.661366.

[6] Peter Cotton. Schur complementary allocation: A unification of hierarchical risk parity and minimum variance portfolios, 2024. URL https://arxiv.org/abs/2411.05807.

[7] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009. doi: 10.1093/rfs/hhm075.

[8] Markus Jaeger, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock, and Peter Schwendner. Interpretable machine learning for diversified portfolio construction. *The Journal of Financial Data Science*, 3(3):31–51, 2021. doi: 10.3905/jfds.2021.1.066.

[9] 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. doi: 10.1111/1540-6261.00580.

[10] J. D. Jobson and Bob Korkie. Estimation for Markowitz efficient portfolios. *Journal of the American Statistical Association*, 75(371):544–554, 1980. doi: 10.1080/01621459.1980.10477507.

[11] J. D. Jobson and Bob M. Korkie. Putting Markowitz theory to work. *The Journal of Portfolio Management*, 7(4):70–74, 1981. doi: 10.3905/jpm.1981.408816.

[12] Raymond Kan and Guofu Zhou. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*, 42(3):621–656, 2007. doi: 10.1017/S0022109000004129.

[13] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, 1999. doi: 10.1103/PhysRevLett.83.1467.

[14] 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. doi: 10.1016/S0927-5398(03)00007-0.

[15] Olivier Ledoit and Michael Wolf. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004. doi: 10.3905/jpm.2004.110.

[16] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

[17] Harald Lohre, Carsten Rother, and Kilian Axel Schäfer. Hierarchical risk parity: Accounting for tail dependencies in multi-asset multi-factor allocations. In Emmanuel Jurczenko, editor, *Machine Learning for Asset Management: New Developments and Financial Applications*, pages 329–368. Wiley–ISTE, Hoboken, NJ, 2020. doi: 10.1002/9781119751182.ch9.

[18] Marcos López de Prado. Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 42(4):59–69, 2016. doi: 10.3905/jpm.2016.42.4.059.

[19] Rosario N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999. doi: 10.1007/s100510050929.

[20] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: 10.2307/2975974.

[21] Richard O. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989. doi: 10.2469/faj.v45.n1.31.

[22] Marat Molyboga. A modified hierarchical risk parity framework for portfolio management. *The Journal of Financial Data Science*, 2(3):128–139, 2020. doi: 10.3905/jfds.2020.1.038.

[23] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, 2(1):86–97, 2012. doi: 10.1002/widm.53.

[24] Johann Pfitzinger and Nico Katzke. A constrained hierarchical risk parity algorithm with cluster-based capital allocation. Working Paper 14/2019, Department of Economics, Stellenbosch University, 2019. URL https://ideas.repec.org/p/sza/wpaper/wpapers328.html.

[25] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral, and H. Eugene Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471–1474, 1999. doi: 10.1103/PhysRevLett.83.1471.

[26] Thomas Raffinot. Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management*, 44(2):89–99, 2017. doi: 10.3905/jpm.2018.44.2.089.

[27] Thomas Raffinot. The hierarchical equal risk contribution portfolio. SSRN Working Paper No. 3237540, 2018.