

Quantifying the Contributions of Clustering to Statistical Arbitrage^{*}

Lunji Zhu[†] Yixuan He[‡] Mihai Cucuringu[§]

This version: September 2026

Abstract

Clustering-based statistical arbitrage makes two decisions, the clustering and the trading model, usually evaluated together. We separate them by trading 12 clusterizers and two controls, no clustering and a random partition, on 27 years of large US equities, under a fixed cross-sectional rule and a neural trading head. Clustering adds alpha and cuts book volatility by 28% to 42%, while the trading head adds primarily alpha. The clusterizers disagree on the groups yet trade closely correlated books; their Sharpe-ratio differences arise mainly through volatility. Most of the 12 clusterizers retain alpha that factors and per-stock peer regressions together cannot explain.

Keywords: statistical arbitrage, clustering, signed graphs, Sharpe ratio, factor models

JEL classification: C58, G11, C38

^{*}We thank NVIDIA Corporation and ASU Research Computing: the computations in this article were performed on an NVIDIA GPU server awarded to Yixuan He and on the Sol supercomputer at Arizona State University. This work was supported by an NVIDIA Academic Grant Award and a Jetstream2 AI Fellowship to Yixuan He. The authors declare that they have no competing interests.

[†]Corresponding author. Department of Mathematics, University of California, Los Angeles. Email: lunji@math.ucla.edu.

[‡]School of Mathematical and Natural Sciences, Arizona State University.

[§]Department of Mathematics, University of California, Los Angeles, and Department of Statistics and Oxford-Man Institute of Quantitative Finance, University of Oxford.

1 Introduction

Machine learning methods are now used at most stages of trading strategies (Kelly and Xiu, 2023), and clustering has become one of the standard ways to form peer groups for relative-value trading. A clustering-based statistical arbitrage strategy makes two decisions that are usually evaluated together: it partitions the cross-section into comoving peer groups, and a trading model builds a book, a set of long and short positions, on that partition, typically from returns residualized against the market. The effectiveness of each component is then inferred from the Sharpe ratio of the two together (Jin et al., 2023; Lee and Chang, 2026; Brutti et al., 2026), even though end-to-end performance does not identify which component produced it. To closely examine the contribution of clustering to statistical arbitrage, we ask three questions: how much the clusterizer and the trading model each contribute, and whether through the mean or the volatility of the payoff; whether different clusterizers recover different information; and whether what the clusterizer contributes could be obtained instead from traditional factor models, from a reversal factor, or from per-stock peer regressions.

The grouping of the cross-section has been studied for far longer than the strategies that trade it, and the groups have come from three sources. An industry taxonomy supplies them in advance (Bhojraj et al., 2003). The groups can instead be estimated from returns, usually from the correlation matrix (Mantegna, 1999), which Tola et al. (2008) put to work in portfolio construction and Jin et al. (2023) cluster for statistical arbitrage. Most recently, the groups have been learned directly against a trading objective: Lee and Chang (2026) fit the grouping and the model that trades it jointly, inside one transformer. Economic taxonomy, the statistical structure of returns, and the trading objective itself have all been brought to bear on how the groups are formed.

A second strand of literature asks what a strategy does with such a partition. Relative-value trading has been extended since Gatev et al. (2006) documented the profitability of trading a single mean-reverting spread. Avellaneda and Lee (2010) replaced the pair with the residual left after removing a systematic component; a partition supplies another such component, the group mean. Clustering-based strategies are a more recent line of work. Jin et al. (2023), Lee and Chang (2026), and Brutti et al. (2026) each contribute clusterizers of their own together with the strategy they support. Jin et al. (2023) compare five spectral clusterizers, four of them signed, under one rule. Their object is the performance of alternative spectral clusterizers; ours is the contribution of clustering itself within the trading pipeline.

To the best of our knowledge, no previous study compares clusterizers from different families under a common trading model, or separates what the partition contributes from what the model contributes. When the two are evaluated together, a higher Sharpe ratio may come from a better partition, from a better model, or from the way the two match, so end-to-end performance alone cannot attribute the gain. Attribution also requires the mean and the volatility of the payoff to be examined separately. Guijarro-Ordóñez et al. (2025) examine both for factor-based statistical arbitrage, reporting the Sharpe ratio, mean, and volatility of every model they fit and varying the factor construction under different signal models, but clustering enters their study only as a remark.

We examine the clusterizers and the trading models separately. For the clustering stage, we compare 12 clusterizers from six families (geometric, hierarchical, eigenstructure-based, learned, trading-driven, and a sector taxonomy) against two controls. The controls are a single all-market group ($K = 1$) and a uniform random partition with the same number of groups as the real partitions. This paper introduces two of the six families: trading-driven clusterizers, which optimize a soft assignment against the payoff of the book the assignment induces and then harden the assignment into a partition, so that a grouping trained on a trading objective can still be compared with one that was not; and SSSNET (He et al., 2022), a signed-graph neural network brought to statistical arbitrage for the first time. Our objective is not to estimate the best attainable performance of each clustering family, but to compare what their partitions contribute under a common downstream pipeline.

For the trading stage, we trade every partition under two trading components: a fixed cross-sectional rule (adapted from Jin et al. (2023)) and a neural trading head in the design of Guijarro-Ordóñez et al. (2025), with the long convolution of Fu et al. (2023) in place of their transformer. Because the fixed rule has no fitted parameter, its performance comes from the partition alone; the neural trading head then shows how the conclusions change when a fitted model is placed downstream of the partition. Both components are run on the same panel, the 500 largest US equities, daily, over 27 years.

The comparison rests on three standards. First, the two controls measure the cost of a missing partition and of a wrong one separately. Second, ranking gaps are judged against the dispersion of the design’s own replications, the trading head refitted across its seeds and the partition redrawn under its own seed. Third, a Sharpe gap is split into a mean part and a volatility part (Section 4.3), because the sample does not measure the two equally (Section 3.7). Together, these standards turn an end-to-end performance comparison into a component-level evaluation: the controls identify whether the component contributes, the replications show whether competing implementations can be distinguished, and the decomposition identifies the channel through which the performance gap arises.

Clustering lowers the volatility of the book and raises its alpha; the neural trading head widens the alpha advantage while narrowing the volatility benefit. First, demeaning the signal within real peer groups removes the comovement those groups share, and the book built on the demeaned signal runs about 40% less volatile than either control under the fixed rule, and about a third less under the trading head (Section 4.1). Second, a return demeaned within its peer group is a sharper relative-value input, and at the rule-based component the two controls hold the lowest alphas of the 14 books (Section 4.1). Under the neural trading head the alpha advantage widens over both controls, especially over the random control, so a wrong partition costs more under the head than under the fixed rule.

The neural trading head raises alpha at all 14 arms while raising volatility by proportionally less (Section 4.4). The ranking it produces largely agrees with the rule-based component’s, so the ranking reflects the quality of the partition as a trading input rather than the trading model that produced it, though the middle of the ranking moves by up to five places and the gap between the two controls closes.

We find that the 12 clusterizers group stocks on different principles and produce partitions that agree little by the adjusted Rand index, yet trade highly correlated books. Whichever way they draw the groups, demeaning within them removes far more comovement than the random control removes, and far more than no clustering, which removes nothing; ten of the 12 remove nearly the same amount (Section 4.2). The resemblance also grows down the pipeline, in the trading positions the partitions induce and again in the payoffs those positions earn, until two different clusterizers differ about as much as one clusterizer differs from its own refit (Section 5.4). Thus, the data clearly distinguish a real partition from the no-clustering and random controls, but provide much weaker separation among the clusterizers themselves. Most adjacent gaps in the Sharpe ranking lie within the run-to-run variation in the gap itself (Section 5.1). A clusterizer is not even stable against itself: redrawing the partition’s own seed rearranges most of what switching the algorithm rearranges, and the neural trading head enlarges rather than damps the Sharpe movement that redraw causes (Section 5.2). The clusterizers’ Sharpe ratios differ mainly through the volatility of their books; their positions are highly similar. Because the books are this similar, re-selecting the trailing best each year does not outperform an equal weight of all 12 (Section 5.4). The inputs a clusterizer is given matter on the scale of the whole span between clusterizers: a longer estimation window moves the Sharpe ratio by as much as that span, and a different number of groups by more than two fifths of it, so the Sharpe ratio separates inputs more sharply than algorithms (Section 5.5).

Neither a factor model nor per-stock peer regressions make a partition redundant. Gujjarro-Ordóñez et al. (2025) remark that latent factor estimation can, under appropriate assumptions, take the place of clustering. Taken as stated, the remark would make the clustering stage redundant, so we test it. After the leading principal components, and again the Fama–French factors, are projected out, every real partition loses at most about half of the variance share it explains, and none falls to the share a random partition with the same number of groups would remove (Section 6.1). In the time series, every book retains positive and significant alpha after controlling for the standard factors, a short-term reversal factor, and the leading principal components; the reversal factor absorbs no more than half of any alpha and more of the controls’ than of the clusterizers’. Regressed on the same pipeline run without clustering, every clusterizer’s book keeps between two fifths and two thirds of its alpha while the random control’s keeps almost none (Section 6.2).

Because clustering is a special case of peer-based trading, in which every stock in a group shares one peer set, we also compare it with per-stock peer regressions, in which each stock has its own. Peer regressions are a stronger substitute than the factors, and in the cross-section they do absorb much of what a partition summarizes, which neither the principal components nor the Fama–French factors do. The peer regressions do not, however, fully absorb what the partitions earn. Against the peer books and the factors together, both controls lose their alpha entirely while most of the 12 clusterizers keep theirs. Regressed the other way, the peer books keep their alpha against the equal weight of all 12 (Section 6.3). The two constructions summarize

similar comovements but trade differently on them.

The article proceeds as follows. Section 2 places the design in the literature. Section 3 sets out the panel, the 12 clusterizers, the two controls, and the two trading components. Section 4 decomposes what the clustering contributes, gives the mechanism, and measures what the trading head adds on top of it. Section 5 compares the 12 clusterizers against the noise of the design. Section 6 tests whether factor models or per-stock peer regressions make the partition redundant, then charges trading costs, under which no daily-rebalanced book is net positive at five basis points one-way. Section 7 concludes. The appendix gives the per-clusterizer construction, the trading head’s architecture and schedule, and the additional results the text refers to.

2 Related Literature

This article combines two strands of previous work: methods for recovering peer-group structure from stock return time series, and modern statistical arbitrage strategies. In this section, we briefly review the literature in each strand, and highlight the gap that our work fills.

Industry taxonomies are the usual benchmark for a similarity-based clustering, and the evaluation criterion is how much contemporaneous return comovement the groups explain. Taxonomies such as the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS), and the Global Industry Classification Standard (GICS) partition the cross-section according to a classifying body’s judgment about firms’ primary lines of business. Bhojraj et al. (2003) show that GICS outperforms SIC, NAICS, and the 48-industry algorithm of Fama and French (1997) on that criterion; the other three perform similarly to one another, and the GICS advantage is sharpest among large firms. Chan et al. (2007) likewise find that within-industry comovement is strongest among large caps, and that finer GICS levels separate comovement further. They also report that industry partitions are more homogeneous out of sample than groups obtained by statistical cluster analysis. Peers can also be chosen firm by firm rather than read from a shared classification (Bhojraj and Lee, 2002), a construction we return to in Section 6.3.

The correlation matrix of stock returns also contains information about market structure: the structure it encodes is hierarchical and tree-like (Mantegna, 1999; Mantegna and Stanley, 1999) and it behaves differently in calm and in crash periods (Onnela et al., 2003). That information can be extracted in several ways: by hierarchical linkage (Tola et al., 2008), by its spectrum (MacMahon and Garlaschelli, 2015; Rohe et al., 2011), and by signed-graph methods that separate positive from negative correlation (Cucuringu et al., 2019; He et al., 2022; Uddin et al., 2024). A partition is also a factor model with grouped loadings (Ando and Bai, 2017), which is the form in which we later ask whether a factor model makes it redundant. A partition of the cross-section is used as a restriction in covariance estimation, where it is normally read from a taxonomy: Alves et al. (2024) restrict the residual covariance of the S&P 500 to ten SIC sectors. The linkage, spectral, and signed-graph methods all start from the same correlation matrix; Section 5 and Section 6 ask whether they recover different information from it, and whether what they recover survives a factor model or a per-stock peer regression. We study what a clusterizer contributes to trading, holding the trading model fixed at each of two models in turn, while 12 clusterizers and two controls read the same panel.

Machine learning has entered statistical arbitrage stage by stage. Gatev et al. (2006) formalize pairs trading as a rules-based relative-value arbitrage and document its profitability over decades of US data. The selection step has been automated by machine learning, from return similarity (Sarmiento and Horta, 2020) and from firm characteristics (Han et al., 2023). Avellaneda and Lee (2010) replace the pair with a factor or ETF decomposition and bet on an Ornstein–Uhlenbeck “ s -score” built on the residual. The trading model itself was then learned from data: Krauss et al. (2017) benchmark deep networks, gradient-boosted trees, and random forests on lagged returns of the S&P 500 cross-section, Fischer and Krauss (2018) test long short-term memory networks on the same data, Guijarro-Ordóñez et al. (2025) learn the trading signal and its allocation jointly on the residuals of a factor model, and Epstein et al. (2025) learn the latent factors themselves with an attention mechanism over firm characteristics. The models come from a wider machine learning program in asset pricing and financial econometrics (Gu et al., 2020; Kelly et al., 2019; Zhang et al., 2024, 2025). Each applies machine learning to a different input of a trading strategy: expected returns, factor loadings, volatility, and covariances. We take one such model, a neural trading head trained on the Sharpe ratio of its own book, and ask what it adds to a partition held fixed and whether it erases the difference between a real partition and a random one.

The signal these strategies trade is a reversal, and the reversal literature has already found that demeaning within a group sharpens it. Short-term reversal itself dates to Jegadeesh (1990) and Lehmann (1990). Da et al. (2014) find that sorting on returns within industries nearly doubles the return of the standard reversal strategy on large stocks, because industry momentum otherwise masks the reversal; Hameed and Mian (2015) document that intra-industry reversals are larger, more stable, and present among large liquid stocks, and decompose the unconditional reversal into intra- and inter-industry parts; and Blitz et al. (2013) obtain a similar gain by reversing the residuals of a factor model. Our within-group demeaning is the same operation with the industry replaced by a statistical partition. These studies establish the value of conditioning reversal on economically or statistically related peers, but do not compare alternative grouping methods against both no-grouping and random-grouping controls; the two controls and the mean–volatility split of this paper supply that comparison. Section 6.2 then adds a market-wide reversal factor to the spanning regressions and finds that it absorbs more of the controls’ alpha than of the clusterizers’, far more at the rule-based component.

Clustering methods can also generate the partition-based residuals a statistical arbitrage strategy trades. Jin et al. (2023) first brought a signed spectral method (SPONGE, Cucuringu et al., 2019) to statistical arbitrage, on the correlation matrix of market residuals of US equities; Lee and Chang (2026) learn the clustering and the model that trades it jointly, end to end, inside a transformer encoder on cryptocurrency futures. Brutti et al. (2026) cluster return deviations with a noise-robust clusterizer, build the partition to be orthogonal to the sector taxonomy, and show over 2012–2024 that the reversion payoff survives residualization against standard benchmarks; they also trade a no-clustering benchmark under the same rule. Comparisons across these studies are few, and those that exist either vary the clusterizer and the trading model together or, as in Jin et al. (2023), compare variants within one family. We separate the two effects, ask whether clusterizers that disagree on the groups recover different information, and judge their ranking against the noise of the design’s own replications.

3 Data and Experimental Design

In this section we hold every element of a statistical arbitrage pipeline fixed except the clusterizer and the trading model, and we state what the resulting comparison estimates. A clustering-based statistical arbitrage strategy has two stages. A *clusterizer* divides the cross-section of N_t stocks into K groups: it maps the day’s panel of residual returns to a partition. A trading model then bets on relative value within each group. We trade every partition under both a fixed cross-sectional rule and a neural trading head. Throughout, *clusterizer* denotes the algorithm, *clustering* the operation it performs, and *partition* the grouping it produces. The two-stage design—extract a relative-value residual, then bet on it—follows the deep statistical arbitrage line of Guijarro-Ordóñez et al. (2025) and Epstein et al. (2025). Comparing partitions inside such a pipeline has precedents in Jin et al. (2023), within the spectral family, and in Brutti et al. (2026), whose no-clustering benchmark our rule-based results corroborate, though not its turnover reduction (Section 6.4).

Write the day- t partition $\mathcal{P}_t$ as a membership matrix

$$\mathbf{Z}_t \in \{0, 1\}^{N_t \times K}, \quad \mathbf{Z}_t \mathbf{1}_K = \mathbf{1}_{N_t}, \quad (1)$$

so that $[\mathbf{Z}_t]_{ik} = 1$ exactly when stock i belongs to group k , and the constraint records that every stock belongs to exactly one. The object we vary is $\mathbf{Z}_t$. It is formed from the panel of residual returns and re-estimated every trading day. Equivalently, $\mathbf{Z}_t$ is a factor loading matrix restricted to be one-hot: every stock loads on exactly one group, with unit weight. An unrestricted $\mathbf{\Lambda}_t \in \mathbb{R}^{N_t \times K}$ (K principal components, say) lets every stock load on every component. Both are rank- K descriptions of the same cross-section, which makes them comparable.

Let $\pi_t(\mathbf{Z})$ denote the payoff of the book the trading component produces from a partition $\mathbf{Z}$. We estimate the map

$$\mathbf{Z} \mapsto (\mu(\mathbf{Z}), \sigma(\mathbf{Z})), \quad \mu(\mathbf{Z}) = \mathbb{E}[\pi_t(\mathbf{Z})], \quad \sigma^2(\mathbf{Z}) = \text{Var}[\pi_t(\mathbf{Z})], \quad (2)$$

and the organizing claim of this article is that clustering moves σ more than it moves μ . Figure 1 draws the pipeline that produces π_t : every arm, one clusterizer or control traded at one component, shares every stage except the partition. Section 6 uses the loading-matrix equivalence of Equation (1) directly, comparing $\mathbf{Z}_t$ against principal components at matched rank K .

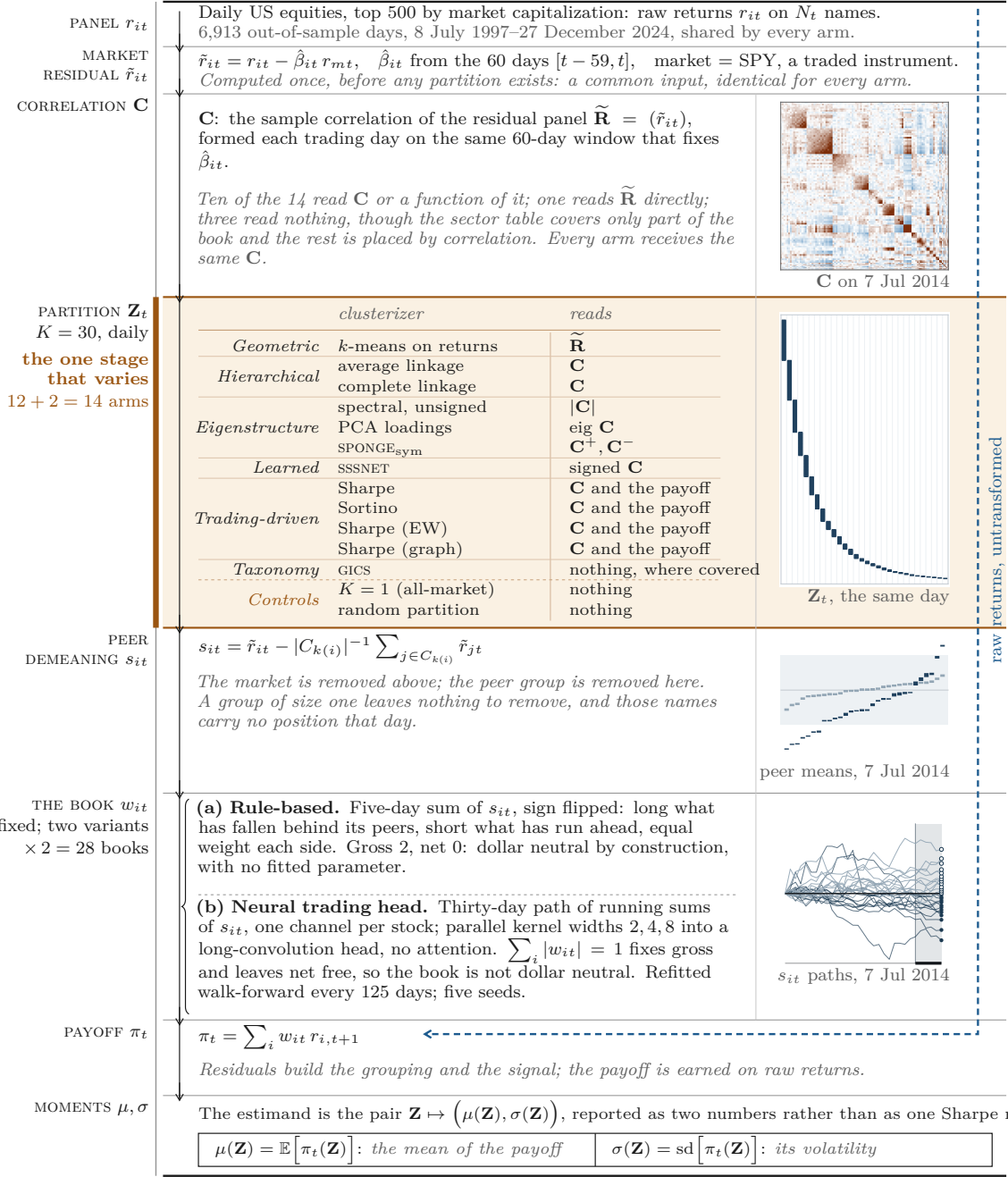


Figure 1: The strategy pipeline. This figure shows the stages of the pipeline as horizontal bands, one per stage. The arms differ only in the partition band, whose second column lists the input each clusterizer uses; every other band is common to all arms. Raw returns bypass the residual construction and re-enter at the payoff. The four insets are computed from the panel and from the k -means partition the runs used ($K = 30$, re-estimated daily) for 7 July 2014, a day with $N_t = 500$ names, all 30 groups occupied, the largest holding 78 names and six holding one, which carry no position. $\mathbf{C}$ is ordered by that day’s groups; $\mathbf{Z}_t$ is the membership matrix with columns sorted by group size. The paths are 26 of the 478 eligible names, 30 days of cumulative peer-demeaned residual return in percent, with the five days the rule-based component reads shaded.

3.1 The Panel

We build the panel from daily CRSP data obtained through WRDS, covering 1990 to 2024. The traded universe is the 500 names with the largest market capitalization as of the *previous* close, reselected daily. This restriction keeps the comparison in a large, liquid-equity universe while allowing membership to evolve using only information available at the previous close. The names must be common stock with a full residual history over the estimation window.¹ The return is CRSP’s close-to-close return adjusted for splits and dividends. Every result in this article is computed on the 6,913 out-of-sample trading days that all 14 arms (the 12 clusterizers and the two controls) share at both trading components, 8 July 1997 to 27 December 2024. The neural trading head is pre-trained on the first 1,000 days on which a book can be formed, and the walk-forward loop begins on the day after the last of them. The panel’s date labels run one trading day ahead of the Fama–French calendar; the offset is corrected before any regression on the factor file, and no statistic computed within the panel depends on it (Appendix A.7).

3.2 Residualization

To remove the market factor, we residualize every stock’s return against the S&P 500 exchange-traded fund (SPY) using a single-factor market model. Writing r_{it} for the raw return on stock i and r_{mt} for the market return,

$$\tilde{r}_{it} = r_{it} - \hat{\beta}_{it} r_{mt}, \quad \hat{\beta}_{it} = \frac{\widehat{\text{Cov}}_{[t-59, t]}(r_i, r_m)}{\widehat{\text{Var}}_{[t-59, t]}(r_m)}, \quad (3)$$

the estimation window running over the 60 trading days ending at t inclusive; the window length follows Avellaneda and Lee (2010). The ratio in Equation (3) is the least-squares slope from a regression of r_i on r_m with an intercept; the intercept is not subtracted, so $\tilde{r}_{it}$ is a market-hedged return, which we call a “residual” for brevity, rather than the regression residual itself. The position formed at the close of day t earns $r_{i,t+1}$, which lies outside every estimation window, so no position uses information about the return it earns; the convention is shared by all 14 arms.²

Equation (3) involves no $\mathbf{Z}_t$ and is computed once. The residual panel $\tilde{\mathbf{R}} = (\tilde{r}_{it})$ is therefore a common input: every clusterizer is the map $\tilde{\mathbf{R}} \mapsto \mathbf{Z}_t$ evaluated at the same argument. The residuals drive the clustering and the trading signal, but the book is settled in raw returns: π_t in Equation (2) accumulates r_{it} , not $\tilde{r}_{it}$.

3.3 The Clusterizers

We compare the 12 clusterizers, using a common 60-day estimation window wherever the partition is estimated from returns or correlations. Table 1 lists them by family, with what each clusterizer reads and how each is constructed.³ Each clusterizer receives the same $\tilde{\mathbf{R}}$, targets the same K , and passes its $\mathbf{Z}_t$ to the same trading stage. K is 30 throughout, the fixed- K benchmark of Jin et al. (2023); Appendix A.6 sweeps it. We do not tune any clusterizer’s settings to its trading results. Each runs at one pre-specified setting, because the question is how much a downstream Sharpe ratio can tell apart different partitions, not which setting of any one method is best. Appendix A.1 gives each clusterizer’s construction, and Appendix A.3 the partitions it produces.

Three features of the list matter later. First, the families differ in the information they use as well as in the algorithm, and the window experiment of Section 5.5 finds sensitivity to the length of the estimation window concentrated in the clusterizers that read individual eigenvectors of the correlation matrix. Second, the four trading-driven clusterizers optimize the partition against the Sharpe or Sortino ratio itself rather than against any other measure of cluster quality. Whether using the payoff helps is then an empirical

¹CRSP share codes 10, 11, and 12.

²Ending the window at $t - 1$ instead, with every partition held fixed and the trained component not rerun, moves each rule-based book’s Sharpe ratio by less than 0.04, well inside the dispersion of Section 5.1. The volatility reduction from clustering is unchanged to the first decimal and the alpha advantage over either control is marginally larger, so the convention retained here is conservative.

³The list omits density-based clustering (DBSCAN, OPTICS), standard in pairs selection since Sarmento and Horta (2020), because it returns a variable number of clusters and discards unclustered names, which a design matching K across arms cannot accommodate; Appendix A.2 finds the cost of discarding names inside the noise of a single refit. Modularity-based community detection also chooses its own K , and single-linkage clustering is omitted for chaining.

Table 1: The 12 clusterizers

Clusterizer	Reads	Window (days)	Construction
<i>Geometric</i>			
<i>k</i> -means on returns	$\tilde{\mathbf{R}}$ directly	60	<i>k</i> -means on the residual return vectors in Euclidean space; forms no correlation matrix
<i>Hierarchical</i> (Mantegna, 1999; Tola et al., 2008)			
Average linkage	$\mathbf{C}$	60	agglomerative clustering on $\mathbf{D} = 1 - \mathbf{C}$
Complete linkage	$\mathbf{C}$	60	the same, complete rather than average linkage
<i>Eigenstructure</i> (MacMahon and Garlaschelli, 2015; Cucuringu et al., 2019)			
Spectral (unsigned)	eigenvectors of $ \mathbf{C} $	60	spectral clustering on the unsigned correlation
PCA loadings	eigenvectors of $\mathbf{C}$	60	<i>k</i> -means on the 10 leading eigenvector coordinates
SPONGE	a signed-Laplacian pencil	60 ^a	the symmetric variant SPONGE _{sym} , on the positive and negative parts of $\mathbf{C}$; unlike the row above it uses the sign ^b
<i>Learned</i> (He et al., 2022)			
SSSNET	$\mathbf{C}$, as a signed graph	60	a signed-graph network trained on a clustering objective, warm-started from SPONGE ^c
<i>Trading-driven</i>			
Sharpe	$\mathbf{C}$ and the payoff	60	a soft assignment optimized against the realized Sharpe ratio of the book it produces, hardened by $\arg \max$
Sortino	$\mathbf{C}$ and the payoff	60	the same, downside deviation replacing the standard deviation
Sharpe (EW)	$\mathbf{C}$ and the payoff	60	the same objective exponentially weighted, decay 0.99 ^d
Sharpe (graph)	$\mathbf{C}$ and the payoff	60	the assignment comes from a signed-graph network, and the objective matches the trading signal’s 30-day horizon
<i>Taxonomy</i> (Bhojraj et al., 2003)			
GICS	nothing, on the names it covers	—	a static sector classification; names outside the table are placed by correlation ^e
<i>Controls</i>			
No clustering ($K = 1$)	nothing	—	one all-market group, so the trading rule demeans against the whole cross-section
Random partition	nothing	—	a uniform random partition redrawn each day, matched to the real partitions in K

Notes. This table lists the 12 clusterizers, the object each reads, its estimation window, and its construction. Every clusterizer receives the same $\tilde{\mathbf{R}}$ and targets the same K . Identifiers used in the runs are given in Appendix A.5.

^a Every clusterizer that forms a correlation matrix reads the same 60 trading days. SPONGE is estimated at 21 days in its original implementation; the arm carried here is its 60-day variant, so that a comparison across clusterizers is not a comparison of windows. The substitution does not disadvantage SPONGE (Appendix A.5). The three arms marked — form no correlation matrix at all.

^b Groups of fewer than two members are handled differently: SPONGE, Sharpe, and Sortino reassign such a member to its most correlated group, while the other nine clusterizers leave it and the trading stage drops it. For the three that reassign this is immaterial; under *k*-means on returns, the one clusterizer with a material singleton rate, the deletion changes the traded book, and Appendix A.2 measures its cost.

^c Run once to produce a fixed set of daily partitions used by all five seeds, like the other stochastic clusterizers (Appendix A.1).

^d Sharpe (EW) differs from Sharpe in more than the weighting, so the clean contrast is Sharpe against Sortino.

^e The table is a single snapshot covering roughly 54% to 73% of the universe by era, so it carries survivorship and coverage bias by construction. The names it does not cover, about 35% of the book on a given day, are assigned to the group they correlate with most strongly, so the arm that is actually traded is a taxonomy–correlation hybrid; the *Reads* column records the taxonomy alone.

question, and with our design it does not: one of the four finishes near the top of the ranking and three in the lower half (Section 5.1). Third, we include GICS to give the comparison a non-statistical reference point, and carry its coverage caveat (Table 1, note e) wherever it appears.

3.4 The Two Trading Components

We trade every partition twice: once by a fixed cross-sectional rule with no fitted parameter, and once by a neural trading head fitted on a Sharpe ratio loss. Both components begin by demeaning within the same partition: on each day we subtract from every stock’s residual the average residual of its own group,

$$s_{it} = \tilde{r}_{it} - \frac{1}{|C_{k(i)}|} \sum_{j \in C_{k(i)}} \tilde{r}_{jt}, \quad (4)$$

where $C_{k(i)}$ is the group holding stock i . Equation (4) is the second of two residualizations in the design: the first, in Equation (3), removes the market; this one removes the peer group.

The rule-based component sums Equation (4) over the past five days and holds every stock with a negative sum long and every stock with a positive sum short, equally weighted within each side:

$$w_{it} = \begin{cases} +1/n_t^+, & \text{if } \sum_{v=t-4}^t s_{iv} < 0, \\ -1/n_t^-, & \text{if } \sum_{v=t-4}^t s_{iv} > 0 \end{cases}, \quad (5)$$

with n_t^+ and n_t^- the number of stocks on each side. The rule is adapted from Jin et al. (2023), who trade the same within-cluster reversal on raw returns, with weights normalized within each cluster rather than across the book. The book is long what has recently fallen behind its peers, the mean-reversion bet. Writing $\sum_i |w_{it}|$ for its gross exposure and $\sum_i w_{it}$ for its net, Equation (5) gives 2 and 0 respectively: dollar neutrality holds by construction here, and will not hold at the trained component. The book is fully determined by the partition.

The trained component replaces that rule with a *neural trading head* in the design of Guijarro-Ordóñez et al. (2025): a convolutional extractor applied to the residual path, followed by a scoring network, with every parameter fitted to maximize the Sharpe ratio of the book itself. We replace their transformer with the long convolution of Fu et al. (2023), as Epstein et al. (2025) do. For each stock the head reads the 30-day path of running sums of Equation (4) and returns a weight through a fixed chain,

$$x_{it} \xrightarrow{\text{Conv}_{2,4,8}} u_{it} \xrightarrow{\text{LongConv}} h_{it} \xrightarrow{\text{FFN}} g_{it} \xrightarrow{c \tanh(\cdot/c)} f_{it} \xrightarrow{/\sum_j |f_{jt}|} w_{it},$$

a multi-scale convolution, a long convolution with one kernel weight per step, a two-layer scoring network, and a soft clamp followed by a rescaling to gross exposure one. The parameters maximize the in-sample Sharpe ratio of the payoff Equation (6), less three regularization penalties, and are refit as the sample rolls forward: pre-trained once on the first 1,000 days, then refit every 125 trading days on the trailing thousand. Every arm uses the same head, with constants fixed in advance. The only choices made within a run, early stopping and the learning-rate schedule, are confined to the trailing training window; the seeds of Section 3.6 vary this fitting and nothing else. Appendix A.4 gives the architecture, the objective, and the schedule. Beyond the fitted model, the head also reads a longer signal window, 30 days against the rule’s five. Section 4.4 separates the model from the window by rerunning the fixed rule on a 30-day signal.

The rescaling fixes gross exposure and leaves $\sum_i w_{it}$ free, so unlike Equation (5) it is not dollar neutral: mean absolute net exposure runs 0.06 to 0.18. Runs are screened for one-sided books, a day holding no long or no short position; no run reported in this article fails the screen. The head’s objective also divides by a standard deviation, so the trained component targets the second moment directly, while the rule-based component does not.

Either component turns weights into the payoff of Equation (2),

$$\pi_t = \sum_i w_{it} r_{i,t+1}, \quad (6)$$

and the return on the right is the *raw* return of Equation (3): residuals build the partition and the signal, but the payoff is earned on raw returns.

3.5 Two Controls

To measure what a partition is worth, we compare every result below against two control arms. The first sets $K = 1$, a single group, so the cluster-relative signal is measured against the whole cross-section rather than a peer group. This control answers whether clustering is worth performing at all. The second assigns each stock to one of K groups uniformly at random, independently across stocks and redrawn every trading day, reading nothing from $\hat{\mathbf{R}}$. The random control removes only the *information*: 30 groups, the same cluster-relative construction, the same trading model, and the same signal window all remain, so it separates the mechanics of grouping from the information in the groups. It differs from the placebo of Brutti et al. (2026), which permutes long and short assignments within clusters and so tests the signal rather than the partition.

3.6 Seeds and Conventions

We train each configuration five times, at seeds $\{1, 2, 3, 4, 5\}$. The seed varies only the initialization of the trading head, so dispersion across seeds is dispersion in the fitted model at a fixed clustering; every dispersion figure below, including the pooled benchmark of Section 5, is computed across a configuration’s own fits.

Nine of the 12 clusterizers are stochastic. Each has its daily partitions generated once and cached, and used by all five runs, so a training seed moves only the fit; SSSNET, which in its original form computes its partition at run time, is run once to produce such a fixed set (Appendix A.1). The partition itself is redrawn only in Section 5.2.

Throughout, the Sharpe ratio of a payoff series is $\sqrt{252}$ times its mean divided by its standard deviation, with no risk-free subtraction: the books are self-financing and long-short, so their return is already an excess return. Tables report either the mean of the per-seed Sharpe ratios or the Sharpe ratio of the seed-averaged payoff, as each note states. Averaging the payoffs first cancels part of the variation between fits, so the second is larger, by 0.16 to 0.22.

3.7 What the Design Can Resolve

We close the section by stating what the design can and cannot resolve. A comparison of trading strategies is a comparison of estimated moments, and the same sample estimates a volatility with $\sqrt{2}$ times the precision of a mean. For a return series of length T and volatility σ ,

$$\text{se}(\hat{\sigma}) = \frac{\sigma}{\sqrt{2T}}, \quad \text{se}(\hat{\mu}) = \frac{\sigma}{\sqrt{T}}. \quad (7)$$

A Sharpe ratio is therefore asymmetric in its own resolution before any data are seen: whatever a study of this kind can say about the denominator, it can say less firmly about the numerator. A comparison carries information only when the spread in a moment is large relative to the precision with which that moment is measured.

The 19 arms of Appendix A.5, the 14 above plus five 21-day window variants, trade the same names on the same days, so their payoffs correlate at $\bar{\rho} = 0.854$ on the seed-averaged books. The common component of the estimation error cancels between them, so the measurement error used below is the idiosyncratic part.⁴

The two moments fall on opposite sides of what the design can resolve. We split the sample at 2001, 2009, and 2016 and measure across the 19 arms of Appendix A.5: differences in volatility run between 13 and 21 times their own measurement error in every period, with no trend across the four periods. Differences in mean return run from 2.2 times theirs in the first period to below one in the last two periods (0.98 and 0.63).

⁴Estimates of a variance inherit the square of the correlation between the underlying series, so the shared component of the volatility error is $\bar{\rho}^2 = 0.729$ and not 0.854; carrying $\bar{\rho}$ through would raise the volatility ratios reported here by about a third. The correlation is measured on seed-averaged payoffs; on single runs it runs from 0.757 to 0.799.

Table 2: Book volatility and factor alpha: 12 clusterizers against the two controls, within each trading component

		Level (%/yr)			Change from clustering	
Trading component		12 clusterizers	$K = 1$	Random	against $K = 1$	against Random
Volatility	Rule-based	4.77–6.18	9.30	8.83	–41.9%	–38.7%
	Trained	2.88–3.80	5.07	4.89	–35.4%	–33.0%
Alpha	Rule-based	8.05–8.81	7.27	7.43	+15.6%	+13.1%
	Trained	5.89–7.02	5.44	5.33	+20.0%	+22.5%

Notes. This table reports the annualized volatility and the annualized alpha of the rule-based and trained books of the 12 clusterizers and the two controls, and the change from each control’s level to the clusterizers’. All 14 arms are measured on the 6,913-day window of Section 3.1 (July 1997 to December 2024), with the common 60-day estimation window wherever the partition is estimated from returns or correlations. Rule-based books carry gross exposure 2 and trained books gross 1, and the two components also differ in signal window, so only within-component comparisons are meaningful. Alphas are against the Fama–French five factors (Fama and French, 2015) plus momentum (Carhart, 1997), FF5+MOM below, with Newey–West(21) standard errors; each payoff is paired with the factor return of the calendar day on which it was realized. The two trading components label their books differently: the rule-based series by decision day, the trained series by a date label that runs one trading day ahead of the factor file’s, so the join is same-day for the rule-based series and lagged one day for the trained series. Trained volatilities are computed per head seed and averaged over the five seeds; trained alphas are on the payoff averaged over the five head seeds. Change is the ratio of the 12 clusterizers’ mean level to the control’s, minus one, computed on unrounded values; those means are 5.41 and 3.27 for volatility and 8.41 and 6.52 for alpha. Averaging the trained payoff series over seeds before measuring volatility gives 2.66–3.48 for the clusterizers against 4.44 and 4.18 for the controls, changes of –32.3% and –28.1%.

4 Decomposing Portfolio Performance

In this section, we show that a real partition has two benefits: it lowers the volatility of the payoff, and it raises its alpha. We measure both benefits within each trading component. We then measure how much comovement each partition removes and show that a random partition removes only the amount that any grouping into K sets removes by construction (Section 4.2). We decompose the Sharpe gap between grouped and control books into a mean part and a volatility part (Section 4.3), and finally turn the comparison around and measure what the trading head adds to a partition held fixed (Section 4.4).

4.1 Volatility and Alpha

Clustering cuts book volatility by 28% to 42% in every combination of trading component, control, and convention (Table 2). Under both components the least volatile control book sits far above the most volatile of the 12 clusterizers’ books. The neural trading head narrows the gap against both controls without closing either, from 41.9% to 35.4% against the all-market control and from 38.7% to 33.0% against the random control.

Every clusterizer produces a reduction in volatility. The 12 trained books span 2.88% (k -means on returns) to 3.80% (spectral clustering), a band of 0.9 points, while either control sits about a point above the band’s top. Across the 12 trained books the highest volatility is 32% above the lowest, the highest alpha only 19% above (Table 2), so their Sharpe ratios differ mainly through volatility. Figure 2 puts the two moments on one plot: the 12 clusterizers form a narrow band in alpha and a wide one in volatility, and both controls fall outside both bands, below in alpha and above in volatility. The rule-based books, at twice the gross exposure, take the same shape. The volatility reduction is a property of clustering, not of a well-chosen partition, and Section 5 returns to how little the choice of clusterizer matters downstream.

Clustering also raises alpha, and unlike the volatility benefit the size of the gain depends on which trading component reads the signal. At the rule-based component both controls sit below every one of the 12 clusterizers (7.27% per year for no clustering and 7.43% for the random partition, against 8.05% to 8.81% for the clusterizers). The same fixed rule therefore earns 15.6% more alpha on the twice-residualized signal of

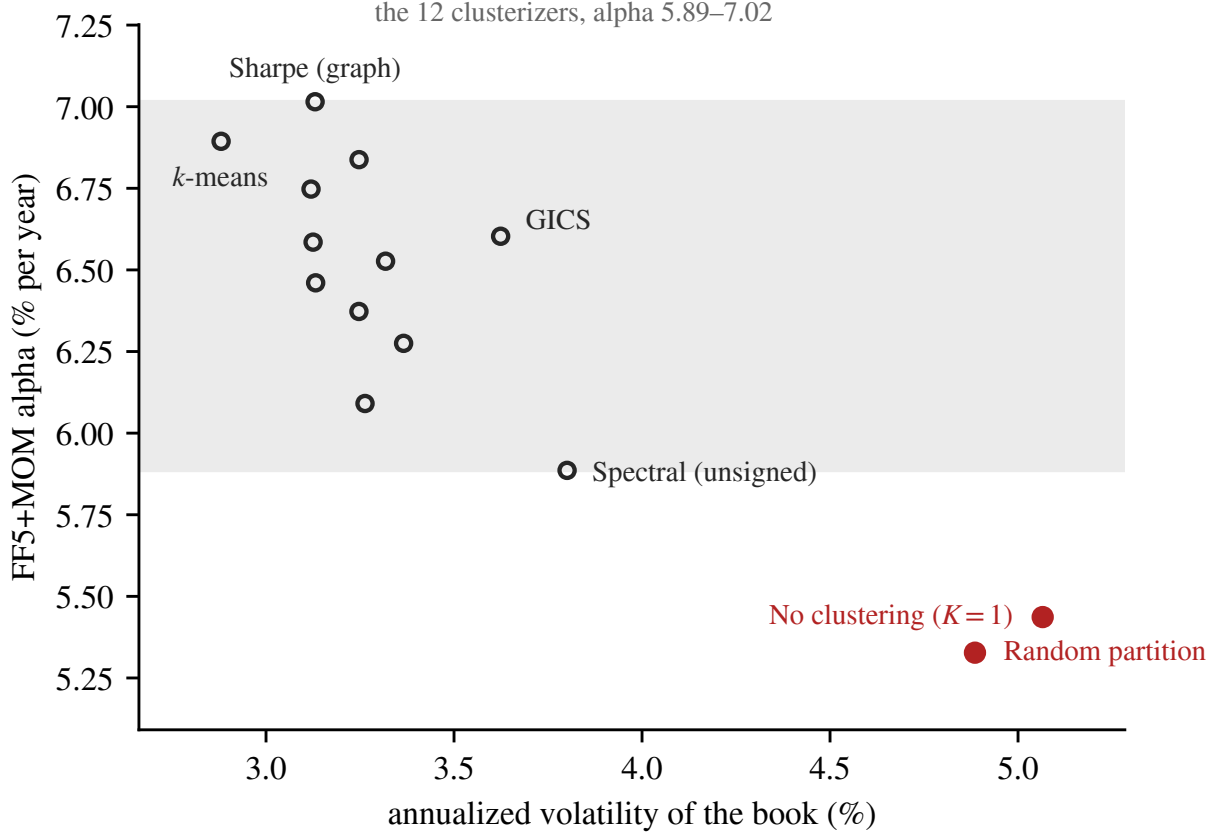


Figure 2: Alpha against volatility at the trained component. This figure plots each arm’s annualized alpha against the annualized volatility of its book, both at the trained component over the 6,913-day sample (July 1997 to December 2024). Alphas are against FF5+MOM with Newey–West(21) standard errors on the seed-averaged payoff, and volatilities are computed per head seed and averaged over the five seeds, the conventions of Table 2. The shaded band spans the 12 clusterizers’ alphas. Four clusterizers are named: *k*-means on returns at the lowest volatility, the graph Sharpe clusterizer at the highest alpha, GICS at the second-highest volatility with a mid-band alpha, and spectral clustering at the highest volatility and the lowest alpha; the eight unnamed circles are the remaining clusterizers, whose coordinates the replication package lists. The two controls are red.

Equation (4) than on the all-market control’s signal, and 13.1% more than on the random control’s. Because the rule-based component has no fitted parameters and the arms differ only in the signal they receive, the extra alpha can come only from the signal the partition supplies.

The neural trading head widens the alpha advantage over both controls, far more over the random control than over the all-market control (against the six factors of Table 2; with a short-term reversal factor added, the rule-based advantage exceeds the trained one, Section 6.2). Against the all-market control the advantage rises from 15.6% to 20.0%; against the random control it rises from 13.1% to 22.5%. The same asymmetry appears arm by arm. Both advantages are positive at every one of the 12 clusterizers. Formed seed by seed, so that the dispersion across a clusterizer’s five head seeds is preserved, the advantage over the all-market control exceeds twice that dispersion at only four of the 12, while the advantage over the random control does so at seven. The count is the same under either dispersion, the clusterizer’s own across its five seeds or the figure pooled over the 12, though not the same four. A wrong partition therefore costs more under the neural trading head than under the fixed rule (Section 4.4). The random control earns less alpha than $K = 1$ under the head (5.33 against 5.44), a gap that plausibly reflects what the two signals contain. Demeaning within a random group subtracts the average of roughly 16 unrelated stocks, so the random signal is the $K = 1$ signal perturbed by a component that carries little systematic peer information, and the trained component reads a

corrupted copy of the signal it reads under $K = 1$.

4.2 Variance Removed by Group Demeaning

Group demeaning removes between a tenth and a quarter of the cross-sectional variance under a real partition, and only the mechanical floor under a random one. We measure directly how much of the cross-sectional variance of the market residuals the group demeaning of Equation (4) removes. For each arm we take each day’s market residuals (Section 3.2), subtract each stock’s own group mean, and record the share of variance that disappears. For the 12 clusterizers the share runs from 10.8% for the sector taxonomy to 26.4% for k -means on returns, and the other 10 lie within about two percentage points of one another (17.0% to 19.1%). Ten of the 12, though they disagree widely on the partition (Section 5.2), remove nearly the same share.

For the $K = 1$ control the same quantity is zero by construction: its one group’s mean *is* the cross-sectional mean, around which the variance is already measured. The $K = 1$ signal is therefore residualized only once, against the market, while a signal after clustering is residualized twice.

The random control measures the part of the removed variance that has nothing to do with structure. Subtracting a group mean removes some variance even when the groups are meaningless, and the amount can be computed: with K random groups over N_t stocks it is $(K - 1)/(N_t - 1)$ in expectation, 5.88% at $K = 30$ on the 494.3 names the rule-based book trades on an average day. The random control measures 5.93%. Everything it removes is therefore this mechanical part, and its book remains one of the two most volatile of the 14 books. Removing variance from the signal helps only when the variance removed is real comovement.

Less comovement in the signal produces less volatility in the book. With the common movement removed, the positions the head builds on the signal comove less, and in the spanning regressions of Section 6.2 the 12 clusterizers’ trained books carry 42% less market beta than the all-market control and 35% less than the random control. Across the three cases (no clustering, random, real), removal share and book volatility rank in the same order at both trading components (Section 4.1). A partition degraded step by step toward a random one traces the same relation continuously: the removal share falls toward the floor and book volatility rises at every step (Appendix A.8).

Balance within groups contributes little to the volatility reduction. The demeaned signal sums to zero within each group, and the book could inherit that balance, but a typical group’s longs and shorts are out of balance by 16% to 23% of the group’s total position, about three quarters of the level that random signs on the same weights would produce. The one exception is GICS, whose imbalance exceeds its own random level.

4.3 Decomposing the Sharpe Gap

Most of a partition’s Sharpe advantage over either control comes through volatility. Because the log Sharpe ratio is the log mean minus the log volatility, the log-Sharpe gap between the clusterizers’ books and a control book splits exactly into a mean part and a volatility part,

$$\log \text{SR}_g - \log \text{SR}_c = (\log \mu_g - \log \mu_c) - (\log \sigma_g - \log \sigma_c). \quad (8)$$

Across the four cells of Table 3, the volatility part is 82.0% of the gap over the $K = 1$ control at the trained component, and between 75% and 83% overall, largest at the rule-based component against the random control. Arm by arm, in the trained cells the share is lowest for the graph Sharpe clusterizer against the $K = 1$ control and for GICS, the coarsest partition, against the random control, down to 68%; no other clusterizer goes below 69% in any cell. In this split, the return the controls earn from factor exposure counts as mean; measured on calendar-aligned alphas instead, the mean part rises to 29.4% against $K = 1$ and 33.6% against random in the two trained cells. Under either measure, the volatility reduction accounts for most of a partition’s Sharpe advantage.

4.4 What the Trading Head Adds

Sections 4.1 to 4.3 held the trading component fixed and varied the partition. We now hold the partition fixed and ask what the neural trading head adds to it. At matched gross exposure it raises alpha at every one of the 14 arms and raises volatility at every arm by proportionally less. The ranking of the 12 clusterizers

Table 3: Mean and volatility parts of the log-Sharpe gap to each control

Trading component	Control	Gap	Mean part	Volatility part	Volatility share
Rule-based	$K = 1$	0.667	0.124	0.543	81.4%
Rule-based	Random	0.590	0.100	0.490	83.0%
Trained	$K = 1$	0.533	0.096	0.437	82.0%
Trained	Random	0.532	0.131	0.401	75.4%

Notes. This table reports the log-Sharpe gap between the 12 clusterizers’ mean book and each control and its split into a mean part and a volatility part, by trading component. All arms use the common 60-day window, on the 6,913-day window of Section 3.1 (July 1997 to December 2024). The clusterizers’ mean book takes the mean over the 12 clusterizers of the annualized mean and of the annualized volatility; each control uses its own. Gap is the left side of Equation (8); the mean part is $\log \mu_g - \log \mu_c$, the volatility part $\log \sigma_c - \log \sigma_g$, and the share is the volatility part over the gap. In the trained cells μ and σ are computed per head seed and averaged over the five seeds; averaging the payoff series over seeds first gives shares of 80.2% and 71.6% in the two trained cells. The rule-based component has no training seed. Arm by arm, in the trained cells the share is lowest for the graph Sharpe clusterizer against the $K = 1$ control and for GICS against the random control, down to 67.8%. Replacing the means with calendar-aligned alphas in the trained cells gives a mean part of 0.182 against $K = 1$ and 0.203 against random, a share of 29.4% and 33.6% for the mean.

Table 4: Alpha and volatility at the two trading components, and the change between them

	Alpha (%/yr)			Volatility (%/yr)		
	Rule-based	Trained	Gain	Rule-based	Trained	Ratio
12 clusterizers	4.03–4.40	5.89–7.02	+1.56 to +2.94	2.38–3.09	2.88–3.80	1.17–1.24
$K = 1$	3.63	5.44	+1.80	4.65	5.07	1.09
Random	3.72	5.33	+1.61	4.41	4.89	1.11

Notes. This table reports each arm’s alpha and volatility at the two trading components and the change from the rule-based to the trained component. All 14 arms at the common 60-day window, on the 6,913-day window of Section 3.1 (July 1997 to December 2024); alphas are against FF5+MOM with Newey–West(21) standard errors, each component joined to the factor file on its own realization calendar (Appendix A.7); rule-based cells are divided by two to match the trained component’s gross exposure of one. Gain is the trained alpha minus the rule-based alpha, in percentage points; Ratio is the trained volatility over the rule-based volatility. Both are computed arm by arm on unrounded values, as are the divided-by-two cells; the 12-clusterizer row is therefore the range of the per-arm values, not the difference or ratio of the printed ranges. Volatility columns are computed per head seed and averaged and alpha columns are on the seed-averaged payoff, both as in Table 2; the rule-based cells here are the rule-based entries of that table halved.

is largely the same under either component; the gap between the two controls also closes, and that closing shows what training does to a partition that carries no information.

The two trading components differ in two ways, the model and the signal window (5 against 30 days, Section 3.4), so any difference between them bundles the two changes. Rerunning the rule-based component with a 30-day window widens the alpha advantage over the controls to +38.7% and +35.8%, and the head still adds alpha at every arm over that rerun, so the longer signal window alone does not account for the head’s alpha improvement. That rerun also sets the base against which the head’s advantage is measured. At the head’s gross exposure, the 30-day window moves the fixed rule’s gap to either control by at most 0.04 alpha points but lowers the alphas themselves to 1.4% to 2.3% a year across the 14 arms, with neither control’s alpha distinguishable from zero ($t = 1.6$ and 1.7). The larger relative advantage therefore comes from the smaller base. In alpha points the head’s gap to either control is roughly twice the fixed rule’s at either signal window, 1.9 to 2.5 times.

Table 4 reports each arm against itself across the two trading components, with the rule-based book divided by two to match the head’s gross exposure. The head adds alpha at every arm, controls included, and raises volatility at every arm, by 9% to 24%, even though its objective divides by the standard deviation.

The rule-based ranking largely remains the same under the neural trading head (Figure 3). Across the

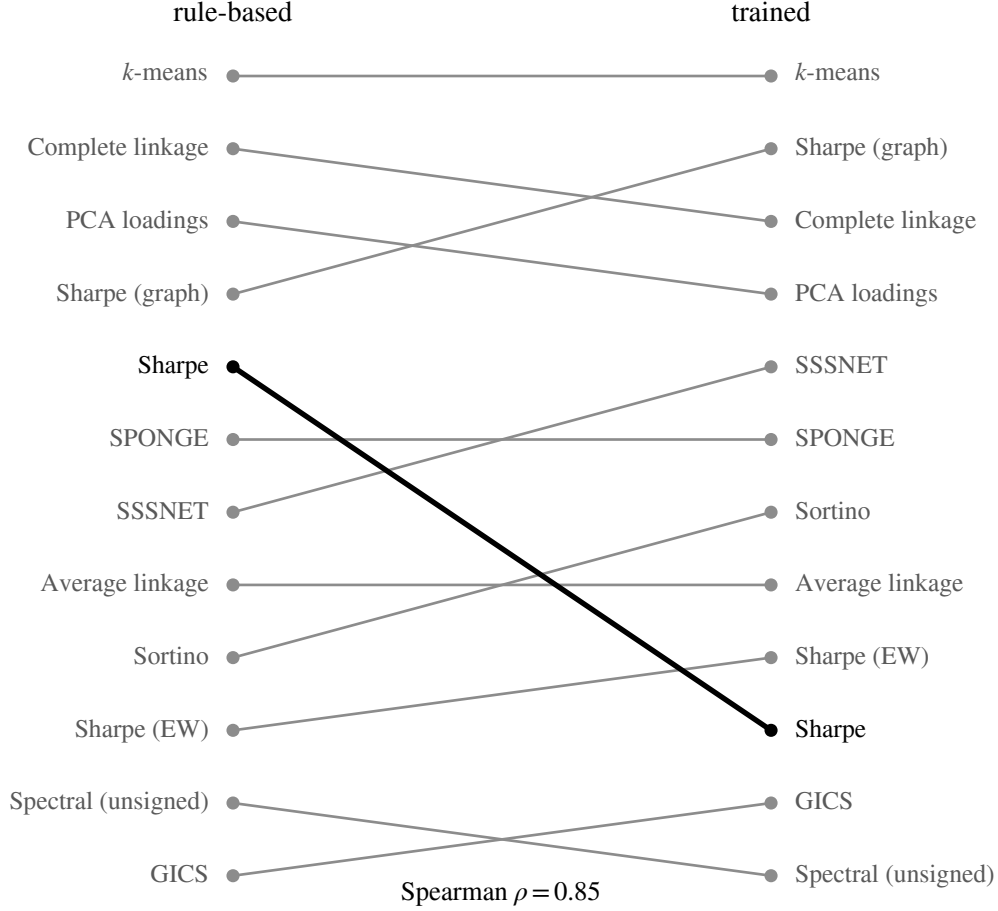


Figure 3: Sharpe-ratio ranks at the two trading components. This figure plots the 12 clusterizers ranked by Sharpe ratio at each trading component, on the matched 60-day estimation window over the 6,913-day sample; the trained column is the mean of the five per-seed Sharpe ratios. A clusterizer that changes rank by three places or more is drawn in black. The Spearman rank correlation between the columns is +0.85.

12 clusterizers the Spearman rank correlation between the two components' rankings is +0.85 ($p = 0.0004$). Three clusterizers keep their rank, the largest move is five places, and both ends of the ranking are the same at both components: *k*-means leads, and spectral clustering and GICS are the two lowest.

The random control's edge over the $K = 1$ control does not survive training. At the rule-based component, where the book is a deterministic function of the partition, the random control earns a higher Sharpe ratio than the $K = 1$ control, 0.89 against 0.83. At the trained component the two are level on the mean-of-per-seed convention, 1.28 against 1.28, and on the seed-averaged book the random control's edge is only 0.04, 1.50 against 1.46, so under the head they are indistinguishable. Paired by seed, the difference is 0.00 with a standard deviation of 0.08 across five fits ($t = 0.00$), and its sign changes from fit to fit. Training therefore makes a wrong partition more expensive. One explanation, consistent with Section 4.1, is that the fixed rule bets small, equal amounts everywhere, so the noise a random partition adds largely cancels, whereas a trained model fits whatever structure the labels imply, even when there is none or it is wrong.

5 Comparing the Clusterizers

This section compares the 12 clusterizers against the variation the design itself produces when a training run or a partition draw is repeated. Their partitions disagree, yet the books they produce move together, and

their Sharpe ratios separate the top of the ranking from the bottom without separating neighbors in the middle. Section 5.1 measures how far the gap between two clusterizers moves when only the training run changes, and uses that dispersion as the benchmark for the ranking. Section 5.2 redraws the seed behind each partition and finds that the leader of the ranking depends on the draw. Section 5.3 runs the formal tests. Section 5.4 follows the differences between clusterizers at three points along the pipeline, from the partition to the payoff, and draws the consequence for choosing among them. Section 5.5 turns to the two inputs that do move the results, the estimation window and the number of groups.

5.1 Variation across Training Runs

Most adjacent gaps in the Sharpe ranking are smaller than the run-to-run variation in the gap itself. A ranking compares two arms, so the dispersion a gap must be judged against is the dispersion of the difference between them, not of either arm’s level. The two are not interchangeable here, and the reason can be measured. In the 12×5 matrix of per-seed Sharpe ratios, 27% of the variance is a shift shared by every arm: a seed that trains one configuration well trains all 12 well. The five seed means span 0.41, half the 0.83 between the best and the worst clusterizer. A shared shift moves every arm together and can reorder nothing. Only the interaction between arm and seed can, and it carries 8% of the variance.

Pooled across the 11 adjacent pairs, the standard deviation of a gap across its five training runs is 0.103, on 44 degrees of freedom, with a 95% interval of [0.085, 0.130]. Measured on the levels instead, the dispersion is 0.168 on 48 degrees of freedom, a fifth of the span between the best and the worst clusterizer. We report both and judge the gaps against the paired figure, which is the error term a comparison across a common set of seeds calls for.⁵

The 12 mean-of-per-seed Sharpe ratios span 1.742 to 2.567, and nine of the 11 adjacent gaps are smaller than 0.103; the count is eight at the interval’s lower end and nine at its upper. Both exceptions sit at the ends of the ranking. *k*-means on returns leads the second clusterizer by 0.139, and GICS leads spectral clustering at the bottom by 0.245, the only gap that also exceeds the dispersion of the levels. The nine gaps between them run from 0.001 to 0.090.

Even the separation at the top does not fix the leader. *k*-means on returns is first at four of the five seeds and complete linkage at the fifth. Only the bottom of the ranking is fixed: spectral clustering is last at every seed. The ordering is therefore not expected to replicate between neighbors (Figure 4).

5.2 Variation across Partition Draws

The training-seed dispersion of Section 5.1 leaves out the variation in the partition itself. Every stochastic clusterizer’s daily partitions were generated once and cached, and all five training runs use the same partitions (Section 3.6). Redrawing the partition measures how much that draw moves the results. This subsection is reported on a different basis from the rest of the paper: five partition draws, each traded by a single training run at head seed 42, so that the partition axis is isolated. The five training seeds of Section 3.6 do not enter here.

For the nine stochastic clusterizers, each redrawn five times (partition seeds 1 to 5; the cached partition of the rest of the paper is a separate draw), the adjusted Rand index (ARI; Hubert and Arabie, 1985) between two draws of the same clusterizer averages 0.30, against 0.19 between two different clusterizers. Expressed as disagreement, one minus the index, a redraw of the seed produces 87% of the disagreement that a change of clusterizer produces. The average hides three levels that do not overlap. The four trading-driven clusterizers agree with their own redraw at 0.21 to 0.28, the four other stochastic clusterizers at 0.32 to 0.45, and SSSNET at 0.18, so stability against the seed is a property of the family.

Redrawing the partition moves the Sharpe ratios as well; the largest range grows from 0.21 under the fixed rule to 0.38 under the neural trading head, though not every clusterizer moves more. At the rule-based component, which has no training noise, the draw already moves the leader (Figure 5a): spectral clustering is last under all five seeds, but *k*-means on returns is first under only three of them, Sharpe (graph) and PCA loadings under one each. The largest range across the five draws is 0.21, for *k*-means on returns, 36% of the

⁵This is a different safeguard from the data-snooping corrections of White (2000), Hansen et al. (2011), and Bailey and López de Prado (2014), which take each measured statistic as given and protect a comparison against selection across many tests. The dispersion measured here is the statistic’s own run-to-run noise, so the two safeguards are complements.

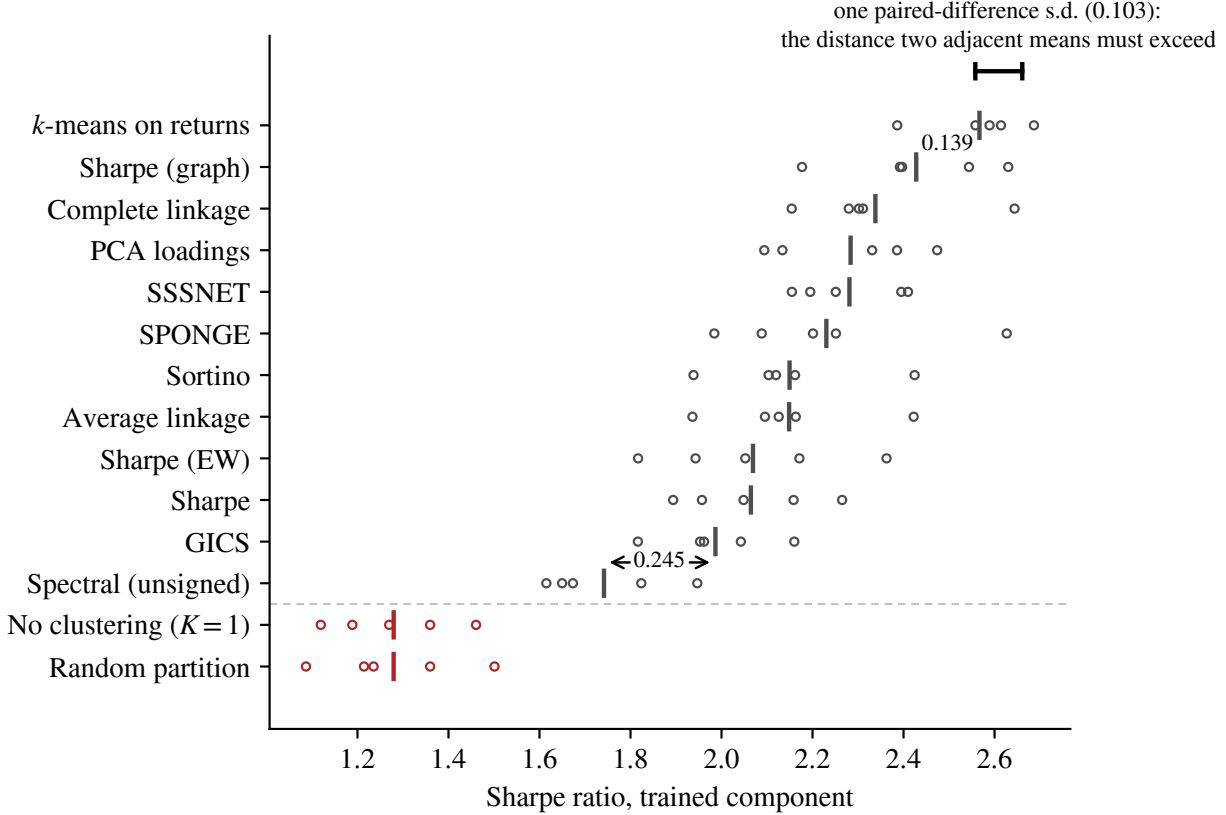


Figure 4: Per-seed Sharpe ratios against the dispersion of a gap. This figure plots the per-seed Sharpe ratios of the 14 arms at the trained component over the 6,913-day window. Each row is one arm: circles are its five per-seed Sharpe ratios, the tick its mean. The bar at the top right is not an interval around a center; its *length* is one pooled paired-difference standard deviation, 0.103, the distance two adjacent means must exceed, and it is meant to be laid against the horizontal distance between two neighboring ticks. Nine of the 11 gaps between adjacent clusterizer means fall inside it; the two that do not are marked, at the top and the bottom of the ranking. Measured on the levels instead the dispersion is 0.168, and by that wider figure 10 of the 11 gaps fall inside. Both controls sit below every clusterizer run.

0.58 between the best and the worst of the 45 rule-based books (nine clusterizers at five partition seeds); the random control, whose labels are redrawn daily under any seed, moves 0.08.

Under the neural trading head the leader still depends on the draw and the movement is larger. When the head is rerun with its own seed pinned on the same redrawn partitions, *k*-means is first at four of the five draws and SPONGE at the other. The largest range across the five draws grows by three quarters, to 0.38 for SPONGE, and the random control moves 0.39 (Figure 5b), as much as it moves across the five head seeds (0.41, Figure 4). The trained component, which might be expected to average the draw away, instead enlarges it, and for a partition that carries no information it converts the draw into as much variation as a refit does. SSSNET, redrawn like the others, moves by 0.15 at the rule-based component and by 0.13 under the head, the fifth-largest and the smallest ranges among the nine. A ranking computed from one partition draw per clusterizer therefore carries a source of variation that rerunning the training leaves in place and enlarges.

5.3 Formal Inference

Formal tests reach the same conclusion, that the ranking separates its top from its bottom but not its neighbors. The comparisons above use no hypothesis test; Table 5 reports tests on the 12 clusterizers' seed-averaged books at the matched 60-day window, over the 6,913 days all arms share. A model confidence set at the 90% level keeps six of the 12 clusterizers, the six with the highest Sharpe ratios, so it divides the

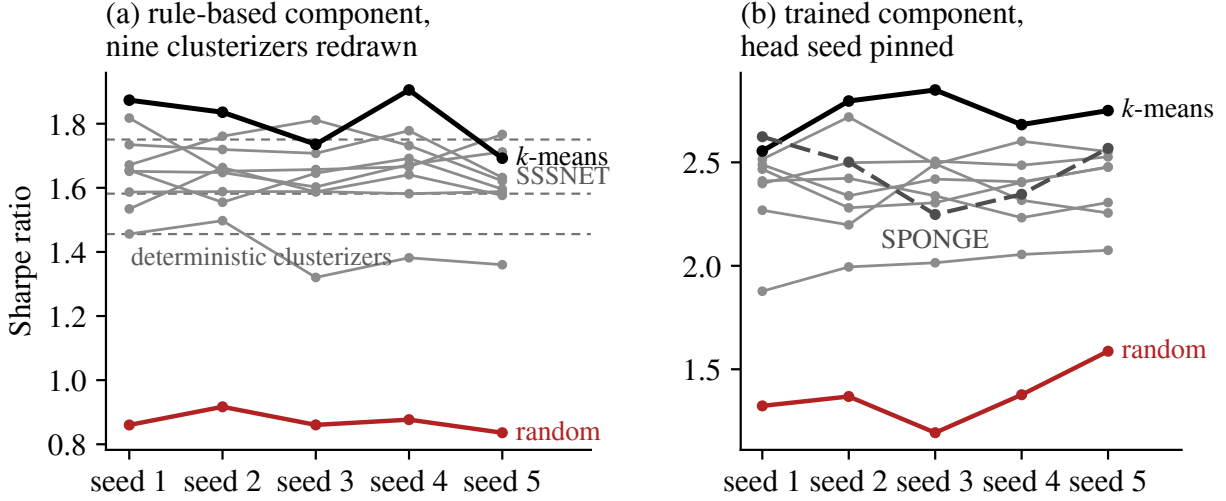


Figure 5: Redrawing the partition seed at each trading component. This figure plots the Sharpe ratio of each stochastic clusterizer under five partition seeds. The labels on the horizontal axis name those partition seeds, not the training seeds of Section 3.6, which are held fixed here. (a) The rule-based component is deterministic, so every movement is the partition draw itself. Spectral clustering is last under all five seeds; *k-means* (black) is first under three of them, and the largest range is 0.21. The panel tracks the nine stochastic clusterizers; the three deterministic ones, with no seed to redraw, are the dashed horizontals. (b) The trained component, with the head seed pinned at 42 and the same nine clusterizers. *k-means* (black) is first at seeds 2 to 5, SPONGE (dashed) at seed 1. The largest range across the five draws is larger here than at the rule-based component, and so is the random control's. The random control is red in both panels.

ranking in the middle rather than removing only the weakest arms; the two elimination statistics differ, T_R keeping two. The set holds the partition draw of Section 5.2 fixed and could only widen with it. Pairwise Ledoit–Wolf tests find 21 of 66 differences that survive a Holm correction (Holm, 1979), none of them between neighbors in the ranking. The adjacent pair closest to significance, the sector taxonomy against spectral clustering at the bottom of the ranking, has a raw bootstrap p -value of 0.004 that the correction raises to 0.15. The Sortino–Sharpe pair, the one clean objective contrast in the design, differs by 0.13 with a standard error of 0.09.

Part of the spread is the annualization itself. Daily Sharpe ratios are scaled by $\sqrt{252}$, which assumes away serial dependence. On non-overlapping 21-day returns, 329 blocks in place of 6,913 days, the spread across the 12 seed-averaged books falls from 0.88 to 0.40, the 90% set keeps 11 of the 12 under $T_{\max}$ and nine under T_R , and one of the 66 pairs survives Holm. The separation between clusterizers is visible day by day and modest at monthly frequency.

The gap to the controls survives at both horizons. Every clusterizer earns a higher Sharpe ratio than both controls, the confidence set discards both, and the pairwise gap is significant for all 12 against each control daily. At 21 days five separate from $K = 1$ and five from the random control, a loss of power rather than a change of sign.

5.4 Between-Clusterizer Differences and the Choice among Clusterizers

Two different clusterizers' books correlate at 0.61 on average in their daily position vectors. Whether 0.61 is close or far requires a reference point, and the design supplies an internal one. Every configuration is fitted five times (Section 3.6), so we can measure how much one clusterizer differs from itself when it is refitted at a fixed partition. We run that comparison at three points along the pipeline (Table 6): the partition, the positions it induces, and the payoff those positions earn. At each point the comparison reports how much two different clusterizers resemble each other, how much one clusterizer resembles itself under a redraw of its own randomness, and the ratio of the two.

Table 5: Statistical tests on the 12 clusterizers

Procedure	Question	Result
Model confidence set, 90%	which clusterizers could be best?	keeps 6 of 12 under $T_{\max}$ and 2 under T_R
Ledoit–Wolf, all pairs	which differences are real?	21 of 66 past Holm; none adjacent
Sortino vs. Sharpe	does the objective matter?	+0.1260 (s.e. 0.0919)
Non-overlapping 21-day returns	does $\sqrt{252}$ inflate the spread?	spread 0.8758 $\rightarrow$ 0.4044; 90% set keeps 11 of 12 under $T_{\max}$, 9 under T_R ; 1 of 66 past Holm
Every clusterizer vs. both controls	do all clusterizers beat both controls?	yes: 12 of 12 above both, at both horizons

Notes. This table reports the outcome of each test on the 12 clusterizers’ seed-averaged payoffs; $T_{\max}$ and T_R are the two elimination statistics of the model confidence set. All five rows are computed on one basis: the 12 clusterizers at the matched 60-day window, over the 6,913 trading days every arm shares (Section 3.1), on seed-averaged payoffs, so the spreads are not comparable with the paired dispersion of Section 5.1. Model confidence sets follow Hansen et al. (2011); pairwise tests follow Ledoit and Wolf (2008), on a circular block bootstrap (block 21 trading days, 5,000 draws, studentized bootstrap p -values). “Past Holm” counts the pairs whose difference survives the Holm step-down correction (Holm, 1979) over the 66 pairs at the 5% family-wise level.

Table 6: Similarity between clusterizers and to their own refits at three stages of the pipeline

	Across clusterizers	Clusterizer against itself	Ratio
Partition (ARI)	0.193	0.296	0.65
Positions, trained component	0.612	0.704	0.87
Daily payoff, trained component	0.815	0.805	1.01

Notes. This table reports how similar two clusterizers are at three stages of the pipeline, how similar a clusterizer is to its own refit, and the ratio of the two. ARI is the adjusted Rand index. In the partition row the across-clusterizer cell covers all 66 pairs of the 12, while the against-itself cell covers the nine stochastic clusterizers, whose partitions were redrawn, on the refit days inside the same window; the position and payoff rows cover all 12 in both columns. The “against itself” column is likewise not the same object in every row: in the first row it is the same algorithm at a redrawn *partition* seed, in the last two the same partition at a redrawn *head* seed. Adjusted Rand indices are computed day by day on the names two partitions share; every pair overlaps on at least 95% of the names except those involving the sector taxonomy, whose snapshot covers about two thirds of the traded universe. Against either control the index is 0.000 over all 24 pairs, which is why the controls are not shown. Positions are the daily weight vectors, correlated across the union of the two books’ names with absent names entered at zero, on all 6,913 common days and averaged over the five seeds. Payoff correlations are per seed and pair arms at the same seed, so the seed is not itself a source of difference; seed-averaging the payoffs first raises the across-clusterizer figure to 0.886 and is not comparable to the row shown.

In the partition itself the ratio of the two indices is 0.65 (0.193 against 0.296), so two clusterizers resemble each other about two thirds as much as a clusterizer resembles its own redraw. Expressed as disagreement, one minus each index, the same two numbers give the result of Section 5.2. In the positions the partition induces the ratio rises to 0.87, and in the payoffs those positions earn it reaches 1.01. Two different clusterizers produce payoffs that correlate as strongly as one clusterizer does with itself refitted.

Figure 6 shows that the resemblance is nearly uniform across pairs. No pair of clusterizers stands out and none falls away: the 12 clusterizers’ pairwise position correlations run from 0.581 to 0.657, and every one of them lies between 81% and 96% of the diagonal of its own row. Sorting the clusterizers by Sharpe ratio produces no block structure, so the arms at the bottom of the ranking hold much the same positions as the leaders, at higher volatility (Section 4.1). The small differences that remain follow the families rather than the Sharpe ranking. The two hierarchical clusterizers are the most similar pair among the 12 (0.657 against an average of 0.61 over all pairs) and the eigenstructure clusterizers are mildly cohesive (0.635). The four trading-driven clusterizers, which share an objective but not a representation, are as similar to each other as to the other clusterizers (0.604 within, 0.604 against the others).

The controls give a reference for how low the correlation falls between books that differ in substance.

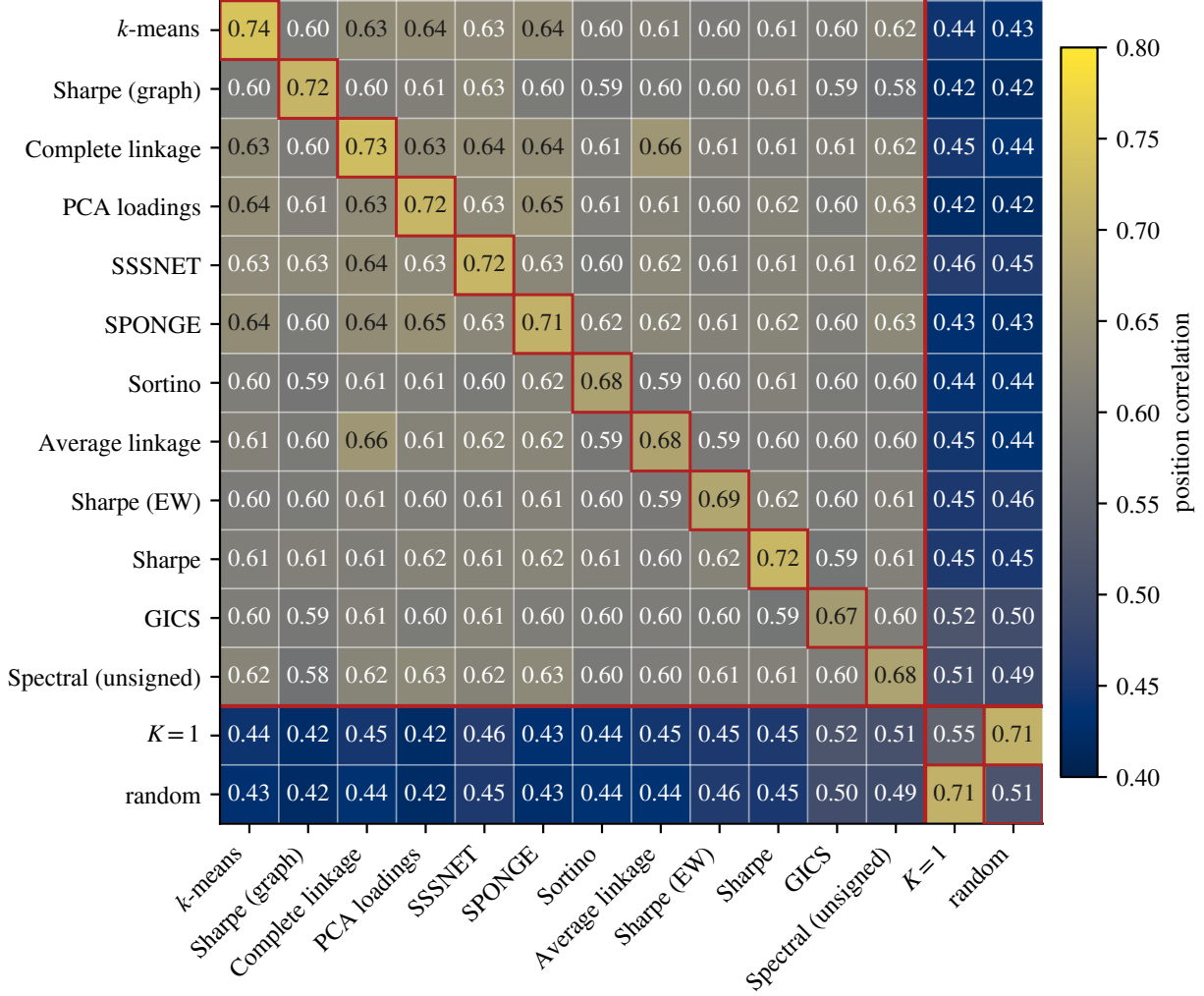


Figure 6: Position-vector correlation between arms. This figure shows the daily position-vector correlation between every pair of the 14 arms at the trained component over the common 6,913-day window, ordered by Sharpe ratio. The diagonal is not one: it is each arm against itself at a different seed, so every cell can be compared against the diagonal of its own row. The two controls are set off by a divider line.

Relative to a clusterizer’s correlation with its own refit, 0.704, the correlation drops by 0.09 between two clusterizers (to 0.612) and by 0.25 between a clusterizer and a control (to 0.450); the first drop is 36% of the second. The two controls correlate with each other at 0.71, more than either correlates with its own refit under a redrawn head seed, 0.55 and 0.51; every clusterizer’s own-refit correlation is 0.67 to 0.74. No clustering and a random partition therefore deliver much the same book, and retraining on a structureless signal is noisier than retraining on any partition. The instability is in the fit rather than the partition. The $K = 1$ control has no partition seed to redraw, and the random control’s labels carry no structure under any seed, so the random control’s Sharpe ratio moves as much when the label stream is redrawn as when the training seed is (0.39 across partition draws against 0.41 across head seeds, Section 5.2).

Because the 12 books are this similar, ex ante selection among individual clusterizers does not improve upon a simple diversified combination of their books. The equal-weight book earns a Sharpe ratio of 2.505 on the seed-averaged payoff, above nine of the 12 books on that convention. Re-selecting the trailing best clusterizer each January leaves a gap to the equal-weight book whose confidence interval covers zero at every look-back considered (Figure 7). The result is familiar from the forecast-combination literature: when candidates are exchangeable up to noise, equal weights are hard to beat (Timmermann, 2006).

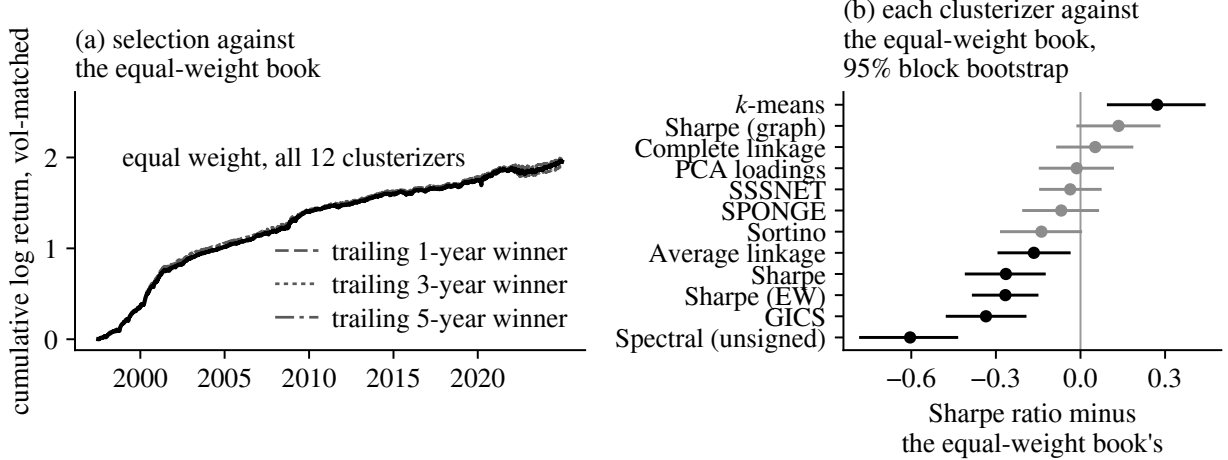


Figure 7: Selecting the trailing best clusterizer against holding all 12. This figure plots two comparisons of the seed-averaged trained books. (a) Cumulative log return of the equal-weight book and of three selection rules, each of which re-selects on the first trading day of each calendar year the clusterizer with the highest Sharpe ratio over the trailing one, three, or five years and holds it for the year; series are scaled to equal volatility. (b) Each clusterizer’s Sharpe ratio minus the equal-weight book’s, with 95% circular block-bootstrap intervals (block 21, 5,000 draws); black marks the clusterizers whose interval excludes zero, gray those whose interval includes it. All books are seed-averaged trained books over the 6,913-day sample, and panel (b) is ordered by that convention’s Sharpe ratio rather than by the mean-of-per-seed ranking of Figure 4.

Five of the 12 clusterizers, however, sit reliably below the equal-weight book, by 0.17 to 0.60 in Sharpe ratio; the intervals are from a circular block bootstrap over days (block 21, $B = 5,000$). One of the five, Sharpe (EW), the exponentially weighted variant, has a gap of 0.27 against a range of 0.30 across its own five partition draws, so its shortfall is not separated from the movement its own partition seed produces; of the other four, two clear their range and two have deterministic partitions and no such range to clear. Three clusterizers score above the equal-weight book, *k*-means, Sharpe (graph), and complete linkage, by 0.27, 0.13, and 0.05; only *k*-means’ interval excludes zero, $[0.10, 0.44]$. Its margin of 0.27 is about the size of the 0.29 range its own five partition draws produce, so the one advantage the day bootstrap resolves is not separated from the partition axis. Differences of this size do not identify a best clusterizer. They do identify the clusterizers to avoid, and the clusterizers furthest below the equal-weight book are the most volatile (Section 4.1).

5.5 The Estimation Window and the Number of Groups

Two upstream inputs move the Sharpe ratio on the scale of the span between clusterizers: the length of a clusterizer’s estimation window, which moves it by as much as the whole span, and the number of groups K , which moves it by up to 0.36 away from the $K = 30$ used throughout, more than two fifths of the span. The comparisons above hold both fixed. Holding the algorithm, the panel, the trading head, and the seeds fixed, we lengthen the window from 21 days to 252. Spectral clustering gains +0.88 in Sharpe ratio and PCA loadings +0.66, each monotone across the intermediate windows and paired by seed, while average linkage, which uses the same correlation matrix, moves 0.00. The difference between those responses is the clean comparison, because the arms use the same input; against average linkage, the window response of spectral clustering is +0.88 and of PCA loadings +0.67, the same sign in every seed. The two clusterizers that gain are the two that use individual eigenvectors of the correlation matrix, and three arms registered in advance to separate that account from its rival (that low starters gain most) move as the eigenvector account predicts (Appendix A.9).

The number of groups is the second input. Of five clusterizers swept at $K \in \{10, 30, 50\}$ (five seeds each, Appendix A.6), three rise monotonically in K , while *k*-means on returns and spectral clustering peak at 30 and fall by 0.09 and 0.14 between 30 groups and 50. For the four other than spectral clustering, coarsening to

10 groups costs 0.20 to 0.36 in Sharpe ratio, and refining to 50 changes it by -0.09 to $+0.22$. The gain from more groups cannot be split between finer structure and the finer demeaning of Section 4.2 without a random control at each K , which was not run; one run per K would not suffice (Appendix A.6). A downstream Sharpe ratio therefore distinguishes changes in the information supplied to the clusterizer and in the number of groups, but discriminates much less sharply among algorithms given the same input.

6 Factors, Peers, and Costs

A partition is a special case of a factor model, and clustering a special case of peer-based trading. As a factor model, it fixes each stock’s loading at 1 on its own group’s mean and 0 on every other group’s mean, whereas the latent factor literature estimates loadings freely (Fan et al., 2013). Gujarro-Ordóñez et al. (2025) remark in passing that “under appropriate assumptions a clustering problem can be solved by a latent factor estimation.” As a choice of peers, it gives every stock in a group the same peer set, whereas the simplest alternative gives each stock its own peer set, the names most correlated with it. This section asks whether either general construction already contains everything a partition contributes. Against factor models the test has two parts: in the cross-section, what a partition still explains after factors are projected out (Section 6.1); in the time series, whether the books the partitions produce are spanned by the standard factors, by a reversal factor, by the leading principal components, or by the same pipeline run without clustering (Section 6.2). Against per-stock peers, Section 6.3 replaces the partition with per-stock peer regressions and asks the same question of the books. Finally, Section 6.4 asks whether the same books’ returns survive trading costs.

6.1 Cross-Sectional Projections

We ask how much of the cross-sectional variance of a day’s residuals a partition can explain, and how much of that it still explains after the leading principal components are projected out. For each of the 6,913 out-of-sample days, we take the day’s cross-section of market residuals (Section 3.2), apply the partition’s same-day labels, and record the share of cross-sectional variance that demeaning within groups removes, the quantity measured in Section 4.2. The names used here are those with a complete 252-day residual history, the window on which the principal components are estimated: 479 to 500 a day for the 11 clusterizers other than the sector taxonomy, and 322 for GICS. The traded book of Section 4.2 has 494 names a day and is not the same set, which is why its numbers differ slightly from those here. We then project out the leading m principal components, estimated on the trailing 252 days and strictly before the day, and record what share the partition explains of what is left. If a low-dimensional latent factor representation captures the same cross-sectional structure as the partition, projecting out a few leading components should substantially reduce the variance share the partition subsequently explains.

Projecting out the five leading components removes only about half of what a real partition explains (Figure 8a). The k -means partition explains 0.265 of the day’s cross-sectional residual variance; after five components are projected out it still explains 0.160 of the remainder, 60% of its unconditional share. Across the 12 clusterizers the retention at five components runs from 44% to 61%, the sector taxonomy at the bottom and spectral clustering at the top. The random control sits at the degrees-of-freedom floor $(K - 1)/(N - 1) = 0.0597$ that K arbitrary groups remove from any cross-section, 0.0600 before the projection and 0.0599 after. The floor is computed on the universe used here; on the traded universe of Section 4.2 the same formula gives 0.0588, the 5.88% reported there. The $K = 1$ control is zero by construction.

Matching the k -means partition’s share of the day’s cross-section takes about 17 principal components on the 79% of days when 30 suffice, and on the other 21% even 30 are not enough (Figure 9). That count is specific to k -means: the other 11 are matched by 7 to 11 components, and 30 fall short on 0.2% to 3.0% of days. Block-bootstrap standard errors on every mean in this subsection are 0.001 to 0.005 and are omitted from the figures.

The Fama–French factors absorb even less than the principal components: after the full six-factor set the 12 clusterizers keep 50% to 68% of their unconditional shares (Figure 8b). We rebuild the same exercise with the stocks’ trailing loadings on nested Fama–French sets in place of principal components (CAPM, then three, five, and six factors). Every real partition’s share declines gently, and none comes close to the floor. For the k -means partition the sequence is 0.246, 0.206, 0.185, 0.178. Against the principal-component sequence, the whole six-factor set removes about as much as three statistical components and less than five, and that

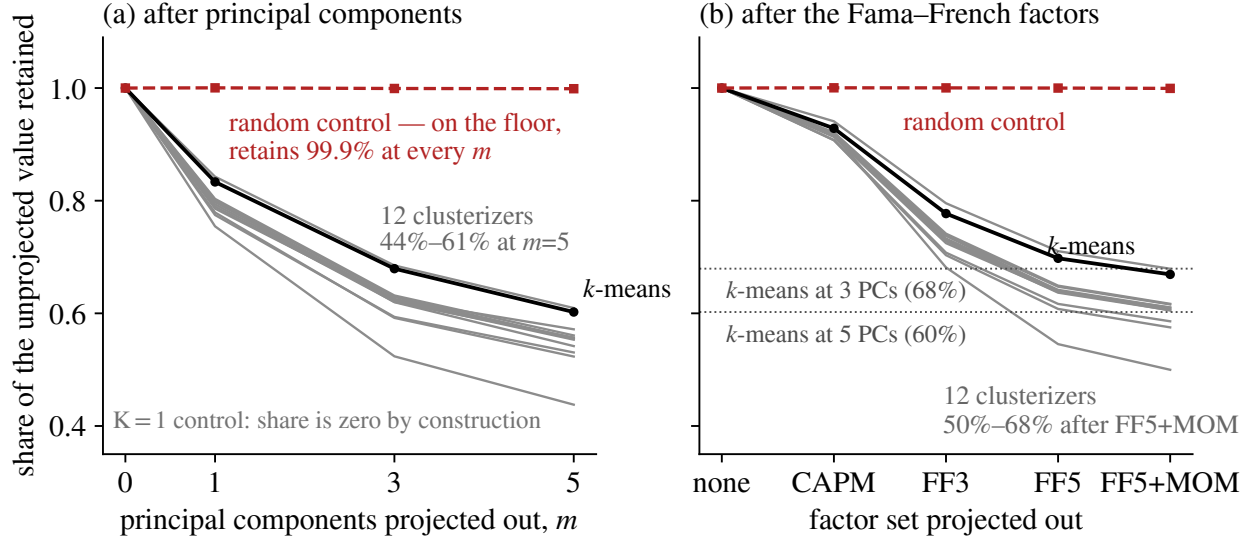


Figure 8: Partition share after projecting out principal components and after projecting out Fama–French loadings. This figure plots the share of the day’s cross-sectional residual variance each partition explains once a set of common factors is projected out, over the 6,913-day sample. The two panels share a vertical axis, so the two projections can be compared directly. (a) The leading principal components, projected out one, three, and five at a time. (b) The stocks’ trailing loadings on nested Fama–French sets, estimated on the same trailing 252 days as the principal components and on the aligned calendar (Appendix A.7); the two dotted horizontals mark where k -means sits in panel (a) at three and at five components. In both panels k -means on returns is black and the other 11 clusterizers gray; the random control is the dashed red line, which sits on the floor $(K - 1)/(N - 1)$ and retains 99.9% or more of its share at every projection, and the $K = 1$ control is zero by construction.

ordering holds for every one of the 12 rather than for k -means alone. The random control keeps 99.9% or more of its share under every one of the seven projections. Its share is only the degrees-of-freedom floor, which no projection can remove, so full retention is the expected result. The $K = 1$ control is again zero.

Loadings conditioned on characteristics absorb the least: removing the span of the four characteristics available on this panel (size, momentum, short-term reversal, and realized volatility) leaves every real partition with 77% to 85% of its share, more than survives three statistical components. Size is the log of market capitalization at the close of day t , momentum the return from day $t - 252$ to day $t - 21$, short-term reversal the return over the 21 days ending on day t , and realized volatility the standard deviation of raw returns over the 60 days ending on day t . Each is standardized across the day’s cross-section, and a missing value is set to zero after standardization. Unlike the principal components, which are estimated strictly before the day, size, reversal, and volatility are measured through day t itself. The same-day information favors the span, not the partition, so the retention reported here is, if anything, understated. The span bounds what any factor model with characteristic-linear loadings could absorb, the instrumented principal components (IPCA) family included (Kelly et al., 2019), whatever its coefficients. The four characteristics explain 0.13 of the day’s cross-sectional residual variance on their own, and the random control again stays at its floor. Accounting characteristics are not available on this panel, so the bound is for these instruments, not for every instrument set. Under every projection, every real partition keeps a positive excess over its own floor; under the principal components, how much it keeps depends on the number projected out and on the floor (Appendix A.10).

6.2 Book Spanning Tests

The time-series part of the factor test asks whether the standard factors span the books’ returns, that is, whether each book’s alpha against those factors is zero (Barillas and Shanken, 2017). Table 7 regresses every arm’s payoff on the Fama–French five factors (Fama and French, 2015) plus momentum (Carhart, 1997), at both trading components, with each component’s payoffs matched to the factor returns of the days on which they are realized (Appendix A.7). No book is spanned. Every alpha is positive, with t between 4.1 and 10.7,

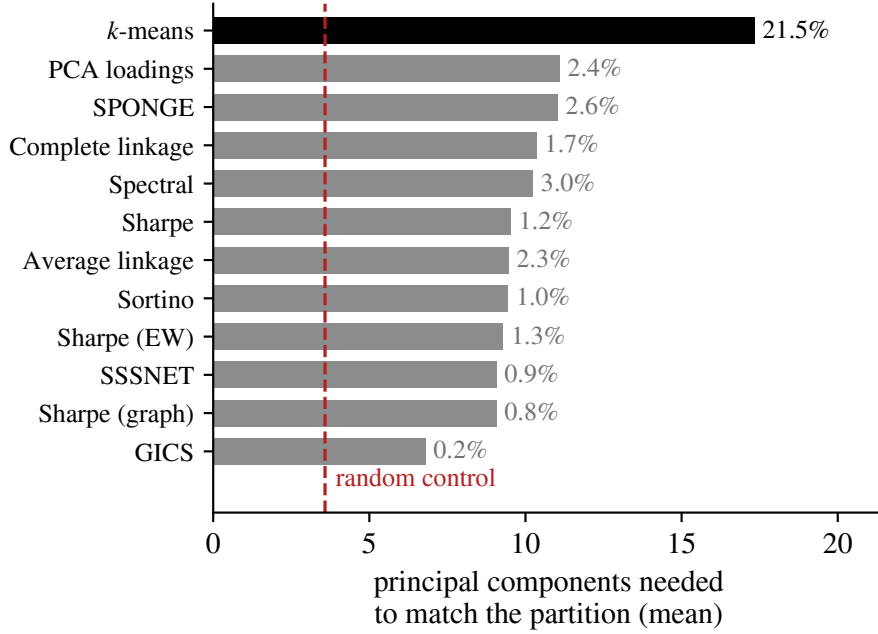


Figure 9: Principal components needed to match a partition. This figure plots, for each clusterizer, the mean number of leading principal components that explain as much of the day’s cross-section as its partition does, over the 6,913-day sample. The search stops at 30 components: the mean is taken over the days on which 30 suffice, and the percentage beside each bar is the share of days on which they do not. *k*-means on returns is black and the other 11 clusterizers gray; the dashed red line is the random control.

and the factors explain little of any of these series (R^2 at most 0.21 for the 12 clusterizers). The alphas of Section 4.1 are therefore not compensation for exposure to these factors.

These rejections are not artifacts of the finite-sample size of the *t*-tests. A Monte Carlo calibrated to each book’s own tails and dependence puts their empirical size at 6.6% at nominal 5%, and the critical value corrected for the worst-behaved book, 2.2 in *t* units, is far below the smallest *t* of 4.1 (Appendix A.11).

The rule-based component is itself a five-day reversal, so the natural factor to add is one that prices reversal. The short-term reversal factor of the French library reverses the prior 20 trading days rather than five, but it is the closest traded factor, and the lower block of Table 7 adds it to FF5+MOM. At the rule-based component the factor absorbs about a quarter of the clusterizers’ alpha and about two fifths of the controls’ (Figure 10): the clusterizers’ alphas, 8.05% to 8.81% under FF5+MOM, fall to 6.18% to 7.00%, the $K = 1$ control’s falls from 7.27% to 4.22%, and the random control’s from 7.43% to 4.57%. The clusterizers’ advantage over the controls therefore widens, from +15.6% and +13.1% to +54.5% and +42.8%. Under the neural trading head every book loads on the factor at only 0.05 to 0.07; the alphas fall by about a tenth and the advantage barely moves. Every one of the 28 books remains unspanned. The reversal the factor prices is the market-wide part; demeaning within a real group removes that part from the signal and keeps the part that is relative to the group. This is the intra- and inter-industry decomposition of Hameed and Mian (2015) in spanning form, with a statistical partition in place of the industry and with the two controls that the industry studies lack. The head loads on the factor lightly whatever the partition, so it collects little of the reversal premium the controls earn under the fixed rule; that is why the rule-based advantage now exceeds the trained one, the reverse of the ordering in Section 4.1.

The leading principal components, which absorb about half of a partition’s cross-sectional share (Section 6.1), absorb almost none of its alpha. Adding the five leading components as daily factors, estimated on the trailing 252 days as in Section 6.1, moves every alpha in Table 7 by at most 0.08 points and every R^2 by

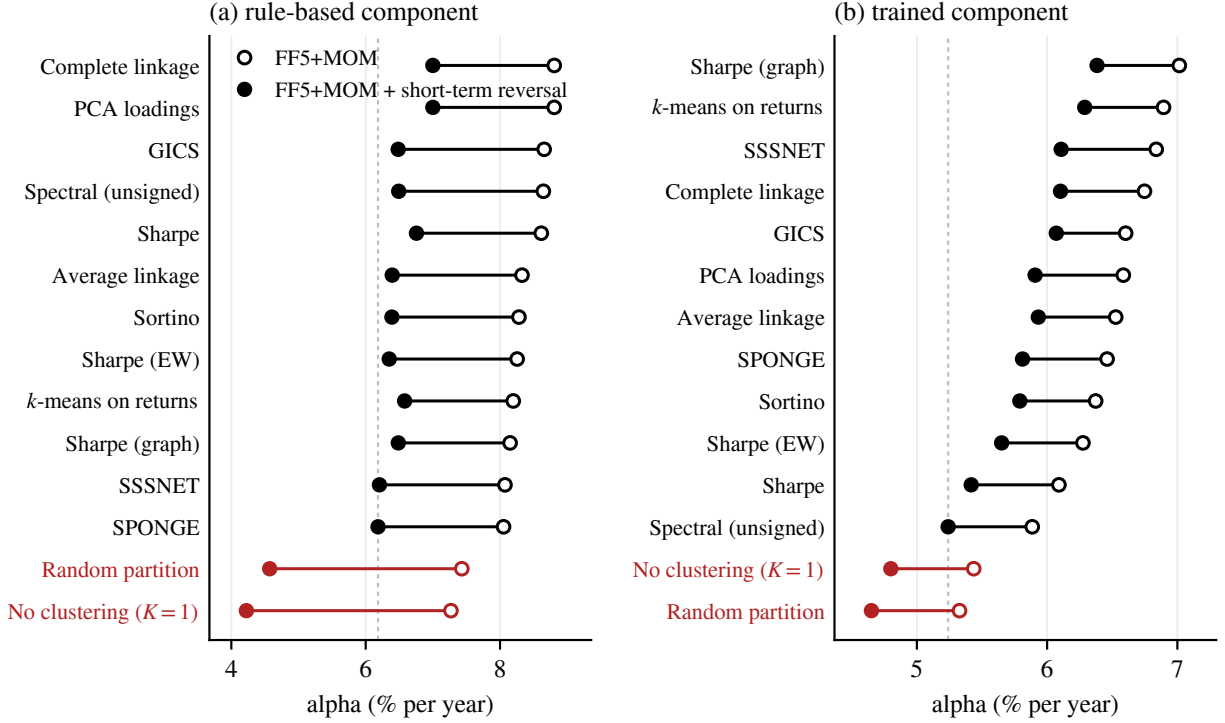


Figure 10: Alpha before and after the short-term reversal factor. This figure plots, for each arm, its alpha on FF5+MOM (open circle) and its alpha once the short-term reversal factor is added (filled circle), on the 6,913-day sample, from the regressions of Table 7. (a) The rule-based component; (b) the trained component. Within each panel the arms are ordered by their FF5+MOM alpha, the two controls are red, and the dotted vertical marks the lowest alpha with the reversal factor among the 12 clusterizers. The rule-based book has gross exposure 2, the head's 1, so the panels carry separate horizontal axes and only within-panel comparisons are meaningful.

at most 0.02.⁶

The closest benchmark of all is the pipeline run without clustering. The $K=1$ book of the same trading component shares the universe and the signal window, and differs from a clusterizer's book only in the grouping, so regressing one on the other asks directly what the grouping adds (Table 8). Every clusterizer keeps alpha against it, 4.48% to 5.73% a year at the rule-based component (t from 6.9 to 8.2) and 2.39% to 4.70% under the head (t from 5.8 to 9.4), with loadings of 0.37 to 0.65, and adding FF5+MOM changes these alphas little. The random control keeps almost none: its book loads on the $K=1$ book at 0.93 and 0.90 with an R^2 of 0.97 and 0.92, and its alpha is 0.69% ($t=2.3$) at the rule-based component and 0.40% ($t=1.7$) under the head. The alpha each clusterizer keeps against the $K=1$ book is 52% to 68% of its FF5+MOM alpha in Table 7 at the rule-based component and 41% to 68% under the head; this alpha is relative value within the group, and a random grouping supplies almost none of it.

Under FF5+MOM the factor loadings also separate the controls from the clusterizers: at the trained component the $K=1$ control loads on the market at 0.105 and the random control at 0.093, against 0.051 to 0.077 for the 12 clusterizers. At the rule-based component, where the book is dollar-neutral by construction, the range is compressed (R^2 0.05 to 0.07 everywhere), and the market loadings sit between 0.05 and 0.06 for the clusterizers and at 0.07 for both controls. This pattern is the volatility reduction of Section 4.1 in factor terms: a real partition leaves the trained book with less factor exposure than a missing or a random one.

The clusterizers' advantage over the controls holds after 2009 as well as before. Split at January 2009 (Table 9), alphas fall everywhere in the later period, from 9%–16% to below 5% a year, but the separation

⁶Each component's daily return is the payoff of the unit-gross book of residuals whose weights are the component's loading vector from the trailing window, sign-aligned from day to day. The five factors have annualized volatilities from 13.5% down to 5.6% and insignificant premia.

Table 7: Spanning against FF5+MOM, with and without a short-term reversal factor, within each trading component

	Rule-based component				Trained component			
	α	t_α	R^2	β_{Rev}	α	t_α	R^2	β_{Rev}
<i>FF5+MOM</i>								
12 clusterizers	8.05–8.81	6.4–7.6	0.05–0.07	—	5.89–7.02	8.2–10.7	0.14–0.21	—
$K = 1$	7.27	4.1	0.05	—	5.44	6.8	0.24	—
Random	7.43	4.4	0.05	—	5.33	7.1	0.22	—
<i>FF5+MOM+Reversal</i>								
12 clusterizers	6.18–7.00	5.3–6.6	0.27–0.31	0.16–0.21	5.24–6.38	7.5–10.4	0.21–0.30	0.05–0.07
$K = 1$	4.22	2.5	0.25	0.30	4.80	6.1	0.28	0.06
Random	4.57	2.8	0.25	0.28	4.65	6.3	0.27	0.07

Notes. This table reports spanning regressions of each arm’s payoff on the six factors, and on the six factors plus a short-term reversal factor, at each trading component, for the 12 clusterizers and the two controls at the common 60-day window, on the 6,913-day window of Section 3.1 (July 1997 to December 2024). Alphas are in percent a year. FF5+MOM is the Fama–French five factors plus momentum. Reversal is the daily short-term reversal factor (ST_Rev) of the Ken French data library (CRSP vintage 202606): the average return of the two low-prior-return portfolios minus the two high-prior-return portfolios from 2×3 independent sorts on size and prior return, with prior return measured over days -20 to -1 and the portfolios formed daily; β_{Rev} is its loading. Newey–West(21) standard errors (Newey and West, 1987); trained cells are on the seed-averaged payoff; each trading component is joined to the factor file on its own realization calendar (Appendix A.7). The rule-based book has gross exposure 2, the head’s 1; only within-component comparisons are meaningful.

Table 8: Spanning against the no-clustering book of the same trading component

	On the $K = 1$ book				With FF5+MOM added	
	α	t_α	β	R^2	α	t_α
<i>Rule-based component</i>						
12 clusterizers	4.48–5.73	6.9–8.2	0.37–0.59	0.53–0.81	4.37–5.67	6.9–8.3
Random	0.69	2.3	0.93	0.97	0.64	2.1
<i>Trained component</i>						
12 clusterizers	2.39–4.70	5.8–9.4	0.42–0.65	0.48–0.69	2.46–4.77	5.8–9.5
Random	0.40	1.7	0.90	0.92	0.39	1.7

Notes. This table reports regressions of each clusterizer’s payoff, and of the random control’s, on the $K = 1$ book of the same trading component, which shares the universe and the signal window and differs only in the grouping; β is the loading on the $K = 1$ book, and alphas are in percent a year. The last two columns add FF5+MOM to the regressors. Sample, calendar, and conventions are those of Table 7, with Newey–West(21) standard errors and trained cells on the seed-averaged payoff. The bands cover the 12 clusterizers.

between clusterizers and controls remains on three measures. First, in both eras and at both trading components, the controls are more volatile than every clusterizer. Second, every alpha t falls, but the ordering is unchanged: over 2009–2024 the lowest clusterizer t (2.6 at the trained component, 2.8 at the rule-based component) exceeds the highest control t (2.5 and 2.1). Third, at the trained component the alpha advantage over the $K = 1$ control is larger in the later era, +29.3% over 2009–2024 against +20.2% before, with all 12 clusterizers above that control in the earlier era and 11 of the 12 in the later, spectral clustering the exception. With the reversal factor added, the rule-based controls are no longer significant after 2009 ($t = 1.5$ and 1.7) while every clusterizer at that component remains so (t from 2.6 to 3.6); at the trained component the controls keep $t = 2.1$ and 2.2, below every clusterizer (2.3 to 5.3).

Table 9: Spanning against FF5+MOM by era, 1997–2008 and 2009–2024

	1997–2008			2009–2024		
	α (%/yr)	t_α	σ (%/yr)	α (%/yr)	t_α	σ (%/yr)
<i>Trained component</i>						
12 clusterizers	10.6–12.1	9.1–10.4	2.6–3.2	2.0–3.2	2.6–5.5	2.7–3.6
$K = 1$	9.41	7.0	4.4	2.15	2.3	4.4
Random	9.29	7.1	4.2	2.09	2.5	4.1
<i>Rule-based component</i>						
12 clusterizers	14.2–16.2	6.4–7.4	5.3–6.5	3.4–4.4	2.8–3.9	4.4–5.9
$K = 1$	12.89	4.0	10.4	3.64	1.9	8.4
Random	13.03	4.3	9.8	3.78	2.1	8.1

Notes. This table reports the spanning regressions of each arm’s payoff on FF5+MOM, split at January 2009, with the same series and calendars as the FF5+MOM block of Table 7, on the 6,913-day window of Section 3.1; σ is the era’s realized annualized volatility of the same series, so the trained σ is that of the seed-averaged payoff, the convention quoted in the notes to Table 2 rather than the per-seed levels in its cells. Only within-component comparisons are meaningful.

6.3 Per-Stock Peer Regressions

The second general construction is the per-stock peer set. We build two per-stock peer books on this panel: for every stock, we take its 20 most correlated names over the trailing 21 days, regress its residual return on theirs (fit windows of 60 and 252 days give the two books), and pass the resulting per-stock residual to the same trading head. The peer sets overlap, are not shared, and involve no partition. Each book is a single head run at seed 42, and peer selection itself is deterministic, so no dispersion is available for these two arms and we make no level comparison against the clusterizers.

In the cross-section, the peer construction absorbs much of what a partition explains. None of the factor projections above comes close. When the measurement of Section 6.1 is rerun with the peer-fitted component removed from each stock first, retention falls to 25% to 32% of the unconditional shares for 10 of the 12 clusterizers, against 44% to 61% under five principal components; k -means retains 52%, inflated by its singletons (Appendix A.10), and GICS 17%. Adding the principal components *after* the peer removal changes almost nothing (at most one point of retention for 10 of the 12 clusterizers, and at most five for the two exceptions), so the joint span of peers and factors is, for this purpose, the peer span. The random control remains at its floor through every stage, retaining 99.7%.

The payoffs separate the clustering and the peer-regression strategies even though the cross-section does not. Table 10 regresses every arm’s payoff on the two peer books, on the six pre-specified factors, and on both together. Against the factors alone every arm keeps its alpha, as in Section 6.2. Against the peer books the results divide. Both controls lose their alpha entirely, the $K = 1$ book to -1.6% a year ($t = -2.2$) and the random book to -1.4% ($t = -2.0$), so whatever this pipeline earns without a real partition, two per-stock peer regressions earn as well. Of the 12 clusterizers, four keep theirs: 0.8% to 1.5% a year, at t between 2.1 and 3.9. The other eight fall between -0.81% and 0.74% (t from -1.7 to 1.8); the weakest is spectral clustering, which is also the arm that gains nothing in K (Section 5.5), has the highest volatility (Section 4.1), and sits furthest below the equal-weight book (Section 5.4). Adding the factors on top of the peer books raises every one of the 12 alphas, by 0.17 to 0.47 points; adding the reversal factor as well moves them by no more than 0.11 points, since the peer books already absorb what it would. The factors raise the alphas and take common variation out of the residual, and both effects sharpen the t -statistics: eight of the 12 clear a t of 1.96 against the peer books and the factors together, against four on the peer books alone.

The two controls show that this alpha comes from the grouping itself and not from the stages around it. The controls use the same residualization, the same signal window, and the same trading head, with a random partition or none. If those stages produced the alpha, the controls would keep alpha against the peer books as well. They keep none.

Regressed the other way, on the equal weight of all 12 clusterizers, the peer books keep their alpha as

Table 10: Spanning against peer books, factors, and both

	On the peer books		On the six factors		On both	
	α (%/yr)	t_α	α (%/yr)	t_α	α (%/yr)	t_α
12 clusterizers ^a	-0.81–1.45	-1.7–3.9	5.89–7.02	8.2–10.7	-0.35–1.68	-0.7–4.5
$K = 1$	-1.63	-2.2	5.44	6.8	-0.82	-1.2
Random	-1.42	-2.0	5.33	7.1	-0.74	-1.1
Peer book, 60-day fit, on the clusterizers' equal weight	2.52	5.5				
Peer book, 252-day fit, on the clusterizers' equal weight	2.32	6.8				

Notes. This table reports spanning regressions of each arm's payoff on the two peer books, on the six factors (the Fama–French five factors plus momentum), and on both, and of each peer book on the equal weight of the 12 clusterizers. Alphas are in percent a year. The six-factor columns repeat the trained-component regressions of the FF5+MOM block of Table 7 for comparison. Clusterizer payoffs are seed-averaged over the five head seeds; the peer books are single runs at seed 42. All series are panel-internal, so the regressions share one calendar; the factor file is joined at a one-day lag, as elsewhere in this section. Newey–West(21) t -statistics.

^a All three ranges cover the 12 clusterizers. On the peer books four of the 12 keep a t above 1.96; on the peer books and the six factors together, eight do.

well: 2.5% and 2.3% a year ($t = 5.5$ and 6.8). The two constructions summarize much the same comovement, yet most clusterizer books earn alpha the peer books do not, and the peer books earn alpha the clusterizer books do not. Since the 12 clusterizer books differ little among themselves (Section 5.4), the alpha that survives appears to attach to the class of peer set, shared or per-stock, rather than to the choice within a class. A clusterizer's closeness to the peer books is not significantly related to its Sharpe ratio at the 5% level (Appendix A.12); the volatility of the books separates the clusterizers (Section 4.1).

6.4 Transaction Costs

The last question is whether the books' returns survive trading costs. The books turn over about two fifths of their notional every day: 38.2% to 46.9% one-way across the 14, a spread of 23% against a spread of 101% in gross Sharpe ratio, both as the top relative to the bottom of the 14. Turnover is therefore much less dispersed across partitions than gross Sharpe ratios. Break-even sits at 2.7 to 3.8 basis points one-way, and at daily rebalancing no configuration is net positive at five.

At costs below break-even, the net Sharpe ranking follows the gross ranking; at costs above break-even, it follows turnover. At one and two basis points every one of the 12 clusterizers holds a higher net Sharpe ratio than either control, the two controls remain the two lowest of the 14 books, and the net ordering correlates with the gross ordering at $\rho = 0.98$ and 0.91 . At five basis points that correspondence largely dissolves. The rank correlation with gross falls to about 0.3; GICS, eleventh of the 12 on gross Sharpe ratio, is the least negative book; k -means on returns falls from first to a tie for seventh of the 14, level with spectral clustering to three decimals; and both controls move up among the clusterizers, $K = 1$ above four of the 12 and the random control above one. Turnover sorts the arms at that level, and turnover is not a measure of clustering quality. The arm that rises furthest, GICS, is also the arm whose book depends least on its own labels: about a third of the names it trades on a given day are assigned by the nearest-correlation fill rather than by the sector table (Table 1, note e), so its low turnover cannot be attributed to the taxonomy. Average linkage shows the same reordering without that qualification, turning over 39.7% and finishing fourth at five basis points. The gross ranking is therefore not an ordering of economic value, and every comparison in Section 4 and Section 5 is a gross comparison and should be interpreted as one. Appendix A.13 gives the cost model, the full table, the same measurement on the rule-based component, and the limitations of the exercise.

7 Conclusion

This paper separates the two decisions a clustering-based statistical arbitrage strategy makes: how to group stocks and what model to trade within the groups. It measures what the clustering contributes against a no-clustering control and a random-partition control, at a fixed trading rule and under a neural trading head. A real partition lowers the volatility of the book, by about 40% relative to either control under the fixed rule, and raises its alpha. The volatility reduction belongs to the clustering and every clusterizer supplies it; the neural trading head widens the alpha advantage, most against the random partition, so a wrong partition costs more under a fitted model.

More broadly, informative clustering is much easier to distinguish from the controls than competing clustering algorithms are from one another. Whether an informative partition is used matters far more than which one. The 12 clusterizers disagree about the partition but largely agree about the book, and their Sharpe ratios differ mainly through the volatility of their books. Their payoffs correlate with one another as strongly as with their own refits, most adjacent clusterizers in the Sharpe ranking are closer together than the run-to-run variation in their gap, and redrawing the partition seed changes the leader at both trading components. A downstream Sharpe ratio distinguishes changes in the information a clusterizer uses and in the number of groups it forms; given the same input, it discriminates much less sharply among algorithms. Ex ante selection on past performance does not outperform a simple equal-weight combination of the 12 books.

The comovement a partition captures is only partly absorbed by principal components or by the Fama–French factors in the cross-section, and the partition’s books are not spanned by either, by a short-term reversal factor, or by the same strategy traded without clustering. Per-stock peer regressions absorb much of what a partition explains in the cross-section, yet most of the 12 clusterizer books keep their alpha against the peer books and the factors together, and the peer books keep theirs against the equal weight of all 12. The two constructions appear to differ in the class of peer set, shared or per-stock, rather than in the choice within a class. None of these books survives a five-basis-point one-way cost at daily rebalancing: break-even sits below four basis points one-way, and at five no book is net positive.

Two rules for evaluating partitions follow. A comparison of partitions should include a trading component with no fitted parameter, so that what the partition contributes is not confounded with what the model fits; a fitted model narrows the volatility benefit and widens the alpha benefit. A ranking of partitions should first check its gaps against seed and partition-draw noise. The two controls, the two noise benchmarks, and the split of the Sharpe gap into its mean and volatility parts apply unchanged to the comparison between shared and per-stock peer sets, which is the open question this design leaves.

Data availability

The CRSP daily data underlying this article were obtained under license through Wharton Research Data Services and cannot be redistributed; the replication package contains the code that rebuilds the panel from the CRSP daily files and the checksums that identify the panel used here. The GICS sector memberships used by one arm are likewise licensed, and the partitions derived from them are provided in their place. The factor returns were derived from a publicly available source, the Kenneth R. French data library (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). The derived data underlying the tables and figures—the daily return of every book, the positions of the trained runs, the partition caches, and the training logs—are available on Zenodo at <https://doi.org/10.5281/zenodo.22241254> under a CC BY 4.0 license (Zhu et al., 2026). The code, with the checks that tie the printed values to these files, is available at <https://github.com/lzwbh/clustering-statarb-replication> under an MIT license.

References

- Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3): 1203–1227, 2013. doi: 10.3982/ECTA8968.
- Rafael P. Alves, Diego S. de Brito, Marcelo C. Medeiros, and Ruy M. Ribeiro. Forecasting large realized covariance matrices: The benefits of factor models and shrinkage. *Journal of Financial Econometrics*, 22 (3):696–742, 2024. doi: 10.1093/jjfinec/nbad013.

- Tomohiro Ando and Jushan Bai. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519):1182–1198, 2017. doi: 10.1080/01621459.2016.1195743.
- Marco Avellaneda and Jeong-Hyun Lee. Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782, 2010. doi: 10.1080/14697680903124632.
- Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002. doi: 10.1111/1468-0262.00273.
- Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005. doi: 10.1214/009117905000000233.
- David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.
- Francisco Barillas and Jay Shanken. Which alpha? *Review of Financial Studies*, 30(4):1316–1338, 2017. doi: 10.1093/rfs/hhw101.
- Sanjeev Bhojraj and Charles M. C. Lee. Who is my peer? A valuation-based approach to the selection of comparable firms. *Journal of Accounting Research*, 40(2):407–439, 2002. doi: 10.1111/1475-679X.00054.
- Sanjeev Bhojraj, Charles M. C. Lee, and Derek K. Oler. What’s my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774, 2003. doi: 10.1046/j.1475-679X.2003.00122.x.
- David Blitz, Joop Huij, Simon Lansdorp, and Marno Verbeek. Short-term residual reversal. *Journal of Financial Markets*, 16(3):477–504, 2013. doi: 10.1016/j.finmar.2012.10.005.
- Rafaël Brutti, Maciej Marowka, and Mihai Cucuringu. Noise-robust orthogonal clustering and applications to equity markets. *Quantitative Finance*, 26(6):903–930, 2026. doi: 10.1080/14697688.2026.2671183.
- Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997. doi: 10.1111/j.1540-6261.1997.tb03808.x.
- Louis K. C. Chan, Josef Lakonishok, and Bhaskaran Swaminathan. Industry classifications and return comovement. *Financial Analysts Journal*, 63(6):56–70, 2007. doi: 10.2469/faj.v63.n6.4927.
- Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. SPONGE: A generalized eigenproblem for clustering signed networks. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 1088–1098, 2019.
- Zhi Da, Qianqiu Liu, and Ernst Schaumburg. A closer look at the short-term return reversal. *Management Science*, 60(3):658–674, 2014. doi: 10.1287/mnsc.2013.1766.
- Elliot Epstein, Rose Wang, Jaewon Choi, and Markus Pelger. Attention factors for statistical arbitrage. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF)*, pages 396–404, 2025. doi: 10.1145/3768292.3770398.
- Eugene F. Fama and Kenneth R. French. Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193, 1997. doi: 10.1016/S0304-405X(96)00896-3.
- Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015. doi: 10.1016/j.jfineco.2014.10.010.
- Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B*, 75(4):603–680, 2013. doi: 10.1111/rssb.12016.

- Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669, 2018. doi: 10.1016/j.ejor.2017.11.054.
- Daniel Y. Fu, Elliot L. Epstein, Eric Nguyen, Armin W. Thomas, Michael Zhang, Tri Dao, Atri Rudra, and Christopher Ré. Simple hardware-efficient long convolutions for sequence modeling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 10373–10391, 2023. arXiv:2302.06646.
- Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006. doi: 10.1093/rfs/hhj020.
- Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020. doi: 10.1093/rfs/hhaa009.
- Jorge Guíjarro-Ordóñez, Markus Pelger, and Greg Zanotti. Deep learning statistical arbitrage. *Management Science*, 2025. doi: 10.1287/mnsc.2022.03132. Articles in Advance, published online December 4, 2025; arXiv:2106.04028.
- Allaudeen Hameed and G. Mujtaba Mian. Industries and stock return reversals. *Journal of Financial and Quantitative Analysis*, 50(1–2):89–117, 2015. doi: 10.1017/S0022109014000404.
- Chulwoo Han, Zhaodong He, and Alenson Jun Wei Toh. Pairs trading via unsupervised learning. *European Journal of Operational Research*, 307(2):929–947, 2023. doi: 10.1016/j.ejor.2022.09.041.
- Peter Reinhard Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2): 453–497, 2011. doi: 10.3982/ECTA5771.
- Yixuan He, Gesine Reinert, Songchao Wang, and Mihai Cucuringu. SSSNET: Semi-supervised signed network clustering. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 244–252, 2022. doi: 10.1137/1.9781611977172.28.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979.
- Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.
- Narasimhan Jegadeesh. Evidence of predictable behavior of security returns. *Journal of Finance*, 45(3): 881–898, 1990. doi: 10.1111/j.1540-6261.1990.tb05110.x.
- Qi Jin, Mihai Cucuringu, and Álvaro Cartea. Correlation matrix clustering for statistical arbitrage portfolios. In *Proceedings of the Fourth ACM International Conference on AI in Finance, ICAIF ’23*, pages 557–564, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702402. doi: 10.1145/3604237.3626894.
- Bryan T. Kelly and Dacheng Xiu. Financial machine learning. *Foundations and Trends in Finance*, 13(3–4): 205–363, 2023. doi: 10.1561/5000000064.
- Bryan T. Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019. doi: 10.1016/j.jfineco.2019.05.001.
- Christopher Krauss, Xuan Anh Do, and Nicolas Huck. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2):689–702, 2017. doi: 10.1016/j.ejor.2016.10.031.
- Olivier Ledoit and Michael Wolf. Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, 15(5):850–859, 2008. doi: 10.1016/j.jempfin.2008.03.002.

- Hyunju Lee and Woojin Chang. An end-to-end deep learning model for clustering-based statistical arbitrage. *IEEE Access*, 14:47301–47318, 2026. doi: 10.1109/ACCESS.2026.3678064.
- Bruce N. Lehmann. Fads, martingales, and market efficiency. *Quarterly Journal of Economics*, 105(1):1–28, 1990. doi: 10.2307/2937816.
- Mel MacMahon and Diego Garlaschelli. Community detection for correlation matrices. *Physical Review X*, 5(2):021006, 2015. doi: 10.1103/PhysRevX.5.021006.
- R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11(1):193–197, 1999. doi: 10.1007/s100510050929.
- Rosario N. Mantegna and H. Eugene Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, Cambridge, UK, 1999. ISBN 978-0-521-62008-6. doi: 10.1017/CBO9780511755767.
- Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.
- Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016, 2010. doi: 10.1162/REST_a_00043.
- J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Physical Review E*, 68:056110, 2003. doi: 10.1103/PhysRevE.68.056110.
- Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011. doi: 10.1214/11-AOS887.
- Simão Moraes Sarmento and Nuno Horta. Enhancing a pairs trading strategy with the application of machine learning. *Expert Systems with Applications*, 158:113490, 2020. doi: 10.1016/j.eswa.2020.113490.
- Allan Timmermann. Forecast combinations. In Graham Elliott, Clive W. J. Granger, and Allan Timmermann, editors, *Handbook of Economic Forecasting*, volume 1, chapter 4, pages 135–196. Elsevier, 2006. doi: 10.1016/S1574-0706(05)01004-9.
- Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N. Mantegna. Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258, 2008. doi: 10.1016/j.jedc.2007.01.034.
- Ajim Uddin, Xinyuan Tao, and Dantong Yu. The network factor of equity pricing: A signed graph Laplacian approach. *Journal of Financial Econometrics*, 22(5):1616–1655, 2024. doi: 10.1093/jjfinec/nbae010.
- Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.
- Chao Zhang, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian. Volatility forecasting with machine learning and intraday commonality. *Journal of Financial Econometrics*, 22(2):492–530, 2024. doi: 10.1093/jjfinec/nbad005.
- Chao Zhang, Xingyue Pu, Mihai Cucuringu, and Xiaowen Dong. Graph-based methods for forecasting realized covariances. *Journal of Financial Econometrics*, 23(2):nbae026, 2025. doi: 10.1093/jjfinec/nbae026.
- Lunji Zhu, Yixuan He, and Mihai Cucuringu. [dataset] Quantifying the contributions of clustering to statistical arbitrage: Replication data. Zenodo, 2026. doi: 10.5281/zenodo.22241254.

A Appendix: Data, Implementation, and Additional Results

This appendix has two parts. Appendix A.1 to Appendix A.7 give the construction of each clusterizer, the treatment of degenerate groups, the realized partition sizes, the trading head’s architecture and schedule, the identifiers under which every arm was run, the sweep over the number of groups, and the calendar offset between the panel and the factor file. Appendix A.8 to Appendix A.13 give the additional results the text refers to, in the order in which the text refers to them: an experiment that degrades a real partition continuously toward a random one, the mechanism behind the estimation-window effect, three robustness checks on the factor and peer comparisons, and the transaction-cost calculation in full. Data and code are available as described in the Data availability statement.

A.1 The Clusterizers in Detail

All clusterizers receive the same input: the day’s universe of N_t stocks (Section 3.1) and the trailing window of market residuals $\mathbf{R}_{tr} \in \mathbb{R}^{60 \times N_t}$, 60 days ending at t . All but GICS target $K = 30$ groups. Where a correlation matrix appears it is $\mathbf{C} = \text{corr}(\mathbf{R}_{tr})$, the $N_t \times N_t$ sample correlation of the window’s columns.

***k*-means on returns.** Each stock is treated as the point of its own 60 residual returns in $\mathbb{R}^{60}$, and *k*-means with Euclidean distance partitions those points (30 clusters, 4 restarts). No correlation matrix is formed. Because the coordinates are unstandardized, distance is dominated by volatility. High-volatility names sit far from everything and often become singletons (about 10 a day), whose members the trading stage later drops (Table 1, note b).

Average and complete linkage. Agglomerative clustering on the dissimilarity $\mathbf{D} = 1 - \mathbf{C}$: start from singletons and repeatedly merge the two closest clusters until 30 remain. The two clusterizers differ only in how the distance between clusters is defined: the average of all cross-pair dissimilarities, or the maximum. Both are deterministic, so the same input gives the same partition and there is no seed to vary.

Spectral, unsigned. Spectral clustering on the affinity $|\mathbf{C}|$: eigenvectors of the graph Laplacian of $|\mathbf{C}|$ embed the stocks, and *k*-means groups the embedded points. Taking absolute values means a strong negative correlation counts as a strong link.

PCA loadings. The 10 leading eigenvectors of $\mathbf{C}$ give each stock a 10-dimensional coordinate, its loadings on the top 10 principal components, and *k*-means groups those coordinates. Stocks are grouped by which factors they load on.

sponge. Cucuringu et al. (2019) give two variants. We use the symmetric one, SPONGE_{sym}, and SPONGE in this paper always means that variant. The correlation matrix is split into its positive part $\mathbf{C}^+ = \max(\mathbf{C}, 0)$ and negative part $\mathbf{C}^- = \max(-\mathbf{C}, 0)$, and the method solves a generalized eigenproblem on the two signed-graph Laplacians (both regularization parameters set to 1), seeking groups with positive correlation inside and negative correlation across. *k*-means on the resulting embedding gives the labels. Unlike in the unsigned spectral clusterizer, the sign of a correlation is information here. A post-processing step reassigns members of groups smaller than two to their most correlated group.

sssnet. SSSNET is the signed-graph neural network of He et al. (2022), trained on a clustering objective. The signed graph is built from the same 60-day correlation matrix, node features are the identity, and training is warm-started from a SPONGE partition. In its original form it computes its partition at run time from a random-number stream shared with the trading head. In this paper it is run once, with daily seeds, to produce a fixed set of daily partitions for the whole sample, 23 July 1993 to 30 December 2024, which the five head seeds then share, exactly as for the other eight stochastic clusterizers; its partitions under seeds 1 to 5 are generated the same way. Across the five partition draws the rule-based book moves by 0.15 in Sharpe ratio and the trained book at head seed 42 by 0.13 (2.398 to 2.527), the fifth-largest and the smallest ranges among the nine stochastic clusterizers (Section 5.2).

The trading-driven family. Four clusterizers replace the clustering criterion with a trading objective: a soft assignment matrix over K groups is optimized directly against the payoff of the rule-based book that the assignment would produce, then hardened by argmax. The objective is evaluated on the trailing 252 trading days of residuals ending the day before the partition is applied, so no return from that day onward enters the partition. All four warm-start from a SPONGE partition, balance group sizes by a Sinkhorn step, and penalize groups whose members are too weakly correlated (a cohesion floor). The four differ in the objective and in the parameterization: *Sharpe* maximizes the plain Sharpe ratio of the induced book; *Sortino* replaces the standard deviation in that objective with the downside deviation; *Sharpe (EW)* is a variant of *Sharpe* that weights recent days exponentially (decay 0.99 over a 252-day window, so the oldest day carries 8% of the newest day’s weight) and differs from *Sharpe* in more than the weighting (Table 1, note d); *Sharpe (graph)* produces the assignment with a signed-graph network rather than free logits, and evaluates the objective on the 30-day horizon the trading signal uses.

gics. A static sector table, from a single S&P 1500 snapshot: the taxonomy itself reads no return data; the correlation fill for the uncovered names does. The table covers 54% to 73% of the universe depending on the era. The remaining names, about 35% of the book on a given day, are assigned to the group with which they correlate most strongly, so the traded arm is a taxonomy–correlation hybrid. Its realized number of groups is the number of sectors in the snapshot, exactly 10 on every day of the sample, not K . The snapshot predates the separation of real estate into an eleventh sector in 2016.

The two controls. $K = 1$ places every stock in one group. Random draws each stock’s label uniformly from the 30 groups, independently each day, reading nothing.

A.2 Two Policies for Degenerate Clusters

The pipeline treats clusters of fewer than two members in two ways, and which treatment an arm receives depends only on which code generated its partition. SPONGE’s generation path repairs them: each orphaned member is merged into its most correlated group, and the stock stays in the book. The Sharpe and Sortino clusterizers inherit the same repair. Every other clusterizer passes its labels on unrepaired, and the trading stage then deletes every member of a sub-minimum cluster from that day’s book.

The asymmetry is nominal for 11 arms and material for one. Where the repair applies it has almost nothing to repair (0.13 groups per day for SPONGE, a handful of days in the whole sample for the other two), and every clusterizer except k -means on returns produces at most 0.6 singletons per day. k -means on returns produces 9.8 singleton labels per day, which the trading stage deletes as 8.65 names per day, and at the trained component its traded book holds 491.3 names against 499.5 to 500.0 for every other arm.

The cost of the asymmetry is measured; on this panel it is too small to resolve. A second arm trades the same k -means partitions with singleton members merged rather than deleted, which is k -means under SPONGE’s policy. That arm earns 0.04 less in Sharpe ratio on the five-seed average and runs 0.05 percentage points more volatile. Neither difference survives its own dispersion: the Sharpe gap carries $t = -0.9$ on four degrees of freedom and changes sign across the five seeds, and the volatility gap runs from 0.00 to +0.24 across them. Set against k -means’ own head-seed standard deviation of 0.112, the gap between the two arms sits inside the noise of a single arm’s refit. Harmonizing the policy in either direction would therefore move the leading clusterizer by less than its run-to-run noise, and imposing the deletion policy on SPONGE would in addition rebuild every one of its partitions to change 0.0 names per day. The repair can also empty a label, so SPONGE realizes 29.9 clusters per day against a nominal 30, as does SSSNET, whose partition is warm-started from a SPONGE partition; no other data-driven arm falls short of its nominal K . We keep each clusterizer’s own policy and state it.

A.3 Partition Size Distributions

Nominally every arm but GICS and the $K = 1$ control targets $K = 30$. In realized terms they differ widely (Table 11), and the differences do not track performance: over the 12 clusterizers, the Spearman correlation between effective K and the mean-of-per-seed Sharpe ratio is -0.15 with $p = 0.63$.

Table 11: Realized partition shape, mean over all 6,913 out-of-sample days

Arm	Effective K	Largest group	Singletons/day	Groups < 5/day
Sharpe (EW)	29.26	22.1	0.00	0.07
Sortino	29.08	23.1	0.00	0.06
Sharpe	28.98	23.5	0.00	0.08
Sharpe (graph)	28.58	24.9	0.01	0.06
Random	28.36	25.5	0.00	0.01
PCA loadings	26.86	30.9	0.01	0.12
Spectral (unsigned)	23.83	40.7	0.01	0.95
SPONGE	23.54	40.7	0.00	1.30
SSSNET	22.36	40.8	0.00	1.81
Complete linkage	22.25	49.8	0.00	0.38
Average linkage	12.09	99.3	0.55	6.90
k -means on returns	9.90	102.0	9.81	14.48
GICS	8.08	55.4	0.00	0.66
$K = 1$	1.00	500.0	0.00	0.00

Notes. This table reports the realized shape of each arm’s partition. Effective K is the inverse-Simpson index of the group-size distribution. All rows are measured on each arm’s main partition over the 6,913-day window; averaging instead over the five partition redraws moves the k -means entry from 9.90 to 9.97 and no other entry by more than 0.02. Effective K is not the label count logged with each run, and for k -means on returns the two diverge sharply: that arm carries 30 labels of which roughly 10 are singletons, whose members the trading stage removes, so the number of groups actually traded is about 20.

A.4 The Trading Head

The head maps the cluster-relative residual of Equation (4) to a weight through the chain displayed in Section 3.4, with clamp constant $c = 0.05$. The stages are as follows. Stock i enters the head on day t as the 30-day path of running sums of the cluster-relative residual Equation (4),

$$x_{it} = \left(\sum_{v=t-29}^{t-29+u} s_{iv} \right)_{u=0}^{29} \in \mathbb{R}^{30}, \quad (9)$$

a single channel, not 30 features and not several channels. Three convolutions of kernel width 2, 4, and 8 run in parallel over that path (the sense in which the extractor is multi-scale). The branches are merged by a data-driven gate. The day’s cross-sectional amplitude spectrum, averaged over names and split into three frequency bands, passes through a softmax with a learned bias, and the resulting weights pair the highest band with the narrowest kernel. The merged output passes to a long-convolution layer: a learned kernel as long as the signal window itself, soft-thresholded toward zero, applied in the frequency domain and combined with a learned skip connection, after which the score is read from the last step of the path. The sequence model is therefore a learned long convolution and not self-attention; the head computes no attention at any point. Writing g_θ for the resulting score and $f_\theta = c \tanh(g_\theta/c)$ for its soft clamp at $c = 0.05$,

$$w_{it} = \frac{f_\theta(x_{it}; \mathbf{Z}_t)}{\sum_j |f_\theta(x_{jt}; \mathbf{Z}_t)|}, \quad \hat{\theta} = \arg \max_{\theta} \frac{\mathbb{E}_{t \in \mathcal{T}} [\pi_t(\theta)]}{\sqrt{\text{Var}_{t \in \mathcal{T}} [\pi_t(\theta)]}}, \quad (10)$$

with $\mathcal{T}$ the current training window and π_t the payoff of Equation (6). The clamp binds: the cross-sectional standard deviation of g_θ is 0.05 to 0.18 across the runs of this article, median 0.10, so a one-standard-deviation score already sits in the nonlinear range of the clamp. The executed loss adds three regularization penalties to the negative of this ratio; Table 12 gives the configuration, penalties included, read from the code as executed.

Reproducibility. Four runs of the k -means-on-returns configuration at head seed 42, on the same code and panel, return Sharpe ratios of 2.8243, 2.8478, 2.8868, and 2.8940 on the common calendar, a standard deviation of 0.033 and a range of 0.07. Floating-point arithmetic on the GPU is not reproducible at that

Table 12: Trading-head configuration and training schedule

Quantity	Value
<i>Input</i>	
Signal window	30 days
Channels	1 (cumulative sum of the cluster-relative residual)
<i>Architecture</i>	
Convolution filters	8
Kernel widths (parallel)	2, 4, 8
Long-convolution kernel length	30 (the signal window)
Soft-threshold on that kernel	0.001
Trading-head hidden width	16
Dropout	0.25
<i>Objective and sizing</i>	
Loss	Negative Sharpe ratio of the book
Regularizers added to the loss	L^1 on raw bets 10^{-3} ; concentration 5×10^{-3} ; margin ranking 0.05 ^b
Soft clamp on scores	± 0.05 ($c \tanh(g/c)$, before normalization)
Normalization	$\ w\ _1 = 1$, gross exposure 1.0 ^a
<i>Walk-forward schedule</i>	
Pre-training sample	first 1,000 days
Pre-training epochs	80
Refit frequency	every 125 trading days
Fine-tuning sample	trailing 1,000 days
Fine-tuning epochs	40
Learning rate, pre-train / fine-tune	5×10^{-4} / 10^{-4}
Weight decay	5×10^{-4}
Batch size	125 days
Early-stopping patience, pre-train / fine-tune	15 / 10 epochs
Validation split	last 15% of each call

Notes. This table lists the trading head’s architecture, objective, and training schedule.

^a The L^1 step fixes gross exposure and preserves whatever long-short ratio the head produces; it does not impose dollar neutrality. Mean absolute net exposure is 0.06 to 0.18 on runs that pass the validity screen of Section 3.4 and 1.0 on those that fail it.

^b The L^1 term is the mean absolute raw bet g_θ , before the soft clamp and the L^1 normalization; the margin term also acts on g_θ . The code carries a hard position clip at ± 0.04 and a volatility-targeting scalar applied at inference before a second L^1 rescaling; the clip never binds on any run-day of this article (largest $|w_{it}|$ is 0.006), so the scalar cancels in the rescaling and both are inert. The concentration term is the Herfindahl index of the long leg plus that of the short leg; the margin term is a hinge of margin 0.01 on the agreement in sign between the predicted and the realized ordering of 64 random stock pairs per day. Gradients are clipped at norm 1.

scale, and differences within that range carry no information. Every weight file used in Section 5.4 and Section 6.4 was verified before use by reconstructing the run’s own logged payoff from it to machine precision; the exported vector’s L^1 norm was also checked against the run’s logged gross leverage and against one, on all 12 clusterizers and both controls at five seeds each.

A.5 Identifiers in the Replication Package

The clusterizers are named for what they optimize rather than by the identifiers under which they were run. The mapping, in the order of Table 1: k -means on returns is `kmeans_ret`; average and complete linkage are `agglo_avg` and `agglo_complete`; spectral (unsigned) is `spectral_unsigned`; PCA loadings is `pca_load`; SPONGE and SSSNET keep their names; Sharpe, Sortino, Sharpe (EW), and Sharpe (graph) are `sharpe_fixed`, `sortino`, `fv2a`, and `fv3sm`; GICS is `gics`. The two controls are `k1` and `random`. The two peer books of Section 6.3 are separate arms, shipped as `peer_w60` and `peer_w252`. The estimation-window variants of the clusterizers carry the window as a suffix, as in `agglo_avg_w252`.

The resolution statement of Section 3.7 is measured across a wider set than the main tables: the 12 clusterizers of Table 1, the two controls, and five window variants (average linkage, complete linkage, k -means on returns, PCA loadings, and spectral (unsigned), each at a 21-day estimation window), 19 in all. SPONGE at 60 days is one of the 12 rather than a variant; at the rule-based component it earns a Sharpe ratio of 1.59 against 1.50 for the 21-day original, so the common window does not disadvantage it. At the trained component the five 21-day variants move the mean Sharpe ratio, relative to the 60-day arms, by +0.11 (average linkage), −0.31 (complete linkage), −0.07 (k -means on returns), −0.39 (PCA loadings), and −0.36 (spectral).

A.6 The Number of Groups

K is fixed at 30 throughout the body. Five clusterizers were rerun at $K = 10$ and $K = 50$, five seeds each, everything else identical (Table 13). Three of the five rise monotonically in K ; k -means and spectral clustering peak at $K = 30$, falling by 0.09 and 0.14 from 30 to 50. The sweep does not include a random control at the new K values. A finer partition removes extra cross-sectional variance from any panel, structured or not: $(K - 1)/(N_t - 1) = 1.8\%$, 5.9% , and 9.9% at $K = 10, 30, 50$ on the 494.3 names of Section 4.2. This mechanical component of the gain therefore cannot be separated from the structural component with the controls that were run. A random partition moves 0.39 between draws of its own randomness at $K = 30$ (Section 5.2, trained component, head seed pinned) and 0.41 across the five head seeds (Figure 4), so a subtraction from one run per K would have no valid replication unit.

Table 13: Sharpe ratio by number of groups, trained component

	$K = 10$	$K = 30$	$K = 50$
k -means on returns	2.21	2.57	2.48
Average linkage	1.95	2.15	2.37
Complete linkage	2.07	2.34	2.55
PCA loadings	1.99	2.28	2.32
Spectral (unsigned)	1.69	1.74	1.61

Notes. This table reports the Sharpe ratio of five clusterizers at three values of K . Mean of per-seed Sharpe ratios, five seeds per cell, on the 6,913-day window of Section 3.1; across-seed standard deviations run 0.09 to 0.25. Differences quoted in the text are computed on unrounded means. The remaining seven clusterizers were not swept.

A.7 The Calendar Offset between the Panel and the Factor File

The factor regressions of Section 6 required fixing an error in the panel’s date labels. The labels run one trading day ahead of the factor file’s: the return the panel books under a date was realized on the previous trading day. The error surfaced as an inconsistency, an R^2 of 0.003 between the panel’s SPY return and the Fama–French market factor, impossible for two series that track the same market, and the lagged cross-correlation identifies the offset. The panel’s SPY return correlates with the Fama–French market factor at -0.06 contemporaneously and at $+0.977$ at a lag of one day. Each of the eight largest SPY moves in the panel is dated one trading day after the Fama–French date of the same move, and every panel date is a Fama–French trading day, so the offset is in the labels, not the holidays.

The two trading components also label their payoffs differently. The rule-based component books a payoff under the day the position was formed; the trained component books it under the label of the day it was realized. With the panel’s labels one day ahead, the same calendar day’s payoff is matched to the factor file same-day at the rule-based component and with a one-day lag at the trained component. Every regression in this paper pairs each payoff with the factor return of the day on which it was realized. Joining either trading component at the wrong phase collapses R^2 to below 0.01 and books the factor premium into alpha: misalignment hides exposure rather than creating it.

The offset changes no internal statistic: Sharpe ratios, volatilities, turnover, and everything in Section 6.1 relabels uniformly. The only affected objects are regressions against external data, so it went unnoticed until one was inconsistent. We have not determined whether the labels were shifted by the data vendor or by a lag convention in the panel’s construction. The offset is recorded and corrected.

A.8 Degrading One Partition toward a Random One

The 12 clusterizers supply only three levels of comovement removal, but intermediate levels can be constructed. We take one partition draw of the PCA-loadings clusterizer and replace a fraction q of each day’s labels with uniform random ones, so the partition degrades continuously from the real one ($q = 0$) toward a random one ($q = 1$). Figure 11 displays the result: the removal share falls from 19.1% to 5.88% and book volatility rises from 5.19% to 8.84%, monotonically at every step. Full corruption reaches the floor to three decimal places, and the two endpoints coincide with the real arm and with the random control. The volatility reduction is a continuous function of how much real comovement the labels carry.

A.9 The Estimation Window

Section 5.5 finds that a longer estimation window raises the Sharpe ratio of spectral clustering and of PCA loadings and leaves average linkage where it was. Random matrix theory supplies a reason the two clusterizers that gain should be the sensitive ones: both read individual eigenvectors of the sample correlation matrix, and at $N = 500$ a population eigenvalue must exceed $1 + \sqrt{N/T}$ to be recoverable at all (Baik et al., 2005). The threshold is 5.9 at $T = 21$, under which only the market mode survives, against 2.4 at $T = 252$, under which sector-scale structure does. A rival explanation, however, fits the same six observations, three clusterizers at two windows. The largest gains go to the clusterizers that started lowest, and a regression of the gain on the 21-day starting level fits with $r = -0.93$. The two explanations make the same prediction on every arm in the sweep. We therefore registered, in advance, three arms on which they differ by construction, and ran them. Average linkage at $K = 10$ starts low, so the rival predicts a large gain and the eigenvector account almost none; the observed gain is $+0.04$. The random control starts lowest of all, so the rival must predict the largest gain in the design, while the labels are drawn by a random number generator that cannot read the window; the observed gain is -0.11 , negative in four of the five seeds. PCA loadings at $K = 50$ reads eigenvectors, so both accounts predict a gain; the observed gain is $+0.33$. The eigenvector account survives, since it predicts which arms move and in which direction, and the two arms that use no eigenvector do not move. The account misses the sign of the interaction with K : for PCA loadings the same 60-to-252-day move is worth $+0.27$ at $K = 30$ against $+0.33$ at $K = 50$, so the window effect strengthens as the partition is refined. We have no account of that.

Figure 12 shows the 60-to-21-day step of the same sweep against seed noise. That step contains the account’s one exception: complete linkage, which reads the same correlation matrix as average linkage and no eigenvector, loses 0.31, outside its across-seed band. No p -value is quoted for the window responses: a paired

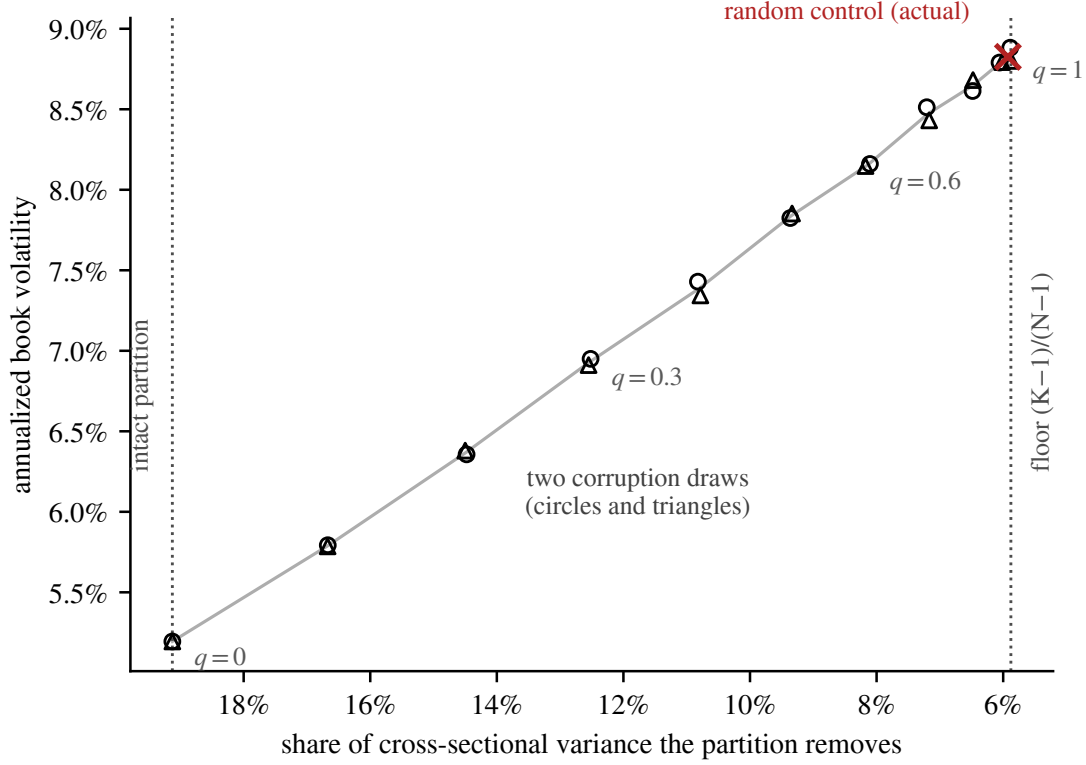


Figure 11: Corrupting a real partition, one fraction of labels at a time. This figure plots book volatility against the share of cross-sectional variance the corrupted partition removes. Each point is a full 6,913-day rule-based backtest of the PCA-loadings partition with a fraction q of each day’s labels redrawn uniformly; two independent corruption draws per q nearly coincide. The dotted verticals mark the degrees-of-freedom floor $(K - 1)/(N - 1)$ and the intact partition’s share. The cross is the random control itself.

test across five fits rests on four degrees of freedom, so each step is judged against its arm’s own across-seed band instead. Average linkage, which moves 0.00 over the whole sweep (Section 5.5), peaks at 120 days, mid-sweep.

A.10 Robustness of the Cross-Sectional Projection

The measurement of Section 6.1 depends on how many principal components to project out and which floor to judge the result against. Estimating the number of factors is a literature of its own (Bai and Ng, 2002; Onatski, 2010; Ahn and Horenstein, 2013), and we do not enter it. At $m = 1, 3, 10$ the k -means partition retains 83%, 68%, 50%, and every real partition keeps a strictly positive excess over its floor at every m tested: the sign of the result is robust to m ; the magnitude is not. The floor must also be matched to the partition’s own shape. Net of it, retention at five components is 28% to 49% across the 12 clusterizers, and GICS, with 10 realized groups on a 322-name universe, has a floor of 0.028, not the 0.060 of Section 6.1. Against the wrong floor its net retention at 10 components would come out negative, an artifact of over-subtraction. Two details of the protocol qualify these numbers. The retention ratio divides two R^2 s with different denominators: the share of the post-projection remainder is not a variance decomposition of the original cross-section. The measurement here also counts every group, singletons included, each of which fits its own member exactly. Under the trading convention of Section 3.3, which deletes singleton members, k -means on returns measures 0.264 on the traded universe with singletons counted (Section 4.2) and 0.216 without them, against the 0.265 here.

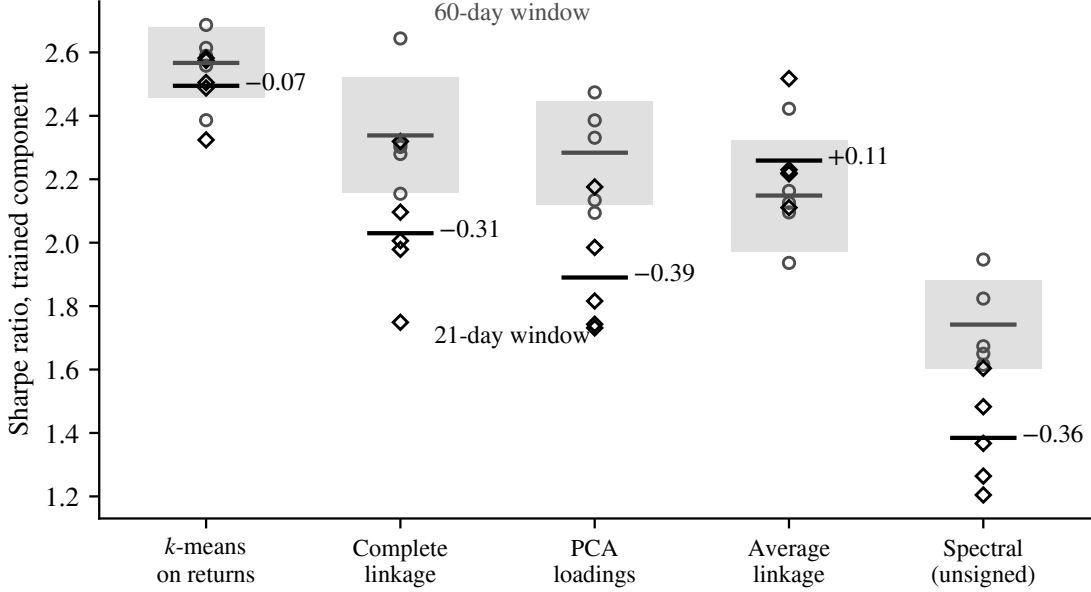


Figure 12: Changing only the estimation window. This figure plots the per-seed Sharpe ratio at the trained component of the 60-day arm (circles) against its 21-day variant (diamonds); the shaded band spans ± 1 across-seed standard deviation of the 60-day arm, and the horizontal lines are the two windows’ seed means. All five 21-day variants of Appendix A.5 are shown. The two clusterizers that read individual eigenvector coordinates lose 0.39 and 0.36, about two and a half times their own band, and complete linkage loses 0.31; *k*-means on returns moves -0.07 and average linkage $+0.11$, both inside their own band.

A.11 Size of the Spanning Tests

The t -statistics of the FF5+MOM block of Table 7 divide by a Newey–West long-run variance at 21 lags, and with heavy tails and serial dependence a test of this form can reject more often than its nominal level. We measure its size directly. For each of the 28 books, the null distribution is rebuilt from block-resampled residuals (21-day blocks, 5,000 replications per book under each of four independent seeds), with the factor matrix held at its sample values and the betas re-estimated on every draw. The empirical size at nominal 5% is 6.6% pooled across the books and 5.3% to 7.6% book by book. The spread comes from the denominator: dependence cumulated over the 21 lags is about 1.7 times as large in the trained books, so the trained books over-reject more. Correcting the size moves the critical value from 1.96 to 2.2 in t units even at the worst-behaved book, and the smallest t in that block is 4.1 (2.5 in the block with the reversal factor), so the rejections are not heavy-tail or dependence artifacts.

A.12 Distance to the Peer Books and Performance

The share of each clusterizer’s payoff variance the two peer books explain runs from 0.52 to 0.57, with a positive relation to the clusterizer’s Sharpe ratio that falls just short of the 5% level (Spearman $\rho = 0.57$, $p = 0.051$). The overlap between a clusterizer’s positions and the peer books’ runs from 0.43 to 0.48, or 0.60 to 0.65 of what that clusterizer shares with its own refit, with a positive relation to its Sharpe ratio that does not reach significance at the 5% level ($\rho = 0.56$, $p = 0.06$). The two peer books share 0.49 of their positions with each other, about as much as either shares with a clusterizer.

A.13 Transaction Costs in Detail

Costs are charged on the positions the runs actually held, so the turnover below is measured from those positions rather than inferred from the payoffs (Appendix A.4); because the L^1 normalization pins the gross book to one (Section 3.4), a position is a share of it. Table 14 reports turnover, the one-way cost at which

Table 14: Turnover, break-even cost, and net Sharpe ratios

	Gross	Return	Turnover	Break-even	Net Sharpe ratio		
	Sharpe	(%/yr)	(%/day)	(bps)	1 bp	2 bps	5 bps
12 clusterizers	1.74–2.57	6.62–7.60	38.2–43.8	3.17–3.79	1.20–1.86	0.67–1.16	–1.24 to –0.67
$K = 1$	1.28	6.47	44.6	2.90	0.83	0.38	–0.96
Random	1.28	6.25	46.9	2.66	0.79	0.31	–1.15

Notes. This table reports each book’s gross Sharpe ratio, annual return, one-way turnover, break-even cost, and net Sharpe ratio at three cost levels. Trained component, five head seeds, each statistic computed per seed and averaged, on the 6,913-day window of Section 3.1 (July 1997 to December 2024). Positions are the weight vectors the runs produced, at gross exposure one. Turnover is one-way, half of $\sum_i |\Delta w_{it}|$; the break-even and the net columns charge the full traded notional at the stated one-way rate. Break-even is the one-way rate, in basis points (bps), at which annual return reaches zero. All arms use the common 60-day estimation window of Table 1; the bands cover the 12 clusterizers, and the two controls are reported on their own rows.

each book’s annual return reaches zero, and net Sharpe ratios at three cost levels. Gross returns run 6.3% to 7.6% a year. At five basis points one-way the same books pay 9.6% to 11.8% a year in cost. Because gross return and cost scale together, the break-even does not depend on leverage.

The reordering reported in Section 6.4 is not an artifact of the trading head. We run the same measurement on the rule-based component, whose book carries gross exposure two rather than one, and normalize both to the same exposure. Its five-day signal window is rebuilt daily, and its positions correlate with the previous day’s at 0.46 on average across the clusterizers, with 27% of names changing sign. Counted one-way and per unit of gross exposure, the rule-based books turn over 23.1% to 30.8% of their exposure a day against the head’s 38.2% to 46.9%, and earn 3.9% to 4.6% a year against the head’s 6.3% to 7.6%. The head trades 1.46 times as much per unit of exposure and earns 1.63 times as much, and the two nearly cancel: break-even at the rule-based component is 2.75 to 3.74 basis points against 2.66 to 3.79 at the trained component. The reordering is sharper without the head: on the rule-based component the two controls finish first and third at five basis points, against ninth and thirteenth at the trained component.

The nearest published comparison finds a much larger turnover reduction from clustering, and we cannot reconcile the two. Brutti et al. (2026) report that clustering cuts turnover sharply in their setup, from 87.9% to between 24.2% and 24.4%, and charge no costs; our configurations show a far smaller gap, the $K = 1$ control turning over 44.6% against a clustered range of 38.2% to 43.8%. Rebalancing frequency is the likely cause (our books are rebuilt every day), but we have not measured it, and we record the divergence as an open question rather than a finding.

Five limitations apply. The cost model is a proportional one-way charge with no market-impact term, and a book rebuilt daily on 500 names is where market impact would be largest. Borrow costs on the short side are not modeled. The panel is US large capitalizations, the most liquid case available, so the break-even here is an upper bound for the same design on a broader universe. The cost figures are at $K = 30$ only; whether the ordering above survives the sweep of Appendix A.6 is unmeasured. Finally, the books are rebuilt daily, and a longer holding period would lower turnover roughly in proportion; its effect on the break-even is left to future work.