Factor Clustering with t-SNE*

Philip Greengard Yukun Liu Stefan Steinerberger Aleh Tsyvinski

September 20, 2020

Abstract

We cluster asset pricing factors using the t-distributed Stochastic Neighborhood Embedding (t-SNE), one of the most empirically successful dimensionality reduction techniques. t-SNE endogenously separates the strategies into six distinct clusters. The first five clusters resemble the standard value, momentum, investment, profitability, and volatility strategies. The sixth cluster is new, and we denote it as the Firm cluster. We show that the first five clusters are low dimensional and are dominated by their corresponding first principal components, while the Firm cluster is intrinsically high dimensional.

*Philip Greengard is with Columbia University. Yukun Liu is with the University of Rochester, Simon Business School. Stefan Steinerberger is with the University of Washington. Aleh Tsyvinski is with Yale University. We thank Stefano Giglio and Stefan Nagel for helpful comments.

1 Introduction

The asset pricing literature has spent much effort to search for factors that can explain the cross-section of expected stock returns. This search lead to tens if not hundreds of potential candidates.[1] Bringing structure to the existing factors is thus critical. In this paper, we discipline the existing factors by grouping them into a small number of clusters with the t-distributed Stochastic Neighborhood Embedding (t-SNE) method.

t-SNE, introduced by van der Maaten and Hinton (2008), is a cutting-edge non-linear dimensionality reduction technique that has shown tremendous success in a variety of areas of natural sciences. Traditional approaches to analysis and visualization often fail in the high dimensional setting. Therefore, it is common to perform dimensionality reduction of datasets in order to make data analysis tractable. t-SNE is perhaps the most popular and successful method to produce two-dimensional embeddings of high dimensional data with the goal of simplifying the identification of clusters.

The basic idea of this paper is to view the daily returns of each strategy as its return profile and compare the return profiles of different strategies. Because a strategy has tens of thousands of historical daily returns, clustering the return profiles of the different high dimensional return profiles is a classic dimensionality reduction problem and an ideal setting for t-SNE.

We now briefly discuss the mathematical problem we are facing, the underpinning behind the nonlinear t-SNE clustering method, and the variations that we introduce to the method. We first note that the classic geometric techniques for organizing data become less useful in high dimensional space—a phenomenon known as the *concentration of measure* problem (Ledoux, 2001). The main issue is that various measures of distance become less informative as the number of dimensions increases. t-SNE tries to avoid working with the underlying geometry in the original high-dimensional space and extracts only information about nearest neighbors to limit the effect of the concentration of measure phenomenon. The idea behind t-SNE can be explained in simple terms: we are given a set of high dimensional vectors and are interested in obtaining a faithful two-dimensional representation as much as this is possible. Here, faithful means that if the original vectors are close to each other, then so are their two-dimensional representation. Conversely, if the original vectors are far apart from each other, then so are their two-dimensional representation. The broad idea of the method is that in the low-dimensional space it attracts the points that are close or similar in the high dimensional space, and repels the points that are far from each other. For a general problem it is, of course, impossible to recreate high-dimensional geometry in two dimensions. However, in a variety of empirical applications in a number of fields of natural sciences, t-SNE is able to successfully capture the structure of the high-dimensional data. Importantly, Linderman and Steinerberger (2019) provide rigorous mathematical foundation for the use of t-SNE as a clustering method. They prove that if the high-dimensional data is formed by the well-defined clusters, t-SNE

[1] See McLean and Pontiff (2016), Feng, Giglio, and Xiu (2019), Hou, Xue, and Zhang (2020), and Kozak, Nagel, and Santosh (2020).

is able to recover them. Moreover, we use a modification of t-SNE proposed by Kobak, Linderman, Steinerberger, Kluger, and Berens (2019) that allows to find a finer cluster structure compared to the traditional t-SNE.

We apply the t-SNE method to a broad set of stock-specific characteristic-sorted long-short strategies in Kozak, Nagel, and Santosh (2020). We show that six distinct clusters endogenously emerge from this process. The first five clusters resemble the familiar value, momentum, investment, profitability, and volatility clusters.

The Value cluster captures all the fundamental-to-market value ratio strategies (e.g., cash flow-to-market value, dividend yield, sales-to-price, book-to-market, and book-to-market monthly) except for the earnings-to-price ratio. The Value cluster also captures another measure that is commonly considered to be a type of a value strategy—the long-term reversal strategy. Several strategies that are not typically identified in the literature as value strategies are categorized as such by t-SNE: duration and leverage strategies. A mixed strategy, momentum-reversal, is also categorized into the Value cluster. The Momentum cluster contains all the classic momentum strategies—industry momentum, momentum (6m), and momentum (1y). Two of the mixed strategies—value-momentum and value-momentum-profitability—are also identified in the Momentum cluster.

The t-SNE procedure also cleanly separates an Investment cluster. The Investment cluster captures most of the commonly believed investment-related strategies, including asset growth, investment growth, investment, and sales growth. It is notable that investment-to-capital, which is commonly regarded as an investment strategy, is excluded from the Investment cluster. The Profitability cluster includes five classic profitability-related anomalies, including return-to-asset monthly, return-to-asset annual, return-to-book equity monthly, return-to-book equity annual, and return-to-market equity. The Profitability cluster also includes the earnings-to-price and price strategies. Interestingly, one of the classic profitability measures, gross profitability, is not in the cluster. The Volatility cluster contains all the classic volatility-related strategies, including beta arbitrage, idiosyncratic volatility, and share volume. The cluster also includes investment-to-capital, which is commonly considered an investment strategy.

The sixth cluster has not been identified in the literature. We label this group the Firm cluster because it contains many strategies that relate to companies' operations. The cluster contains many strategies that relate to companies' operating activities. For example, the cluster includes strategies based on accounting practices (Accruals), asset turnover speed (Aturnover), debt issuance (Debtiss), changes in dividend policies (Divg), and share repurchasing decisions (Repurch). The cluster also includes several operation measures, including Piostroski's F-score (Fscore), growth in LTNOA (Gltnoa), gross margin (Gmargins), net operating assets (Noa), and gross profitability (Prof). Lastly, the Firm cluster contains a measure of the seasonality of the firm (Seaon).

Moreover, t-SNE allows us to visualize the relations of the elements in two-dimensional space as the six distinct clusters. Strategies within a given cluster are close to each other and strategies in different clusters cleanly separate from each other in the two-dimensional space.

We further compare the t-SNE procedure with other methods in separating and clustering objects that were used in finance. The other methods we consider include both linear methods, such as principal component analysis, and nonlinear methods, such as K-means and a kernel principal component analysis. We also compare to the results of another nonlinear method—the spectral clustering technique that has not been widely used in the finance literature so far. Overall, we show that t-SNE outperforms standard applications of these methods in separating the broad set of factors.

Next, we investigate the statistical properties of each cluster. We first calculate the principal components of each group. We show that, for the first five clusters, their corresponding first principal components already explains a large fraction of the variations within the groups. The first principal components can already explain 51.7 percent, 70.7 percent, 63.0 percent, 63.0 percent, and 73.4 percent of the variations of the strategy returns in the Value, Momentum, Investment, Profitability, and Volatility clusters, respectively. The decreases in magnitude from the first eigenvalue to the next eigenvalue are large for these five clusters. However, the first principal component only explains 22.8 percent of the variations of the strategy returns in the Firm cluster. The second and third principal components explain 15.9 percent and 14.2 percent of the return variations, respectively. These results show that the first five clusters have dominating principal components, but the Firm cluster is intrinsically high dimensional. We reach similar conclusions based on the results of explaining individual long-short strategies using the first principal components of their corresponding clusters.

The high dimensionality of the Firm cluster leads us to examine the cross-sectional pricing power of the principal components of the Firm cluster. We sort the cross-section of stocks based on the return exposures to each of the principal components of the Firm cluster. We find that exposures to the first principal component command sizable risk premia, while exposures to the other principal components do not. Therefore, although the Firm cluster is intrinsically high dimensional, its pricing power comes from the first principal component, allowing us to summarize its pricing relevant information with one factor.

Furthermore, we regress each long-short strategy on the first principal components of all the groups. The goal of the exercise is to determine whether the additional principal components of other clusters explain much of the return variation in the individual long-short strategies. For all of the six clusters, the model using all the first principal components do not substantially increase the explanatory power over using only the first principal component of the corresponding clusters. The increases in R-squareds are less than 10 percent for eight of the nine strategies in the Value cluster, for four of the five strategies in the Momentum cluster, for all the strategies in the Investment cluster, for five of the seven strategies in the Profitability cluster, for half of the volatility cluster, and for seven of the eleven strategies in the Firm cluster.

We also test whether including the various factor models proposed in the asset pricing literature can substantially increase the explanatory power above using only the corresponding first principal components of the clusters. The factor models we consider include the CAPM, Fama-French 3-

factor, Carhart 4-factor, Fama-French 5-factor, Fama-French 6-factor, q-factor, and the Daniel-Hirshleifer-Sun behavioral factor models. For all the six clusters, the various factor models do not substantially increase the R-squareds when they are included with the first principal components of the corresponding clusters. Based on the Fama-French 5-factor model, the increases in R-squareds are less than 10 percent for eight of the nine strategies in the Value cluster, for four of the five strategies in the Momentum cluster, for all the strategies in the Investment cluster, for five of the seven strategies in the Profitability cluster, for half of the volatility cluster, and for seven of the eleven strategies in the Firm cluster.

This paper relates to several strands of the literature. First, our paper is most related to the recent literature on exploring the high dimensionality of cross-sectional asset returns. McLean and Pontiff (2016) show that performances of the discovered anomalies tend to worsen post-publications. Harvey, Liu, and Zhu (2016) use a multiple testing framework common in medical sciences to evaluate existing asset pricing factors and argue that a higher hurdle is needed for future research. Kozak, Nagel, and Santosh (2020) find that, although a small number of characteristics-based factors cannot adequately summarize the cross-section of expected stock returns, the principal components of the universe of these potential characteristics-based factors approximate the stochastic discount factor well. Freyberger, Neuhierl, and Weber (2020) use a group LASSO method to choose characteristics and nonparametrically estimate their contribution to the expected returns. Feng, Giglio, and Xiu (2019) propose a model selection method to evaluate the contribution of any new factor. Lettau and Pelger (2020b) propose a new method to estimate latent asset pricing factors that fit both the time-series and the cross-section of expected returns. Our paper differs from the existing literature in that we focus on clustering the existing factors into a small number of groups.

Our paper is also related to the finance literature in grouping and clustering. For example, researchers have attempted to group companies into different industries (e.g., Bhojraj and Lee, 2002; Bhojraj, Lee, and Oler, 2003; Hoberg and Phillips (2016)), locations (e.g., Garcia and Norli, 2012), and labor markets (e.g., Liu and Wu, 2020). Ludvigson and Ng (2007) and Ludvigson and Ng (2009) group a broad set of macroeconomic variables into several categories based on prior information. K-means is a popular clustering method in the literature. Brown and Goetzmann (1997) and Brown and Goetzmann (2003) use K-means clustering methods to group mutual funds and hedge funds into broad categories, respectively. A key input to K-means clustering is the number of clusters to employ. Several selection methods have been proposed in the literature, including the "gap" statistic of Tibshirani, Walther, and Hastie (2001), the information criteria of Fraley and Raftery (2002), cross-validation methods of Tibshirani and Walther (2005), and the general "T/2" test of Patton and Weller (2019). There is also a growing literature that aims to group high dimensional textual data using machine learning methods in economics. For a survey of the literature, see Gentzkow, Kelly, and Taddy (2019).

The rest of the paper is structured as follows. Section 2 discusses the methodology we use. Section 3 presents the t-SNE groups and compare them with other methods. Section 4 examines the properties of the t-SNE groups. Section 5 presents additional results and Section 6 concludes

the paper.

2 Methodology

The purpose of this section is to discuss the mathematical problem that we are facing, the underpinning behind the t-SNE clustering method that we study, and the variations that we introduce on it.

2.1 The Problem

The problem can be summarized as follows: let $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a set of n points in $\mathbb{R}^d$ dimensions. These are n data points in d dimensions, we can think of the outcome of the returns of $n = 55$ long-short strategies on $d = 11328$ days – the number of trading days in the sample. That is, we have much fewer strategies, than even the information contained in a single strategy.

Observe that even representing a basic shape like the boundary points of a unit cube $[0,1]^d$ does already require $n = 2^d$ points. To make things more complicated, classical Euclidean distance, the way we measure distances between two points in $\mathbb{R}^d$ dimensions via Pythagoras' theorem

$$\|x - y\| = \left(\sum_{k=1}^{d} (x_k - y_d)^2 \right)^{1/2} \tag{1}$$

becomes a lot less informative in high dimensions. This is a subtle phenomenon that was only properly understood in the 1970s and is now known as the *concentration of measure* phenomenon (Ledoux, 2001). We illustrate it with a simple example: we take $n = 100$ points in $\mathbb{R}^2$ (uniformly and identically at random) in Figure 1.

We observe that the points are fairly regularly distributed. Moreover, as may seem tautological, some points are close to each other than others. In fact, the maximum distance between any pair of points in the figure is 1.214 and the minimum distance between any pair of points is 0.012. The mean distance is 0.528 with standard deviation 0.246.

We now repeat the same experiment with $n = 100$ points chosen uniformly at random in $[0,1]^{10000}$. The minimum distance between any two points is 40.005, the maximum distance is 41.643, the mean is 40.833 and the standard deviation is 0.2422, very close to the earlier value. Even though we have a lot more information about each point (10000 coordinates instead of 2), it actually becomes harder to distinguish them using nothing but the Euclidean distance. That is, as the dimension of the data increases Euclidean distance becomes less informative. A similar problem is present with using correlation in high dimensions to measure distance. One can show that two random vectors from the unit cube are basically decorrelated when the dimension is large (see e.g. Steinerberger, 2011).

We have seen that in the problem of trying to understand the distribution of points, classic geometric techniques becomes less useful: the Euclidean distance scales in an unfortunate way. This

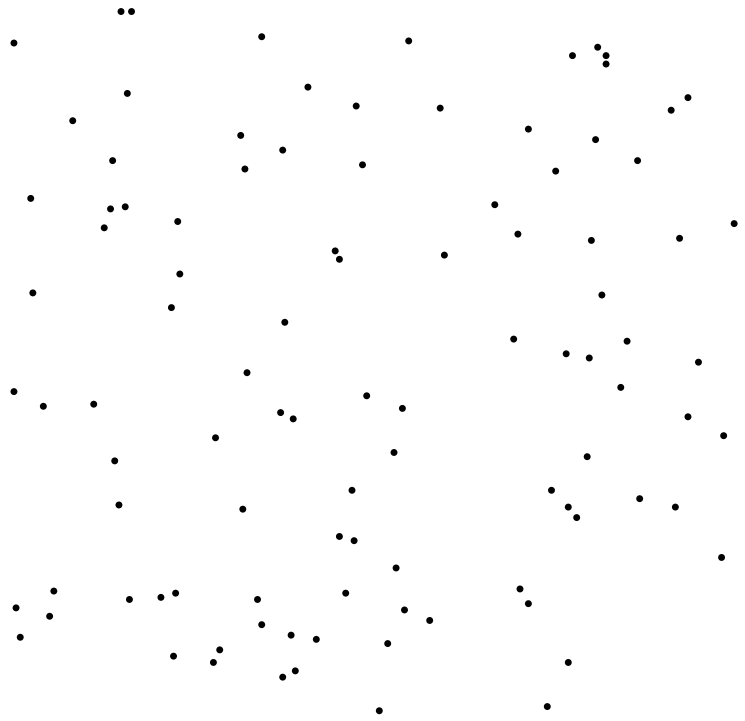
is a problem affecting all of data science and clearly shows the necessity of a different approach. Many of the leading techniques, including the one that we use in this paper, try to avoid working with the underlying geometry in the original space and reduce things very quickly to a more abstract setting: that of a graph $G = (V, E)$. Here, $V = \{i : 1 \leq i \leq n\}$ is the original data set which is now used to describe the vertices of a graph. These vertices are connected by $E \subseteq V \times V$, where

$$E = \left\{(i, j) : \text{ if } x_i \text{ and } x_j \text{ are close}\right\}, \tag{2}$$

where "close" depends on the method. It is common to connect each point to its k nearest neighbors (i.e. k points having the smallest Euclidean distance) where k is a parameter.

Figure 1: 100 points chosen uniformly at random from $[0,1]^2$

This figure plots $n = 100$ points in $\mathbb{R}^2$ (uniformly and identically at random).



2.2 t-SNE

The idea behind t-SNE can be explained in simple terms: we are given a set $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and are interested in obtaining a faithful two-dimensional representation $\{y_1, \ldots, y_n\} \subset \mathbb{R}^2$ as much as this is possible. Here, faithful means that if x_i and x_j are close to each other, then so are y_i and y_j and, conversely, if x_i and x_j are far apart from each other, then so are y_i and y_j. The broad idea

of the method is that in the low-dimensional space it attracts the points that are close or similar in the high dimensional space, and repels the points that are far from each other.

2.2.1 Summary of the t-SNE (van der Maaten and Hinton, 2008)

1. We create a probability distribution $P = (p_{ij})_{i,j=1}^{n}$ on the set $\{x_1, \ldots, x_n\} \times \{x_1, \ldots, x_n\}$. Essentially, the number p_{ij} is large if x_i and x_j are close to each other and small otherwise:

$$p_{ij} = \frac{p_{\{j|i\}} + p_{\{i|j\}}}{2} \tag{3}$$

where

$$p_{i|j} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{\{k \neq l} \exp(-\|x_k - x_l\|^2/2\sigma_i^2)}. \tag{4}$$

Since dense regions of data require smaller values of σ_i than more sparsely concentrated regions, the value of σ_i is determined, for each x_i, according to the density of points in the region around x_i. As a consequence of setting σ_i separately for each x_i, we are able to deal with a data set that consists of, for example, two well-separated regions of points, one dense and the other more sparse. For this sort of data set, σ_i will be larger for x_i is the sparse region than x_j in the dense region. As a result, our probability distribution P will consider the two separated regions to be far from each other, but points within the dense region will be close to each other and points within the sparse region will also be close to one another.

2. We then start with a random set of points $\{y_1, \ldots, y_n\} \subset \mathbb{R}^2$.

3. Our goal is to have the set of y_i in the two-dimensional space behave as much as possible as the set of x_i in the high-dimensional space. We create, for any given set $\{y_1, \ldots, y_n\}$ a probability distribution $Q = (q_{ij})_{i,j=1}^{n}$ on $\{1, \ldots, n\} \times \{1, \ldots, n\}$. Again, q_{ij} should be large if and only if y_i and y_j are close to each other in two dimensions. The original stochastic neighborhood embedding (SNE) method (Hinton and Roweis, 2003) used the Gaussian kernel $k(l) = e^{-l^2}$, where l is the Euclidean distance between two points in the low-dimensional space. The main issue with this method is the so called "crowding" problem. It is difficult to represent the points with the moderate distance in the high-dimensional space in the low-dimensional space if one wants to preserve the small distances. In simple terms, objects are being mapped into the correct clusters but the clusters are too close to one another to allow for an easy distinction where one cluster starts and the other one ends. The t-SNE method instead proposes to use the heavier-tailed t-distribution $k(l) = 1/(1 + l^2)$ that allows the distance in the low-dimensional space to decrease more slowly.

4. We now proceed as follows: we introduce a measure of similarity between two probability distributions P and Q: the *Kullback-Leibler divergence*

$$D_{KL}(P||Q) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \tag{5}$$

This notion is a classical proper notion of distances between probability distributions. This notion is motivated by information theory and intuitively tells that if you know one distribution, how many bits of information are needed to update your knowledge to the other. It also has the following crucial property, Gibbs inequality: $D_{KL}(P||Q) \geq 0$ with equality if and only if $P = Q$.

5. The probability distribution P is given by the data and fixed. The probability distribution Q is generated by $\{y_1, \ldots, y_n\}$ and something that we change.

6. We now start moving the points $\{y_1, \ldots, y_n\}$ around in such a way that the KL-divergence between P and Q decreases: we do this by computing the gradient of the Kullback-Leibler divergence with respect to each y_i

$$\nabla|_{y_i} D_{KL}(P||Q) \tag{6}$$

and then use this to do a gradient descent.

7. After convergence, we are left with a set of two-dimensional points, $\{y_1, \ldots, y_n\} \subset \mathbb{R}^2$ such that if their corresponding high-dimensional points x_i and x_j are also close together, then y_i and y_j are close to each other. Our final step is to cluster those points $\{y_1, \ldots, y_n\} \subset \mathbb{R}^2$. We do this using DBSCAN (Ester, Kriegel, Sander, and Xu, 1996), which works by assigning points y_i and y_j to the same cluster if they are connected by sufficiently dense regions of points. In particular, DBSCAN has two inputs n and ϵ. For each point y_i, DBSCAN evaluates the number of points in $\{y_1, \ldots, y_n\}$ that are within distance ϵ of y_i. If there are n such points, then y_i and all neighbors within distance ϵ are put into the same cluster.

We conclude this section noting that while there is tremendous empirical success of t-SNE in a variety of fields, the mathematical theory behind it is in its nascence. Shaham and Steinerberger (2017) is the first paper to provide such foundation and shows that the optimal SNE embedding separates the well-defined clusters. Their results are also broadly applicable to the family of other methods including t-SNE. Linderman and Steinerberger (2019) demonstrate that t-SNE can be rigorously analyzed and prove the connection of t-SNE to fundamental ideas in partial differential equation. Importantly, they give conditions under which t-SNE provably recovers well-separated clusters. Of course, this is not always feasible as high-dimensional structures may contain complexities that cannot be represented in low dimensions. However, if the high-dimensional data has the underlying defined cluster structure, t-SNE is able to separate them.

2.2.2 Our Approach

We deviate from the classical t-SNE approach in the following important way. The first mathematical analysis of the t-SNE algorithm Linderman and Steinerberger (2019) showed that one can also interpret it as a dynamical many-body particle system: essentially we have n particles in the plane where each particle is being attracted to all their particles with high similarity in high

dimensions while being repelled by all other particles else. This interpretation, however, invites to study different types of attraction-repulsion potentials. Kobak, Linderman, Steinerberger, Kluger, and Berens (2019) proposed to use the kernel $k(l) = \frac{1}{\left(1+l^2/\alpha\right)^{\alpha}}$ that corresponds to the scaled t-distribution with $\nu = 2\alpha - 1$. In particular, this kernel allows to use $\alpha \in (0, 1/2)$ corresponding to negative ν which have tails that are heavier than any t-distribution. They show that this allows to better separate the clusters and reveal the finer, or "hidden", cluster structure. We follow this methodology with $\alpha = 0.3$.

2.2.3 Comparing with Other Techniques

Given the importance of the problem, unsurprisingly, there is a very large number of techniques for dimensionality reduction and, presumably, many of them could be used to study the finance data under consideration here. t-SNE is special for a variety of reasons: it is relatively easy to use and does not require a careful selection of many parameters. It is also agnostic towards the data and does not need expert knowledge (in comparison, for example, to k-means which requires to have a good understanding of the number of clusters to expect). Additionally, t-SNE is able to represent outputs in two dimensions thus making it easy to visualize the structure of the data. Many classical dimensionality reduction techniques allow for a reduction of the dimensionality of the data by many orders of magnitude but not, typically, down to two dimensions. The Johnson-Lindenstrauss method, for example, will be able to reduce dimensionality but often to a relatively large number of dimensions. Therefore, while the complexity of the data has been tremendously reduced, one does frequently end up with data in 10-100 dimensions which then requires further analysis. t-SNE is very different insofar as it was constructed from the very beginning to result in a two-dimensional embedding. Indeed, it was initially conceived as a visualization technique, a tool that allows one to look at abstract high-dimensional data, however, it started being used as a clustering technique which, as discussed by Linderman and Steinerberger (2019), is in accordance with the underlying theoretical mechanism (where t-SNE can be understood as a spectral method with an additional repulsion term). The difficulty of embedding a high-dimensional structure into only two dimensions is counteracted by having a highly adaptive nonlinear energy functional that actively promotes separation despite the measure concentration phenomenon.

There are a variety of other techniques for dimensionality reduction. van der Maaten and Hinton (2008) discuss the relationship of them to t-SNE. Linear techniques such as PCA aim on keeping the low-dimensional representation of dissimilar points apart. When the high dimensional data lies on the low dimensional, nonlinear manifold, t-SNE aims to keep similar points to represent this low-dimensional structure which is difficult with the linear techniques. A variety of nonlinear methods have also been proposed.[2] While many of them perform well on the artificial datasets, t-SNE achieved an impressive record in dimensionality reduction in a variety of fields due to its ability to retain both the local and the global structure of complex high-dimensional data and often signif-

[2]See Lee and Verleysen (2007) and van der Maaten, Postma, and van den Herik (2009) for a comprehensive survey.

icantly outperforms other methods (van der Maaten and Hinton, 2008; van der Maaten, Postma, and van den Herik, 2009). In the Appendix, we further discuss the properties and performance of t-SNE, in comparison with PCA.

3 Empirical Results

3.1 Data

We consider a broad set of stock-specific characteristic-sorted long-short strategies in Kozak, Nagel, and Santosh (2020).[3] Kozak, Nagel, and Santosh (2020) show that this set of portfolios effectively summarizes heterogeneity in expected returns. They find that, although a small number of characteristics-based factors cannot adequately summarize the cross-section of expected stock returns, the principal components of the universe of these potential characteristics-based factors approximate the stochastic discount factor well. We examine portfolios rather than individual stocks to reduce the effect of idiosyncratic component in the individual stock returns. We focus on long-short strategies to approximately remove the strong market common component.

We briefly describe the construction of the strategies. There are 55 long-short strategies in total. Each strategy is constructed from the universe of CRSP and COMPUSTAT stocks, where individual stocks are sorted into value-weighted portfolios. The portfolios include stocks from all major stock exchanges (NYSE, AMEX, and NASDAQ) but the portfolio breakpoints are based only on NYSE firms as in Fama and French (2016). The long-short strategies are at the daily frequency and range from February 03, 1975 to December 31, 2019.[4] The long- and short-sides of the strategies are determined based on the original papers.

Table 1 shows the average returns of each of the long-short strategies. Consistent with the asset pricing anomaly literature, most of the long-short strategies exhibit positive and significant excess returns and CAPM alphas. In Section 5, we also show results based on the monthly frequency.

3.2 t-SNE

In this section, we first describe the results generated from the t-SNE procedure described in Section 2. We apply the t-SNE procedure to the daily returns of the long-short strategies discussed above.

[3]The data are available at www.serhiykozak.com/data. Other papers that tackle the factor zoo include Bryzgalova (2015), Lettau and Pelger (2020b), Feng, Giglio, and Xiu (2019), and Freyberger, Neuhierl, and Weber (2020).

[4]February 03, 1975 is the first day that all long-short strategies are available.

Table 1: Strategy Description

This table reports the summary statistics of the 55 long-short strategies in Kozak, Nagel, and Santosh (2020). The average strategy returns and the CAPM alphas are reported. The numbers correspond to the t-SNE group numbers. *, **, *** denote significance levels at the 10%, 5%, and 1% based on the standard t-statistics.

Num	Strategy		$\overline{Ret}$	α	Num	Strategy		$\overline{Ret}$	α
1	Cfp	Cash Flow-Market Value	0.016**	0.021***	5	Shvol	Share Volume	-0.005	0.021**
1	Divp	Dividend Yield	0.010	0.020**	6	Accruals	Accruals	0.018***	0.016***
1	Dur	Cash Flow Duration	0.019**	0.021**	6	Aturnover	Asset Turnover	0.020***	0.021***
1	Lev	Leverage	0.009	0.005	6	Debtiss	Debt Issuance	0.007**	0.006*
1	Lrrev	Long-term Reversals	0.007	0.010	6	Divg	Dividend Growth	0.001	-0.000
1	Momrev	Momentum-Reversal	0.022**	0.025***	6	Fscore	Piotroski's F-score	0.003	0.005*
1	Sp	Sales-Price	0.021***	0.025***	6	Gltnoa	Growth in LTNOA	0.001	0.002
1	Value	Book-to-Market	0.017**	0.021**	6	Gmargins	Gross Margins	0.006	0.006
1	Valuem	Book-to-Market (monthly)	0.011	0.012	6	Noa	Net Operating Assets	0.024***	0.021***
2	Indmom	Industry Momentum	0.013	0.019	6	Prof	Gross Profitability	0.019***	0.016**
2	Mom	Momentum (6m)	0.009	0.012	6	Repurch	Share Repurchases	0.005	0.007**
2	Mom12	Momentum (1 year)	0.055***	0.059***	6	Season	Seasonality	0.033***	0.030***
2	Valmom	Value-Momentum	0.015*	0.020**	7	Age	Firm Age	-0.000	-0.007
2	Valmomprof	Value-Momentum-Profitability	0.032***	0.035***	7	Ciss	Composite Issuance	0.026***	0.034***
3	Growth	Asset Growth	0.012*	0.017**	7	Exchsw	Exchange Switch	0.007	0.008
3	Igrowth	Investment Growth	0.011*	0.016***	7	Indmomrev	Industry Momentum-Reversal	0.039***	0.040***
3	Inv	Investment	0.017**	0.022***	7	Indrrev	Industry Relative Reversals	0.034***	0.029***
3	Sgrowth	Sales Growth	-0.004	0.002	7	Indrrevlv	Industry Relative Reversals (low vol)	0.047***	0.044***
4	Ep	Earnings-Price	0.028***	0.035***	7	Invaci	Abnormal Corporate Investment	0.004	0.003
4	Price	Price	0.014	0.017*	7	Ipo	Initial Public Offering	-0.005	-0.011
4	Roa	Return to Assets (m)	0.026***	0.032***	7	Nissa	Share Issuance (a)	0.031***	0.038***
4	Roaa	Return to Assets (a)	0.015**	0.022***	7	Nissm	Share Issuance (m)	0.025***	0.033***
4	Roe	Return to Book Equtiy (m)	0.013	0.022***	7	Shortint	Short Interest	-0.006	0.003
4	Roea	Return to Book Equity (a)	0.030***	0.039***	7	Size	Size	0.003	0.014**
4	Rome	Return to Market Equity	0.052***	0.062***	7	Strev	Short-term Reversal	0.014	0.006
5	Betaarb	Beta Arbitrage	0.001	0.038***	7	Sue	PEAD (SUE)	0.023***	0.025***
5	Invcap	Investment-Capital	0.006	0.021**	7	Valprof	Value-Profitability	0.030***	0.036***
5	Ivol	Idiosyncratic Volatility	0.032***	0.052***					

A typical implementation of t-SNE starts from random initial conditions. We run the t-SNE procedure 500 times with different initializations and record whether two strategies are in the same group in a given run. Then, for each pairs of strategies, we calculate the ratio that they are in the same cluster across all runs. We denote these ratios as the similarity ratios between strategies. For example, Cfp and Divp are in the same group 435 times out of the 500 runs, so the ratio between Cfp and Divp is 435/500 = 0.87. A high ratio indicates that the two strategies are consistently categorized in the same cluster, and a low ratio indicates that the two strategies belong to different clusters.

We summarize the results in Figure 2 as a heatmap. Each entry in the heatmap shows a similarity ratio of the corresponding column strategy and row strategy. A darker color means higher similarity ratio (compare, for example, the similarity ratio of 0.87 between Cfp and Divp and the similarity ratio of 0.00 between Cfp and Indmom).[5] The diagonal of the heatmap represents the similarity ratio of the same strategy, and thus the entries on the diagonal are always one. We sort the strategies into seven groups where the pairwise similarity ratios within the clusters are high.

Figure 2 shows that the t-SNE separates the strategies in seven groups. We report the strategies of the seven groups in Table 2. The strategies are as follows: (1) Cash Flow-to-Market Value (Cfp), Dividend Yield (Divp), Cash Flow Duration (Dur), Leverage (Lev), Long-Term Reversals (Lrrev), Momentum-Reversal (Momrev), Sales-to-Price (Sp), Book-to-Market (Value), and Book-to-Market monthly (Valuem); (2) Industry Momentum (Indmom), Momentum 6m (Mom), Momentum 1y (Mom12), Value-Momentum (Valmom), and Value-Momentum-Profitability (Valmomprof); (3) Asset Growth (Growth), Investment Growth (Igrowth), Investment (Inv), and Sales Growth (Sgrowth); (4) Earnings-to-Price (Ep), Price (Price), Return-to-Assets monthly (Roa), Return-to-Assets annual (Roaa), Return-to-Book Equity monthly (Roe), Return-to-Book Equity annual (Roea), and Return-to-Market Equity (Rome); (5) Beta Arbitrage (Betaarb), Investment-to-Capital (Invcap), Idiosyncratic Volatility (Ivol), and Share Volume (Shvol); (6) Accruals (Accruals), Asset Turnover (Aturnover), Debt Issuance (Debtiss), Dividend Growth (Divg), Piotroski's F-score (Fscore), Growth in LTNOA (Gltnoa), Gross Margins (Gmargins), Net Operating Assets (Noa), Gross Profitability (Prof), Share Repurchases (Repurch), and Seasonality (Season); and (7) Firm Age (Age), Composite Issuance (Ciss), Exchange Switch (Exchsw), Industry Momentum Reversal (Indmomrev), Industry Relative Reversals (Indrrev), Industry Relative Reversals Low Volatility (Indrrevlv), Abnormal Corporate Investment (Invaci), Initial Public Offering (Ipo), Share Issuance annual (Nissa), Share Issuance monthly (Nissm), Short Interest (Shortint), Firm Equity (Size), Short-Term Reversal (Strev), PEAD (Sue), and Value-Profitability (Valprof). We report the average similarity ratio for each strategy in the group and the average similarity ratio of each group in Table 3. For example, Cfp has average similarity ratios of 0.86, 0.00, 0.00, 0.01, 0.02, 0.00, and 0.03 with the each of the seven t-SNE groups, respectively.

[5]The entries of a column (or a row) are pairwise probabilities and do not need to sum up to one because multiple strategies can fall in the same cluster.

Table 2: t-SNE Groups

This table reports the descriptions of the seven groups generated by the t-SNE procedure. The first six groups are distinct clusters based on the t-SNE procedure and the Residual group contains the long-short strategies that are left out by the t-SNE procedure.

Group	Num	Strategy							
Value	1	Cfp Valuem	Divp	Dur	Lev	Lrrev	Momrev	Sp	Value
Momentum	2	Indmom	Mom	Mom12	Valmom	Valmomprof			
Investment	3	Growth	Igrowth	Inv	Sgrowth				
Profitability	4	Ep	Price	Roa	Roaa	Roe	Roea	Rome	
Volatility	5	Betaarb	Invcap	Ivol	Shvol				
Firm	6	Accruals Prof	Aturnover Repurch	Debtiss Season	Divg	Fscore	Gltnoa	Gmargins	Noa
Residual	7	Age Nissa	Ciss Nissm	Exchsw Shortint	Indmomrev Size	Indrrev Strev	Indrrevlv Sue	Invaci Valprof	Ipo

Figure 2: Heatmap

This figure reports the similarity ratios between any pairwise of strategies. Each entry in the heatmap shows a similarity ratio of the corresponding column strategy and row strategy. A darker color means higher similarity ratio

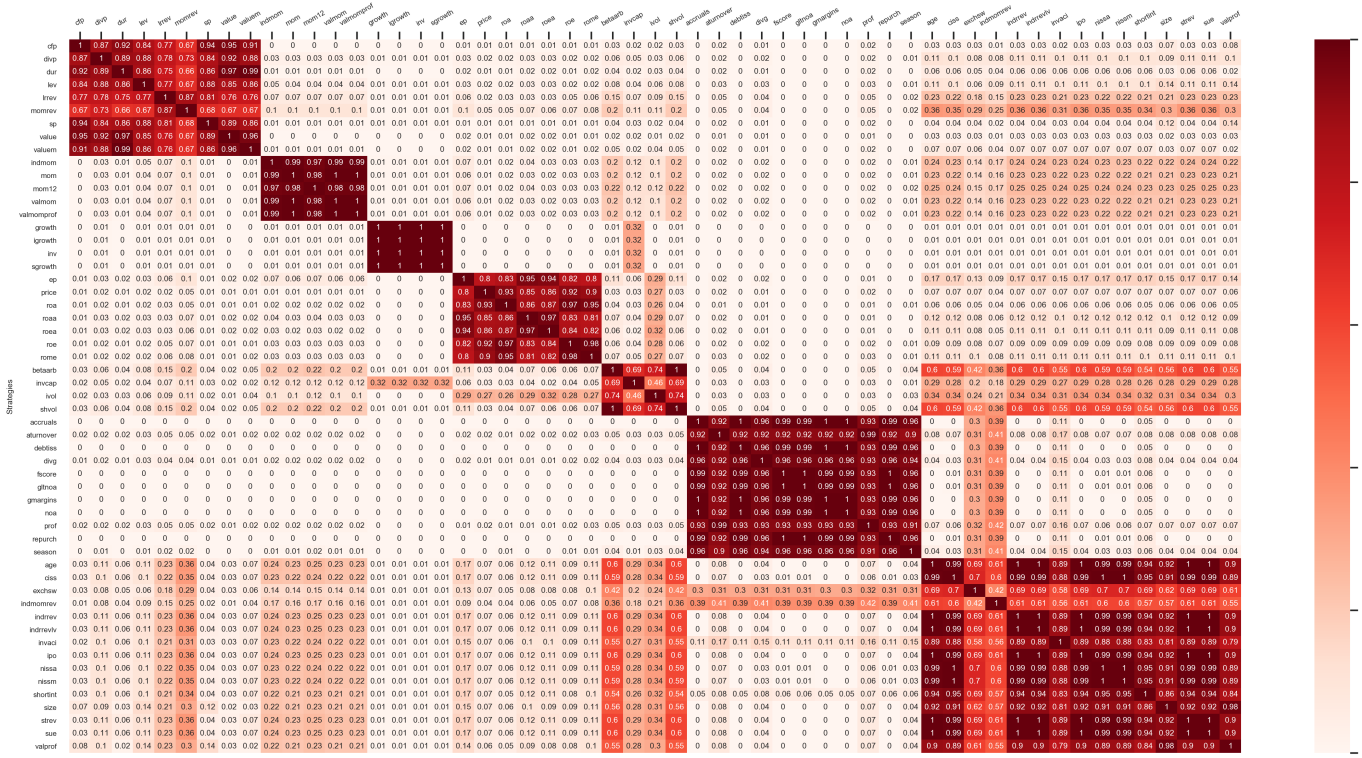


Table 3: Average Similarity Ratio

This table reports the average similarity ratios. Panel A reports the average similarity ratios of each long-short strategy to the t-SNE groups. Panel B reports the average similarity ratios between any pairwise of t-SNE groups.

Panel A		1	2	3	4	5	6	7
		Value	Momentum	Investment	Profitability	Volatility	Firm	Residual
Value	Cfp	0.86	0.00	0.00	0.01	0.02	0.00	0.03
	Divp	0.85	0.03	0.01	0.02	0.05	0.01	0.10
	Dur	0.86	0.01	0.00	0.02	0.03	0.00	0.05
	Lev	0.83	0.04	0.01	0.02	0.06	0.01	0.11
	Lrrev	0.78	0.07	0.01	0.04	0.12	0.01	0.22
	Momrev	0.70	0.10	0.01	0.07	0.16	0.01	0.33
	Sp	0.84	0.01	0.01	0.01	0.03	0.00	0.05
	Value	0.87	0.00	0.00	0.01	0.02	0.00	0.03
	Valuem	0.86	0.01	0.00	0.02	0.04	0.00	0.06
Momentum	Indmom	0.03	0.99	0.01	0.03	0.16	0.01	0.22
	Mom	0.03	0.99	0.01	0.03	0.16	0.01	0.21
	Mom12	0.03	0.98	0.01	0.03	0.17	0.01	0.23
	Valmom	0.03	0.99	0.01	0.03	0.16	0.01	0.21
	Valmomprof	0.03	0.99	0.01	0.03	0.16	0.01	0.21
Investment	Growth	0.01	0.01	1.00	0.00	0.08	0.00	0.01
	Igrowth	0.01	0.01	1.00	0.00	0.08	0.00	0.01
	Inv	0.01	0.01	1.00	0.00	0.08	0.00	0.01
	Sgrowth	0.01	0.01	1.00	0.00	0.08	0.00	0.01
Profitability	Ep	0.03	0.06	0.00	0.86	0.14	0.00	0.16
	Price	0.02	0.01	0.00	0.88	0.09	0.00	0.07
	Roa	0.02	0.02	0.00	0.90	0.09	0.00	0.06
	Roaa	0.03	0.03	0.00	0.88	0.12	0.00	0.11
	Roe	0.03	0.02	0.00	0.88	0.11	0.00	0.10
	Roea	0.02	0.03	0.00	0.89	0.11	0.01	0.09
	Rome	0.03	0.03	0.00	0.88	0.12	0.01	0.11
Volatility	Betaarb	0.07	0.20	0.01	0.06	0.81	0.02	0.56
	Invcap	0.04	0.12	0.32	0.04	0.61	0.01	0.27
	Ivol	0.05	0.10	0.00	0.28	0.65	0.01	0.32
	Shvol	0.07	0.20	0.01	0.06	0.81	0.02	0.56
Firm	Accruals	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Aturnover	0.03	0.02	0.00	0.02	0.04	0.93	0.12
	Debtiss	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Divg	0.02	0.02	0.00	0.01	0.03	0.95	0.09
	Fscore	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Gltnoa	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Gmargins	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Noa	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Prof	0.03	0.02	0.00	0.02	0.04	0.93	0.11
	Repurch	0.00	0.00	0.00	0.00	0.00	0.97	0.06
	Season	0.01	0.01	0.00	0.00	0.03	0.95	0.09

Table 3 continued								
Residual	Age	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Ciss	0.11	0.23	0.01	0.10	0.45	0.02	0.92
	Exchsw	0.09	0.14	0.01	0.08	0.32	0.31	0.65
	Indmomrev	0.08	0.16	0.01	0.06	0.28	0.40	0.58
	Indrrev	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Indrrevlv	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Invaci	0.10	0.23	0.01	0.10	0.42	0.13	0.82
	Ipo	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Nissa	0.11	0.23	0.01	0.10	0.45	0.02	0.92
	Nissm	0.11	0.23	0.01	0.10	0.45	0.02	0.92
	Shortint	0.11	0.22	0.01	0.10	0.42	0.06	0.88
	Size	0.11	0.22	0.01	0.10	0.43	0.02	0.86
	Strev	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Sue	0.12	0.24	0.01	0.10	0.46	0.02	0.92
	Valprof	0.12	0.22	0.01	0.09	0.42	0.02	0.85
Panel B		1	2	3	4	5	6	7
		Value	Momentum	Investment	Profitability	Volatility	Firm	Residual
	Value	0.83	0.03	0.01	0.02	0.06	0.01	0.11
	Momentum	0.03	0.99	0.01	0.03	0.16	0.01	0.22
	Investment	0.01	0.01	1.00	0.00	0.08	0.00	0.01
	Profitability	0.02	0.03	0.00	0.88	0.11	0.00	0.10
	Volatility	0.06	0.16	0.08	0.11	0.72	0.01	0.43
	Firm	0.01	0.01	0.00	0.00	0.01	0.96	0.07
	Residual	0.11	0.22	0.01	0.10	0.43	0.07	0.86

The first five groups map closely to the findings in the large anomaly literature. We label the first group the "Value" cluster because it contains many strategies that are considered to be variations of value strategies. A general definition of value strategy is the strategy that buys assets that are cheap relative to their fundamental value and shorts the assets that are expensive relative to their fundamental value. In the equity market, value strategies are often measured by ratios of the "fundamental" value to the market value of the firm. The average similarity ratio of this cluster is 0.83. These fundamental-to-market value ratios are nicely captured in the Value cluster (e.g., Cfp, Divp, Sp, Value, and Valuem). The average similarity ratio of these strategies are 0.90. Another measure that is commonly considered to be a version of value and is highly correlated with fundamental-to-market value ratios is the negative of the long-term past return of the assets, or the long-term reversal anomaly (e.g., De Bondt and Thaler, 1985; Fama and French, 1993; Asness, Moskowitz, and Pedersen, 2013; and Moskowitz, 2015). The Value cluster also captures the long-run reversal strategy, or Lrrev. Importantly, several strategies that are not typically identified in the literature as value strategies are categorized as such by our methodology: the duration (Dur)

and leverage (Lev) strategies. A mixed strategy, Momrev, is categorized into the Value group. It is also notable that one of the classic value strategies, earning-to-price, is consistently not categorized in the Value cluster—its average similarity ratio with the Value cluster is only 0.03.

We label the second group the "Momentum" cluster. Momentum is the idea that assets that had superior performance in the past tend to perform better in the near future. The t-SNE procedure assigns all the classic momentum strategies in the Momentum cluster, including Indmom, Mom, and Mom12. Two of the mixed strategies, Valmom and Valmomprof, are also identified in the Momentum cluster. Average similarity ratio of the Momentum cluster is 0.99, suggesting that the Momentum cluster is highly robust.

The third group is labeled the "Investment" cluster. The investment strategies are generally defined as the strategies that long firms with little investment and short firms with significant investment. In the asset pricing literature, there is a debate on whether value and investment are capturing the same underlying factor (e.g., Fama and French, 2016 and Hou, Xue, and Zhang, 2015). The idea is that growth companies, or companies with low fundamental-to-market value ratios, also tend to invest more. Importantly, the t-SNE procedure cleanly separates the two broad sets of strategies from each other. The Investment cluster captures most of the commonly believed investment-related strategies and does not contain any of the value strategies, and vice versa. The average similarity ratio between the Value and Investment clusters is 0.01. Looking at Figure 2, both the Value and Investment clusters are highly robust clusters, where the similarity ratios within the clusters are usually above 0.90. The similarity ratios between strategies in the Value cluster and the Investment clusters are always lower than 0.05, suggesting that the t-SNE procedure rarely confuses the two clusters with each other. It is notable that invcap strategy that is commonly regarded as an investment strategy has an average similarity ratio of 0.32 with this group.

We label the fourth group the "Profitability" cluster. Profitability can be broadly defined as the strategy that buys good "fundamental" companies and shorts bad "fundamental" companies. This group includes five classic profitability-related anomalies, including Roa, Roaa, Roe, Roea, and Rome. The average similarity ratios of the five strategies with the Profitability cluster are 0.90, 0.88, 0.88, 0.89, and 0.88. Importantly, it also includes the Ep and Price strategies. The Ep strategy is a traditional value strategy but its average similarity ratio with the Profitability cluster is 0.86. Price also has a high average similarity ratio with the Profitability cluster—0.88. It is notable that one of the classic profitability measures, Prof, only has an average similarity ratio of 0.02 with the Profitability cluster.

The fifth group is labeled the "Volatility" cluster. This group contains all of the classic volatility-related anomalies (Betaarb, Ivol, and Shvol). These volatility measures were originally constructed to measure different components of stock volatility—Betaarb on systematic volatility, Ivol on idiosyncratic volatility, and Shvol on trading volatility. These rather different in purpose volatility measures all fall into the same cluster. The average similarity ratio among these three strategies is 0.83, somewhat lower than the average across the other clusters. Interestingly, the Invcap is also included in this cluster, albeit with relatively lower average similarity ratio with the Volatility

cluster at 0.61.

The sixth cluster is a cluster that has not been identified in the anomaly literature. We label this group the "Firm" cluster. The cluster contains many strategies that relate to companies' operating activities. For example, the cluster includes strategies based on accounting practices (Accruals), asset turnover speed (Aturnover), debt issuance (Debtiss), changes in dividend policies (Divg), and share repurchasing decisions (Repurch). The Firm cluster has an average similarity ratio of 0.96. We show later in Section 4 that the Firm cluster is intrinsically high-dimensional, while the other clusters are low-dimensional with one dominating component. Perhaps, this is one reason why this cluster was not previously identified as a stand-alone group by the anomaly literature which focused on the low-dimensional clusters. In contrast, we show later that the Firm cluster is high-dimensional making it more difficult to identify with the traditional methods. This difficulty is again related to two facts discussed above: (1) many of the classical methods of dimensionality reductions result in an intermediate output in 10-100 dimensions and (2) the identification of higher-dimensional structures even in "relatively low" dimensions 10-100 suffers from the measure concentration phenomenon (a different instance of which is labeled "the curse of dimensionality"). If one is given an abstract data set of 5 clusters, 4 of which are close to 1-dimensional and well described by the leading principal vector while the last one is intrinsically eight-dimensional, this will be missed by many methods. While linear methods, such as PCA, have wonderfully clean mathematical properties and rely on Linear Algebra for which fast algorithms are available, it has become clear in recent years that the analysis of genuinely high-dimensional data will have to incorporate and rely on nonlinear techniques such as the one we are using which are designed to combat the measure concentration phenomenon.

The remaining strategies are in the "Residual" group that contains all the strategies the t-SNE procedure leaves out. In general, there can be two primary reasons that a strategy enters the Residual group. Firstly, a strategy can be a mix of some of the six main clusters and thus the procedure is uncertain where to assign the strategy. Secondly, a strategy can be distinct from all other strategies. In Section 4, we investigate why the strategies enter the Residual group.

t-SNE Graphs

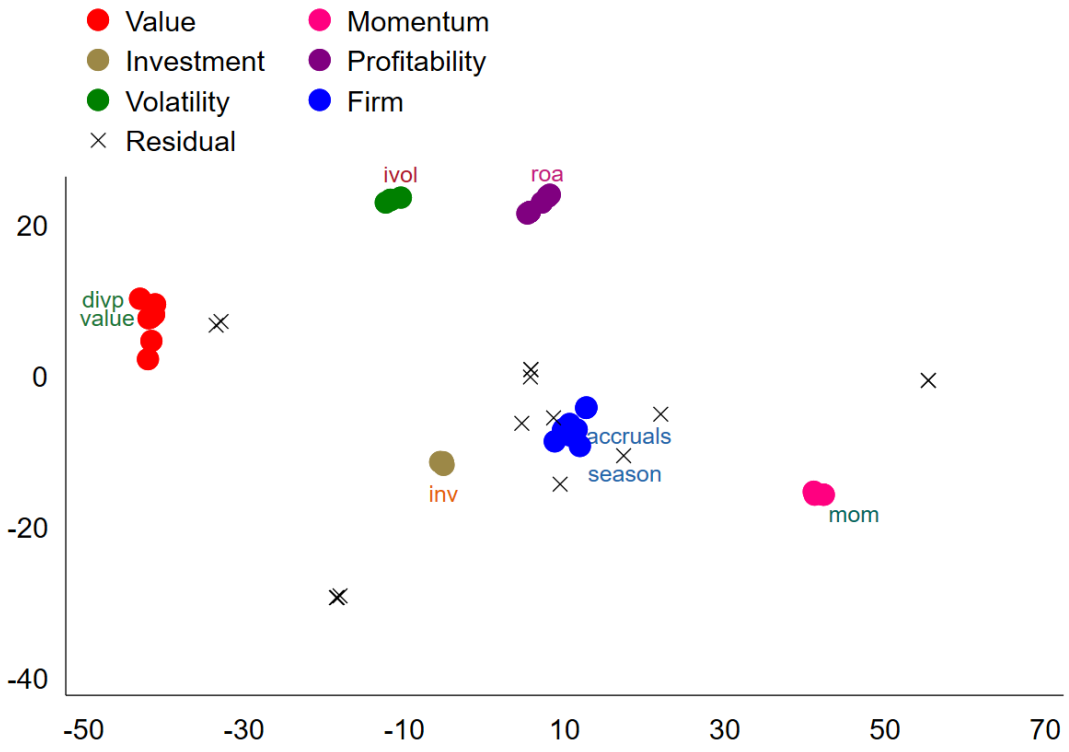
A major advantage of t-SNE is that the procedure effectively summarizes the relations of high-dimensional objects in low dimensions (commonly in two dimension). In this section, we show the two dimensional representations of the 55 strategies.

For the main presentation of the results, we follow Kobak, Linderman, Steinerberger, Kluger, and Berens (2019). They propose to set the initial position for the t-SNE procedure as the first two eigenvalues from the spectral clustering method. They argue that this representation allows to better capture both the local and also the global structure of the high-dimensional objects when they are mapped to the low-dimensional representations. Figure 3 shows the t-SNE graph of the embedding of the high-dimensional strategies in two dimensions. We color the strategies based on their clusters as in Table 2. The first six clusters are represented in circles of different colors and

the Residual group is shown in crosses. The blue dots in the center are the strategies in the Firm cluster. The brown dots on the left side of the Firm cluster are in the Investment cluster. The pink dots on the right side of the Firm cluster are in the Momentum cluster. The Volatility and the Profitability clusters are colored in green and purple, respectively. They are on the top of the graph. The red dots are the strategies in the Value cluster and they are on the left of the graph. Strategies in the same clusters are much closer to each other than to other clusters. Strategies in different clusters cleanly separate from each other in the 2-dimensional space. The strategies in the Residual group scatter in the graphs, and therefore the t-SNE procedure leaves the strategies as uncategorized.

Figure 3: t-SNE Graph

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The initial condition is set as the first two eigenvalues from the spectral clustering method following Kobak, Linderman, Steinerberger, Kluger, and Berens (2019).



Next, we apply the t-SNE procedure to the daily returns of the 55 strategies based on different random initializations to study the robustness of t-SNE procedure. We report the results from 10 runs in the Appendix to visually illustrate the results of Section 2. Similarly, we color the strategies based on their clusters as in Table 2. Again, strategies in the same clusters are much closer to each other across the runs. Strategies in different clusters tend to cleanly separate from each other in the 2-dimensional space.

3.3 Comparison with Other Clustering Methods

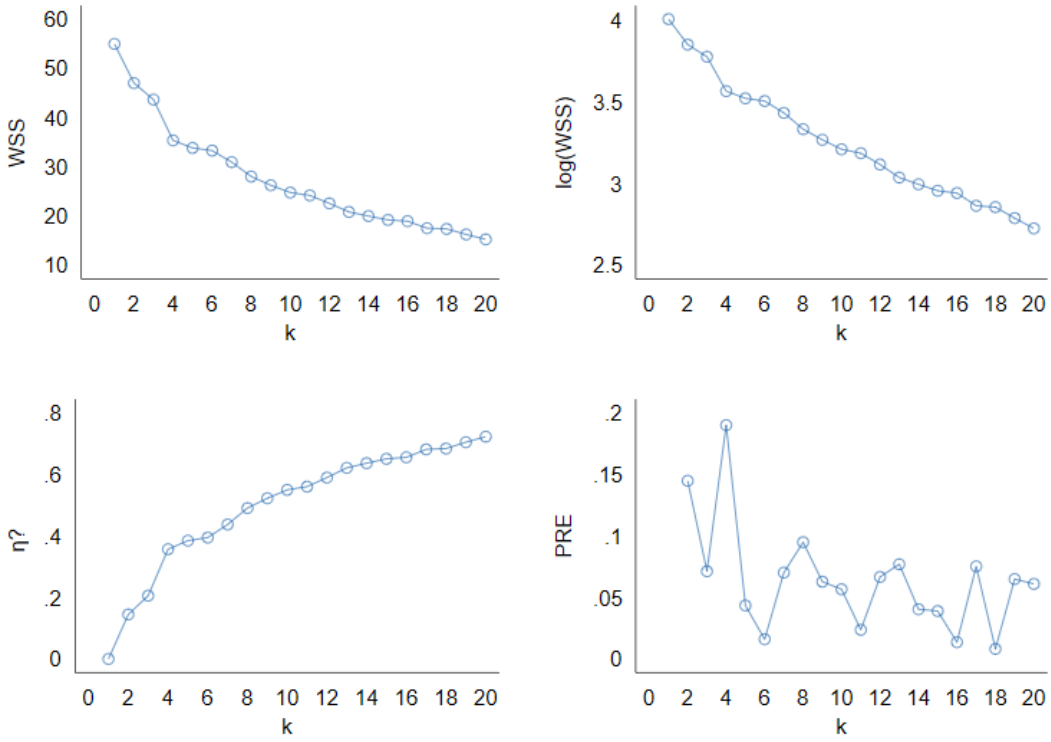
In this section, we examine the performance of traditional methods in separating and clustering objects. We examine four methods that have been employed in the asset pricing literature – K-Means, principal component analysis, kernel principal component analysis, and spectral method.

K-Means

We start with the K-Means clustering method. The goal of the K-Means is to partition N objectives into K clusters in which each object belongs to the cluster with the nearest mean. K-Means clustering and its variations are widely used and studied in finance (e.g., Brown and Goetzmann, 1997; Brown and Goetzmann, 2003; Patton and Weller, 2019).

A key input of K-Means clustering is the number of clusters, K. There is no consensus on the method to select the number of clusters, and the choice may lead to very different outputs. Several methods have been proposed to select the optimal number of groups K^* in the literature. To detect the optimal number of groups K^*, the most common method is to use a screeplot and search for a kink from the within sum of squares (WSS) or the logarithm of the within sum of squares ($\log(WSS)$) for all cluster solutions. Other criteria for selecting the optimal K^* include the η^2 coefficient and the proportional reduction of error (PRE) coefficient (see Hastie, Tibshirani, and Friedman, 2009). We calculate the four statistics using the daily returns of the 55 long-short strategies and report them in Figure 4. The four panels in Figure 4 correspond to WSS, $\log(WSS)$, η^2, and PRE, respectively. For each panel, we plot the corresponding statistics from $K = 1$ to $K = 20$. The curves of the graphs are generally smooth and there are not obvious kinks for the WSS, $\log(WSS)$, and η^2 panels. For the PRE panel, the statistic is not stable across the cluster solutions. Overall, it is difficult to determine K^* as an input of K-Means clustering.

Figure 4: K-Means Graphs



Next, we test whether the K-Means clustering method delivers reasonable clusters if we assume $K = 7$. Table 4 reports the elements in each of the seven clusters. Clusters 3, 4, and 6 only contain three strategies each and do not form consistent themes. Cluster 1 mainly consists of strategies in the Firm cluster and the Residual group, and also has some investment-related strategies. Cluster 2 has many of the elements in the Profitability cluster, but also includes many volatility-related strategies. Cluster 5 closely matches the Momentum cluster. Cluster 7 has many elements in the Value cluster, but also includes two investment-related strategies. In summary, even if we can determine the optimal number of clusters, the K-Means clustering method does not cleanly separate the strategies.

Table 4: K-Means Clusters

Group Num	Strategy							
1	Accruals Gltnoa Prof	Age Igrowth Repurch	Aturnover Indmomrev Size	Ciss Inv Sue	Debtiss Invaci	Divg Ipo	Exchsw Nissa	Fscore Noa
2	Ep Rome	Invcap Shvol	Ivol	Nissm	Roa	Roaa	Roe	Roea
3	Gmargins	Season	Shortint					
4	Indrrev	Indrrevlv	Strev					
5	Indmom	Mom	Valmom	Valmomprof				
6	Betaarb	Mom12	Price					
7	Cfp Sp	Divp Valprof	Dur Value	Growth Valuem	Lev	Lrrev	Momrev	Sgrowth

Principal Component

Principal component analysis is a widely used technique in the asset pricing literature. It is a global dimensionality reduction method and we have discussed the advantages and disadvantages in Section 2. Among the main advantages is the linear nature of PCA which makes it well suited for the detection of linear structures that have been subjected to some noise, even when those structures are embedded in very high dimensions. PCA is thus, unsurprisingly, the most fundamental way to analyze linear structures (or close-to-linear structures). However, it is not designed to handle truly high-dimensional clusters (especially when these clusters are not defined by one principal component but may require dozens or hundreds to be accurately captured); it also suffers from the fact that forcing it to produce an embedding in very low dimensions may lead to sizable distortions.

To summarize the strategies in 2-dimensional space, we plot the loading graph with the first

and second principal components from the principal component analysis. The result is documented in Figure 5. Panel A of Figure 5 is the baseline loading plot and Panel B of Figure 5 is colored based on the t-SNE clusters. From Panel A, we see that the principal component analysis finds it difficult to determine the clusters. Perhaps, one can argue that the four dots in the upper right corner are distinct – these are, in fact, the four volatility strategies.

We now color the strategies in the loading plot based on the t-SNE clusters and show it in Panel B. Although difficult to separate the strategies apart in Panel A, we see that strategies within the same t-SNE clusters are, in general, closer to each other than strategies in different t-SNE clusters. That is, PCA captures some notion of distance in the high-dimensional space.

Kernel Principal Component

Another method that has been used in the literature is the kernel principal component analysis (e.g., Kozak, 2019). Similar to the principal component analysis, we plot the loading graph with the first and second principal components from the kernel principal component analysis. The result is documented in Figure 6. Panel A of Figure 6 is the baseline loading plot and Panel B of Figure 6 is colored based on the t-SNE clusters. From Panel A, similar to the results of principal component analysis, and there is no clear separation in the loading plot. In other words, it is difficult to determine the clusters of the strategies from the kernel principal component analysis. Kernel principal component analysis does not seem to do much better in separating out strategies than simple principal component analysis. Again, the colored graph in Panel B reveals that the strategies within the same t-SNE clusters are in the same neighborhood in the KPCA loading plot, even though it is difficult to separate the clusters from each other in Panel A.

Spectral Method

Spectral clustering methods are another widely used family of methods in dimensionality reduction,[6] though not extensively used in finance. They are intrinsically nonlinear and similar to t-SNE in that they try to avoid working with the high-dimensional data set more than absolutely necessary. Indeed, Linderman and Steinerberger (2019) identified t-SNE as a type of spectral method that has a built in repulsion effect; this repulsion effect allows for the embedding into two dimensions in contrast to classical spectral methods that end up with an output in the 10-100 dimensional range.

To summarize the strategies in 2-dimensional space, we plot the graph with the first and second eigenvectors from the spectral method. The result is documented in Figure 7. Panel A of Figure 7 is the baseline loading plot and Panel B of Figure 7 is colored based on the t-SNE clusters. From Panel A, we see that the spectral method makes it difficult to determine the clusters. One can argue that the four dots in the upper left corner are distinct – these are, in fact, the five momentum strategies.

[6]For example, these methods include: ISOMAP (Tenenbaum, De Silva, and Langford, 2000); Locally Linear Embedding (Roweis and Saul, 2000); Laplacian Eigenmaps (Belkin and Niyogi, 2002); Hessian LLE (Donoho and Grimes, 2003); and Diffusion Maps (Coifman and Lafon, 2006).

Figure 5: Principal Component Analysis Graphs

This figure plots the first two principal component loadings of the 55 long-short strategies. Panel A is uncolored and Panel B is colored based on the t-SNE groups.

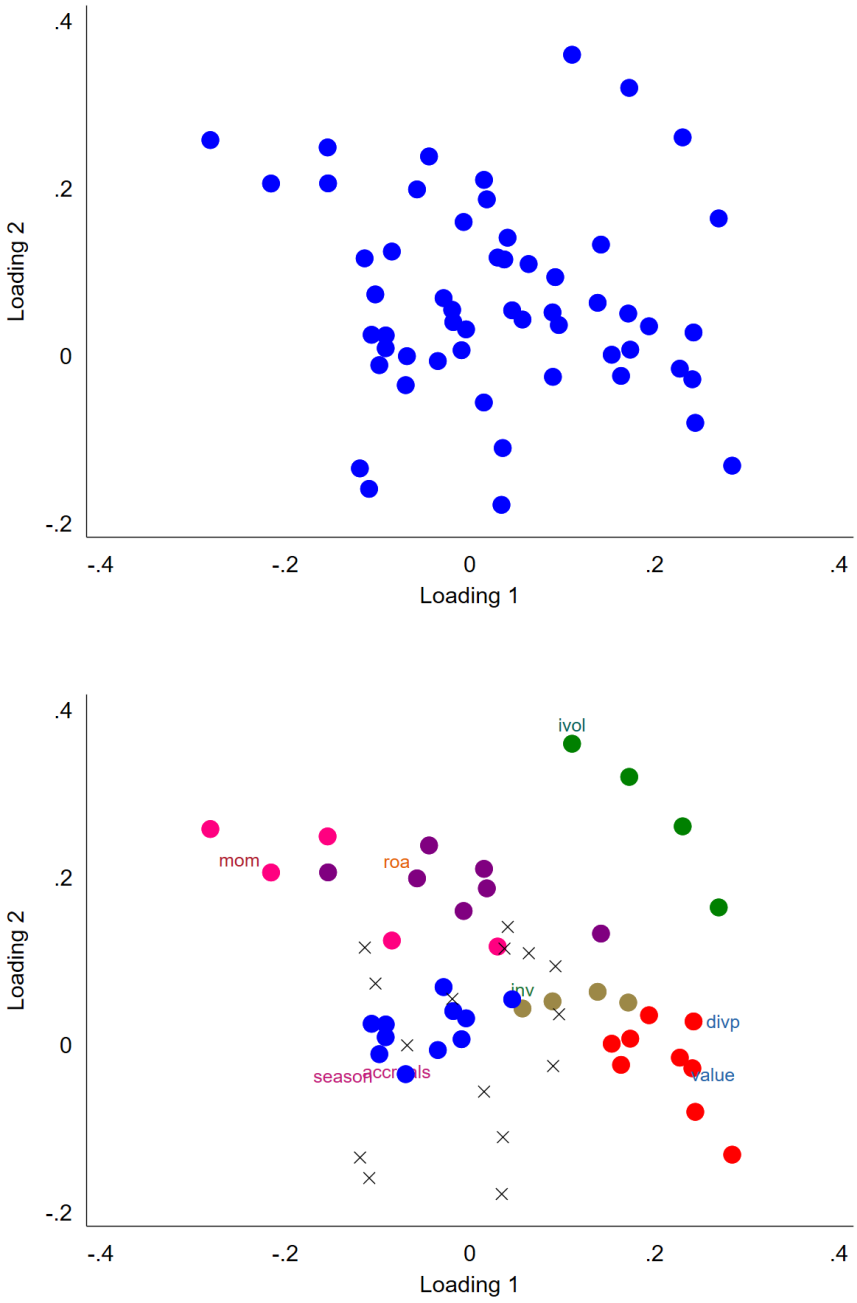


Figure 6: Kernel Principal Component Analysis Graphs

This figure plots the first two kernel principal component loadings of the 55 long-short strategies. Panel A is uncolored and Panel B is colored based on the t-SNE groups.

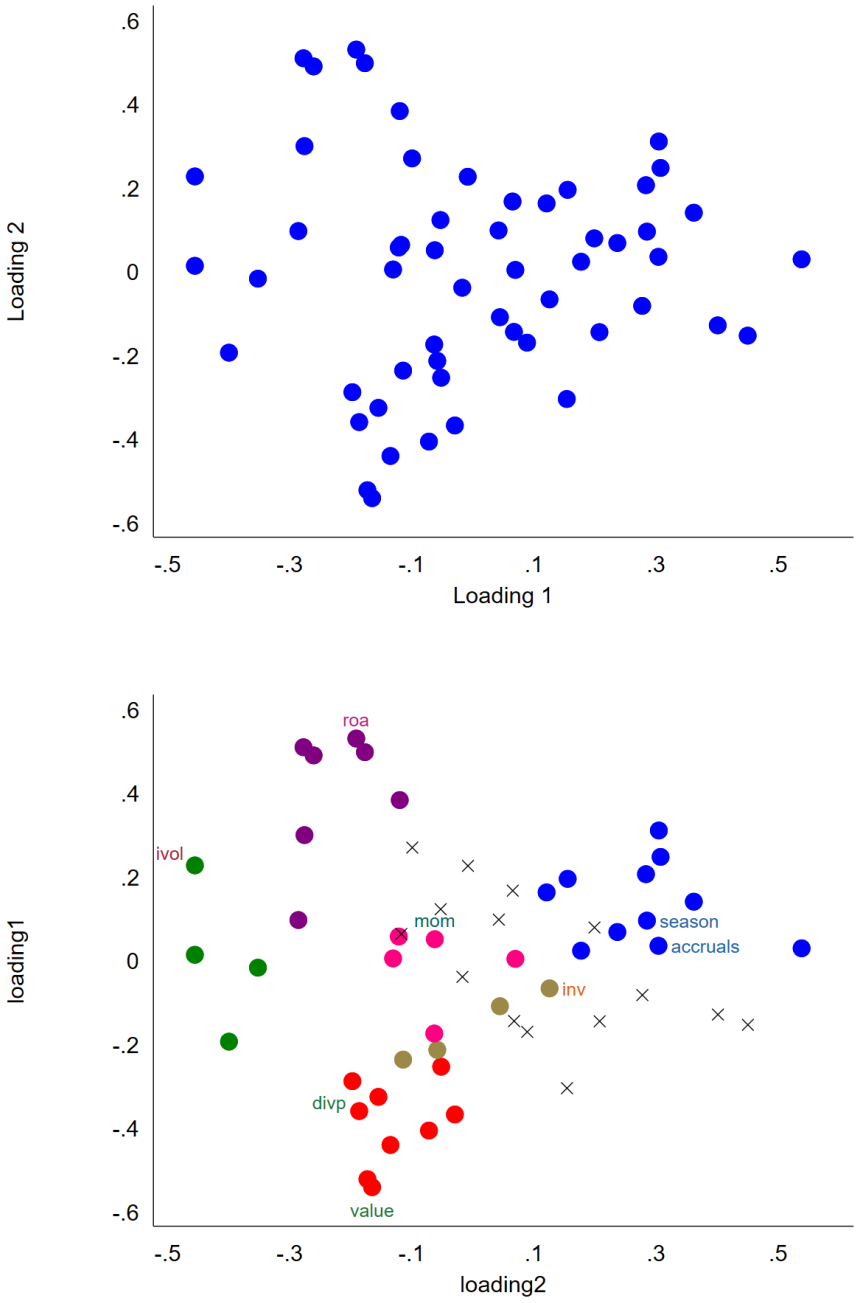
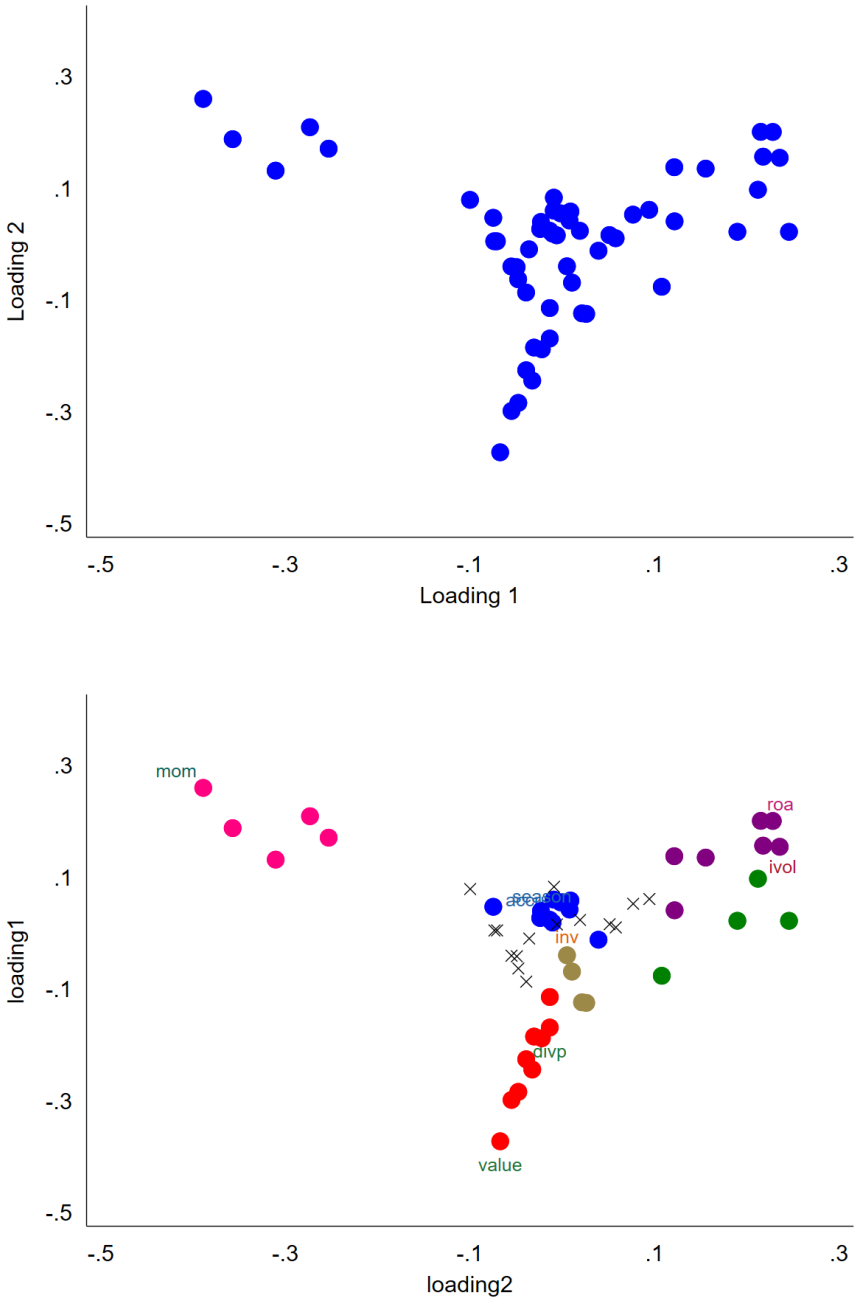


Figure 7: Spectral Method Graphs

This figure plots the first two eigenvectors of the 55 long-short strategies from the spectral method. Panel A is uncolored and Panel B is colored based on the t-SNE groups.



We now color the strategies in the loading plot based on the t-SNE clusters and show it in Panel B. Although difficult to separate the strategies apart in Panel A, we see that strategies within the same t-SNE clusters are, in general, closer to each other than strategies in different t-SNE clusters. That is, spectral method captures some notion of distance in the high-dimensional space.

4 Properties of the t-SNE Groups

In the previous sections, we showed that the t-SNE is able to categorize the factors into distinct clusters. In this section, we further examine the properties of the t-SNE groups.

4.1 Principal Component Analysis Within Clusters

We start by examining the general structure of the t-SNE clusters. Within each t-SNE cluster, we calculate the first three principal components and report results in Table 5. In Table 5, we also report the results based on the first five principal components of all the 55 long-short strategies.

Panel A of Table 5 documents the results based on covariances. For the first five clusters, the first principal component already explains a large fraction of the variations within the group. The first principal component can already explain 51.7 percent, 70.7 percent, 63.0 percent, 63.0 percent, and 73.4 percent of the variations of the strategy returns in Value, Momentum, Investment, Profitability, and Volatility clusters, respectively. The decreases from the first eigenvalue to the next two eigenvalues are large for these five clusters. The first three principal components explain the vast majority of the return variations within clusters. However, the first principal component only explains 22.8 percent of the variations of the strategy returns in the Firm cluster. The second and third principal components explain 15.9 percent and 14.2 percent of the return variations, respectively. These fractions are similar to the results based on all the 55 long-short strategies. For all the 55 strategies, the first three principal components explain 21.7 percent, 16.7 percent, and 10.1 percent of the return variations, respectively.

Panel B of Table 5 reports the results based on correlations. Similar to Panel A, we find that the first principal components dominate the first five clusters, but not the Firm cluster. For the Firm cluster, the first principal component only explains 19.8 percent of the return variations, while the second and third principal components explain 17.1 percent and 10.6 percent, respectively. These results show that the first five clusters (Value, Momentum, Investment, Profitability, and Volatility) are low-dimensional, even 1-dimensional, cluster, but the Firm cluster is intrinsically high-dimensional.

We summarize these findings as follows. The literature successfully identifies the five 1-dimensional clusters but has difficulty in detecting the high-dimensional Firm cluster. The Firm cluster is high-dimensional and the elements in the cluster are close in this high-dimensional space. As we argued, the methodology of t-SNE is well suited to successfully recover it.

We now look into the individual strategies within the clusters. We regress the daily returns of each strategy on the first principal component of the corresponding cluster. For example, because

Cfp belongs to the Value cluster, we regress the daily returns of Cfp on the first principal component of the nine strategies in the Value cluster. The results are reported in Table 6. Based on R-squareds, for the first five clusters, the first principal components explain sizable fractions of the returns of the long-short strategies. For the Value cluster, the first principal component explains 29.1 percent (Momrev) to 76.2 percent (Value) of the variations in the returns of the nine strategies. For the Momentum cluster, the first principal component explains 49.6 percent (Valmom) to 86.6 percent (Mom) of the variations in the returns of the five strategies. For the Investment cluster, the first principal component explains 48.4 percent (Igrowth) to 74.3 percent (Sgrowth) of the variations in the returns of the four strategies. For the Profitability cluster, the first principal component explains 35.3 percent (Ep) to 30.5 percent (Roea) of the variations in the returns of the seven strategies. For the Volatility cluster, the first principal component explains 54.5 percent (Invcap) to 85.4 percent (Shvol) of the variations in the returns of the four strategies. The coefficient estimates on the first principal components are all positive and statistically significant for the strategies in these five clusters.

For the Firm cluster, consistent with the findings in Table 5, we show that the first principal component does not explain much of the variations in the strategy returns. The first principal component explains less than 10 percent of the variations in returns for Accruals (5.1 percent), Debtiss (5.9 percent), Divg (0.0 percent), Fscore (7.6 percent), Gltnoa (1.8 percent), Repurch (4.1 percent), and Season (6.9 percent). The R-squareds are higher for Aturnover (34.2 percent), Gmargins (30.0 percent), Noa (46.8 percent), and Prof (67.9 percent).

What is even more important is that the coefficient estimates on the first principal component are all positive and significant (except Divg). This is in sharp contrast to the Residual group. For the Residual group, the coefficient estimates on the first principal components range from -0.220 to 0.646. These results suggest that, though there is not a dominating principal component for either the Firm cluster or the Residual group, the structure of the two clusters are different. For the Firm cluster, elements are close to each other, despite that the elements are high-dimensional, which can be seen from the fact that the coefficient estimates on the first principal components are consistently positive. However, elements in the Residual group can be very far apart from each other, consistent with the notion that these strategies are left out by the t-SNE procedure. The residual group is mainly a consequence of our very conservative way of obtaining clusters. On the one hand, the clusters have a well-defined block structure affinity and the residual group is also intrinsically consistent.

4.2 Cross-Sectional Pricing Power of Firm Cluster

As discussed earlier, the Firm cluster is intrinsically high dimensional. Many principal components are needed to capture the variations in returns of the elements in the Firm cluster. In this section, we examine the cross-sectional pricing power of the principal components of the Firm cluster.

Table 5: Principal Component Analyses

This table reports the results of principal component analysis within each of the seven t-SNE group. Panel A and Panel B report results based on covariance and correlation, respectively.

Panel A: Based on Covariance					
Group		Eigenvalue	Difference	Proportion	Cumulative
Value					
	Comp1	0.00041	0.00032	0.517	0.517
	Comp2	0.00009	0.00001	0.114	0.631
	Comp3	0.00008	0.00002	0.098	0.728
Momentum					
	Comp1	0.00051	0.00042	0.707	0.707
	Comp2	0.00009	0.00002	0.128	0.835
	Comp3	0.00007	0.00004	0.094	0.929
Investment					
	Comp1	0.00013	0.00010	0.630	0.630
	Comp2	0.00003	0.00001	0.152	0.782
	Comp3	0.00003	0.00001	0.124	0.906
Profitability					
	Comp1	0.00036	0.00028	0.630	0.630
	Comp2	0.00008	0.00002	0.136	0.766
	Comp3	0.00006	0.00002	0.106	0.871
Volatility					
	Comp1	0.00048	0.00040	0.734	0.734
	Comp2	0.00008	0.00001	0.122	0.856
	Comp3	0.00006	0.00004	0.100	0.956
Firm					
	Comp1	0.00009	0.00003	0.228	0.228
	Comp2	0.00007	0.00001	0.165	0.393
	Comp3	0.00006	0.00001	0.135	0.528
Residual					
	Comp1	0.00026	0.00008	0.229	0.229
	Comp2	0.00018	0.00002	0.159	0.388
	Comp3	0.00016	0.00007	0.142	0.530
All Strategies					
	Comp1	0.00098	0.00022	0.217	0.217
	Comp2	0.00075	0.00030	0.167	0.384
	Comp3	0.00046	0.00025	0.101	0.485
	Comp4	0.00020	0.00001	0.045	0.530
	Comp5	0.00019	0.00002	0.042	0.572

Panel B: Based on Correlation					
Group		Eigenvalue	Difference	Porportion	Cumulative
Value					
	Comp1	4.741	3.817	0.527	0.527
	Comp2	0.924	0.079	0.103	0.630
	Comp3	0.845	0.255	0.094	0.723
Momentum					
	Comp1	3.472	2.843	0.695	0.695
	Comp2	0.630	0.188	0.126	0.820
	Comp3	0.442	0.140	0.089	0.909
Investment					
	Comp1	2.473	1.858	0.618	0.618
	Comp2	0.615	0.047	0.154	0.772
	Comp3	0.568	0.223	0.142	0.914
Profitability					
	Comp1	4.481	3.550	0.640	0.640
	Comp2	0.931	0.235	0.133	0.773
	Comp3	0.696	0.221	0.099	0.873
Volatility					
	Comp1	2.902	2.405	0.726	0.726
	Comp2	0.497	0.082	0.124	0.850
	Comp3	0.415	0.230	0.104	0.954
Firm					
	Comp1	2.183	0.301	0.198	0.198
	Comp2	1.881	0.712	0.171	0.369
	Comp3	1.169	0.090	0.106	0.476
Residual					
	Comp1	3.521	1.406	0.235	0.235
	Comp2	2.115	0.528	0.141	0.376
	Comp3	1.586	0.166	0.106	0.482
All Strategies					
	Comp1	10.290	1.625	0.187	0.187
	Comp2	8.665	3.942	0.158	0.345
	Comp3	4.723	2.269	0.086	0.431
	Comp4	2.454	0.347	0.045	0.475
	Comp5	2.107	0.112	0.038	0.513

We form portfolios based on pre-ranking betas on the each of the principal components of the firm cluster. The betas are computed using past 3-year rolling window regressions. We form quintile portfolios based on ex-ante principal component betas and show the average excess returns of the resulting portfolios. Our sample includes all the stocks that are in the major stock exchanges: NYSE, AMEX, and NASDAQ. We further exclude firms in the financial industry (SIC 6000 – SIC 6999) and the utility firms (SIC 4900 – SIC 4999. For each quarter, we sort the stocks based on their estimated betas over the past 3 years, which are denoted as pre-ranking betas. We compute the

betas by regressing daily firm-level excess returns on one of the principal components, controlling for the Fama-French 3-factor model. We require at least 500 non-missing returns for each stock. We rank stocks based on the estimated pre-ranking betas and form value-weighted portfolios. We perform this exercise each quarter and report the average excess returns of each group.

We present the results in Table 7. Panel A of Table 7 shows results based on betas estimated using the each of the first three principal components of the Firm cluster. Panel A reports the average excess returns and t-statistics for each of the quintile portfolio as well as the long-short strategy. For the results based on the first principal component, the average excess returns monotonically increase from the first portfolio to the fifth portfolio. The long-short strategies based on the first principal component generate positive and statistically significant spreads. The magnitude of the spread is 0.522 percent per month or 6.264 percent per year. For the results based on the second and third principal component, the spreads of the long-short strategies are not statistically significant, suggesting that exposures to the second and third principal components do not command risk premia.

Panel B of Table 7 further examines the factor exposures of the quintile portfolios sorted on the first principal component of the Firm cluster. Controlling for the Fama-French 5-factor model, the alphas remain monotonic from the first portfolio to the fifth portfolio. The alpha of the long-short strategy is 0.767 percent per month or 9.204, suggesting that the factor model cannot account for the long-short strategy spread based on the first principal component.

Additionally, we conduct Fama-MacBeth two-pass regressions to test the cross-sectional pricing power of the first principal component of the Firm cluster. In the first pass, the excess returns of each test portfolio are regressed on the factors at the daily frequency to estimate beta loadings using the entire sample. In the second pass, we run a cross-sectional regression of the average excess returns of the test portfolios on the beta loadings estimated in the first pass. We do not include an intercept in the second pass. The test portfolios are the 10 size, 10 book-to-market, and 10 momentum portfolios, together with the strategies in the Firm cluster that have significant risk premia over the sample–10 Accral, 10 Aturnover, 10 Debtiss, 10 Noa, 10 Prof, and 10 Season. The results of the second pass are reported in the Appendix.

The adjusted R-squared using the Fama-French 5-factor model is 10.82 percent. With the addition of the first principal component of the Firm cluster, the R-squared increases to 32.09 percent. The adjusted R-squared using the first five principal components of the 55 strategies is 6.73 percent. With the addition of the first principal component of the Firm cluster, the R-squared increases to 26.15 percent. The first principal component of the Firm cluster is statistically significant at the 5-percent level for both specifications. The results of the Fama-MacBeth two-pass regressions confirm that the first principal component command positive risk premia in the cross-section.

In summary, this section shows that exposures to the first principal component of the Firm cluster command risk premia in the cross-section of firms, while exposures to the higher principal components do not. In other words, although the Firm cluster is intrinsically high dimensional,

its pricing power all comes from the first principal component. Therefore, we can summarize the relevant information for pricing of the Firm cluster with one factor.

4.3 Using All Groups

In this section, we regress each long-short strategy on the first principal components of all seven groups. The goal is to determine whether the additional principal components of other clusters explain much of the return variations in the strategies. For example, whether the returns on Cfp are better explained by adding the first principal components of six other clusters. We document the R-squareds from the baseline model and the model using all the first principal components in Table 8.

For all of the six clusters, the model using all the first principal components does not substantially increase R-squareds over using only the first principal component of the corresponding clusters. For the Value cluster, the increases in R-squareds are less than 10 percent for eight of the nine strategies. The only exception is Valem—the R-squared increases from 59.7 percent to 74.1 percent. For the Momentum cluster, the increases in R-squareds are less than 10 percent for four of the five strategies. The only exception is Valmom—the R-squared increases from 49.6 percent to 73.8 percent. For the investment cluster, the increases in R-squareds for all the strategies are less than 10 percent. For the profitability cluster, five of the seven strategies experience less than 10 percent of increases in R-squareds. The R-squareds of Ep and Price increase by 24.1 percent and 14.6 percent, respectively. For the Volatility cluster, the strategies that experience more than 10 percent increases in R-squared are Invcap and Ivol, but the increases are only 17.4 percent and 12.4 percent, respectively. For the Firm cluster, the model using all the first principal components also does not improve the model fitness for most of the long-short strategies. The increases in R-squareds are less than 10 percent for seven of the eleven strategies – Accruals by 2.7 percent, Aturnover by 7.3 percent, Divg by 3.2 percent, Gltnoa by 3.7 percent, Noa by 8.8 percent, Prof by 2.5 percent, and Season by 6.3 percent. Two additional strategies experience less than 20 percent increases—Fscore by 17.2 percent and Gmargins by 13.9 percent. Debtiss and Repurch experience relatively large increases in R-squareds—Debtiss by 35.9 percent and Repurch by 45.0 percent. These results show that the additional principal components from the other clusters do not generally contain much new information in explaining the return variations of the strategies in the first six clusters.

For the Residual group, the increases in R-squareds tend to be substantial. Twelve of the fifteen strategies experience more than 10 percent increases in R-squareds. Still, there are a few strategies that do not have large increases in goodness-of-fit—Exchsw by 1.9 percent, Indrrevlv by 5.1 percent, and Invaci by 3.3 percent. The results suggest that many of the strategies in the Residual group are mixes of principal components of the other clusters. However, three strategies (Exchsw, Indrrevlv, and Invaci) in the Residual group are unique strategies that do not belong to the six robust t-SNE clusters. This is one explanation for the behavior of the t-SNE clustering which we discussed in Section 3.

Table 6: Individual Strategies Regressions

Regressions of individual strategies on their corresponding first principal components:

$$ret_{i,t} = \alpha + \beta \times PC1_{j,t} + \epsilon_{i,t} \tag{7}$$

where ret_i is the return of strategy i; $PC1_j$ is the first principal component of group j and strategy i is an element of group j. *, **, *** denote significance levels at the 10%, 5%, and 1% based on the standard t-statistics.

Group	Strategy	Beta		Alpha		R^2
Value	Cfp	0.294***	(112.269)	0.004	(0.691)	0.527
	Divp	0.332***	(92.303)	-0.004	(-0.542)	0.429
	Dur	0.399***	(168.340)	0.002	(0.368)	0.714
	Lev	0.370***	(99.907)	-0.007	(-0.911)	0.468
	Lrrev	0.281***	(83.512)	-0.005	(-0.706)	0.381
	Momrev	0.247***	(68.099)	0.011	(1.555)	0.291
	Sp	0.269***	(113.565)	0.010**	(2.030)	0.532
	Value	0.370***	(190.200)	0.001	(0.182)	0.762
	Valuem	0.397***	(129.467)	-0.006	(-0.919)	0.597
Momentum	Indmom	0.511***	(160.015)	-0.015**	(-2.148)	0.693
	Mom	0.506***	(270.197)	-0.019***	(-4.578)	0.866
	Mom12	0.533***	(177.307)	0.026***	(3.784)	0.735
	Valmom	0.299***	(105.584)	-0.001	(-0.232)	0.496
	Valmomprof	0.330***	(135.940)	0.014**	(2.492)	0.620
Investment	Growth	0.552***	(174.314)	0.003	(0.758)	0.728
	Igrowth	0.393***	(103.090)	0.004	(1.012)	0.484
	Inv	0.441***	(107.714)	0.010**	(2.147)	0.506
	Sgrowth	0.588***	(180.745)	-0.013***	(-3.505)	0.743
Profitability	Ep	0.282***	(78.610)	0.009	(1.280)	0.353
	Price	0.367***	(103.987)	-0.010	(-1.519)	0.488
	Roa	0.402***	(203.429)	-0.001	(-0.337)	0.785
	Roaa	0.355***	(167.464)	-0.009**	(-2.123)	0.712
	Roe	0.448***	(215.961)	0.000	(0.019)	0.805
	Roea	0.392***	(182.240)	-0.014***	(-3.341)	0.746
	Rome	0.379***	(125.032)	0.027***	(4.644)	0.580
Volatility	Betaarb	0.590***	(203.394)	-0.009	(-1.408)	0.785
	Invcap	0.362***	(116.472)	-0.000	(-0.004)	0.545
	Ivol	0.506***	(160.808)	0.024***	(3.478)	0.695
	Shvol	0.515***	(257.705)	-0.013***	(-3.031)	0.854
Firm	Accruals	0.152***	(24.622)	0.011*	(1.820)	0.051
	Aturnover	0.436***	(76.757)	0.000	(0.059)	0.342
	Debtiss	0.095***	(26.635)	0.003	(0.934)	0.059
	Divg	-0.005	(-0.641)	0.001	(0.134)	0.000
	Fscore	0.089***	(30.441)	-0.000	(-0.172)	0.076
	Gltnoa	0.077***	(14.209)	-0.002	(-0.382)	0.018
	Gmargins	0.360***	(69.626)	-0.010*	(-1.932)	0.300
	Noa	0.497***	(99.741)	0.002	(0.310)	0.468
	Prof	0.582***	(154.933)	-0.007**	(-2.002)	0.679
	Repurch	0.076***	(22.023)	0.002	(0.514)	0.041
	Season	0.206***	(28.969)	0.024***	(3.478)	0.069
Residual	Age	0.290***	(71.844)	-0.004	(-0.539)	0.313
	Ciss	-0.172***	(-47.523)	0.027***	(4.718)	0.166
	Exchsw	-0.131***	(-24.963)	0.009	(1.008)	0.052
	Indmomrev	-0.039***	(-8.374)	0.040***	(5.250)	0.006
	Indrrev	0.433***	(134.852)	0.029***	(5.649)	0.616
	Indrrevlv	0.226***	(65.104)	0.045***	(8.035)	0.272
	Invaci	0.003	(0.378)	0.004	(0.336)	0.000
	Ipo	0.278***	(62.637)	-0.008	(-1.184)	0.257
	Nissa	-0.190***	(-58.851)	0.033***	(6.401)	0.234
	Nissm	-0.220***	(-58.261)	0.027***	(4.511)	0.231
	Shortint	-0.138***	(-30.727)	-0.004	(-0.607)	0.077
	Size	0.093***	(20.310)	0.002	(0.273)	0.035
	Strev	0.646***	(170.598)	0.007	(1.160)	0.720
	Sue	-0.143***	(-28.116)	0.025***	(2.990)	0.065
	Valprof	-0.051***	(-12.660)	0.030***	(4.676)	0.014

Table 7: Portfolio Sorting on Firm Cluster's PC Betas

This table reports the portfolio sorting results based on the firms' beta loadings of the first three principal components of the Firm cluster. Panel A reports the average excess returns and t-statistics of the quintile portfolios as well as long-short strategy based on exposures of each of the first three principal components. Panel B reports the results of the quintilte portfolios as well as long-short strategy based on exposures of the first principal component, adjusted for the Fama-French 5-factor model.

Panel A: Mean Excess Returns						
PC1	1	2	3	4	5	5–1
Mean	0.320	0.731	0.731	0.697	0.841	0.522
t-Stat	(1.246)	(3.338)	(3.627)	(3.449)	(3.627)	(2.688)
PC2	1	2	3	4	5	5–1
Mean	0.719	0.715	0.793	0.732	0.552	-0.167
t-Stat	(3.582)	(3.659)	(3.837)	(3.209)	(1.879)	(-0.778)
PC3	1	2	3	4	5	5–1
Mean	0.565	0.747	0.722	0.697	0.702	0.137
t-Stat	(2.230)	(3.739)	(3.591)	(3.339)	(2.877)	(0.772)

Panel B: Factor Exposures of PC1 Beta Sorted Portfolios							
	Alpha	MKTRF	SMB	HML	RMW	CMA	R^2
1	-0.440	1.100	0.005	0.014	-0.120	0.144	0.710
	(-2.898)	(29.556)	(0.098)	(0.196)	(-1.718)	(1.377)	
2	-0.126	1.086	0.096	0.060	0.087	0.266	0.885
	(-1.550)	(54.686)	(3.257)	(1.618)	(2.328)	(4.766)	
3	0.025	0.980	0.075	-0.047	0.066	0.198	0.899
	(0.366)	(57.685)	(2.995)	(-1.472)	(2.071)	(4.161)	
4	0.046	0.965	0.038	-0.172	0.069	0.167	0.914
	(0.714)	(61.614)	(1.636)	(-5.904)	(2.340)	(3.798)	
5	0.328	0.987	-0.014	-0.252	0.002	-0.184	0.879
	(3.752)	(46.108)	(-0.434)	(-6.325)	(0.047)	(-3.063)	
5–1	0.767	-0.113	-0.019	-0.266	0.122	-0.328	0.084
	(3.772)	(-2.259)	(-0.260)	(-2.860)	(1.301)	(-2.342)	

Table 8: Other Cluster Regression R-Squareds

The "Correspond" columns report the regression R-squareds of individual strategies on their corresponding first principal components:

$$ret_{i,t} = \alpha + \beta \times PC1_{j,t} + \epsilon_{i,t} \tag{8}$$

where ret_i is the return of strategy i; $PC1_j$ is the first principal component of group j and strategy i is an element of group j.
The "All" columns report the regression R-squards of individual strategies on the first principal components of all seven groups:

$$ret_{i,t} = \alpha + \sum_{\forall j} \beta_j \times PC1_{j,t} + \epsilon_{i,t} \tag{9}$$

where ret_i is the return of strategy i; $PC1_j$ is the first principal component of group j.

Group	Strategy	Correspond	All	Group	Strategy	Correspond	All
Value	Cfp	0.527	0.609	Firm	Accruals	0.051	0.078
	Divp	0.429	0.476		Aturnover	0.342	0.415
	Dur	0.714	0.732		Debtiss	0.059	0.418
	Lev	0.468	0.527		Divg	0.000	0.032
	Lrrev	0.381	0.464		Fscore	0.076	0.248
	Momrev	0.291	0.324		Gltnoa	0.018	0.055
	Sp	0.532	0.558		Gmargins	0.300	0.439
	Value	0.762	0.795		Noa	0.468	0.556
	Valuem	0.597	0.741		Prof	0.679	0.704
Momentum	Indmom	0.693	0.712		Repurch	0.041	0.491
	Mom	0.866	0.879		Season	0.069	0.132
	Mom12	0.735	0.797	Residual	Age	0.313	0.587
	Valmom	0.496	0.738		Ciss	0.166	0.327
	Valmomprof	0.620	0.707		Exchsw	0.052	0.071
Investment	Growth	0.728	0.739		Indmomrev	0.006	0.318
	Igrowth	0.484	0.489		Indrrev	0.616	0.728
	Inv	0.506	0.561		Indrrevlv	0.272	0.323
	Sgrowth	0.743	0.765		Invaci	0.000	0.033
Profitability	Ep	0.353	0.594		Ipo	0.257	0.448
	Price	0.488	0.634		Nissa	0.234	0.402
	Roa	0.785	0.807		Nissm	0.231	0.422
	Roaa	0.712	0.751		Shortint	0.077	0.270
	Roe	0.805	0.817		Size	0.035	0.387
	Roea	0.746	0.792		Strev	0.720	0.822
	Rome	0.580	0.637		Sue	0.065	0.277
Volatility	Betaarb	0.785	0.835		Valprof	0.014	0.369
	Invcap	0.545	0.719				
	Ivol	0.695	0.819				
	Shvol	0.854	0.870				

4.4 Factor Model

The asset pricing literature has proposed various factor models to account for the cross-section of equity returns (e.g., Fama and French, 1993; Carhart, 1997; Hou, Xue, and Zhang, 2015; Fama and French, 2016; Stambaugh and Yuan, 2017; Daniel, Hirshleifer, and Sun, 2020). We test whether including the various factor models substantially increase the R-squareds above using the corresponding first principal components of the clusters. The factor models we consider include the CAPM, Fama-French 3-factor, Carhart 4-factor, Fama-French 5-factor, Fama-French 6-factor, q-factor, and the Daniel-Hirshleifer-Sun behavioral factor models. We report the R-squareds of the various models in Table 9. The first column in Table 9 reports the R-squareds from the PC1 only specification. The remaining columns document the R-squareds from the regression specifications with PC1 and one of the factor models. We focus on the Fama-French 5-factor model in discussing the results because it is currently one of the widest used factor models, and the results based on the other factor models are similar.

For all the first six clusters, the Fama-French 5-factor model does not substantially increase R-squareds when it is included with the first principal component of the corresponding clusters. For the Value cluster, the increases in R-squareds are less than 10 percent for eight of the nine strategies. The only exception is Lev—the R-squared increases from 46.8 percent to 64.7 percent. For the Momentum cluster, the increases in R-squareds are less than 10 percent for four of the five strategies. The only exception is Valmom—the R-squared increases from 49.6 percent to 69.2 percent. For the investment cluster, the increases in R-squareds for all the strategies are less than 10 percent. For the profitability cluster, five of the seven strategies experience less than 10 percent of increases in R-squareds. The R-squareds of Ep and Price increase by 27.2 percent and 23.2 percent, respectively. For the Volatility cluster, the strategies that experience more than 10 percent increases in R-squared are Invcap and Ivol, but the increases are only 17.7 percent and 16.3 percent, respectively. For the Firm cluster, the Fama-French 5-factor model also does not improve the model fitness for most of the strategies. The increases in R-squareds are less than 10 percent for seven of the eleven strategies—Accruals by 4.2 percent, Aturnover by 9.1 percent, Divg by 2.2 percent, Gltnoa by 1.3 percent, Noa by 8.4 percent, Prof by 3.2 percent, and Season by 6.0 percent. Two additional strategies experience less than 20 percent increases—Fscore by 15.2 percent and Gmargins by 15.0 percent. Debtiss and Repurch experience relatively large increases in R-squareds—Debtiss by 45.8 percent and Repurch by 45.9 percent. These results suggest that the factor models do not contain much new information in explaining the return variations of the strategies in the first six clusters.

For the Residual group, the increases in R-squareds tend to be substantial. Ten of the fifteen strategies experience more than 10 percent increases in R-squareds. Still, there are a few strategies that do not have large increases in goodness-of-fit—Exchsw by 7.7 percent, Indmomrev by 6.0 percent, Indrrevlv by 5.6 percent, Invaci by 1.3 percent, and Sue by 8.7 percent. The results

suggest that many of the strategies in the Residual group are mixes of the existing factors.

Table 9: Factor Model R-Squared Improvement

q-factor and DHS behavioral factor models end in 2018.

Group	Strategy	PC1	+CAPM	+FF 3	+Carhart 4	+FF 5	+FF 6	+q-fac	+DHS
Value	Cfp	0.527	0.532	0.550	0.569	0.590	0.606	0.565	0.552
	Divp	0.429	0.469	0.507	0.507	0.508	0.508	0.498	0.498
	Dur	0.714	0.715	0.732	0.740	0.738	0.743	0.713	0.712
	Lev	0.468	0.513	0.640	0.647	0.662	0.664	0.531	0.522
	Lrrev	0.381	0.381	0.418	0.433	0.480	0.486	0.417	0.398
	Momrev	0.291	0.292	0.311	0.314	0.315	0.316	0.293	0.287
	Sp	0.532	0.535	0.552	0.572	0.575	0.589	0.574	0.564
	Value	0.762	0.763	0.772	0.787	0.773	0.787	0.760	0.764
	Valuem	0.597	0.600	0.609	0.737	0.636	0.742	0.657	0.642
Momentum	Indmom	0.693	0.695	0.697	0.723	0.699	0.725	0.701	0.698
	Mom	0.866	0.867	0.871	0.871	0.874	0.874	0.872	0.870
	Mom12	0.735	0.736	0.783	0.827	0.783	0.827	0.765	0.745
	Valmom	0.496	0.499	0.692	0.692	0.699	0.700	0.624	0.561
	Valmomprof	0.620	0.620	0.661	0.671	0.662	0.672	0.650	0.649
Investment	Growth	0.728	0.729	0.733	0.735	0.743	0.744	0.742	0.737
	Igrowth	0.484	0.487	0.489	0.491	0.494	0.497	0.495	0.486
	Inv	0.506	0.506	0.532	0.532	0.558	0.561	0.527	0.518
	Sgrowth	0.743	0.743	0.751	0.758	0.752	0.760	0.755	0.747
Profitability	Ep	0.353	0.360	0.614	0.625	0.639	0.643	0.456	0.460
	Price	0.488	0.510	0.647	0.720	0.654	0.725	0.636	0.644
	Roa	0.785	0.787	0.798	0.798	0.801	0.801	0.806	0.798
	Roaa	0.712	0.712	0.718	0.745	0.741	0.757	0.730	0.721
	Roe	0.805	0.805	0.827	0.828	0.842	0.842	0.819	0.815
	Roea	0.746	0.749	0.751	0.787	0.772	0.798	0.767	0.762
	Rome	0.580	0.584	0.622	0.627	0.643	0.644	0.623	0.616
Volatility	Betaarb	0.785	0.838	0.877	0.878	0.884	0.885	0.857	0.860
	Invcap	0.545	0.564	0.712	0.712	0.722	0.723	0.677	0.637
	Ivol	0.695	0.738	0.839	0.842	0.858	0.862	0.850	0.829
	Shvol	0.854	0.854	0.862	0.870	0.867	0.875	0.864	0.861
Firm	Accruals	0.051	0.054	0.063	0.064	0.093	0.094	0.085	0.066
	Aturnover	0.342	0.357	0.373	0.376	0.433	0.434	0.401	0.400
	Debtiss	0.059	0.066	0.379	0.381	0.517	0.526	0.383	0.431
	Divg	0.000	0.002	0.003	0.003	0.022	0.022	0.002	0.003
	Fscore	0.076	0.122	0.160	0.164	0.228	0.229	0.191	0.152
	Gltnoa	0.018	0.020	0.022	0.034	0.031	0.041	0.052	0.028
	Gmargins	0.300	0.301	0.422	0.437	0.450	0.458	0.445	0.417
	Noa	0.468	0.469	0.471	0.472	0.552	0.552	0.505	0.478
	Prof	0.679	0.682	0.695	0.696	0.711	0.713	0.693	0.695
	Repurch	0.041	0.089	0.347	0.350	0.500	0.515	0.396	0.477
	Season	0.069	0.084	0.106	0.109	0.129	0.136	0.123	0.125
Residual	Age	0.313	0.316	0.506	0.514	0.683	0.694	0.546	0.592
	Ciss	0.166	0.239	0.264	0.264	0.335	0.335	0.314	0.388
	Exchsw	0.052	0.053	0.123	0.124	0.129	0.130	0.129	0.127
	Indmomrev	0.006	0.006	0.061	0.241	0.066	0.242	0.037	0.064
	Indrrev	0.616	0.627	0.682	0.682	0.744	0.744	0.707	0.728
	Indrrevlv	0.272	0.274	0.301	0.301	0.328	0.329	0.310	0.329
	Invaci	0.000	0.001	0.012	0.013	0.013	0.013	0.009	0.010
	Ipo	0.257	0.258	0.448	0.462	0.507	0.526	0.398	0.446
	Nissa	0.234	0.265	0.278	0.278	0.372	0.372	0.335	0.385
	Nissm	0.231	0.270	0.300	0.301	0.428	0.431	0.357	0.380
	Shortint	0.077	0.141	0.262	0.262	0.262	0.263	0.189	0.217
	Size	0.035	0.286	0.823	0.829	0.832	0.837	0.807	0.789
	Strev	0.720	0.725	0.798	0.798	0.844	0.844	0.820	0.837
	Sue	0.065	0.065	0.150	0.257	0.152	0.257	0.191	0.116
	Valprof	0.014	0.082	0.300	0.301	0.326	0.327	0.284	0.304

5 Additional Results

5.1 Recession Periods

Recession periods are commonly regarded as being different from the norm. Furthermore, several papers (e.g., Ang and Chen, 2002; Ang, Chen, and Xing, 2006; Lettau, Maggiori, and Weber, 2014) argue that the characteristics and correlation structures of factors are revealed during downturns. In this section, we apply the t-SNE procedure on the 55 strategies for the recession sample, where recessions are defined as the NBER recession periods.

We again follow Kobak, Linderman, Steinerberger, Kluger, and Berens (2019) and set the initial position for the t-SNE procedure as the first two eigenvalues from the spectral clustering method. They argue that this representation allows to better capture both the local and also the global structure of the high-dimensional objects when they are mapped to the low-dimensional representations. Figure 8 plots the two-dimensional visual representation of the result. We color the strategies based on their clusters as in Table 2. The strategies in the same clusters are in general much closer to each other than to other clusters. The blue dots in the center are the strategies in the Firm cluster. The brown dots on the on top of the Firm cluster are in the Investment cluster. The pink dots on the left side of the Firm cluster are in the Momentum cluster. The Volatility and the Profitability clusters are colored in green and purple, respectively. They are on the top of the graph. The red dots are the strategies in the Value cluster and they are on the right of the graph. Some strategies in the Value and Profitability clusters leave the their corresponding main clusters on the figure. Overall, we conclude that the t-SNE clusters are generally stable during the NBER-defined recession periods.

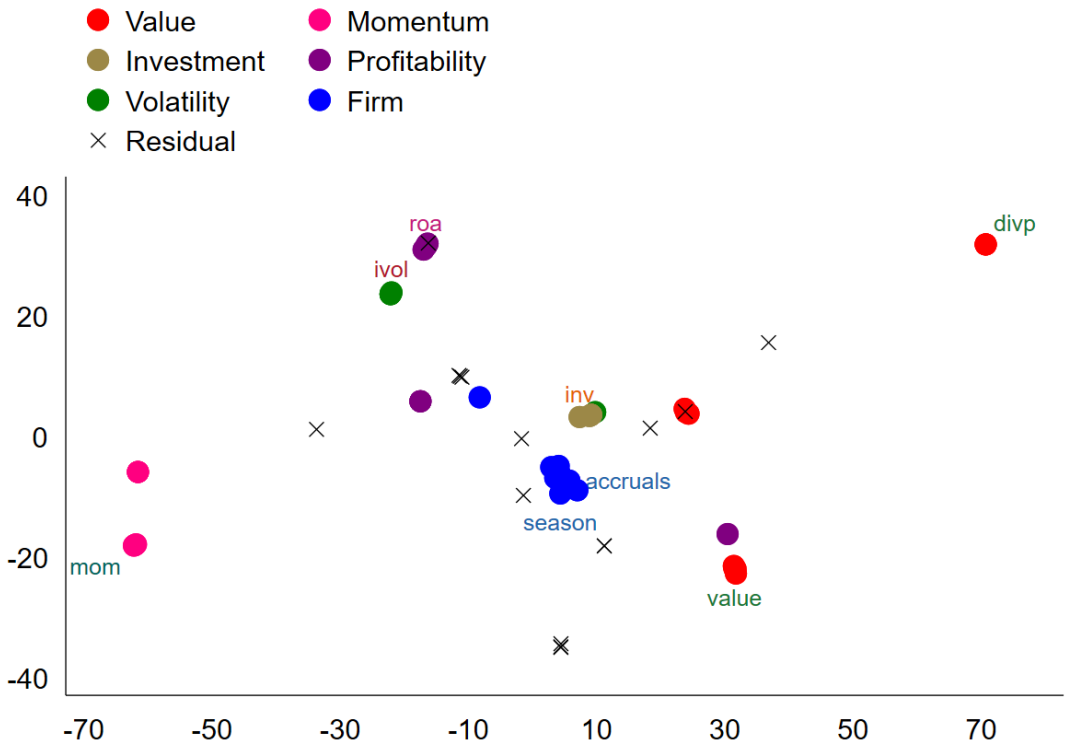
5.2 Subsample

In this section, we break the full sample into two halves, and apply the t-SNE procedure on each subsample.

We follow Kobak, Linderman, Steinerberger, Kluger, and Berens (2019) and set the initial position for the t-SNE procedure as the first two eigenvalues from the spectral clustering method. We document the results in the Appendix. We color the strategies based on their clusters as in Table 2. For both the subsamples, strategies in the same clusters are much closer to each other than to other clusters. The strategies in the Residual group scatter in the graphs, and therefore the t-SNE procedure leaves the strategies as uncategorized. We also show that the clusters are better separated in the second half of the sample, relative to the first half of the sample.

Figure 8: Graphical Illustration of t-SNE – Recession Period

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The return profiles are based on the recession period defined by NBER. The initial condition is set as the first two eigenvalues from the spectral clustering method following Kobak, Linderman, Steinerberger, Kluger, and Berens (2019). The groups are defined in the table based on the t-SNE procedure on the whole sample.



5.3 Results Based on Monthly Returns

We use daily returns for our main specifications. In this section, we apply the t-SNE procedure on the monthly, instead of daily returns, and test whether the results are similar.

Instead of daily return profiles, we apply the t-SNE procedure on the monthly return profiles. We again follow Kobak, Linderman, Steinerberger, Kluger, and Berens (2019) and set the initial position for the t-SNE procedure as the first two eigenvalues from the spectral clustering method of the monthly return profiles. We report the results in the Appendix. We color the strategies based on their clusters as in Table 2. Most of the clusters remain cleanly separated from one another. All but two strategies in the Value cluster are at the bottom of the graph. The momentum strategies form a distinct cluster at the bottom left of the graph. The Investment and Profitability clusters

remain distinct from all the other clusters. The Firm cluster is centered in the middle of the graph. The only exception is the Volatility cluster—elements of the Volatility cluster scattered at different places of the graph. This is consistent with the earlier finding that the Volatility cluster is relatively unstable, compared with the other distinct clusters. Overall, we show that we reach similar conclusion using monthly returns. We also note that the results show that using daily returns allow us to uncover finer structures of the data, relative to using monthly returns.

6 Conclusion

We introduce to finance a new method of dimensionality reduction t-SNE that enjoyed tremendous popularity for its empirical successes in a variety of natural sciences. Importantly, this method was shown recently to have a strong mathematical foundation. We use it to group a broad set of asset pricing factors into a small number of clusters. The t-SNE procedure endogenously generate six distinct clusters. The first five clusters resemble the familiar value, momentum, investment, profitability, and volatility strategies. The sixth cluster is new and we label it the Firm cluster. We show that the first five clusters are low dimensional and have dominating principal components. However, the Firm cluster is intrinsically high dimensional and does not have one dominating principal component.

References

Ang, Andrew and Joseph Chen (2002). "Asymmetric correlations of equity portfolios". In: *Journal of Financial Economics* 63.3, pp. 443–494.

Ang, Andrew, Joseph Chen, and Yuhang Xing (2006). "Downside risk". In: *The Review of Financial Studies* 19.4, pp. 1191–1239.

Asness, Clifford S, Tobias J Moskowitz, and Lasse Heje Pedersen (2013). "Value and momentum everywhere". In: *The Journal of Finance* 68.3, pp. 929–985.

Belkin, Mikhail and Partha Niyogi (2002). "Laplacian eigenmaps and spectral techniques for embedding and clustering". In: *Advances in neural information processing systems*, pp. 585–591.

Bhojraj, Sanjeev and Charles Lee (2002). "Who is my peer? A valuation-based approach to the selection of comparable firms". In: *Journal of Accounting Research* 40.2, pp. 407–439.

Bhojraj, Sanjeev, Charles Lee, and Derek K Oler (2003). "What's my line? A comparison of industry classification schemes for capital market research". In: *Journal of Accounting Research* 41.5, pp. 745–774.

Brown, Stephen J and William N Goetzmann (1997). "Mutual fund styles". In: *Journal of Financial Economics* 43.3, pp. 373–399.

— (2003). "Hedge funds with style". In: *The Journal of Portfolio Management* 29.2, pp. 101–112.

Bryzgalova, Svetlana (2015). "Spurious factors in linear asset pricing models". In: *Working Paper* 1, p. 3.

Bybee, Leland, Bryan T Kelly, Asaf Manela, and Dacheng Xiu (2020). *The structure of economic news*. Tech. rep. National Bureau of Economic Research.

Carhart, Mark M (1997). "On persistence in mutual fund performance". In: *The Journal of Finance* 52.1, pp. 57–82.

Cochrane, John H (2011). "Presidential address: Discount rates". In: *The Journal of Finance* 66.4, pp. 1047–1108.

Coifman, Ronald R and Stéphane Lafon (2006). "Diffusion maps". In: *Applied and Computational Harmonic Analysis* 21.1, pp. 5–30.

Daniel, Kent, David Hirshleifer, and Lin Sun (2020). "Short-and long-horizon behavioral factors". In: *The Review of Financial Studies* 33.4, pp. 1673–1736.

De Bondt, Werner FM and Richard Thaler (1985). "Does the stock market overreact?" In: *The Journal of Finance* 40.3, pp. 793–805.

Donoho, David L and Carrie Grimes (2003). "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data". In: *Proceedings of the National Academy of Sciences* 100.10, pp. 5591–5596.

Ester, Martin, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: AAAI Press, pp. 226–231.

Fama, Eugene F and Kenneth R French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of Financial Economics*.

— (2016). "Dissecting anomalies with a five-factor model". In: *The Review of Financial Studies* 29.1, pp. 69–103.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu (2019). *Taming the factor zoo: A test of new factors*. Tech. rep. National Bureau of Economic Research.

Fraley, Chris and Adrian E Raftery (2002). "Model-based clustering, discriminant analysis, and density estimation". In: *Journal of the American Statistical Association* 97.458, pp. 611–631.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber (2020). "Dissecting characteristics nonparametrically". In: *The Review of Financial Studies* 33.5, pp. 2326–2377.

Garcia, Diego and Øyvind Norli (2012). "Geographic dispersion and stock returns". In: *Journal of Financial Economics* 106.3, pp. 547–565.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as data". In: *Journal of Economic Literature* 57.3, pp. 535–74.

Giglio, Stefano, Yuan Liao, and Dacheng Xiu (2019). "Thousands of alpha tests". In: *Chicago Booth Research Paper* 18-09, pp. 2018–16.

Giglio, Stefano and Dacheng Xiu (2019). "Asset pricing with omitted factors". In: *Chicago Booth Research Paper* 16-21.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Han, Yufeng, Ai He, David Rapach, and Guofu Zhou (2018). "What firm characteristics drive us stock returns". In: *Available at SSRN*.

Harvey, Campbell R and Yan Liu (2019). "Lucky factors". In: *Available at SSRN 2528780*.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "âŠ and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hinton, Geoffrey E and Sam T Roweis (2003). "Stochastic neighbor embedding". In: *Advances in neural information processing systems*, pp. 857–864.

Hoberg, Gerard and Gordon Phillips (2016). "Text-based network industries and endogenous product differentiation". In: *Journal of Political Economy* 124.5, pp. 1423–1465.

Hou, Kewei, Chen Xue, and Lu Zhang (2015). "Digesting anomalies: An investment approach". In: *The Review of Financial Studies* 28.3, pp. 650–705.

— (2020). "Replicating anomalies". In: *The Review of Financial Studies* 33.5, pp. 2019–2133.

Kelly, Bryan T, Seth Pruitt, and Yinan Su (2019). "Characteristics are covariances: A unified model of risk and return". In: *Journal of Financial Economics* 134.3, pp. 501–524.

Kobak, Dmitry, George Linderman, Stefan Steinerberger, Yuval Kluger, and Philipp Berens (2019). "Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 124–139.

Kozak, Serhiy (2019). "Kernel Trick for the Cross-Section". In: *Available at SSRN 3307895*.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2018). "Interpreting factor models". In: *The Journal of Finance* 73.3, pp. 1183–1223.

— (2020). "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2, pp. 271–292.

Ledoux, Michel (2001). *The concentration of measure phenomenon*. 89. American Mathematical Soc.

Lee, John A and Michel Verleysen (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Lettau, Martin, Matteo Maggiori, and Michael Weber (2014). "Conditional risk premia in currency markets and other asset classes". In: *Journal of Financial Economics* 114.2, pp. 197–225.

Lettau, Martin and Markus Pelger (2020a). "Estimating latent asset-pricing factors". In: *Journal of Econometrics*.

— (2020b). "Factors that fit the time series and cross-section of stock returns". In: *Review of Financial Studies*.

Linderman, George C and Stefan Steinerberger (2019). "Clustering with t-SNE, provably". In: *SIAM Journal on Mathematics of Data Science* 1.2, pp. 313–332.

Liu, Yukun and Xi Wu (2020). "Labor Links and Shock Transmissions". In: *Working Paper*.

Ludvigson, Sydney C and Serena Ng (2007). "The empirical risk–return relation: A factor analysis approach". In: *Journal of Financial Economics* 83.1, pp. 171–222.

— (2009). *A factor analysis of bond risk premia*. Tech. rep. National Bureau of Economic Research.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" In: *The Journal of Finance* 71.1, pp. 5–32.

Moskowitz, Tobias J (2015). "Asset pricing and sports betting". In: *Chicago Booth Research Paper* 15-26.

Novy-Marx, Robert (2013). "The other side of value: The gross profitability premium". In: *Journal of Financial Economics* 108.1, pp. 1–28.

Patton, Andrew J and Brian M Weller (2019). "Testing for Unobserved Heterogeneity via k-means Clustering". In: *arXiv preprint arXiv:1907.07582*.

Roweis, Sam T and Lawrence K Saul (2000). "Nonlinear dimensionality reduction by locally linear embedding". In: *Science* 290.5500, pp. 2323–2326.

Shaham, Uri and Stefan Steinerberger (2017). "Stochastic neighbor embedding separates well-separated clusters". In: *arXiv preprint arXiv:1702.02670*.

Stambaugh, Robert F and Yu Yuan (2017). "Mispricing factors". In: *The Review of Financial Studies* 30.4, pp. 1270–1315.

Steinerberger, Stefan (2011). "Extremal uniform distribution and random chord lengths". In: *Acta Mathematica Hungarica* 130.4, pp. 321–339.

Tenenbaum, Joshua B, Vin De Silva, and John C Langford (2000). "A global geometric framework for nonlinear dimensionality reduction". In: *Science* 290.5500, pp. 2319–2323.

Tibshirani, Robert and Guenther Walther (2005). "Cluster validation by prediction strength". In: *Journal of Computational and Graphical Statistics* 14.3, pp. 511–528.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411–423.

van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov, pp. 2579–2605.
van der Maaten, Laurens, Eric Postma, and Jaap van den Herik (2009). "Dimensionality reduction: a comparative". In: *J Mach Learn Res* 10.66-71, p. 13.

Appendix: Tables & Graphs

Table A.1: Individual Strategies Regressions Using Five Principal Components

Regressions of individual strategies on the five principal components of all strategies.

Group	Strategy	1 PCs	2 PCs	3 PCs	4 PCs	5 PCs
Value	Cfp	0.017	0.435	0.498	0.532	0.566
	Divp	0.007	0.422	0.438	0.459	0.462
	Dur	0.009	0.482	0.658	0.671	0.684
	Lev	0.053	0.422	0.431	0.577	0.630
	Lrrev	0.007	0.241	0.407	0.407	0.423
	Momrev	0.000	0.205	0.253	0.283	0.284
	Sp	0.001	0.405	0.555	0.560	0.566
	Value	0.003	0.526	0.728	0.739	0.756
	Valuem	0.155	0.709	0.712	0.712	0.717
Momentum	Indmom	0.313	0.407	0.681	0.713	0.714
	Mom	0.272	0.505	0.805	0.806	0.858
	Mom12	0.329	0.633	0.778	0.796	0.814
	Valmom	0.145	0.152	0.694	0.702	0.769
	Valmomprof	0.167	0.228	0.616	0.625	0.662
Investment	Growth	0.068	0.327	0.383	0.383	0.393
	Igrowth	0.060	0.202	0.220	0.233	0.254
	Inv	0.035	0.082	0.097	0.114	0.185
	Sgrowth	0.039	0.396	0.405	0.406	0.411
Profitability	Ep	0.209	0.394	0.447	0.513	0.570
	Price	0.410	0.588	0.659	0.688	0.705
	Roa	0.514	0.547	0.669	0.679	0.695
	Roaa	0.386	0.386	0.661	0.687	0.705
	Roe	0.608	0.624	0.716	0.718	0.725
	Roea	0.452	0.455	0.730	0.741	0.752
	Rome	0.476	0.478	0.493	0.493	0.508
Volatility	Betaarb	0.467	0.572	0.576	0.827	0.880
	Invcap	0.226	0.701	0.712	0.717	0.719
	Ivol	0.712	0.765	0.852	0.852	0.857
	Shvol	0.444	0.713	0.723	0.807	0.821
Firm	Accruals	0.001	0.022	0.024	0.031	0.031
	Aturnover	0.086	0.098	0.110	0.111	0.113
	Debtiss	0.086	0.336	0.400	0.464	0.489
	Divg	0.001	0.002	0.002	0.019	0.032
	Fscore	0.153	0.178	0.236	0.244	0.246
	Gltnoa	0.029	0.030	0.031	0.031	0.032
	Gmargins	0.014	0.223	0.426	0.427	0.430
	Noa	0.001	0.127	0.128	0.131	0.151
	Prof	0.011	0.144	0.185	0.221	0.221
	Repurch	0.209	0.326	0.453	0.501	0.505
	Season	0.002	0.127	0.131	0.131	0.135
Residual	Age	0.357	0.486	0.631	0.691	0.709
	Ciss	0.251	0.317	0.328	0.345	0.346
	Exchsw	0.034	0.037	0.078	0.095	0.098
	Indmomrev	0.079	0.202	0.334	0.341	0.376
	Indrrev	0.152	0.164	0.189	0.421	0.791
	Indrrevlv	0.064	0.068	0.075	0.163	0.360
	Invaci	0.000	0.022	0.022	0.035	0.100
	Ipo	0.230	0.368	0.492	0.551	0.565
	Nissa	0.319	0.345	0.363	0.364	0.365
	Nissm	0.351	0.374	0.430	0.446	0.447
	Shortint	0.131	0.231	0.276	0.279	0.279
	Size	0.010	0.105	0.404	0.568	0.623
	Strev	0.208	0.213	0.253	0.471	0.891
	Sue	0.160	0.280	0.316	0.322	0.324
	Valprof	0.026	0.169	0.250	0.252	0.256

Table A.2: Fama-MacBeth Two-Pass Regression

Table table shows the results from the Fama-MacBeth two-pass regressions. In the first pass, the excess return of each test portfolio is regressed on the factors at the daily frequency to estimate beta loadings using the entire sample. In the second pass, we run a cross-sectional regression of the average excess returns of the test portfolios on the beta loadings estimated in the first pass without intercept. The time-series average slope coefficients and t-statistics from the second pass regression are reported in the table. Parentheses contain the Fama-Macbeth adjusted t-statistics and brackets contain the Shanken adjusted t-statistics. The 51 test portfolios are the 10 Size, 10 BM, 10 Momentum, 10 Accrual, 10 Asset Turnover, 10 Debtiss, 10 Noa, 10 Gross Profitability, and 10 Seasonality. *, **, *** denote significance levels at the 10%, 5%, and 1% based on the Shanken adjusted t-statistics.

	MKTRF	SMB	HML	RMW	CMA	$PC1^{Firm}$	Adj R^2
1	0.0343**	0.0060	-0.0045	0.0062	0.0027		0.1082
	(3.5300)	(1.0948)	(-0.6653)	(1.0706)	(0.4904)		
	[2.5013]	[0.7962]	[-0.5355]	[0.9040]	[0.4150]		
2	0.0357**	0.0025	0.0083	-0.0046	0.0107	0.0352**	0.3209
	(3.6730)	(0.4660)	(1.2301)	(-0.8051)	(2.0213)	(3.6237)	
	[2.6022]	[0.3387]	[0.9887]	[-0.6790]	[1.6904]	[2.6398]	
	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC1^{Firm}$	Adj R^2
1	-0.0022	-0.0317	0.0829**	-0.1776***	-0.0947**		0.0673
	(-0.0652)	(-1.1373)	(3.1230)	(-4.8863)	(-2.1861)		
	[-0.0496]	[-0.8358]	[2.4925]	[-4.5841]	[-2.0947]		
2	-0.0655	0.0092	0.0177	-0.0216	0.1057**	0.0366***	0.2615
	(-1.7932)	(0.3203)	(0.6768)	(-0.5450)	(2.2557)	(3.7273)	
	[-1.3977]	[0.2383]	[0.5368]	[-0.5161]	[2.1743]	[2.7275]	

Figure A.1: Graphical Illustration of t-SNE Clusters

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The initial condition is generated randomly for each of the runs. The groups are defined in the table based on the t-SNE procedure on the whole sample.

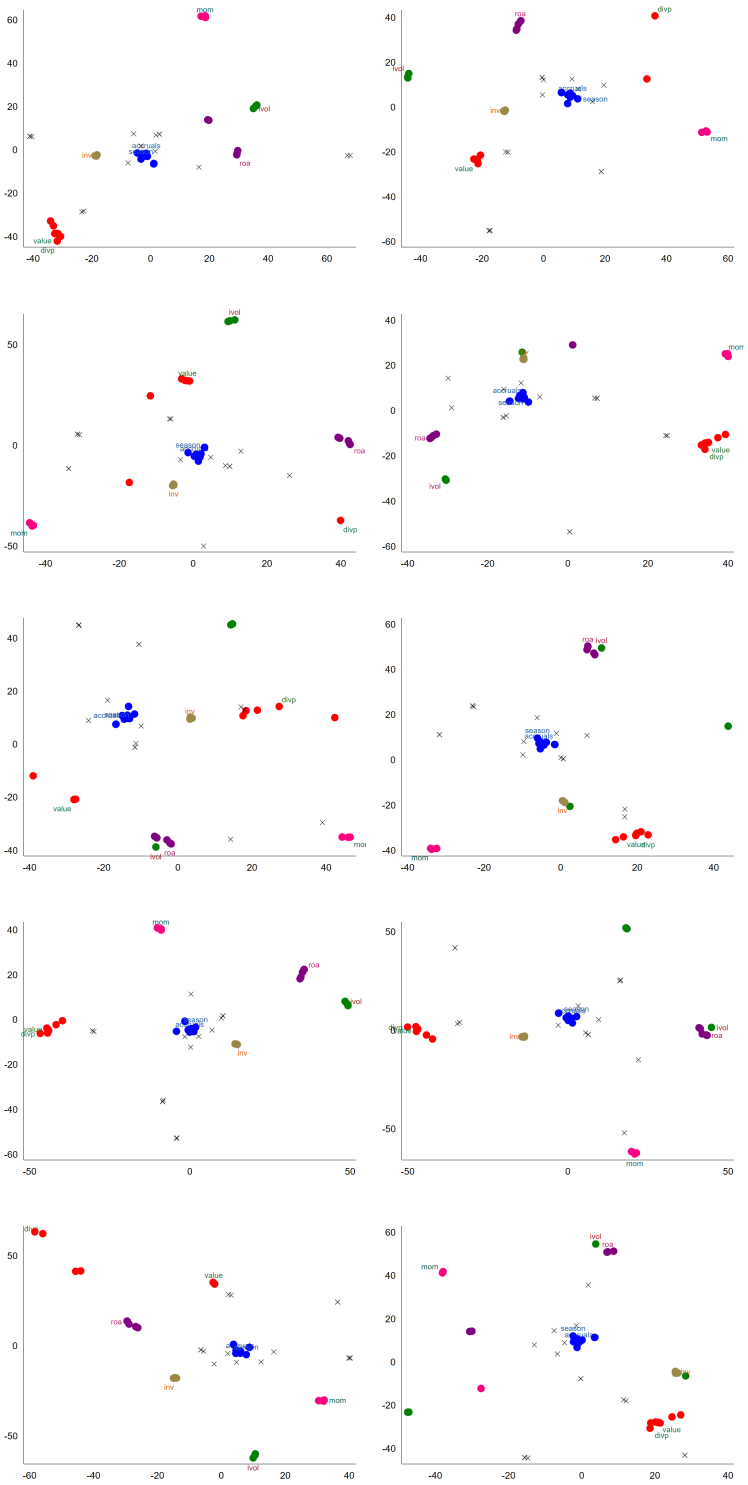


Figure A.2: Graphical Illustration of t-SNE – First Half

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The return profiles are based on the first half of the sample. The initial condition is set as the first two eigenvalues from the spectral clustering method following Kobak, Linderman, Steinerberger, Kluger, and Berens (2019). The groups are defined in the table based on the t-SNE procedure on the whole sample.

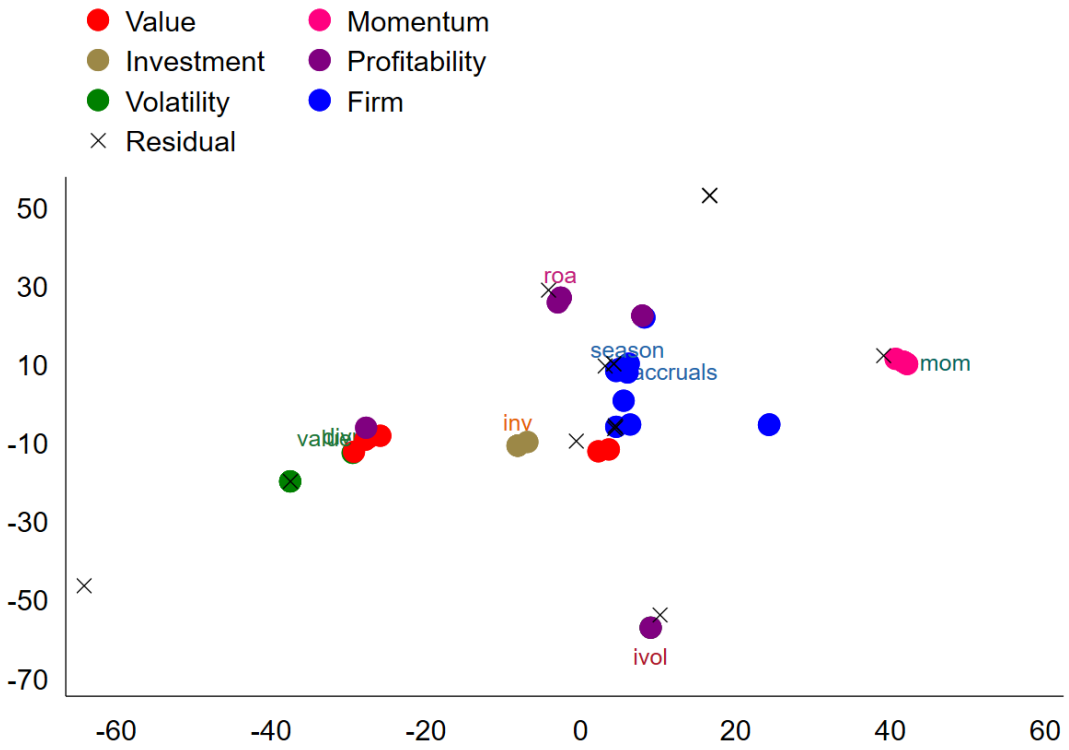


Figure A.3: Graphical Illustration of t-SNE – Second Half

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The return profiles are based on the second half of the sample. The initial condition is set as the first two eigenvalues from the spectral clustering method following Kobak, Linderman, Steinerberger, Kluger, and Berens (2019). The groups are defined in the table based on the t-SNE procedure on the whole sample.

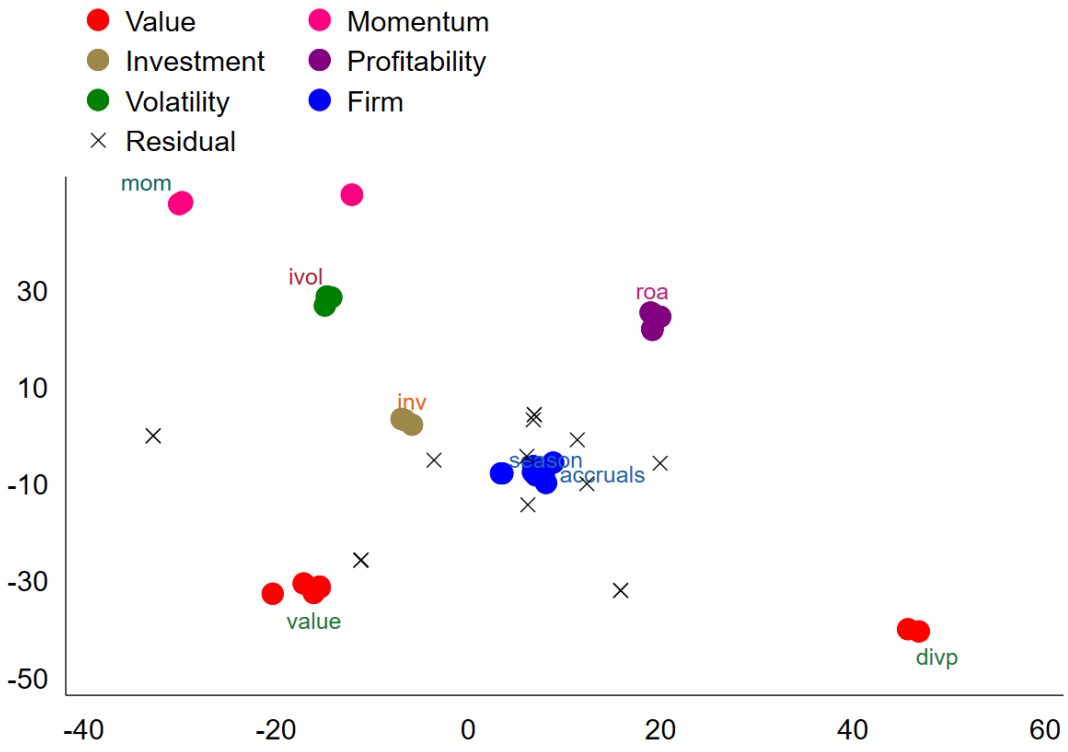
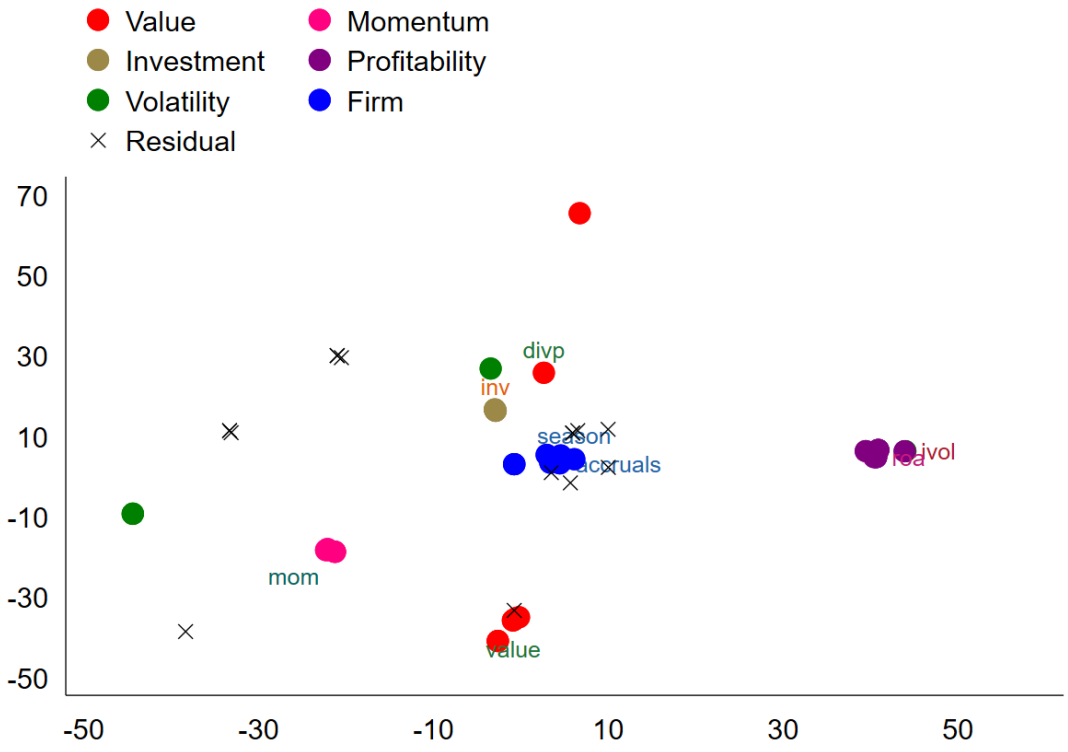


Figure A.4: Graphical Illustration of t-SNE – Monthly

This figure plots the two-dimensional representation of the 55 long-short strategies based on the t-SNE procedure. The return profiles are based on monthly returns of the full sample. The initial condition is set as the first two eigenvalues from the spectral clustering method following Kobak, Linderman, Steinerberger, Kluger, and Berens (2019). The groups are defined in the table based on the t-SNE procedure on the whole sample.



Appendix: Mathematical Aspects of PCA/t-SNE

In this Appendix, we briefly describe some of the issues that may arise with the application of the PCA to cluster high-dimensional data and broadly describe the mathematical background behind the use of t-SNE as a clustering tool. For more details, we refer to Shaham and Steinerberger (2017) and Linderman and Steinerberger (2019).

PCA

PCA is, for many reasons, a canonical tool in the analysis of high-dimensional data. A non-exhaustive list of its strengths are

1. **Linearity.** The entire method is based on linear structures which allows a complete analysis in terms of Linear Algebra. This has resulted in a complete, rigorous and substantial theoretical understanding.

2. **Simplicity.** The process of computing PCA is fast and robust.

3. **Geometry.** The underlying idea of PCA approximating existing data by the best subspaces leads to a natural interpretation of the underlying process. In particular, projecting onto low-dimensional approximating subspaces is natural way of reducing the dimensionality of existing data.

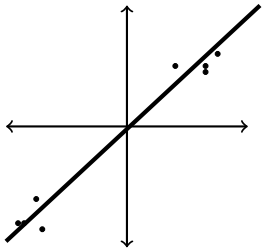


Figure A.5: PCA tries to find the best approximation by a hyperplane of a fixed size.

Given these (and many other) reasons, the reason for why it may be preferable to use other methods than PCA is subtle: high-dimensional objects, even fairly well-behaved ones, are subjected to an underlying distortion of lengths thats simply do not possess an accurate representation in two dimensions that arises from linear methods – indeed, if one is allowed to use PCA with a large number of embedding dimensions, then it remains a valuable and often used tool.

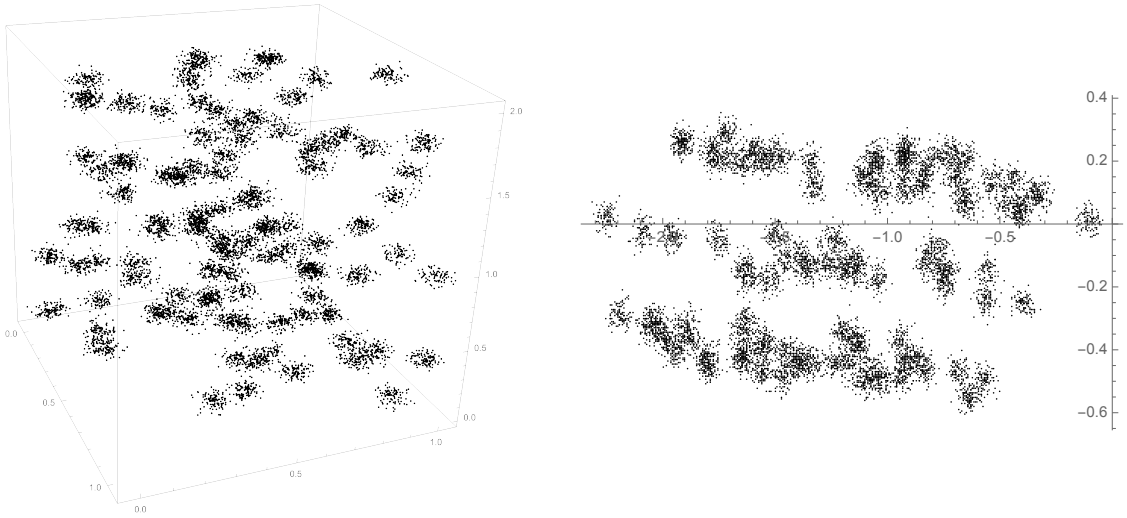


Figure A.6: 100 clusters in the three-dimensional unit cube (left) embedded by PCA into two dimensions (right).

However, it is not designed to detect clusters and may fail to do so for a variety of reasons. The simplest general reason is that high-dimensional structures (even clusters) cannot be well approximated by low-dimensional structures. A simple example is shown in Fig. 2: we have 100 well-defined clusters in $\mathbb{R}^3$, using PCA to embed it into $\mathbb{R}^2$ gives rise to continuous segments. We emphasize that this example is actually *very* benign: three dimensions are not exactly high-dimensional.

PCA applied to a high-dimensional sphere

We now describe one striking such example that can be done in closed form. Let us suppose we use PCA to embed the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ into one dimension. By rotational symmetry, we can assume without loss of generality, that this is given by the map

$$\pi : (x_1, \ldots, x_d) \to x_1. \tag{10}$$

Suppose now furthermore that we are given N points $\{x_1, \ldots, x_d\}$ chosen randomly from the sphere $\mathbb{S}^{d-1}$, what is the distribution of $\{\pi x_1, \ldots, \pi x_d\} \subset \mathbb{R}^d$? A moments consideration shows that the density at z is given by the volume of the codimension one slice $\pi x = z$. Thus, the density $f_d(x)$ of embedding the unit sphere in $d-$dimensions into $\mathbb{R}$ is given by

$$f_d(x) = c_d \left(1 - x^2\right)^{\frac{d-1}{2}}, \tag{11}$$

where c_d is a normalization constant. When projecting the two-dimensional sphere into one dimension, we obtain a familiar density. However, when embedding the unit sphere in 100 dimensions,

the shape is strikingly different.

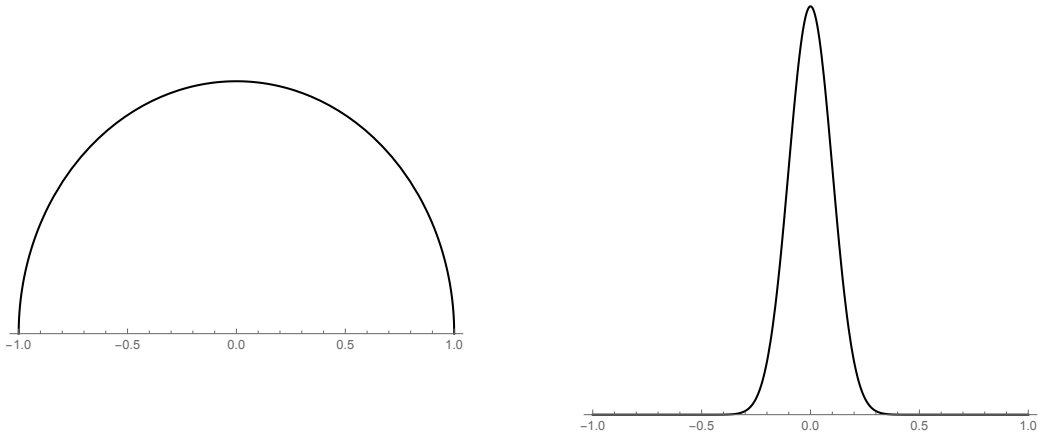


Figure A.7: 100 clusters in the three-dimensional unit cube (left) embedded by PCA into two dimensions (right).

We see that the shape changes from the semicircular shape to something that looks more like a Gaussian. Indeed, recalling the definition

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n, \tag{12}$$

we see that the Gaussian does indeed naturally arise in the limit

$$\lim_{d\to\infty} \frac{1}{c_d} f_d\left(x/\sqrt{d}\right) = \lim_{d\to\infty} \left(1 - \frac{x^2}{d}\right)^{\frac{d-1}{2}} = e^{-x^2/2}. \tag{13}$$

We also observe that a typical coordinate of an element on a d-dimensional sphere is $\sim 1/\sqrt{d}$ which starts being much smaller than 1. These and other phenomena in high dimensions have, in recent years, led to a renewed understanding in the analysis of high-dimensional objects: *it is exceedingly difficult to work with the global geometry of high-dimensional data.* An even more striking fact is that even for fairly clustered high-dimensional data, the distance between elements within one cluster may be hard to distinguish from the distance to elements from different clusters – the standard Euclidean distance is subject to strong distortion and loses accuracy. This leads us to the following broad conclusion that emerged that to get a good representation of data as little as possible on the high-dimensional data itself.

t-SNE

A Sketch of t-SNE

t-SNE is a method that is firmly grounded in this broad conclusion. Denote the d-dimensional input dataset by $\mathcal{X} = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$, the only data that t-SNE uses is the distances between these elements to create a sketch of the local connection. More precisely, t-SNE computes the joint probability p_{ij} measuring the similarity between x_i and x_j via

$$p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)} \qquad \text{and} \qquad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}. \tag{14}$$

We observe that this is a truly non-linear process: σ is chosen so that most $p_{i|j}$ are close to 0: customarily, we want for each i to only have ~ 90 nonzero connections. This serves as a way of representing the local structure and this is the only thing that is being used: no global geometry is ever being considered, the method tries to get an accurate view of each single data point and its immediate neighbors. The main idea is then to aim to represent a two-dimensional set of points by the same philosophy: for any set $\mathcal{Y} = \{y_1, y_2, ..., y_n\} \subset \mathbb{R}^2$ we define

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{15}$$

and then aim to have the q_{ij} be as close as possibly to the p_{ij}. We note that while the p_{ij} are defined via a very quickly decaying kernel, the q_{ij} are defined via a rather slowly decaying the kernel (slower still in the variant that we use). This is one way that we can try to represent high dimensions into low dimensions. T-SNE then proceeds by moving the points in the two-dimensional space around in such a way that the distributions become as similar as possible, i.e. the goal is to have

$$\sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \rightarrow \min \tag{16}$$

While this was the original motivation, more recent results have underlined that one could easily assume a different view that comes from a suitable interpretation of the energy functional. Differentiating the energy functional with respect to a point y_i, we obtain

$$-\frac{1}{4}\frac{\partial C}{\partial y_i} = \sum_{j\neq i} p_{ij}q_{ij}Z(y_j - y_i) - \sum_{j\neq i} q_{ij}^2 Z(y_j - y_i), \tag{17}$$

where

$$Z = \sum_{k\neq l}\left(1 + \|y_k - y_l\|^2\right)^{-1} \tag{18}$$

is a constant depending on the global geometry of the embedding. We emphasize that this gradient has a geometric interpretation: using the gradient descent, t-SNE will move y_i in the direction

$A + B$, where

$$A = \sum_{j \neq i} p_{ij} q_{ij} Z(y_j - y_i) \qquad \text{is the attraction term and}$$
$$B = -\sum_{j \neq i} q_{ij}^2 Z(y_j - y_i) \qquad \text{is the repulsion term.}$$

We observe that A is an average of nearby points j that are connected to i in the original high-dimensional data (this is achieved by including the p_{ij} term which is typically very close to 0 or exactly 0 unless x_i and x_j are close in high dimensions). In short, the term A ensures that y_i moves towards points y_j if x_i and x_j are close to each other in high-dimensional space. The second term is different: there is no p_{ij}, it is term that is solely determined by $\{y_1, \ldots, y_n\} \subset \mathbb{R}^2$ and it ensures that each point would like to be far away from other points. What ensues is an interplay between attraction between points that truly are similar in an environment where points, generally, do not want to be close to one another.

The Abstract Framework

We can interpret this in a more abstract framework as suggested by (Linderman and Steinerberger, 2019). More precisely, t-SNE belongs to the following class of dynamical systems on point sets in $\mathbb{R}^2$. Let $z_1, \ldots, z_n \in \mathbb{R}^2$. We use them as initial values for a time-discrete dynamical system that is defined via

$$z_i(t+1) = z_i(t) + \sum_{j=1}^{n} \alpha_{i,j,t}(z_j(t) - z_i(t)) - \sum_{j=1}^{n} \beta_{i,j,t}(z_j(t) - z_i(t))$$
$$z_i(0) = z_i$$

The coefficients $\alpha_{i,j,t}$ and $\beta_{i,j,t}$ are allowed to depend both on the pairs of particles involved as well as time (to allow the use of, say, q_i or Z). At this level of generality, there is no way of analyzing the behavior of the system: it is dramatically under-determined, there are more free variables than particles. However, it is possible to deduce some general rules by studying a special case. Let us assume that this dynamical system is created out of high-dimensional data that is comprised of one single cluster. A way to make this precise is to make the following assumptions

1. There is a uniform lower bound on the coefficients for all $t > 0$ and all $i \neq j$

$$|\alpha_{i,j,t}| \geq \delta > 0. \tag{19}$$

2. There is a uniform upper bound on the coefficients

$$\sum_{j=1}^{n} \alpha_{i,j,t} \leq 1. \tag{20}$$

3. There is a uniform upper bound on the error term

$$\left| \sum_{j=1}^{n} \beta_{i,j,t}(z_j(t) - z_i(t)) \right| \leq \varepsilon. \tag{21}$$

Under these assumptions, we see that there is a strong attractive force between any pair of points and, possibly, some repulsive force whose total size never exceeds ε. Under these assumptions, it is possible to show (see Linderman and Steinerberger 2019), the following two statements. The first shows that systems of this type can never expand into big regions: their convex hull is shrinking (up to the repulsion factor).

Lemma 1. *[Stability of the convex hull] With the assumptions above, we have*

$$\text{conv}\,\{z_1(t+1), z_2(t+1), \ldots, z_n(t+1)\} \subseteq \text{conv}\,\{z_1(t), z_2(t), \ldots, z_n(t)\} + B(0, \varepsilon), \tag{22}$$

However, it is also the case that if the points of the dynamical system are spread out over a vast region, then there is a contractive force.

Lemma 2. *[Contraction inequality] With the notation above, if the diameter is large*

$$\text{diam}\,\{z_1(t), z_2(t), \ldots, z_n(t)\} \geq \frac{10\varepsilon}{n\delta}, \tag{23}$$

then

$$\text{diam}\,\{z_1(t+1), z_2(t+1), \ldots, z_n(t+1)\} \leq \left(1 - \frac{n\delta}{20}\right) \text{diam}\,\{z_1(t), z_2(t), \ldots, z_n(t)\}. \tag{24}$$

These two results combined show that dynamical systems of this type concentrated in a certain area whose size is governed by the scales of attraction (δ) and repulsion (ε). As for the more general case, we observe that the case of more than one cluster essentially decouples into dynamical systems comprised of individual clusters trying to contract in a small region coupled with repulsion between these clusters. We conclude by emphasizing that while t-SNE has received tremendous attention as a valuable tool for clustering and visualization, a fully rigorous mathematical analysis is still in its nascency (though, as shown in these remarks, there is a clearer and clearer picture emerging).