Department of Economics- FEA/USP

Risk Factor Centrality and the Cross-Section of Expected Returns

FERNANDO MORAES

RODRIGO DE-LOSSO

WORKING PAPER SERIES Nº 2020-17

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2020-17

Risk Factor Centrality and the Cross-Section of Expected Returns

Fernando Moraes (fernandotm@al.insper.edu.br)

Rodrigo De-Losso (delosso@usp.br)

Research Group: NEFIN

Abstract: The Factor Zoo phenomenon calls for answers as to which risk factors are in fact capable of providing independent information on the cross-section of expected excess returns, while considering that asset-pricing literature has produced hundreds of candidates. In this paper, we propose a new methodology to reduce risk factor predictor dimensions by selecting the key component (most central element) of their precision matrix. Our approach yields a significant shrinkage in the original set of risk factors, enables investigations on different regions of the risk factor covariance matrix, and requires only a swift algorithm for implementation. Our findings lead to sparse models that pose higher average in samples !" and lower root mean square out of sample error than those attained with classic models, in addition to specific alternative methods documented by Factor Zoo-related research papers. We base our methodology on the CRSP monthly stock return dataset in the time frame ranging from January 1981 to December 2016, in addition to the 51 risk factors suggested by Kozak, Nagel, and Santosh (2020).

Keywords: Risk factors, factor zoo, graph lasso, network analysis.

JEL Codes: G12, C55, D85.

Risk factor centrality and the cross-section of expected returns

Fernando Moraes[1] Rodrigo De-Losso[2]

September 11, 2020

Abstract

The Factor Zoo phenomenon calls for answers as to which risk factors are in fact capable of providing independent information on the cross-section of expected excess returns, while considering that asset-pricing literature has produced hundreds of candidates. In this paper, we propose a new methodology to reduce risk factor predictor dimensions by selecting the key component (most central element) of their precision matrix. Our approach yields a significant shrinkage in the original set of risk factors, enables investigations on different regions of the risk factor covariance matrix, and requires only a swift algorithm for implementation. Our findings lead to sparse models that pose higher average in samples R^2 and lower root mean square out of sample error than those attained with classic models, in addition to specific alternative methods documented by Factor Zoo-related research papers. We base our methodology on the CRSP monthly stock return dataset in the time frame ranging from January 1981 to December 2016, in addition to the 51 risk factors suggested by Kozak, Nagel, and Santosh (2020).

Keywords: Risk factors, factor zoo, graph lasso, network analysis.

JEL Codes: G12, C55, D85.

[1] E-mail: fernandotm@al.insper.edu.br.

[2] E-mail: delosso@usp.br.

1. Introduction

The Factor Zoo phenomenon calls for answers as to which risk factors are in fact capable of providing independent information on the cross-section of expected excess returns, while considering that asset-pricing literature has produced over three hundred different potential risk factors in the last decades (Harvey, Liu, and Zhu (2016)). Solving this "factor zoo" (Cochrane (2011)) implies studying, on one side, which of these 350+ factors provide independent information about the cross-sectional variation of expected returns and, on the other, which are redundant. In fact, this increasing number of potential factors also brought about new methodological challenges for empirical research (i.e. overfitting, data mining, and design matrix dimension reduction)[3]. Solutions for these methodological challenges were constrained oftentimes by excessively costly approaches resulting from their need for increased computational power. Nonetheless, as witnessed in the last years, computational power's decreasing costs, in addition to the rise in alternative statistical methods geared towards handling high-dimensional problems[4] has allowed researchers to start addressing the high-dimensional[5] challenges that relate to the factor zoo.[6]

In this paper, we propose a new methodology to reduce risk factor predictor dimensions by selecting the key component[7] of their precision matrix[8]. In order to implement this procedure, we use a graph to represent the risk factor precision matrix, which describes how information networks and risk factors are related. After attaining the graph precision matrix, we proceed to select its key component in accordance with a specific centrality measure. Those key components become our new risk factor candidates set to explain the cross-sectional of expected returns.

Our findings attain sparse models that pose better results than classic models documented in the literature as well as specific alternative methods proposed by factor

[3] Hastie, Tibshirani, and Friedman (2009) and Abu-Mostafa, Magdon-Ismail, and Lin (2012).

[4] Some of those methods are dubbed 'Machine Learning' techniques. To learn more, see Hastie, Tibshirani, and Friedman (2009).

[5] For definition purposes, we consider high-dimensional environmental models yielding more than 10 predictors.

[6] Harvey, Liu, and Zhu (2016), Green, Hand, and Zhang (2017), Yan and Zheng (2017), Feng, Giglio, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020) and Kozak, Nagel, and Santosh (2020).

[7] A key component of the precision matrix is the most central element, according to some centrality measure, of the adjacent matrix whom is computed by the precision matrix.

[8] The precision matrix is given by the inverse of the covariance matrix for any given vector of random variables.

zoo papers. Our model achieves a higher in-sample average adjusted R^2 from the Fama and MacBeth procedure and more significant factors risk premia parameters than the Principal Components – PC – peer's methodology. Regarding out-of-sample results, our paper presents the lowest root mean square errors than all other tested alternative factor zoo methodologies presented in this study.

It is important to point out two things from this framework: First, by assuming that $\boldsymbol{f} = (f_1, \dots, f_P)' \sim MN\left(\boldsymbol{\mu}_f, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}\right)$, where $\boldsymbol{f}$ is a $(\mathrm{P} \times 1)$ risk factor vector, and $\boldsymbol{\Sigma}$ is a positive semidefinite matrix, the $\boldsymbol{\Theta}$ (precision matrix) has the following property[9]:

$$\theta_{i,j} = 0 \Leftrightarrow f_i \perp f_j \left| \{\boldsymbol{f}\} \setminus \{f_i, f_j\} \right. \quad (1)$$

Equation (1) shows that f_i is independent (orthogonal) to f_j, conditional upon every other risk factor if, and only if, the $\theta_{i,j}$ element of $\boldsymbol{\Theta}$ is zero. Thus, $\boldsymbol{\Theta}$ yields the conditional dependence in risk factors, and can be represented by a graph that illustrates a network. Second, as pointed out by Borgatti (2005), a network's key element is the component that best condenses information about that network individually. Thus, since our network is the precision matrix, its key component is best suited to summarize information about the risk factor covariance matrix. It therefore yields the highest conditional dependence among every other factor, reason why it is analogous to the $PC1$ (first principal component) and, consequently, is a natural candidate to explain the cross-section of expected returns. Furthermore, this methodology poses an advantage since we are able to select a specific risk factor which, unlike PCs, entails an economic interpretation. Another benefit from this methodology stems from the fact that it enables us to clusterize the graph into sub regions, in addition to selecting the key component of each of these regions. The precision matrix is therefore divided into partitions, allowing us to compare factors selected globally to factors selected regionally. This also enables us to infer which regions of the covariance matrix are best suited to explain the cross-section of expected returns.

In our methodology, high-dimensional problems are solved using two shrinkage steps. First, we estimate the precision matrix with the Graphical Lasso methodology proposed by Friedman, Hastie, and Tibshirani (2008) aimed at avoiding Markowitz's curse, which is a well-known documented phenomenon in empirical asset-pricing

[9] For the proof, see Meinshausen and Bühlmann (2006).

literature[10]. Basically, as the number of factors increases, the conditional number[11] of $\boldsymbol{\Theta}$ also rises, and $\boldsymbol{\Theta}$ becomes singular. Since the covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}$, whereas $\boldsymbol{\Sigma}$ must be a positive semidefinite, we therefore need a sparse estimation for $\boldsymbol{\Theta}$[12]. When looking at equation (1), as Friedman, Hastie, and Tibshirani (2008) argue, we see that enforcing the L_1 penalty results in a $\boldsymbol{\Theta}$ sparsity estimation. Second, selecting the key component of the precision matrix reduces dimensions of $\boldsymbol{f}$ from P to 1. The alternative methodology, on the other hand, consists of selecting the component of each precision matrix cluster whereby lowering dimensions of $\boldsymbol{f}$ sets P to m, where m is the number of clusters.

Our paper resonates very closely with this new literature about high-dimensional cross-sectional asset-pricing models. This research field applies a wide range of statistical methods such as bootstrapping[13], lasso, multiple-test corrections[14] and principal component analysis[15] to achieve robust estimators in a high-dimensional environment, in addition to evaluating which risk factors are in fact capable of explaining the cross-section of expected returns. Kozak, Nagel, and Santosh (2020) find that a stochastic discount factor – SDF – with a small number of principal components from large-set zero-cost portfolio returns is capable of explaining cross-sectional returns. This research applies elastic net estimator to perform dimensionality reduction on the set of risk factors, and finds that a sparsity model only achieves satisfactory results to explain cross-sectional returns when the principal components of portfolio returns are used as risk factors. The authors also point out that a non-sparse model (up to 15 explanatory variables) is needed whenever risk factors are applied to explain cross-sectional returns. Feng, Giglio, and Xiu (2020) and Freyberger, Neuhierl, and Weber (2020) use a lasso-type estimator to reduce their respective set of risk factors, also reaching a non-sparse result whenever the issue at

[10] Ledoit and Wolf (2004), Engle, Shephard, and Sheppard (2008) and Brito, Medeiros, and Ribeiro (2018).

[11] Given a matrix **A,** the conditional number can be expressed as $\text{cond}(\mathbf{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}\|^{-1}$, whereas $(\mathbf{A}) = \infty$ if **A** is singular.

[12] See De Prado (2018) for a thorough discussion on empirical instability results concerning the covariance matrix estimator for a large set of securities.

[13] Harvey and Liu (2019) and Yan and Zheng (2017) apply bootstrapping techniques to evaluate models.

[14] Harvey, Liu, and Zhu (2016) and Green, Hand, and Zhang (2017) implement a threshold adjustment in the empirical test with the purpose of avoiding false-positive discoveries as well as data mining.

[15] Kelly, Pruitt, and Su (2019) develop a latent factor model with time-varying loadings instrumented by a large set of characteristics called Instrumented Principal Component Analysis – IPCA. The authors come to a sparse model after setting a low dimension for the factor vector, consequently concluding that only five latent factors provide satisfactory results in regards to explaining average cross-sectional returns. Gu, Kelly, and Xiu (2019) delve even deeper into this matter and implement a non-linear IPCA that improves out-of-sample results.

stake concerns explaining the cross-section of expected returns. Hence, we are witnessing the emergence of a well-known fact whereby the cross-section of expected returns can only be adequately described by PCs in a sparse representation. Whenever we apply risk factors, it becomes increasingly challenging to explain cross-sectional expected returns in a satisfactory manner by a sparse model.

This paper also aims to contribute to the applied network theory documented in financial literature[16]. Network analysis were applied to several fields to tackle economic problems, such as teaching methods, labor markets and banking and investment decisions[17], and it has become an increasingly popular subject. Considering asset-pricing research, we must acknowledge some prominent papers, such as Herskovic (2018), which uses input and output transactions to explain network relationships among firms and, furthermore, recognizes that more central firms require less risk premium.

In short, this paper seeks to add a new method to the existing factor zoo-related literature, thereby enabling a significant shrinkage in the original set of risk factors and allowing investigations on different regions of the risk factor covariance matrix, all of which can be implemented with a swift algorithm. To the best of our knowledge, this is also the first paper that uses a graph model to describe joint risk factor distribution, in addition to using precision matrix network analysis to select risk factor candidates.

2. Methodology

Our research method is described in four steps. First, we estimate the risk factor precision matrix by applying the Graphical Lasso algorithm. Second, we use a graph to represent the estimated precision matrix and to select the key component risk factor as our candidate to explain the cross-section of expected returns. Third, we proceed to partition the precision matrix into regions by clustering the graph, after which we pick the key component of each cluster as the risk factor candidate. Fourth, and last, we employ the Fama-MacBeth (1973) procedure to verify whether risk factors selected by both the second (global models) and third (cluster models) steps entail better in and out-of-sample results compared to certain "classic models" from the asset- pricing literature, in addition

[16] Allen and Babus (2009) provide an in-depth survey on financial network research and discuss potential network analysis tools to address financial problems.

[17] Calvo-Armengol and Jackson (2004), Golub and Jackson (2010), Elliott, Golub, and Jackson (2014), Gofman (2017), Hochberg, Ljungqvist, and Lu (2007) and Cohen, Frazzini, and Malloy (2008).

to other methodologies proposed by research papers addressing the factor zoo that we describe ahead.

2.1.Estimating the precision matrix with the graph lasso

There are countless papers on methods used to estimate a sparse inverse covariance matrix with L_1 (lasso) regularization[18]. Oftentimes, the basic model employed assumes a multivariate Gaussian distribution as $\boldsymbol{f} = (f_1, \dots, f_P)' \sim MN(\boldsymbol{\mu}_f, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$, whereas, in our case, $\boldsymbol{f}$ is a $(\mathrm{P} \times 1)$ vector of observed risk factors. Consequently, the graph lasso estimator for the precision matrix can be estimated by maximizing the penalized log-likelihood expressed as (2):

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{f}) = ln[det(\boldsymbol{\Theta})] - tr[\boldsymbol{S}(\boldsymbol{f})\boldsymbol{\Theta}] - \tau\|\boldsymbol{\Theta}\|_1 \qquad (2)$$

where $\boldsymbol{S}(\boldsymbol{f})$ is the empirical covariance matrix[19], $\|\boldsymbol{\Theta}\|_1$ is the L_1 norm of $\boldsymbol{\Theta}$, and τ is the regularization parameter.

We apply the procedure proposed by Friedman, Hastie, and Tibshirani (2008), which is a simple yet swift algorithm used to solve (2), though it limits the solution to the positive semidefinite space parameter. Consequently, we need to embed the following assumption: $\boldsymbol{f} = (f_1, \dots, f_P)' \sim MN(\boldsymbol{\mu}_f, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$ which implies in $\boldsymbol{\Theta}$ being positive semidefinite. Furthermore, as equation (1) suggests, we are able to analyze factorization properties of $\boldsymbol{f}$ directly with the sparsity pattern of the precision matrix $(\boldsymbol{\Theta})$. We use network analysis to investigate these factorization properties since $\boldsymbol{\Theta}$ can be illustrated by a graph, as we describe in the following section.

Liu, Roeder, and Wasserman (2010) the *stability approach to regularization selection* method[20] (StARS) to set the τ parameter. According to the authors, whenever we consider the maximization issue that equation (2) poses, the StARS approach yields a better performance both in simulated and real data when compared to k-fold cross-validation, AIC and BIC methods. This can be explained by the fact that AIC and BIC assume a fixed number of parameters as the sample size increases, reason why results tend not to be suitable when the number of parameters is large in comparison to the sample size. Additionally, Wasserman and Roeder (2009) show that k-fold cross-

[18] Chapter 9 of Hastie, Tibshirani, and Wainwright (2015) brings a description on several methods, in addition to a review on the literature relating to the subject.

[19] $\boldsymbol{S}(\boldsymbol{f}) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t'$ for $t = 1, \dots, T$.

[20] Stability approach used for regularization selection.

validation tends to overfit data. The StARS methodology is described in three steps. First, we set a regularization parameter grid as $\mathcal{G}_\Lambda = (\Lambda_1, \ldots, \Lambda_K)$, where $\Lambda_k = 1/\tau_k$. Second, we generate N subsample s_n with $\boldsymbol{f}$, yielding $B = (s_1, \ldots, s_N)$. Third, we evaluate the grid in each subsample and select the regularization parameter that produces a sparse result for $\boldsymbol{\Theta}$, though subsamples fail to show high variability results. Appendix A1 describes the StARS approach in detail.

When we put it all together, our sparse precision matrix estimator can be written as:

$$\widehat{\boldsymbol{\Theta}}_{GLASSO} = argmin_{\boldsymbol{\Theta}\epsilon\boldsymbol{\Theta}^+} \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{f}, \tau_{\text{StARS}}) \tag{3}$$

where $\boldsymbol{\Theta}^+$ is the positive semidefinite space matrix, $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{f}, \tau)$ is defined as (2), and τ_{StARS} is the regularization parameter selected with the StARS approach.

2.2. Precision matrix representation by graph

A graph is represented as $\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$, where $\boldsymbol{\mathcal{V}}$ is the vertex set, and $\boldsymbol{\mathcal{E}}$ is the edge set. Elements of $\boldsymbol{\mathcal{V}}$ represent random variables, whereas elements of $\boldsymbol{\mathcal{E}}$ are pairs of distinct vertices (u, v) so that $u, v \in \boldsymbol{\mathcal{V}}$. We label graphs whose edge pairs have different orderings ((u, v) is different from (v, u)) as *directed*. If, however, the edge pairs do not have different orderings ((u, v) is not different from (v, u)), we classify the graph as *undirected*. An edge can describe any measurable characteristic across a pair of vertices; therefore, graph $(\mathcal{G})$ illustrates each specific network's relation to random variables pursuant to the definition given to the edge.

It is important to note that $\mathcal{G}$ can be grouped into an adjacency matrix $\boldsymbol{A}$ such that:

$$A_{i,j} = \begin{cases} 1 \; if \; (u, v) \in \boldsymbol{\mathcal{E}} \\ 0 \; \; otherwise \end{cases} \tag{4}$$

Since $\widehat{\boldsymbol{\Theta}}_{GLASSO}$ is symmetric ($\hat{\theta}_{GLASSOi,j} = \hat{\theta}_{GLASSOj,i}$ for any given $i, j \in P$), we show $\widehat{\boldsymbol{\Theta}}_{GLASSO}$ as an undirected[21] graph $\mathcal{G}_u = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ so that $\boldsymbol{f} = \boldsymbol{\mathcal{V}} \in \mathbb{R}^P$, and $(i, j) \in \boldsymbol{\mathcal{E}}$, if $\hat{\theta}_{GLASSOi,j} \neq 0$ for any $f_i, f_j \in \boldsymbol{f}$. Thus, $\widehat{\boldsymbol{A}} \in \mathbb{R}^{\mathrm{P}\times\mathrm{P}}$ can be written as:

$$\hat{A}_{i,j} = \begin{cases} 1 \; if \; \hat{\theta}_{GLASSOi,j} \neq 0 \\ 0 \quad otherwise \end{cases} \tag{5}$$

[21] It's worth pointing out that undirected graphs denote a symmetric adjacency matrix.

Once we are able to have $\widehat{\boldsymbol{\Theta}}_{GLASSO}$ represented as $\mathcal{G}_u$, we can proceed to compute centrality measures and select key component risk factor candidates to explain the cross-section of expected returns.

2.3. Centrality measures

Given any $\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$, a centrality measure is a function $\boldsymbol{c}$ represented as:

$$\boldsymbol{c}: \mathcal{G} \rightarrow \mathbb{R}^N \tag{6}$$

where $\mathcal{G}$ is a graph, N is the number of elements of $\boldsymbol{\mathcal{V}}$, c_p is the centrality measure of vertex p, $p \in \boldsymbol{\mathcal{V}}$, and $p = 1, \dots, N$.

Centrality measures is a well-studied concept in the network analysis literature. Borgatti (2005) seminal paper stresses that different measures entail different assumptions in regards to how information flows into the network, represented by $\mathcal{G}$. The concept of vertex position depends on the research focus and the context of the problem; thus, a specific application requires a specific centrality measure.

In the case at hand, we estimate $\mathcal{G}_u$, which represents the risk factor network stated as $\hat{\mathcal{G}}_u = \mathcal{G}_u\left(\boldsymbol{\mathcal{V}} = \boldsymbol{f}, \boldsymbol{\mathcal{E}}\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right)\right)$, meaning that vertex set variables are the risk factors, while adjacency matrix $\boldsymbol{A}$ is computed with (5). The only information that we have concerning estimated risk factor network $\left(\hat{\mathcal{G}}_u\right)$ is that risk factors f_i and f_j are conditional dependent (independent) if $A_{i,j} = 1$ ($A_{i,j} = 0$), given every other risk factor. Nevertheless, this does not allow us to infer any causality relationships between f_i and f_j. Consequently, the way information flows into the network remains unknown, and there are no *a priori* assumptions about it. This forces us to compute different types of centrality measures in order to select our risk factor candidate, while also comparing results among them.

In this paper, we select four well-known centrality measures in the network analysis field: Eigenvector; Degree; Closeness; and Betweenness. As demonstrated by Bloch, Jackson, and Tebaldi (2019), said measures have the same logical structure[22], the only difference being which vertex attributes are taken into consideration to compute the

[22] The authors also prove that each of the four measures features the three following axioms: i) monotonicity, implying that higher statistical figures lead to higher centrality; ii) symmetry, meaning that vertices' centrality relies only on their chosen attributes as opposed to other traits; and iii) additivity, denoting that centrality measures are computed in an additively separable manner.

results. Consequently, calculating (6) since $\boldsymbol{\mathcal{V}} = \boldsymbol{f}$, and $\boldsymbol{f} \in \mathbb{R}^P$ implies that $N = P$, whereas, in our case, $\boldsymbol{c}(\mathcal{G}) \in \mathbb{R}^P$. Each centrality measure is described below.

For each centrality measure we have $\boldsymbol{c}(\hat{\mathcal{G}}_u) = (c_1, \dots, c_P)^{\boldsymbol{T}}$, where c_p is the centrality measure of risk factor f_p, while the selected risk factor (f_S) (key component) is expressed as $f_S = \left\{f_p : c_p = max\left(\boldsymbol{c}(\hat{\mathcal{G}}_u)\right)\right\}$. Thus, we are able to reduce the original vector dimension of $\boldsymbol{f}$ from P to 1.

2.3.1. Eigenvector centrality

$$c_p^e = \frac{1}{\lambda_e}\sum_{j=1}^{P} A_{p,j}c_j^e \tag{7}$$

where $\boldsymbol{A}$ is the adjacent matrix for $\mathcal{G}$, and λ_e is the large eigenvalue of $\boldsymbol{A}$. It measures the influence a risk factor has on its neighborhood, adjusted by the neighborhood's influence on the network.

2.3.2. Degree centrality

$$c_p^d = \frac{d_p}{P-1} \tag{8}$$

where $d_p = \sum_{j=1}^{N} A_{p,j}$ is the degree of vertices p. It measures the influence a risk factor has on its neighborhood, disregarding the factor position in the network architecture.

2.3.3. Closeness centrality

$$c_p^c = \frac{P-1}{\sum_{j=1}^{P} \rho_g(p.j)} \tag{9}$$

where $\rho_g(p.j)$[23] is the distance between vertices p and node j. It measures how well a risk factor is positioned to obtain novel (random) information early on from the network.

2.3.4. Betweenness centrality

$$c_p^b = \frac{2}{(P-1)(P-2)}\sum_{j=1}^{P}\sum_{i=1}^{P}\frac{\nu_g^{(p)}(j,i)}{\nu_g(j,i)} \tag{10}$$

[23] The distance from vertex p to j in a graph is the number of edges in the shortest path connecting them. A path from vertex p to j is a sequence of distinct edges, which joins a sequence of vertices from vertex p to j. See Bavelas (1950) for a formal definition.

where $v_g(j,i)$ is the number of paths from vertices i to vertices j, and $v_g^{(p)}(j,i)$ is the number of paths from vertices i to vertices j, passing through p. It measures the importance a risk factor has in conveying information across the network.

2.4. Clustering graph by modularity

A common feature in the network analysis is the search for community structures across all random variables. In our graph, such community structures can be described as a "cohesive" subset of vertices, which occurs inside dense connections as well as in sparse associations with other vertices. In other words, finding community structures consists of clustering $\mathcal{G} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ into a partition set $\mathcal{M} = (g_1, \dots, g_M)$, where $g_m = (\boldsymbol{\mathcal{V}}^m, \boldsymbol{\mathcal{E}}^m)$ is also a graph; M is the number of clusters; $\bigcap_{m=1}^{M} \boldsymbol{\mathcal{V}}^m = 0$, $\bigcup_{m=1}^{M} \boldsymbol{\mathcal{V}}^m = \boldsymbol{\mathcal{V}}$; and the number of edges of vertices inside $\boldsymbol{\mathcal{V}}^m$ should ideally be higher than the number of edges of vertices between $\boldsymbol{\mathcal{V}}^m$ and $\boldsymbol{\mathcal{V}}^n$ for $m, n = 1, \dots, M$, and $m \neq n$.

Among the numerous existing methods used to clusterize[24] $\mathcal{G}$, we favor the modularity approach proposed by Newman and Girvan (2004a) since it is a very popular and well-established procedure. The number of clusters in this method is endogenous, and therefore, an advantage, since we lack prior information about them. Moreover, this method has been successfully applied across a broad range of different problems and networks[25].

The approach consists of maximizing the modularity function:

$$Q = \frac{1}{2n} \sum_{i=1}^{P} \sum_{j=1}^{P} \left[\boldsymbol{A}_{i,j} - \frac{k_i k_j}{2n} \right] \delta\left(\boldsymbol{\mathcal{V}}^i, \boldsymbol{\mathcal{V}}^j\right) \tag{11}$$

where $n = \#(\boldsymbol{\mathcal{E}})$ is the number of edges; $k_i = \sum_{j=1}^{P} A_{i,j}$ is the degree number of f_i, $\delta\left(\boldsymbol{\mathcal{V}}^i, \boldsymbol{\mathcal{V}}^j\right) = 1$ if i $= j$ and, otherwise, equal to 0; $\boldsymbol{\mathcal{V}}^i$ is a partition from $\boldsymbol{\mathcal{V}}$; and $f_i \in \boldsymbol{\mathcal{V}}^i$. Equation (11) allows us to verify that the modularity function measures how far any given network community's structure is from the randomized structure. The first term $(\frac{1}{2n} \sum_{i=1}^{P} \sum_{j=1}^{P} [\boldsymbol{A}_{i,j}] \delta\left(\boldsymbol{\mathcal{V}}^i, \boldsymbol{\mathcal{V}}^j\right))$ relates to the fraction of the edges that belong within the communities. Since $\frac{k_i k_j}{2n}$ is the probability of the edge, (i,j) only exists if connections are randomly made, taking into account vertices f_i, and f_j degrees (k_i and k_j). The second

[24] See Lancichinetti and Fortunato (2009) for a survey on cluster graph methods.
[25] See Khan and Niazi (2017) for a survey on network community detection.

term $\frac{1}{2n}\sum_{i=1}^{P}\sum_{j=1}^{P}\left[\frac{k_i k_j}{2n}\right]\delta(\boldsymbol{\mathcal{V}}^i, \boldsymbol{\mathcal{V}}^j)$ concerns the fraction of the edges that belong within the communities for an expected randomized network due to the vertex degree. Hence, a high Q denotes deviation from randomness and signals community structures.

By setting $\boldsymbol{\mathcal{V}}^* = (\boldsymbol{\mathcal{V}}^1, \dots, \boldsymbol{\mathcal{V}}^m)$[26], the cluster solution based on modularity can be expressed as $\widehat{\boldsymbol{\mathcal{V}}}^* = argmax_{\{\boldsymbol{\mathcal{V}}^*, M\}}Q$, and $\widehat{\mathcal{M}} = (\widehat{\mathcal{G}}_1, \dots, \widehat{\mathcal{G}}_M)$, where $\widehat{\mathcal{G}}_m = \left(\widehat{\boldsymbol{\mathcal{V}}}^m, \boldsymbol{\mathcal{E}}^m\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right)\right)$[27]. To maximize Q with respect to $\boldsymbol{\mathcal{V}}^*$ and M**,** we apply the Clauset, Newman, and Moore (2004) algorithm, which is swift and enables large-scale dimensions for both $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{E}}$ (large networks)**.**

In regards to our research, the graph's distribution seeks to divide the precision matrix into regions. This, in turn, enables us to investigate the role of different covariance regions and explain the cross-section of expected returns, in addition to comparing risk factors selected globally to risk factors chosen regionally. After maximizing (11) into $\widehat{\mathcal{M}} = (\widehat{\mathcal{G}}_1, \dots, \widehat{\mathcal{G}}_M)$, and consequently clustering $\hat{\mathcal{G}}_u$ in accordance with each centrality measure, we proceed to select the risk factor vector with $\boldsymbol{f}_{SK} = \left(f_{1,p}, \dots, f_{M,p}\right)'$, where $f_{m,p} = \left\{f_p : c_p = max\left(\boldsymbol{c}(\widehat{\mathcal{G}}_m)\right)\right\}$ for $m = 1, \dots, M$. Thus, our risk factor vector $\boldsymbol{f}_{SK}$ has, by definition, m and $m \leq P$ dimensions[28].

2.5. Fama-MacBeth procedure (FM)

Using the APT model introduced by the Stephen (1976) paper, we are able to represent the cross-section of expected returns as:

$$r_{i,t} = a_i + \boldsymbol{\beta}_i \boldsymbol{f}_t + \varepsilon_t \tag{12}$$

$$\boldsymbol{E}\left(r_{i,t}\right) = \boldsymbol{\beta}_i \boldsymbol{\lambda} \tag{13}$$

where $r_{i,t}$ is the excess asset returns i observed for period t; $\boldsymbol{f}_t$ is the vector of selected risk factors observed in period t; $\boldsymbol{\beta}_i$ stands for the loading's i asset matrix for risk factors; and $\boldsymbol{\lambda}$ is the vector of selected factor risk premia. As demonstrated by Cochrane (2009), risk factors occur in the stochastic discount factor (inside investors' marginal utility)

[26] Note that $\cap_{m=1}^{M}\boldsymbol{\mathcal{V}}^m = 0$, $\cup_{m=1}^{M}\boldsymbol{\mathcal{V}}^m = \boldsymbol{\mathcal{V}}$ by definition.

[27] $\boldsymbol{\mathcal{E}}^m\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right) = \boldsymbol{\mathcal{E}}\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right) \setminus \left\{(i,j): j \in \widehat{\boldsymbol{\mathcal{V}}}^{m\,C}\right\}$, where $\boldsymbol{\mathcal{V}}^{m\,C}$ is the complement of set $\boldsymbol{\mathcal{V}}^m$.

[28] In some cases that use this method, m may be close to P, in which case it will not be suitable for use as a shrinkage procedure.

whenever factor risk premia are different from zero. Thus, in order to verify whether our methodology is in fact capable of selecting risk factors (f_S and $\boldsymbol{f}_{SK}$) that explain the cross-section of expected returns ($\boldsymbol{E}(r_{i,t})$), we choose to focus on the $\boldsymbol{\lambda}$ parameters.

Using the methodology described in sections 2.1. to 2.3., we reduce the original dimension of $\boldsymbol{f}$ from P to 1 for f_S (global models), and to m for $\boldsymbol{f}_{SK}$ (cluster models). Since our empirical findings point to a m close to 3 in each scenario that we examined, we can therefore estimate the model expressed as (12) and (13) using regular and very well-known econometric methods[29]. As a result, we estimate factor risk premia using the Fama-MacBeth (FM) procedure, which, in addition to accommodating unbalanced panels, allows for time-varying betas and runs swiftly for a large number of assets.

It is a two-pass regression where the first-pass entails estimating (12) with a time-rolling window procedure whose length[30] equals t^*. The first-pass therefore yields a sequence of estimated betas like $\{\widehat{\boldsymbol{\beta}}_{i,t}\}_{t=t^*}^{T}$, whereas the second-pass estimate (13) using the following sequence of cross-sectional regressions:

$$r_{i,t} = \lambda_0 + \widehat{\boldsymbol{\beta}}_{i,t}\boldsymbol{\lambda}_f + e_t \tag{14}$$

By setting $\hat{\boldsymbol{\lambda}}_t = \left(\hat{\lambda}_0, \hat{\boldsymbol{\lambda}}_f^T\right)^T$, $\hat{\boldsymbol{\lambda}}_t$ can be estimated for each cross-section, hence, the second-pass from FM produces a sequence of estimated risk premia factors like $\{\hat{\boldsymbol{\lambda}}_t\}_{t=t^*}^{T}$, while the final factor risk premia estimator can be expressed as:

$$\hat{\boldsymbol{\lambda}} = \frac{\sum_{t=t^*}^{T}\hat{\boldsymbol{\lambda}}_t}{(T-t^*)} \tag{15}$$

In order to make inferences about $\hat{\boldsymbol{\lambda}}$, and resulting from the fact that betas from the first-pass are pre-estimated, this procedure consequently generates errors. To correct this bias, we follow Shanken (1992) to compute the $\hat{\boldsymbol{\lambda}}$ covariance matrix by:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}shanken} = \left(1 + \hat{\boldsymbol{\lambda}}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}^{-1}\hat{\boldsymbol{\lambda}}\right)\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}} - \frac{\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}}{(T-t^*)}\right) + \frac{\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}}{(T-t^*)} \tag{16}$$

[29] Campbell et al. (1997), Cochrane (2009) and Goyal (2012) bring extensive descriptions on available methods used to estimate factor risk premia when $\boldsymbol{f}$ yields a low dimension.

[30] We set $t^* = 60$ in our research paper, using the same value employed by the original Fama and MacBeth (1973) article.

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{f}}$ is the estimated risk factor covariance matrix[31], and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}}$ is the regular estimated factor risk premia covariance matrix[32].

2.6. Classic Models and Alternative Methodologies

As pointed out before, our methodology selects risk factor vectors (f_S and $\boldsymbol{f}_{SK}$) that yield lower dimensions from the very beginning, consequently, applying the FM procedure allows us to compare results directly to certain commonly-cited (classic) models that also operate with a low set of risk factors. For our classic models, we choose to employ the Fama-French three-factor model (FF3)[33], the Novy-Marx four-factor model (NM4)[34], and the Carhart four-factor model (C4)[35].

It is worth noting that we cannot perceive comparisons between our method and classic models as being fair since we begin our research based on a large set of potential risk factor candidates aimed at explaining the cross-section of expected returns. Hence, with the purpose of evaluating our methodology's empirical performance, we choose to compare it to other approaches documented by research papers on the factor zoo literature.

As already mentioned, a well-known fact in this literature, as endorsed by Kozak, Nagel, and Santosh (2020), is that it is not possible to explain the cross-section of expected returns in a satisfactory manner with a small number of risk factors, even though only a few principal components (*PC*s) of risk factors are capable of achieving highly satisfactory results. From this standpoint, and seeking to compare our findings to results attained with the principal component analysis[36], we also test an FM procedure using two alternative risk factor sets. The first one stems from the first four *PC*s from $\boldsymbol{f}$. This makes for an interesting comparison since *PC*s represent orthogonal regions of the risk factor covariance matrix, and even though our clusters from the precision matrix are not necessarily orthogonal to each other, risk factors selected by $\boldsymbol{f}_{SK}$ also denote key

---

[31] $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{f}} = \frac{\sum_{t=t^*}^{T}(f_t-\bar{f})(f_t-\bar{f})'}{(T-t^*)^2}$.

[32] $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}} = \frac{\sum_{t=t^*}^{T}(\hat{\lambda}_t-\hat{\lambda})(\hat{\lambda}_t-\hat{\lambda})'}{(T-t^*)^2}$.

[33] Risk factors are Mkt, SMB, HML. See Fama and French (1993) for a reference on this model.

[34] Risk factors are Mkt, SMB, HML and GP. The GP risk factor is proposed by Novy-Marx (2013).

[35] Risk factors are Mkt, SMB, HML and MOM. For a description of the MOM risk factor, see Carhart (1997). In this paper, we apply the six-month MOM risk factor, as described by Jegadeesh and Titman (1993).

[36] See Kelly, Pruitt, and Su (2019) and Gu, Kelly, and Xiu (2019) for principal component analysis applications regarding the factor zoo.

components from different regions. As we pointed out before, the advantage of our methodology is that we are able to recover risk factors that pose economic meanings differently from PCs which do not present direct economic interpretation. The second entails the four-risk factor that yields the highest factor loadings for the first principal component of $\boldsymbol{f}$ $(PC1)$[37]**.** Since the first principal component is the latent factor with the largest variance, each of these four selected risk factor candidates may help to make a satisfactory proposal compared to the risk factor covariance matrix, thereby being a good set of candidates to explain the cross-section of expected returns.

Another prominent segment of factor-zoo[38] literature applies the elastic net model (L_1 and L_2 regularization) to estimate either factor risk prices or premia, depending on the research focus. In light of the foregoing, and in order to have a lasso-type estimator that we can use to make direct comparisons to our method, we estimate an alternative FM procedure with an original high-dimensional vector of risk factors ($P = 51$ in or sample) as well as an elastic net penalization loss function in both the first and second-passes. Consequently, we can express the loss function in the time-series regression as:

$$\mathcal{L}(\boldsymbol{\beta}_i, \boldsymbol{f}) = \frac{1}{2}\sum_{t=1}^{T}\left(r_{i,t} - \alpha_i - \boldsymbol{\beta}_i^T \boldsymbol{f_t}\right)^2 + \lambda_{P1}\left[\frac{1}{2}(1-\alpha_{P1})\|\boldsymbol{\beta}\|_2^2 + \alpha_{P1}\|\boldsymbol{\beta}\|_1\right] \qquad (17)$$

where λ_{P1} (regularization parameter) is chosen by a three-fold cross-validation method, and α_{P1} is set to 0.5 to have both L_1 and L_2 penalizations (elastic net estimator). We apply the same rolling-window estimation procedure for (17), using the exact same t^* time-window length to estimate betas that we use on (12) for the regular FM approach. Thus, this alternative FM procedure also yields a sequence of betas such as $\{\widehat{\boldsymbol{\beta}}_{\boldsymbol{en},i,t}\}_{t=t^*}^{T}$, whereas subscription $\boldsymbol{en}$ refers to elastic net estimators. For the second-pass cross-sectional regression, we employ the following loss function:

$$\mathcal{L}(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{en}}) = \frac{1}{2}\sum_{i=1}^{I}\left(r_{i,t} - \lambda_0 - \boldsymbol{\lambda}_f^T \widehat{\boldsymbol{\beta}}_{\boldsymbol{en},\boldsymbol{i},t}\right)^2 + \lambda_{P2}\left[\frac{1}{2}(1-\alpha_{P2})\|\boldsymbol{\lambda}_f\|_2^2 + \alpha_{P2}\|\boldsymbol{\lambda}_f\|_1\right] \qquad (18)$$

where λ_{P2} is also selected by a three-fold cross-validation method, while α_{P2} is likewise set to 0.5 in order to generate an elastic net estimator for factor risk premia for each cross-sectional regression. Using the FM procedure in equation (18) as the loss function in the second-pass, we also achieve a sequence of estimated factor risk premia such as

[37] Factor loading is the correlation coefficient between the principal component and the original random variable.

[38] Feng, Giglio, and Xiu (2020), Freyberger, Neuhierl, and Weber (2020) and Kozak, Nagel, and Santosh (2020).

$\{\hat{\boldsymbol{\lambda}}_{\boldsymbol{en},t}\}_{t=t^*}^{T}$. From there on, we proceed to calculate R^2 and the adjusted R^2 for each cross-section, in addition to comparing averages of both measures to results attained with the regular FM procedure.

Last but not least, we compute the one-step-ahead forecast for all models with the aim of evaluating and comparing out-of-sample (OOS) results among them. Since we are able to calculate the estimator $\{\widehat{\boldsymbol{\beta}}_{i,t}\}_{t=t^*}^{T}$ and $\{\hat{\boldsymbol{\lambda}}_t\}_{t=t^*}^{T}$ for each methodology we propose, and our interest lies in explaining the cross-section of expected returns, we estimate the one-step-ahead forecast as follows:

$$\hat{r}_{i,t+1} = \hat{\lambda}_{0,t} + \widehat{\boldsymbol{\beta}}_{i,t}\hat{\boldsymbol{\lambda}}_{f,t} \tag{19}$$

It is worth noting that the forecast computed with (19) is entirely out-of-sample. For each cross-section, we calculate the root-mean-square-error ($RMSE_t$)[39] for $t = t^*, \dots, T-1$, and then compare root-mean-square-error averages ($AV.RMSE$)[40] and square-error medians ($M.RMSE$)[41] across all models.

3. Database

In our research, we look at the factor zoo dataset compiled by Kozak, Nagel, and Santosh (2020) with monthly data ranging from January 1981 to December 2016[42]. It consists of 51 risk factors, the first being the *Excess Market Return*[43] gathered from the French Library[44], while the remaining 50 are zero-investment, long-short portfolios constituted by well-known traits described in the asset-pricing literature. In accordance with Feng, Giglio, and Xiu (2020), we split each risk factor into six types of groups: *Value vs Growth*; *Investment*; *Profitability*; *Momentum*; *Intangibles*; and *Trading Frictions*[45]. This *a priori* rating system provides us with some interesting tools to assess whether said

[39] $RMSE_t = \sqrt{\frac{\sum_{i=1}^{I}(\hat{r}_{i,t}-r_{i,t})^2}{I}}$, where I is the number of assets presents on the cross-section are period t. Since we estimate the one-step-ahead forecast, we lose one observation; thus, we are able to compute OOS results only with $t = t^*, \dots, T-1$ as opposed to $t = t^*, \dots, T$.

[40] $AV.RMSE = \frac{\sum_{t=t^*}^{T-1} RMSE_t}{T-t^*-1}$.

[41] $M.RMSE = median(RMSE_t)\ t = t^*, \dots, T-1$.

[42] Data can be downloaded from: https://www.serhiykozak.com/data.

[43] Sharpe (1964).

[44] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[45] Risk factor dataset results show nine risk factors from the Momentum group; 11 from the Value vs. Growth group; eight from the Investment group; 13 from the Profitability group; one from the Intangibles group; and nine from the Trading Frictions group.

categories are conditionally dependent, in addition to whether precision matrix clusters concentrate risk factors from specific category types. Table A1 of the Appendix summarizes risk factor descriptions and statistics.

In regards to cross-sectional returns, there is a trade-off between the choice of portfolio and individual assets. Although portfolios do not produce missing data by construction (balance panel), they do have a tendency of showing a bias towards traits used to build them as highlighted by Harvey and Liu (2019). As a result, and due to the fact that the FM procedure supports a large unbalanced panel, we choose to focus on individual assets from the CRSP stock return dataset. To compose excess asset returns, we set one-month maturity USD LIBOR interest rates as risk-free. Since we adopt a 60-month time window for the first-pass of our FM procedure, we consequently disregard assets with less than 60 observations. Additionally, we remove stocks from the financial sector. Using these procedures, our dataset accounts for 14,317 individual stocks (CSRP). Considering that this first dataset also consists of small-caps, which may yield significantly illiquid stocks, we create a second stock-free excess return set with prices lower than USD 5.00, thereby leading to a dataset with 10,221 individual assets (CRSP without small-caps).

4. Results

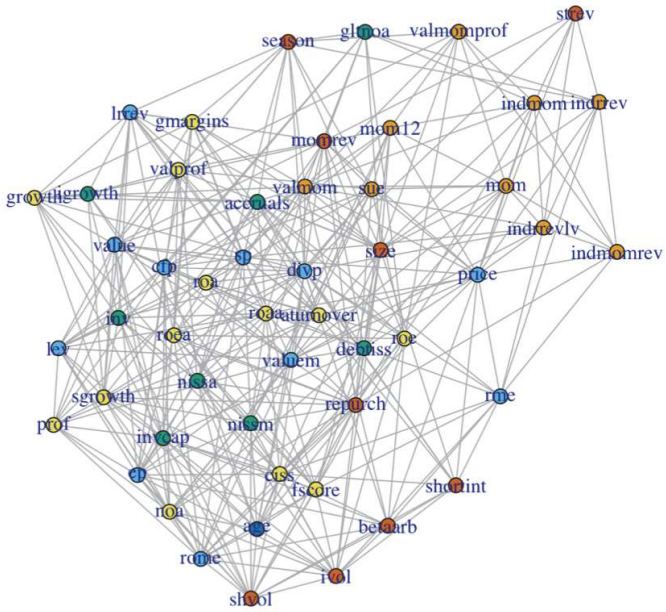
4.1.Network result and selected risk factors

Figure 1 displays the graphic representation of the estimated risk factor network described as $\hat{\mathcal{G}}_u$. After using our sparse precision matrix estimator $\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right)$ to compute the set edge of $\hat{\mathcal{G}}_u$, we can deem $\hat{\mathcal{G}}_u$ a sparse estimated representation for joint risk factor distribution $P(\boldsymbol{f})$. Results attained with $\hat{\mathcal{G}}_u$ lead to some interesting features of $P(\boldsymbol{f})$. First, our method selects an optimal regularization parameter (τ) close to 0.11. As expected, as τ increases, the estimated conditional dependence among risk factors decreases, as we can see in Figure A1 at the Appendix. Since $\hat{\tau}_{StARS} \cong 0.11$, average risk factor conditional dependence numbers correspond to 14.2. In other words, a risk factor is, on average, linked to 14.2 other risk factors from an universe of 51 risk factors. With 21 links, *Share Repurchases*[46] is the risk factor with the most connections, in stark

[46] Described in Ikenberry, Lakonishok, and Vermaelen (1995).

contrast to the *Short-term Reversal*[47], which is the risk factor with the smallest number of relations across all of them with only six conditional dependent links. As we see in Figure 1, the risk factor categories, illustrated according to the node color, is not necessarily grouped into clusters since that a coloring pattern does not emerges. Therefore, Figure 1 suggests that risk factors from different categories interact within each other too, thereby implying that systematic risks across different kinds of sources are also related.

Figure 1: Graphic representation of the estimated risk factor network $(\widehat{\mathcal{G}}_u)$



Note: The figure displays graphic representation of the estimated risk factor network $(\hat{\mathcal{G}}_u)$. *As described in section* 2.1., *we estimate the risk factor joint distribution precision matrix by graph lasso in order to obtain* $\widehat{\boldsymbol{\Theta}}_{GLASSO}$. *By StARS procedure, we find an optimal regularization parameter* $\hat{\tau}_{StARS} \cong 0.11$. *As described in section* 2.2., *we compute our estimated risk factor network by* $\hat{\mathcal{G}}_u = \mathcal{G}_u\left(\boldsymbol{\mathcal{V}} = \boldsymbol{f}, \boldsymbol{\mathcal{E}}(\widehat{\boldsymbol{\Theta}}_{GLASSO})\right)$. *In this picture, each node represents a risk factor and the edge between than indicates* $\hat{\theta}_{GLASSO,i,j} \neq 0$, *which means conditional dependence among the risk factors* i *and* j *given all others risk factors. The node color represents the risk factor category (blue for Value vs Growth; green for Investment; yellow for Profitability; orange for Momentum; dark blue for Intangibles; and red for Trading Frictions).*

We describe risk factors (before $\hat{\mathcal{G}}_u$ is clusterized) selected pursuant to each centrality measure in Table 1. We dub these models by global centrality selection. As we have already pointed out, *Share Repurchases* is the element with the most connections within this network, even though its choice as the main risk factor in the precision matrix

[47] See Jegadeesh (1990).

(only Betweenness selected *Dividend Yield*[48]) is merely an empirical coincidence. This result is a consequence of high correlations among centrality measures, which we can observe in Figure A2 at the Appendix. We note that the smallest correlation among centrality measures is between the Eigenvector and Betweenness models, which is still 0.52, and, consequently, significant.

Table 1: Risk factor global centrality selection

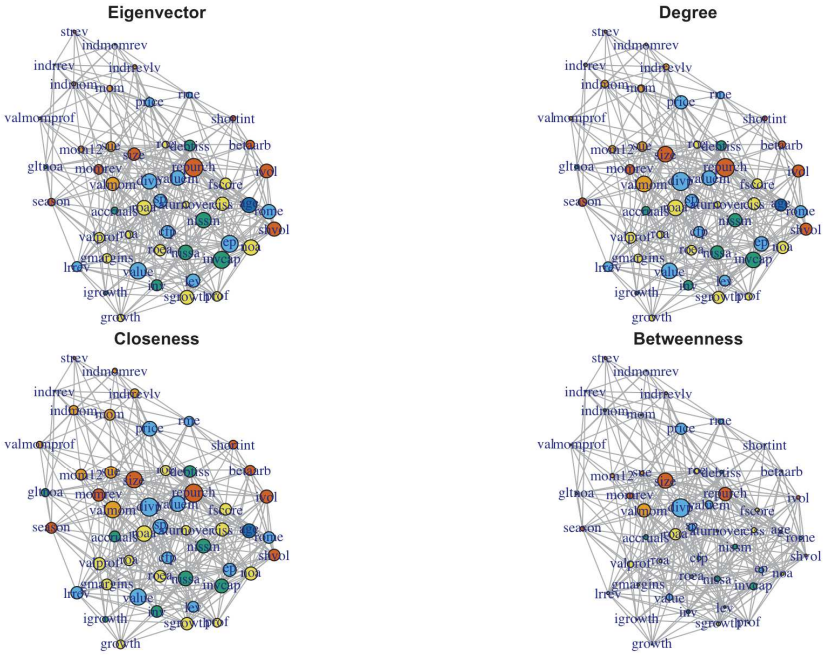
Description	Ret.	S.R.	Category	Reference	Code
Eigenvector					
Share Repurchases	0.029	0.172	Trading Frictions	Ikenberry, Lakonishok, and Vermaelen (1995)	repurch
Degree					
Share Repurchases	0.029	0.172	Trading Frictions	Ienberry, Lakonishok, and Vermaelen (1995)	repurch
Closeness					
Share Repurchases	0.029	0.172	Trading Frictions	kenberry, Lakonishok, and Vermaelen (1995)	repurch
Betweenness					
Dividend Yield	0.022	0.155	Value vs Growth	Naranjo, Nimalendran, and Ryngaert (1998)	divp

Note: The table displays the selected risk factor pursuant to each centrality measure (f_S)*, where* $f_S = \left\{f_p : c_p = max\left(\boldsymbol{c}\left(\hat{\mathcal{G}}_u\right)\right)\right\}$ *and* $\boldsymbol{c}$ *is a centrality measure function. The selection is done before* $\hat{\mathcal{G}}_u$ *is clusterized. For each selected risk factor, the table includes annualized average excess returns, annualized Sharpe ratios, a priori category classification, literature reference and code name.*

Figure 2 exhibits the $\hat{\mathcal{G}}_u$ risk factor network in which the size of the node measures each risk factor centrality according to each type of centrality measure. Figure 2 results allows us to acknowledge that each of the four types of centrality measures poses very similar results. A direct consequence is that, in spite of the $\hat{\mathcal{G}}_u$ clusterization, which we will examine ahead, models comprised of the network's key component are considerably similar, and present a high number of risk factors in common. Consequently, we expect a similar performance in explaining the cross-section of expected returns to that observed in these models, in addition to inferring scarce data as to how information spreads within the risk factor network. Fortunately, risk factors selected with our methodology have zero intersection with every other alternative model proposed, therefore making it possible for us to compare our findings to specific results documented by the literature on the factor zoo.

[48] See Naranjo, Nimalendran, and Ryngaert (1998).

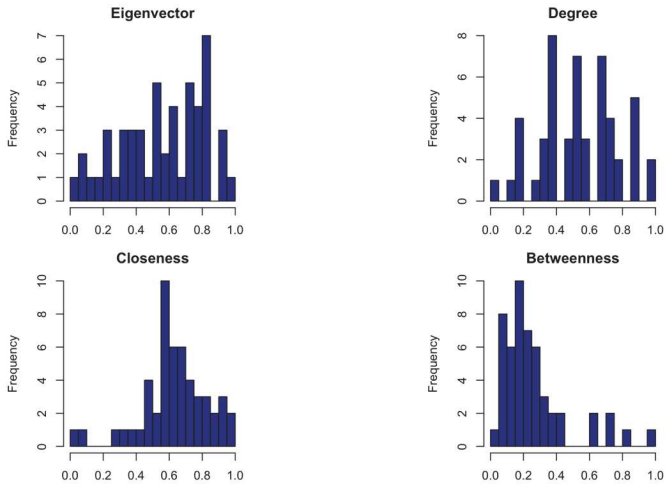
Figure 2: Graph representation of global centrality measures



Note: The figure displays graphic representation of our estimated risk factor network $(\hat{\mathcal{G}}_u)$ *according to each centrality measure. In this picture, each node represents a risk factor and the edge between than indicates* $\hat{\theta}_{GLASSO,i,j} \neq 0$. *The size of each vertex represent value of the its centrality measure (the bigger is the size the higher is the measure). The node color represents the risk factor category (blue for Value vs Growth; green for Investment; yellow for Profitability; orange for Momentum; dark blue for Intangibles; and red for Trading Frictions).*

Another very important observation to validate our research entails the fact that risk factors present variabilities in degree centralities for each measure. Figure 3 computes the histogram for centrality measures, enabling one to observe how only a few risk factors yield a high degree centrality when we consider each measure. It is important to point out that our methodology cannot be applied to homogenous networks where each element has the same degree of centrality due to the fact that the element with the highest centrality would fail to stand apart from the rest of the set.

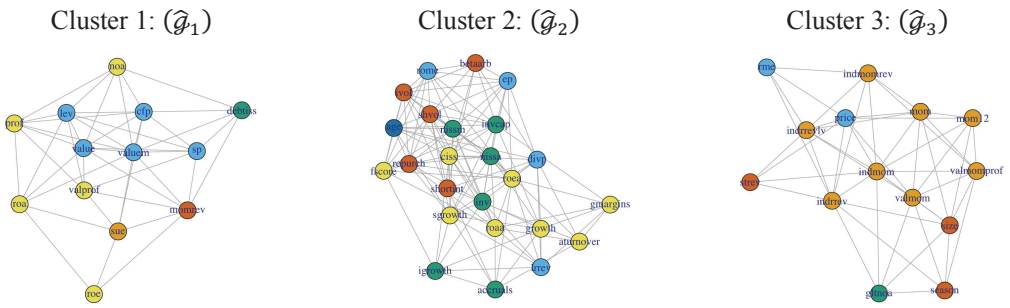
Figure 3: Global centrality measures histograms



Note: This figure displays the histograms from different centrality measures of our estimated risk factor network described by $\hat{\mathcal{G}}_u$.

Applying the Clauset, Newman, and Moore (2004) algorithm to maximize the modularity function (11) results in three clusters shown in Figure 4, suggesting that the risk factor precision matrix can be described in three different regions, as represented by $\widehat{\mathcal{M}} = (\hat{g}_1, \hat{g}_2, \hat{g}_3)$ with $\hat{g}_m$ being defined as in section 2.3. It is worth mentioning that models that select the key component for each cluster are sparse in this empirical case since we only have three clusters, and can therefore proceed with our methodology.

Figure 4: Graphic representation of the modularity cluster from our estimated risk factor network $(\widehat{\mathcal{G}}_u)$



Note: This figure displays the graphic representation of the clusters $(\widehat{\mathcal{M}} = (\hat{g}_1, \ldots, \hat{g}_M))$ *resulted from the modularity function maximization for our estimated risk factor network described by* $\hat{\mathcal{G}}_u$. *The optimal number of clusters is three* $(M = 3)$. *In this picture, each node represents a risk factor and the edge between than indicates* $\hat{\theta}_{GLASSO,i,j} \neq 0$. *The node color represents the risk factor category (blue for Value vs Growth; green for Investment; yellow for Profitability; orange for Momentum; dark blue for Intangibles; and red for Trading Frictions).*

The first cluster $(\hat{g}_1)$ entails 13 risk factors, five of which are classified as *Value vs Growth*, and five as *Profitability*. The second cluster $(\hat{g}_2)$ is the biggest one, featuring

24 risk factors, and encompasses over 50% of all risk factors from the *Investment*, *Profitability* and *Trading Frictions* models. With 14 factors to account for, the third cluster ($\widehat{\mathcal{G}}_3$) concerns *Momentum* distribution since it incorporates more than 80% of all risk factors from such a category. When looking at Figure 4, we find significant interaction taking place across each risk factor group, considering that *Momentum* alone stands for over 50% of every risk factor within a cluster, while every other cluster is concentrated in risk factors spanning two or three groups. Figure A3 at the Appendix shows the risk factor correlation matrix tidy according to the clusters order. As expected, risk factor within clusters present higher correlations than risk factors outside clusters.

Table 2 summarizes the selected risk factor vector pursuant to each centrality measure for each cluster result given by $\widehat{\mathcal{M}} = (\widehat{\mathcal{G}}_1, \widehat{\mathcal{G}}_2, \widehat{\mathcal{G}}_3)$ graphs described in Figure 4. Since the optimal number of clusters is three ($M = 3$), each centrality measure is able to selects three risk factors. We dub these models as cluster models. As highlighted above, said models are very similar when compared to the high correlations existing across the aforementioned centrality measures. In fact, Eigenvector and Closeness models are exactly the same, and the only difference from the Degree model is that *Composite Issuance*[49] is the main risk factor from cluster 2 instead of *Share Volume*[50]. In regards to the Betweenness model, it also poses *Value*[51], a common risk factor across all cluster models, as a key component for cluster 1.

[49] See Daniel and Titman (2006).
[50] See Datar, Naik, and Radcliffe (1998).
[51] See Asness and Frazzini (2013).

Table 2: Risk factor cluster centrality selection

Description	Ret.	S.R.	M	Category	Reference	Code
Eigenvector						
Value (monthly)	0.027	0.161	1	Value vs Growth	Asness and Frazzini (2013)	valuem
Share Volume	-0.037	-0.222	2	Trading Frictions	Datar, Naik, and Radcliffe (1998)	shvol
Industry Momentum	0.042	0.244	3	Momentum	Moskowitz and Grinblatt (1999)	indmom
Degree						
Value (monthly)	0.027	0.161	1	Value vs Growth	Asness and Frazzini (2013)	valuem
Composite Issuance	-0.086	-0.543	2	Profitability	Daniel and Titman (2006)	ciss
Industry Momentum	0.042	0.244	3	Momentum	Moskowitz and Grinblatt (1999)	indmom
Closeness						
Value (monthly)	0.027	0.161	1	Value vs Growth	Asness and Frazzini (2013)	valuem
Share Volume	-0.037	-0.222	2	Trading Frictions	Datar, Naik, and Radcliffe (1998)	shvol
Industry Momentum	0.042	0.244	3	Momentum	Moskowitz and Grinblatt (1999)	indmom
Betweenness						
Value (monthly)	0.027	0.161	1	Value vs Growth	Asness and Frazzini (2013)	valuem
Industry Relative Reversals	-0.133	-0.808	2	Momentum	Da, Liu, and Schaumburg (2014)	indrrev
Investment-to-Capital	-0.053	-0.302	3	Investment	Xing (2008)	invcap

Note: The table displays the selected risk factor vector pursuant to each centrality measure for each cluster result $(\boldsymbol{f}_{SK})$*, where* $\boldsymbol{f}_{SK} = (f_{1,p}, \dots, f_{M,p})$*,* $f_{m,p} = \{f_p : c_p = max(\boldsymbol{c}(\hat{\mathcal{G}}_m))\}$ *for* $m = 1, \dots, M$*,* $M = 3$ *and* $\boldsymbol{c}$ *is a centrality measure function. The selection is done after* $\hat{\mathcal{G}}_u$ *is clusterized. For each selected risk factor, the table includes annualized average excess returns, annualized Sharpe ratios, cluster origin, a priori category classification, literature reference and code name.*

4.2. Fama-MacBeth Results

After selecting risk factors for our global and cluster models, we are able to verify whether this methodology is in fact capable of explaining the cross-section of expected returns. Table 3 shows the factor risk premia estimated with FM procedure global models.

Table 3: FM Results for global models

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0039*	0.0039*	0.0039*	0.0055***	0.0058***	0.0058***	0.0058***	0.006***
	S. Error	0.0022	0.0022	0.0022	0.0021	0.0018	0.0018	0.0018	0.0016
	p-value	0.0832	0.0832	0.0832	0.0080	0.0014	0.0014	0.0014	0.0002
Share Repurchases	Estimate	(0.0041)	(0.0041)	(0.0041)		(0.0057)***	(0.0057)***	(0.0057)***	
	S. Error	0.0031	0.0031	0.0031		0.0028	0.0028	0.0028	
	p-value	0.1888	0.1888	0.1888		0.0435	0.0435	0.0435	
Dividend Yield	Estimate				(0.0030)				(0.0063)***
	S. Error				0.0027				0.0024
	p-value				0.2690				0.0095
Av. adjusted R^2		0.043	0.043	0.043	0.039	0.054	0.054	0.054	0.052

Note: The table displays the Fama and MacBeth (FM) results for the global models. The FM complete procedure is described in Section 2.5. *Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done before* $\hat{\mathcal{G}}_u$ *is clusterized (global models described in Table* 1*). The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is* 4,041 *and* 2,885 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0: \lambda = 0$ *against the alternative hypothesis* $H_1: \lambda \neq 0$. *The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%*,* 5% *and* 1% *level of significance, respectively.*

Table 4 summarizes the same results for cluster models. By comparing such results, we observe that the average adjusted R^2 for cluster models is higher (almost two-fold) than global models for both CRSP and CRSP without small-cap samples. For global models, on the other hand, none of the factor risk premium parameter is significant in the full CRSP sample. If we ignore small-caps, though, every factor risk premium parameter becomes significant, an expected result when we consider that small-cap stocks are known to yield more idiosyncratic elements than other stocks.

In regards to cluster models, we have significant factor risk premium parameter in both samples (with and without small-caps). Nevertheless, it is important to point out that the only significant risk premium factor, in each case[52], comes from cluster 2. All other factor risk premium parameters are insignificant. In summary, cluster models are superior to global models since they yield a higher adjusted R^2 as well as significant factor risk premia, even in the full CRSP sample. Nonetheless, this result is driven by information originating from the second region of the precision matrix.

[52] The *Share Volume* risk factor applies to Eigenvector and Closeness models; the *Composite Issuance* risk factor applies to the Degree model; and the *Investment to Capital* ratio risk factor apples to the Betweenness model.

Table 4: FM Results for cluster models

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0020	0.0019	0.0020	0.0034***	0.0033***	0.0037***	0.0033***	0.0049***
	S. Error	0.0015	0.0015	0.0015	0.0016	0.0012	0.0013	0.0012	0.0013
	p-value	0.1752	0.2231	0.1752	0.0313	0.0076	0.0043	0.0076	0.0002
Value (monthly)	Estimate	(0.0006)	(0.0009)	(0.0006)	(0.0012)	(0.0022)	(0.0028)	(0.0022)	(0.0035)
	S. Error	0.0029	0.0030	0.0029	0.0029	0.0027	0.0027	0.0027	0.0027
	p-value	0.8276	0.7592	0.8276	0.6806	0.4260	0.3068	0.4260	0.2057
Industry Momentum	Estimate	0.0010	0.0012	0.0010		0.0007	0.0018	0.0007	
	S. Error	0.0031	0.0031	0.0031		0.0028	0.0028	0.0028	
	p-value	0.7341	0.6952	0.7341		0.8067	0.5262	0.8067	
Share Volume	Estimate	0.0047*		0.0047*		0.007***		0.007***	
	S. Error	0.0028		0.0028		0.0027		0.0027	
	p-value	0.0972		0.0972		0.0093		0.0093	
Composite Issuance	Estimate		0.0052*				0.008***		
	S. Error		0.0030				0.0027		
	p-value		0.0809				0.0030		
Investment-to-Capital	Estimate				0.0054*				0.0077***
	S. Error				0.0030				0.0029
	p-value				0.0721				0.0078
Industry Relative Reversals	Estimate				0.0017				0.0003
	S. Error				0.0028				0.0027
	p-value				0.5555				0.9061
Av. adjusted R^2		0.098	0.096	0.098	0.096	0.115	0.111	0.115	0.112

Note: The table displays the Fama and MacBeth (FM) results for the cluster models. The FM complete procedure is described in Section 2.5. *Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done after* $\hat{\mathcal{G}}_u$ *is clusterized (cluster models described in Table* 2*). The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is* 4,041 *and* 2,885 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0: \lambda = 0$ *against the alternative hypothesis* $H_1: \lambda \neq 0$*. The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%*,* 5% *and* 1% *level of significance respectively.*

Classic model results are displayed in Table 5. The average adjusted R^2 is slightly better than in our global models, though lower in every case when we compare it to our cluster models. When we consider the full CRSP sample, classic models do not show any significant estimated factor risk premia. And when small-caps are removed from the CRSP dataset, the risk premium for the *Excess Market Return* becomes significant for the FF3 and NM4 models, although it remains insignificant in the C4 model. *Gross Profitability* is significant in the NH4 model, whereas *Momentum* is not significant in the C4 model.

In the next step, we compute the FM procedure for the first four PCs of $\boldsymbol{f}$ whose results are described in Table 6. The first four PCs account for approximately 80% of the total cumulative variance of $\boldsymbol{f}$, as we observe in Figure A4 at the Appendix, resulting in promising risk factor candidates to explain cross-sectional returns.

Table 5: FM Results for classic models

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**FF3**	**NM4**	**C4**	**P5**	**FF3**	**NM4**	**C4**	**P5**
Intercept	Estimate	0.0032**	0.0035***	0.0029**	0.0031**	0.0039***	0.0038***	0.0037**	0.0036**
	S. Error	0.0013	0.0013	0.0012	0.0012	0.0009	0.0009	0.0008	0.0008
	p-value	0.0182	0.0070	0.0197	0.0106	0.0000	0.0000	0.0000	0.0000
Excess Market Return	Estimate	0.0026	0.0025	0.0028	0.0026	0.0046*	0.0046*	0.0045	0.0044
	S. Error	0.0027	0.0027	0.0027	0.0026	0.0025	0.0025	0.0025	0.0025
	p-value	0.3372	0.3610	0.2963	0.3213	0.0715	0.0697	0.0690	0.0717
Size	Estimate	(0.0025)	(0.0029)	(0.0031)	(0.003)	-0.0059**	-0.0065**	-0.006**	(0.006)
	S. Error	0.0028	0.0028	0.0027	0.003	0.0026	0.0026	0.0026	0.003
	p-value	0.3823	0.3002	0.2571	0.217	0.0253	0.0124	0.0199	0.012
Value	Estimate	(0.0024)	(0.0025)	(0.0021)	(0.0024)	(0.0037)	(0.0037)	(0.0038)	(0.0039)
	S. Error	0.0028	0.0028	0.0027	0.0027	0.0027	0.0027	0.0026	0.0026
	p-value	0.3986	0.3700	0.4321	0.3886	0.1640	0.1609	0.1475	0.1441
Gross Profitability	Estimate		0.0034		0.0035		0.0047*		0.0047
	S. Error		0.0028		0.0027		0.0026		0.0026
	p-value		0.2205		0.1981		0.0707		0.0703
Momentum (6m)	Estimate			(0.0002)	0.0000			0.0003	0.0005
	S. Error			0.0030	0.0029			0.0028	0.0028
	p-value			0.9343	0.9987			0.9101	0.8476
Av. adjusted R^2		0.048	0.060	0.060	0.071	0.058	0.071	0.066	0.082

Note: The table displays the Fama and MacBeth (FM) results for the classic models. The FM complete procedure is described in Section 2.5*. Columns FF3, NM4, C4 and P5 refers to classic models described in Section 2.6. The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is* 4,041 *and* 2,885 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0: \lambda = 0$ *against the alternative hypothesis* $H_1: \lambda \neq 0$*. The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%*,* 5% *and* 1% *level of significance respectively.*

Table 6, in turn, shows that none of the PCs are significant for full CRSP samples. In regards to the dataset without small-caps, only the first principal component is significant across all models, while the fourth principal component is significant in the PC4 model. The average adjusted R^2 for the PC3 and PC4 models is attuned to the results from our cluster models.

Table 6: FM Results for PCs models

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**PC1**	**PC2**	**PC3**	**PC4**	**PC1**	**PC2**	**PC3**	**PC4**
Intercept	Estimate	0.0042**	0.0035**	0.0021	0.0024*	0.0057***	0.0046***	0.0035***	0.0035***
	S. Error	0.0019	0.0016	0.0015	0.0014	0.0017	0.0013	0.0013	0.0011
	p-value	0.0286	0.0296	0.1735	0.0912	0.0007	0.0007	0.0053	0.0011
PC1	Estimate	0.0191	0.0167	0.0148	0.0148	0.0263**	0.0263**	0.0243**	0.0237**
	S. Error	0.0129	0.0129	0.0126	0.0125	0.0119	0.0122	0.0121	0.0120
	p-value	0.1397	0.1948	0.2410	0.2372	0.0278	0.0319	0.0455	0.0499
PC2	Estimate		(0.0047)	(0.0018)	(0.000)		(0.0056)	(0.0027)	0.000
	S. Error		0.0080	0.0076	0.007		0.0073	0.0072	0.007
	p-value		0.5582	0.8148	0.986		0.4390	0.7055	0.962
PC3	Estimate			0.0045	0.0035			0.0050	0.0044
	S. Error			0.0066	0.0065			0.0058	0.0057
	p-value			0.4985	0.5916			0.3866	0.4441
PC4	Estimate				(0.0063)				-0.0098**
	S. Error				0.0050				0.0044
	p-value				0.2081				0.0273
Av. adjusted R^2		0.042	0.074	0.100	0.124	0.055	0.088	0.116	0.143

Note: The table displays the Fama and MacBeth (FM) results for the PCs models. The FM complete procedure is described in Section 2.5. *Columns PC1, PC2, PC3 and PC4 refers to principal components' models described in Section* 2.6. *The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is* 4,041 *and* 2,885 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0: \lambda = 0$ *against the alternative hypothesis* $H_1: \lambda \neq 0$. *The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%, 5% *and* 1% *level of significance respectively.*

Considering the PC1 (first principal component) risk premium significance, we also test models with four-risk factors that yield the highest factor loadings for the first principal component of $\boldsymbol{f}$, leading to models that we dub as PC1 loadings**.** Figure A5 at the Appendix shows factor loadings for the first principal component of $\boldsymbol{f}$**.** It is interesting to note that all four factors with the highest loadings on the first principal component (*Firm Age*, *Investment to Capital*, *Share Issuance* [*Monthly* and *Annual*]) come from the second cluster. Table 7 summarizes estimated results for PC1 loadings models. As we can see, unlike previous models with the exception of cluster models, PC1 loadings models achieve significant factor risk premium estimators for both CRSP and CRSP samples without small-caps. Model 4 presents an average adjusted R^2 slightly worse than in cluster models' statistics. Consequently, Table 7 also supports specific evidence that favors risk factors from the second region of the precision matrix.

In general terms, we determine that our cluster models as well as the PC1 loadings models are the only ones that present significant estimated factor risk premia parameters in the full CRSP sample. After comparing each cluster model to each PC1 loadings model, we observe that cluster models achieve a higher average adjusted R^2 in both samples.

When, in turn, we look at the in-sample result, only PC3 and PC4 models display average adjusted R^2 numbers higher than cluster models; nevertheless, they are unable to provide significant factor risk premia for the full CRSP dataset. Thus, FM results support cluster models when we take into consideration each sparse model tested so far. In regards to in-sample average adjusted R^2 numbers, they range from approximately 0.09 for the CRSP full sample, to 0.11 for the CRSP with small-cap samples, a satisfactory outcome according to Campbell and Thompson (2008).

Table 7: FM Results for PC1 loadings models

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**Model1**	**Model2**	**Model3**	**Model4**	**Model1**	**Model2**	**Model3**	**Model4**
Intercept	Estimate	0.0032*	0.0031*	0.0027	0.0033**	0.0049***	0.005***	0.0047***	0.0052***
	S. Error	0.0018	0.0017	0.0017	0.0016	0.0015	0.0014	0.0014	0.0014
	p-value	0.0771	0.0728	0.1086	0.0454	0.0014	0.0007	0.0011	0.0002
Firm Age	Estimate	(0.0053)	(0.0051)*	(0.0046)	(0.0047)	(0.006)***	(0.0071)**	(0.0068)**	(0.0071)**
	S. Error	0.0032	0.0031	0.0030	0.0030	0.0029	0.0028	0.0027	0.0028
	p-value	0.1011	0.0942	0.1272	0.1155	0.0187	0.0119	0.0138	0.0108
Investment-to-Capital	Estimate		0.007**	0.0065**	0.0064**		0.0081***	0.0077***	0.008***
	S. Error		0.0031	0.0031	0.003		0.0029	0.0028	0.003
	p-value		0.0258	0.0350	0.037		0.0048	0.0057	0.005
Share Issuance (monthly)	Estimate			0.0051*	0.0051*			0.0057**	0.006**
	S. Error			0.0030	0.0031			0.0028	0.0028
	p-value			0.0967	0.0965			0.0437	0.0347
Share Issuance (annual)	Estimate				0.0050				0.0058**
	S. Error				0.0031				0.0029
	p-value				0.1130				0.0459
Av. adjusted R^2		0.043	0.065	0.089	0.107	0.057	0.081	0.106	0.125

Note: The table displays the Fama and MacBeth (FM) results for the PC1 loadings models. The FM complete procedure is described in Section 2.5. Columns Model1, Model2, Model3 and Model4 refers to PC1 loadings models described in Section 2.6. The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is 4,041 and 2,885 for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis $H_0: \lambda = 0$ against the alternative hypothesis $H_1: \lambda \neq 0$. The subscription $*, **$ and $***$ indicates that the null hypothesis is rejected at 10%, 5% and 1% level of significance respectively.

Results from using our lasso-type estimator (the alternative FM method described in section 2.5) are summarized in Table 8 from which we observe that the average adjusted R^2 is around 0.17 and 0.24, respectively, for CRSP and CRSP datasets without small-caps, higher than the average adjusted R^2 exhibited by every other sparse model tested. However, the number of average risk factor predictors in each cross-sectional regression is 11.7 and 19.7 for the respective datasets. Consequently, this entails that we start moving away from sparse models.

Table 8: FM Results for lasso model

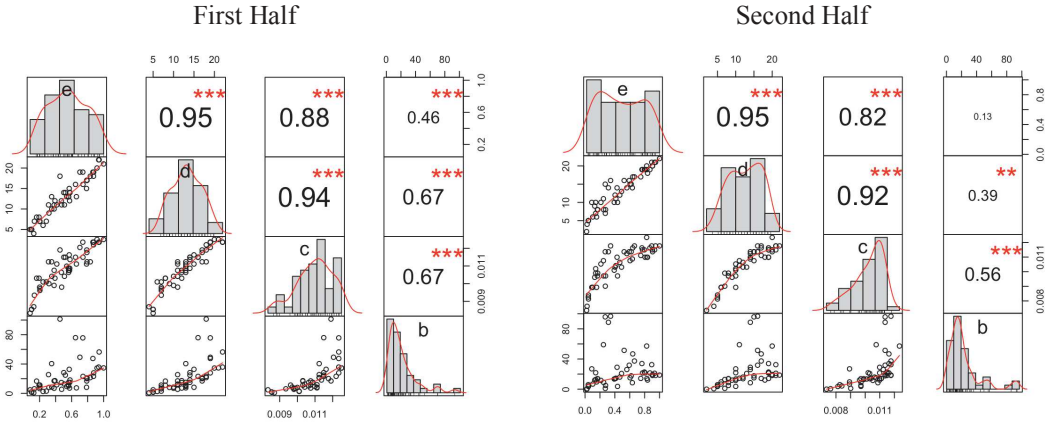
Base	CRSP	CRSP without Small Caps
Av. adjusted R^2	0.172	0.244
Av. number predictors	11.7	19.7

Note: The table displays the Fama MacBeth (FM) results for the lasso model which is described in Section 2.6. The results are reported both for complete and without small caps CRSP datasets. The average number of securities in each cross-sectional regression is 4,041 and 2,885 for complete and without small caps CRSP datasets respectively. For each dataset, the table includes the average adjusted R^2 and the average number of non-zero predictors from the FM second-pass procedure describe by equation (18).

4.3.Time-Varying Robustness

Variations regarding the explanatory power of individual factors in the cross-section of expected returns over time is something commonly described in research papers on the factor zoo literature[53]. Taking such an observation into consideration, and seeking to examine this well-known fact, we proceed to split our samples in half so that the first part ranges from January 1981 to December 1998, while the second encompasses the time frame from January 1999 to December 2016. For these samples we apply the same methodology described in Sections 2.1. to 2.4.

Figure 7: Risk factor`s global centrality measures histograms and correlations



Note: This figure displays the histograms and correlations from different centrality measures of our estimated risk factor network for different samples periods. The first half ranges from January 1981 to December 1998 and the second half ranges from January 1999 to December 2016. Letter e, d, c and b refers to Eigenvector, Degree, Closeness and Betweenness centrality measure respectively.

Figure 7 shows correlations among centrality measures for both time frames, which enables us to verify high correlations between Eigenvector, Degree and Closeness

[53] See Freyberger, Neuhierl, and Weber (2020) and Green, Hand, and Zhang (2017).

models in both samples. Although Betweenness accounts for lower correlations in regards to the remaining models, it remains positive and only becomes insignificant in the second half of the sample, after we compute it against the Eigenvector centrality measure. Table 9 describes selected risk factors according to each centrality measure (f_S) for different sampling periods.

Table 9: Risk factor global centrality selection

All samples	First Half	Second Half
	Eigenvector	
Share Repurchases	Investment-to-Capital	Firm Age
	Degree	
Share Repurchases	Value	Firm Age
	Closeness	
Share Repurchases	Investment-to-Capital	Share Repurchases
	Betweenness	
Dividend Yield	Price	Price

Note: The table displays the selected risk factor pursuant to each centrality measure (f_S) *for different samples periods, where* $f_S = \left\{f_p : c_p = max\left(\boldsymbol{c}(\hat{\mathcal{G}}_u)\right)\right\}$ *and* $\boldsymbol{c}$ *is a centrality measure function. The selection is done before* $\hat{\mathcal{G}}_u$ *is clusterized. The first half ranges from January* 1981 *to December* 1998 *and the second half ranges from January* 1999 *to December* 2016.

After maximizing modularity (11), the number of optimal clusters remains three for all sampling periods, and Table 10 illustrates selected risk factor vectors according to each centrality measure and precision matrix cluster for different sampling periods. We note that selected risk factors usually differ across different periods, though oftentimes they are the same within certain periods of both global and cluster models. In fact, when it comes to cluster models, all four centrality measures attain the same model for the first half of the sampling period. This reflects high correlations found among centrality measures shown in Figure A2 and Figure 7.

For the FM procedure, and given that we are still using the 60-month time window for the first-pass, results from second-pass regressions apply cross-sections estimated from January 1986 to December 1998 in the first half, and from January 2004 to December 2016 for the second half. Table 11 and Table 12 display FM results for global and cluster models, respectively, for the first half period. This allows us to conclude that *Investment to Capital* is the only risk factor from global models that poses a significant estimator for the CRSP without small-cap samples.

Table 10: Risk factor cluster centrality selection

All samples	First Half	Second Half
	Eigenvector	
Value (monthly)	Firm Age	Composite Issuance
Share Volume	Value	Return on Assets (annual)
Industry Momentum	Price	Price
	Degree	
Value (monthly)	Firm Age	Beta Arbitrage
Composite Issuance	Value	Return on Assets (annual)
Industry Momentum	Price	Price
	Closeness	
Value (monthly)	Firm Age	Beta Arbitrage
Share Volume	Value	Return on Assets (annual)
Industry Momentum	Price	Price
	Betweenness	
Value (monthly)	Firm Age	Beta Arbitrage
Industry Relative Reversals	Value	Return on Assets (annual)
Investment-to-Capital	Price	Price

Note: The table displays the selected risk factor vector pursuant to each centrality measure for each cluster result $(\boldsymbol{f}_{SK})$ *for different samples periods, where* $\boldsymbol{f}_{SK} = (f_{1,p}, \dots, f_{M,p})$, $f_{m,p} = \{f_p : c_p = max(\boldsymbol{c}(\hat{g}_m))\}$ *for* $m = 1, \dots, M$, *and* $\boldsymbol{c}$ *is a centrality measure function. The selection is done after* $\hat{\mathcal{G}}_u$ *is clusterized.* $M = 3$ *for all the three different sample periods. The first half ranges from January* 1981 *to December* 1998 *and the second half ranges from January* 1999 *to December* 2016.

Table 11: FM Results for global models for the first half sample

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0054	0.0049	0.0054	0.0026	0.0034	0.0057**	0.0034	0.005*
	S. Error	0.0037	0.0035	0.0037	0.0034	0.0021	0.0024	0.0021	0.0026
	p-value	0.1505	0.1732	0.1505	0.4424	0.1180	0.0186	0.1180	0.0559
Investment-to-Capital	Estimate	0.0017		0.0017		0.0067**		0.0067**	
	S. Error	0.0044		0.0044		0.0033		0.0033	
	p-value	0.7030		0.7030		0.0463		0.0463	
Value	Estimate		(0.0012)				(0.0051)		
	S. Error		0.0035				0.0033		
	p-value		0.7401				0.1202		
Price	Estimate				(0.0025)				(0.0035)
	S. Error				0.0029				0.0024
	p-value				0.3859				0.1476
Av. adjusted R^2		0.029	0.029	0.029	0.030	0.055	0.044	0.055	0.040

Note: The table displays the Fama and MacBeth (FM) results for the global models. The FM complete procedure is described in Section 2.5. Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done before $\hat{\mathcal{G}}_u$ *is clusterized (global models described in Table* 1*). The results are reported both for complete and without small caps CRSP datasets. The FM second-pass regressions results apply cross-sections estimated from January* 1986 *to December* 1998*. The average number of securities in each cross-sectional regression is* 3,825 *and* 2,528 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0 : \lambda = 0$ *against the alternative hypothesis* $H_1 : \lambda \neq 0$*. The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%*,* 5% *and* 1% *level of significance respectively.*

Table 12: FM Results for cluster models for the first half sample

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0029	0.0029	0.0029	0.0029	0.0046**	0.0046**	0.0046**	0.0046**
	S. Error	0.0025	0.0025	0.0025	0.0025	0.0020	0.0020	0.0020	0.0020
	p-value	0.2484	0.2484	0.2484	0.2484	0.0253	0.0253	0.0253	0.0253
Firm Age	Estimate	(0.0031)	(0.0031)	(0.0031)	(0.0031)	(0.0050)	(0.0050)	(0.0050)	(0.0050)
	S. Error	0.0032	0.0032	0.0032	0.0032	0.0031	0.0031	0.0031	0.0031
	p-value	0.3386	0.3386	0.3386	0.3386	0.1059	0.1059	0.1059	0.1059
Value	Estimate	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0042)	(0.0042)	(0.0042)	(0.0042)
	S. Error	0.0033	0.0033	0.0033	0.0033	0.0032	0.0032	0.0032	0.0032
	p-value	0.4745	0.4745	0.4745	0.4745	0.1840	0.1840	0.1840	0.1840
Price	Estimate	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0020)	(0.0020)	(0.0020)	(0.0020)
	S. Error	0.0029	0.0029	0.0029	0.0029	0.0025	0.0025	0.0025	0.0025
	p-value	0.6852	0.6852	0.6852	0.6852	0.4249	0.4249	0.4249	0.4249
Av. adjusted R^2		0.081	0.081	0.081	0.081	0.097	0.097	0.097	0.097

Note: The table displays the Fama and MacBeth (FM) results for the cluster models. The FM complete procedure is described in Section 2.5. Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done after $\hat{\mathcal{G}}_u$ *is clusterized (cluster models described in Table* 2*). The results are reported both for complete and without small caps CRSP datasets. The FM second-pass regressions results apply cross-sections estimated from January* 1986 *to December* 1998*. The average number of securities in each cross-sectional regression is* 3,825 *and* 2,528 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis* $H_0: \lambda = 0$ *against the alternative hypothesis* $H_1: \lambda \neq 0$*. The subscription* $*, **$ *and* $***$ *indicates that the null hypothesis is rejected at* 10%*,* 5% *and* 1% *level of significance respectively.*

Table 13 and Table 14 feature global and cluster model results for the second half sample. In this case, results are slightly better with *Price* risk factor premium being significant in the Betweenness global model, and three other cluster models for the CRSP without small-cap samples.

Splitting the dataset into half enables us to verify the factor zoo evidence on the time-varying explanatory power of individual factors in the cross-section of expected returns. Nevertheless, results are considerably poorer compared to those obtained using the full sampling period, thereby suggesting that 13-year monthly data time frames for FM second-pass regressions may be too short a sample to reach satisfactory results[54].

[54] See Cavalcante Filho et al. (2020) for interesting observations on sample sizes required to obtain a robust risk premium estimator.

Table 13: FM Results for global models for the second half sample

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0051	0.0051	0.0078	0.0019	0.0053**	0.0053**	0.0082***	0.0027
	S. Error	0.0024	0.0024	0.0036	0.0027	0.0021	0.0021	0.0031	0.0024
	p-value	0.0379	0.0379	0.0311	0.4783	0.0129	0.0129	0.0092	0.2553
Firm Age	Estimate	(0.0004)	(0.0004)			(0.0022)	(0.0022)		
	S. Error	0.0023	0.0023			0.0022	0.0022		
	p-value	0.8752	0.8752			0.3328	0.3328		
Share Repurchases	Estimate			0.0025				0.0004	
	S. Error			0.0025				0.0024	
	p-value			0.3292				0.8745	
Price	Estimate				(0.0043)				(0.0065)**
	S. Error				0.0032				0.0030
	p-value				0.1739				0.0338
Av. adjusted R^2		0.034	0.034	0.034	0.038	0.044	0.044	0.038	0.046

Note: The table displays the Fama and MacBeth (FM) results for the global models. The FM complete procedure is described in Section 2.5. *Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done before $\hat{\mathcal{G}}_u$ is clusterized (global models described in Table* 1*). The results are reported both for complete and without small caps CRSP datasets. The FM second-pass regressions results apply cross-sections estimated from January* 2004 *to December* 2016*. The average number of securities in each cross-sectional regression is* 4,128 *and* 3,103 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis $H_0: \lambda = 0$ against the alternative hypothesis $H_1: \lambda \neq 0$. The subscription $*, **$ and $***$ indicates that the null hypothesis is rejected at* 10%, 5% *and* 1% *level of significance respectively.*

Table 14: FM Results for global models for the second half sample

Base		**CRSP**				**CRSP without Small Caps**			
Coefficient		**E**	**D**	**C**	**B**	**E**	**D**	**C**	**B**
Intercept	Estimate	0.0019	0.0028	0.0028	0.0028	0.0022	0.0026	0.0026	0.0026
	S. Error	0.0021	0.0021	0.0021	0.0021	0.0017	0.0017	0.0017	0.0017
	p-value	0.3617	0.1837	0.1837	0.1837	0.1820	0.1204	0.1204	0.1204
Composite Issuance	Estimate	0.0009				0.0027			
	S. Error	0.0028				0.0027			
	p-value	0.7429				0.3292			
Return on Assets (annual)	Estimate	(0.0009)	(0.0016)	(0.0016)	(0.0016)	(0.0029)	(0.0047)	(0.0047)	(0.0047)
	S. Error	0.0039	0.0038	0.0038	0.0038	0.0038	0.0037	0.0037	0.0037
	p-value	0.8114	0.6630	0.6630	0.6630	0.4424	0.2002	0.2002	0.2002
Price	Estimate	(0.0027)	(0.0034)	(0.0034)	(0.0034)	(0.0050)	(0.0064)**	(0.0064)**	(0.0064)**
	S. Error	0.0031	0.0029	0.0029	0.0029	0.0031	0.0029	0.0029	0.0029
	p-value	0.3881	0.2430	0.2430	0.2430	0.1040	0.0257	0.0257	0.0257
Beta Arbitrage	Estimate		0.0020	0.0020	0.0020		0.0041	0.0041	0.0041
	S. Error		0.0032	0.0032	0.0032		0.0031	0.0031	0.0031
	p-value		0.5406	0.5406	0.5406		0.1828	0.1828	0.1828
Av. adjusted R^2		0.088	0.086	0.086	0.086	0.106	0.106	0.106	0.106

Note: The table displays the Fama and MacBeth (FM) results for the cluster models. The FM complete procedure is described in Section 2.5. *Columns E, D, C and B refers to selected risk factor pursuant to eigenvector, degree, closeness and betweenness centrality measure respectively. The selection is done after $\hat{\mathcal{G}}_u$ is clusterized (cluster models described in Table* 2*). The results are reported both for complete and without small caps CRSP datasets. The FM second-pass regressions results apply cross-sections estimated from January* 2004 *to December* 2016*. The average number of securities in each cross-sectional regression is* 4,128 *and* 3,103 *for complete and without small caps CRSP datasets respectively. For each model and each dataset, the table includes the estimated risk premia, standard error and the p-value for null hypothesis $H_0: \lambda = 0$ against the alternative hypothesis $H_1: \lambda \neq 0$. The subscription $*, **$ and $***$ indicates that the null hypothesis is rejected at* 10%, 5% *and* 1% *level of significance respectively.*

4.4. Out-of-sample Results

As Abu-Mostafa, Magdon-Ismail, and Lin (2012) point out, as the model's complexity increases, the in-sample performance (IN) will also permanently increase. On the other hand, out-of-sample performances (OOS) will start to decrease after the training model stops fitting the data structure, and starts fitting the data noise. Thus, we can state that a desirable result is a model posing both IN and OOS satisfactory results. Hence, we analyze out-of-sample (OOS) results from the one-step-ahead forecast computed, as described in section 2.5.

Table 15 displays statistics for the $AV.RMSE$ (root-mean-square-error averages) and $M.RMSE$ (root-mean-square-error medians) of every model tested in the previous sections. The Degree model from the cluster model group shows the lowest figures for $AV.RMSE$ and $M.RMSE$ in both samples. Furthermore, all cluster models also scored the lowest $AV.RMSE$ and $M.RMSE$ averages compared to every other model in both samples. This outperform is statistically significant in both samples as we observe the Diebold-Mariano test results in Table A2 and A3 on the Appendix. Thus, the OOS analysis favors our cluster model methodology when compared to every other approach.

Table 15: All models out-of-sample results

Base		**CRSP**		**CRSP without Small Caps**	
Model/Metric		$AV.RMSE$	$M.RMSE$	$AV.RMSE$	$M.RMSE$
Global	E, D and C	0.1730	0.1599	0.1279	0.1166
	B	0.1725	0.1606	0.1275	0.1162
Cluster	E and C	0.1673	0.1543	0.1234	0.1116
	D	**0.1672**	**0.1540**	**0.1233**	**0.1113**
	B	0.1673	0.1546	0.1235	0.1117
Classic	FF3	0.1763	0.1620	0.1298	0.1173
	NM4	0.1775	0.1627	0.1306	0.1185
	C4	0.1776	0.1632	0.1307	0.1183
	P5	0.1787	0.1632	0.1314	0.1192
PCs	PC1	0.1726	0.1600	0.1278	0.1171
	PC2	0.1747	0.1611	0.1290	0.1174
	PC3	0.1763	0.1626	0.1299	0.1177
	PC4	0.1776	0.1629	0.1307	0.1182
PC1 Loadings	Model1	0.1728	0.1596	0.1278	0.1169
	Model2	0.1740	0.1610	0.1286	0.1174
	Model3	0.1755	0.1618	0.1296	0.1185
	Model4	0.1767	0.1628	0.1303	0.1188
Lasso type (Elastic Net FM)		0.1732	0.1599	0.1293	0.1175

Note: The table displays the root mean square error averages ($AV.RMSE$) and medians ($M.RMSE$) across all models relating to the out-of-sample one-step-ahead forecast described in Section 2.6.

5. Conclusion

Factor risk premia papers have produced hundreds of potential candidates to explain the cross-section of expected returns, resulting in the factor zoo problem. Unfortunately, the question posed by Cochrane (2011) concerning which risk factor "(…) really provides independent information about average returns" remains unanswered. Nevertheless, thanks to the advanced computational power, superior dataset as well as novel econometric methods, economists are now able to start addressing this issue.

With this paper, we propose a new methodology to reduce risk factor predictor dimensions by employing several tools/resouces from different fields, such as high-dimensional statistics and network analysis. We estimate the risk factor precision matrix using an elastic net penalization. Moreover, we apply risk factor precision matrices' inherent conditional dependence to develop a risk factor network, and argue that its key component is best suited to condense information stemming from it. We consequently achieve sparsity after selecting its key component (global models). Nevertheless, since we are able to clusterize the estimated precision matrix by maximizing modularity, we go even further and select a group of risk factors that are the key components within each cluster (cluster models). And since the number of clusters is much smaller than the dimension of the original risk factor vector, we are also able to achieve a sparse model in this second approach. To the best of our knowledge, this is the first time that both penalized risk factor precision matrix estimators and network analysis are applied together to solve the factor zoo. After selecting these low-dimensional risk factor vector candidates, we then apply the FM procedure to evaluate our methodology based on the CRSP monthly stock return dataset ranging from January 1981 to December 2016, in addition to Kozak, Nagel, and Santosh (2020) factor datasets.

Our findings imply that cluster models yield better both in and out-of-sample results when compared to classic models or specific alternative methods documented in the literature about the factor zoo. Cluster model outperformances of global models may suggest that investors are more concerned with a smaller and specific set of risk factors than with a globally systematic risk summarized from a large set of different risk factors. Additionally, the fact that cluster models produced satisfactory results shows that applying network analysis may be an interesting method that economists can use to better understand joint risk factor distribution, in addition to how its properties can be applied to select risk factors with the purpose of explaining the cross-section of expected returns.

Appendix

A1. Stability approach for regularization selection (StARS)

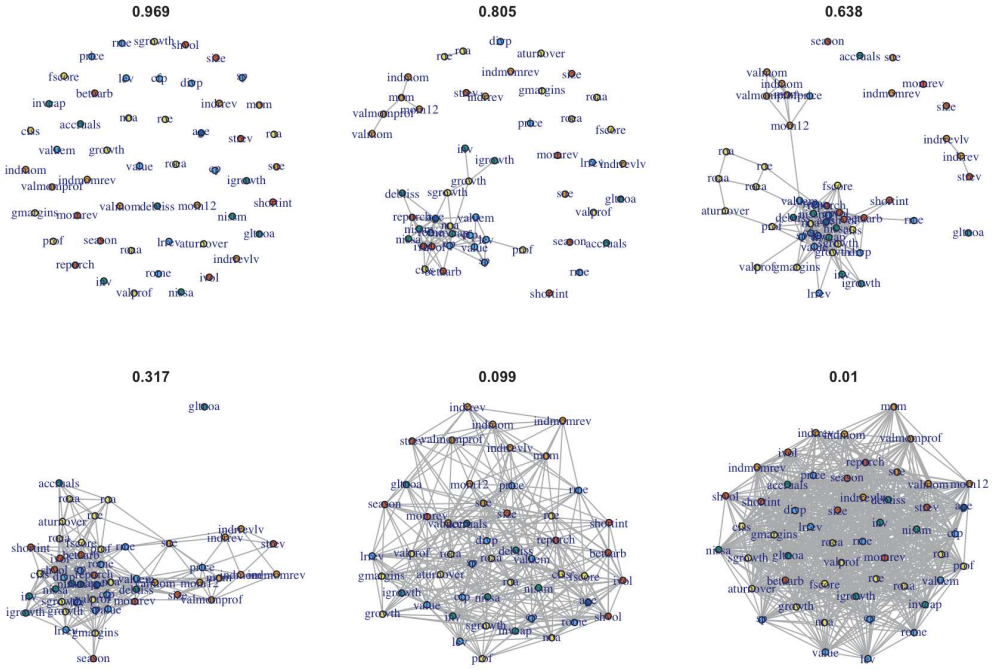
We seek to generate N subsample s_n with $\boldsymbol{f}$ without replacements with size $b = t/T$ so that $B = (s_1, \dots, s_N)$, and $0 < b < 1$. Next, we define a regularization parameter grid $\mathcal{G}_\Lambda = (\Lambda_1, \dots, \Lambda_K)$, setting $\Lambda_k = 1/\tau_k$. For each s_n and Λ_k, we define $A_{i,j}(s_n, \Lambda_k)$, where $\boldsymbol{A}(s_n, \Lambda_k)$ is the adjacent matrix resulting from the $\widehat{\boldsymbol{\Theta}}_{GLASSO}$ estimator described as (3) with the s_n as the observed data, and Λ_k as the regularization parameter. Next, we set $\psi_{i,j}(\Lambda_k) = P\big(A_{i,j}(s_1, \dots, s_N|\Lambda_k) = 1\big)$ and estimate it with $\hat{\psi}_{i,j}(\Lambda) = \frac{1}{N}\sum_{n=1}^{N} A_{i,j}(s_n, \Lambda)$. We then proceed to write $\varphi_{i,j}(\Lambda) = 2\Big(\psi_{i,j}(\Lambda)\Big)\Big(1 - \psi_{i,j}(\Lambda)\Big)$, and estimate it with $\hat{\varphi}_{i,j}(\Lambda) = 2\Big(\hat{\psi}_{i,j}(\Lambda)\Big)\Big(1 - \hat{\psi}_{i,j}(\Lambda)\Big)$. The $\varphi_{i,j}(\Lambda)$ relation measures edge instability, represented as $A_{i,j}$ across all subsamples. Thus, we compute the total instability by averaging $\hat{\varphi}_{i,j}(\Lambda)$ across all edges, resulting in $\hat{\varphi}(\Lambda) = \sum_{i=1}^{P}\sum_{j=1}^{P}\hat{\varphi}_{i,j}(\Lambda)$. Since a large Λ tends to generate a dense graph with low $\hat{\varphi}(\Lambda)$, and our interest lies on a sparse result for $\widehat{\boldsymbol{\Theta}}_{GLASSO}$, we monotonize $\hat{\varphi}(\Lambda)$ by $\bar{\varphi}(\Lambda) = sup_{0\le t\le\Lambda}\hat{\varphi}(t)$ using $\hat{\tau}_{\mathrm{StARS}} = sup\{\Lambda : \bar{\varphi}(\Lambda) \le \beta\}$ to estimate τ_{StARS}. Finally, following the recommendation made by Liu et al. (2010), we set $N = 100$, $b = 0.90$, $K = 100$, and $\beta = 0.05$ for the StARS procedure.

Table A1: Risk factor descriptive statistics

Description	Ret.	S.R.	Category	Reference	Code
Excess Market Return	0.064	0.418	Value vs Growth	(Sharpe 1964)	Rme
Size	-0.027	-0.170	Trading Frictions	(Fama and French 1993)	Size
Value	0.039	0.242	Value vs Growth	(Fama and French 1993)	value
Gross Profitability	0.024	0.151	Profitability	(Novy-Marx 2013)	prof
Value-Profitability	0.131	0.845	Profitability	(Novy-Marx 2013)	valprof
Piotroski's F-score	0.038	0.221	Profitability	(Piotroski 2000)	fscore
Debt Issuance	0.016	0.098	Investment	(Spiess and Affleck-Graves 1999)	debtiss
Share Repurchases	0.029	0.172	Trading Frictions	(Ikenberry, Lakonishok, and Vermaelen 1995)	repurch
Share Issuance (annual)	-0.100	-0.563	Investment	(Pontiff and Woodgate 2008)	nissa
Accruals	-0.029	-0.188	Investment	(Sloan 1996)	accruals
Asset Growth	-0.090	-0.556	Profitability	(Cooper, Gulen, and Schill 2008)	growth
Asset Turnover	0.036	0.246	Profitability	(Soliman 2008)	aturnover
Gross Margins	-0.011	-0.068	Profitability	(Novy-Marx 2013)	gmargins
Dividend Yield	0.022	0.155	Value vs Growth	(Naranjo, Nimalendran, and Ryngaert 1998)	divp
Earnings/Price	0.052	0.313	Value vs Growth	(Basu 1977)	Ep
Cash Flow / Market Value of Equity	0.048	0.295	Value vs Growth	(Lakonishok, Shleifer, and Vishny 1994)	Cfp
Net Operating Assets	0.004	0.026	Profitability	(Hirshleifer et al. 2004)	Noa
Investment	-0.098	-0.594	Investment	(Chen, Novy-Marx, and Zhang 2011)	Inv
Investment-to-Capital	-0.053	-0.302	Investment	(Xing 2008)	invcap
Invetment Growth	-0.107	-0.630	Investment	(Xing 2008)	igrowth
Sales Growth	-0.071	-0.427	Profitability	(Lakonishok, Shleifer, and Vishny 1994)	sgrowth
Leverage	0.033	0.195	Value vs Growth	(Bhandari 1988)	Lev
Return on Assets (annual)	0.010	0.067	Profitability	(Chen, Novy-Marx, and Zhang 2011)	roaa
Return on Equity (annual)	0.038	0.235	Profitability	(Haugen and Baker 1996)	roea
Sales-to-Price	0.066	0.394	Value vs Growth	(Barbee Jr, Mukherji, and Raines 1996)	Sp
Growth in LTNOA	-0.017	-0.132	Investment	(Fairfield, Whisenant, and Yohn 2003)	gltnoa
Momentum (6m)	0.004	0.023	Momentum	(Jegadeesh and Titman 1993)	mom
Industry Momentum	0.042	0.244	Momentum	(Moskowitz and Grinblatt 1999)	indmom
Value-Momentum	0.029	0.178	Momentum	(Novy-Marx 2013)	valmom
Value-Momentum-Profitability	0.051	0.308	Momentum	(Novy-Marx 2013)	valmomprof
Short Interest	-0.006	-0.048	Trading Frictions	(Dechow, Kothari, and Watts 1998)	shortint
Momentum (1y)	0.043	0.259	Momentum	(Jegadeesh and Titman 1993)	mom12
Momentum-Reversal	-0.074	-0.451	Trading Frictions	(Jegadeesh and Titman 1993)	momrev
Long-term Reversals	-0.058	-0.374	Value vs Growth	(De Bondt and Thaler 1985)	lrrev
Value (monthly)	0.027	0.161	Value vs Growth	(Asness and Frazzini 2013)	valuem
Share Issuance (monthly)	-0.096	-0.533	Investment	(Pontiff and Woodgate 2008)	nissm
PEAD (SUE)	0.072	0.481	Momentum	(Foster, Olsen, and Shevlin 1984)	sue
Return on Book Equity	0.084	0.541	Profitability	(Chen, Novy-Marx, and Zhang 2011)	roe
Return on Market Equity	0.073	0.441	Value vs Growth	(Chen, Novy-Marx, and Zhang 2011)	rome
Return on Assets	0.047	0.316	Profitability	(Chen, Novy-Marx, and Zhang 2011)	roa
Short-term Reversal	-0.069	-0.413	Trading Frictions	(Jegadeesh 1990)	strev
Idiosyncratic Volatility	-0.054	-0.323	Trading Frictions	(Ang, Chen, and Xing 2006)	ivol
Beta Arbitrage	-0.034	-0.207	Trading Frictions	(Cooper, Gulen, and Schill 2008)	betaarb
Seasonality	0.076	0.482	Trading Frictions	(Heston and Sadka 2008)	season
Industry Relative Reversals	-0.133	-0.808	Momentum	(Da, Liu, and Schaumburg 2014)	indrrev
Industry Relative Reversals (Low Vol.)	-0.225	-1.542	Momentum	(Da, Liu, and Schaumburg 2014)	indrrevlv
Industry Momentum-Reversal	0.143	0.885	Momentum	(Moskowitz and Grinblatt 1999)	indmomrev
Composite Issuance	-0.086	-0.543	Profitability	(Daniel and Titman 2006)	ciss
Price	-0.015	-0.092	Value vs Growth	(Blume and Husic 1973)	price
Firm Age	0.013	0.074	Intangibles	(Barry and Brown 1984)	age
Share Volume	-0.037	-0.222	Trading Frictions	(Datar, Naik, and Radcliffe 1998)	shvol

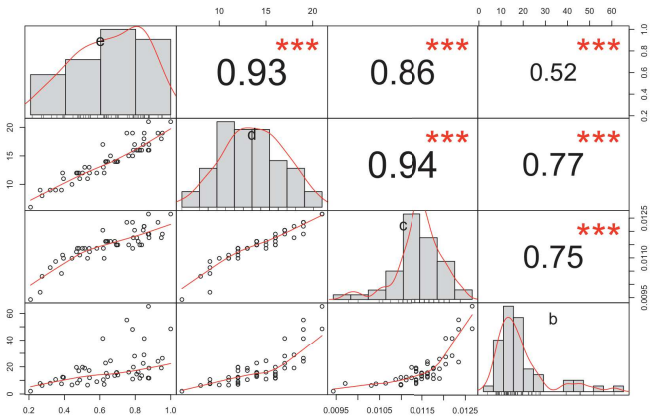
Note: The table displays the descriptive statistics for our risk factors dataset compiled by Kozak, Nagel, and Santosh (2020) with monthly data ranging from January 1981 to December 2016. For each risk factor, the table includes annualized average excess returns, annualized Sharpe ratios, a priori category classification, literature reference and code name.

Figure A1: Graphic representation of our estimated risk factor network $(\widehat{\mathcal{G}}_u)$ for a selected parameter regularization (τ)



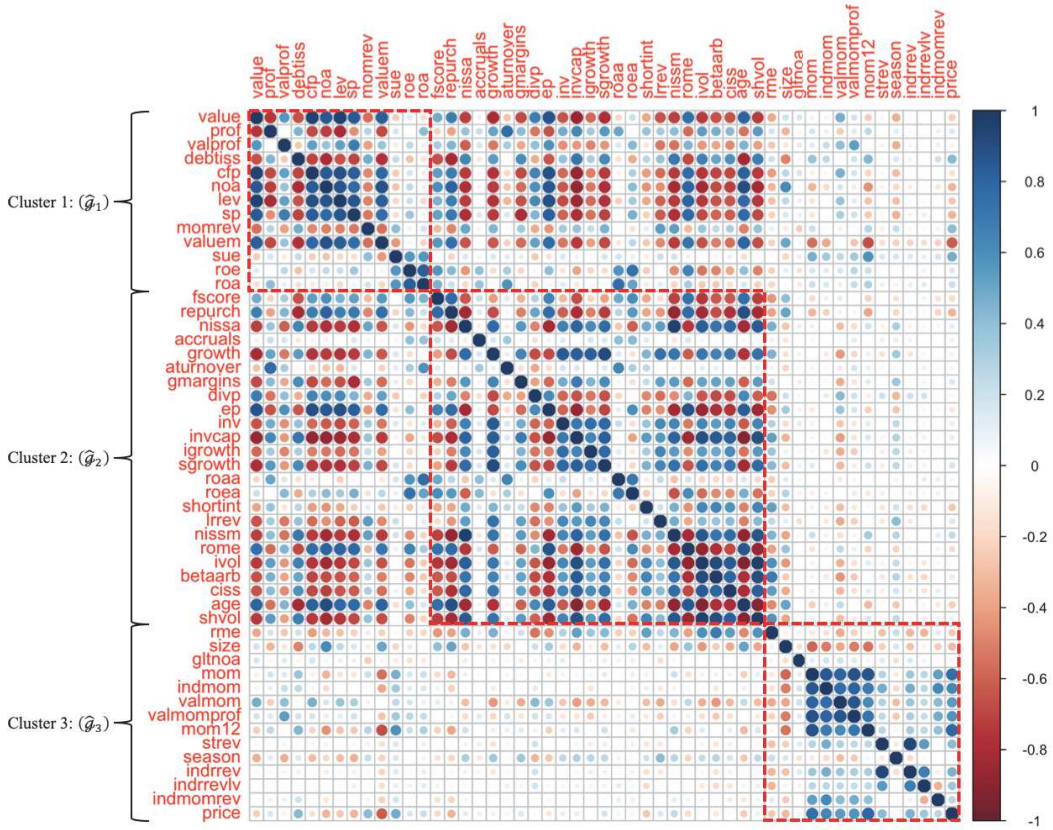
Note: The figure displays graphic representation of our estimated risk factor network $(\hat{\mathcal{G}}_u)$. As described in section 2.1., we estimate the risk factor joint distribution precision matrix by graph lasso in order to obtain $\widehat{\boldsymbol{\Theta}}_{GLASSO}$.. As described in section 2.2., we compute our estimated risk factor network by $\hat{\mathcal{G}}_u = \mathcal{G}_u\left(\boldsymbol{\mathcal{V}} = \boldsymbol{f}, \boldsymbol{\mathcal{E}}\left(\widehat{\boldsymbol{\Theta}}_{GLASSO}\right)\right)$. In this picture, each node represents a risk factor and the edge between than indicates $\hat{\theta}_{GLASSO,i,j} \neq 0$, which indicates conditional dependence among the risk factors i and j given all others risk factors. The node color represents the risk factor category (blue for Value vs Growth; green for Investment; yellow for Profitability; orange for Momentum; dark blue for Intangibles; and red for Trading Frictions). We select six different regularization parameters (τ). As τ increases, the estimated conditional dependence among risk factors also decreases.

Figure A2: Risk factor`s global centrality measures histograms and correlations



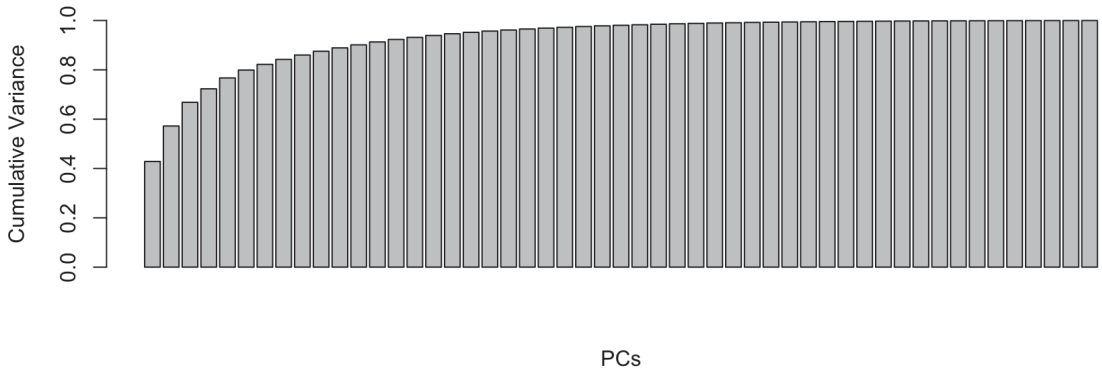
Note: This figure displays the histograms and correlations from different centrality measures of our estimated risk factor network described by $\hat{\mathcal{G}}_u$ in Figure 2. Letters e, d, c and b refers to Eigenvector, Degree, Closeness and Betweenness centrality measure respectively.

Figure A3: Risk factor`s correlations tidy by the modularity clusters order



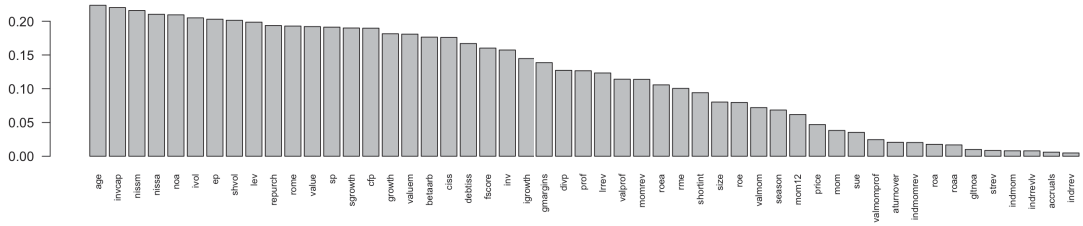
Note: This figure displays the risk factors correlations matrix tidy according to the modularity clusters order. The red dotted square area defines within cluster risk factor correlations matrix. Each risk factors cluster group is highlighted on right label.

Figure A4: Principal component cumulative risk factor variance



Note: This figure displays the cumulative risk factors variance explained by its principal components (PCs).

Figure A5: First principal component risk factor loadings absolute values



Note: This figure displays the absolute values from the first principal component risk factor loadings order by value.

Table A2: Modified Diebold-Mariano test results for the CRSP dataset

	Model 1		Global		Cluster			Classic				PCs				PC1 Loadings				Elastic
			E D and C	B	E and C	D	B	FF3	NM4	C4	P5	PC1	PC2	PC3	PC4	Model1	Model2	Model3	Model4	Net FM
Model 2	Global	E, D and C		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.4648
		B			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Cluster	E and C				0.0001	0.1873	0.0734	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000
		D					0.0007	0.0138	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		B						0.1522	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000	0.0000	0.0000	0.0068	0.0000
	Classic	FF3							0.0000	0.0000	0.0000	0.0000	0.0000	0.9364	0.0000	0.0000	0.0000	0.0000	0.6749	0.0000
		NM4								0.0048	0.0000	0.0000	0.0000	0.0000	0.3342	0.0000	0.0000	0.0000	0.0002	0.0000
		C4									0.0000	0.0000	0.0000	0.0000	0.5765	0.0000	0.0000	0.0000	0.0000	0.0000
		P5										0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	PCs	PC1											0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		PC2												0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		PC3													0.0000	0.0000	0.0000	0.0000	0.3281	0.0000
		PC4														0.0000	0.0000	0.0000	0.0000	0.0000
	PC1 Loadings	Model1															0.0000	0.0000	0.0000	0.0365
		Model2																0.0000	0.0000	0.0000
		Model3																	0.0000	0.0000
		Model4																		0.0000
	Elastic Net FM																			

Note: The table displays the p-value from the modified Diebold-Mariano test for predictive accuracy among two models proposed by D. Harvey, Leybourne, and Newbold (1997)*. The null hypothesis is given by* $H_0: h(e_{1,t}) = h(e_{2,t})$ *against the alternative hypothesis* $H_0: h(e_{1,t}) \neq h(e_{2,t})$ *(model 1 is less accurate than model 2), where* $e_{i,t}$ *is the out-of-sample one-step-ahead forecast error described in Section 2.6 from model* i *and* h *is a quadratic function.*

Table A3: Diebold-Mariano test results for the CRSP dataset without small caps

Model 1		Global		Cluster			Classic				PCs				PC1 Loadings				Elastic
		E D and C	B	E and C	D	B	FF3	NM4	C4	P5	PC1	PC2	PC3	PC4	Model1	Model2	Model3	Model4	Net FM
Global	E, D and C		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000
	B			0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Cluster	E and C				0.0168	0.2237	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0207	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	D					0.4684	0.0714	0.0000	0.0000	0.0000	0.0000	0.0000	0.0031	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	B						0.0021	0.0000	0.0000	0.0000	0.0000	0.0000	0.0029	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Classic	FF3							0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	NM4								0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
	C4									0.0000	0.0000	0.0000	0.0000	0.4080	0.0000	0.0000	0.0000	0.0000	0.0000
	P5										0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PCs	PC1											0.0000	0.0000	0.0000	0.7863	0.0000	0.0000	0.0000	0.0000
	PC2												0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0157
	PC3													0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	PC4														0.0000	0.0000	0.0000	0.0000	0.0000
PC1 Loadings	Model1															0.0000	0.0000	0.0000	0.0000
	Model2																0.0000	0.0000	0.0001
	Model3																	0.0000	0.0000
	Model4																		0.0000
Elastic Net FM																			

Model 2

Note: The table displays the p-value from the modified Diebold-Mariano test for predictive accuracy among two models proposed by D. Harvey, Leybourne, and Newbold (1997)*. The null hypothesis is given by* $H_0: h(e_{1,t}) = h(e_{2,t})$ *against the alternative hypothesis* $H_0: h(e_{1,t}) \neq h(e_{2,t})$*, where* $e_{i,t}$ *is the out-of-sample one-step-ahead forecast error described in Section 2.6 from model* i *and* h *is a quadratic function.*

References

Abu-Mostafa, Y S, M Magdon-Ismail, and H T Lin. 2012. "Learning from Data Vol. 4: AMLBook New York." NY, USA.

Allen, Franklin, and Ana Babus. 2009. "Networks in Finance." The network challenge: strategy, profit, and risk in an interlinked world 367.

Ang, Andrew, Joseph Chen, and Yuhang Xing. 2006. "Downside Risk." The review of financial studies 19(4): 1191–1239.

Asness, Clifford, and Andrea Frazzini. 2013. "The Devil in HML's Details." The Journal of Portfolio Management 39(4): 49–68.

Barbee Jr, William C, Sandip Mukherji, and Gary A Raines. 1996. "Do Sales--Price and Debt--Equity Explain Stock Returns Better than Book--Market and Firm Size?" Financial Analysts Journal 52(2): 56–60.

Barry, Christopher B, and Stephen J Brown. 1984. "Differential Information and the Small Firm Effect." Journal of financial economics 13(2): 283–94.

Basu, Sanjoy. 1977. "Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis." The journal of Finance 32(3): 663–82.

Bavelas, Alex. 1950. "Communication Patterns in Task-Oriented Groups." The journal of the acoustical society of America 22(6): 725–30.

Bhandari, Laxmi Chand. 1988. "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence." The journal of finance 43(2): 507–28.

Bloch, Francis, Matthew O Jackson, and Pietro Tebaldi. 2019. "Centrality Measures in Networks." Available at SSRN 2749124.

Blume, Marshall E, and Frank Husic. 1973. "Price, Beta, and Exchange Listing." The Journal of Finance 28(2): 283–99.

De Bondt, Werner F M, and Richard Thaler. 1985. "Does the Stock Market Overreact?" The Journal of finance 40(3): 793–805.

Borgatti, Stephen P. 2005. "Centrality and Network Flow." Social networks 27(1): 55–71.

Brito, Diego, Marcelo C Medeiros, and Ruy Ribeiro. 2018. "Forecasting Large Realized Covariance Matrices: The Benefits of Factor Models and Shrinkage." Available at SSRN 3163668.

Calvo-Armengol, Antoni, and Matthew O Jackson. 2004. "The Effects of Social Networks on Employment and Inequality." American economic review 94(3): 426–54.

Campbell, John Y et al. 1997. The Econometrics of Financial Markets. princeton University press.

Campbell, John Y, and Samuel B Thompson. 2008. "Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?" The Review of Financial Studies 21(4): 1509–31.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." Journal of Finance: 57–82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x.

Cavalcante Filho, Elias, Fernando Chague, Rodrigo De-Losso, and Bruno Giovannetti. 2020. "US Risk Premia under Emerging Markets Constraints." Available at SSRN 3600947.

Chen, Long, Robert Novy-Marx, and Lu Zhang. 2011. "An Alternative Three-Factor Model."

Available at SSRN 1418117.

Clauset, Aaron, Mark E J Newman, and Cristopher Moore. 2004. "Finding Community Structure in Very Large Networks." Physical review E 70(6): 66111.

Cochrane, John H. 2009. Asset Pricing: Revised Edition. Princeton university press.

———. 2011. "Presidential Address: Discount Rates." The Journal of finance 66(4): 1047–1108.

Cohen, Lauren, Andrea Frazzini, and Christopher Malloy. 2008. "The Small World of Investing: Board Connections and Mutual Fund Returns." Journal of Political Economy 116(5): 951–79.

Cooper, Michael J, Huseyin Gulen, and Michael J Schill. 2008. "Asset Growth and the Cross-Section of Stock Returns." the Journal of Finance 63(4): 1609–51.

Da, Zhi, Qianqiu Liu, and Ernst Schaumburg. 2014. "A Closer Look at the Short-Term Return Reversal." Management science 60(3): 658–74.

Daniel, Kent, and Sheridan Titman. 2006. "Market Reactions to Tangible and Intangible Information." The Journal of Finance 61(4): 1605–43.

Datar, Vinay T, Narayan Y Naik, and Robert Radcliffe. 1998. "Liquidity and Stock Returns: An Alternative Test." Journal of Financial Markets 1(2): 203–19.

Dechow, Patricia M, Sagar P Kothari, and Ross L Watts. 1998. "The Relation between Earnings and Cash Flows." Journal of accounting and Economics 25(2): 133–68.

Elliott, Matthew, Benjamin Golub, and Matthew O Jackson. 2014. "Financial Networks and Contagion." American Economic Review 104(10): 3115–53.

Engle, Robert F, Neil Shephard, and Kevin Sheppard. 2008. "Fitting Vast Dimensional Time-Varying Covariance Models."

Fairfield, Patricia M, J Scott Whisenant, and Teri Lombardi Yohn. 2003. "Accrued Earnings and Growth: Implications for Future Profitability and Market Mispricing." The accounting review 78(1): 353–71.

Fama, Eugene F, and Kenneth R French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." Journal of financial economics 33(1): 3–56. https://doi.org/10.1016/0304-405X(93)90023-5.

Fama, Eugene F, and James D MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." Journal of political economy 81(3): 607–36.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu. 2020. "Taming the Factor Zoo: A Test of New Factors." The Journal of Finance 75(3): 1327–70.

Foster, George, Chris Olsen, and Terry Shevlin. 1984. "Earnings Releases, Anomalies, and the Behavior of Security Returns." Accounting Review: 574–603.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber. 2020. "Dissecting Characteristics Nonparametrically." The Review of Financial Studies 33(5): 2326–77.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." Biostatistics 9(3): 432–41.

Gofman, Michael. 2017. "Efficiency and Stability of a Financial Architecture with Too-Interconnected-to-Fail Institutions." Journal of Financial Economics 124(1): 113–46.

Golub, Benjamin, and Matthew O Jackson. 2010. "Naive Learning in Social Networks and the Wisdom of Crowds." American Economic Journal: Microeconomics 2(1): 112–49.

Goyal, Amit. 2012. “Empirical Cross-Sectional Asset Pricing: A Survey.” Financial Markets and Portfolio Management 26(1): 3–38.

Green, Jeremiah, John R M Hand, and X Frank Zhang. 2017. “The Characteristics That Provide Independent Information about Average Us Monthly Stock Returns.” The Review of Financial Studies 30(12): 4389–4436.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu. 2019. “Autoencoder Asset Pricing Models.”

Harvey, Campbell R, and Yan Liu. 2019. “Lucky Factors.” Available at SSRN 2528780.

Harvey, Campbell R, Yan Liu, and Heqing Zhu. 2016. “… and the Cross-Section of Expected Returns.” The Review of Financial Studies 29(1): 5–68.

Harvey, David, Stephen Leybourne, and Paul Newbold. 1997. “Testing the Equality of Prediction Mean Squared Errors.” International Journal of forecasting 13(2): 281–91.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC press.

Haugen, Robert A, and Nardin L Baker. 1996. “Commonality in the Determinants of Expected Stock Returns.” Journal of Financial Economics 41(3): 401–39.

Herskovic, Bernard. 2018. “Networks in Production: Asset Pricing Implications.” The Journal of Finance 73(4): 1785–1818.

Heston, Steven L, and Ronnie Sadka. 2008. “Seasonality in the Cross-Section of Stock Returns.” Journal of Financial Economics 87(2): 418–45.

Hirshleifer, David, Kewei Hou, Siew Hong Teoh, and Yinglei Zhang. 2004. “Do Investors Overvalue Firms with Bloated Balance Sheets?” Journal of Accounting and Economics 38: 297–331.

Hochberg, Yael V, Alexander Ljungqvist, and Yang Lu. 2007. “Whom You Know Matters: Venture Capital Networks and Investment Performance.” The Journal of Finance 62(1): 251–301.

Ikenberry, David, Josef Lakonishok, and Theo Vermaelen. 1995. “Market Underreaction to Open Market Share Repurchases.” Journal of financial economics 39(2–3): 181–208.

Jegadeesh, Narasimhan. 1990. “Evidence of Predictable Behavior of Security Returns.” The Journal of finance 45(3): 881–98.

Jegadeesh, Narasimhan, and Sheridan Titman. 1993. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” The Journal of finance 48(1): 65–91.

Kelly, Bryan T, Seth Pruitt, and Yinan Su. 2019. “Characteristics Are Covariances: A Unified Model of Risk and Return.” Journal of Financial Economics 134(3): 501–24.

Khan, Bisma S, and Muaz A Niazi. 2017. “Network Community Detection: A Review and Visual Survey.” arXiv preprint arXiv:1708.00977.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh. 2020. “Shrinking the Cross-Section.” Journal of Financial Economics 135(2): 271–92.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny. 1994. “Contrarian Investment, Extrapolation, and Risk.” The journal of finance 49(5): 1541–78.

Lancichinetti, Andrea, and Santo Fortunato. 2009. “Community Detection Algorithms: A

Comparative Analysis." Physical review E 80(5): 56117.

Ledoit, Olivier, and Michael Wolf. 2004. "Honey, I Shrunk the Sample Covariance Matrix." The Journal of Portfolio Management 30(4): 110–19.

Liu, Han, Kathryn Roeder, and Larry Wasserman. 2010. "Stability Approach to Regularization Selection (Stars) for High Dimensional Graphical Models." In Advances in Neural Information Processing Systems, , 1432–40.

Meinshausen, Nicolai, Peter Bühlmann, and others. 2006. "High-Dimensional Graphs and Variable Selection with the Lasso." The annals of statistics 34(3): 1436–62.

Moskowitz, Tobias J, and Mark Grinblatt. 1999. "Do Industries Explain Momentum?" The Journal of finance 54(4): 1249–90.

Naranjo, Andy, M Nimalendran, and Mike Ryngaert. 1998. "Stock Returns, Dividend Yields, and Taxes." The Journal of Finance 53(6): 2029–57.

Newman, Mark E J, and Michelle Girvan. 2004. "Finding and Evaluating Community Structure in Networks." Physical review E 69(2): 26113.

Novy-Marx, Robert. 2013. "The Other Side of Value: The Gross Profitability Premium." Journal of Financial Economics 108(1): 1–28.

Piotroski, Joseph D. 2000. "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers." Journal of Accounting Research: 1–41.

Pontiff, Jeffrey, and Artemiza Woodgate. 2008. "Share Issuance and Cross-Sectional Returns." The Journal of Finance 63(2): 921–45.

De Prado, Marcos Lopez. 2018. Advances in Financial Machine Learning. John Wiley & Sons.

Shanken, Jay. 1992. "On the Estimation of Beta-Pricing Models." The review of financial studies 5(1): 1–33.

Sharpe, William F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." The journal of finance 19(3): 425–42.

Sloan, Richard G. 1996. "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows about Future Earnings?" Accounting review: 289–315.

Soliman, Mark T. 2008. "The Use of DuPont Analysis by Market Participants." The Accounting Review 83(3): 823–53.

Spiess, D Katherine, and John Affleck-Graves. 1999. "The Long-Run Performance of Stock Returns Following Debt Offerings." Journal of Financial Economics 54(1): 45–73.

Stephen, Ross. 1976. "The Arbitrage Theory of Capital Asset Pricing." Journal of Economic Theory 13(3): 341–60.

Wasserman, Larry, and Kathryn Roeder. 2009. "High Dimensional Variable Selection." Annals of statistics 37(5A): 2178.

Xing, Yuhang. 2008. "Interpreting the Value Effect through the Q-Theory: An Empirical Investigation." The Review of Financial Studies 21(4): 1767–95.

Yan, Xuemin, and Lingling Zheng. 2017. "Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach." The Review of Financial Studies 30(4): 1382–1423.